COMPUTING ASPECTS OF PROBLEMS IN NON-LINEAR PREDICTION AND FILTERING

by

I. G. CUMMING

A thesis submitted for the degree of Ph.D. in Engineering

Centre for Computing and Automation,
Imperial College,
University of London,
London, S.W.7.

MAY 1967

ABSTRACT

This thesis discusses some of the computational problems arising in the application of modern stochastic control theory. We deal with continuous time systems where two typical problems are the prediction of the future statistical behaviour of systems and the synthesis of systems designed from stochastic theory such as filters. In each case, computational problems occur if the system is non-linear or non-Gaussian.

The non-linear prediction problem involves solving a parabolic partial differential equation, the Fokker-Planck equation, and we discuss two numerical methods of solving this equation. Finding that these methods can only handle a restricted class of low-dimensional systems, we study Monte Carlo methods in the hopes of finding a more general solution procedure. We find these to be successful if we allow for accuracy limitations, and find that the theory of the Fokker-Planck equation can be extended to include the Monte Carlo solution of a wider class of parabolic equations than previous methods would accommodate.

Monte Carlo methods involve the simulation of a stochastic system on a computer, and as the problem of synthesising a given system is the same as simulating it on a computer, the rest of the thesis centres around the theoretical and practical aspects of simulation techniques. We find in each case that the system we must simulate is a continuous Markov (diffusion) process, and that diffusion processes have properties which distinguish them from any process which can be constructed on a computer or in the physical world (we call these physical processes).

Thus we discuss the statistical equivalence of physical and diffusion processes, and show how and under what conditions a physical process can be chosen to approximate a diffusion process, and vice versa. In this way, we clarify a controversy on the interpretation of limiting forms of physical processes, and give an example which confirms the accuracy of the approximation and illustrates that diffusion approximations can provide a useful method of analysing the transient statistics of physical processes. On the practical side of the simulation problem, we discuss the choice of noise source and its proper characterisation on the analogue computer, and the convergence rates and efficiencies of discrete simulation formulae on the digital computer.

Acknowledgements

This thesis has resulted from research carried out in the Department of Electrical Engineering*, Imperial College, from October 1963 to November 1966. The author is deeply indebted to his supervisor, J. H. Westcott, Professor of Control Systems, for providing the facilities for research, and considerable personal encouragement during this period. Discussions with colleagues were helpful in clarifying the aims and certain specific points of the project, and the author would particularly like to thank J. M. C. Clark, G. S. Marliss and D. Q. Mayne in this regard.

The financial support of the Ministry of Education, Province of Quebec (1963/64), and the National Research Council of Canada (NATO Scholarship 1964/65 and 1965/66) is gratefully acknowledged.

Finally, the author is very thankful for the ever-willing help of Miss M. Stocker (typing), Miss M. Lancaster (diagrams) and Miss L. Hawkins (duplicating).

---

* The Control Systems Group is now in the Centre for Computing and Automation

CONTENTS PAGE

List of Figures

Glossary of Principal Symbols

$A(t)$, $A^*(t)$	characteristic matrices of physical noise process (4.1.21, 30)
$a(x,t)$	second incremental moment of diffusion process (1.2.2)
a	coefficient of linear system (6.1.4)
$B(t)$	noise scaling matrix $B\,B^T = A + A^*$ (4.1.39)
$b(x,t)$	first incremental moment of diffusion process (1.2.2)
b_i	coefficient of linear system (4.3.1)
b	coefficient of linear system (6.1.4)
C	noise scaling matrix (4.4.6)
$c(t)$	forcing function of linear system (4.3.1)
c	piecewise constant random coefficient (6.1.6), (6.2.13)
$2D(t)$	intensity matrix of noise process = $A + A^*$
$d\cdot$	ordinary differential or Ito stochastic differential operator
$\bar{d}\cdot$	Stratonovich stochastic differential operator
$E[\cdot]$	expectation (ensemble average)
$E[\cdot \mid \cdot]$	conditional expectation
$F(x,t)$	n x m noise coefficient matrix
$f(x,t)$	non-random part of system dynamics n vector
f_u	effective upper frequency of physical noise
$G(X,t)$	n x m coefficient matrix
$g(X,t)$	non-random part of system dynamics n vector
Hz	cycles per second
$H(t,j\omega)$	linear system frequency function (5.1.2)
$h(t,u)$	non-stationary impulse response of linear system (5.1.1)
$h(s-u)$	stationary impulse response of linear system (5.2.4)
I	the unit or identity matrix
$J(x,t)$	flux vector in Fokker-Planck equation (1.2.3)
j	the imaginary constant $(-1)^{\frac{1}{2}}$
K	a variance parameter (5.2.5)
L	number of bits in one cycle of PRBS
m	dimension of noise process
$N(a,b)$	normal or Gaussian density with mean a and variance b
n	dimension of system

$O(\cdot)$	of order equal to
$o(\cdot)$	of order greater than
$P(x,t)$	probability density function of $x(t)$
$P(x,t \mid y,s)$	conditional probability of $x(t)$ given $y(s)$
$P(t)$	output state of physical non-linear filter (5.2.23)
$p(t)$	output state of ideal non-linear filter (5.2.17)
$Q_{kl}(x,t)$	part of bias term relating physical and diffusion processes n-vector (2.2.2a)
R	the entire space domain of $x(t)$
$R(t,\tau)$	m x m matrix correlation function (4.1.5)
$S(t,\omega)$	non-stationary spectral density (5.1.2)
$S(t)$	discrete model of integral of PRBS
$s(t)$	diffusion model of $S(t)$
s	backward time (1.2.4)
T	filter time constant
t	time
V	variance of random numbers (6.1.3)
v	switching rate of random telegraph wave (5.2.17)
$w(t)$	m vector independent unit parameter Wiener process
$\dot{w}(t)$	m vector independent unit parameter white noise
$X(t)$	n vector physical process described by an o.d.e.
$x(t)$	n vector diffusion process described by an s.d.e.
$\bar{x}(t)$	noisy measurement (5.2.13)
$Y(t)$	integral of $y(t)$ (5.2.36)
$y(t)$	n vector physical noise process
$z(t)$	high bandwidth noise source (5.1.1)
$\bar{z}(t)$	measurement state in filtering problem (5.2.15)
α	cross-coupling coefficient in asymmetrical noise source (5.1.14)
α	coefficient of linear equation (6.2.4)
β	noise scaling parameter (5.2.13)
$\delta(\tau)$	Dirac (symmetrical) delta function
$\delta_A(\tau)$	modified delta function (4.2.16)
$\delta\cdot$	forward difference increment operator

Δ	PRBS bit interval
Δt	discrete time interval on digital computer
ΔX	increment of $X(t)$ over time Δt
σ	standard deviation parameter
θ	parameter in definition of stochastic integral (4.2.21)
τ	time shift
τ_{cor}	memory or correlation time of physical noise $y(t)$ (4.1.5)
τ_{rel}	response or relaxation time of system (2.3.3)
τ'_{rel}	memory time of system (4.1.44)
ω	angular frequency
$(\cdot)^T$	transpose of matrix or vector argument
$\sum_{i}^{n}$ $\sum_{i,j}^{n}$	sums with an implied lower limit of 1
i j k l	as subscripts denote elements of a vector or matrix
t x	as subscripts denote partial derivatives
$(\cdot,\cdot)$	open interval
$[\cdot,\cdot]$	closed interval
= ISP =	is proportional to
FP	Fokker-Planck
l.h.s.	left hand side (also r.h.s.)
PRBS	a maximum length pseudo random binary sequence
RK	Runge-Kutta
o.d.e.	ordinary differential equation
s.d.e.	stochastic differential equation

Symbols local to Section 2.3

c_i	coefficients of non-linear example (2.3.2)
D_k E_k G_k	constants used in numerical solution procedure (2.3.24)
E'_k G'_k	constants used in dual solution procedure (2.3.39)
$h, h_1(k), h_2(k)$	finite difference model parameters (2.3.9)
K	number of unknown points in space grid (2.3.8)
k_{eq}	Booton's equivalent gain (2.3.12)
m_x	mean value of $x(t)$
n_x	standard deviation of $x(t)$
$P_j(k)$	discrete solution point at $t = j\Delta t$ and $x = x' + k\Delta x$
x'	left hand edge of space grid
V_x	variance of $x(t)$
α	semi-permeable boundary parameter (2.3.27)

Symbols local to Section 2.4

$\underline{a}(x,t)$	second incremental moment of standardised process $y(t)$ (2.4.39)
$\underline{b}(y,t)$	first incremental moment of standardised process $y(t)$ (2.4.39)
c_r	rescaled Hermite expansion coefficient (2.4.28)
Error (n)	an error function (2.4.15)
$f(x)$	ar arbitrary function (2.4.20)
$G(x)$	normalised Gaussian density (2.4.3)
$H_r(x)$	r:th Hermite polynomial (2.4.1)
h_r	Hermite quadrature weights (2.4.20)
k_r	r:th Hermite expansion coefficient (2.4.6)
$m(t)$	mean of $x(t)$
$m_i(x)$	i:th moment of $x(t)$
$P_n(x,t)$	reconstructed density function using n terms of Hermite expansion (2.4.16)
$Q(y,t)$	density of standardised variable $y(t)$
$q_i(t)$	functions of m v σ (2.4.40)
$v(t)$	mean square of $x(t)$
x_r	roots of the Hermite polynomials (2.4.20)
y	standardised variable x (2.4.17)
$\sigma(t)$	standard deviation of $x(t)$

Symbols local to Chapter 3 and Section 6.3

c	specific heat of material (3.2.62)
D	density of particles or trajectories in the simulation (3.2.50)
$G(x)$	an arbitrary function (3.2.60)
$g(t)$	a solution related to the parabolic equation (3.2.60)
$\underline{g}(t)$	an estimate of $g(t)$ (3.2.61)
$g(x)$	non-negative weighting function (6.3.3)
$h_r(x)$	a polynomial of degree r (6.3.3)
$\bar{K}_i$	various bounded constants in positivity proof
^oK	degrees Kelvin (3.2.62)
K_i	diffusivity of material in direction x_i (3.2.62)
k_i	thermal conductivity of material in direction x_i (3.2.62)
k_r	orthogonal polynomial expansion coefficients (6.3.3)
$\underline{L}(\cdot)$	a general n dimension elliptic operator (3.2.2)
$L(\cdot)$	an elliptic operator of the FP type (3.2.7)
N	the original number of trajectories in the simulation
N'	the net number of trajectories in the simulation
n_i	the number of particles in cell i
$\bar{P}(x,t)$	an estimate of the density $P(x,t)$
p	density of material (3.2.62)
$q(t)$	proportion of total number of net particles in cell (3.2.51)
$\underline{q}(t)$	an estimate of $q(t)$ (3.2.53)
R'	a subset of x space (3.2.50)
T	final time in range of parabolic equation
$U(x,t)$	the dependent variable in a general linear parabolic equation (3.2.2)
u_j	chi-square variate of j:th experiment (3.2.101)
Var $[\cdot]$	variance (6.3.2)
$V(x,t)$	a proportional term in general parabolic equation (3.2.7)
$W(x,t)$	a constant term in general parabolic equation (3.2.7)

α	n x n matrix of coefficients of second order term of general parabolic equation (3.2.2)
α	flux reflection coefficient (3.2.78, 85)
β	space scaling factor (3.2.49)
β	flux magnification coefficient (3.2.81, 87)
β_i	coefficient of parabolic equation (3.2.22)
Δ	time increment (3.2.39)
ϵ	a small positive constant (3.2.15)
γ	coefficient of parabolic equation (3.2.22)
μ	a non-zero constant
ϕ	a normalisation quantity (3.2.33)
θ	flux attenuation coefficient (3.2.83)
$\Omega_\alpha(t)$	trajectory weighting factor (3.2.43)
Ω^-, Ω^+	flux of particles hitting boundary in time Δ (3.2.77)

CHAPTER 1

INTRODUCTION

1.1 Scope

This is an engineering thesis, and is, broadly speaking, an attempt to overcome some of the practical difficulties connected with the simulation, synthesis and computing problems of modern statistical control theory. Our aim is to develop methods, and when we do introduce new theoretical ideas, the tone of our presentation will lean towards an intuitive understanding of the ideas rather than mathematical conciseness or precision. The examples given are kept in their simplest form to ease presentation while adequately illustrating the point at hand. In most cases, more elaborate examples were worked out as computing exercises and in the text, we state any limitations found in the various methods presented.

The motivation of the project has come from the recent increase in interest in statistical control problems. Following the attention given to deterministic optimal control of the Bellman-Pontryagin-Calculus of variations approach in the years around 1960, it is clear that the emphasis of research is once again on statistical systems, for hardly a system is studied nowadays which does not incorporate random factors into its structure. The papers presented at the Third IFAC Congress (London, 1966) are a good indication of this trend, where the majority of the papers discussed some aspect of identification, prediction, filtering, stability or control of stochastic systems.

As in the field of deterministic optimal control, when analysing or constructing statistical systems, the researcher or engineer soon finds that large computational problems must be surmounted before results can be obtained or systems synthesised. The major emphasis of this thesis is placed on the simulation of statistical systems on analogue and digital computers, and we see that this covers a general solution to the prediction problem as well as to

the synthesis problem of constructing systems such as non-linear filters.

This thesis does not discuss the complex identification problem,and assumes that the systems under discussion have deterministic parameters which are known exactly, and random parameters whose probabilistic structure is known. We formulate our systems as differential equations where the random and the non-random parts are separated into different factors. We point out that there is clear distinction between random processes which exist in the physical world (i.e. processes which are physically realisable: we call them physical processes $X(t)$) and some which exist only in the mathematical world. In the latter category we consider continuous Markov processes (diffusion processes $x(t)$), which differ from physical processes in that the band-limited, finite-power noise of the physical process is replaced by an infinite power white noise. These diffusion processes are used almost exclusively in all branches of stochastic control theory formulated in continuous time, as Markov processes are much more convenient to manipulate mathematically than non-Markovian processes. This distinction between the processes which we deal with on a theoretical level and those which we meet in practice means that we must be concerned with the relation between these types of processes whenever theoretical results are to be implemented (or a theoretical problem is formulated from a physical situation). This thesis discusses the statistical aspects of this relation.

1.2 Preliminaries - the Fokker Planck Equation

This thesis begins with the prediction problem: given the past and present behaviour of the system, what is its future statistical behaviour? This problem only has a concise formulation if the system is Markovian, in which case the present probability density is all that is needed of the system's past and present behaviour to determine the future behaviour. Restricting our attention to continuous Markov

processes, we find that the future statistical behaviour of the system is given by the solution of a parabolic partial differential equation called the Fokker-Planck (FP) equation. We state the FP equation as follows:

The function $P(x,t)$, the probability density of the n-vector Markov process $x(t)$, is the solution of the FP equation

$$\frac{\partial P(x,t)}{\partial t} = - \sum_{i}^{n} \frac{\partial}{\partial x_i} \left[b_i(x,t) P(x,t) \right] + \frac{1}{2} \sum_{i,j}^{n} \frac{\partial^2}{\partial x_i \partial x_j} \left[a_{ij}(x,t) P(x,t) \right] \qquad (1.2.1)$$

satisfying the initial condition $P(x,t_o)$ subject to the following conditions:

(1) the first and second incremental moments of the Markov process $x(t)$ exist and are given by

$$b_i(x,t) = \underset{\delta t \downarrow 0}{\text{limit}} \frac{1}{\delta t} E\left[\delta x_i \,\middle|\, x,t \right],$$

$$\text{and} \quad a_{ij}(x,t) = \underset{\delta t \downarrow 0}{\text{limit}} \frac{1}{\delta t} E\left[\delta x_i \delta x_j \,\middle|\, x,t \right], \qquad (1.2.2)$$

where $\delta x_i = x_i(t + \delta t) - x_i(t)$.
These quantities are sometimes called the drift coefficient and the diffusion coefficient respectively.

(2) the higher incremental moments are zero, which means that the process $x(t)$ is continuous with probability one.

(3) the partial derivatives with respect to x_i and t in equation (1.2.1) exist and are piecewise continuous in their arguments.

As $P(x,t)$ is a probability density function, it must satisfy the conditions

$$P(x,t) \geq 0,$$

$$\text{and} \quad \int_R P(x,t)\, dx = 1,$$

where R is the entire x space. Usual derivations of the FP equation proceed from the assumption that $P(x,t_o) = \delta(x - x_o)$, in which case the solution of the FP equation is the conditional or transition probability density $P(x,t \mid x_o,t_o)$. As the FP equation is linear, the superposition principle shows that the solution of the equation is valid for a general functional initial condition $P(x,t_o)$, in which case the solution is not interpreted as a transition probability density.

The diffusive character of the Markov process assures that $P(x,t)$ is continuous in both arguments with probability one for $t > t_o$. A second continuity condition relates to the flux of the Markov process. If we interpret $P(x,t)$ as the density of particles whose trajectories are an ensemble of realisations of $x(t)$, then the principle of conservation implied in the continuity of the trajectories gives us, by Green's theorem, that the density $J(x,t)$ of the flow of particles across any arbitrary $n - 1$ dimension surface in R is continuous, normal to that boundary. The component of flow $J_i(x,t)$ in the x_i direction is given by

$$J_i(x,t) = b_i(x,t)P(x,t) - \tfrac{1}{2} \sum_j^n \frac{\partial}{\partial x_j} \left[a_{ij}(x,t)P(x,t) \right] \qquad (1.2.3)$$

and the flow density normal to any surface is given by the inner product of $J(x,t)$ and the unit normal of the surface. The FP equation is then written as

$$\frac{\partial P(x,t)}{\partial t} = - \sum_i^n \frac{\partial J(x,t)}{\partial x_i} = - \text{div } J(x,t).$$

These continuity conditions are necessary to construct solutions of the FP equation at points of discontinuity of $b(x,t)$ and $a(x,t)$. Examples of the use of these conditions in piecewise linear systems are given by Khazen [8, 19] and Merklinger [1], and we use a modification of these conditions in Chapter 3.

The FP equation (1.2.1) is the forward Kolmogorov equation of the process $x(t)$, and we can write down the backward Kolmogorov equation as

$$\frac{\partial P(x,s)}{\partial s} = - \sum_{i}^{n} b_i(x,s) \frac{\partial P(x,s)}{\partial x_i}$$

$$- \tfrac{1}{2} \sum_{i,j}^{n} a_{ij}(x,s) \frac{\partial^2 P(x,s)}{\partial x_i \partial x_j} , \qquad (1.2.4)$$

where s is the backward time parameter and equation (1.2.4) has the"initial" conditions $P(x,s_f)$, where $s \leqslant s_f$. As we are interested in the prediction problem and seek to evaluate the evaluation of the system's probability density in real (forward) time, we shall restrict out attention to the forward or FP equation.

Solutions of the FP Equation

An extensive historical background of the FP equation is given by Merklinger [1], and in the classic papers collected by Wax [11]. Most of the known solutions of the FP equation have been of the associated elliptic equation obtained by setting the left hand side of (1.2.1) to zero. With the added condition that $b(x,t)$ and $a(x,t)$ are not functions of time, the solution $P(x)$ gives the steady state statistical behaviour of stationary systems. The known stationary solutions fall into three categories:

(1) explicit analytic solutions for the statistics of linear systems [10],

(2) explicit analytic solutions for a restricted class of non-linear systems, including piecewise linear systems [3-9],

and (3) a numerical approximation approach to general non-linear systems, with attention concentrated on relay (piecewise linear) systems [1, 2].

Results of less generality have been obtained for the transient solution of the FP equation, for the time dependence introduces further analytic and computational difficulties. The known transient

solutions can be divided between

(1) explicit analytic solutions for the statistics of linear systems, conveniently expressed as linear differential equations for the moments of the system [11-16],
and
(2) analytic but usually approximate solutions for a few first order non-linear systems [17, 18, 21].

Reading the above references makes it clear that the challenge lies in the analysis of non-linear systems. The reason for this is seen in the derivation of differential equations for the moments of stochastic systems [74], where the simplifications afforded by system linearity become apparent. The infinite set of system moments are an equivalent statistical system description to the density function $P(x,t)$, and if we can obtain an explicit solution for them we have obtained an explicit solution of the transient FP equation. If the system is linear, the equations for the moments are linear and first order, and are arranged in a recursive order so that the solutions for each moment can be obtained in a straightforward fashion. If the system is not linear, the moment equations may not be linear, but more important is the fact that the unknowns in the equations are no longer arranged in a recursive order, and all equations must be solved simultaneously. As there are an infinite number of equations, this cannot be done, unless we can set all moments beyond a certain order to zero. The convergence properties of the moments in general do not allow this, but in some cases, other statistical descriptions of systems do (see Section 2.4). But in general, analytic solutions for arbitrary non-linear systems cannot be obtained.

Also, for linear systems, spectral analysis and correlation techniques are available, and are equivalent to Fokker-Planck methods. They are often more convenient to use, particularly in the well-known analysis of stationary linear systems. Again, the simplicity of this approach does not apply to non-linear systems, and we are faced with using approximate numerical techniques to obtain the solution of the FP equation for non-linear systems.

1.3 Outline of Thesis

As well as listing the main topics of the thesis, this section presents the origin of, and some reflection on, the more interesting of the points discussed in the thesis. In this way we state the originality of the ideas presented, and point out which topics are considered to be of greatest interest.

Finite Difference Methods (Section 2.3) This section and the next one are two approaches to the numerical solution of the transient FP equation. Implementing the finite difference method required a fair amount of initial study and involved little originality. However, certain problems peculiar to the FP equation were met and overcome by somewhat arbitrary methods. The dual solution method was new and proved to be very helpful. The design of the adaptive sampling and alarm scheme given as an example is an interesting application of little used transient FP techniques.

Hermite Transform Methods (Section 2.4) The idea of using the Hermite Transform is a result of an attempt to get the most concise description possible of a system's statistics. To this end, the Gram-Charlier series was proposed for near-Gaussian systems, and when applied to the FP equation, turned out to be a neat integral transform method. However, the form of system amenable to solution by this method was rather restricted, as only one-dimensional systems with smooth non-linearities could be handled efficiently. Thus this method was considered to be more of academic than practical interest.

In general, this method and the finite difference method were felt to be rather unrewarding, in that a large effort was put into obtaining solutions for a small class of low dimensional systems. Thus methods of more generality were sought, and as a result, Monte Carlo methods were investigated.

General Simulation Results (Section 3.2) When looking at Monte Carlo methods for solving high dimensional FP equations, it was discovered that the methods could be modified to solve a wide class of parabolic equations. A general simulation result was developed by associating

the dependent variable of a parabolic equation with the density of trajectories of a simulated diffusion process. Analogies were drawn between the simulated trajectories and physical properties of the problem which the parabolic equation describes, and concepts such as flux and continuity were used to specify the nature of the simulation. We found that if the parabolic equation had the form of a FP equation, the simulation had conserved trajectories (agreeing with the law of conservation of probability density), and if the simulation was modified to allow the number of trajectories to be time varying, other parabolic equations could be solved.

Previously developed Monte Carlo methods for parabolic equations are based on the connection between simulated diffusion processes and the backward Kolmogorov equation (1.2.4), while our new method is based on the forward Kolmogorov or FP equation (1.2.1). There are conceptual differences between the two methods with the result that the new method can solve a more general type of parabolic equation than the previous method. Not enough examples were solved to fully test the efficiency (and hence usefulness) of the new method, but it may prove to be the only method of solving certain high dimensional parabolic equations. In addition we present several theorems relating to the positivity of the solutions of parabolic equations which support the new simulation method of solving the equations.

The Relation between Continuous Markovian and Non-Markovian Processes (Section 2.2, Chapter 4) To implement the results of Chapter 3, we have to simulate (Markovian) diffusion processes on a physical computer where all processes are non-Markovian. In addition, many of the results of modern stochastic control theory such as the design of non-linear filters are developed in the stochastic calculus (where processes are Markovian), but must be implemented in the ordinary (non-Markovian) calculus. Thus the relation between these two types of processes had to be studied.

This was an interest developed jointly with J. M. C. Clark, when the paper of Wong and Zakai [24] came to our attention, and was later promoted by discussions with K. J. Astrom. Once the differences between these types of processes become apparent, it became clear

that much of the previous work using FP techniques could be misinterpreted when applied to physical situations, and a unified approach was needed to clarify the application of continuous stochastic control theory.

The approach of Clark [22] was to study the convergence of the trajectories of physical processes to those of diffusion processes, which is called convergence in the mean. As we are primarily interested in the ensemble statistics of processes as opposed to the properties of the individual sample paths, we confine our attention to the less demanding (mathematically and conceptually) concept of convergence in distribution. In this way we are able to show the convergence of more general physical processes (with non-stationary and non-Gaussian noises) to diffusion processes than those considered by Clark.

The purpose of both Clark's work and ours is to show how diffusion processes can be approximated by physical processes, and vice versa. Stratonovich [21] also studies this approximation, and our approach is similar to Stratonovich's except that our approach more clearly shows under which conditions we may approximate one process by another.

Both Clark's approach and ours point out the need of characterising a physical noise process by its characteristic matrix. This characterisation is more complete than that used by most previous authors, and using this characterisation, we can resolve a controversy which has existed in the interpretation of physical systems with random coefficients.

The main contribution of the present approach over Clark's and Stratonovich's centres around the convenience of matching physical and diffusion processes by evaluating and equating their incremental statistics over a small but non-vanishing time increment. The PRBS example of Appendix D shows that our method can provide a very simple method of choosing equivalent processes without, in fact, having to directly evaluate the characteristic matrix of the physical noise involved.

Computing Problems of Simulation (Chapters 5 & 6) The work of these chapters was done to illustrate the theory of Chapter 4 and to investigate the practical difficulties of simulating diffusion processes on a computer.

On an analogue computer (Chapter 5), we did not wish to get involved with the practicalities of computing, recording and general accuracy analysis, as this would have been a time-consuming task detracting attention from the main points of interest. Instead, we concentrated on those points peculiar to our problem - the choice of noise source which simulates the white noise of the diffusion process, and the proper characterisation of the noise source by experimental means. The analysis of Chapter 4 showed that different noise sources introduced different biases into simulations in a manner which depends on the characteristic matrix of the noise, and we illustrate this in a qualitative fashion with several examples.

On a digital computer (Chapter 6), the noise source is usually confined to pseudo random numbers generated on the computer, and so the interest centred around the choice of computing method (digital formula). Previous computing methods for simulating diffusion processes were based on a very simple formula which was not very efficient in terms of accuracy achieved in relation to computing time. Thus more efficient formulae were sought, and it turned out that the results of Chapter 4 allowed us to specify a general high order computing method of simulating diffusion processes. The convergence rates of these formulae were investigated, and found to be different from those rates met in classical numerical analysis problems. An example verified the theoretical convergence rates and illustrated the increased efficiency of the new method.

Of the data reduction and smoothing methods used, one was interesting and apparently novel. It was based on the concise statistical description afforded by the orthogonal functional expansions of Section 2.4, and was found to preserve the lower order moments of the statistical data.

The Stochastic Calculus (Appendix A) This appendix gives an expository account of the properties of the stochastic calculus in a form which is not generally available. The stochastic calculus has been used recently in the control literature (e.g. Wonham's papers), yet its properties have only been presented in journals and books of a mathematical flavour (of these the book by Doob [20] seems to be the most readable reference, yet even it is rather formidable in tone to one with an engineering background). The appendix and the separate note [74] illustrates the ease of analysis which the stochastic calculus affords of continuous Markov processes, and the PRBS analysis in Appendix D is a good example of the power of the analysis.

The author was introduced to the properties of the stochastic calculus by J. M. C. Clark, and in particular, the properties of the Stratonovich stochastic calculus were discovered through a translation of [50] obtained with the help of F. Domnin. Clark pointed out that the Stratonovich calculus has advantages of representation over the usual Ito stochastic calculus when dealing with limiting forms of physical processes, and this prompted the use of the Stratonovich form of writing diffusion processes in Chapters 4 to 6.

Model of the PRBS (Appendix D) The pseudo random binary sequence (PRBS) has many advantages as a random noise source, and is coming into widespread use as a test signal for system simulation and on-line identification. But of the many recent papers on the theory and practice of the PRBS, none have fully exploited the known deterministic properties of the PRBS and shown how these effect the statistical properties of the PRBS. Our analysis is a step in this direction, and shows that systems driven by a PRBS can have some unusual transient properties.

Historically, the PRBS analysis holds an important place in the development of the latter half of the thesis and it is interesting to trace the evolution of ideas. It was noticed that the PRBS had some unusual transient statistical properties, and as a basis of comparison, we sought to compare the integral of the PRBS with the Wiener process. We knew, for example, that the known number of positive and negative bits in one period of the PRBS made the integral of the PRBS come back

to near zero every period while the Wiener process did not. Thus we wished to see how the variance of the integrated PRBS varied with time compared with the linear increase of the variance of the Wiener process. Realising that the transient variance could only be evaluated conveniently for continuous Markov (diffusion) processes, it was decided to construct a diffusion model of the integrated PRBS which incorporated the interesting deterministic properties of the PRBS. Knowing that a diffusion process is specified by two incremental moments, and prompted by the inherently discrete nature of the PRBS, we decided to evaluate the incremental properties of the integrated PRBS over a finite time interval equal to the discrete time quantisation of the PRBS switching points. Then the diffusion process which has approximately the same incremental properties over the same time increment was chosen as the model of the integrated PRBS (a physical process). This is done in equations (3) to (8) of Appendix D, and is the genesis of the approach used in the analysis of Section 4.1.

Being a pseudo random process with a limited number of possible outcomes, the statistics of functions of the PRBS could easily be evaluated exactly on a digital computer, and in this way we could evaluate the exact error of the diffusion model. The fact that the error was found to be quite small encouraged the analysis of Section 4.1. This analysis was a departure from the earlier analysis given by Clark [22], and in fact, the need of characterising a physical noise source by Clark's characteristic matrix is then illustrated in a manner independent of Clark's.

CHAPTER 2

THE DIRECT APPROACH TO PREDICTION PROBLEMS - SOLUTION OF THE FOKKER-PLANCK EQUATION

2.1 Fokker-Planck Equation for a Diffusion Process

The Fokker-Planck (FP) equation for the first order probability density of a Markov diffusion process $x(t)$ was introduced in Section 1.2. From the discussion of Appendix A, it is apparent that the diffusion process must be described by a stochastic, as opposed to an ordinary, integral or differential equation. The Ito stochastic differential equation (s.d.e.) for an arbitrary diffusion process can be written in the state vector form

$$dx(t) = f(x, t)\, dt + F(x, t)\, dw(t), \qquad (2.1.1)$$

where $x(t)$ is an n-column vector of the state of the Markov process,

$f(x, t)$ is an n-column vector representing the deterministic part of the state dynamic equations,

$F(x, t)$ is an $n \times m$ matrix representing the random part of the dynamics,

and $dw(t)$ is the Ito stochastic differential of an m-column vector of unit parameter independent Wiener processes $w(t)$.

The state vector equation (2.1.1) would be in a more familiar form if divided through by dt, but the properties of $dw(t)$ do not allow us to do this (see Appendix A). Nevertheless, it is useful

to consider $\frac{dw(t)}{dt}$ or $\dot{w}(t)$ as a concept when comparing it with the physical noise $y(t)$ introduced in the next section, as $\dot{w}(t)$ can be considered as a limiting form of $y(t)$, and is, in fact, the usual definition of white noise.

The FP equation (1.2.1) for the process (2.1.1) is written down from the incremental moments (1.2.2) of the process. These are

$$b_i(x, t) = f_i(x, t) , \qquad i = 1, n, \tag{2.1.2}$$

and $$a_{ij}(x, t) = \sum_{k}^{m} F_{ik}(x, t) F_{kj}^{T}(x, t), \quad i, j = 1, n. \tag{2.1.3}$$

The diffusion process (2.1.1) can also be expressed as a Stratonovich stochastic differential equation (see Appendix A) whose i:th component is

$$\bar{d}x_i = \left[f_i(x, t) - \tfrac{1}{2} \sum_{jk} F_{jk}(x, t) \frac{\partial F_{ik}(x,t)}{\partial x_j} \right] dt + \sum_{k} F_{ik}(x, t) \bar{d}w_k(t), \tag{2.1.4}$$

where $\bar{d}\cdot$ denotes a stochastic differential in the Stratonovich sense. It should be emphasised that (2.1.1) and (2.1.4) are precise but differing definitions of the same diffusion process, and they differ in bias term only because of the different rules of defining the Ito and Stratonovich stochastic integral. Because of the simple relation between the coefficients of the Ito equation (2.1.1) and the incremental moments (2.1.2, 3), we will use the Ito definition of the diffusion process in the sequel. We will also introduce the Stratonovich form (2.1.4) when it provides a convenient representation.

2.2 Fokker-Planck Equation for a Non-Markovian Process

A continuous Markov process (synonymous with diffusion process) differs from a continuous non-Markovian process by the properties of the noise source which generates the random process. In particular, in comparing the Markov process noise $\dot{w}(t)$ of equation (2.1.1) with the noise $y(t)$ of the non-Markovian process (2.2.1), successive values of the Markov noise (i.e. set $t = t_1, t_2, t_3 \ldots$ for $t_1 < t_2 < t_3 \ldots$) will be independent of each other for arbitrarily small time increments, $t_i - t_{i-1}$, whereas successive values of the non-Markovian noise do lose their independence as the time increments are arbitrarily refined. In the stochastic process literature, this property of the Markov noise is expressed by stating that the process $w(t)$ has independent increments (or is infinitely divisible) [20, pages 96 and 273]. The only continuous process with independent increments is the Wiener process $w(t)$ and thus the diffusion process noise is Gaussian white noise. It is because of this property of infinite divisibility that the rules of the stochastic calculus differ from those of the ordinary calculus.

This property can also be seen from the correlation function $R(\tau)$ of the noise. White noise has a delta function as a correlation function, which is exactly zero for any non-zero time shift τ. In contrast, the non-Markovian noise $y(t)$ has a correlation function which is non-zero for some arbitrarily small but non-zero time shift τ. In the frequency domain, white noise has a continuous flat power density spectrum at all frequencies (including infinite frequency), while non-Markovian noise does not have a flat spectrum at all frequencies.

Now consider what kind of noise can exist in the physical world. Proceeding from the premise that no existing noise source can have an infinite power, and as power is the infinite integral of the power density spectrum, then no existing noise source can have a power density spectrum which is flat at all frequencies. In order for the power to be finite, the noise source must have an upper frequency f_u, above which the noise signal has no significant power. Also, the correlation function of such a noise source will have a non-zero value for some non-zero time shift τ.

These considerations lead us to the conclusion that any noise source existing in the physical world cannot be the Markovian white noise discussed above, but must be a non-Markovian noise source. We shall call such a noise physical noise $y(t)$ and a random process involving physical noise we shall call a physical process $X(t)$ (this terminology is not new: see for example [41], [22]). This terminology for physical processes parallels white noise $\dot{w}(t)$ for diffusion processes $x(t)$.

By the same argument, diffusion processes cannot exist in the physical world ($\dot{w}(t)$ has infinite power, and an infinite magnitude at all times) but are introduced as a mathematical concept convenient for analysis. An example is the analysis of the pseudo random binary sequence [57] using diffusion processes, an analysis which could not be conveniently carried out by other means.

Another indication of the difference between physical and diffusion processes is that all the derivatives of continuous physical processes are finite almost everywhere with probability 1 [21, p.124], whereas the n:th order Markov process $x(t)$ (2.1.1)

does not possess a derivative (that is, the derivative is always infinite. Certain components only of $x(t)$ will have a derivative if the corresponding rows of $F(x, t)$ are void).

Let us consider the physical process $X(t)$ described by the ordinary differential equation

$$\dot{X}(t) = g(X, t) + G(X, t)\, y(t), \qquad (2.2.1)$$

where $y(t)$ is an approximately Gaussian m-vector physical noise process which, as discussed above, must have a finite spectrum with which we can associate an upper frequency f_u or correlation time $\tau_{cor} \triangleq 1/f_u$ [21, p.22]. The forcing function or dynamics $g(X, t)$ and noise coefficient $G(X, t)$ are analogous to f and F of the diffusion system (2.1.1). We wish to have an equation which describes the statistical dynamics of the physical process (2.2.1), but unfortunately the FP equation cannot be directly applied to this non-Markovian process, as the derivation of the FP equation requires that the process be Markov*.

However, if we can find a diffusion process $x(t)$ which possesses some of the properties of the physical process $X(t)$, then by writing down the FP equation for $x(t)$, we will obtain a differential equation for some of the statistical properties of $X(t)$. Later in this section and in Sections 4.1, 2, we will discuss what properties of $x(t)$ and

---

* A Markov process can be defined as a process $x(t)$ for which all the information of the conditional or transition probability $P[x(t_1) \mid x(t_2), x(t_3) \ldots x(t_n)]$ is contained in $P[x(t_1) \mid x(t_2)]$ where $t_1 > t_2 > t_3 \ldots > t_n$ are points in time which can be arbitrarily close together. This "no after-effect" property is a result of the independence of the infinitely divisible noise increments mentioned earlier, and is essential to the derivation of the FP equation (for example, see [21, chapter 4]).

$X(t)$ we must match in order that $x(t)$ can be called a "diffusion approximation to $X(t)$", and we will discuss how good the approximation is in representing certain statistical properties.

Generally speaking, there are two ways of proposing a diffusion approximation to the physical process $X(t)$. On one hand we can represent $y(t)$ as the output of a filter driven by white noise, and the extra state variables, corresponding in number to the order of the filter (i.e. the order of the differential equation describing it) are appended to the state variables of the system (2.2.1). The resultant system in extended state space constitutes a Markov process, and the FP equation can be applied to this system. This method seems to have been first discussed in the control systems literature by Khazen [23, 8], who shows that if $y(t)$ has a rational spectral density with a denominator of degree $2k$, then $y(t)$ can be generated by a k:th order linear filter with white noise input. However, this method is at best an approximation as no physical process possesses a rational spectral density (or is exactly Gaussian), and the order of the filter generating $y(t)$ would have to be extended indefinitely before $y(t)$ would be modelled exactly [21, p.124]. In practice a few extra state variables would likely give an adequate approximation, but we shall see later that the addition of state variables greatly complicates the solution of the FP equation by the methods of this chapter. Thus we shall in most cases prefer the method given below which does not involve the addition of state variables to the physical process.

The second method of obtaining a diffusion approximation to $X(t)$ is to directly replace $y(t)$ in equation (2.2.1) by white

noise of equal zero frequency cross spectral density, and alter the dynamics $g(X, t)$ by the addition of a bias term which ensures that the resulting process $x(t)$ has equivalent statistics to $X(t)$ over time increments which are substantially greater than the correlation time τ_{cor} of the noise $y(t)$. This method can be deduced from the work of Stratonovich [21, Chapter 4, Sections 7-9] who derives an approximate FP equation directly from the statistics of the increments of $X(t)$, but the form of the recent (and equivalent) results of Clark [22, Chapter 2] on the approximation of physical processes by diffusion processes are more appropriate to the present argument. In Sections 4.1, 2, we will derive and extend Clark's results in a fashion which is more relevant to the present problem.

Clark shows that, given a family of physical processes $X(t, f_u)$ parameterised by the noise upper frequency f_u, as the upper frequency parameter f_u is extended to infinity the members of the family converge to a diffusion process $x(t)$ in such a way that the second moment of the error between $X(t, f_u)$ and $x(t)$ is of order f_u^{-1}.* This limiting process will be called the equivalent diffusion process to $X(t)$, and is given by the following Ito stochastic differential equation

$$dx(t) = g(x,t)dt + \sum_{k,l}^{m} Q_{kl}(x,t)A_{kl}dt + G(x,t)Bdw(t), \qquad (2.2.2)$$

where Q_{kl} is an n-column vector with i:th component

$$(Q_{kl})_i = \sum_{j}^{n} G_{jl} \frac{\partial G_{ik}}{\partial x_j}, \qquad (2.2.2a)$$

* This convergence is a sample path convergence, as opposed to the incremental statistical convergence considered in Sections 4.1, 2.

A_{kl} is an element of the constant m x m matrix given by

$$A = \lim_{f_u \to \infty} \int_{-\infty}^{t} E[y(t)\, y^T(s)]ds, \qquad (2.2.2b)$$

B is an m x m constant matrix introduced to relate the independent unit parameter components of dw(t) to the arbitrarily scaled and cross-correlated components of y(t) in such a way that the integral of y(t) converges in the mean to Bw(t) as f_u tends to infinity. (2.2.2c)

Notes on Equation (2.2.2) and Clark's Results

(a) The term of (2.2.2) involving Q_{kl} distinguishes the form of the equivalent diffusion process from the form of the original physical process (2.2.1), and is called the bias term. It will be zero if the noise y(t) is additive noise, that is, G(X, t) = G(t), or if the noise y(t) is only multiplied by those components of the state vector X(t) which are relatively smooth (then $Q_{kl} = 0$).

(b) The matrix A represents the total information that we must have of the physical noise process y(t) in order to obtain the equivalent diffusion process, and has been called the characteristic matrix of the physical noise by Clark. The definition (2.2.2b) differs slightly from that given by Clark who gives

$$A = \lim_{f_u \to \infty} \frac{1}{t} \int_o^t \int_o^r E[y(r)\, y^T(s)]\, ds\, dr.$$

The extra smoothing obtained by the $\frac{1}{t} \int_o^t [\,\cdot\,]\, dr$ operator in Clark's definition is included to account for "quasi-stationary" noise processes. These are processes which are not non-stationary

in the normal sense (where the statistics of $y(t)$ depend explicitly on t), but are processes which are non-stationary in the sense that statistical variations are allowed over a time period whose maximum is of the order of f_u^{-1}. Thus the non-stationarity disappears as f_u tends to infinity, but the extra smoothing is needed to make the limiting operation smooth. An example of such a quasi-stationary process is the piecewise constant process discussed by Clark [22, pages 24 and 84]. This process is constant over time increments f_u^{-1}, and hence is non-stationary only over time increments f_u^{-1}, and as f_u tends to infinity, this non-stationarity disappears. Without modification, Clark's analysis does not include noise processes $y(t)$ which are non-stationary in the normal sense, for example, a noise process with a time varying variance.

Clark defines his physical and diffusion processes in the time interval $[0, T]$, while we consider ours in the interval $[-\infty, T]$. There is no conceptual difference between these approaches, except that Clark's is more convenient for sample path comparisons. He must specify initial conditions for his processes at $t = 0$, while we do not. The effect on the definition of the characteristic matrix A is that we use the lower integration limit of $-\infty$ in (2.2.2b) while he uses a zero lower limit. This is inconsequential as f_u tends to infinity, as the integral from $-\infty$ to zero of (2.2.2b) contributes nothing to the overall integral, and our definitions coincide for all positive t.

If we exclude the quasi-stationary noise processes allowed by Clark, the noise processes considered are stationary in the wide sense [20, p.95], and then we can define the characteristic matrix A as

$$A = \underset{f_u \to \infty}{\text{limit}} \int_0^{\infty} R(\tau)\, d\tau, \qquad (2.2.3)$$

where $R(\tau)$ is the correlation function of the stationary noise $y(t)$

$$R(\tau) = E\left[y(t)\, y^T(t - \tau)\right]. \qquad (2.2.4)$$

In Chapter 4, we will broaden this definition of A in two ways. Instead of discussing a family of physical processes parameterised by the noise upper frequency f_u, we will be considering one particular physical process where the limit $f_u \to \infty$ is not taken. Then the characteristic matrix A will be taken simply as the integral on the right hand side of (2.2.3) without the limiting operation. This represents a change from the concept of the equivalent diffusion process proposed by Clark, for the A matrix of this definition does not necessarily remain constant under the limiting operation on the physical noise used by Clark. However, it is known that at least for some physical noise sources that the A matrix of this definition does remain constant as its spectrum is extended to infinity, and it is proposed in Section 4.2 that the constancy of the A matrix be a condition of the limiting operation on a physical noise source.

The second extension of the definition will include non-stationary noise processes, where the left hand side of (2.2.4) becomes $R(t, \tau)$, and so the left hand side of (2.2.3) becomes $A(t)$. Although the concept of a non-stationary correlation function $R(t, \tau)$ may not be familiar, it can be defined by interpreting the $E[\cdot]$ operator of (2.2.4) as an ensemble average instead of a time average, in which case the non-stationarity of $y(t)$ poses no difficulties.

(c) The definitions (2.2.2b, c) imply that

$$A + A^T = B B^T, \tag{2.2.5}$$

and as far as the FP equation of (2.2.2) is concerned, B need not be specified as only $B B^T$ is needed. Further, if we wish the system (2.2.2) to model the statistics of the physical process (2.2.1), then B can be chosen to satisfy (2.2.5) without regard to (2.2.2c), which is not a unique specification of B. However, if we wish the system (2.2.2) to model the sample paths of (2.2.1), then B must be chosen exactly as in (2.2.2c). It is noted that B in (2.2.2) is merely a noise scaling factor, and could be incorporated in the coefficient matrix G(x, t) (as in the form (2.1.1)), or in the Brownian increment dw(t) (see the form used in [74]). The B matrix has not been incorporated in these terms, but has been kept separate here to keep G(x, t) of (2.2.2) the same function as G(X, t) of (2.2.1), and to keep w(t) a unit parameter independent Wiener process. In Section 4.1, the relation (2.2.5) is modified when the physical noise y(t) is non-stationary.

(d) Clark's proof has required that y(t) be a Gaussian process, but the results of Stratonovich [21] and the more particular results of Wong and Zakai [24, 25] and Astrom [26] indicate that the approximation result of Clark is valid for a wide class of piecewise continuous noise processes y(t). In Sections 4.1, 2, we present an analysis which derives Clark's results using only the assumption that the physical noise process y(t) possesses a (non-stationary) correlation function R(t, τ). The latter assumption seems to cover most noise processes found in practice.

Fokker-Planck Equation for the Diffusion Approximation

We can write down the FP equation (1.2.1) for the diffusion process (2.2.2) by noting that the incremental moments (1.2.2) are given by [see Appendix A, or compare with (2.1.1, 2)]

$$b(x, t) = g(x, t) + \sum_{k,l}^{m} Q_{kl}(x, t)\, A_{kl}, \qquad (2.2.6)$$

and

$$a(x, t) = G(x, t)\, B\, B^{T}\, G^{T}(x, t)$$

$$= G(x, t)\, [A + A^{T}]\, G^{T}(x, t). \qquad (2.2.7)$$

Concerning the choice of diffusion approximation (2.2.2) to the physical process (2.2.1) by the method of Clark, it will be useful to offer the following intuitive statement. The diffusion process $x(t)$ will be a process which will appear to have properties similar to the physical process $X(t)$ to an observer who can only detect frequencies substantially below f_u. Then the solution of the FP equation for $x(t)$ will accurately give the transition probabilities of $X(t)$ over time increments substantially greater than τ_{cor}. This statement cannot be deduced from the results of Clark who speaks of the order of convergence of the family $X(t, f_u)$ to $x(t)$ but does not give a bound on the error. However, the related statements of Stratonovich [21, pages 89, 94 and 122-126] substantiate the above statement. Moreover, the main purpose of the analysis given later in Sections 4.1, 2, apart from the extensions to non-Gaussian and non-stationary noise processes, is to derive the equivalent diffusion process in such a way that the approxi-

mations made in the derivation give an intuitive feeling for the scope and validity of the approximation. At that stage, the intuitive statement given above will be enlarged upon, and its plausibility should become evident.

In contrast, the first method given for modelling non-Markovian processes by adding extra state variables to account for the non-zero correlation time τ_{cor} of the physical noise $y(t)$, does model the physical process for time increments of the order of, and smaller than, τ_{cor}. Thus if this statistical information is required, the first method of modelling must be used.

The results of Clark, their extensions, and their application to simulation problems will be discussed more thoroughly in subsequent chapters. The rest of this chapter will discuss two methods of solving the FP equation numerically, and present a simple example. The FP equations involved will describe the statistics of the diffusion process $x(t)$ which is an approximation to a physical process $X(t)$. In the light of the comments of this section, and Sections 4.1, 2, the processes $x(t)$ and $X(t)$ will be interchanged without further comment, and the implicit approximations involved should be kept in mind.

2.3 Numerical Solution by Finite Differences

It is well known that analytical solution of partial differential equations is feasible only in special cases. Stratonovich [21, Ch. 4, Section 4] discusses the separation of variables method for one dimensional cases, but the resulting second-order equation in x can only be solved for simple cases. Other analytical methods are discussed by Merklinger [1, Chapter 4] and as he comments, it becomes clear that our attention must be devoted to approximate numerical methods if non-linear examples of some generality are to be studied.

Although semi-analytic methods can be used to aid a numerical solution (e.g. transform or orthogonalisation methods, Section 2.4), the most direct approach to numerical solution is by finite differences, and is the approach most commonly used for partial differential equations in general [27, Chapter 15]. Although much research has been devoted to finite difference methods by numerical analysts, the solution of partial differential equations on a computer is still by no means a straightforward operation.

The FP equation is a linear parabolic partial differential equation, and an account of the problems associated with the application of finite difference methods to parabolic equations in one space variable is given by Richtmyer [28]. As the FP equation is linear, we avoid many problems associated with solving non-linear partial differential equations, but the nature of the space dimensions of the FP equation, which are the spatial domains of the system (2.1.1), present some interesting problems.

The author has had experience with solving some one- and two-dimensional examples, and the method pertaining to a one-dimensional example has been reported earlier [29]. As the finite difference methods used are fairly standard ones in the field of numerical analysis the presentation of this section will be confined to a discussion of a few special problems associated with the FP equation, such as the treatment of spatial boundaries during a transient solution. To this end the material in [29] will be reviewed, as the one-dimensional case is sufficient to illustrate the techniques involved, and the

design example presented in [29] will be discussed as it illustrates an interesting application of transient FP techniques.

In connection with the discussion of the previous section, the FP equation presented in [29] was only valid for systems with additive noise (e.g. $G(X, t) = G(X)$ in (2.2.1)), in which case it did not matter if the noise $y(t)$ were strictly white or not. The quantity $\sigma^2(t)$ introduced in [29, page 257] should have been called the zero frequency cross-spectral density, which is the $d\tau$ integral of the variance-covariance matrix $R(t, \tau) = \sigma^2(t)\delta(\tau)$. In Section 4.1 we shall see that $\sigma^2(t)$ is given by

$$\sigma^2(t) = A(t) + A^*(t) = \int_{-\infty}^{\infty} R(t, \tau)\, d\tau. \qquad (2.3.1)$$

In this thesis we shall allow forms of the FP equation which are more general than that of [29], in that we shall allow state-dependent (non-additive) noise processes, and so have to distinguish between white noise and physical noise as outlined in Sections 2.1 and 2.2.

2.3.1 Example of a Noisy Control System

Let us consider the first order regulating system given by the following differential equation

$$\dot{X}(t) = -\underline{c}_1 X - \underline{c}_2 X^3 + \underline{c}_3(t) \qquad (2.3.2)$$

as shown in Figure (2.3.1). Here $X(t)$ is the state of the system, $\underline{c}_3(t)$ is the command input, and $\underline{c}_1$, $\underline{c}_2$ are the feedback coefficients. The system has the response or relaxation time [21, P. 99]

$$\tau_{rel} = -g_X(X)^{-1} = (\underline{c}_1 + \underline{c}_2 X^2)^{-1} \qquad (2.3.3)$$

in which X^2 should be replaced by the mean square value of X to give a realistic value to τ_{rel}. The relaxation time is analogous to the time constant of a first order linear system, and τ_{rel}^{-1} gives an estimate of the upper frequency response of the system.

Further, we shall assume the feedback coefficients can have random components along with a constant part

$$\underline{c}_i = c_i + y_i(t) \quad , \quad i = 1, 2, \tag{2.3.4a}$$

and the command signal has a deterministic part $c_3(t)$ which is constant or slowly moving with respect to τ_{rel}^{-1}, and a random part $y_3(t)$

$$\underline{c}_3(t) = c_3(t) + y_3(t). \tag{2.3.4b}$$

We shall assume that $y_i(t)$, $i = 1, 3$ are wide band physical noise sources of zero mean and correlation function $R(t, \tau)$. The noise source $y(t)$ will then be characterized by the matrices $A(t)$ and $A^*(t)$ as in Chapter 4.

When the system (2.3.2) is written out in the form (2.2.1), we find $g(X, t)$ has the single component

$$g(X, t) = - c_1X(t) - c_2X^3(t) + c_3(t) \tag{2.3.5a}$$

and $G(X, t)$ has the single row

$$G(X, t) = [\ -X(t), \quad -X^3(t), \quad 1\ ]. \tag{2.3.5b}$$

In order to obtain the diffusion process $x(t)$ which is equivalent to the physical system $X(t)$ (2.3.2) in the manner described earlier, we will need the terms of equation (2.2.2). From (2.3.5b) we obtain the matrix

$$Q(X, t) = \begin{bmatrix} X(t) & X^3(t) & -1 \\ 3X^3(t) & 3X^5(t) & -3X^2(t) \\ 0 & 0 & 0 \end{bmatrix}, \tag{2.3.6}$$

which, in conjunction with the noise characteristic matrices, gives the equivalent diffusion process (2.2.2). The FP equation is then obtained via the incremental moments (2.2.6, 7). The solution of the FP equation for the equivalent diffusion process will give us the statistics of the physical system over time increments substantially greater than τ_{cor}, the largest time shift τ at which any element of $R(t, \tau)$ has a significant non-zero value.

In Chapter 5 we shall see what form the characteristic matrix A takes for some common noise sources, but for this chapter we will only consider the example given in [29]. In this case, where c_1, c_3, $y_1(t)$ and $y_2(t)$ are all zero, the noise $y(t) = y_3(t)$ is scalar and additive, and the bias term of (2.2.2) given by A and Q is zero. The diffusion process equivalent to (2.3.2) is then simply

$$dx(t) = - c_2 x^3(t)\, dt + \sigma(t)\, dw(t)$$

as $y(t)$ has the property

$$A(t) + A^*(t) = B^2(t) = \sigma^2(t) = 2D(t),$$

where B is introduced in equation (2.2.2c) and later modified to $B(t)$ in equation (4.1.26), $\sigma(t)$ is as used in [29], and $2D(t)$ is commonly called the intensity of the noise. The associated FP equation is given by

$$\frac{\partial P}{\partial t}(x, t) = 3c_2 x^2 P(x, t) + c_2 x^3 \frac{\partial P}{\partial x}(x, t) + D(t) \frac{\partial^2 P}{\partial x^2}(x, t) \qquad (2.3.7)$$

with $P(x, t_o)$ given as an initial condition at $t = t_o$.

Although this example adequately illustrates the points of this section, it does not bring out problems arising when $g(X, t)$ is discontinuous in X, or when the derivative $g_X(X, t)$ cannot be explicitly obtained. The case of a discontinuous $g(X, t)$ has been extensively discussed by Merklinger [1]. The FP equation does

not hold at such discontinuities, and the solution must be obtained separately in regions divided by the discontinuity and then pieced together at the boundaries using continuity and conservation conditions as discussed in Chapter 3. If $g_X(X, t)$ is not explicit, it must be obtained by the finite differencing method below, which will add to the finite differencing errors in the space variable.

2.3.2 Choice of Finite Difference Model

Characteristically, finite difference methods divide the domains of the independent variables x and t of (2.3.7) into a finite number of discrete cells parameterised by the cell dimensions Δx and Δt. Algebraic methods are then used to solve for the dependent variable P at discrete points on the cell boundaries. The mechanics of the continuous to discrete transformation are not unique and are summed up in the finite difference model. In [28, Table I, page 93], Richtmyer gives an extensive list of possible finite difference models for the basic parabolic equation.

Parabolic equations differ from elliptic equations by the presence of the time variable t in the set of independent variables. Parabolic equations are "initial value problems" in the sense that only boundary conditions for $t \leq t_1$ are needed to obtain the solution $P(x, t_1)$ at time t_1. This means that the solution is naturally obtained in a step-wise fashion, solving at all x for t successively equal to $t_1, t_2, t_3 \ldots (t_{i+1} - t_i = \Delta t)$, whereas for elliptic equations, all solution points may be found simultaneously. This property, although common to all parabolic equations, comes from the "Markov property" inherent in the problem: the solution at any space point at a given time t is completely specified by the solution at all space points at the last available time (usually $t - \Delta t$). We will be exploiting the Markov nature of parabolic equations in Chapter 3. The steady state FP equation ($P_t = 0$ in (2.3.7)) is inherently an elliptic

equation, although it degenerates to an ordinary differential equation in the one-dimensional case, and to an equation similar to the parabolic type for the multidimensional case where the noise does not excite all states.

When the solution is obtained at successive time steps, the problem of solution instability may be encountered. Instability occurs if an error existing at one time step is magnified at subsequent time steps, for then the solution will eventually be ruined by the growing error. As the equations involved are linear, error terms will grow or decay exponentially, and a stability analysis usually proceeds by finding the eigenvalues of the linear operator which propogates the solution (and error) forward one time step (see [28], Chapter 4). If the eigenvalues are less than one in magnitude, the finite difference model will be stable.

The stability criterion is usually the main consideration in the choice of finite difference model. Some models will be stable for all values of the parameters Δt, Δx, but some will only be stable for restricted ranges of these parameter values, and in particular for unreasonably small values of Δt. The simplest models computationally are called explicit models, as the solution at each point in the new time step is expressed explicitly in terms of known solution values at old time steps. By contrast implicit models are those in which the solution at each point in the new time step involves adjacent unknown values in the same time step, as well as those from previous time steps, and so simultaneous equations must be solved to obtain the solution at each new time step. This involves techniques similar to those used for elliptic equations, but special considerations will allow us to use direct solution methods and avoid iterative techniques. In general we will prefer an implicit model as they are unconditionally stable, whereas the simple explicit model requires a severe upper bound on Δt to make them stable.

The most commonly used implicit model is that of Crank and Nicholson [28, Table I, p.93, model number 2; 27, p.402; or 30 for the original reference]. Among the implicit models it gives a

reasonable compromise between (computational) complexity and truncation error, and is stable for all values of Δt and Δx.

In introducing the discrete formulation of the FP equation we shall use the following notation and formulae:

Continuous	Discrete
t	$j\Delta t$, $j = 0, 1, 2 \ldots$
x	$x' + k\Delta x$, $k = 0, 1, 2, \ldots K + 1$
$P(x,t)$	$P_j(k)$
$\frac{\partial P(x,t)}{\partial t}$	$\frac{P_{j+1}(k) - P_j(k)}{\Delta t}$
$\frac{\partial P(x,t)}{\partial x}$	$\frac{P_j(k+1) - P_j(k-1)}{2\Delta x}$
$\frac{\partial^2 P(x,t)}{\partial x^2}$	$\frac{P_j(k+1) - 2P_j(k) + P_j(k-1)}{\Delta x^2}$

(2.3.8)

The quantity x' represents the left hand edge of the space grid, and the space derivatives given are the usual central differences. The time $t = j\Delta t$ is the present time at which we know the values $P_j(k)$, $k = [0, K+1]$, and the time $(j+1)\Delta t$ is the next time point, at which we must solve for the values $P_{j+1}(k)$. If the coefficient functions $g(x, t)$ and $G(x, t)$ occurring in the FP equation do not possess analytic derivatives, as they have in the present example, differencing formulae analogous to (2.3.8) would have been applied to these functions. Applying these formulae to the FP equation (2.3.7) of the system of our example, and grouping the unknown values on the left hand side, we have the discrete equation

$$- \tfrac{1}{2} h \, P_{j+1}(k+1) + (1+h) \, P_{j+1}(k) - \tfrac{1}{2} h \, P_{j+1}(k-1) \; =$$

$$P_j(k) + h_1(k)\, P_j(k) + h_2(k)\,[P_j(k+1) - P_j(k-1)]$$

$$+ \tfrac{1}{2} h\,[P_j(k+1) - 2\,P_j(k) + P_j(k-1)], \quad k = [1, K], \qquad (2.3.9)$$

where we have introduced the following constants and variables for convenience

$$h = \frac{D\,\Delta t}{\Delta x^2}\ , \qquad (2.3.9a)$$

$$h_1(k) = 3\,\Delta t\, c_2\,(x' + k\Delta x)^2\ , \qquad (2.3.9b)$$

and $$h_2(k) = \frac{\Delta t}{2\,\Delta x}\, c_2\,(x' + k\Delta x)^3\ . \qquad (2.3.9c)$$

It is noted that we have applied the Crank and Nicholson formula only to the second derivative term (with coefficient $\frac{1}{2}h$) which is all that is necessary for stability considerations [28, p.98]. Young remarks that better accuracy is obtained by applying the time centered Crank and Nicholson formula to all terms [31, p.424], but it is felt that this is a small point unless the magnitude of the low order terms are very large compared with the second order term.

2.3.3 Solution Procedure and Boundary Conditions

It is noted that equation (2.3.9) is in reality K simultaneous equations for K + 2 unknowns. In most parabolic equations, the "unknowns" $P_{j+1}(0)$ and $P_{j+1}(K+1)$ would be supplied as boundary conditions, but for the FP equation, these conditions can only be deduced from the behaviour of the system (2.3.2) at the extremes of its space variable X(t). Reflection on this question soon leads us to agree that the structure of physical systems never allows us to specify the probability of the systems' state near its extremes,

although some information may be available on physical (or practical) constraints on the system which would be incorporated into the forcing function $g(X, t)$. We are left then to select arbitrary methods of choosing a boundary and applying boundary conditions which will not affect the solution accuracy in regions in which we are interested. This means we must devise a procedure for specifying $P_{j+1}(0)$ and $P_{j+1}(K + 1)$ at each time stage so the solution at the interior points of interest $P_{j+1}(k)$ are not distorted by the arbitrary choice of boundary conditions.

Choice of Space Grid - Δx Parameter

To choose the extent and nature of the discretisation of the space variable x, we must have an a priori estimate of the range of the significant solution, and in a manner depending on the use we have for the solution, accuracy requirements will influence our choice. For a stable control system, the probability density $P(x, t)$ will vanish at values of the space variable x sufficiently far removed from the system's mean value, and we shall apply a truncation condition at "edge" points which are beyond the area of interest, or sufficiently far from the mean value to represent an essentially-zero solution value. Merklinger [1] discusses a transformation which gives a space discretisation with a non-uniformly spaced grid, which tends to concentrate solution points in the region of interest. This involves some extra programming and computation, but seems to improve accuracy in special cases. In this study, we are concentrating on transient analysis of system statistics, and the grid requirements may change during the course of the solution, so optimising the grid via transformations will be impractical in the general case.

To proceed to a specific example, let us set $c_2 = 0.1$ and let the zero frequency noise spectrum $S_y(0) = 2D = 1.0$. We note that the output $X(t)$ will be symmetrically distributed about the mean value $m_x = 0$.* To obtain an a priori estimate of the steady state variance of $X(t)$, we shall use the technique of statistical linearisation. The system

---

* In the steady state

$$\dot{X}(t) = -0.1\, X^3(t) + y_3(t) , \qquad (2.3.10)$$

will be replaced by an equivalent linear system

$$\dot{X}(t) = -k_{eq}\, X(t) + y_3(t) , \qquad (2.3.11)$$

where k_{eq} is an equivalent gain chosen by the method of Booton [32]. This technique assumes that the input to the zero-memory non-linearity ($-0.1\, X^3(t)$ of (2.3.10)) is approximately Gaussian, and essentially minimises the mean square error at the non-linearity's input caused by replacing it with a constant gain. If the input to the non-linearity is $X(t)$, and the resultant output $g(X(t))$, then

$$k_{eq} = \frac{E\,[-X \,.\, g(X)]}{E\,[X^2]} . \qquad (2.3.12)$$

If we assume $X(t)$ is Gaussian with zero mean and variance V_x, then

$$k_{eq} = \frac{0.1\, E\,[X^4]}{V_x} = 0.3\, V_x . \qquad (2.3.13)$$

The variance V_x now, can be found for the linear system (2.3.11) whose system function [33, p.328] is

$$H(j\omega) = \frac{1}{j\omega + k_{eq}} = \frac{1/k_{eq}}{1 + j\,\frac{\omega}{k_{eq}}} . \qquad (2.3.14)$$

We will then assume noise input $y(t)$ is exponentially correlated with cut-off frequency ω_o (the approximation made later (2.3.18) applies equally well to any band-limited signal with cut-off frequency ω_o) so that

$$R_y(\tau) = V_y\, e^{-|\tau|\,\omega_o} , \qquad (2.3.15)$$

and

$$S_y(\omega) = \frac{2\, V_y/\omega_o}{1 + (\frac{\omega}{\omega_o})^2} \qquad (2.3.16)$$

is the power spectral density of the noise $y(t)$.*

The spectrum of $X(t)$ is then given by

$$S_x(\omega) = H(-j\omega)\, S_y(\omega)\, H(j\omega)$$

$$= \frac{2 V_y}{\omega_o k_{eq}^2} \cdot \frac{1}{[1 + (\frac{\omega}{\omega_o})^2][1 + (\frac{\omega}{k_{eq}})^2]} , \qquad (2.3.17)$$

and as we have assumed earlier that $\omega_o \gg k_{eq}$ (as $k_{eq} = O(\tau_{rel}^{-1})$ and $\omega_o = O(\tau_{cor}^{-1})$), the integral of $S_x(\omega)$ will be well approximated by putting $\omega_o = \infty$. Then $X(t)$ corresponds to an exponentially correlated signal

$$R_x(\tau) = \frac{V_y}{\omega_o k_{eq}} e^{-k_{eq}|\tau|} , \qquad (2.3.18)$$

which has a variance $V_x = \frac{V_y}{\omega_o k_{eq}}$. (2.3.19)

Substituting this in (2.3.13) and letting $y_3(t)$ have a unit spectral density at low frequencies (that is, $2 V_y \omega_o^{-1} = 1$ in (2.3.16)), then

$$k_{eq}^2 = (0.3)\, 0.5 = 0.15$$

and $$k_{eq} = 0.4 . \qquad (2.3.20)$$

Then from (2.3.19), we can say that the variance of $X(t)$ will approximately

$$V_x = \frac{0.5}{0.4} = 1.25 . \qquad (2.3.21)$$

---

* The Fourier transforms used in this thesis will be the symmetrical f transform [34, p.66] defined in Chapter 5, although it will often be convenient to use ω as a parameter.

This is the variance of a Gaussian distribution whose tails will extend to about $X = \pm 4$, but it is also an estimate of the variance (as k_{eq} was derived from a mean square criterion) of the non-linear system (2.3.10) which, because of its "hard" cubic feedback, will have an output distribution with much shorter tails than the Gaussian. We shall estimate, then, that the system will hardly ever go beyond $X = \pm 3$, provided the initial conditions specify the system well inside this range. We see later from Figure (2.3.5) that these a priori estimates are quite accurate, which lends weight to the method of statistical linearisation as a rough solution check.

In the present example, we are studying the transient statistics of the system from an initial condition given by a low-variance peaked distribution, and the solution will expand to its steady state variance estimated above. We shall see later that the boundary condition sensitivity is reduced if the solution is near zero at the boundary, and thus we shall choose the boundary at $X = \pm 3$.

The number of points is chosen with regard to the accuracy and resolution desired in the X domain, and in the case of higher dimension problems storage capacity and computing effort will be a limitation. In the present example, $\Delta x = 0.2$ (31 points) was found to give sufficient accuracy for the purpose at hand. If extra accuracy was desired in the initial stages of the transient, the grid could be made adaptive to expand with the solution, but we shall see later that this may accentuate difficulties in regions of a rapidly varying solution. An adaptive grid was tried, but not found efficient, as all the coefficients of the linear equations (2.3.9) had to be recomputed at each time stage. Although one can often make a good a priori guess at the X grid used, it should be emphasized that the finite difference method does not give us any explicit estimate of solution accuracy, and the programmer must always try different values of Δx to obtain confidence in the solution. A good rule of thumb for experimental determination of accuracy is that halving or doubling the grid size will produce a change in solution which is of the order of the accuracy of either solution.

Solution Procedure

Owing to the structure of the left hand side of the linear simultaneous difference equations (2.3.9), when they are written in matrix form the coefficient matrix is tri-diagonal. Because of this special form, where all the matrix elements are zero except those on and directly adjacent to the main diagonal, iterative techniques need not be used to solve the large matrix, and the following elimination type method is very efficient [28, p.101].

Let the solution at time $j + 1$ be expressed as a recursion relation following from right to left across the grid:

$$P_{j+1}(k) = E_k P_{j+1}(k + 1) + G_k , \quad k = 0, K, \qquad (2.3.22)$$

where the recursion coefficients E_k and G_k are obtained by substituting (2.3.22) into (2.3.9) with the k index reduced by one. This has the effect of eliminating $P_{j+1}(k - 1)$ from (2.3.9) and, letting the right hand side of (2.3.9) equal the known quantity D_k, we have

$$P_{j+1}(k) = \frac{\frac{1}{2}h}{1 + h - \frac{1}{2}h E_{k-1}} \cdot P_{j+1}(k+1) + \frac{D_k + \frac{1}{2}h G_{k-1}}{1 + h - \frac{1}{2}h E_{k-1}} . \qquad (2.3.23)$$

Comparing this expression for $P_{j+1}(k)$ with (2.3.22), and equating coefficients, we have the following recursive formula for the E_k, G_k coefficients, the recursion this time travelling from left to right:

$$E_k = \frac{\frac{1}{2}h}{1 + h - \frac{1}{2}h E_{k-1}} , \qquad (2.3.24a)$$

$$G_k = \frac{D_k + \frac{1}{2}h G_{k-1}}{1 + h - \frac{1}{2}h E_{k-1}} , \quad k = 1, K + 1. \qquad (2.3.24b)$$

The form that the boundary conditions must take is now apparent. The E_k, G_k series is determined when E_o and G_o are supplied,

and then the solution $P_{j+1}(k)$ can be obtained when $P_{j+1}(K+1)$ is supplied. A method described below will make the solution quite insensitive to choices of these constants.

Choice of Boundary Conditions

Of the series (2.3.24) for E_k and G_k, we see that G_k depends on the solution $P_j(k)$ by way of D_k, whereas the E_k series is solution-independent, h being a constant. This means that unless other considerations apply (e.g. a symmetry condition on the solution), E_o can be chosen independently, and indeed we note that for all positive h, the series (2.3.24a) converges to a value E_s independent of E_o. This value is found as the lesser solution of the quadratic equation obtained by replacing E_k and E_{k-1} by E_s in (2.3.24a). The limit E_s is well behaved for all positive h, is always between zero and one, and is reproduced in Figure (2.3.2) for convenience.

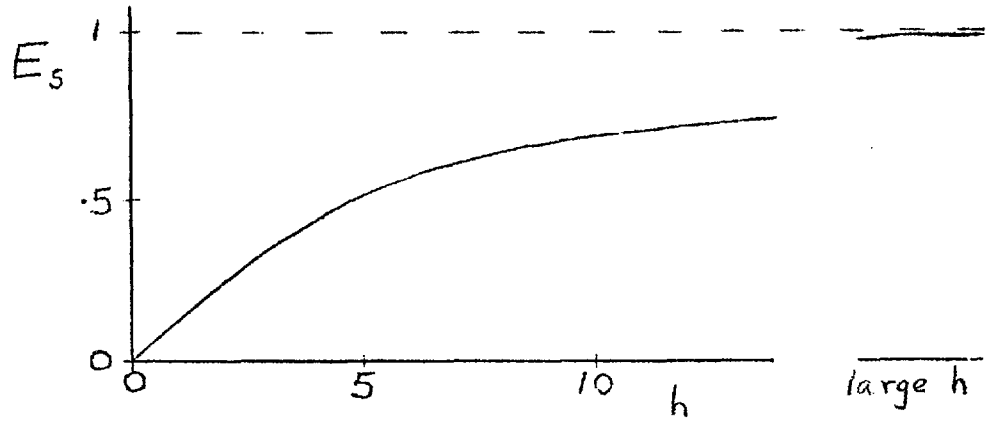


Figure 2.3.2
Variation of limiting boundary condition E_s with h

The suggested approach is to let E_o equal this limit E_s, which will make all $E_k = E_s$. Disregarding the effect of G_o, this operation removes the effect of the left boundary altogether, and was found to work well with the choice of G_o given below. Thus

$$E_k = \frac{1 + h - \sqrt{1 + 2h}}{h} \quad , \quad k = 0, K. \qquad (2.3.25)$$

As the coefficients of the G_k series (2.3.24b) are not constant, we cannot apply a similar treatment to this series, but we note the following points:

(1) The G_k series loses its dependence on G_o as k increases, as from (2.3.24b) we have

$$G_k = \text{function } (h, D_1, \ldots D_k) + \left[\frac{\frac{1}{2}h}{1 + h - \frac{1}{2}h E_s} \right]^k G_o \,, \qquad (2.3.26)$$

and the quantity in braces will always be less than one. For convenience this quantity is reproduced in Figure (2.3.3) below.

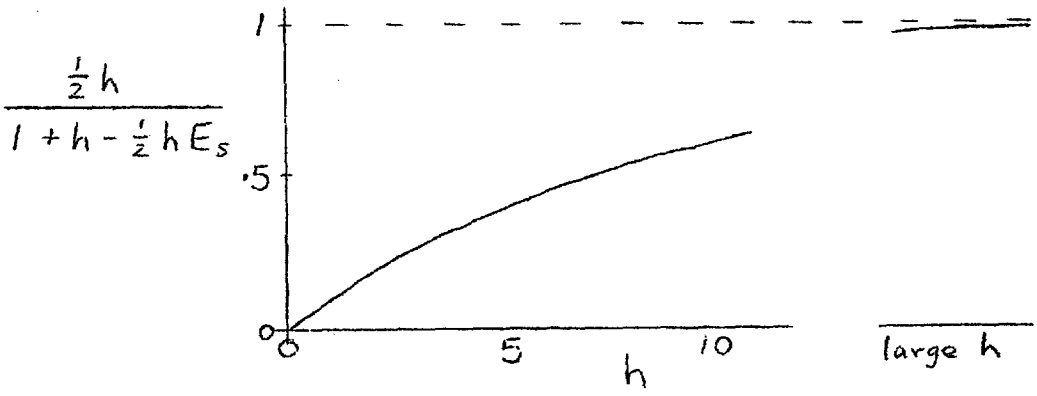


Figure 2.3.3
Variation of a parameter of G_k series with h

We see that for values of h used in our example ($h < 5$, say) G_k loses its dependence on G_o very quickly, and at most only one or two points of the solution adjacent to the left boundary $P_{j+1}(1)$ or $P_{j+1}(2)$, will be affected by a wrong choice of G_o.

(2) Owing to the relative magnitudes of the solution at the edge and at the centre of the grid, it turns out that G_o is much less than G_k near the centre of the grid. This fact lends weight

to the argument of point (1) above, and as the solution $P_{j+1}(0)$ will be very small, we shall set G_o to an arbitrary small constant. This can be done in a fashion similar to a semi-permeable boundary by setting

$$G_o = \alpha G_1 \quad , \quad o \leqslant \alpha < 1, \tag{2.3.27}$$

where the G_1 is taken from the last time stage. This procedure is useful during the transient solution as it allows G_o to change in a stable manner. In practice, α is quite small, 0.1 or 0.2, and the solution $P_{j+1}(1)$ depended little on this choice of α.

(3) The technique of a dual solution outlined below will further reduce the solution sensitivity to G_o.

Dual Solution

The choice of $P_{j+1}(K + 1)$ at the right hand boundary is somewhat more critical as it has a more direct effect of the solution adjacent to it. As $E_s < 1$, errors will not be propagated too far, but $P_{j+1}(K)$ may have a significant error caused by the choice of $P_{j+1}(K + 1)$. There is also the need to allow $P_{j+1}(K + 1)$ to change during the transient solution. As this boundary condition is more critical, we will avoid expressing $P_{j+1}(K + 1)$ as a factor of $P_j(K)$, the semi-permeable boundary method, but we note that the solution procedure outlined above computes $P_j(0)$ with reasonable accuracy (provided E_s is not too small or G_o is chosen reasonably). A dual method for solving the tri-diagonal matrix is proposed which will compute $P_j(K + 1)$ with similar accuracy.

The right to left recursion series (2.3.22) will be supplemented by a left to right recursion:

$$P'_{j+1}(k) = E'_k P'_{j+1}(k - 1) + G'_k \quad , \; k = 1, K + 1. \tag{2.3.28}$$

Paralleling the relations (2.3.24) we find

$$E'_k = \frac{\frac{1}{2}h}{1 + h - \frac{1}{2}h E'_{k+1}} , \qquad (2.3.29a)$$

$$G'_k = \frac{D_k + \frac{1}{2}h G'_{k+1}}{1 + h - \frac{1}{2}h E'_{k+1}} , \quad k = 0, K. \qquad (2.3.29b)$$

The same remarks apply to these series, that is, E'_k has a steady state solution $E'_s = E_s$, and G'_{k+1} must be chosen arbitrarily (expressed as a factor α of G'_k of the previous stage).

The solution at time $j + 1$ is now taken as an average of the two dual solutions (2.3.23) and (2.3.28). The edge probabilities $P_{j+1}(K + 1)$ and $P'_{j+1}(0)$ are taken as the corresponding values of the <u>averaged</u> solution at time j. This procedure was found to greatly minimise the errors occurring at the boundary and allows the boundary values to adjust themselves in a stable manner during the transient solution.

This method of dual solution has the added advantage that any errors caused by the application of arbitrary boundary conditions will be symmetrically distributed through the solution, an important factor if an accurate estimation of the solution mean value (or other odd-order moments) is desired.

The effect of this method of applying space truncation conditions was checked in the example to follow, by truncating at several points further in from the original boundary, and noting the change in the shape of the residual distribution. This shape was found to be virtually unchanged even for truncation points very near the centre of the distribution, but in this case the normalization operation described below was difficult to perform unless some information were available about the truncated tails of the distribution. Thus the method of applying boundary conditions via the dual solution was felt to be very successful.

Normalisation

A property of the probability density function $P(x, t)$ is that its integral over the space x must be unity at all times. The continuous Fokker-Planck equation preserves this property, but the effect of discretization and boundary condition errors in the numerical solution is to make the integral drift away from unity, by a small amount at each time stage. It is therefore expedient to normalize the distribution at each stage in the solution using the rectangular integration rule, multiplying $P_j(k)$, $k = 0, K + 1$, by a constant factor to make

$$\sum_{k} P_j(k) \cdot \Delta x = 1 . \qquad (2.3.30)$$

The accuracy of this operation depends on Δx and on the amount of distribution lying outside the space grid, but an error in this operation will only alter the P scale and not change the fundamental shape of the solution. In the procedure described earlier, we cannot use this normalisation condition to arbitrarily choose a boundary condition and compensate by normalising, as an arbitrary boundary condition will alter the shape of the solution, which normalising cannot correct.

An important feature of the normalising operation is that the normalising factor used is an excellent indication of the average size of discretisation and boundary condition errors. If there were no errors, the factor would be unity, and the percentage difference from unity gives an average indication of the relative size of the numerical errors introduced at each solution stage. In the example to follow, this normalising factor was kept to 0.001 of unity.

2.3.4 An Application of the Transient Solution

As an example of the use of the FP equation for dynamic prediction in noisy non-linear systems, an adaptive sampling and alarm scheme will be designed.

Consider the system of Figure (2.3.1), with the parameter values given earlier, as a regulating system to keep the system output $X(t)$ at zero in the face of the noise disturbance $y_3(t)$. Suppose the system becomes dangerously overloaded if $X(t)$ exceeds $\pm$ 2, and we wish to estimate $X(t)$ to determine the extent of this danger. Suppose that measurements of $X(t)$ are difficult or costly to make, and so we will minimise the number of measurements taken by using the FP equation to predict the system's output between samples, which will determine when the next sample should be taken.

We recall that the solution of the Fokker-Planck equation is the system's output probability density function, evolving with time, conditional on an initial value or distribution being given. Thus, unless prior information is to be considered, the measurement gives the initial distribution $P(x, 0)$ and the solution predicts future system probabilities $P(x, t)$. The measurement could be exact, $P(x, 0)$ being a delta function, or could contain some known error or expected deviation. A common model of sampling statistics is to assume a normal error distribution so that the initial probability distribution is given by

$$P(x, 0) = N(m_x, n_x^2) , \qquad (2.3.31)$$

where the right hand side is the normal distribution with mean m_x, the measured value, and standard deviation n_x, the expected measurement deviation.

Following each measurement, the Fokker-Planck equation is solved, and the probability that the system's output exceeds the given bounds is found. In particular, the time at which the probability exceeds a limit, say 5 per cent, is found, and it is proposed to make another measurement when this time has passed. This "safe time to go" can be computed for a representative set of measurement conditions and

values, and the sampling mechanism could be pre-programmed to adjust its sampling period depending on the most recent measurement. If the measured value at any instant exceeds the given bounds, or is such that the safety time is less than the minimum practical sampling period, the alarm could be given to initiate alternate control action or safety measures.

Choice of Δt

We have left the choice of Δt until this stage as it depends to a certain extent on the use we will make of the transient solution. If only the steady state solution is of interest, the choice of Δt does not affect accuracy but only the rate of convergence to the steady state solution, as the solution of the parabolic equation (transient FP equation) in the manner described in this section can be thought of as an iterative method of solving the associated elliptic type equation obtained by setting P_t equal to zero (the steady state FP equation). When the transient solution is of interest, there are two main considerations, the resolution and accuracy desired in the time step variable, and the prevention (if necessary) of oscillations caused by a solution which is rapidly changing in the x direction.

For the time resolution consideration, we want an estimate of the speed of response of the system. For this the earlier linear analysis is useful where we found the first order system had an equivalent gain $k_{eq} = 0.4$ (2.3.20) with which we can associate the relaxation time $\tau_{rel} = k_{eq}^{-1} = 2.5$. By analogy with the linear system's time constant, we can deduce that a useful increment over which to obtain the transient solution would be of the order of $0.1\ \tau_{rel}$ or about $\Delta t = 0.2$. This estimate is meant only as an initial guide, as Δt must always be varied to test solution accuracy, as no explicity estimate of accuracy is available. All we can say about accuracy with respect to Δt is that the Δt and Δx discretisation errors vary with $(\Delta t)^2$ and $(\Delta x)^2$ for the Crank-Nicholson formula. This gives us an estimate of the effect of refining the Δt or Δx grid, but does not allow us to separate the errors due

to Δt and Δx. Thus we must resort to numerical experimentation to obtain confidence in the accuracy of the solution.

To observe the effect of Δt on a rapidly varying solution, let us look at a typical solution of our regulator example. Considering the case of a perfect measurement $X = 0$ we use the initial distribution $P(x, o) = \delta(o)$. Choosing the parameters $\Delta t = 0.2$, $\Delta x = 0.2$ we find that the solution $P(x, 0.2)$ dips below zero at $x = 0$. This occurs because of the poor discrete representation of $\frac{\partial^2 P}{\partial x^2}$ and $\frac{\partial P}{\partial x}$ by formula (2.3.8) when $P(x, o)$ has the sharp discontinuity represented by

$$P_o(k') = \frac{1}{\Delta x} \quad , \quad k' \text{ at centre of grid,}$$

$$\text{and} \qquad P_o(k) = 0 \qquad \text{elsewhere.} \qquad (2.3.32)$$

This leads to the discrete values of the space derivatives

k	k' - 2	k' - 1	k'	k' + 1	k' + 2
$\frac{\partial P_o(k)}{\partial x}$	0	12.5	0	-12.5	0
$\frac{\partial^2 P_o(k)}{\partial x^2}$	0	125	-250	125	0

which clearly do not form an adequate representation of the properties of a delta function. This leads to a negative value for $P_1(k')$ and overestimated values for $P_1(k' - 1)$ and $P_1(k' + 1)$. Furthermore, at the next time stage we find $P_2(k')$ is overestimated, with $P_2(k' - 1)$ and $P_2(k' + 1)$ underestimated. This oscillation is stable, in the sense that it eventually dies out, but may be undesirable if an accurate solution is needed during the initial time stages.

To reduce the size of this oscillation we note that the change in the solution at a discrete point k over one time step is proportional to Δt and has contributions proportional to $(\Delta x)^{-1}$

and $(\Delta x)^{-2}$. To reduce the oscillations, the solution change must be lessened by reducing Δt or increasing Δx. In the solution procedure outlined for parabolic equations, it is easier to alter Δt, and so Δx is usually not changed during the course of a solution unless an awkward size of Δt occurs.

Figure (2.3.4) shows the transient solution for $\Delta t = 0.1$, $\Delta x = 0.2$. Here the solution point $P_1(k')$ has not gone negative and the oscillation is quite small by the third time step, and negligible by the fifth. It is interesting that this initial oscillation has no observable effect on the smoothness of the variance curve Figure (2.3.7), and so may not be undesirable in some cases. Also, the oscillation has no effect on solution accuracy once the oscillation has died away, and so in most cases

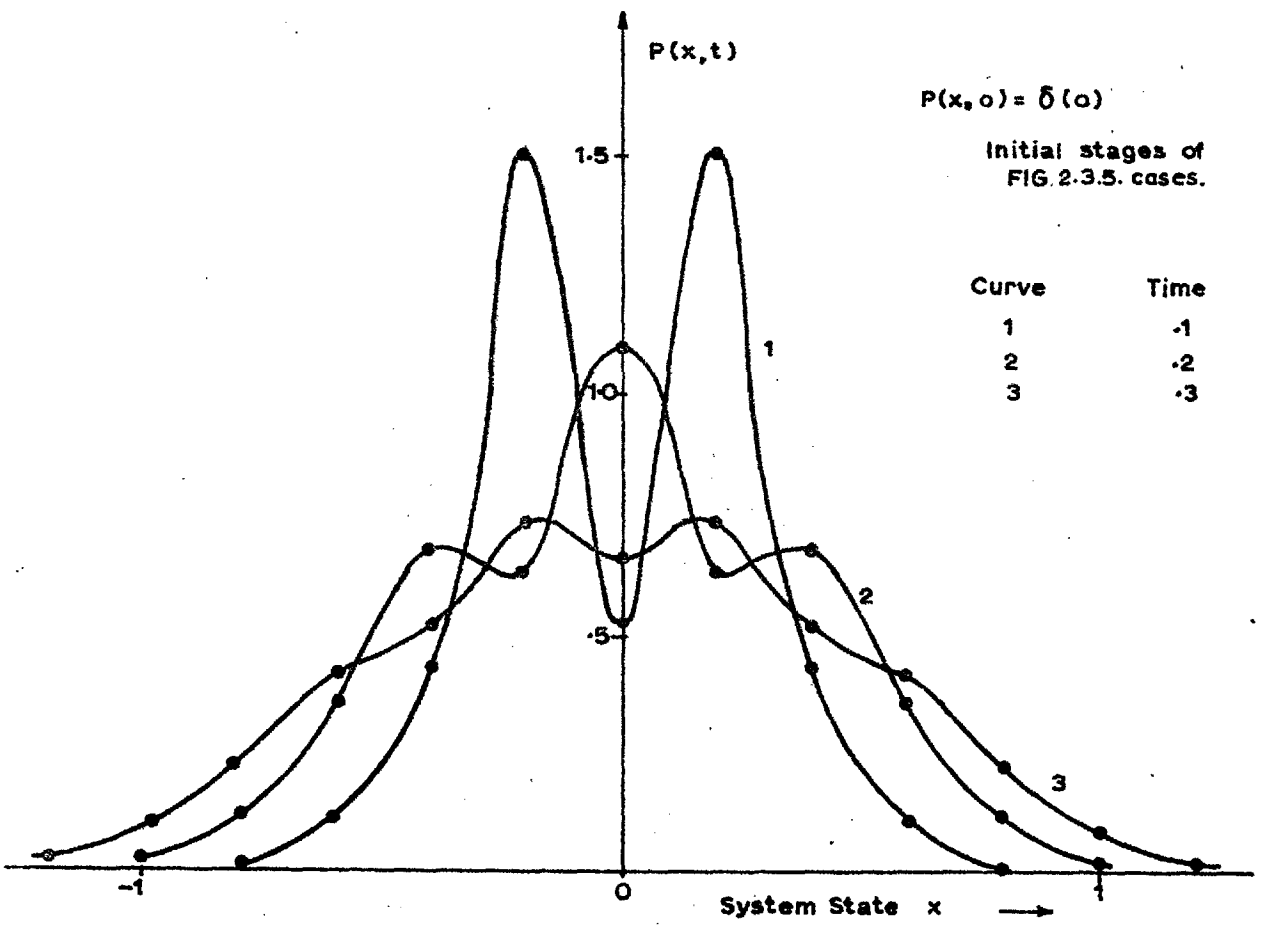


Figure (2.3.4) Transient solution from a delta function

it will not be necessary to take a great deal of care in reducing the oscillation.

Typical Solutions

To determine the growth of uncertainty of the variable $X(t)$ following a measurement, the FP equation (2.3.9) was solved for a set of initial conditions corresponding to measurements of $X(o) = m_x = 0, 0.4, 0.8, 1.2$ and 1.6, each with a measurement noise of standard deviation $n_x = 0, 0.3, 0.6$ and 0.9. For the present purpose, solutions for negative measurements are not necessary, as the space symmetry of the example about $X = 0$ make the negative measurements equivalent to the reflected positive ones. Typical solutions are shown in Figures (2.3.5, 6), the curves representing the evolution of the probability density function with time.

Figure (2.3.5) corresponds to a noiseless measurement of $m_x = 0$ giving a delta function at $x = 0$ (shown as coinciding with the vertical axis of the graph). The solution is shown to spread out at $t = 0.5$ and again at $t = 2.0$, showing the increased uncertainty of the whereabouts of the system caused by the input noise. A measure of this uncertainty is the probability that the system is outside the range (-2, 2) shown shaded in the graph. This corresponds to the tolerance band mentioned above and the probability that the system output exceeds this tolerance is computed and shown in Figure (2.3.8) for the cases shown in Figure (2.3.5) and Figure (2.3.6). Also shown in Figure (2.3.5) (dotted line) is a linear Gaussian distribution of the same mean and variance as the solution at $t = 2.0$. This shows the non-linear character of the system which exhibits a large force constant for large deviations from zero, and a weak force near the origin.

Figure (2.3.6) shows the solution for the case of a noisy initial measurement, the measurement having a non-zero value of $m_x = 0.8$. Thus the initial distribution is centered about $x = 0.8$, and is Gaussian with a standard deviation, given by the measurement noise, of 0.3. It is noticed that the symmetry of the solution is no longer preserved, another indication of the non-linearity of the system. It was found that for moderate measurement noise, the time scale of the solution is little affected by this noise, for the solution tends to expand very rapidly from a delta function initial condition in any case. For example, in the Figure (2.3.5) case, the solution had expanded to as large a standard deviation by $t = 0.1$ as the initial distribution in the Figure (2.3.6) case (see Figure (2.3.7)). This is to be expected as the initial growth is very rapid, being caused by a relatively high diffusive force acting on the system.

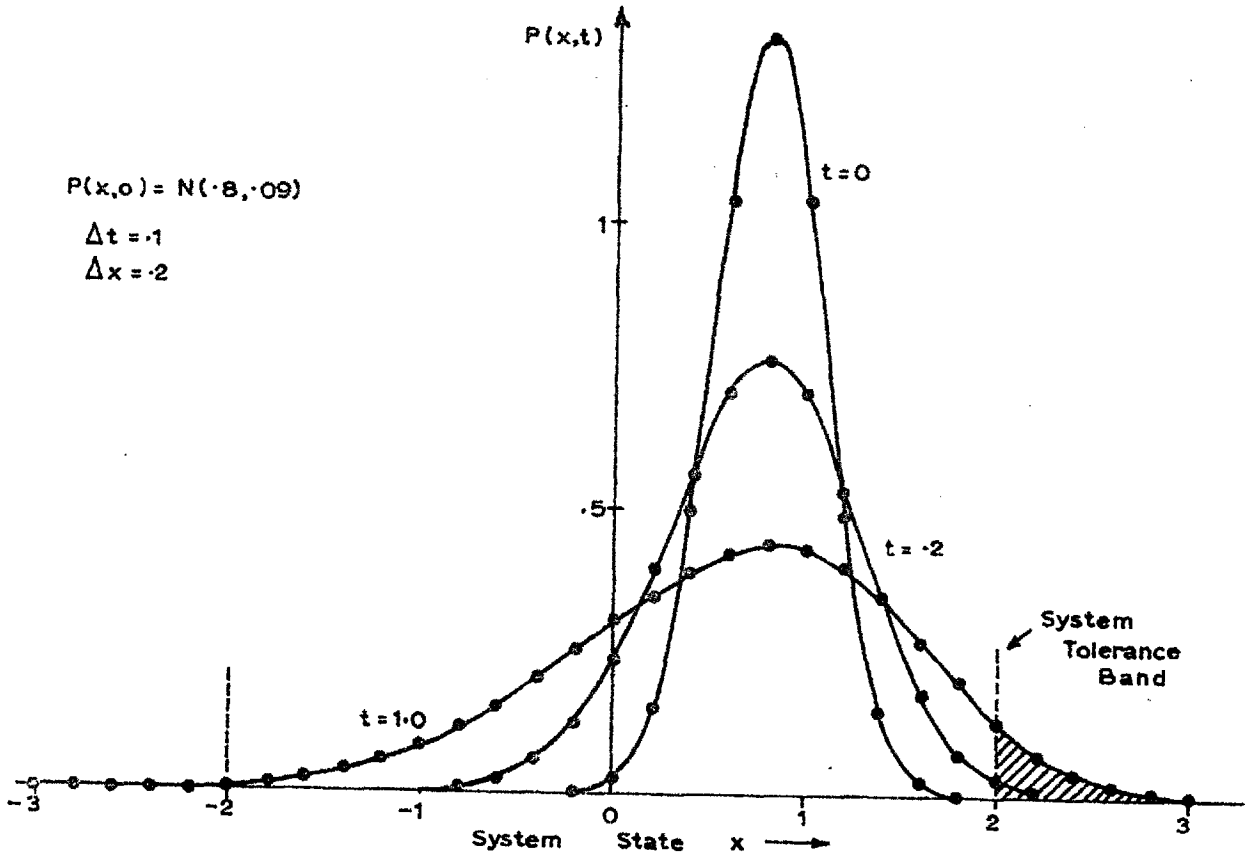


Figure (2.3.6)

Time solution of Fokker-Planck equation, noisy non-zero measurement

As discussed above, the parameter values at $\Delta t = 0.1$ and $\Delta x = 0.2$ were considered reasonable for the present application. By refining the finite difference mesh size, the accuracy was found to be better than 1% in root mean square, the errors not being large enough to be visible in Figures (2.3.5, 6).

From this numerical representation of the solution, it is easy to calculate functions of the system's probability density function. An example is shown in Figure (2.3.7), being the variance of the distribution for the solutions shown in Figures (2.3.5, 6). The noiseless measurement case begins with zero variance and the noisy measurement case begins with the measurement noise variance (0.09) and in each case the growth is smooth with time. That the growth pattern is somewhat different in each case, however, is further indication of the non-linearity of the control system.

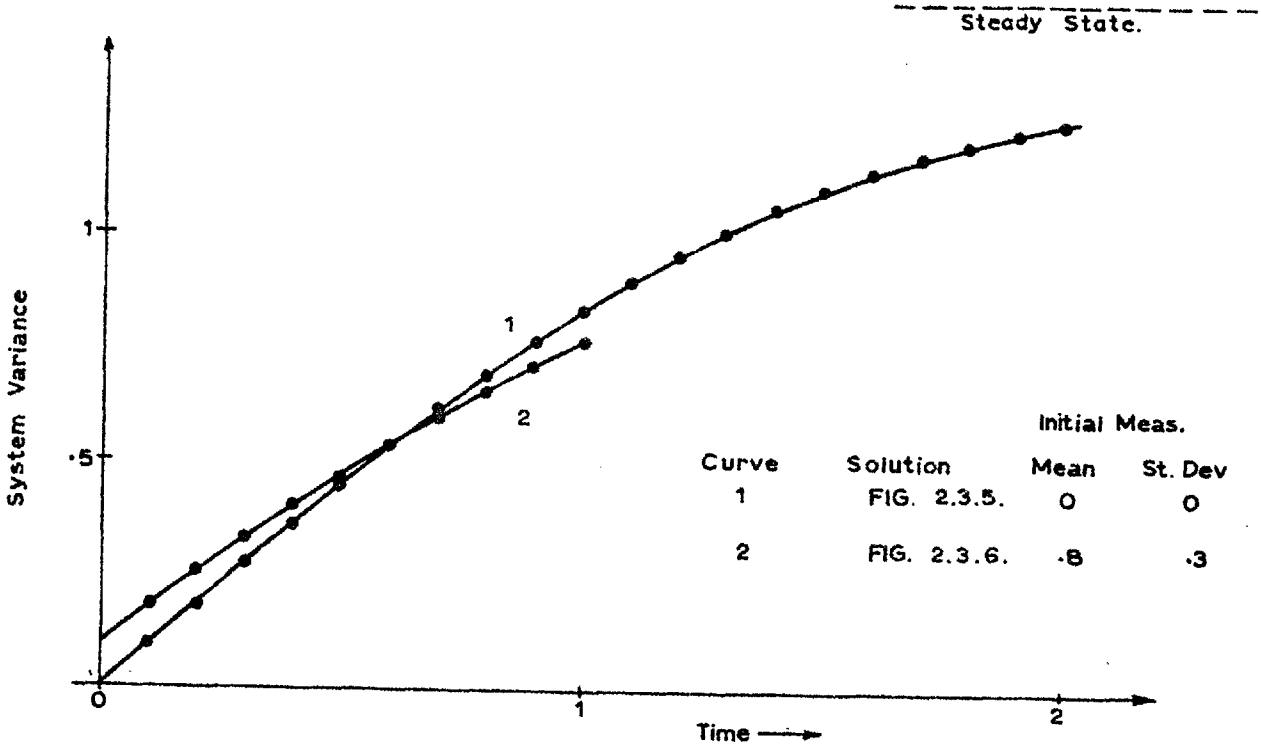


Figure (2.3.7)

Growth of system uncertainty with time (output variance).

Design of Adaptive Sampling Scheme

The adaptive sampling scheme will be designed so that when the probability that the system's output X(t) exceeds $\pm$ 2, given the last measurement, is greater than 5%, another measurement will be taken. This time between samples depends only on the last measurement and the known measurement error, as illustrated in Figures (2.3.8, 9).

To obtain the probability that the system state exceeds $\pm$ 2 as a function of time, the appropriate section of the solution of the transient FP equation must be integrated, as shown in the shaded portions of Figures (2.3.5, 6). The resultant function is shown in Figure (2.3.8) for the two measurement cases of Figures (2.3.5, 6). The accuracy of this function depends on the number of solution points outside the tolerance band, especially as the two outside grid points have the lowest accuracy. In the example, one third of the points were outside the $\pm$ 2 band, and the curves of Figure (2.3.8) were obtained to 2% of scale accuracy.

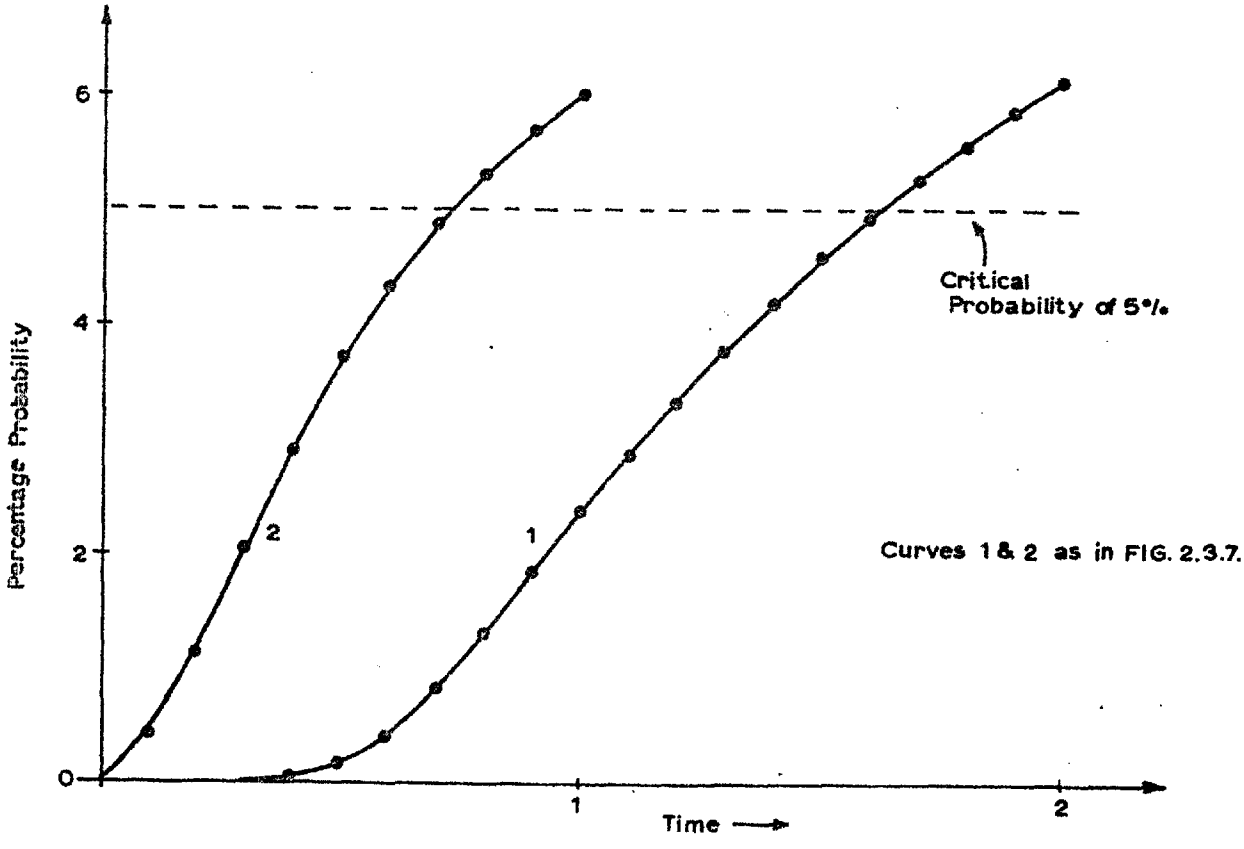


Figure (2.3.8)

Probability that system exceeds tolerance band of $\pm$ 2

From Figure (2.3.8) we obtain the time from the last measurement that the system can run before the probability of the system output X(t) exceeding $\pm$ 2 goes beyond the critical level of 5%. This time has the same units as the time in the derivative of the dynamic system (2.3.2). As most instruments will have a constant measurement error deviation, it is convenient to keep this error deviation n_x as a parameter, and plot the elapsed time to the next measurement as a function of the value of the last measurement m_x.

This is shown in Figure (2.3.9), and one, or a family, of these curves could be used in conjunction with a timer and incorporated into the measuring apparatus. Alternatively, if an on-line digital computer is used on the process, these curves could be stored and used in conjunction with the computer's priority interrupt system to

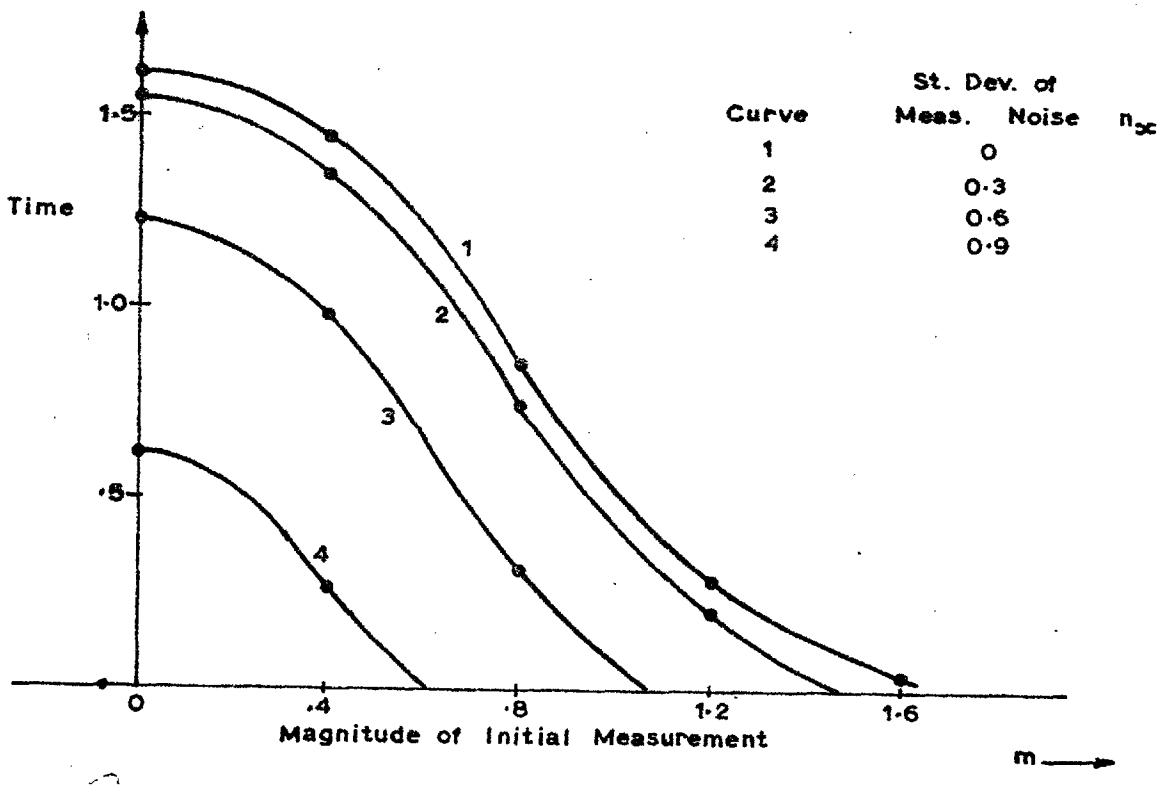


Figure (2.3.9)

Time until Pr(|X| > 2) exceeds 5%

time the next measurement. As a further refinement, if the on-line computer is updating the system model, the FP equation can be solved after each measurement using the best current estimate of the system dynamics (2.3.2).

The curves of Figure (2.3.9) bring out one interesting and important point which is probably valid for a wide class of noisy systems. They show that for prediction purposes in stochastic control systems there is no advantage to using elaborate and expensive techniques to obtain accurate measurements, for the added information obtained is soon swamped by the future uncertainty caused by noise in the system. This is shown in Figure (2.3.9) by the fact that the safety time indicated by a measurement with moderate noise (0.3) is very little different from that obtained with a noiseless instrument. For larger measurement noises of 0.6 and 0.9, this difference is appreciable, and a reasonable instrument accuracy can be chosen on a basis of the system dynamics and the magnitude of noise in the system.

2.3.5 Solution for Higher Order Systems

A solution method using finite differences has been presented which is a reasonably accurate and efficient one for the one-dimensional problem. The numerical analysis and stability of the one dimension equation is well treated in the literature, but the infinite space range of the FP equation presented special difficulties in applying the appropriate boundary conditions. It was found that by applying arbitrary properties of the solution, an empirical method gave satisfactory space truncation properties. It was also noted that solution accuracy depended on the system non-linearity being in a form amenable to finite difference representation.

The numerical analysis of parabolic equations in more than one space variable is not as well known, and the literature is usually confined to examining a few simple examples. The FP equation (1.2.1) will have as many first order derivatives $\partial P/\partial x_i$ as states x_i of the dynamic system considered, and as many second

order derivatives as there are states x_i directly excited by the noise vector, given by the existing elements of the $n \times m$ matrix $F\,F^T(x, t)$ (2.1.3). A simplification is gained if the second order derivatives do not involve all the space variables as will be seen from the two dimensional examples given below.

Two-dimensional Example : Case 1

Consider the noisy system represented by the diffusion process (2.1.1) of the form

$$dx_1(t) = f_1(x_1, x_2, t)\,dt$$

$$dx_2(t) = f_2(x_1, x_2, t)\,dt + F_{22}(x_1, x_2, t)\,dw_2(t). \qquad (2.3.33)$$

Here $dw_1(t) = 0$ and there is only a single noise input $dw_2(t)$. Dropping the x_1, x_2, t parameters, this system has the FP equation

$$\frac{\partial P}{\partial t} = -\frac{\partial}{\partial x_1}[f_1\,P] - \frac{\partial}{\partial x_2}[f_2\,P] + \tfrac{1}{2}\frac{\partial^2}{\partial x_2^2}[F_{22}^2\,P], \qquad (2.3.34)$$

which only has a single second derivative term as opposed to two first derivative terms $\partial P/\partial x_1$ and $\partial P/\partial x_2$. The same situation would exist if a second noise term $F_{21}\,dw_1$ were added to the dx_2 equation of (2.3.33) only (that is, F_{11} and F_{12} remaining zero), as the 2×2 matrix $F\,F^T$ only has a non-zero (2,2) element.

The numerical analysis of an equation of this form has not been found in the literature. The equation (2.3.34) is parabolic in the x_2 variable, but has a term similar to those in hyperbolic equations in the x_1 variable. Basically there are two possible approaches to this problem:

(1) assume a second derivative term in x_1 exists and treat the equation as a parabolic equation in two space variables, as in Case 2 below, or

(2) treat the problem primarily as a parabolic equation in one space variable, as in Section 2.3.2, 3, and then choose the

free finite difference model parameters to assure overall stability.

Preliminary enquiries [35] have indicated that there is no theoretical justification to prefer either approach, and the latter was used on an example of the form (2.3.33), as it is simpler to implement than the first approach. The implicit Crank-Nicholson scheme was used on the x_2 variable as in equation (2.3.9) where the known variables on the right hand side now include a finite difference approximation to the first derivative $\partial P/\partial x_1$ in the x_1 variable. The implicit scheme means that no restriction is placed on the size of Δt on account of the $\partial^2 P/\partial x_2^2$ term, and the finite difference model parameters can be chosen to assume stability with respect to the hyperbolic terms. The finite difference formula used should be shown to satisfy the necessary conditions for stability of von Neumann [28, p.59], which are in practice usually sufficient conditions as well.

The above procedure means we have a tri-diagonal matrix to solve for each <u>row</u> of the two dimensional finite difference grid (that is, a row for each discrete value of x_1). This was done using the solution procedure of Section 2.3.3, the boundary values at the extremes of the x_2 grid being found by the methods of that section. The solution for the two extreme rows required boundary conditions in the x_1 variable, on account of the presence of the $\partial P/\partial x_1$ term. The choice of these conditions was not found to be critical, and a simple reflection coefficient method was satisfactory.

Although the above method worked satisfactorily, it is clear that considerably more effort was needed than the one dimensional example. The number of solution points and finite difference model constants (2.3.9a, b, c) to be stored was the square of those stored in the one-dimensional example, and computing time was increased by this factor as well. In addition, the choice of the finite difference model parameters and boundary conditions were more difficult (and somewhat more empirical) than earlier, and the computer programming much more involved.

Two-dimensional Example : Case 2

Consider now the more general example of the diffusion process

$$dx_1(t) = f_1(x_1, x_2, t)dt + F_{11}(x_1, x_2, t)dw_1(t) + F_{12}(x_1, x_2, t)dw_2(t),$$

$$dx_2(t) = f_2(x_1, x_2, t)dt + F_{21}(x_1, x_2, t)dw_1(t) + F_{22}(x_1, x_2, t)dw_2(t),$$

(2.3.35)

where each component of the state vector x is now disturbed directly by noise. The associated FP equation is now

$$\frac{\partial P}{\partial t} = -\frac{\partial}{\partial x_1}[f_1 P] - \frac{\partial}{\partial x_2}[f_2 P] + \tfrac{1}{2}\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{\partial^2}{\partial x_i \partial x_j}[(F F^T)_{ij} P],$$

(2.3.36)

which contains, in general, three second derivative terms (as $\partial^2/\partial x_1 \partial x_2$ and $\partial^2/\partial x_2 \partial x_1$ are equivalent). This is the two space dimensional parabolic equation which has been treated in the literature, with the exception that the constant and first derivative terms of (2.3.36) usually do not appear in quoted examples.

For this equation, a direct extension of the Crank-Nicholson implicit method to all second order derivatives is possible, but this would result in linear equations to solve with a matrix containing elements off the three diagonals of the one-dimensional example or Case 1 above. The iterative routine needed to solve this large matrix is much more time-consuming than the method described below which maintains the tri-diagonal form of the linear equation matrix.

The suggested solution procedure is called the alternating-direction method developed by Peaceman and Rachford [36] and Douglas [37]. In this method, the second derivatives of (2.3.36) are made implicit in one direction only (say the x_1 direction), while the derivatives in the other direction (the x_2 direction) are explicitly expressed in terms of known values of the solution. Then the simultaneous equations involved have a tri-diagonal matrix

as before and can be solved easily without iteration. The procedure must then be repeated at the next time step of equal size, with the difference equations implicit in the x_2 direction, but explicit in the x_1 direction, and the overall procedure for the two time is stable for any size of time step Δt.

The computation time for this alternating-direction method is the same as for Case 1 above, but the programming associated with the implementation of the alternating directions and the boundary conditions is somewhat more complicated. A modification of the alternating-direction method [38] allows parabolic equations in three space variables to be solved by the implicit method, but this was not attempted.

2.3.6 Summary of Finite Difference Methods

In Sections 2.3.2, 3 and 4 we have outlined an implicit finite difference method of solving the FP equation of a noisy control system of one dimension and presented an example. Although the numerical method involved was well treated in the literature, each problem encountered usually presents special difficulties, and some of these have been discussed for a typical FP equation. Few specific details have been given about the choice of the finite difference model parameters and the resultant accuracy obtained, as for any particular application, these will have to be experimentally adjusted to arrive at a suitable compromise involving accuracy, storage capacity and computing time.

The choice of boundary conditions was the problem most unique to the FP equation, where in most cases, no natural boundary presents itself. Arbitrary conditions must be imposed to make the artificial boundary as unobtrusive as possible. It was found that for best results the boundary should be placed where the solution is smooth and small. This is a reasonable prospect for stable regulatory-type control systems, and a method was described to achieve a boundary which did not adversely affect solution accuracy near the boundary, particularly when the boundary values were time-varying.

In Section 2.3.5 it was shown how a modified version of the one-dimensional implicit scheme could be used to solve two and three-dimensional problems. Although a few two-dimensional examples were solved, they were not presented, as they involved essentially the same methods as the one-dimension example.

The purpose of this section has not been to illustrate new numerical analysis techniques, but to show some of the modifications which must be applied to standard techniques when specific examples are tackled. For example, the dual solution procedure and normalisation operation were found particularly helpful in solving the FP type of parabolic equation. To the uninitiated, the numerical analysis literature concerned with partial differential equations is rather formidable. Recent articles treat the simplest examples by methods of increasing subtlety, while the engineer is left with the older methods to solve his more complicated examples. That is, he becomes involved in the "folklore" of the art, which never appears in printed form, and finds himself developing special routines to meet the specific problem at hand. This is essentially what we have done in this section.

Although a three-dimensional example was not attempted, this scale of problem would just be feasible on a modern, high-speed, large-memory digital computer. At this point, however, we seem to reach the limit of numerical analysis knowledge and digital computer capacity [39], and we shall have to seek alternative methods for higher dimensional problems. Recent developments in the application of hybrid computers to the solution of partial differential equations [40] seem to afford an interesting alternative to digital computer methods. However, the manner in which the equation dimensionality affects the speed and storage requirements is the same as before, and it seems that the scale of problem amenable to solution on hybrid computers is limited to the same order of magnitude as on digital computers.

2.4 Numerical Solution by Hermite Transforms

2.4.1 Characterization of a Random Process

When considering alternative methods of obtaining a numerical solution for the statistical description of the vector system $x(t)$ of (2.1.1), it is fundamental to enquire into the ways of expressing this statistical description, in the hope of finding a characterization which is more convenient for computing purposes.

The characterization used in the FP equation is the first order probability density function $P(x(t), t)$, and if the initial conditions are a delta function, the solution of the FP equation gives the second order transition probabilities $P(x(t),t \mid x(t_o) = x_o)$. The complete characterization of the random process $x(t)$ requires the infinite order probability density function (p.d.f.) $P(x(t_1), x(t_2), \ldots ; t_1, t_2 \ldots)$ for all $t_1, t_2 \ldots$ in the domain of the process $x(t)$. If the process is stationary, one of these time parameters t_i may be removed, and the rest replaced by time differences referred to this arbitrary parameter t_i. Clearly a p.d.f. of arbitrarily large order can be reconstructed from a sufficiently large set of solutions of the FP equation, but it will not be very convenient to do this. There is no real impetus to construct large order p.d.f.'s, as the added information gained about the process $x(t)$ by increasing the order diminishes rapidly with the order. Indeed, if the process $x(t)$ is Gaussian, then the second order p.d.f. completely describes the process. Further, for many applications we will only want the information of the first order p.d.f. $P(x(t), t)$ for all t of interest.

Let us only consider the first order p.d.f. and look at different ways of writing it. The characteristic function is often used, and is the Fourier Transform of the probability density function. The characteristic function is defined for p.d.f.'s of all orders, but again only the low order ones will have significant physical meaning. The characteristic function has the advantage that it contains the moments of x in a convenient form (as coefficients of a power series). Thus the moments of x form a representation of the

random process, but corresponding to the first order characteristic function, there is an infinite set of moments. For the vector process $x = [x_1, x_2, \ldots x_n]$ there are n or nC_1* first moments $m_1(x_i) = E[x_i]$, nC_2 second moments $m_2(x_i, x_j) = E[x_i\, x_j]$ etc. Although the significance of the moments m_i decreases as i increases, their magnitudes usually do not diminish very rapidly, if at all, and many of them will be needed to form a good reconstruction of the characteristic function.

An equivalent description to the moments are the cumulants which are the coefficients of a power series for the logarithm of the characteristic function. The cumulants have the advantage over the moments in that, if we are only going to compute the first n of them, we can set the higher ones to zero with less consequence to the convergence or accuracy of the series for the characteristic function. Indeed, if the process x has a Gaussian distribution, then only the first two cumulants are non-zero, and thus need be computed. Also, if the distribution of x is near Gaussian, then only a small number of the first n cumulants will give an adequate description of the process x. Clearly this is very desirable from the computing point of view, but we can go one step better, for the probability density function itself** can be expressed as a series whose coefficients are rapidly diminishing (and the representation rapidly converging) for distributions which are near Gaussian.

This representation is the Gram-Charlier Type A Series or the orthogonal polynomial series of Hermite, as discussed by Cramer [51, p.131 and 221]. In the presentation and example to follow, we will assume that the system $x(t)$ is one-dimensional. Although the presentation becomes more involved, there is no fundamental reason why the procedure of this section cannot be applied to multi-dimensional systems (see [52] and [53] for a discussion of

---

* nC_i is the combination operator "n choose i"

** This is advantageous, as we will usually want the system's statistical description in p.d.f. form in the end, and it is often difficult to obtain the p.d.f. from the characteristic function, particularly if the latter has only been obtained in approximate form.

multi-dimensional Hermite polynomials. In an interesting recent paper, Kolosov and Stratonovich [58] illustrate an approximate solution to the optimal stochastic control problem for a two-dimensional example, by expanding the loss function in two-dimensional Hermite polynomials.)

2.4.2 Hermite Polynomial Expansions

Consider the r:th Hermite polynomial given by

$$H_r(x) = (-1)^r e^{\frac{1}{2}x^2} \frac{d^r}{dx^r} (e^{-\frac{1}{2}x^2}), \quad r = 0, 1, 2, \ldots \qquad (2.4.1)$$

which satisfy the orthogonality relationships

$$\int_{-\infty}^{\infty} H_r(x) H_s(x) G(x) dx = \begin{cases} r! & , r = s, \\ 0 & , r \neq s. \end{cases} \qquad (2.4.2)$$

The quantity $$G(x) = (2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}x^2} \qquad (2.4.3)$$

in the integral (2.4.2) is called the base of the orthogonal polynomials, and is seen to be the one-dimensional zero-mean, unit variance Gaussian distribution.* $H_r(x)$ is a polynomial of degree r, and the first few are given below:

$$\begin{aligned} H_0(x) &= 1 \\ H_1(x) &= x \\ H_2(x) &= x^2 - 1 \\ H_3(x) &= x^3 - 3x \\ H_4(x) &= x^4 - 6x^2 + 3 \\ H_5(x) &= x^5 - 10x^3 + 15x \end{aligned} \qquad (2.4.4)$$

---

* Hermite's original definition used the base e^{-x^2}, but we will prefer this form when considering expansions related to the Gaussian distribution.

These can be obtained from the recursion formula

$$H_{r+1}(x) = x\,H_r(x) - r\,H_{r-1}(x) \quad , \quad r \geqslant 1 \tag{2.4.5}$$

which is obtained later (2.4.26c).

Consider the expansion of the probability density function $P(x, t)$ in the infinite function series

$$P(x, t) = \sum_{r=0}^{\infty} k_r(t)\,(r!)^{-\frac{1}{2}}\,H_r(x)\,G(x), \tag{2.4.6}$$

where $k_r(t)$, $r = 0, 1, 2 \ldots$ are the time varying coefficients of the expansion. To obtain an expression for the $k_r(t)$, we multiply both sides of (2.4.6) by $(s!)^{-\frac{1}{2}} H_s(x)$ and integrate with respect to x over the infinite range:

$$\int_{-\infty}^{\infty} P(x,t)(s!)^{-\frac{1}{2}} H_s(x)dx = \sum_{r=0}^{\infty} k_r(t) \int_{-\infty}^{\infty} (r!)^{-\frac{1}{2}} (s!)^{-\frac{1}{2}} H_r(x)H_s(x)G(x)dx. \tag{2.4.7}$$

From the orthogonality relationships (2.4.2), we find that only one term of the right hand side series is non-zero, and that is when $r = s$. As the integral then is unity, we have an explicit formula for $k_r(t)$:

$$k_r(t) = \int_{-\infty}^{\infty} P(x,t)\,(r!)^{-\frac{1}{2}} H_r(x)dx, \quad r = 0, 1, 2 \ldots \tag{2.4.8}$$

Note. The Hermite polynomial expansion for the probability density $P(x, t)$ of (2.4.6) is somewhat different from Cramer [51, p.223] who includes a factor $(-1)^r$ in both (2.4.6) and (2.4.8), and places the scaling factor $(r!)^{-\frac{1}{2}}$ of (2.4.8) into (2.4.6). The $(-1)^r$ factor merely alters the sign of our odd coefficients $k_{2r+1}(t)$ compared with Cramer's, but the $(r!)^{-\frac{1}{2}}$ factor changes the scale of the $k_r(t)$ coefficients. The scaling we have chosen is convenient as the functions $(r!)^{-\frac{1}{2}} H_r(x)\,G(x)$ appearing in

the expansion (2.4.6) have extrema which are the same order of magnitude.* This helps us judge the convergence of the series (2.4.6) by the relative magnitudes of the $k_r(t)$ series alone, which is an aid to deciding where to truncate the series.

Cramer [51, p.223] has shown that the expansion (2.4.6) is convergent whenever the integral

$$\int_{-\infty}^{\infty} e^{\frac{1}{4}x^2} P(x, t)\, dx \qquad (2.4.9)$$

exists. If $P(x, t)$ is a zero mean Gaussian distribution with variance σ^2, then the expansion will converge for any $\sigma^2 < 2$. However, in practical applications, the important question is not whether the series will converge or not, but whether the series will give an adequate representation of $P(x, t)$ for a small number of the lower order terms of the series.

The expansion (2.4.6) can be rewritten as

$$P(x, t) = \sum_{r=0}^{\infty} k_r(t)\, (r!)^{-\frac{1}{2}} (-1)^r \frac{d^r}{dx^r} [G(x)], \qquad (2.4.10)$$

and so the individual expansion functions are scaled versions of the successive derivatives of the standardized Gaussian distribution. To become familiar with the expansion and its convergence properties, the coefficients $k_r(t)$ were calculated and the distribution reconstructed using a finite number of the series terms for a selection of distribution functions $P(x, t)$:

(a) $P(x, t) = (2\pi\sigma^2)^{-\frac{1}{2}} e^{-x^2/2\sigma^2}$ σ^2 varying,

(b) $P(x, t) = (2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}(x-m)^2}$ m varying,

(c) $P(x, t) = \frac{1}{2a}$ for $|x| < a$, $a = (3)^{\frac{1}{2}}$

---

* For $r = 0, 10$ these extrema are $.399, \pm.242, -.282, \pm.223, .244, \pm.210, -.223, \pm.197, .209, \pm.184, -.198$.

$$\text{(d)}\quad P(x,\,t) \;=\; \frac{1}{a}\left[1 - \frac{1}{a}\,|x|\,\right],\ |x| < a, \qquad a = (6)^{\frac{1}{2}}. \tag{2.4.11}$$

Some general comments are as follows.

The coefficient $k_o(t)$ is equal to one for a true distribution function $P(x, t)$, as from (2.4.8) we have

$$k_o(t) \;=\; \int_{-\infty}^{\infty} P(x,\,t)\,dx. \tag{2.4.12}$$

The coefficient $k_1(t)$ is the mean of the distribution $P(x, t)$, as from (2.4.8) we have

$$k_1(t) \;=\; \int_{-\infty}^{\infty} x\,P(x,\,t)\,dx. \tag{2.4.13}$$

The coefficient $k_2(t)$ times $(2)^{\frac{1}{2}}$ is one less than the mean square of $x(t)$, as from (2.4.8) we have

$$k_2(t) \;=\; \int_{-\infty}^{\infty} (2)^{-\frac{1}{2}}\left[x^2 - 1\right] P(x,\,t)\,dx. \tag{2.4.14}$$

The equations (2.4.12-14) imply that if $P(x, t)$ is a zero mean, unit variance distribution, then $k_o(t) = 1$, and $k_1(t) = k_2(t) = 0$. In this case, the coefficients $k_3(t)$ and $k_4(t)$ are proportional to the well known measures of "skewness" and "kurtosis". It is also noted that the convergence of the series (2.4.6) is most rapid for distributions of zero mean and unit variance.

If the distribution $P(x, t)$ has a zero mean and is symmetrical, then all the odd coefficients k_{2r+1}, $r = 0, 1, 2 \ldots$ are identically zero. This is seen from the polynomial relations (2.4.4, 5) as the polynomial $H_{2r+1}(x)$ only involves odd powers of x. If it is known a priori, that the distribution $P(x, t)$ has zero mean and is symmetrical, then only the even coefficients $k_{2r}(t)$ need be computed.

Some specific comments on the expansions of the distributions (2.4.11) follow. We will use the error function

$$\text{Error}\ (n) \;=\; \int_{-\infty}^{\infty} \left| P(x,\,t) \;-\; P_n(x,\,t) \right| dx, \tag{2.4.15}$$

where $P(x, t)$ is the true distribution, and $P_n(x, t)$ is the reconstructed version using expansion coefficients up to the n:th

$$P_n(x, t) = \sum_{r=0}^{n} k_r(t) (r!)^{-\frac{1}{2}} H_r(x) G(x). \qquad (2.4.16)$$

Error (n) can be considered as a relative error (relative to one) as

$$\int_{-\infty}^{\infty} P(x, t)\, dx = 1.$$

(a) Zero mean Gaussian, variance σ^2.

The coefficients $k_2, k_4 \ldots k_{10}$ were computed, and are given along with the error function in Table (2.4.1) below, for various values of the standard deviation .

σ	k_2	k_4	k_6	k_8	k_{10}	Error (10)
.4	-.59	.42	-.32	.24	-.18	.259
.5	-.53	.34	-.23	.16	-.11	.102
.6	-.45	.25	-.14	.09	-.05	.031
.8	-.25	.079	-.026	.009	-.003	.0007
1.0	0	0	0	0	0	0
1.1	.15	.027	.005	.001	.0002	.0000
1.2	.31	.19	.048	.020	.008	.0011
1.3	.49	.29	.18	.12	.077	.013
1.4	.68	.56	.49	.44	.40	.083
1.6	1.1	1.5	2.1	3.1	4.5	1.13

Table (2.4.1)

It is noted that the series is divergent for $\sigma \geq (2)^{\frac{1}{2}} \doteq 1.414$, but Error (10) was not too high even for $\sigma = 1.4$. It is also noted

that the accuracy deteriorated rapidly for σ below 0.6.

(b) Unit variance Gaussian, mean m

The coefficients k_2 to k_{10} were computed, and are shown in Table (2.4.2) below along with the error function for various values of the mean m.

m	k_2	k_3	k_4	k_5	k_6	k_7	k_8	k_9	k_{10}	Error (10)
0	0	0	0	0	0	0	0	0	0	0
.5	.18	.051	.013	.0028	.0006	.0001	10^{-5}	10^{-5}	10^{-6}	10^{-6}
1.0	.71	.41	.20	.091	.037	.014	.005	.0016	.0005	.00006
1.5	1.59	1.38	1.03	.69	.42	.24	.13	.06	.03	.0055
2.0	2.83	3.27	3.27	2.92	2.39	1.80	1.27	.85	.54	.13
2.5	4.4	6.4	8.0	8.9	9.1	8.6	7.6	6.3	5.0	1.48

Table (2.4.2)

It is interesting that the series for m = 2.5 is convergent*, but many more terms than we have calculated would have been necessary to give an acceptable reproduction of P(x, t). The cases m = 1.5 or 2.0 could be taken as the rough limits of acceptability.

It is clear that the convergence is only rapid if P(x, t) has a mean value near zero, and a standard deviation near one. This is reasonable when you consider the shapes of the curves (2.4.10) used in the expansion. The basic curve r = 0 is the zero mean, unit variance Gaussian, and the subsequent curves are its r:th derivatives, scaled by the factor $(-1)^r(r!)^{-\frac{1}{2}}$. Thus if it is anticipated that the distribution to be expanded has a mean value greater than 1.5 in magnitude, or a standard deviation

* The integral (2.4.9) exists for all values of m, and thus the series for case (b) [i.e. $\sigma = 1$] is always convergent.

outside the range [0.6, 1.4]*, then it will be expedient to apply the expansion (2.4.16) to the standardized variate

$$y = \frac{x - m}{\sigma} , \qquad (2.4.17)$$

where m and σ are the true or estimated values of the mean and standard deviation of the random variable $x(t)$.

Then the expansion becomes

$$P_n(x, t) = \sum_{r=0}^{n} k_r(t)(r!)^{-\frac{1}{2}} \sigma_{(t)}^{-1} H_r(\tfrac{x-m}{\sigma}) G(\tfrac{x-m}{\sigma}), \qquad (2.4.18)$$

where the coefficient series $k_r(t)$ are evaluated from

$$k_r(t) = \int_{-\infty}^{\infty} P(x, t) (r!)^{-\frac{1}{2}} H_r(\tfrac{x-m}{\sigma}) dx. \qquad (2.4.19)$$

To test the convergence of the series for distributions with unusual shape, expansions were carried out for a flat and a triangular distribution. In each case they were standardized to zero mena and unit variance.

(c) Flat distribution

The even order coefficients were evaluated as follows.

k_o	k_2	k_4	k_6	k_8	k_{10}
.80	-.22	-.005	.12	-.16	.17

Table (2.4.3)

The coefficients k_o and k_2 were included as they should have been $k_o = 1.0$ and $k_2 = 0$; the inaccuracy being due to the quadrature**

---

* It appears from experimental evidence that a non-zero mean and a non-unit variance have independent effects on the convergence of the coefficient series.

** See "Hermite quadrature" a few pages ahead.

used to evaluate the integrals in (2.4.8). The flat distribution which is contained within $|x| < (3)^{\frac{1}{2}}$ only spanned four points of the 20 point quadrature, and a more suitable integration formula could have been chosen for this case. Even so, the reconstruction $P_{10}(x, t)$ still contained the essence of the flat distribution and the error function Error (10) = .26. As no odd order coefficients were used, the reconstructed distribution is also an even function, and the left halves of the distribution and its reconstruction are given in Table (2.4.4) below.

x	-3.0	-2.5	-2.0	-1.5	-1.0	-.5	0
P(x,t)	0	0	0	.29	.29	.29	.29
$P_{10}(x,t)$	.001	- .011	- .007	.10	.26	.31	.29

Table (2.4.4) Reconstruction of Flat Distribution

For much the same reasons that the square wave is the most difficult waveform to expand in Fourier series from the convergence viewpoint, the flat distribution is likely the worst example to be met in expansions using Hermite polynomials. This is because the flat distribution is quite unlike the expansion's basic distribution $G(x)$, and also unlike the higher expansion curves $d^r/dx^r\ [G(x)]$. That this accuracy was achieved using only four coefficients k_4, k_6, k_8 and k_{10} , was considered quite promising.

(d) Triangular Distribution

The triangular distribution was also difficult to handle because of its sharp peak. The even coefficients are given in Table (2.4.5).

k_o	k_2	k_4	k_6	k_8	k_{10}
.98	-.04	-.13	.10	-.03	-.03

Table (2.4.5)

Compared with the flat distribution, the coefficients k_o and k_2 are much nearer their proper values of 1 and 0, which is an

indication of the quadrature accuracy. The higher coefficients have decreased more quickly than those of the flat distribution [Table (2.4.3)] and the error is better at Error (10) = .09. The reconstructed distribution is given in Table (2.4.6) below, again giving only the left hand parts.

x	-3.0	-2.5	-2.0	-1.5	-1.0	-0.5	0
P(x,t)	0	0	.07	.16	.24	.32	.41
$P_{10}(x,t)$	-.002	.006	.05	.15	.26	.33	.35

Table (2.4.6)

Accuracy of Hermite Polynomial Expansions

It was found from the examples above, and others attempted, that the accuracy of the reconstructed distribution $P_n(x, t)$ (2.4.16) is closely related to the magnitude and the convergence (i.e. tendency to zero) of the coefficient $k_n(t)$ and its neighbours. This is seen from the fact that the error is directly given by the sum of those terms of the series (2.4.6) involving the coefficients $k_{n+1}(t)$, $k_{n+2}(t)$ ... etc. It was found that if $P(x, t)$ had a mean value near zero, and a variance near one, the convergence of the higher coefficients to zero was quite reliable, the main exception being the difficult flat distribution.

On the basis of the author's experience, the following "rule of thumb" can be stated. If a relative error as defined by (2.4.15) is desired to be in the region of 1 or 2 per cent, then the series should be truncated when $k_n(t)$ is around 0.05. Generally speaking, if the convergence of successive coefficients to zero is rapid, this value can be relaxed somewhat to $k_n(t) = .10$, but if the convergence is slow, then this value must be smaller, say $k_n(t) = 0.02$ or 0.03. By average rates of convergence we are referring to those experienced in case (a) above, for $\sigma = 0.6$ or 1.3. As is usual with numerical analysis procedures, however, it is best to treat each example as a special case, and to experiment with different truncation points to

gain an estimate of accuracy. In practical examples (to follow), the true distribution $P(x, t)$ is not available, and we must experiment to determine the accuracy.

Hermite Quadrature

As well as leading to an expansion formula, the properties of the Hermite orthogonal polynomials lead to a numerical integration formula called Hermite quadrature which is very efficient compared with formulae derived from Taylor series expansions if the argument of the integral is near Gaussian in shape. The quadrature formula is given by

$$\int_{-\infty}^{\infty} f(x)\, e^{-\frac{1}{2}x^2}\, dx \doteq \sum_{r=1}^{N} h_r\, f(x_r), \qquad (2.4.20)$$

where the x_r, $r = 1, N$ are the roots of the N:th Hermite polynomial and the h_r are a set of weighting coefficients. The reader is referred to Lanczos [54] for a discussion of quadrature methods, where it is shown that the weighting coefficients are chosen to make the numerical integration (2.4.20) exact if $f(x)$ is a polynomial of degree $2N - 1$ or less. This shows the connection of the quadrature with the polynomial expansions, for if

$$P(x, t) = f(x)\, e^{-\frac{1}{2}x^2} \qquad (2.4.21)$$

then the finite series (2.4.16) for $P_n(x, t)$ represents $P(x, t)$ exactly if $f(x)$ is a polynomial of degree n or less.*

Thus the Hermite quadrature formula (2.4.20) is particularly suitable for evaluating integrals of the type (2.4.8) for the $k_r(t)$ coefficients, for if $P(x, t)$ is a density function which is suitable for expansion in the Hermite polynomial series then it will be approximately of the form (2.4.21), and the multiplication by $H_r(x)$ in (2.4.8), increases the degree of the polynomial

---

* As, in this case, $f(x)$ can always be expressed as a linear combination

$$f(x) = \sum_{i=0}^{n} \alpha_i\, H_i(x).$$

$f(x)$ by r. As we have been evaluating $k_r(t)$ up to $r = 10$, then we should use a quadrature formula for (2.4.8) which is accurate for polynomials of at least the 20:th degree. As the quadrature (2.4.20) is very fast computationally, we can afford to be conservative, and the quadrature formula for $N = 20$ was used which integrates (2.4.20) exactly if $f(x)$ is a polynomial of degree 39 or less.

Comparing (2.4.20) with (2.4.8) it is noted that the factor $e^{-\frac{1}{2}x^2}$ of (2.4.20) does not appear explicitly in (2.4.8), but we have assumed $P(x, t)$ is approximately of this form. Thus the $f(x)$ of the quadrature formula (2.4.20) must be taken as

$$f(x) = P(x, t)\,(r!)^{-\frac{1}{2}}\,H_r(x)\,e^{\frac{1}{2}x^2} \qquad (2.4.22)$$

when applied to the integral (2.4.8). Again it is noted that this quadrature formula will only be efficient if $P(x, t)$ is the distribution of an approximately standardized variate $x(t)$ [see (2.4.17)] with mean near zero and variance near one. If this is not the case, the integral (2.4.8) can be transformed by the change of variable (2.4.17) to make the quadrature efficient.

The weights h_r and polynomial zeroes x_r are listed in many numerical analysis textbooks (for example, Kopal [55, p.530]) but are usually given for the Hermite quadrature derived from orthogonal polynomials with an e^{-x^2} base. If this is the case, then it is easily shown that the h_r and x_r needed in formula (2.4.20) are obtained by multiplying both the h_r and the x_r found in the tables by the factor $(2)^{\frac{1}{2}}$.

2.4.3 Hermite Transformation of the Fokker-Planck Equation

We have seen how, under certain conditions, the probability density function $P(x, t)$ can be described to a high accuracy by a very small number of parameters which are the coefficients of the orthogonal polynomial expansion of Hermite. Although these parameters

can be obtained directly from an equal number of the moments of the variable $x(t)$ [51, p.223], and hence contain no more explicit information than the variable's moments, they are a more efficient representation than the moments as they contain the implicit information of the orthogonal polynomial base $G(x)$ and the other expansion functions $d^r/dx^r\ [G(x)]$. Indeed, if a priori information leads us to expect that $P(x, t)$ will have a distribution other than near Gaussian, then there is a wide range of orthogonal polynomials associated with well known distributions which are derived from the Sturm-Liouville differential equations [56] and one of these may form a more efficient expansion series.

In the previous section we showed how a given distribution $P(x, t) = P(x)$ could be expanded in the coefficient series $k_r(t) = k_r$. We will now apply this expansion to the differential equation for $P(x, t)$, the FP equation, and we see that this operation is in effect an integral transform on the partial differential equation, for the space derivates $\partial P/\partial x$ and $\partial^2 P/\partial x^2$ can be expressed as algebraic operations on the Hermite polynomials. A separation of variables then occurs, and we are left with an infinite set of linear first order ordinary differential equations to solve. The infinite set can be reduced to a small finite set by truncation of the expansion series as in the last section, and readily solved.

Consider the FP equation (1.2.1) in single dimension form

$$\frac{\partial P(x, t)}{\partial t} = -\frac{\partial}{\partial x}\left[b(x,t)P(x,t)\right] + \tfrac{1}{2}\frac{\partial^2}{\partial x^2}\left[a(x,t)P(x,t)\right]. \tag{2.4.23}$$

We will assume that the incremental moments $b(x,t)$ and $a(x,t)$ can be written as, or well approximated by, a finite power series in x, with time varying coefficients if necessary. Then they can be written as

$$b(x,t) = \sum_i b_i(t)\, x^i,$$

and

$$a(x,t) = \sum_i a_i(t)\, x^i. \ * \tag{2.4.24}$$

---

* For the summation operators $\sum_i$ of this section, the index i will run from zero to the truncation point.

For linear systems, this representation will only involve two terms, e.g. $b_o(t)$ and $b_1(t)$, and for many non-linear systems an adequate representation should be obtained with only a few more terms. The main exception is a system with discontinuities, such as a relay system, which, as in the finite difference approach, will have to be handled by special methods. This could involve separate solutions in the continuous domains, piecing them together by appropriate continuity conditions across the system discontinuity, but it is unlikely that this could be achieved as easily by the present approach than by the finite difference approach. We will say no more about this except to state that the success of this method will depend on how accurate the representation (2.4.24) is for the given dynamic system.

As before, we shall write $P(x, t)$ as

$$P(x, t) = \sum_r k_r(t) \, (r!)^{-\frac{1}{2}} H_r(x) \, G(x). \qquad (2.4.25)$$

When the relations (2.4.24, 25) are substituted into the FP equation (2.4.23), the operations in the equation become space and time derivatives of products of series involving Hermite polynomials $H_i(x)$ and the Gaussian distribution $G(x)$. It is in the fact that these operations reduce to simple recursion relations on the Hermite polynomial series that the main power of the present approach lies. We will present a summary of the necessary relations below.

Relations (2.4.26)

(a) By an induction argument on the definition (2.4.1) of $H_r(x)$ it can be shown that

$$H_r(x) = r! \sum_{k=0}^{I(\frac{r}{2})} \frac{(-1)^k \, x^{r-2k}}{2^k \, k! \, (r-2k)!}$$

$$= x^r - \frac{r(r-1)}{2} \, x^{r-2} + \frac{r(r-1)(r-2)(r-3)}{2.4} \, x^{r-4} - \ldots$$

where $I(\cdot)$ is the integer part of the argument.

(b) Differentiating (2.4.26a) gives

$$\frac{dH_r(x)}{dx} = r\, H_{r-1}(x)$$

(c) Differentiating (2.4.1) directly gives

$$\frac{dH_r(x)}{dx} = x\, H_r(x) - H_{r+1}(x).$$

Equating the right hand sides of these last two relations gives the relation (2.4.5).

(d) Taking the partial *time* derivative of $P(x, t)$ of (2.4.25) gives

$$\frac{\partial P(x,t)}{\partial t} = \sum_r \dot{k}_r(t)\,(r!)^{-\frac{1}{2}} H_r(x)\, G(x),$$

where the dot denotes the time derivative.

(e) Taking the partial space derivative of $P(x, t)$ of (2.4.25) gives

$$\frac{\partial P(x,t)}{\partial x} = \sum_r k_r(t)(r!)^{-\frac{1}{2}} \left[H_r(x)\,\frac{dG(x)}{dx} + \frac{dH_r(x)}{dx}\, G(x)\right],$$

$$= -\sum_r k_r(t)\,(r!)^{-\frac{1}{2}} H_{r+1}(x)\, G(x),$$

as $\frac{dG(x)}{dx} = -x\, G(x)$ and $\frac{dH_r(x)}{dx}$ is taken from (c).

(f) Repeating the operation in (e), we have

$$\frac{\partial^2 P(x,t)}{\partial x^2} = \sum_r k_r(t)\,(r!)^{-\frac{1}{2}} H_{r+2}(x)\,G(x).$$

It is noted that the operation $\frac{\partial}{\partial x}$ advances the H index one integer, and changes the sign of H.

(g) From the formula (2.4.5) for $x\,H_r(x)$ we have

$$x\,P(x,t) = \sum_r k_r\,(r!)^{-\frac{1}{2}} [H_{r+1}(x) + r\,H_{r-1}(x)]\,G(x),$$

where $H_{-1}(x)$ is to be taken as zero.

(h) Similarly

$$x^2 P(x,t) = \sum_r k_r (r!)^{-\frac{1}{2}} [H_{r+2}(x) + (2r+1)H_r(x) + r(r-1)H_{r-2}(x)]G(x)$$

This operation can be repeated to obtain $x^s P(x,t)$. It is noted that each multiplying by x shifts the H index up one and down one integer, and the H_{r-i} with negative indices are set to zero.

(i) Multiplying (e) successively by x we obtain

$$x\frac{\partial P}{\partial x} = -\sum_r k_r (r!)^{-\frac{1}{2}} [H_{r+2}(x) + (r+1)H_r(x)]G(x),$$

$$x^2\frac{\partial P}{\partial x} = -\sum_r k_r (r!)^{-\frac{1}{2}} [H_{r+3}(x) + (2r+3)H_{r+1}(x) + r(r+1)H_{r-1}(x)]G(x)$$

$$x^3\frac{\partial P}{\partial x} = -\sum_r k_r (r!)^{-\frac{1}{2}} [H_{r+4}(x) + 3(r+2)H_{r+2}(x) + 3(r+1)^2 H_r(x) + (r-1)r(r+1)H_{r-2}(x)]\,G(x).$$

(j) The operations above for multiplying by x^s can be summed up as

$$x\, H_r(x) = H_{r+1}(x) + r\, H_{r-1}(x)$$

$$x^2 H_r(x) = H_{r+2}(x) + (2r+1)H_r(x) + r(r-1)H_{r-2}(x)$$

$$x^3 H_r(x) = H_{r+3}(x) + 3(r+1)H_{r+1}(x) + 3r^2 H_{r-1}(x) + r(r-1)(r-2)H_{r-3}(x).$$

Expanding the differentials in the FP equation (2.4.23) and dropping the (x, t) argument, we have

$$\frac{\partial P}{\partial t} = -\frac{\partial b}{\partial x} P - b \frac{\partial P}{\partial x} + \tfrac{1}{2} \frac{\partial^2 a}{\partial x^2} P + \frac{\partial a}{\partial x} \frac{\partial P}{\partial x} + \tfrac{1}{2} a \frac{\partial^2 P}{\partial x^2} . \qquad (2.4.27)$$

Using the relations (2.4.26) and the rescaled coefficient

$$c_r(t) = (r!)^{-\frac{1}{2}} k_r(t) \quad , \quad r = 0, 1, 2 \ldots \qquad (2.4.28)$$

we have

$$\sum_r \dot{c}_r H_r G = \left[-\frac{\partial b}{\partial x} + \tfrac{1}{2} \frac{\partial^2 a}{\partial x^2} \right] \sum_r c_r H_r G$$

$$+ \left[-b + \frac{\partial a}{\partial x} \right] \left[-\sum_r c_r H_{r+1} G \right]$$

$$+ \tfrac{1}{2} a \sum_r c_r H_{r+2} G. \qquad (2.4.29)$$

The multiplication of the polynomials b, $\frac{\partial b}{\partial x}$, a, $\frac{\partial a}{\partial x}$ and $\frac{\partial^2 a}{\partial x^2}$ by the H_r series is carried out by the relations (2.4.26j). The equation (2.4.29) involving the series is separated into simultaneous equations, involving single coefficients $\dot{c}_r$ on the left hand side, by multiplying both sides by H_s and integrating over the infinite range. Then by orthogonality relations (2.4.2), only that term in each series involving H_s^2 remains, and we obtain

the differential equation for $c_s(t)$. Repeating for $s = 0, 1, 2 \ldots$ up to the truncation point, we obtain simultaneous equations for the $c_s(t)$ which are linear, and do not involve $H_r(x)$, $G(x)$ or even x itself. The elimination of x only occurs when $a(x,t)$ and $b(x,t)$ are polynomials in x. This is a considerable convenience and explains the use of the polynomial form for a and b earlier (2.4.24). One difficulty is that the equation for $c_s(t)$ will in general involve $c_{s+1}(t)$, $c_{s+2}(t)$ ... (unless the system $x(t)$ is linear), and the series can only be truncated at $c_s(t)$ by setting the higher ones to zero. But this is what we have done in the finite expansions (2.4.16) of the previous section, and the same remarks as stated there will describe the accuracy of the truncation, except that in this case c_s $(= k_s\,(s!)^{-\frac{1}{2}})$ is considerably smaller than k_s for large s, and so will converge to zero more rapidly than the k_r series.

In order to avoid an excessive number of terms in the polynomial multiplication in the general case, we will continue our discussion with the specific example given in Section 2.3.1 with the parameters of Section 2.3.4. In this case

$$b(x,\,t) \;=\; b_3\,x^3,$$

and

$$a(x,\,t) \;=\; a_o. \qquad (2.4.30)$$

Then (2.4.29) becomes

$$\sum_r \dot{c}_r\,H_r\,G \;=\; -\,3\,b_3\,x^2 \sum_r c_r\,H_r\,G$$

$$+\; b_3\,x^3 \sum_r c_r H_{r+1} G + \tfrac{1}{2}\,a_o \sum_r c_r H_{r+2} G. \qquad (2.4.31)$$

Carrying out the substitutions for $x^s H_r(x)$ of (2.4.26j), and grouping the terms with the same H index together, we have

$$\sum_r \dot{c}_r H_r G = b_3 \sum_r c_r H_{r+4} G + \sum_r [-3b_3 + 3b_3(r+2) + \tfrac{1}{2}a_0] c_r H_{r+2} G$$

$$+ \sum_r [-6b_3(r+1) + 3b_3(r+1)^2] c_r H_r G + \sum_r [-3b_3 r(r-1) + b_3(r+1)r(r-1)] c_r H_{r-2} G. \quad (2.4.32)$$

Now multiply (2.4.32) by $(s!)^{-1} H_s(x)$ and integrate over the infinite x range. As an example of the mechanism involved, the series in the first term of the right hand side becomes

$$b_3 \sum_r c_r (s!)^{-1} \int_{-\infty}^{\infty} H_s H_{r+4} G \, dx, \quad (2.4.33)$$

whose elements are zero except when $r = s - 4$. This term then becomes $b_3 c_{s-4}$. The other terms of (2.4.32) involving H_{r+i} are treated by replacing r by $s - i$ and eliminating the H and G factors. Thus (2.4.32) becomes

$$\dot{c}_s = b_3 c_{s-4} + [-3b_3 + 3b_3 s + \tfrac{1}{2}a_0] c_{s-2} + [-6b_3(s+1) + 3b_3(s+1)^2] c_s$$

$$+ [-3b_3(s+2)(s+1) + b_3(s+3)(s+2)(s+1)] c_{s+2}, \quad (2.3.34)$$

where $s = 0, 1, 2 \ldots$, and those coefficients with negative indices are set to zero.

This is an infinite set of linear simultaneous differential equations for the Hermite expansion coefficients. It is noted that a separation of variables has occurred, as (2.3.34) does not involve x, and so our partial differential equation has been reduced to a greater extent than some integral transform techniques do.

As mentioned earlier, there is little difficulty in truncating the set of equations to obtain an easily soluble set for the first N Hermite coefficients. However, it was noted in the previous section that the Hermite expansion for $P(x, t)$ was only useful if $x(t)$ had a mean value near zero and a variance near one. For

the transient solution of the FP equation, the statistics of $x(t)$ will vary away from this normalized condition in general, and so we cannot rely upon the Hermite expansion being valid during the course of the transient solution. Thus we must apply the Hermite transform (i.e. expansion) to a normalized version of the FP equation.

2.4.4 Hermite Transformation of Normalized FP Equation

Consider the one-dimension FP equation (2.4.23) which is the FP equation of the diffusion system

$$dx(t) = b(x,t)dt + F(x,t)dw(t), \tag{2.4.35}$$

where

$$a(x,t) = F^2(x,t). \tag{2.4.36}$$

Define a normalized variate

$$y(t) = \frac{x(t) - m(t)}{\sigma(t)}, \quad \sigma(t) \neq 0, \tag{2.4.37}$$

where $m(t)$ and $\sigma^2(t)$ are the mean and variance of $x(t)$. To effect the transformation (2.4.37) we must obtain $m(t)$ and $\sigma(t)$. Differential equations for $m(t)$ and $v(t)$ have been derived in Appendix B, where $v(t)$ is the mean square of $x(t)$ and

$$\sigma(t) = [v(t) - m^2(t)]^{\frac{1}{2}}. \tag{2.4.37}$$

The differential equations are

$$\dot{m}(t) = E\,[b(x,t)], \tag{2.4.38a}$$

and

$$\dot{v}(t) = 2\,E\,[x\,b(x,t)] + E\,[a(x,t)]. \tag{2.4.38b}$$

Because of the $E[\cdot]$ operators on the right hand sides, these equations are non-random, and are not functions of $x(t)$. It is

noted that if $b(x,t)$ and $F(x,t)$ are linear functions of x (making $x(t)$ a linear system), the right hand sides of (2.4.38) can be expressed explicitly in terms of linear functions of $m(t)$ and $v(t)$, and the equations can be solved separately. For non-linear systems, however, the right hand side must be expressed as integral functions of $P(x,t)$, and thus the $m(t)$ and $v(t)$ equations must be solved simultaneously with the (normalized) $P(x,t)$ equation. In this case, the equations for $m(t)$ and $v(t)$ are no longer linear.

To solve (2.4.38) and to complete the solution for the statistics of $x(t)$, we must obtain $P(x,t)$ by solving for the distribution $Q(y,t)$ of the standardized variable $y(t)$. In Appendix B we derive the stochastic differential equation for $y(t)$, showing it is a diffusion process with the incremental moments

$$\underline{b}(y,t) = q_1(t)\, b(\sigma y + m, t) + q_2(t) y + q_3(t), \tag{2.4.39a}$$

and

$$\underline{a}(y,t) = q_1^2(t)\, a(\sigma y + m, t), \tag{2.4.39b}$$

where q_1, q_2 and q_3 are time varying functions depending on $m(t)$, $v(t)$ and $\sigma(t)$:

$$q_1(t) = \sigma(t)^{-1}, \quad \sigma(t) \neq 0, \tag{2.4.40a}$$

$$q_2(t) = -\tfrac{1}{2}\sigma(t)^{-2}\left[\dot{v}(t) - 2\, m(t)\, \dot{m}(t)\right], \tag{2.4.40b}$$

and

$$q_3(t) = -\sigma^{-1}\, \dot{m}(t). \tag{2.4.40c}$$

We can now write down the FP equation for the normalized system $y(t)$ as

$$\frac{\partial Q(y,t)}{\partial t} = -\frac{\partial}{\partial y}\left[\underline{b}(y,t)Q(y,t)\right] + \tfrac{1}{2}\frac{\partial^2}{\partial y^2}\left[\underline{a}(y,t)Q(y,t)\right]. \tag{2.4.41}$$

We can now solve this equation numerically by the Hermite transform method of the previous section, where we expand $Q(y,t)$ as

$$Q(y,t) = \sum_r c_r(t)\, H_r(y)\, G(y). \qquad (2.4.42)$$

The distribution of $x(t)$ is then obtained by

$$P(x,t) = \sigma^{-1}(t)\, Q(\tfrac{x-m}{\sigma}, t)$$

$$= \sigma^{-1}(t) \sum_r c_r(t)\, H_r(\tfrac{x-m}{\sigma})\, G(\tfrac{x-m}{\sigma}), \qquad (2.4.43)$$

as we had given earlier (2.4.18).

In Appendix B the differential equations for $m(t)$, $v(t)$ and $c_r(t)$ are derived in a form suitable for numerical computation, and, using the parameters (2.4.30) of our previous example, the equations are

$$\dot{m}(t) = (2\pi)^{\frac{1}{2}} \sum_r c_r(t) \sum_{s=1}^{N} h_s b_3 (\sigma(t) y_s + m(t))^3 H_r(y_s), \qquad (2.4.44a)$$

$$\dot{v}(t) = (2\pi)^{\frac{1}{2}} \sum_r c_r(t) \sum_{s=1}^{N} h_s [2b_3 (\sigma(t) y_s + m(t))^4 + a_o] H_r(y_s), \qquad (2.4.44b)$$

$$\sigma(t) = [v(t) - m^2(t)]^{\frac{1}{2}}, \qquad (2.4.44c)$$

$$\dot{c}_s(t) = [s(s+2)(s+1)b_3\sigma^2(t)]c_{s+2}(t) + [3s(s+1)b_3 m(t)\sigma(t)]c_{s+1}(t)$$

$$+ [3s^2 b_3\sigma^2(t) + s(3b_3 m^2(t) + q_2(t))]c_s(t)$$

$$+ [3(2s-1)b_3 m(t)\sigma(t) + q_1(t)b_3 m^3(t) + q_3(t)]c_{s-1}(t)$$

$$+ [3(s-1)b_3\sigma^2(t) + 3b_3 m^2(t) + q_2(t) + \tfrac{1}{2} q_1^2(t) a_o]c_{s-2}(t)$$

$$+ [3b_3 m(t)\sigma(t)]c_{s-3}(t) + [b_3\sigma^2(t)]c_{s-4}(t), \quad s = 0, 1, 2 \ldots, \qquad (2.4.44d)$$

remembering that the $c_s(t)$ with negative indices, or indices above the truncation point are set to zero. The h_s are the

Hermite quadrature weights (2.4.20) for the function evaluations at the points y_s.

Solution Example

In Figure 2.4.1 we show an example of the transient solution of equations (2.4.44), for initial conditions corresponding to those of the previous example, Figure 2.3.6. These conditions are the Gaussian (0.8, 0.09) curve and so we have

$$\begin{aligned} m(o) &= 0.8 \\ \sigma(o) &= 0.3 \\ v(o) &= 0.73 \\ c_o(o) &= 1.0 \\ c_r(o) &= 0.0 \quad , \quad r = 1, 10. \end{aligned} \tag{2.4.45}$$

The coefficient series was truncated at $r = 10$, but we shall see later that a lower truncation point could be used.

The equations (2.4.44) were integrated using Gill's version of the Runge-Kutta routine [55, p.213], with a time step $\Delta t = 0.1$. Although the Runge-Kutta method is not the most efficient of differential equation solution procedures, it has the advantages of

(a) being a self-starting one-step method (and thus structurally like a Markov process),
(b) being in most computers' program libraries, which, combined with (a) means it is very simple and automatic to implement,
(c) being reasonably stable and accurate for the non-linear equations (2.4.44), although one must be careful not to generalise on this point.

From (2.4.44) we have that

$$\dot{c}_o(t) = 0,$$

and hence

$$c_o(t) = 1.0, \tag{2.4.46}$$

which is the normalization condition (2.3.30), one of the conditions for $Q(y,t)$ [and hence $P(x,t)$] to be a probability density function.*

From (2.4.44) we see that $\dot{c}_1(t)$ and $\dot{c}_2(t)$ will be zero [see (2.4.13, 14)] if $m(t)$ and $v(t)$ are determined exactly so that $y(t)$ is precisely standardized to zero mean and unit variance. Of course this will not be so during the solution of the equations (2.4.44), as the $m(t)$ and $v(t)$ equations are not exact owing to the finiteness of the c_r series of (2.4.44a, b) and of the c_r equations (2.4.44d). The non-zero values of $c_1(t)$ and $c_2(t)$ occurring during the solution will be one indication of the errors caused by truncating the c_r series.

In Figure 2.4.1 we have plotted $k_r(t)$ instead of $c_r(t)$ because of their desirable scaling properties [see note following equation (2.4.8)]. The curves are shown for $t < 3.0$, at which point their rate of change was small. The steady state values are shown as dotted lines, and these are zero in the case of the odd order coefficients. These steady state values took a long time to be approached ($t > 10$), which is mainly due to the lethargy of the system near the origin.

The error, as measured by the error function Error (n) of (2.4.15), was computed using a very accurate finite difference solution as a reference. This is shown in Table (2.4.7)† for different truncation points of the coefficient series. Also given in the table is a guide to the typical sizes of $k_1(t)$ and $k_2(t)$ achieved during the solution.

The advantage of using the size of $k_1(t)$ and $k_2(t)$ as an indication of accuracy is that they are necessarily obtained with the solution, and thus are always available, whereas other error functions must be computed separately, and are generally not available if a known solution is not available. After some computation, some confidence was gained in the use of $k_1(t)$ and $k_2(t)$ as an indication of the error magnitude, and those values of Table (2.4.7) were typical values.

---

* The other main condition, the non-negativity of $Q(y,t)$, is not deducible from the c_r coefficients.

†On page 98.

Physical Interpretation of the Expansion Coefficients

The principal advantage of using the Hermite Transform method of solution, is that the coefficient series $k_r(t)$ helps to relate the shape of the solution $P(x,t)$ to the structure of the non-linearity of the dynamic system $x(t)$. We know that for linear systems in which $b(x,t)$ of (2.4.35) is linear in x, and $a(x,t)$ of (2.4.36) does not depend on x, the coefficient series $k_r(t)$, $r > 1$, are identically zero as the distribution of $x(t)$ is Gaussian.* The existence of the $k_r(t)$ elements, then, is due to the system non-linearity, and we can give the lower order coefficients $k_3(t)$ and $k_4(t)$ some physical meaning related to the form of the non-linearity of $b(x,t)$ of (2.4.36). The effect of any non-linearity in $a(x,t)$ on the coefficient series is not clear, and will not be discussed. In the comments below we will assume $a(x,t) = a(t)$.

As the odd order Hermite polynomials (2.4.4) contain only odd powers of x, and the even order ones contain only even powers, the expansion functions (2.4.10) of odd order will be odd functions of x, and those of even order will be even functions of x. Thus if the distribution $P(x,t)$ is an even (or symmetrical) function about the mean value $x = m$, then all coefficients $k_r(t)$ of odd order will vanish as they express the asymmetry of the distribution referred to its mean value.** Indeed if $b(x,t)$ is an odd function of x around the mean value, the distribution of x will be symmetrical. Although $b(x,t)$ had this odd function property in the example we have quoted, the non-zero condition we have used caused an asymmetry to appear in the transient solution. The odd order $k_r(t)$ are initially zero, as the initial condition is the Gaussian curve, but they grow initially as the non-linearity of the

---

* The exception is when there are non-Gaussian initial conditions, in which case the elements of the coefficient series will tend to zero as several time constants of the system elapse.

** Remember that the coefficients $k_r(t)$ are the expansion coefficients of the standardised variable $y(t)$.

of the system distorts the distribution as it tends towards its steady state mean value of zero. This effect disappears in our example as the steady state is reached, and the final distribution is symmetrical about zero.

If $b(x,t)$ were not an odd function about the mean value of $x(t)$, then a skewness will appear in the steady state distribution evidenced by a non-zero value of $k_3(\infty)$. Roughly speaking, the value of $k_3(t)$ will be proportional to how severely $b(x,t)$ is away from an odd function. A distribution which has an extended right hand tail will have a positive $k_3(t)$ coefficient. Referring to $P(x,t)$ of our example (see $t = 1.0$ of Figure (2.3.6)), it is noted that the distribution has an extended left hand tail (relative to $m(1.0) \doteq 0.6$), and thus $k_3(t)$ has a significant negative value.

The non-symmetry of the distribution is also reflected in non-zero values of the higher odd order coefficients, $k_5(t)$, $k_7(t)$ ... . These relate to subtler properties of the distribution and their physical meaning becomes difficult to interpret. The expansion function of (2.4.10) corresponding to the coefficient $k_r(t)$ has r zero crossings, and as r increases, these functions describe the "higher frequency" behaviour of the distribution. Because of the degree r of the power series associated with the r:th expansion function, the higher expansion coefficients have a relatively greater effect on the tails of the distribution than do the lower order coefficients.

The even order coefficients $k_4(t)$, $k_6(t)$ ... refer to the even part of the distribution $P(x,t)$ which is different from the Gaussian. The most significant of these, $k_4(t)$, describes the flatness or peakedness of the distribution compared to the Gaussian. This is seen in Figure (2.3.5) for $t = 2.0$ where $P(x, 2.0)$ is shown with a Gaussian curve of equal mean and variance. It is noted that $P(x, 2.0)$ is much flatter than the Gaussian, which indicates a significant negative value of $k_4(2.0)$. This was also noted in the example of Figure (2.4.1).

The existence of the even coefficients can be traced to the odd function part of the system non-linearity, $b(x,t)$, which is

different from $b_1 x$. This is most easily seen when $b(x,t)$ is expressed as the power series (2.4.24), for then it is the coefficients $b_3(t)$, $b_5(t)$ ... which influence the even order expansion coefficients $k_4(t)$, $k_6(t)$ ... . In our example, b_3 was negative, which caused a negative $k_4(t)$, or a flattened distribution. If $b(x,t)$ were a sign function (an ideal relay) then in the expansion (2.4.24), the coefficient b_3 would be positive, $k_4(t)$ would be positive, and the resultant distribution highly peaked. This is confirmed by many of Merklinger's examples [1]. Expressing this effect in control systems terminology, if the system non-linearity is of the saturating or "soft" type, then $k_4(t)$ will be positive; if the non-linearity is of the "hard" type, then $k_4(t)$ will be negative. Again, it is more difficult to place a physical interpretation on the higher even coefficients $k_6(t)$, $k_8(t)$ ... as these are related to the "higher frequency" components of the even part of $P(x,t)$, and especially to the tails of the distribution.

Truncation of the Expansion Coefficients

The smallness of the higher expansion coefficients (especially $k_9(t)$ and $k_{10}(t)$ - see Figure (2.4.1)) and their lack of physical meaning suggested that we might get by with fewer expansion coefficients than we originally retained. Solutions were obtained by successively omitting (i.e. setting to zero) the highest two expansion coefficients, and the effect on solution accuracy is summarized in Table (2.4.7).

Highest order of c_r coefficient kept in eqn.(2.4.44)	Error (n)	Typical size of $k_1(t)$, $k_2(t)$ during solution
n = 10	.010	.004
8	.013	.005
6	.04	.010
4	.12	.100

Table (2.4.7) Effect of Hermite Coefficient Series Truncation on Solution Accuracy

As anticipated, the elimination of $k_9(t)$ and $k_{10}(t)$ had very little adverse effect on the error function, and gave a reconstructed distribution $P_8(x,t)$ which was hardly discernable from $P_{10}(x,t)$. The elimination of subsequent coefficients did have an appreciable effect on accuracy, but depending on accuracy requirements, the retention of only $k_3(t)$ and $k_4(t)$ may well provide an acceptable solution in conjunction with $m(t)$ and $\sigma(t)$.

We have noted that the coefficient series can be severely truncated if the distribution $P(x,t)$ is near Gaussian in shape. We have also noted that for separation of variables in the Hermite transform to be successful, the system functions $a(x,t)$ and $b(x,t)$ must be expressible as polynomials, and the effect of this power series is to cause coupling between the coefficient equations [c.f. Relation (2.4.26j) and the substitution leading to eqn. (2.4.32)]. Thus the lower the degree of the power series for $b(x,t)$ and $a(x,t)$ the less coupling there is between the $c_r(t)$ equations. This likely infers that the $c_r(t)$ series could be truncated earlier with less ill effects (and that $P(x,t)$ is nearer Gaussian) if the power series degree is smaller, but it would be difficult to follow up this inference by analytical means.

2.4.5 Summary of Hermite Transform Method

We have presented an infinite series functional representation (2.4.6) of a probability density function $P(x,t)$ which is very economical of parameters when $P(x,t)$ is similar in shape to a Gaussian distribution. This representation is well known in applied statistics as the Gram-Charlier series of Type A, but the present novelty lies in applying this representation of $P(x,t)$ to the FP equation. This substitution is in effect a linear integral transform of the partial differential FP equation, which results in a complete separation of variables (due to the orthogonality property of the functional representation) and a rather neat reduction to an infinite set of first order ordinary differential equations (2.4.34). These

can be truncated, under convergence assumptions, to form a readily solvable set of equations.

The convergence properties of the functional representation were tested for a variety of arbitrary distributions, and it was found that rapid convergence of the series representation depended not only on the nearness of $P(x,t)$ to the Gaussian distribution, but on the nearness of $x(t)$ to being a standardized variable (zero mean and unit variance).

As the standardization of $x(t)$ could not be assured during the general transient solution of the FP equation, a preliminary transformation (2.4.37) had to be made to the $x(t)$ variable prior to the Hermite integral transform. The combined transformations still resulted in an infinite set of simultaneous ordinary differential equations to solve, but they were no longer linear.

It was noted that the separation of variables required that the system non-linearity $b(x,t)$ and $a(x,t)$ be of a form suitable for representation as power series in x. This eliminated certain discontinuous systems such as relay systems from consideration by the Hermite transform method. However the simple example solved earlier by finite differences was well suited to the Hermite transform method, and was presented in detail. It was found that solution accuracies of the order of one to two per cent could be obtained by solving only ten equations (for $m(t)$, $v(t)$ and $k_1(t)$ to $k_8(t)$), which was very economical of storage space and computer time. Although somewhat better than the finite difference method in these respects, the initial posing of the problem and reconstruction of the solution were more difficult with the Hermite transform method, and so no preference could be stated for solution methods for the present example.

Although system discontinuities could not be handled by Hermite transforms, there were no arbitrary space truncation conditions to apply in this case, these being replaced by the straightforward functional expansion truncation. However, if the distribution $P(x,t)$ had strange tails such that the space truncation would be difficult to apply in the finite difference method, a large number of terms in the functional expansion would have to be retained in

order to model the tails accurately (that is, if the tails were of interest, as they were in our alarm design application).

Although extra computation was needed to reconstruct the distribution $P(x,t)$ from (2.4.43) in the Hermite transform method, it was noted that infinite integral functions of $P(x,t)$ (i.e. $E[g(x,t)]$) could easily and accurately be evaluated using the Hermite quadrature method of numerical integration, as this was done directly from the coefficients of the functional expansion of $P(x,t)$.

Although the form of the solution $P(x,t)$ could not be deduced from the Hermite expansion coefficients without some experience, it was interesting that rough information about the lower order expansion coefficients could be deduced from the shape of the system non-linearity. Most of the physically important information was contained in $m(t)$, $\sigma(t)$, and the first odd and even order expansion coefficient, $k_3(t)$ and $k_4(t)$, and if accuracy requirements were limited, a useful solution could be obtained by solving for only those quantities. The existence of the coefficients $k_3(t)$, $k_4(t)$, $k_5(t)$ ... tell us explicitly how the statistics of the non-linear system differ from those of the linearized system having the same mean and variance.

Compared with the finite difference case, the problem of system dimensionality would seem to be even more of an obstacle. If the variable $x(t)$ did not need to be standardized, then the linear transform technique of Section 2.4.3 could be carried out using multi-dimensional generalizations of the basic definitions (2.4.1-8). We would obtain sets of linear differential equations involving approximately $(N)^n$ unknowns, where N is the number of coefficients in one dimension, and n is the dimension of the system.

However, the variable $x(t)$ will have to be transformed to standardized form, and the combination of the two transforms would be quite formidable for higher dimension systems. Further, a large set of non-linear simultaneous differential equations would have to be solved, which would be very difficult, particularly from

the stability point of view. For this reason, the method of Hermite transforms was held to be sufficiently unpromising for multi-dimensional systems that no examples were attempted. It was noted earlier that finite difference methods were feasible for systems of up to three space variables, and thus are to be preferred to the Hermite transform method for problems of two and three dimensions.

Thus for these direct methods of obtaining a system's statistical behaviour by solving the FP equation, a law of diminishing returns applies to system dimensionality, where the effort required for solution increases out of proportion with the knowledge gained about the system when the dimensionality increases.

CHAPTER 3

Simulation and the Monte Carlo Solution of Parabolic Equations

Generally speaking, there were two main motives for studying the use of simulation techniques to determine the statistical behaviour of noisy systems. First of all, it became clear during the work of the previous chapter, that the direct method of solution using Fokker-Planck techniques could not handle problems of a very wide generality. Secondly, it appeared that the simulation of diffusion processes on a digital or analogue computer was not a routine matter, and would merit investigation in itself. This chapter will elaborate on this motivation, and then present theorems connecting simulation techniques to the solution of general linear parabolic equations, which are extensions to previous results on the Monte-Carlo solution of parabolic equations. Subsequent chapters will be devoted to the problem of the simulation of physical and diffusion processes on analogue and digital computers.

3.1 Motivation for Simulation Techniques

Limitations of the Direct Method

The direct method of obtaining a system's statistical behaviour is to solve a partial differential equation, the Fokker-Planck equation, for the system's transition probability density function. In Chapter 2, we discussed two techniques for obtaining approximate numerical solutions of the FP equation. The finite difference method is the classical method of numerical solution, and has been well proven on a large variety of parabolic, elliptic and hyperbolic partial differential equations. With a little study, an engineer can learn the subtleties of finite difference techniques, and apply them to the FP equation. The second method, the Hermite transform method

was rather novel in the context of the numerical solution of partial differential equations, and was developed specifically for application to the FP equation. Both of these methods, however, have limitations which restrict the form of statistical system that can be studied.

The FP equation has as many independent space variables as states in the Markov representation of the physical system (c.f. Section 2.2), and this number can be quite high for practical systems. We have seen how the numerical solution of the FP equation becomes difficult when there was more than one space variable involved. In particular we have found the Hermite transform method to be suitable only for systems of single dimension, and that with liberal amounts of computer storage capacity, execution time, and programming effort available, problems as high as three-dimensional ones could be handled by finite differences.

Clearly then, system dimensionality is a severe restriction on the variety of systems studied. The law of diminishing returns prompts us to consider ways of reducing the system's dimensionality, whilst retaining the pertinent characteristics of the system's statistical behaviour. Davison [59] has discussed a method of reducing system dimensionality by retaining the prominent eigenvalues of the system matrix, which has been shown to provide a significant reduction in state variables for very large systems with little loss of accuracy. His technique could be extended to stochastic systems, but is applicable only to linear systems. By methods analogous to those of Section 2.2, "equivalent" diffusion processes of lower dimension could be derived for a given non-linear diffusion process which would model some properties of the original process and be more amenable to numerical solution. This prospect, however, needs further study before its practicality could be assessed, and likely would not be as successful as the linear case.

Another limitation on the system studied is on the form of the non-linearity present in the system. Once again the Hermite transform method seems to be more sensitive to this limitation as we have noted that the non-linearity must be in the form of a power series

in x of reasonably low degree for the method to be successful. This requires that the non-linearity be quite smooth, hence eliminating discontinuous (relay) systems, unless a crude approximation were allowed.

For finite difference methods, it was noted that the system discontinuity must be amenable to finite difference representation. Here, smoothness will be desirable, but will not be as critically important as in the Hermite transform case.

Advantages of Monte Carlo Techniques when the above limitations are present.

By Monte Carlo or simulation techniques, we will be referring to a direct simulation of the statistical system $x(t)$, using a representative noise driving signal. If the statistics of $x(t)$ are collected in such a way as to form an estimate of $P(x,t)$, then this can be considered as a Monte Carlo solution of the FP equation. In contrast to the earlier numerical methods for the FP equation, we will call the simulation method the indirect method of solving the prediction problem.

When studying the n:th order or n state variable dynamic system represented by the diffusion system

$$dx(t) = f(x,t)\,dt + F(x,t)\,dw(t), \qquad (3.1.1)$$

the Monte Carlo method requires the simulation on a computer of n simultaneous first order stochastic differential equations. This problem will be discussed in the subsequent chapters, but here we will compare the scale of this problem with that of the direct methods described earlier.

Essentially the Monte Carlo solution proceeds in two parts which can be considered separately. First there is the actual simulation of (3.1.1), and then there is the data reduction operation which forms the required estimates of the statistical parameters of interest. This separation of computing operations is the main simplicity of the Monte Carlo method, for, recalling the discussion of the characterization of a random process in Section 2.4.1, the

direct methods of solving the FP equation require the computation of the rather complete statistical description $P(x,t)$ in more or less one operation, with no short cuts to obtain a simplified description involving less effort.

The main significance of this statement concerns the problem of system dimensionality, for in the multi-dimension problem, the dependence of the statistical parameters of each independent space variable $x_i(t)$, $i = 1, n,$ cannot be separated in the direct methods, and a complete solution must always be obtained involving the dependence on the entire state vector $x(t)$. Only after the entire solution $P(x,t)$ is obtained can a simpler statistical parameter such as $E[x_1(t)]$ be obtained. By contrast, in the second (data reduction) step of the Monte Carlo method, the data reduction operations are only carried out on the independent space variables of interest, and the rest are ignored. For example, as far as the data reduction operation is concerned, it is just as easy to obtain $E[x_1(t)]$ for a ten-dimensional system, as for a one- or two-dimensional system.

Thus if only a limited statistical description of the system is required a considerable saving is afforded by the Monte Carlo method for high dimensional systems, for it is the complete representation of $P(x,t)$ which requires the high storage capacity and long computing time. The first operation of the Monte Carlo technique of simulating the system (3.1.1) does not have the same magnitude of dimensionality problem, for the effort of simulation, roughly speaking, will only increase linearly with the number of system dimensions, whereas for the direct methods the effort increased geometrically. Furthermore, the number of trials needed in the simulation depends on the possibly limited solution requirements, and not on the dimensionality of the system.

It was mentioned in Section 2.4.1 that a complete characterization of a random process requires the time correlation information $P(x(t_1), x(t_2), \ldots ; t_1, t_2 \ldots)$. This can only be obtained by the direct method by the tedious procedure of piecing together successive solutions obtained from delta function initial conditions, [i.e. we must solve for $P(x(t_2),t_2 \mid x(t_1) = x_1)$ for as many

x_1, t_1 and t_2 of interest]. By contrast, this information could be obtained with relatively little extra work in the data reduction part of the Monte Carlo solution.

As an added advantage, the simulation required in the Monte Carlo method is not influenced too greatly by the complexity of the non-linearity of the statistical system under study. What is more to the point is that the system should be reasonably stable so that its simulation on the computer will not present undue stability problems.

The main disadvantage of the Monte Carlo method is the property of sampling statistics that the accuracy of estimators only increases as the square root of the number of samples taken. For example, if $x(t)$ is a random variable with standard deviation σ, and $x_\alpha(t)$ is the α:th realization generated by a simulation procedure, then the estimated quantity

$$E_N [x(t)] = \frac{1}{N} \sum_{\alpha=1}^{N} x_\alpha(t) \qquad (3.1.2)$$

has a mean value equal to $E [x(t)]$ and a standard deviation $(N)^{-\frac{1}{2}}\sigma$, provided the Monte Carlo procedure generates uncorrelated samples $x_\alpha(t)$. Thus to obtain an estimate of $E [x(t)]$ with a relative error of 10% (i.e. $= 0.1\sigma$) we will need 100 independent trials, but to reduce the error to 1%, we would need 10,000 trials. These remarks apply to the effort of simulation, for the data reduction operation (which may be the most time-consuming) need not be increased in proportion, as quantising methods make the final part of the data reduction operation independent of the number of trajectories. Nevertheless, the law of diminishing returns applies, and it will not be practical to obtain very accurate solutions by the Monte Carlo method which is more suited to obtaining quick rough answers.

Thus we have replaced the limitations on system complexity by a limitation on the accuracy of solution. While for many simple systems,

the direct method will be preferable to Monte Carlo methods, it does seem that many of the complex systems met in engineering will not be amenable to solution by the direct method, and Monte Carlo methods will have to be used.

Explicit Motives for Studying Simulation Techniques

In addition to the reasons above, there are motives for the study and use of Monte Carlo methods in their own right.

Apart from computational considerations, the main advantage of using simulation techniques is the insight one can gain by studying individual system trajectories in detail. The direct method of obtaining $P(x,t)$ only examines the behaviour of trajectories in a collective fashion, and may miss some of the subtler points of interest. In other words, the solution $P(x,t)$ is a deterministic function and insight into the random nature of the system can only be inferred through an interpretation of the function $P(x,t)$. By contrast, Monte Carlo methods allow us to examine the statistical behaviour of the system directly. This, of course, has always been the power of analogue computers, but these remarks apply to simulation on digital computers as well.

The need to study simulation techniques as an academic exercise became apparent when recent papers indicated a difference between systems arising from physical situations and truly stochastic systems, when the noises involved were non-additive [18, 24]. Some computation was done to test the results of the paper by Wong and Zakai [24], but the extensions indicated by the results of Clark [22] added a greater impetus to the study of simulation methods. Further, preliminary enquiries among those aware of the theory indicated that little computation had been done to test the theory.*

The essential computational difficulty of simulating diffusion processes is that white noise cannot be represented on an analogue

---

* Some computational experience was mentioned in a private communication to Wonham [49] in July, 1965, by J. Ternan of the Defence Standards Laboratories, Victoria, Australia, but this does not appear to be documented. In his thesis [60, Appendix 10], Ternan derives a result which is equivalent to Clark's in the scalar case.

or digital computer exactly, and must be approximated by a signal of finite bandwidth and power. This means we are simulating a diffusion process by a physical process, and the implications of this are discussed in Chapter 4. It turns out that the choice of noise source is the principal problem on analogue computers, and the choice of integration formula is the main problem on digital computers, and these problems will be discussed in Chapters 5 and 6.

The data reduction part of the Monte Carlo solution will not be discussed, although an interesting smoothing technique was developed for the digital computer solution, and will be mentioned in Chapter 6.

A further impetus to studying simulation techniques comes from some recent suggestions for the practical implementation of optimal stochastic control theory [75, 76]. The suggestion is to use Monte Carlo techniques to hill climb in control policy space, and hence iteratively converge towards an optimal control policy. Essentially their approach is motivated by the same problem as ours - the partial differential equations of optimal stochastic control theory are too difficult to solve (particularly in an "on-line" mode), and a direct simulation of the system is used as an alternative method of solution.

3.2 The Monte Carlo Solution of Parabolic Equations

In this section we will discuss in more detail the connection between the simulation of a system and the solution of its FP equation (forward Kolmogorov equation), and emphasize the principle of conservation of trajectories of the Markov process inherent in the FP equation. We will show that if we allow a violation of the principle of conservation, then simulation techniques can be used to solve a class of parabolic equations of wider generality than those fitting into the FP format (1.2.1).

Monte-Carlo methods for obtaining approximate solutions to differential and integral equations first became prominent when the advent of the modern electronic computer allowed large-scale statistical experiments to be performed, and a method of handling elliptic differential equations was developed as early as 1949 by Metropolis and Ulam [61]. Later, as research in numerical analysis proceeded, Monte Carlo methods generally gave way to more efficient and accurate methods such as that of finite differences, and Monte Carlo methods were considered as brute force or last resort methods, if all else failed. Interest in Monte Carlo methods has never died out, however, and recent developments in hybrid computer facilities have brought Monte Carlo techniques back into a respectable position among numerical methods (see, for example, [62]). Also, there is still a large variety of more difficult problems which direct methods of numerical analysis have not been able to solve satisfactorily.

The common method of solving elliptic and parabolic differential equations by Monte Carlo methods has stemmed from the backward Kolmogorov equation (1.2.4) of a diffusion process [63]. From the theory of integration over Wiener measure (or function space integration) [65, 66] has come an extension which allows the appearance of proportional terms in the parabolic equation not existing in the backward Kolmogorov equation. These same terms could be handled by stopping or boundary conditions on the random walks used [64, 67 Chapter 13].

The essential feature of Monte Carlo methods based on the backward equation is that the time scales of the differential equation and of the simulated diffusion process are in the reversed direction to each other. That is, if the time parameter t of the process $x(t)$ increases to the right on the real time axis, the time parameter s of the backward Kolmogorov equation (1.2.4) governing $x(t)$ increases to the left on the same axis. The main advantage of this method is that solutions to the differential equation $P(x,t_f)$ are obtained pointwise [i.e. $P(x(t_f) = x_a, t_f)$, $P(x(t_f) = x_b, t_f)$ are obtained for various x_a, x_b] by letting the simulated trajectories run backwards from the (present)time t_f, starting at the desired solution point $x(t_f)$. The solution $P(x(t_f), t_f)$ is obtained from averaging functions of the stopping conditions of the trajectories.

This method is very efficient if solutions are required at a few specific points only, but will be inefficient if general solutions $P(x,t)$ or solutions of the form Prob $[x(t) \in A]$ are required, where A is a given subset of the state space of x. To avoid this difficulty, we will propose a method based on the forward Kolmogorov or FP equation which inherently obtains the general solution $P(x,t)$, but may not be as efficient as the previous method for point solutions.

As the time scale of the Monte Carlo trajectories is now coincident with that of the parabolic differential equation, we are in a position to draw analogies between the trajectories and underlying principles of the differential equation. Indeed, we may exploit these physical analogies to help us solve the Monte Carlo trajectories. For example, in solving the heat conduction equation, the Monte Carlo trajectories can be associated with a modified form of a quantum of heat energy, an analogy which will help us choose the correct form of flux type boundary conditions. In addition, the method to be presented will handle parabolic equations of more generality than the previous methods, as constant terms can now be accommodated. Although the method to be presented will handle elliptic equations by allowing steady state distribution of trajectories to be achieved, our discussion will be directed towards the solution of parabolic equations where transient aspects of the solution are of interest.

3.2.1 The Fokker-Planck Equation as an Equation of Conservation*

The FP equation is derived [6, 7] directly from the Chapman-Kolmogorov integral equation

$$P(x_2, t_2 \mid x_o, t_o) = \int_R P(x_2, t_2 \mid x_1, t_1)\, P(x_1, t_1 \mid x_o, t_o)\, dx_1, \tag{3.2.1}$$

where the domain of integration R is the entire phase space of the Markov process $x(t)$, and t_1 is an arbitrary time where $t_2 > t_1 > t_o$. This integral relation of conditional probability densities is a definitive feature of Markov processes, but viewed another way, it is a strong statement of continuity** and conservation of the individual trajectories of $x(t)$. Figure 3.2.1 will illustrate this aspect of the Chapman-Kolmogorov equation.

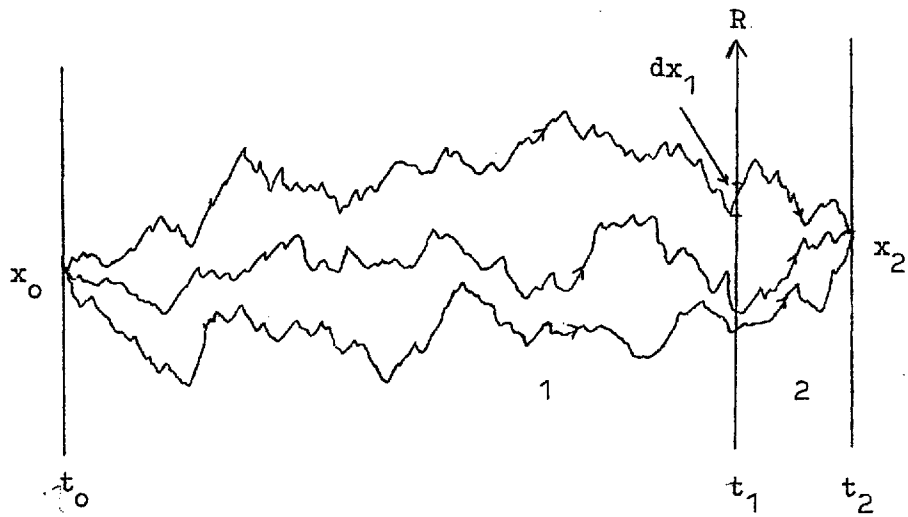


Figure 3.2.1 Continuous Trajectories of a Markov Process

With reference to Figure 3.2.1, equation (3.2.1) can be expressed in words as follows.

The probability that the Markov process $x(t)$ has the state (position) x_2 at time t_2 given that it began in the arbitrary

---

* A summary of the methods proposed in this and the next section is given at the beginning of Section 3.2.3. It may be helpful to refer to this summary during the course of reading the next two sections.

** continuity in time only.

state x_o at time t_o is given by the sum (over all dx_1) of the product of

(a) the probability of the process reaching a point x_1 in dx_1 at time t_1 after beginning at x_o at time t_o $[= P(x_1, t_1 \mid x_o, t_o)dx_1]$ and

(b) the probability that the process then follows on from x_1 at t_1 to arrive at point x_2 at time t_2 $[= P(x_2, t_2 \mid x_1, t_1]$.

The intermediate intervals dx_1 in the sum must span the reachable space R, and when the magnitudes of the intervals dx_1 are reduced to zero, we obtain the integral (3.2.1).

With reference to Figure 3.2.1, the product of these probabilities is only meaningful if

(i) the probabilities refer to independent events. This is the Markov property, and means that the trajectory in region 2 only depends on its initial point x_1 and not on the previous part of the trajectory in region 1;

and (ii) both the events (a) and (b) do occur. The first event is the trajectory in region 1 which reaches x_1, and the second event is the trajectory in region 2 which begins at x_1. If either of these events do not take place (that is, a trajectory terminates or begins at time t_1), then the product of the probabilities to give an element in the sum for the probability of the total trajectory (from t_o to t_2) is invalid. ~~The same argument shows that the product of probabilities is not meaningful if a trajectory reaches R at time t_1 at point x_1, and leaves at a different point x_1'~~

The argument above holds for all t_1 in (t_o, t_2) and all x_1 in R, and so if the integral relation (3.2.1) is to be a property of the Markov process $x(t)$, then $x(t)$ must be <u>continuous</u> in $[t_o, t_2]$ with probability one. From another viewpoint, we can integrate both sides of (3.2.1) with respect to x_2 over the domain R. If the arguments of the integral are genuine probability density functions, then the integral equals <u>unity</u> and we have a statement of <u>conservation of trajectories</u>:

"All trajectories leaving the arbitrary point x_o at time t_o arrive at some point x_2 in R at time t_2 with probability one". The arguments of the integrals were shown to be valid probability density functions by the discussion above concerning the continuity of trajectories at an arbitrary intermediate time t_1.

Although Figure 3.2.1 has shown a one-dimensional diffusion process, the Chapman-Kolmogorov equation (3.2.1) and the discussion above apply equally well to the general n-dimensional diffusion process (3.1.1). Thus we have shown that the trajectories of a process whose statistics are described by a Fokker-Planck equation are continuous and obey a conservation principle with probability one. This principle, and modifications of it, is fundamental to the Monte Carlo methods discussed in the rest of this chapter.

Application to Parabolic Equations

Consider the general linear parabolic partial differential equation

$$U_t = \underline{L}(U) \quad , \quad t \text{ in } [0, T], \tag{3.2.2}$$

where $U = U(\mathbf{x}, t)$

$$U_t = \frac{\partial U}{\partial t}(x, t),$$

and $\underline{L}(\cdot)$ is a general linear second order elliptic operator*, involving constant, proportional, and first and second partial derivative terms with respect to the space variables x_i, $i = 1, n$.

---

* A definition of an n-dimensional elliptic operator could not be found,** but we will take it to be an operator whose matrix of coefficients $\alpha_{ij}(x,t)$ of the second order terms $U_{x_i x_j}$ is real and positive definite. This is consistent with the usual definition of the two-dimensional elliptic operator, and it ensures that the matrix $[\alpha_{ij}]$ can be factored into the form $F F^T$. As $U_{x_i x_j}$ is the same as $U_{x_j x_i}$, the matrix α can be made symmetric with no loss of generality.

** See Duff "PDE's" P. 72

Then if $U(x,o)$ is given, this is a properly posed initial value problem defining the solution $U(x,t)$, and other boundary conditions are optional. For the time being we will assume there are no other boundary conditions, and will later discuss the treatment of a case where space boundary conditions are present.

Compare the form of the equation (3.2.2) with the general form of the FP equation (1.2.1)

$$P_t = - \sum_{i}^{n} \frac{\partial}{\partial x_i} [b_i(x,t)P] + \tfrac{1}{2} \sum_{i,j}^{n} \frac{\partial^2}{\partial x_i \partial x_j} [a_{ij}(x,t)P]. \quad (3.2.3)$$

Expansion of the derivatives on the right hand side of (3.2.3) will give terms involving $\partial^2 P/\partial x_i \partial x_j = P_{x_i x_j}$, P_{x_i}, and P but no constant terms (independent of P).

By judicious choice of the coefficients $a_{ij}(x,t)$ and $b_i(x,t)$ of (3.2.3), it may be possible to equate the FP equation (3.2.3) term by term to the parabolic equation (3.2.2) for a given operator $\underline{L}(U)$. If this is the case, then the given parabolic equation can be put in the FP form (3.2.3) and it is said to be an equation of the FP type. The Monte Carlo solution of such an equation is discussed in this section.

If no choice of $a_{ij}(x,t)$ and $b_i(x,t)$ can equate all the terms of (3.2.3) to those of the given parabolic equation (3.2.2), the parabolic equation is not of the FP form. The Monte Carlo solution of such equations is discussed in Section 3.2.2.

Consider a given parabolic equation (3.2.2).

Theorem 3.2.1 If the parabolic equation is of the FP type, then it can be solved by a Monte Carlo method involving the conservation of trajectories.

The proof of this theorem is heuristic, and relies upon the discussion given earlier which showed that a process $x(t)$ which was governed by the Chapman-Kolmogorov integral relation (3.2.1), and hence also by a FP equation, underwent trajectories in state space which obeyed a principle of conservation.

In particular, if the given parabolic equation is of the FP type, then it is the FP equation for some continuous Markov or diffusion process $x(t)$ (2.1.1) defined by the Ito s.d.e.

$$dx(t) = f(x,t)\,dt + F(x,t)\,dw(t). \qquad (3.2.4)$$

The coefficients of this diffusion process $f(x,t)$ and $F(x,t)$ are related to those of the parabolic equation by the relations (3.2.5, 6). As the FP equation describes the probability density of the process (3.2.4), a Monte Carlo solution of the parabolic equation is obtained by simulating the process (3.2.4) and forming suitable estimates of the probability density of the simulated trajectories.

Many simulated trajectories will be needed to obtain estimates with useful accuracies, and it is helpful to think of the state space R filled with "particles", each one of which follows a trajectory defined by the s.d.e. (3.2.4) with a different and independent noise vector $dw(t)$ driving each particle. The initial distribution of the particles is given by $U(x,o)$, and the subsequent distribution of the particles at time t gives an unbiased estimate of the solution functional $U(x,t)$ provided the trajectories can be simulated without statistical bias. The estimate will have to be suitably scaled as described in the paragraph headed "Normalization of Solution" later in this section.

The conservation principle can be thought of as a statement of the conservation of the individual particles in the simulation, where each particle is an entity of matter in the state space. That the particles are conserved during a Monte Carlo solution of a parabolic equation of the FP type can be seen from a converse argument.

Consider a diffusion process $x(t)$ whose trajectory is allowed to terminate (or begin) within the time interval $(0, T)$ of the parabolic equation. Then from the earlier discussion, this diffusion process violates the conservation principle inherent in the Chapman-Kolmogorov integral equation (3.2.1). Thus the conditional probabilities of the terminating diffusion process are not described by the Chapman-Kolmogorov equation, and as the FP equation is derived

from the Chapman-Kolmogorov equation, the statistics of the terminating diffusion process are not described by the FP equation of the diffusion process. The converse then must be true: if we are using simulation methods to obtain a statistical solution to a parabolic equation of the FP type, then the system $x(t)$ simulated must have continuous trajectories which do not terminate or begin during the time interval of interest. That is, the particles representing the realisations of $x(t)$ obey the conservation principle.

Matching of Coefficients to Find the Underlying Diffusion Process of the Parabolic Equation

Essentially we must match the coefficients of the terms of the parabolic equation (3.2.2) directly with the corresponding terms of (3.2.3). It is convenient to begin with the highest derivative and work down.

It was noted that the matrix of second order terms $[U_{x_i x_j}]$ has as coefficients the positive definite matrix $[\alpha_{ij}(x,t)]^{i}_{j}$ and we can simply set the terms $a_{ij}(x,t)$ of (3.2.3) equal to twice these quantities. The terms a_{ij} represent the diffusion coefficients of the system $x(t)$ and are directly related to the noise coefficient of the diffusion process (3.2.4) by the relation (2.1.3) which is

$$\sum_{k}^{m} F_{ik}(x,t)\, F^{T}_{kj}(x,t) \;=\; a_{ij}(x,t) \;=\; 2\,\alpha_{ij}(x,t). \qquad (3.2.5)$$

As $\alpha(x,t)$ is positive definite, then a real matrix $F(x,t)$ can always be found which satisfies (3.2.5), although the functional dependence on x may not be simple.

If $\alpha_{ij}(x,t)$ is a function of x, then the second derivative on the right hand side of (3.2.3) will contribute terms containing P_{x_i} and P to the equation, and these will have to be accounted for when we match the lower order terms of (3.2.3) and (3.2.2). Remembering this, we will choose $b_i(x,t)$, $i = 1, n$, of (3.2.3) to match all the first order terms P_{x_i} of (3.2.3) with the U_{x_i}

term of (3.2.2). These $b_i(x,t)$ then give us the drift terms of the simulated diffusion process (3.2.4)

$$f_i(x,t) = b_i(x,t), \qquad i = 1, n. \tag{3.2.6}$$

Having chosen b_i and a_{ij}, there is the possibility that proportional and constant terms of (3.2.2) will not be matched by the FP equation (3.2.3). If b_i is a function of x or a_{ij} has a second x partial derivative, then the FP equation will have a term proportional to P, which may, or may not, match that of (3.2.2), as we have no more free coefficients to choose. Further, if (3.2.2) has a constant term (i.e. a term not depending on U) then it cannot be matched by an equivalent term of the FP equation. This outlines the restrictions on the form of the parabolic equation such that it can be written in the FP form (3.2.3) and solved by a simulation method with conserved Monte Carlo trajectories. To sum up:

(i) If $\underline{L}(U)$ of (3.2.2) has coefficients of the second order derivatives which have a zero second x partial derivative, and $\underline{L}(U)$ has coefficients of the first order derivatives which do not depend on x, and $\underline{L}(U)$ has no proportional or constant terms, then (3.2.2) can always be put in FP form.

(ii) If $\underline{L}(U)$ has any constant terms, then (3.2.2) cannot be put in FP form.

(iii) Otherwise, it is a matter of chance whether the choice of a_{ij} and b_i will correctly match all terms of equation (3.2.3) to equation (3.2.2).

If the linear elliptic operator $\underline{L}(U)$ cannot be put in FP form by the matching of coefficients (3.2.5, 6) then it can be written as

$$\underline{L}(U) = L(U) + V(x,t)U + W(x,t), \tag{3.2.7}$$

where $L(U)$ are those terms of $\underline{L}(U)$ which are matched by the

choice of coefficients (3.2.5, 6) and VU and W are residual unmatched terms (called proportional and constant terms respectively). Thus if $\underline{L}(U)$ is of the FP form, then $\underline{L}(U) = L(U)$. If $\underline{L}(U)$ is not of the FP form, then V and/or W will be non-zero, and if a_{ij} or b_i are functions of x, then L(U) and VU may even have terms not in the original operator $\underline{L}(U)$ (e.g. see equation (3.2.69).

From limited experience, the author feels that many of the parabolic equations arising from physical phenomena of a diffusive nature will be of the FP form. A physical analogy for the heat conduction equation discussed later, will help us visualize what type of physical situations lead to parabolic equations not of this form. We will see later how to solve equations not of this form by Monte Carlo methods.

Positivity of Solutions

As we will be obtaining an approximate solution of U(x,t) of (3.2.2) by forming the probability density function of trajectories of a simulated diffusion process, it will be useful to know under what conditions the true solution U(x,t) is positive, given that the initial solution U(x,o) is positive everywhere in the finite subset R* of phase space R.[¤]

Theorem 3.2.2.

If $\left[\frac{\partial}{\partial t} - L\right]$ is a linear parabolic operator of the FP form, defined in R and time [0, T], it is an operator which preserves the positivity of functions U(x,t) in R* when

---

[¤] R* is that subset of R which includes all of R except for its infinite reaches where the solution U(x,t) is assumed to be zero. This allows us to bound certain functions in the proof to follow.

$$[\frac{\partial}{\partial t} - L]U = 0,$$

with the condition that $U(x,o)$ is positive everywhere in R^*, and that $U(x,t)$ has the smoothness property that U_t and U_{xx} exist everywhere in R for t in $(0, T]$.

Feller proves this in the one-dimensional case using semi-group theory of functional analysis [68]. We prove the theorem in the multi-dimensional case using the assumption of solution smoothness. The elegance of the semi-group approach is that this assumption is not necessary, but it is not known whether Feller's proof can be extended to more than one dimension.

Proof. We first show that the solution can never go negative in R^*, and then show it always is positive in R^*.

If the solution $U(x,t)$ is to become zero or negative sometime, there must be a unique time t_p and place x_p in R^* where the solution first becomes zero:[*]

$$U(x_p, t_p) = 0 \qquad (3.2.8)$$

We first show that if this situation exists, the solution has a positive increase and thus does not go negative, and later show that this situation can never exist.

Consider the value of the derivative U_t at (x_p, t_p). As x_p is the first solution point to become zero, the solution $U(x, t_p)$ is still positive in an arbitrarily small neighbourhood of x_p. Then as we have assumed that $U_{xx}(x_p, t_p)$ exists (i.e. it is not infinite), we have the simple conditions of a smooth local minimum of the function $U(x,t)$:

---

[*] There is no difficulty in extending the following argument to multiple points x_p, or even to a continuous manifold of points in R^*. In the latter case the argument to follow is applied to the edge of the manifold.

$$U_{x_i}(x_p, t_p) = 0, \quad i = 1, n, \tag{3.2.9}$$

and

$$U_{x_i x_i}(x_p, t_p) > 0, \quad i = 1, n. \tag{3.2.10}$$

However, there is a less obvious but more specific necessary condition for the existence of a smooth local minimum, and this involves the matrix of second partial derivatives

$$U_{xx} = [U_{x_i x_j}]. \tag{3.2.11}$$

The condition is that the determinants of U_{xx} and its $n - 1$ minors be all positive [69, p. 6], which is also a sufficient condition that the matrix U_{xx} be positive definite [70, p. 74].

Now, owing to the conditions (3.2.8, 9), the parabolic equation at (x_p, t_p) becomes

$$U_t(x_p, t_p) = \sum_{i,j}^{n} \alpha_{ij} U_{x_i x_j}, \tag{3.2.12}$$

where the matrix

$$[\alpha_{ij} U_{x_i x_j}] \tag{3.2.13}$$

is recognized as the Schur product of the matrices α and U_{xx}. It is known [71] that if α and U_{xx} are real, positive definite, symmetric matrices, their Schur product is real, positive definite and symmetric.* But the definition of the parabolic operator implies that α is real, positive definite and symmetric in R, and thus the matrix (3.2.13) is positive definite, and the sum of its elements is positive.**

Then from (3.2.12), $U_t(x_p, t_p)$ is positive, the solution increases at (x_p, t_p), and does not go negative. As the conditions (3.2.8-12) must immediately precede the first appearance of a

---

* The author is indebted to W. A. Murray of the Mathematics Division, National Physical Laboratory, for bringing this result to his attention.

** The sum of the elements is also positive if one of the matrices (eg. U_{xx}) is only positive semi-definite.

negative solution, a negative solution can never occur. We now go one step further and show that the zero solution (3.2.8) can never occur.

To see if a zero solution can occur at (x_p, t_p), consider the behaviour of U_t at $x = x_p$ for $t < t_p$. If the initially positive solution is to become zero at (x_p, t_p), the derivative $U_t(x_p, t)$ must be negative for at least a small time immediately preceding t_p:

$$U_t(x_p, t) < 0 , \qquad t_q \leq t < t_p, \tag{3.2.14}$$

as U_t is finite. The basis of our argument is to show that the variation of U_t is sufficiently bounded that there exists some $\epsilon > 0$, $\Delta > 0$, such that for

$$U(x_p, t_p - \Delta) = \epsilon , \tag{3.2.15a}$$

we have $$L(U(x_p, t_p - \Delta)) > 0 , \tag{3.2.15b}$$

which violates (3.2.14). This states that $U_t(x_p, t)$ must become positive before the zero solution is reached, in which case the zero solution is never reached. The property (3.2.15) requires the time continuity of U, U_x, U_{xx} and the eigenvalues of U_{xx} evaluated at $x = x_p$.

We assume the existence of the condition (3.2.14) leading to the zero solution (3.2.8). As U_t is finite, we can put

$$|U_t(x_p, t)| < \bar{K}_1, * \tag{3.2.16}$$

and we have $$U(x_p, t_p - \Delta) = \epsilon < \Delta \bar{K}_1, \tag{3.2.17}$$

---

* The symbols $\bar{K}_i$ all refer to positive bounded constants.

which states that the solution $U = \epsilon$ is continuous near zero. As U_t is also bounded in the neighbourhood of x_p, the elements of U_x and U_{xx} evaluated at $x = x_p$ are also continuous in time. That is, from (3.2.9) we have

$$|U_{x_i}(x_p, t_p - \Delta)| < \epsilon \bar{K}_2 \quad , i = 1, n, \tag{3.2.18}$$

and also $|U_{x_i x_j}(x_p, t_p - \Delta) - U_{x_i x_j}(x_p, t_p)| < \epsilon \bar{K}_3$,

$$i,j = 1,n. \tag{3.2.19}$$

Now the continuity condition (3.2.19) on the elements of U_{xx} implies that the eigenvalues of U_{xx} are continuous, for they are polynomial functions of the elements. Then the positive definiteness of U_{xx} is assured for U in a small region

$$\bar{\epsilon} > U \geqslant 0$$

for some positive $\bar{\epsilon}$.

We now consider ϵ in the range

$$\bar{\epsilon} > \epsilon > 0. \tag{3.2.20}$$

As α and U_{xx} are positive definite in this range of ϵ, there is some positive constant μ independent of ϵ such that

$$\sum_{i,j}^{n} \alpha_{ij} U_{x_i x_j} > \mu \tag{3.2.21}$$

for all ϵ in the range (3.2.20). Now consider the value of $U_t = L(U)$ at $(x_p, t_p - \Delta)$. We write $L(U)$ as

$$L(U) = \gamma U + \sum_{i}^{n} \beta_i U_{x_i} + \sum_{i,j}^{n} \alpha_{ij} U_{x_i x_j}. \tag{3.2.22}$$

The coefficients γ and β_i are functions of x and t and may not be continuous, but we do know they are finite in R^* as we have assumed U_t is finite in R^*. Then we can put

$$|\gamma| \quad < \quad \bar{K}_4, \tag{3.2.23}$$

and

$$|\beta_i| \quad < \quad \bar{K}_5 \qquad i = 1, n. \tag{3.2.24}$$

Now, using the inequalities (3.2.21, 17, 23, 18, 24), we can write (3.2.22) as the inequality

$$L(U(x_p, t_p - \Delta)) \; > \; \mu - \epsilon \bar{K}_4 - \epsilon \bar{K}_2 \; n \; \bar{K}_5, \tag{3.2.25}$$

for all ϵ in the range (3.2.20). The important point is that μ is positive and independent of ϵ, so that as we decrease ϵ ($= U(x_p, t)$) in the range (3.2.20), there comes a point where the r.h.s. of (3.2.25) is positive. Then (3.2.15) is true for some $\epsilon > 0$ and the condition (3.2.14) for a zero solution to occur is violated. Thus the initially positive solution can never go to zero at the arbitrary point x_p in R^*, and Theorem 3.2.2 stating

$$U(x, t) \; > \; 0 \; , \quad \text{all} \quad x \quad \text{in} \quad R^* \quad \text{and} \quad t \quad \text{in} \quad (0, T] \tag{3.2.26}$$

is proved.

Remarks on Theorem 3.2.2

(a) The theorem does not discuss the case where the initial conditions $U(x,o)$ in certain regions of phase space R are zero (but nowhere negative). Then assuming the initial conditions are not everywhere zero (the solution is then trivially $U(x,t) = 0$, all t), the region $\bar{R}$ where $U(x,o) = 0$ is enclosed by an $n - 1$ dimensional surface S where $U(x,o) > 0$ beyond S*. Then along S the conditions

$$U(S,o) = 0 \tag{3.2.27}$$

$$U_x(S,o) = 0 \tag{3.2.28}$$

and $U_{xx}(S,o)$ – positive (semi-) definite (3.2.29)

hold, and $U_t(S,o) > 0$. Then the surface S which bounds the zero solution region shrinks and the region $\bar{R}$ disappears, and

---

* Again we do not discuss the part of S belonging to the infinite reaches of R.

as the differential equation is first order in time, this happens instantaneously, as long as α is positive definite and U_{xx} exists everywhere.

Physically, the FP equation $U_t = L(U)$ describes the probability density of a continuous diffusion process $x(t)$. Zero initial conditions are possible for we may exclude the possibility of $x(o)$ being in $\bar{R}$ (e.g. $x(o)$ may be known exactly and $U(x,o) = \delta(x - x(o))$ then). Then as long as α is positive definite everywhere, the diffusive forces acting upon the process are nowhere zero, and the infinite character of the diffusive force ($x(t)$ always has an infinite velocity) assures that $x(t)$ can reach all regions of R instantaneously. Then the regions $\bar{R}$ which have an initially zero solution (i.e. probability density) develop a positive solution at $t = 0^+$ (e.g. see the solution (3.2.93)). Then Theorem 3.2.2 still holds as we have discussed the open time interval, t in $(0, T]$. This is an ideal situation, of course, and does not hold when $x(t)$ is simulated on a physical computer. Then a zero solution can exist for a finite time.

(b) The theorem does not discuss problems where U_{xx} does not exist everywhere in R for t in $(0, T]$. Then we can divide R into regions where U_{xx} exists in the interior of the regions, and show via the arguments of Theorem 3.2.2 that the solution is positive in the interiors of all the regions for all t in $(0, T]$. The solution $U(x,t)$ can be zero at the boundaries, for there we are free to specify certain conditions which may include specifying the solution itself (see Section 3.2.5 for examples).

(c) We have shown the positivity of solutions of n-dimensional parabolic equations of the FP type, given non-negative initial conditions, with the assumption that all the second order partial derivatives of the dependent variable exist everywhere. In Theorem 3.2.1 we stated that such an equation could be solved by a Monte Carlo method by forming the probability density of particles which

underwent simulated trajectories in phase space R while obeying a conservation principle. But a probability density function is, by definition, always non-negative, and as the trajectories always enjoy complete mobility (as α is positive definite and the lower order coefficients (3.2.23, 24) are bounded everywhere) the probability density of the trajectories is nowhere zero for $t > 0$. This agrees with Theorem 3.2.2 which shows that the true solution is positive for $t > 0$.

The two following positivity theorems follow directly from Theorem 3.2.2 and are relevant to the forms of parabolic equations discussed in Section 3.2.2. Once again we assume the smoothness conditions that U_t and U_{xx} exist everywhere in R.

Theorem 3.2.3 If $[\frac{\partial}{\partial t} - L]$ is a linear parabolic operator which preserves positivity in the manner described in Theorem 3.2.2, then the linear parabolic operator $[\frac{\partial}{\partial t} - L - V]$ also preserves positivity, where V is bounded in R and adds a proportional term $V(x,t)U(x,t)$ to the r.h.s. of the parabolic equation.

The operator L already has a proportional term (c.f. γU in (3.2.22)) and as V satisfies the condition (3.2.23), the proof of Theorem 3.2.2 covers Theorem 3.2.3 as well.

Theorem 3.2.4 If $[\frac{\partial}{\partial t} - L - V]$ is a linear parabolic operator which preserves positivity in the manner described in Theorem 3.2.3, then the linear parabolic operator $[\frac{\partial}{\partial t} - L - V - W]$, where W adds a non-homogeneous term $W(x,t)$ to the r.h.s. of the parabolic equation, preserves positivity if $W(x,t)$ is non-negative for all x in R and t in $[0, T]$.

The proof of Theorem 3.2.2 depended on showing that U_t was positive for some region arbitrarily close to $U = 0$. Then if W is non-negative, the addition of W to U_t cannot decrease U_t,

and the arguments of Theorem 3.2.2 still hold. Furthermore, the following corollary is clear.

Corollary If t_p is the first time in $[0, T]$ when $W(x,t)$ is negative for some x in R, then the solution $U(x,t)$ is positive for t in $(0, t_p]$, and may (but will not necessarily) be negative for t in $(t_p, T]$.

Normalization of Solution

In Theorem 3.2.1 we have stated that an approximate solution of

$$\left[\frac{\partial}{\partial t} - L \right] U = 0, \qquad (3.2.30)$$

where $\left[\frac{\partial}{\partial t} - L \right]$ is a linear parabolic operator of the FP type, is obtained by forming the probability density function of simulated trajectories of the underlying diffusion process. But an operator of this type is a linear homogeneous operator, and so $U(x,t)$ is also given by

$$\left[\frac{\partial}{\partial t} - L \right] \mu U = 0, \qquad (3.2.31)$$

where μ is an arbitrary non-zero constant. This means that solutions for $U(x,t)$ are equivalent to an arbitrary scaling factor. To pursue this matter, consider

Theorem 3.2.5

If $U(x,t)$ is defined by the linear transformation (3.2.30) of the FP type, then the normalization quantity $\phi(t)$ given by the integral

$$\phi(t) = \int_R U(x,t)\, dx \quad , \qquad t \text{ in } [0, T], \tag{3.2.32}$$

is independent of time, where R is the entire domain of x.

The proof follows directly from the principle of conservation of trajectories, and the interpretation of $U(x,t)$ as the probability density of trajectories, for then ϕ is proportional to the number of trajectories used in the simulation, which does not vary for t in $[0, T]$.

Corollary

The normalization quantity ϕ is given by the initial condition

$$\phi = \int_R U(x,o)\, dx. \tag{3.2.33}$$

This follows from the fact that the definition (3.2.32) holds for $t = o$, and the function $U(x,o)$ is an imposed condition on the solution of (3.2.30). This means that the initial conditions set the scale of the subsequent solution, and the quantity ϕ combined with the total number of particles used in the simulation determine the constant weighting factor we must give each particle when we are reducing the statistics of the trajectories to form an estimate of $U(x,t)$. That is, $U(x,t)$ is a smoothed version of

$$\sum_{\alpha=1}^{N} \phi\, N^{-1}\, \delta(x - x_\alpha(t)),$$

where N is the number of trajectories $x_\alpha(t)$ in the simulation with conserved trajectories, and $\delta(\cdot)$ is the Dirac delta function. More details of the solution procedure are given in Section 3.2.3.

3.2.2 Simulation of Parabolic Equations not of Fokker-Planck Form

We have seen from equations (3.2.5, 6) that in matching an arbitrary parabolic equation to the FP equation, we had not enough free parameters of the FP equation to match a general proportional term, or any constant term of the parabolic equation. Thus many parabolic equations cannot be put in FP form, and cannot be solved by a Monte Carlo method involving conservation of trajectories.

To solve the more general parabolic equation $U_t = \underline{L}(U)$ by Monte Carlo methods, we match as many terms of the parabolic equation as we can to the FP equation by the method of equations (3.2.5, 6). This operation defines the "nearest" FP equation of the parabolic equation (3.2.7); the parabolic equation differs from its nearest FP equation by the residual terms VU and W.

The nearest FP equation defines a diffusion process (3.2.4, 5, 6) which is called the underlying diffusion process of the parabolic equation. If the underlying diffusion process is simulated with conserved trajectories, the probability density of the simulated particles gives a Monte Carlo solution of the nearest FP equation but not of the given parabolic equation. However, if we modify the simulation by allowing the birth and death of particles to simulate the effects of the residual terms (violating the principle of conservation of trajectories), the probability density of the simulated particles then gives a Monte Carlo solution of the given parabolic equation. This idea stems from a suggestion of King [72] who seems to be the only person to previously solve parabolic equations by simulation in forward time. This section describes how the birth and death of the trajectories of the underlying diffusion process can simulate the effects of the residual terms.

Using the notation of equation (3.2.7) and Theorem 3.2.4, the general linear parabolic operator can be written as

$$\left[\frac{\partial}{\partial t} - L - V - W\right], \qquad (3.2.34)$$

where $\left[\frac{\partial}{\partial t} - L \right]$ is that part of the operator which can be put in FP form, and the residual operator terms V and W give rise to right hand side terms

$$V(x,t)\, U(x,t) + W(x,t). \tag{3.2.35}$$

The treatment of the V and W terms will be discussed separately below, but as their effects are additive, there is no problem with combining their treatment by superimposing the techniques discussed below.

Treatment of Proportional Term $V(x,t)\, U(x,t)$

We are concerned with the parabolic equation

$$U_t = L(U) + VU . \tag{3.2.36}$$

To observe the effect of the added term VU, consider the ordinary differential equation (with x a constant)

$$\frac{d\, U(t)}{dt} = V(t)\, U(t) \quad , \quad U(0) = U_o , \tag{3.2.37}$$

whose solution is

$$U(t) = U_o\, e^{\int_0^t V(s)\, ds} \tag{3.2.38}$$

However, as U is also affected by the operator L during the solution, we are not allowed to isolate the effect of V over the whole solution by using (3.2.38), but we can look at the incremental effect of the V term. Over a time increment Δ we have

$$U(t + \Delta) = U(t)\, e^{[V(t)\Delta + o(\Delta)]}$$

$$= U(t)\, [1 + V(t)\Delta + o(\Delta)],$$

or $U(t + \Delta) - U(t) = V(t)\, U(t)\Delta + o(\Delta)$. (3.2.39)

Over this interval, the change in $U(t)$ caused by the L operator is proportional to Δ, and thus only contributes an $O(\Delta^2)$ or $o(\Delta)$ term to the $VU\Delta$ term of (3.2.39) and thus its effect can be incorporated into the existing error term $o(\Delta)$. Thus (3.2.39) is a valid expression for the change in U caused by the V term over a time increment Δ. The above argument holds for all x, and so we may write (3.2.39) as

$$U(x, t + \Delta) - U(x,t) = V(x,t)\, U(x,t)\Delta + o(\Delta). \quad (3.2.40)$$

Now, how do we modify our simulation of $U_t = L(U)$ involving conserved trajectories of the underlying diffusion process to include the contribution of (3.2.40)? We note that the dependence of $U(x,t)$ on x is given by the density of simulated trajectories $x(t)$ in the space R. In particular, each trajectory $x_\alpha(t)$, $\alpha = 1, N$, contributes the constant function

$$\phi\, N^{-1}\, \delta(x - x_\alpha) \quad (3.2.41)$$

to $U(x,t)$, as the sum of all contributions to $U(x,t)$ must equal ϕ from (3.2.32).

Thus each conserved trajectory of weight $\phi\, N^{-1}$ can be thought of as representing a proportion of the solution $U(x,t)$, as the solution is proportional to local density of particles. But the effect of the VU term (3.2.40) is proportional to $U(x,t)$, and so we can associate the incremental change (3.2.40) with a change in each trajectory in the simulation. Thus the solution change $VU\Delta + o(\Delta)$ of (3.2.40), which is the cumulative effect of all trajectories, can be simulated by allowing each trajectory to contribute

$$V(x_\alpha, t)\, \phi\, N^{-1}\, \Delta + o(\Delta)$$

to this change. Thus a trajectory $x_\alpha(t)$ which at time t has a

weight ϕN^{-1}, should have a new weight of

$$[1 + V(x_\alpha, t)\Delta + o(\Delta)]\, \phi N^{-1} \qquad (3.2.42)$$

at time $t + \Delta$.

In the Monte Carlo simulations, this growth process can be mechanised in two ways. The first involves conservation of the N trajectories, and allows the weighting factor to vary with time during the course of the solution as suggested by equation (3.2.42). The weighting factor will now be different for each trajectory $x_\alpha(t)$, and we will denote it as $\Omega_\alpha(t)$. Then

$$\Omega_\alpha(o) = \phi N^{-1}, \qquad (3.2.43a)$$

and

$$\Omega_\alpha(t) = \Omega_\alpha(o)\, e^{\int_o^t V(x_\alpha, s)ds}. \qquad (3.2.43b)$$

This expression can be deduced by tracing back the growth of $U(t)$ to the contribution (3.2.38). The exponent in (3.2.43b) is essentially a path integral over the trajectory of $x_\alpha(s)$, and the fact that $x_\alpha(s)$ varies over the trajectory accounts for the effect of the L operator on the contribution of the VU term, which we had neglected in deriving (3.2.38). Note that the approximation given by the error $o(\Delta)$ in (3.2.39, 40, 42) has disappeared, but will reappear when the integral in (3.2.43b) is evaluated numerically.

This path integral method is essentially the same as that proposed by Gelfand and Yaglom [65] (and implemented by Little [64]) in connection with the backward operator method for handling proportional terms in the Schroedinger equation of mathematical physics. However the path integral may be difficult to evaluate, especially if $V(x,t)$ is non-linear in x, and, depending on the equipment at our disposal, the following method involving a violation of the principle of conservation of trajectories may be preferred.

Instead of allowing the weight to grow to the amount (3.2.42) during time Δ, we could keep the weight constant and allow the number of trajectories to grow at the same rate. As $U(x,t)$ is given by the sum of the weighted contributions of each trajectory, we are just transferring the change in U caused by VU from the weights to the trajectories themselves. The new trajectories have the same dynamics (3.2.4) as all other trajectories, and the weighting factor associated with each trajectory remains at ϕN^{-1}, where N is the original number of trajectories in the simulation. It is due to the fact that new trajectories undergo dynamic movement once they are initiated, that coupling occurs between the contribution of the VU term and the other terms of the parabolic operator L.

The main difficulty with adding trajectories is that trajectories are integral entities, and if (3.2.42) calls for a growth by the factor 1.02 during time Δ, we cannot add 0.02 of a trajectory to our simulation. We can, however, allow a new trajectory to begin at (x_α, t) with a probability 0.02, and repeat this every time interval Δ for every trajectory $x_\alpha(t)$, or we could allow a new trajectory to begin at (x_β, t) for every $(0.02)^{-1} = 50$ trajectories, where x_β is the centre of gravity of a localized group of 50 trajectories, and repeat this for every time interval Δ, and every group of 50 trajectories. There are obvious accounting difficulties with this last procedure (e.g. forming the localized groups and finding their centre of gravity), especially as the growth factor

$$1 + V(x_\alpha, t)\,\Delta \qquad (3.2.44)$$

varies with time, and with every trajectory.

Thus the first method of these two will be preferred, as it involves a single (and independent) probabilistic operation on each trajectory. This is to be compared with the previous method of evaluating the path integral of $V(x_\alpha, s)$ along each trajectory.

Note that if $V(x_\alpha, t)$ is negative, the probabilistic method must allow a probability of $-V(x_\alpha, t)\Delta$ that the trajectory $x_\alpha(t)$

should expire during time $[t, t + \Delta]$. This procedure must be applied at every time interval Δ and to all trajectories x_α existing at the present value of time. As this procedure is only applied to existing trajectories, it cannot result in the appearance of trajectories with negative weights (as when $W(x,t)$ terms exist in the parabolic equation; see later) and the positivity of the solution is preserved in agreement with Theorem 3.2.3.

The main disadvantage of the new probabilistic method is that it introduces a further degree of randomness to the Monte Carlo method, thus increasing the sampling error and possibly necessitating an increase in sample size. On hybrid computer implementations*, the path integral method will be more convenient unless the function $V(x_\alpha, t)$ is difficult to generate. On the digital computer installation which was available to the author (an IBM 7090), the allowance for the change in the number of trajectories involved only a simple extension to the existing accounting and data reduction operations, and was considerably simpler than the evaluation of path integrals. Furthermore, although modern hybrid computers, such as the statistical repetitive special purpose machine (ASTRAC II) built by Korn [62], seem very suited to the Monte Carlo solution of partial differential equations, it is unlikely that this type of computer will become widely available for the time being. This is compared with the present availability of large, fast, digital computers, and it is suggested that the method proposed in this section is more suitable for digital computers.

Treatment of Constant Term $W(x,t)$

We are concerned with the parabolic equation

---

* The trajectory simulation is carried out on the analogue part, and the control of initial conditions and collection of statistics is done on the digital part of the hybrid computer [62]. In addition, the generation of $V(x_\alpha, t)$ and the evaluation of its integral, can be done on the analogue part [64]. The allowance of the probabilistic change in the number of trajectories during the simulation interval $[0, T]$ may result in an unreasonable increase in the control function work of the digital part.

$$U_t = L(U) + W, \tag{3.2.45}$$

and our argument will closely follow the treatment of the proportional term VU. The main difference is that the contribution of W to U_t is not proportional to U(x,t), and thus cannot be associated with the existing trajectories of the simulation.

Analogous to (3.2.38) we find that the contribution to U(t) of W(t) is

$$U(t) = \int_0^t W(s)\, ds, \tag{3.2.46}$$

but this contribution cannot be simply evaluated for U(x,t) and W(x,t) at a given x, as this would ignore the coupling effect of the other terms in the parabolic equation. As the contribution (3.2.46) is independent of the trajectories, we cannot use a path integral method of its evaluation, but we can use a method involving a change in the number of trajectories as before. In [67, Ch.13, Section 4], Dynkin shows that a path integral type method can be used to handle constant (non-homogeneous) terms in the solution of elliptic differential equations by Monte Carlo methods derived from the backward operator. The implementation of this technique for elliptic equations has been discussed by Little [64] and Handler [83], but there is no indication in these current research works that the backward simulation method can be adapted to handle parabolic equations with non-homogeneous terms.

Analogous to equation (3.2.40), the contribution of W(x,t) to U(x,t) over time $[t, t + \Delta]$ is given by

$$U(x, t + \Delta) - U(x,t) = W(x,t)\Delta + o(\Delta). \tag{3.2.47}$$

We will simulate this increase (or decrease) of U(x,t) by beginning a trajectory at (x,t) with a positive (or negative) weight ϕN^{-1} attached to it. But a trajectory $x_\alpha(t)$ is really a contribution to $U(x_\alpha, t)$ in the region around x_α (as the trajectory exists at a discrete x point, but U(x,t) is continuous over x),

and so we must include a space scaling factor in the contribution of a new trajectory to (3.2.47).

Consider a unit element in the x space, R, and let $W(x,t) = W(t)$ over this element. Let there be β new trajectories which originate in the unit element in unit time. Then from (3.2.47) the increase in $U(x,t)$ during Δ is given by

$$U(x,\ t + \Delta) - U(x,t) \doteq W(t)\Delta = \phi\, N^{-1}\, \beta\, \Delta. \qquad (3.2.48)$$

Thus the space scaling factor β equals

$$\beta = N\, W(x,t)\, \phi^{-1} \qquad (3.2.49)$$

trajectories per unit time per unit space, where N is the original number of trajectories in the simulation and ϕ is the initial condition scaling factor (3.2.33).

Thus the simulation of the contribution of $W(x,t)$ to the parabolic equation (3.2.45) is achieved by the addition of $N\, W(x,t)\, \phi^{-1}$ trajectories per unit time and per unit space to the simulated trajectories. This operation can take place with a variety of space-time discretizations (that is, we can add particles infrequently to points in a fine space mesh δx, or more frequently to points in a coarser space mesh), with the resultant errors being $o(\delta x)$ and $o(\Delta)$.* The compromise involved must be chosen with regard to the trajectory dynamics (for example if the diffusive force is high, then the space discretisation will be less important), and the space-time resolution required of the solution.

If $W(x,t)$ is negative, then the trajectories added must be given negative weight factors, $-\phi\, N^{-1}$. Thus we can have negative solutions $U(x,\ t_2)$ only if $W(x,\ t_1)$ is negative for some $t_1 < t_2$, in agreement with Theorem 3.2.4. These negative trajectories can be cancelled with positive ones when they come arbitrarily close to each other, but this may or may not be convenient depending on the mechanics of the simulation and data collection procedures (3.2.50).

---

* The $o(\delta x)$ arises from taking $W(x,t)$ as $W(x_\mu,\ t)$, where x_μ is an arbitrary point in the appropriate space mesh.

3.2.3 Solution Procedure

In the last two sections, we discussed the Monte Carlo solution of a general linear parabolic equation. In this section we will summarize the methods already presented, and be more specific about the mechanics of obtaining the solution. Later we will discuss some computing experience related to two interesting aspects of solution - the treatment of spatial discontinuities and boundary conditions.

We are concerned with solving the general linear parabolic equation (3.2.2) where the right hand side contains first and second partial space derivative terms, and proportional and constant terms as well. The procedure is to relate this equation as far as possible to the FP equation (3.2.3) by choosing the coefficients a_{ij} and b_i to match the first and second partial derivative terms of the parabolic equation to those of the FP equation. The coefficients a_{ij} and b_i define the underlying diffusion process (3.2.4,5,6), which is closely related to the physical structure generating the parabolic equation.

If, following this matching operation, there are no terms left over in the parabolic equation (we call these residual terms), then we say the parabolic equation is of FP form, or more specifically, the parabolic equation is the FP equation of the underlying diffusion process. As the solution of the FP equation is the first order probability density of the diffusion process (with appropriate initial conditions), then the parabolic (FP) equation can be solved by collecting the appropriate statistics of a simulation of the diffusion process. Furthermore we have shown that the implications of transition probabilities inherent in the derivation of the FP equation require that the simulated trajectories of the diffusion process be continuous in space, and continuous over the time interval $[0, T]$. This has been called the principle of conservation of trajectories, and means that of all N trajectories which begin the simulation at $t = 0$, these N and no others exist at the end. We have seen that this implies that the normalization quantity $\phi(t)$ defined in (3.2.32) remains constant in $[0, T]$.

If, after the matching operation, residual terms of the parabolic equation do exist, then we must alter our method to take account of the residual terms by allowing the non-conservation of trajectories. These residual terms will be terms proportional to $U(x,t)$ or constant terms. We still simulate the underlying diffusion process given by the matching coefficients a_{ij} and b_i, and begin the simulation with N trajectories, but we allow these trajectories to grow or expire in $[0, T]$ to account for the residual proportional term, and/or allow independent trajectories to be added (with positive or negative weights) in $[0, T]$ to account for the residual constant term.

In this case, the number of trajectories existing for $t > 0$ will not in general equal N and the normalizing quantity $\phi(t)$ will vary over $[0, T]$. The weight associated with each trajectory will, however, remain constant over $[0, T]$ and is given by $\pm \phi N^{-1}$, where $\phi = \phi(0)$ is given by the normalizing quantity of the initial conditions (3.2.33).

This treatment of the residual terms was derived by assuming that the residual terms had independent effects on the solution over incremental time steps. The implementation of the treatment of residual terms involved discretization errors, which are $o(\Delta)$ for the proportional term, and $o(\Delta)$ and $o(\delta x)$ for the constant term, and it was noted that extra statistical errors are introduced by the implementation. It was also noted that the interdependence of the residual and other terms was represented by the subsequent diffusion of the introduced trajectories.

Justification was added to the methods presented by showing that they satisfied certain positivity theorems. In essence, the only possibility of a negative solution appearing is when the residual constant term $W(x,t)$ is negative in some (x,t) region of the solution.

The Formation of Point and Functional Solutions

We have already stated that the solution $U(x,t)$ is obtained from statistical estimates of the simulated Monte Carlo trajectories. More details will be given in this sub-section.

The simulation begins by taking N particles (often several thousand), and placing them in the space R so that their density corresponds to the initial solution $U(x,o)$. Then each particle is allowed to move according to the dynamics (3.2.4-6), where an independent noise vector $dw(t)$ is used for the trajectory of each particle. The simulation of these trajectories is discussed in Chapters 4, 5 and 6. If residual terms exist in the parabolic equation, then particles are added and deleted according to the methods discussed in Section 3.2.2., and the data collection routine is adjusted to cover all existing trajectories.

Whenever information on $U(x,t)$ is required, we note the positions of the particles at time t, and collect the appropriate statistics. For example, if an estimate of $U(x_p,t)$ is required for a particular point x_p, then we choose a small region R' around x_p such that the solution $U(x',t)$, x' in R', is relatively uniform, and we have

$$U(x_p,t) \doteq \frac{\sum_{x_\alpha} \pm \phi N^{-1}}{\int_{R'} dx} \qquad (3.2.50a)$$

where the sum of the x_α includes all those trajectories $x_\alpha(t)$ in R', and $\int_{R'} dx$ is the n-dimensional "volume" of the region R'. The $\pm$ weight is normally positive, but the negative sign is needed to allow for the possibility of negative trajectories introduced by a negative $W(x,t)$ term (3.2.47).

Thus $U(x_p,t)$ is taken as a local (in the neighbourhood R') approximation to the scaled density of trajectories at x_p. In particular we have

$$U(x_p,t) = \phi N^{-1} D(R') \qquad (3.2.50b)$$

where $D(R')$ is the average density of net particles* in the region R'. By taking R' large we lose solution resolution, but by taking R' small, we lost solution accuracy, and so a compromise is chosen. If $U(x_p,t)$ is obtained at a number of adjacent (often equally spaced) points x_p, then the resolution/accuracy trade-off can be further influenced by smoothing the data points obtained. **

In particular, the accuracy of $U(x_p,t)$ is given explicitly by the binomial distribution in terms of the true solution $U(x,t)$. Suppose for a given neighbourhood R' of x_p, we have

$$\frac{\int_{R'} U(x,t)\,dx}{\int_{R} U(x,t)\,dx} = q(t), \qquad (3.2.51)$$

where $q(t)$ is the proportion of the solution density in R', at time t, and so equals the expected proportion of simulated particles in R'. Assuming the simulation to have an unbiased error, then each trajectory can be considered as an independent trial with a probability $q(t)$ of being in R' at time t. Then out of N' trials*, the expected number of trajectories in R' at time t is

$$q(t)\,N', \qquad (3.2.52a)$$

and the variance of the number is

$$q(t)\,[1 - q(t)]\,N'. \qquad (3.2.52b)$$

The Monte Carlo solution (3.2.50) for $U(x_p,t)$ obtained by an averaging operation in the region R' is really forming an estimate $\underline{q}(t)$ of $q(t)$ by finding the proportion of net trajectories in R'. In fact, $\underline{q}(t)$ is a maximum likelihood estimate of $q(t)$, and is in general the best estimate of $q(t)$ if no prior information of the solution is available [79, Ch.8] (in discussing filtering or smoothing

---

* The number of trajectories or trials in the simulation (or in a given region R') is always to be interpreted as the net number: the number of negative trajectories arising from a negative $W(x,t)$ is to be subtracted from the number of positive trajectories. The following error analysis assumes that the number of negative trajectories is small compared with the number of positive ones.

** See Section 6.3.

techniques later, we shall see that the operation of filtering data is similar to including prior information).

The estimate of $\underline{q}(t)$ is unbiased, for

$$E\left[\underline{q}(t)\right] = q(t), \tag{3.2.53}$$

and has a variance

$$\text{Var}\left[\underline{q}(t)\right] = \sigma_q^2 = \frac{q(t)\left[1 - q(t)\right]}{N'} . \tag{3.2.54}$$

Assuming N' is large, and $q(t)$ not too small so $q(t)\,N'$ is also large, the binomial distribution for $\underline{q}(t)$ can be approximated by a Gaussian distribution with the same mean and variance (3.2.53, 54), and confidence limits for $q(t)$ given a particular estimate $\underline{q}(t)$ can be readily evaluated [79, Ch.11]. These are found from the tails of the Gaussian distribution, and for example we have

$$\text{Prob}\left[\left|q(t) - \underline{q}(t)\right| < 1.96\,\sigma_q\right] = 0.95, \tag{3.2.55}$$

or
$$\left[\underline{q}(t) - 1.96\,\sigma_q\ ,\ \underline{q}(t) + 1.96\,\sigma_q\right] \tag{3.2.56}$$

is the 95% confidence interval for the true solution $q(t)$ given the estimate $\underline{q}(t)$. Of course, σ_q of (3.2.54) is not known exactly as $q(t)$ is not known, but the use of the estimate $\underline{q}(t)$ for $q(t)$ in (3.2.54) is justified under the assumption of large sample sizes, N' [79, p.262].

The confidence interval can be used to choose the number of trajectories N' in the simulation and/or the size of the region R', for the accuracy depends on the expected number of trajectories in R' (3.2.52a). As an accuracy estimate, the confidence interval can be used as an absolute error interval

$$\pm\ 1.96\ \sigma_q\ \frac{N' \not\oslash N^{-1}}{\int_{R'} dx} \tag{3.2.57}$$

or as a relative accuracy measure

$$\frac{1.96\,\sigma_q}{q(t)} \quad . \tag{3.2.58}$$

The following table gives these accuracy estimates for a varying number of trajectories N', and for two neighbourhood sizes which have values of q(t) of 0.1 and 0.01. For the non-conservation simulations, the value of N' may be difficult to estimate beforehand, but using N in its place should give an adequate accuracy estimate.

N'	q(t) = .1			q(t) = .01		
	σ_q	$1.96\,\sigma_q$	$\frac{1.96\,\sigma_q}{q}$	σ_q	$1.96\,\sigma_q$	$\frac{1.96\,\sigma_q}{q}$
100	.03	.06	.6	.01	.02	2
400	.015	.03	.3	.005	.01	1
1,600	.0075	.015	.15	.0025	.005	.5
5,000	.004	.008	.08	.0013	.0026	.26
20,000	.002	.004	.04	.0006	.0013	.13

Table 3.2.1 Estimates of Point Accuracies

It is noted that the accuracy estimates are proportional to $(N')^{-\frac{1}{2}}$, and are also proportional to $[q(t)]^{-\frac{1}{2}}$ for small q(t). The latter factor illustrates the resolution/accuracy compromise, for maximum accuracy (for a given N') is obtained when $q(t) = \frac{1}{2}$, at which point the resolution is a minimum.

A variation on the point solution (3.2.50) is the regional solution

$$\text{Prob}\,[x(t) \text{ in } R'] \text{ or } \int_{R'} U(x,t)\,dx. \tag{3.2.59}$$

The solution procedure is identical to the point solution with similar accuracy considerations except that no space quantization

error exists in this case. However, the type of solution to which the forward simulation method is best suited, is a functional solution of the form

$$g(t) = \int_R G(x)\, U(x,t)\, dx, \qquad (3.2.60)$$

for a given function $G(x)$. An example would be the solution for the moments of $x(t)$. The value $g(t)$ is estimated by

$$\underline{g}(t) = \sum_{x_\alpha} \pm \phi\, N^{-1}\, G(x_\alpha(t)) \qquad (3.2.61)$$

where the sum is taken over all trajectories and as before the sign $(\pm)$ of the weight $\phi\, N^{-1}$ is determined by the origin of the trajectory. As mentioned previously (3.1.2), the standard deviation of $\underline{g}(t)$ is $(N')^{-\frac{1}{2}}$ times the standard deviation of $G(x(t))$. This means that for a relative error of 10%, only 100 trajectories need be simulated, whereas for 10% relative error in the point solutions (see Table 3.2.1), many thousand trajectories will often be needed.

In comparison, the backward simulation method [63, 64] is inefficient for calculating functional solutions (3.2.60), but quite efficient for calculating point solutions (3.2.50). In fact, the efficiencies of the methods are complementary: the backward method is as efficient for single point solutions as the forward method is for functional solutions, and the backward method is as inefficient for functional solutions as the forward method is for single point solutions. As the number of desired solution points increases in the point type solution, the efficiency of the forward simulation method increases relative to that of the backward method, and the break-even point is likely at about ten solution points.

Computing Experience

The methods of this chapter were tested on some FP type equations such as the example of Chapter 2, as well as some variations of the heat conduction parabolic equation. The interesting aspects of the simulation exercise are given in Chapters 5 and 6.

The accuracy of the estimates agreed with the theoretical ones given earlier. This was shown by Chi-squared tests on the variance of the estimates obtained. This is one indication that the trajectories used are independent and unbiased. In the case of the example of Chapter 2, the solutions were found to be unbiased, with the error deviation as discussed above.

It is difficult to discuss computing times for complete solutions with the Monte Carlo method as the solution time depends on many factors. To simulate the one-dimensional system of Chapter 2, approximately 30 seconds were needed on the IBM 7090 to integrate 1000 trajectories over the time interval $[0, 2]$ using steps of 0.1.* This gave 3% accuracy for functional solutions of the form (3.2.60) but only 20% accuracy for point solutions of the form (3.2.50).

This computing time is only given as a rough example, but what is more interesting is the effect of system dimensionality and data reduction requirements on computing time. The system dimensionality or complexity only affects computing time insomuch as the time needed to integrate the differential equations are concerned. On an analogue (or hybrid) computer, the system equations are integrated "in parallel", and so provided the analogue computer has sufficient capacity, the computing time for the simulation is independent of system complexity. On a digital computer, each term of the system differential equation is handled "in series", and so the computing time is proportional to the number of additions, multiplications, etc. involved in the system equations. However, in some installations

---

* In contrast, a modern hybrid computer such as Astrac II [62] could integrate the 1000 trajectories in one second.

the simulation may be limited by the magnetic tape manipulation involved in processing large numbers of Gaussian random numbers, in which case the total computing time will be somewhat less than proportional to the system complexity, provided the number of system noise inputs remains constant. On the IBM 7090, about one third of the simulation time was used for tape manipulation for the one dimensional example of Chapter 2, but this ratio would be higher for computers with less high speed storage capacity than the IBM 7090's 32,000 word capacity.

Thus in contrast with the methods of Chapter 2, the Monte Carlo simulation method of determining a system's statistical behaviour is not seriously affected by the dimensionality or complexity of the system equations. From a limited computing experience, it was found that Monte Carlo methods were not competitive with direct methods for one-dimensional examples, but were slightly better than finite difference methods in two dimensions. It follows that Monte Carlo methods would have a substantial advantage over finite difference methods for three-dimensional problems, although the amount of data reduction necessary in the Monte Carlo method must be considered. As mentioned before, Monte Carlo methods are likely to be the only ones at our disposal for systems of dimensionality higher than three.

3.2.4 Example: The Heat Conduction Equation

To illustrate the physical analogies which can be drawn between forward trajectory simulation methods and the solution of parabolic equations, the heat conduction equation will be discussed. The basic equation governing the conduction of heat in isotropic solids is [77]

$$U_t = \frac{1}{pc} \sum_{i}^{3} \frac{\partial}{\partial x_i} \left[k_i U_{x_i} \right], \qquad (3.2.62)$$

where $U = U(x,t)$ = temperature, degrees Kelvin

x = three dimensional space variable

$p = p(x,t)$ = density of the solid, gm/cm^3

$c = c(x,t)$ = specific heat of the solid, calories/(gm ^oK)

$k_i = k_i(x,t)$ = thermal conductivity in direction of x_i, calories/(sec.cm^2 (^oK/cm))

$K_i = K_i(x,t) = k_i(pc)^{-1}$ = diffusivity, cm^2/sec.

We will solve this equation by simulating a diffusion process, and interpreting the solution $U(x,t)$ as the density of simulated trajectories $x(t)$. This means that the particles have units of temperature times volume ($^oK\ cm^3$), the scaling factor being ϕN^{-1}.

As a physical analogy, the trajectories or particles can be considered as specific energy particles, for each particle raises the temperature of one cm^3 of material ϕN^{-1} degrees K. Note that this simulated particle is not the same as a quantum of heat energy (a calorie, for example), for the rise in temperature caused by the addition of one calorie to one cm^3 of material is equal to $\frac{1}{pc}\ ^oK$, whereas the effect of one of the simulated particles on temperature is independent of pc. Thus it is useful to think of one particle at (x,t) as being equivalent to ϕN^{-1} pc calories, although, in reality, the relationship between heat (calories) and temperature is only an incremental relationship. This analogy will be used later to choose or check conservation and continuity conditions, for the only physical law at our disposal is the conservation of heat energy, and the dependence of heat flux on temperature gradient.

To match the heat conduction equation (3.2.62) to the FP form (3.2.3), we expand the differential of the right hand side of (3.2.62) and get

$$U_t = \sum_i^3 \left[\frac{1}{pc} \frac{\partial k_i(x)}{\partial x_i} U_{x_i} + \frac{k_i}{pc} U_{x_i x_i} \right]. \qquad (3.2.63)$$

Matching the second order terms as in (3.2.5) we find

$$\begin{aligned} a_{ij}(x,t) &= 2\,K_i(x,t) \quad , \quad i = j = 1,\ 3, \\ &= 0 \quad , \quad i \neq j, \end{aligned} \tag{3.2.64}$$

where $K_i = k_i(pc)^{-1}$. As the matrix $a(x,t)$ is diagonal in this case, the noise coefficient $F(x,t)$ of the underlying diffusion process (3.2.4) is just the square root of (3.2.64).

Now we must match the first order terms of (3.2.63) with those of the FP equation (3.2.3), which are

$$-\sum_{i}^{3} b_i(x,t)\,P_{x_i} + \sum_{i}^{3} \frac{\partial a_{ii}(x,t)}{\partial x_i}\,P_{x_i}. \tag{3.2.65}$$

We do this by setting

$$b_i(x,t) = 2\frac{\partial K_i(x)}{\partial x_i} - \frac{1}{pc}\frac{\partial k_i(x)}{\partial x_i}, \quad i = 1,3, \tag{3.2.66}$$

which specifies the drift terms (3.2.6) of the underlying diffusion process (3.2.4).

We have now matched the parabolic equation (3.2.62) to the FP equation (3.2.3) as best we could, and found the diffusion process underlying the parabolic equation to be

$$\begin{aligned} dx_i(t) &= \left[\, 2\frac{\partial K_i(x)}{\partial x_i} - \frac{1}{pc}\frac{\partial k_i(x)}{\partial x_i} \,\right] dt \\ &\quad + (2\,K_i)^{\frac{1}{2}}\, dw_i(t), \quad i = 1,\ 3. \end{aligned} \tag{3.2.67}$$

Following this matching, however, the FP equation (3.2.3) has the following residual terms left over:

$$\left[\, \frac{\partial}{\partial x_i}\left(\frac{1}{pc}\frac{\partial k_i(x)}{\partial x_i}\right) - \frac{\partial^2 K_i(x)}{\partial x_i^{\,2}} \,\right] P \quad , \quad i = 1,3. \tag{3.2.68}$$

In Section 3.2.2, we presented a first order analysis, which discussed the MonteCarlo treatment of proportional terms VU (3.2.36) which were residual terms existing in the parabolic equation. As the analysis was first order (that is, in Δ), the analysis is valid

if the residual terms (3.2.68) are taken from the FP equation, and placed in the parabolic equation with a change of sign. This gives us the residual term coefficient

$$V(x,t) = \sum_i^3 \left[\frac{\partial^2 K_i(x)}{\partial x_i^2} - \frac{\partial}{\partial x_i} \left(\frac{1}{pc} \frac{\partial k_i(x)}{\partial x_i} \right) \right]. \qquad (3.2.69)$$

In the above analysis, we have allowed the generality of p, c and k being functions of x (the t dependence does not complicate the analysis). For most heat conduction problems, all or some of these parameters will be independent of x, in which case the following simplifications occur (we will write down the equations equivalent to (3.2.64, 66 and 69):

Case I: p or c depends on x; k does not.

$$a_{ii}(x,t) = 2 K_i(x,t), \qquad i = 1, 3, \qquad (3.2.70a)$$

$$b_i(x,t) = 2 \frac{\partial K_i(x,t)}{\partial x_i}, \qquad i = 1, 3, \qquad (3.2.70b)$$

$$V(x,t) = \sum_i^3 \frac{\partial^2 K_i(x,t)}{\partial x_i^2}. \qquad (3.2.70c)$$

Case II: k depends on x; p and c do not.

$$a_{ii}(x,t) = 2 K_i(x,t), \qquad i = 1, 3, \qquad (3.2.71a)$$

$$b_i(x,t) = \frac{\partial K_i(x,t)}{\partial x_i}, \qquad i = 1, 3, \qquad (3.2.71b)$$

$$V(x,t) = 0. \qquad (3.2.71c)$$

Case III: k, p and c independent of x.

$$a_{ii}(x,t) = a_{ii} = 2K_i, \qquad i = 1, 3, \qquad (3.2.72a)$$

$$b_i(x,t) = 0, \qquad i = 1, 3, \qquad (3.2.72b)$$

$$V(x,t) = 0. \qquad (3.2.72c)$$

Note that when p and c are independent of x, no residual terms exist, and the conduction equation is solved by a simulation involving conserved trajectories. But we have seen earlier that in this case there is a constant relationship between the trajectories and quantums of heat energy, and so the conservation of trajectories in this case agrees with the physical law of conservation of heat energy (assuming the material has no sources or sinks of heat energy).

Also note that the equations above are only valid for sufficiently smooth functions $K_i(x,t)$ so that the first and second x_i partial derivatives exist and are bounded. If $K_i(x,t)$ is discontinuous with respect to an x_i variable, then the parabolic equation does not hold along the discontinuity. The simulation can be carried out in the separate regions of continuity, and special boundary conditions (to be discussed later) can be applied at the discontinuity to make the simulation consistent with the interpretation of the parabolic equation.

In equation (3.2.67) and the subsequent special cases, we have shown what diffusion processes must be simulated in order to solve the heat conduction equation by the Monte Carlo method of this chapter. We have also seen that if p or c is dependent on x, then the simulation will have to be modified to account for the residual VU term (3.2.69). Apart from the aspects mentioned in Chapters 5 and 6, the mechanics of the simulation are quite

straightforward and will not be mentioned here, but it will be worthwhile to discuss the effects on the simulation of auxiliary properties of the parabolic equation: spatial discontinuities and spatial boundary conditions.

3.2.5 The Treatment of Spatial Discontinuities and Boundary Conditions in the Parabolic Equation

Spatial Discontinuities

Although spatial discontinuities in the solution of the parabolic equation $U(x,t)$ cannot exist inside the domain R of the equation (as we have assumed that the second order coefficient $\alpha(x,t)$ vanishes nowhere in R), discontinuities in U_{x_i} and $U_{x_i x_j}$ can exist in R, and these occur where the coefficients of the parabolic equation, or their space derivatives which appear in (3.2.67 or 69), are discontinuous.

First consider genuine FP equations which describe the statistics of a given diffusion process (3.2.4). In this case any discontinuities in the FP equation come directly from discontinuities in the dynamics of the system (3.2.4). An example would be the relay system

$$dx(t) = f(\text{sign } (x), t)dt + dw(t). \qquad (3.2.73)$$

When a trajectory $x_\alpha(t)$ reaches the boundary $x = 0$, the dynamics of the particle undergo a step change according to (3.2.73), but physical considerations tell us that the trajectory $x_\alpha(t)$ is conserved at the boundary, and no special conditions are to be applied there. Thus the parabolic (FP) equation is solved by simply simulating the system (3.2.73) and collecting the appropriate statistics of the simulation.

Now consider a general parabolic equation whose origin has not been a specific dynamic system, but by the method of this chapter,

we have derived a dynamic system which underlies the parabolic equation. Although simulation of this underlying system (with modification if residual terms are present) does give us a Monte Carlo solution to the parabolic equation, discontinuities in the coefficients of the parabolic equation do not specify the behaviour of the simulated trajectories when a boundary of discontinuity is reached, in the same way as with the FP equation. This is because the physical meaning of the general parabolic equation is not as directly connected with the simulated trajectories as in the case of the FP equation. Thus in the general case we must rely upon whatever physical principles are inherent in the parabolic equation to construct suitable conditions for the trajectories to satisfy at the discontinuities. In the heat conduction example, the conservation of thermal energy or the continuity of thermal flux will furnish these conditions.

In the heat conduction example, we will assume a discontinuity occurs along a boundary which is a two-dimensional surface in the solid. To simplify our argument, we will rotate the x coordinates (if necessary) so that the x_1 coordinate is locally normal to the surface of the discontinuity. We shall regulate the flow of simulated particles across this boundary so that thermal energy is conserved, and thermal flux (heat flow) is continuous and the correct magnitude.

Consider the material discontinuity in the x_1 direction occurring across the boundary shown in Figure 3.2.2. We assume that the solution derivative U_{x_1} and the material parameters k_1, p and c to be independent of x_1 in a small region on either side of the discontinuity (this assumption is consistent with the first order analysis of particle flux to follow). These values do of course change abruptly at the discontinuity boundary, and we use the superscripts $-$ and $+$ to denote the values to the left of and to the right of the boundary respectively.

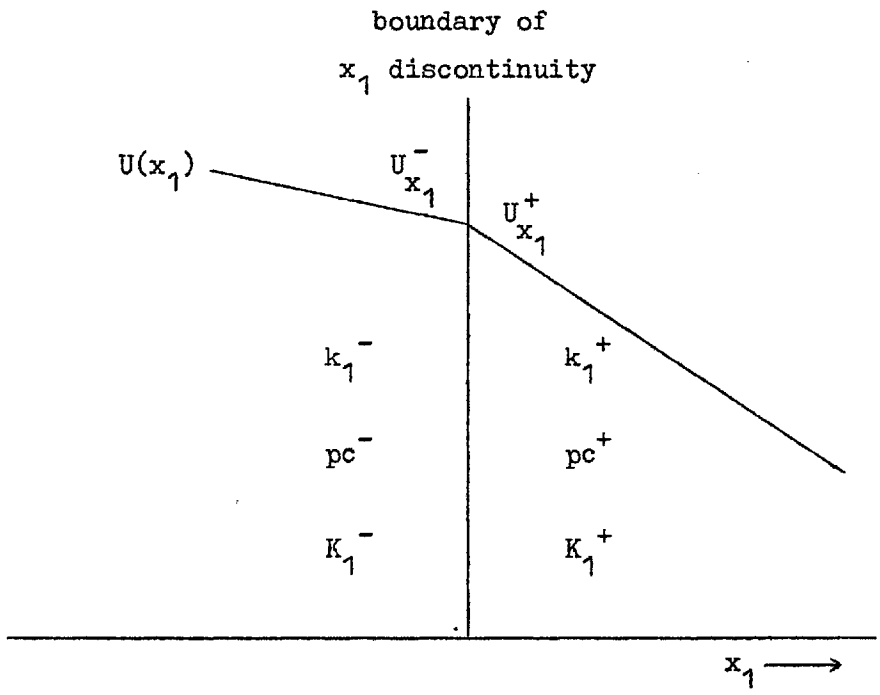


Figure 3.2.2 Solution of Heat Conduction Equation Across a Discontinuity

From the basic principles of heat conduction [77], the heat flux in the direction of x_1 is proportional to the x_1 temperature gradient, and is given by $-k_1 U_{x_1}$ calories/(sec.cm^2).* Assuming no sources of sinks of heat exist at the boundary, then this flux vector must be continuous across the boundary and we have

$$- k_1^- U_{x_1}^- = - k_1^+ U_{x_1}^+ \text{ calories/(sec.cm}^2\text{).} \qquad (3.2.74)$$

Thus the ratio of the thermal conductivities across the discontinuities specifies the ratio of the temperature gradients on either side of the boundary.

---

* In Figure 3.2.2, U_{x_1} is negative, and hence $-k_1 U_{x_1}$ is a positive heat flux to the right (the positive x_1 direction). Thus the heat flow is <u>down</u> the gradient of temperature.

Using our physical analogy of one calorie of heat energy equalling $\phi^{-1} N(pc)^{-1}$ simulated trajectories, the thermal flux (3.2.74) is simulated by the particle flux of

$$- \phi^{-1} N(pc^-)^{-1} [k_1^- U_{x_1}^-] = - \phi^{-1} N(pc^+)^{-1} [k_1^+ U_{x_1}^+]$$

$$\text{particles/(sec.cm}^2\text{)} \qquad (3.2.75)$$

Now consider the simulation of the situation of Figure 3.2.2. The solution $U(x_1)$ and its gradients $U_{x_1}^-$ and $U_{x_1}^+$ is simulated by a trajectory density $D(x_1)$ particles/cm^3, and local density gradients $D_{x_1}^-$ and $D_{x_1}^+$ on either side of the boundary, through the relationship (3.2.50b), where the units of ϕ are [$^\circ$K cm^3] and of N are [particles]. Thus from (3.2.75), the flux of trajectories across the boundary should be

$$- (pc^-)^{-1}[k_1^- D_{x_1}^-] = - K_1^- D_{x_1}^- \text{ particles/(sec.cm}^2\text{).} \qquad (3.2.76)$$

$$\text{or} \qquad = - K_1^+ D_{x_1}^+ \text{ particles/(sec.cm}^2\text{).}$$

In Appendix C we have derived the behaviour of the particles in the simulation in the x_1 direction adjacent to the boundary, with the assumptions that the material parameters p, c and k are constant for a small region on either side of the boundary. In particular, we derived the flux of particles hitting the boundary per cm^2 in Δ seconds from the left [Ω^-, equation (C7)] and from the right [Ω^+, equation (C8)]. We find that these fluxes have a component due to the density of particles at the boundary D and a component due to the x_1 gradient of density of particles on either side of the boundary, $D_{x_1}^-$ and $D_{x_1}^+$ (there are components due to the higher x_1 derivatives of the particle density and the derivatives of k and pc, but these depend on $\Delta^{3/2}$, Δ^2 etc.,

and hence contribute a vanishingly small amount to the rate of particles hitting the boundary as $\Delta \downarrow 0$).

Thus under natural flow conditions, the net number of particles crossing the boundary per cm^2 from left to right in Δ seconds is

$$\Omega^- - \Omega^+ = \left(\frac{\Delta}{\pi}\right)^{\frac{1}{2}} D\left[(K_1^-)^{\frac{1}{2}} - (K_1^+)^{\frac{1}{2}}\right] - \tfrac{1}{2}\Delta\left[K_1^- D_{x_1}^- + K_1^+ D_{x_1}^+\right], \tag{3.2.77}$$

where we shall assume for the moment that particles are conserved as they cross the boundary (i.e. $pc^- = pc^+$). Then as $K_1^- D_{x_1}^- = K_1^+ D_{x_1}^+$ the second term of (3.2.77) contributes exactly the desired boundary flux (3.2.76) (the flux is obtained by dividing (3.2.77) by Δ), but the first term, which is non-zero if $K_1^- \neq K_1^+$, contributes an unwanted component to the net flux. This component can be eliminated by arbitrary conditions imposed upon those trajectories which try and cross the boundary from either side.

Suppose that K_1^- is greater than K_1^+. Then

$$\alpha = \left(\frac{K_1^+}{K_1^-}\right)^{\frac{1}{2}} < 1 \tag{3.2.78}$$

is the ratio of the Ω^+ flux to the Ω^- flux due to the boundary particle density D. If we reduce the Ω^- flux to $\alpha\,\Omega^-$, and keep the Ω^+ flux constant, then the contributions of the left and right flux due to D will cancel each other*. The net number of particles crossing the boundary per cm^2 in Δ seconds is (from (3.2.77))

$$\alpha\,\Omega^- - \Omega^+ = -\tfrac{1}{2}\Delta\left[\alpha\, K_1^- D_{x_1}^- + K_1^+ D_{x_1}^+\right]. \tag{3.2.79}$$

We have now upset the flux due to the density gradient D_{x_1}. We can restore the proper flow by magnifying the flow of particles

---

* If $K_1^+ > K_1^-$, the coefficient α is applied to particles arriving from the right, with $\alpha = (K_1^-/K_1^+)^{\frac{1}{2}}$.

in both directions (3.2.79) by the factor β. Equating the modified flux (3.2.79) (divided by Δ to get particles/second) to the desired flux (3.2.76) we have

$$-\tfrac{1}{2}\,\alpha\,\beta\,K_1^-\,D_{x_1}^- - \tfrac{1}{2}\,\beta\,K_1^+\,D_{x_1}^+ = -K_1^-\,D_{x_1}^- .$$

Then substituting $\dfrac{k_1^-}{k_1^+}\,D_{x_1}^-$ for $D_{x_1}^+$, we have

$$-\tfrac{1}{2}\,\alpha\,\beta\,K_1^-D_{x_1}^- - \tfrac{1}{2}\,\beta\,K_1^+\,\frac{k_1^-}{k_1^+}\,D_{x_1}^- = -K_1^-\,D_{x_1}^-$$

$$\text{whence}\quad \beta = 2\,K_1^-\left[\alpha\,K_1^- + K_1^+\,\frac{k_1^-}{k_1^+}\right]^{-1}. \qquad (3.2.80)$$

Now as we have assumed $pc^- = pc^+$, β becomes

$$\beta = \frac{2}{1+\alpha} > 1. \qquad (3.2.81)$$

As β is greater than one, the flow is magnified, and this is implemented during the simulation be forcing $(\beta - 1)$ more particles to cross the boundary than would naturally go after the reflection coefficient α has been applied to the Ω^- flow. As in the case of the treatment of VU and W terms given earlier, this magnification (and the reflection coefficient α) could be implemented deterministically or probabilistically, and either method was found to be equally convenient on the digital computer. As a summary, the final particle flows per cm^2 in Δ seconds are shown disgrammatically in Figure 3.2.3. The factor θ shown does not enter in the constant pc case just discussed,(i.e. $\theta = 1$ here).

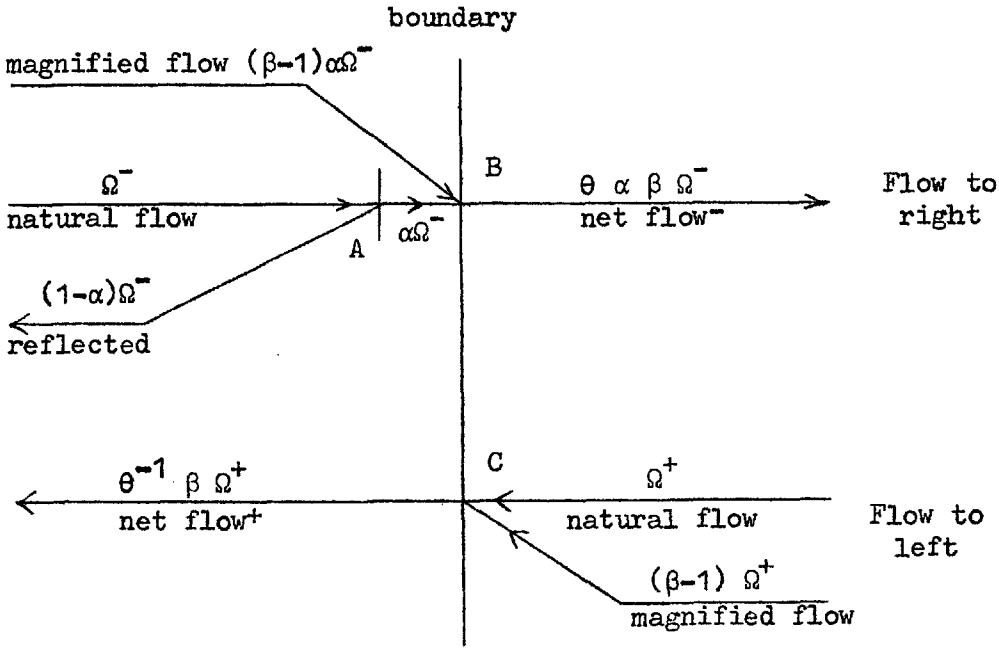


Figure 3.2.3 Imposed Boundary Conditions for K_1 Discontinuity

It can be verified that the particles are conserved at each of the operation points, A, B and C, and that the net left to right flux

$$\alpha\ \beta\ \Omega^- - \beta\ \Omega^+ \qquad (3.2.82)$$

is the correct value.

Now consider the case where pc does vary across the boundary of the material discontinuity. When a particle crosses the boundary, we must ensure that it obeys the physical laws inherent in the parabolic equation. Recall that a particle is a quantity which raises one cm^3 of material $\not{o}\ N^{-1}\ ^oK$ no matter what the pc of the material is, while a calorie raises one cm^3 of material $(pc)^{-1}\ ^oK$. Thus if a particle in a material with the parameter pc^- moves into a material with the parameter pc^+, it must change its value by a factor

$$\theta = \frac{pc^-}{pc^+}\ , \qquad (3.2.83)$$

so that it represents the same amount of heat energy in each material.

This value change is implemented in the simulation by violated the conservation of trajectories principle (we noted earlier, at equation (3.2.70), that if p or c depended on x, the simulation would involve non-conserved trajectories). This is achieved by further operating on all trajectories which undergo the operations B and C in Figure 3.2.3: particles passing to the right of B are amplified by a factor θ, and those passing to the left at C are amplified by a factor θ^{-1}. This is very similar to the procedure we described earlier for handling a residual proportional term VU of the parabolic equation, and is implemented statistically by allowing existing trajectories to expire or new trajectories to begin with the appropriate probability.

With the addition of this extra operation, we must rederive the flux balance equations. Consider the net flow of particles entering the right hand region in Figure 3.2.3 per cm^2 in Δ seconds. We have from (C7), (C8)

$$\theta\,\alpha\,\beta\,\Omega^- - \beta\,\Omega^+ = \theta\,\alpha\,\beta\,\left(\frac{\Delta}{\pi}\right)^{\frac{1}{2}}(K_1^-)^{\frac{1}{2}}\,D$$

$$- \theta\,\alpha\,\beta\,\tfrac{1}{2}\,K_1^-\,\Delta\,D_{x_1}^- - \beta\,\left(\frac{\Delta}{\pi}\right)^{\frac{1}{2}}\,(K_1^+)^{\frac{1}{2}}\,D$$

$$- \beta\,\tfrac{1}{2}\,K_1^+\,\Delta\,D_{x_1}^+ \qquad (3.2.84)$$

As before, we choose α so that the contribution from the particle density D vanishes. Thus equating the sum of the first and third terms on the right hand side of (3.2.84) to zero, we have

$$\theta\,\alpha\,(K_1^-)^{\frac{1}{2}} = (K_1^+)^{\frac{1}{2}}\,,$$

and so

$$\alpha = \theta^{-1}\left(\frac{K_1^+}{K_1^-}\right)^{\frac{1}{2}}. \qquad (3.2.85)$$

If this α is greater than one, then the reflection operation should be applied to particles hitting the boundary from the other side, for then the resulting α will be less than one.

Now we choose β so that the contribution from the particle density gradient D_{x_1} is the correct flux (3.2.76). Thus from the second and fourth terms on the right hand side of (3.2.84), we have (dividing them by Δ)

$$- \theta \alpha \beta \tfrac{1}{2} K_1^- D_{x_1}^- - \beta \tfrac{1}{2} K_1^+ D_{x_1}^+ = - K_1^+ D_{x_1}^+.$$

But the left and right density gradients at the boundary are related by the ratio of the conductivities, and so we can replace $D_{x_1}^-$ by $(k_1^+/k_1^-)D_{x_1}^+$ giving

$$\theta \alpha \beta \tfrac{1}{2} K_1^- k_1^+ (k_1^-)^{-1} + \beta \tfrac{1}{2} K_1^+ = K_1^+. \qquad (3.2.86)$$

Now from (3.2.83), $\theta = K_1^+ (k_1^+)^{-1}(K_1^-)^{-1}k_1^-$, and so (3.2.86) becomes

$$\tfrac{1}{2} \alpha \beta + \tfrac{1}{2} \beta = 1,$$

or

$$\beta = \frac{2}{1 + \alpha} < 2. \qquad (3.2.87)$$

It can be verified that these values of α and β also give the correct value to the net flux leaving the left hand side. Compared with the constant pc case studied earlier, the values of β found are the same (3.2.81 and 87), but the values of α (3.2.78 and 85), differ by the factor θ. As $\theta = 1$ in the constant pc case, the coefficients derived earlier (3.2.78, 81) can be considered as special cases of the ones just derived (3.2.83, 85, 87).

Boundary Conditions

Parabolic equations are basically initial value problems, and often no boundary conditions are imposed on the solution after the initial time $t = 0$. This was the case for the example studied in Chapter 2, and indeed for most FP equations describing the statistics of dynamic systems, the systems concerned do not have boundary conditions imposed upon them which can be readily translated into boundary conditions for the FP equation. Other parabolic equations, however, often do have imposed boundary conditions. For example, heat conduction problems usually do, as the material under study is finite, and edge conditions are imposed on the material. We will continue our discussion on the heat conduction example, and show how boundary conditions are handled.

From the discussion on the treatment of spatial discontinuities, the treatment of boundary conditions follows directly. Using the physical analogies given previously, we transfer the boundary conditions directly into conditions on the simulated trajectories.

Boundary conditions for heat conduction problems take two forms [77], the specification of the solution U on the boundary (Dirichlet type boundary conditions), and the specification of the normal heat flux $(kU_x)_{nor}$ across the boundary (Neumann type boundary conditions). In each case, we regulate the flow of particles across the boundary to satisfy the given conditions.

For the case where the solution U is given on the boundary, we consider a region directly adjacent to the boundary, and continually estimate the solution in this region by keeping track of the number of particles in the region [see equation (3.2.50)]. Then at each time step at which accounting is done, trajectories are forced across the boundary to keep the number of trajectories in the region at the proper level. This time step may be the same Δt which is the basic discretization time of the differential equation solution (see Chapter 6) on the digital computer, or multiples of it. This operation sets the number of trajectories in the boundary region to the exact value at each step, but still possesses the inherent Δt and Δx discretization errors of the

Monte Carlo method. An example of this type of boundary condition is given later.

For the case where the normal heat flux is specified at the boundary, we regulate the flow of trajectories across the boundary so that the particle flux at the boundary equals the required heat flux (remember, 1 particle = ϕN^{-1} pc calories). In principle, the implementation of these flux conditions is not complicated by arbitrarily shaped boundaries, for only those particles crossing the boundary contribute to the normal flux at the boundary. However, in practice, the analogue or digital circuitry or logic needed to detect when a particle crosses the boundary is complicated by an unusual boundary shape.

If the flux out of the material is larger than the desired flux, then particles will have to be reflected back into the material at the boundary. If the flux out of the material is too low, then extra trajectories have to be taken out of the material from a region near the boundary. If the boundary conditions call for a net flux in at the boundary, then all trajectories hitting the boundary must be reflected back, and extra trajectories (corresponding in rate to the specified flux) have to be created and set free just inside the boundary. An example of this type of boundary condition is given in Section 3.2.6.

In addition, mixed boundary conditions of the form $\mu_1 U + (U_x)_{nor} = \mu_2$ can be handled by our method. The technique is to measure U adjacent to the boundary, and then regulate the flow of particles across the boundary so that the normal flux across the boundary equals $(U_x)_{nor} = \mu_2 - \mu_1 U$. No method has so far been proposed for handling mixed boundary conditions in the backward simulation method.

With the backward simulation method, Dirichlet type boundary conditions are handled somewhat more conveniently than by our method, but only certain Neumann type boundary conditions involving geometrically simple boundaries can so far be handled by the backward simulation method [83]. In this latter case, our methods are more conveniently implemented, and can handle a general Neumann condition.

This is essentially because of the physical analogy which can be drawn between the flow of particles in the simulation and the physical flux which is specified in the Neumann conditions.

Physical Analogies

We have seen that in the case of the equation of heat conduction, physical analogies could be drawn between the simulated particles and the parameters of the physical situation which the parabolic equation describes. These analogies allowed us to translate certain given or necessary conditions of the physical situation (e.g. conservation of thermal energy and specification of thermal flux) into conditions on the simulated trajectories. This allowed us to specify the behaviour of the simulated trajectories at certain boundaries where the parabolic equation did not necessarily hold.

Let us look more closely at the origin of the physical analogy. The particles in the simulation are quantities whose density in the state space R gives the value of the dependent variable $U(x,t)$. This fact will often give a physical meaning to the particles. In addition, Green's theorem connects the instantaneous flux of the simulated particles to the space gradient of the dependent variable of the Fokker-Planck equation of the underlying diffusion process governing the trajectories of the particles. This flux is given in equation (1.2.3) for the FP equation (1.2.1). The main point is that parabolic equations governing physical situations of a diffusive character (e.g. weather prediction equations) are usually derived from physical principles embodying flux concepts, and these physical principles connect the flux (1.2.3) to the forward Komogorov equation (1.2.1). Through this connection, it is felt that empirical methods such as given in this section can be found to handle unusual conditions on most parabolic equations.

It is not known whether such a simple connection exists between the particle flux and the backward Kolmogorov equation (1.2.4). Authors who solve parabolic equations by Monte Carlo methods based on the backward simulation method [63, 64, 83] do not mention such

analogies, and hence these methods may not be able to cope with the discontinuity conditions discussed in this section. This may be a significant advantage of the Monte Carlo methods described in this Chapter, but the present author does not have sufficient experience with the backward simulation method or with the parabolic equations other than the heat conduction equation to comment further on this point.

3.2.6 Numerical Results of the Heat Conduction Example

As the purpose of these examples is to illustrate the treatment of material discontinuities and boundary conditions, we discuss what is essentially a one-dimensional example. This is because the example has boundaries and a discontinuity lying in the x_2-x_3 plane, and the material parameters, initial conditions and boundary conditions are all independent of x_2 and x_3, so that the solution is always uniform in x_2 and x_3. Thus our simulation can be considered to be confined to a 1 cm^2 cross-section of the x_2-x_3 plane, and our remarks will only concern the dependence of the solution on x_1 and time.

Consider a material which is homogenous in two separate regions as shown in Figure 3.2.4. The material $^-$ is copper extending from $x_1 = 0$ to 5 cms., and material $^+$ is steel extending from $x_1 = 5$ to 10 cms. From standard tables [77] we find the material parameter values for copper and steel shown in Table 3.2.2.

We assume that a general solution is desired as a function of x_1 and t. For accounting purposes, the x_1 axis is divided into 10 equal parts or cells, labelled 1 to 10. No significant quantisation occurs in the individual trajectories, which are obtained to the accuracy of the computer, but the trajectories are quantised into the cells 1 to 10 when an estimate of the solution is desired. The solution is then estimated at the mid-points of the cells, $x_1 = 0.5, 1.5, 2.5, \ldots 9.5$ cms.

Transient Solution

A transient solution was obtained under the following conditions:

(a) Initial temperature $U(x_1, o) = 300^oK$ for all x_1. Assuming each cell to have a 1 cm^2 cross-section in the x_2-x_3 plane, the volume of each cell is 1 cm^3. Then from (3.2.33), $\phi = 3000$. If we let 100 particles in each cell equal a temperature of 300^oK, then $N = 1000$ and in (3.2.50b),

$$\phi N^{-1} = 3\ ^oK/(\text{particles}/cm^3). \qquad (3.2.88)$$

(b) A thermal flux of 500 watts/cm^2 = 119.5 calories/(sec.cm^2) enters at the left hand boundary 1. In terms of simulated particles, this flux represents $119.5\ N\ (\phi pc^-)^{-1} = 48.8$ particles/sec., entering cell 1 from the left boundary. If Δt is chosen as a multiple of $(48.8)^{-1} = 0.02047$ seconds, this flux can be simulated conveniently by adding the necessary number of particles every Δt seconds. In the present example we have chosen $\Delta t = 0.2047$ seconds, and added 10 particles to cell 1 every Δt. Each new particle so added is given a position $x_1 = 0.0$, and has the same dynamics and weight ϕN^{-1} as the original particles.

(c) The right hand boundary 3 is kept at a constant temperature of 300^oK. This was simulated by keeping 100 trajectories in cell 10 (by adjusting the number every Δt seconds) but this involves a Δx quantisation error, as the number of particles in cell 10 is an approximation to the solution at $x_1 = 9.5$, whereas the boundary is actually at 10.0. This error could be eliminated by modifying the cell spacings so that the rightmost cell was centred around $x_1 = 10.0$, but this was not done in the present example for the convenience of retaining simple quantised values of x_1.

(d) The dynamics of the simulated trajectories are given by (3.2.72a)

$$dx_1(t) = (2\,K_1)^{\frac{1}{2}}\, dw(t). \qquad (3.2.89)$$

From Table 3.2.2, $(2\,K_1)^{\frac{1}{2}}$ equals 1.51 for copper and 0.49 for steel. The dynamics which affect the particle in the x_2 and x_3 directions do not affect the trajectories in the x_1 direction, and for the present purpose they need not be simulated.

Equation (3.2.89) is simulated digitally by adding Gaussian increments to $x_1(t)$:

$$x_1(t + \Delta t) = x_1(t) + N(0,\, 2K_1\Delta t), \qquad (3.2.90)$$

where $N(0, 2K_1\Delta t)$ is a Gaussian random number with mean zero and variance $2K_1\Delta t$. As mentioned in Appendix A, and discussed in Chapter 6, this is a valid incremental interpretation of the stochastic differential equation (3.2.89). Indeed, as the s.d.e. (3.2.89) has no drift term and the noise $dw(t)$ is independent of $x_1(t)$, then $x_1(t)$ is a simple Brownian motion (with variance $2K_1t$ in this case), and the discrete simulation (3.2.90), simulates the statistical behaviour of $x_1(t)$ exactly at the discrete sample points $t = n\,\Delta t$. Also, the random numbers used were obtained by a log-trigonometric transformation of a uniform variate, and were from an exactly Gaussian distribution [78] (the usual practice is to add 10 or 12 uniformly distributed numbers together, but this procedure does not represent the tails of the Gaussian distribution well. Also, most generators of uniform numbers are quite non-uniform in the short term, and our method is less sensitive to such distortions than the common method.). A typical trajectory is shown in Figure 3.2.5.

(e) The treatment of particles when they reach the material discontinuity at $x_1 = 5$ has been discussed earlier in connection with Figure 3.2.3. In particular, we have the following discontinuity coefficients:

(3.2.83) $\theta = .826$

$\theta^{-1} = 1.21$

(3.2.85) $\alpha = .393$

$1-\alpha = .607$

(3.2.87) $\beta = 1.435$

$\beta-1 = .435$

Referring to Figure 3.2.3, a particle hitting the boundary from the copper side undergoes the following operations:

A: It is reflected back into the copper material with probability $(1 - \alpha) = .607$. This decision is made by choosing a random number from a uniform (0, 1) distribution, and it is reflected back a distance such that its total path length equals the distance (3.2.90) it would have travelled had no discontinuity been present (this prevents particles piling up at the boundary and ensures a natural distribution of particles near the boundary). A particle has been reflected back into the copper material at 1 in Figure 3.2.5.

B: If the particle is not reflected back into the copper region, it passes into the steel region, and with probability $(\beta - 1) = .435$, brings a second particle with it. Then, as $pc^- \neq pc^+$, this particle (and the second one if it comes) must undergo a value change by a factor $\theta = .826$ as it crosses the boundary. As θ is less than 1 in this case, this value change is implemented by allowing the particle(s) to pass through freely with a probability θ, or to become extinct with probability $(1 - \theta)$. In Figure 3.2.5, a particle has passed through freely at 2 (and not brought another particle with it), but has become extinct at 3. Note that when a particle does pass into another material, as at 2, 4 or 5, its speed changes by the ratio $(K_{new}/K_{old})^{\frac{1}{2}} = 0.32$ or 3.1 in much the same way as a light ray passing into a new material with a different refractive index. The dotted lines in Figure 3.2.5 show where the particle trajectory would lie had the material discontinuity not been present.

A particle which hits the boundary from the steel^{+} side does not undergo a reflection operation, but only two magnification operations:

C: The particle passes into the copper region, and as at B, brings a second particle with it with probability $(\beta - 1) = .435$. As the particle(s) crosses the boundary, it (they) must undergo a value change by a factor $\theta^{-1} = 1.21$. As θ^{-1} is greater than one, new particles must be invented and set free in the copper region (at 4 or 5 in Figure 3.2.5). This is implemented by deterministically adding $I(\theta^{-1} - 1)$* new particles, and adding another one with probability $(\theta^{-1} - 1) - I(\theta^{-1} - 1)$. This operation must be carried out on the original particle and the one added with probability $(\beta - 1)$ above. Figure 3.2.5 shows particles added at 5, following the transmission of the original particle through the steel-copper interface. If the particle had been added with the β operation, the particle would be one arbitrarily chosen and removed from the steel material (near where the original particle was, if possible) and set free at the new location of the original particle. If the particle had been added with the θ operation, it would be a new particle.

A graph of the initial part of the temperature transient is given in Figure 3.2.6. Individual solution points are shown as being the average temperature in each quantised cell of 1 cm. length. The standard deviation of each point estimate can be approximately obtained from equation (3.2.52b). Two extreme cases are worked out:

(a) Cell 1, $t = 45$ secs., $N' = 2900$ particles, $q \doteq 0.15$.
Standard deviation $\doteq$ 19 particles
$= 57^{\circ}K$ (4%). (3.2.91a)

---

* $I(\cdot)$ is the integer part of $(\cdot)$

(b) Cell 9, $t = 10$ secs., $N' = 1500$ particles, $q \doteq 0.07$.
Standard deviation $\doteq$ 10 particles
$= 30^{\circ}K$ (10%). (3.2.91b)

The standard deviation of other points in the transient shown lies between these values. Assuming the distribution of error to be Gaussian, the 95% confidence interval for these points is roughly plus or minus twice the standard deviation. Of 100 solution points tested, 93 fell within this confidence interval (the true solution was taken as the smoothed one described below), which confirms our method of estimating error.

The statistical errors (3.2.91) are rather high for some purposes, and higher accuracy may be desired. In lieu of adding more points to the simulation (storage problems may preclude this), we may sacrifice time and space resolution to obtain a smoother solution. As the space quantisation of 1 cm is already quite coarse, let us look at the time resolution Δt. The Δt of about 0.2 seconds was chosen with regard to the mobility of particles in the copper region (the particles are less mobile in the steel region by a factor of 0.32 - see equation (3.2.90)). With this Δt, the particles in the copper region were sufficiently mobile so that successive evaluations (at each Δt) of the number of particles in each cell had statistical fluctuations with low correlation, and yet sufficiently immobile so that the probability of a particle jumping two cell boundaries was kept low. The latter point is not important but makes accounting simpler, but the first point is important for the following reason. For the given Δx quantisation, this is the minimum value of Δt for which successive cell particle densities are reasonably independent, and so while this Δt may be much smaller than needed for solution time resolution, it generates the maximum amount of independent statistical information for the given number of trajectories N'.

Thus if a lower time resolution can be tolerated, a smoother solution can be obtained by averaging successive cell densities. This operation introduces no bias if the time variation of the solution

is linear over the interval of the time smoothing. In the present example, solutions were obtained by averaging over 10 successive time intervals, and from Figure 3.2.7, we see that the solution is quite linear over an interval of $\pm 5\,\Delta t \doteq \pm 1$ seconds. Although Figure 3.2.7 shows the solution variation averaged over the entire steel region, the variation shown is typical of the point variations encountered.

This smoothing reduces the statistical errors by about a factor of $(10)^{\frac{1}{2}}$ or 3, and the smooth curves of Figure 3.2.6 were drawn from these points - the points are not shown, but are all within two line thickness of the curves drawn. Thus drawing the curve has entailed further smoothing, which was justified by the known information:

(a) The slope at the left hand edge of the curve is given by the imposed thermal flux: 119.5 calories/(sec.cm^2) $= k_1 U_{x_1}$ at $x = 0$, which for copper gives a slope of 128.5°K per cm.

(b) As thermal flux is conserved at the material boundary the slope ratio $U^+_{x_1}/U^-_{x_1}$ is given by $k_1^-/k_1^+ \doteq 8$.

If smoothing is considered as a sequential operation, that is, every Δt a new solution estimate is obtained using only past data, the operation can be likened to a Bayesian filtering operation. A proper sequential Bayesian filter works as follows:

(1) An estimate $\underline{q}(t)$ is obtained at time t with an associated probability density $P[\underline{q}(t) \mid t]$.

(2) Knowledge of how $q(t)$ is likely to vary in the interval $(t, t + \Delta t)$ is used to update $P[\underline{q}(t) \mid t]$ to the a priori probability density for $\underline{q}(t + \Delta t)$, $P[\underline{q}(t + \Delta t) \mid t]$.

(3) The likelihood function obtained from data collected at time $t + \Delta t$ is used to update $P[\underline{q}(t + \Delta t) \mid t]$ into the a posteriori probability density for $\underline{q}(t + \Delta t)$, $P[\underline{q}(t + \Delta t) \mid t + \Delta t]$. Using

this density and perhaps an associated risk function, the optimal estimate $\underline{q}(t + \Delta t)$ is obtained.

Compared with the maximum likelihood method considered earlier (3.2.50), the difference of the Bayesian method is to include past data and inherent knowledge of the behaviour of $q(t)$ via the step (2) above([80, Ch.3] gives a lucid comparison of maximum likelihood and Bayesian estimators). The Bayesian Kalman filter [80] achieves step (2) by incorporating an exact model of the dynamics of $q(t)$ into the structure of the estimator.

However, in many situations of applied statistics, the information needed to carry out step (2) exactly is not available, yet there is often good, if empirical, motivation for incorporating some a priori information on $q(t + \Delta t)$ into the maximum likelihood estimator for $q(t + \Delta t)$. It is felt that it is from this reasoning that many empirical smoothing or filtering operations are derived in practice. For example, if it were known (or guessed) that $q(t)$ had a variation with a maximum frequency content of ω_o, then the data $\underline{q}(t)$ could be filtered by an exponential filter of time constant ω_o^{-1} to reduce the statistical fluctuations of higher frequency than ω_o in the data (c.f. in the Kalman filter, if $q(t)$ is generated by a first order system, then $\underline{q}(t)$ is smoothed by the same first order filter). In our example above, we have allowed $q(t)$ to have a linear variation over a range $t \pm 5\Delta t$, and thus have used a smoothing filter which gives equal weighting to data $\underline{q}(t)$ over this range. This operation differs from that of the recursive filter in that future data is incorporated into the current estimate, but the main analogy to be drawn is that information other than that of the current data is used to obtain the current estimate. This is justified by proposing a structure for the change of the parameter $q(t)$.

As mentioned earlier, point solutions (3.2.50) are not the forte of the Monte Carlo method, but rather functional solutions of the form (3.2.60) are better suited to evaluation by these statistical methods. This is essentially because a point solution uses the information of only a subset of the simulated trajectories, whereas the functional solution uses all the trajectories, and the statistical fluctuations involved depend mainly on the number of points used in the solution. In between these two situations is the regional solution (3.2.59) in which a large proportion, but not all, of the simulated points are used in the solution.

An example of this latter type of solution is given in Figure 3.2.7, where the average temperature in the steel region, $x_1 = (5, 10)$, is evaluated. The error statistics are again given by (3.2.52) where at $t = 0$, $N' = 1000$, $q = .5$, and at $t = 40$, $N' = 2750$, and $q \doteq .3$. The resultant standard deviation of the temperature error is 10° K near $t = 0$ rising to 14° K near $t = 40$. Of course at $t = 0$ there is no error, and this error analysis only applies when individual trajectories have travelled long enough so that their position is relatively independent of their initial position (about $t = 5$ seconds in the steel region). In any case, the standard deviations given are to be taken as upper limits. Only every 10th solution point is shown on the graph, and so there was adequate opportunity for smoothing. However the graph shown was just drawn by eye, and the points shown are given as typical points. The error estimate was verified as about 2/3 of the solution points were with a standard deviation of the final curve (theoretically, an average proportion of 0.68 of the points should be within the standard deviation).

Consistency of the Monte Carlo Solution with Respect to Time Dependence

It appears that owing to the material discontinuity of the present example, an analytic solution for the transient temperature is not readily available. However, the steady state solution is easily obtained from flux balance considerations. Thus as we cannot directly compare our Monte Carlo solution with a known solution, it is useful to consider the following argument which shows that the Monte Carlo method does not introduce any errors due to the transient nature of the problem. This implies that the accuracy of the Monte Carlo method during a transient solution is the same accuracy with which the steady state solution is obtained (for an equal number N' and proportion q of simulated points). We can then confine our accuracy discussion to the steady state Monte Carlo solution.

Consider an infinite homogeneous rod of 1 cm^2 cross-section with no heat transfer across the surface of the rod. Then we have linear heat flow in the x direction (dropping the subscript 1 of our example) which is governed by the differential equation

$$U_t = K U_{xx}, \tag{3.2.92}$$

where K is the diffusivity of the material. This is the same situation of our example of Figure 3.2.4 if the material discontinuity and boundaries were removed from our example.

Assume the rod is initially at zero temperature, $U(x,o) = 0$, and a certain quantity of heat energy, pc calories, is released at $x = 0$, $t = 0$. Then by differentiation and substitution, it is easily verified that

$$U(x,t) = (4\pi Kt)^{-\frac{1}{2}} \exp(-x^2/4Kt) \tag{3.2.93}$$

is a solution of (3.2.92) satisfying the given initial conditions [77, p.50]. It is noted that (3.2.93) is a Gaussian distribution of zero mean and variance $2Kt$.

Now consider how (3.2.92) could be solved by Monte Carlo methods. The zero initial conditions means that no particles are present initially, and the release of pc calories at $x = 0$, $t = 0$, corresponds to the release of $\phi^{-1}N$ particles at $x = 0$, $t = 0$. As in (3.2.89), the parabolic equation (3.2.92) dictates that the particles undergo a Brownian motion given by the S.D.E.

$$dx(t) = (2K)^{\frac{1}{2}} dw(t). \qquad (3.2.94)$$

Thus the distribution of simulated particles at time t is given by the Gaussian distribution with zero mean and variance $2Kt$. It remains now to check the scaling factor of the Monte Carlo solution. Recall that the contribution of each particle $x_\alpha(t)$ to the solution is $\phi N^{-1} \delta(x - x_\alpha)$. Thus the integral of the Monte Carlo solution

$$\int_{-\infty}^{\infty} u(x,t)\, dx = \sum_{\alpha} \phi N^{-1}$$

$$= 1$$

as there are $\phi^{-1} N$ particles. But the integral of the true solution (3.2.93) is also one, and so the <u>expected value</u> of the solution obtained by the Monte Carlo simulation is given by

$$U(x,t) = (4\pi Kt)^{-\frac{1}{2}} \exp(-x^2/4Kt)$$

which is the true solution (3.2.93). Thus our Monte Carlo method provides an <u>unbiased</u> estimate of the true solution.

This argument can be extended to the case where the insulated rod has an arbitrary initial temperature

$$U(x,o) = f(x). \qquad (3.2.95)$$

The solution is then given by [77, p.53]

$$U(x,t) = (4\pi Kt)^{-\frac{1}{2}} \int_{-\infty}^{\infty} f(x') \exp(-(x - x')^2/4Kt)dx' \qquad (3.2.96)$$

Concerning the Monte Carlo simulation of this situation, from (3.2.33) we have the normalising constant ϕ as

$$\phi = \int_{-\infty}^{\infty} f(x)\, dx.$$

The number of calories in the rod is then ϕ pc which is simulated by N particles. Thus the initial density of particles is

$$D(x) = \phi^{-1} N f(x) \quad \text{particles/cm.} \qquad (3.2.97)$$

Now each particle in the simulation has a weight ϕN^{-1} and undergoes the Brownian motion defined by (3.2.94). Then a particle starting at $x = x'$ at $t = 0$, has a Gaussian distribution of mean x' and variance $2Kt$ at time t. Thus the expected value of the contribution of a particle beginning at $x = x'$ to $U(x,t)$ is

$$\phi N^{-1} (4\pi Kt)^{-\frac{1}{2}} \exp(-(x - x')^2/4Kt). \qquad (3.2.98)$$

Now taking into account the initial density of particles (3.2.97), the expected value of the Monte Carlo solution equals the expected contribution of all simulated particles to $U(x,t)$, and is given by the integral of (3.2.98) times (3.2.97) over all x, which is

$$(4\pi Kt)^{-\frac{1}{2}} \int_{-\infty}^{\infty} f(x') \exp(-(x - x')^2/4Kt)dx'.$$

But this is exactly equal to the true solution (3.2.96), and so our Monte Carlo method provides an unbiased estimate of the true solution for the arbitrary initial conditions (3.2.95).

The situation we have just described is the same as our example of Figure 3.2.4, except that we have imposed boundary conditions at $x_1 = 0$, 5 and 10 in our example. However these imposed conditions do not involve time dynamics as they were derived (e.g. Figure 3.2.3) to satisfy instantaneous flux balance conditions. While the imple-

mentation of these conditions involve statistical fluctuations inherent in the Monte Carlo method, the boundary conditions were derived so that the error in implementing the boundary conditions has a zero mean value at all times. Thus the imposition of boundary conditions in our simulation does not introduce any errors which are specifically due to the time dependence of the transient solution.

With this argument we suggest that the error in the transient Monte Carlo solution depends only on the same simulation parameters N' and q as the associated steady state Monte Carlo solution (e.g. equation (3.2.52) for point or regional solutions). With this assumption, we will proceed to analyse the error of two steady-state solutions in more detail.

Steady State Solution

Consider the example of Figure 3.2.4, with the heat flux entering at the left boundary reduced to 50 watts/cm^2 = 11.95 calories/(sec.cm^2), and the solution $U = 300^{\circ}K$ is specified at $x_1 = 95$ cms. instead of 10.0 cms. (this latter point eliminates the quantisation error in implementing the right hand boundary condition). The steady state solution is obtained by setting $U_t = 0$ in (3.2.62) (remember $U_{x_i} = 0$, $i = 2, 3$ in this example) and we have

$$(pc)^{-1} \frac{\partial}{\partial x_1} [k_1 U_{x_1}] = 0$$

whence
$$k_1 U_{x_1} = \text{a constant} = \text{the heat flux in the } x_1 \text{ direction} = 11.95 \text{ calories/(sec.cm}^2\text{).} \qquad (3.2.99)$$

Thus (3.2.99) specifies the temperature gradient in terms of the known heat flux in the x_1 direction (the heat conduction equation (3.2.62) is derived from the basic principle that the rate of heat flow is in the direction of, and proportional to, the local temperature gradient). As no heat is lost in the material between

$x_1 = 0$ and 10, the heat flux (3.2.99) is constant in this range. In particular we have from Table 3.2.2

$$U_{x_1}(x_1) = 12.85\ ^{\circ}K/cm, \qquad x_1 = (0, 5),$$

and

$$U_{x_1}(x_1) = 103.9\ ^{\circ}K/cm, \qquad x_1 = (5,10). \tag{3.2.100}$$

Then as $U(9.5) = 300^{\circ}K$,

$$U(x_1) = 300 + 103.9(9.5 - x_1)\ ^{\circ}K, \quad x_1 = (5, 9.5),$$

whence $U(5.0) = 767.5^{\circ}K$. Then we have

$$U(x_1) = 767.5 + 12.85(5.0 - x_1)\ ^{\circ}K, \quad x_1 = (0, 5),$$

whence $U(o) = 831.8\ ^{\circ}K$.

Thus the particles in the Monte Carlo simulation should be distributed according to a ramp distribution, with one change of slope at $x_1 = 5$.

Dividing up the x_1 space into 10 evenly spaced cells, and letting $\phi N^{-1} = 3$ as before (3.2.88), the average temperature and the expected number of particles in each cell is given in Table 3.2.3.

Cell	1	2	3	4	5	6	7	8	9	10
Average Temperature	825.4	812.5	799.7	786.8	774.0	715.6	611.7	507.8	403.9	300.0
E[No. of Particles]	275.1	270.8	266.5	262.3	258.0	238.5	203.9	169.3	134.6	100.0

Table 3.2.3 Steady State Temperature and Particle Distribution

To test the Monte Carlo solution, and particularly the implementation of boundary and discontinuity conditions, the steady state conditions described above were simulated, and 100 independent trials were recorded, each time tabulating the number of particles in each of the 10 cells. We would like to test the null hypothesis H_o: that the expected number of particles in each cell is exactly as given in Table 3.2.3. To test this hypothesis, we use Pearson's chi-square test for goodness-of-fit [79, p.309]. We form the test criterion

$$u_j = \sum_{i=1}^{10} \frac{(n_i - E[n_i])^2}{E[n_i]} \qquad (3.2.101)$$

where n_i is the number of particles in cell i at the j:th trial, and $E[n_i]$ is the expected number given in Table 3.2.3 assuming H_o is true. Then if H_o is true, u_j will be a chi-square variate with nine degrees of freedom (i.e. $E[u_j] = 9$), and from standard tables [79, p.432] we can find cumulative intervals such as: "50% of u_j should be in the range (5.90, 11.4)". Also, if H_o is true, then the variance of n_i is given by (3.2.52b), as this assumption is used in the derivation of the chi-square distribution of u_j.

In fact, in our 100 trials, 55 of the u_j found fell within this range (5.90, 11.4). The success or failure of all 100 trials forms a binomial distribution ($n = 100$, $p = \frac{1}{2}$) with mean 50 and standard deviation 5, and so our 100 trials are quite consistent with the null hypothesis H_o being true.

Another way of doing the chi-square test is to sum the number of particles in cell i over the 100 tests, and form the test criterion (3.2.101) using these total n_i's. The chi-square variate so formed equalled 8.23 which again shows that the experimental results are quite consistent with H_o being true. If H_o were not true, then the chi-square variate (3.2.101) would have a mean value considerably larger than 9.0, and H_o would have a very low probability of being true. Furthermore, if H_o were not true, then the variate (3.2.101) would increase indefinitely as the

number of trials is increased, and so the reliability of the chi-square test increases with the number of trials. The 100 trials we have taken was considered to be a suitable number to give a high significance to the chi-square test.

Thus we have shown statistically that it is very plausible that our Monte Carlo solution is an unbiased estimate of the true solution in the steady state. We can do no better than this, except to take more trials and reconfirm our statistical conclusion. Furthermore, we have argued that our simulation technique introduces no errors specifically due to the transient nature of the solution provided that the stochastic equations for the trajectories and the instantaneous boundary conditions are implemented correctly, which implies that the error analysis just made in the steady state solution also applies to the transient solution.

The examples we have considered so far have involved the simulation of trajectories with the simple dynamics (3.2.89). Our discussion on the error of the Monte Carlo technique will conclude with a look at a simulation with a more complicated dynamic equation. We consider heat conduction in a material whose thermal conductivity k_1 varies continuously with x_1, but p and c are constant, as in Case II, equation (3.2.71). Again we consider only linear flow of heat in the x_1 direction, and from (3.2.71) the particle dynamics are

$$dx_1(t) = \frac{\partial K_1(x,t)}{\partial x_1}\, dt + (2K_1(x,t))^{\frac{1}{2}}\, dw(t). \qquad (3.2.102)$$

We consider the region $x_1 = [0, 5]$ with no heat flux across the boundaries, and the x_1 region is quantised into 5 equally spaced cells for data reduction purposes. We assume that the thermal conductivity varies linearly so that

$$K_1(x_1) = 0.1 + 0.18\, x_1.$$

The particle dynamics are then given by

$$dx_1(t) = 0.18\, dt + (2K_1(x_1))^{\frac{1}{2}}\, dw(t). \qquad (3.2.103)$$

The first term on the right hand side of (3.2.103) represents a constant drift of particles to the right, and the second term represents a dispersion which increases from left to right. It is this first term and the variation of the second term which distinguishes this example from the previous one, and we wish to know whether the simulation of (3.2.103) still represents the flux balance conditions correctly.

It will be sufficient to consider the steady state, uniform temperature, situation, and see whether the simulation of (3.2.103) with reflective boundary conditions at $x_1 = 0$ and 5 (as no heat flux passes these points) preserves the uniform temperature distribution. Unfortunately the integrals in the analysis of Appendix C are difficult to evaluate in this case, and we will have to be content with an experimental test to show that the net passage of particles past any point is zero on the average.

A simulation was initiated with a uniform distribution of 500 particles per cell, and the trajectories were integrated forward in time with the dynamics (3.2.103) until 100 independent steady state solutions were recorded. Our first concern is that the drift term of (3.2.103) should not bias the trajectories too much to the left or right, and this can be quickly checked by finding the mean value $E[x_1]$ of the trajectories in the simulation. This is equivalent to the functional solution (3.2.60) with $G(x) = \phi^{-1} x_1$, and as $U(x,t)$ should be uniform in $x_1 = [0, 5]$, the answer should be 2.5.

Of the 500 x 5 x 100 = 250,000 solution points obtained, the mean value was 2.4963, an error of $-.0037$. But the standard deviation of the error is $(250{,}000)^{-\frac{1}{2}}$ times the standard deviation of x_1 (= 1.442), and is 0.0029. Thus the assumption that the mean value of $x_1 = 2.5$ is quite plausible, as 20% of points from a Gaussian distribution of standard deviation 0.0029 have deviations from the mean greater than 0.0037.

Another check on our simulation is the chi-square test for a uniform distribution of cell densities. A typical calculation of the Chi-square variate u_j (3.2.101) is shown in Table 3.2.4.

Cell	1	2	3	4	5
n_i	487	511	494	531	477
$E[n_i]$	500	500	500	500	500
$n_i - E[n_i]$	-13	11	-6	31	-23
$(n_i - E[n_i])^2$	169	121	36	961	529
$\div E[n_i]$	.338	.242	.072	1.922	1.058

Sum of last row = u_j = 3.632

Table 3.2.4 Calculation of Chi-square Variate

Of the 100 trials, 43 of them fell within the 50% mid-range of the chi-square distribution of 4 degrees of freedom (50% mid-range = [1.92, 5,39]). Alternatively, as before, a chi-square variate can be formed from the sum of particles in each cell over the 100 trials. This equalled 5.07, which is not too far from the expected value of 4.0 and within the 50% mid-range. Thus the statistical results obtained are quite consistent with the null hypothesis of a uniform particle distribution being true. This means that there is no statistical evidence that our Monte Carlo solution is biased in the present case.

This sort of statistical test could be carried out under a wide variety of conditions on the heat conduction equation, and other parabolic equations as well. However, the orientation of the project did not justify further experimental work along these lines, particularly as we wish to study the simulation of more complicated equations than are met in the solution of parabolic equations.

3.2.7 Summary of Monte Carlo Solutions

In this chapter we have discussed the solution by Monte Carlo methods of a wide class of linear parabolic partial differential equations. If the parabolic equation is a Fokker-Planck equation of a particular system, then an obvious way of solving the parabolic equation is to simulate the given system. The collected statistics of the system then constitutes a Monte Carlo solution of the parabolic equation.

In Section 3.2.2 it is shown how this method can be extended to solve parabolic equations of a more general form than FP equations. As FP equations describe the statistics of a real Markov system, then the simulation in the Monte Carlo solution involved forming realisations of trajectories of the system, each of which began at $t = t_o$ and existed during the entire time range of the parabolic equation. This is to say that the simulation involved the conservation of trajectories.

Although some parabolic equations like some special cases of the equation of heat conduction can be put in the form of the FP equation and solved by the same Monte Carlo method, parabolic equations in general cannot be put in the FP form and thus must be solved by a modified Monte Carlo method. The modification involves simulating a system which has a FP equation which is as similar as possible in term by term comparison to the given parabolic equation, and then allowing trajectories to originate and expire during the time range of parabolic equation, in such a way as to simulate the differences between the given parabolic equation and its "nearest" FP equation. This latter operation means that trajectories are not conserved during the simulation, and thus only those parabolic equations which are of the FP form can be solved by a simulation method involving conserved trajectories.

If the parabolic equation has spatial discontinuities or boundaries, the equation does not hold at these points, and the normal rules of simulation may not be sufficient to define the behaviour of the trajectories at these points. In such cases, we must rely on auxiliary conditions suggested by the physical nature

of the problem to determine the behaviour of the simulated trajectories at edge or discontinuity boundaries. These auxiliary conditions may introduce a violation of the principle of conservation of trajectories at the boundaries, even though trajectories are conserved in the main part of the simulation. In the heat conduction equation, the physical laws of conservation of heat energy and the specification of thermal flux by the temperature gradient were used to construct the auxiliary conditions at the edge and discontinuity boundaries.

Also in Section 3.2.2 we presented several theorems which gave the conditions under which the solution of the parabolic equation remained positive. The Monte Carlo solution methods subsequently presented agreed with the positivity theorems, for the Monte Carlo solution involved the simulation of trajectories with positive weights except when a constant term $W(x,t)$ which was negative existed in the parabolic equation. Then trajectories with negative weights appeared which could lead to a negative Monte Carlo solution. Theorem 3.2.4 showed that this was the only condition under which the parabolic equation could have a negative solution. Although the positivity theorems did not consider the case where edge or discontinuity boundaries were present, the auxiliary conditions imposed on the simulated trajectories at these boundaries did not introduce trajectories with negative weights if these were not already present in the simulation.

Comparing Monte Carlo methods with the direct methods for solving parabolic equations discussed in Chapter 2, the most striking differences are provided by the effect of dimensionality and accuracy on solution effort. By dimensionality we refer to the number of independent variables of the parabolic equation other than time, and we recall that dimensionality was a severe restriction on the size of problem which could be handled by the direct methods. The most versatile of the two direct methods considered was the finite difference method, and in this case the storage requirements and computing time were proportional to the power of the dimensionality.

This was principally because the finite difference method required that the solution be obtained at all space points at each time stage, regardless of whether we are interested in them or not. By contrast, in the Monte Carlo method, the solution which is obtained by data reduction operations on the simulated trajectories is only obtained at those points or over those space variables which are of interest. The solution effort, for a given accuracy, is roughly proportional to the number of solution points obtained, but it is likely that very few more solution points would be desired for a typical large dimension problem than for one of low dimension. Thus a Monte Carlo solution of a high dimension problem may entail little more effort than a low dimension one, provided the mechanics of simulation are not unduly complicated for the high dimension problem.

On the question of solution accuracy, the Monte Carlo method has different considerations from the finite difference method. In the Monte Carlo method, the standard deviation of the solution error was inversely proportional to the square root of the number of trajectories in the simulation, while in the finite difference method, the solution error was proportional to the square of the independent variable mesh quantisation, Δx and Δt. Thus to reduce the solution error by a factor of 4, we must simulate 16 times as many trajectories in the Monte Carlo method, but in the finite difference method we must halve each independent variable mesh size. This means a factor of 2^{n+1} more solution points to compute and store, where n is the number of space dimensions in the problem. Thus for a 1-dimensional finite difference solution, we need 4 times as many solution points to reduce the error by a factor of 4; for a two-dimensional problem, 8 times as many points; for a three-dimensional problem, 16 times as many points, and so on. Thus only for the one-dimensional finite difference method is an increase in effort directly rewarded by a proportional increase in accuracy. For higher dimension finite difference problems and for Monte Carlo methods for all problems, increase of effort is not rewarded by a proportional increase of accuracy, and it will not in general be feasible to obtain solutions of arbitrarily high accuracy. Note also that the accuracy considerations of the Monte Carlo method are independent

of dimensionality, whereas the finite difference method is again plagued by the curse of dimensionality.

This chapter has only briefly looked at Monte Carlo methods for solving parabolic equations. The original intention was to look at methods of solution for the FP equation, and the Monte Carlo method involved a direct simulation of the continuous Markov system whose statistics the FP equation was describing. However, this idea indicated a link between diffusion processes and more general types of parabolic equations, and in exploring this link, the contents of this chapter developed.

As a result, a new method of solving linear parabolic equations was developed based on the FP or forward Kolmogorov equation of a diffusion process. Previous Monte Carlo methods for parabolic (or elliptic) equations had been based on the backward Kolmogorov equation and are called backward simulation methods. It was noted that the backward simulation method is complementary to our forward simulation method in some ways, but also the forward simulation method seems to handle problems of a wider generality than the backward method. Some points of comparison are listed in Table 3.2.5. The question marks indicate restrictions on the backward simulation method which may be removed by future research.

Some experimental work was presented using examples of the equation of heat conduction. The results supported the theory given earlier, but more work has to be done before the viability of the method is fully checked.

In the latter phases of this thesis topic, interest has centred around the relation between physical processes and diffusion processes. This relation is particularly relevant to the simulation of the diffusion processes of this chapter on a physical computer, but it turned out that the examples of this chapter were not convenient to illustrate the points of interest. Some recent developments in non-linear filtering did provide some useful examples, and our discussion of simulation techniques will treat non-linear filtering and other examples in Chapters 5 & 6.

CHAPTER 4

The Relation Between Physical and Diffusion Processes with Applications to Simulation

In the last chapter, we presented a Monte Carlo or simulation method of obtaining an approximate numerical solution to a wide class of parabolic partial differential equations. The method involved an extension of the relation between the solution of the Fokker-Planck equation of a diffusion process and a direct simulation of the diffusion process.

Thus the Monte Carlo methods require the simulation on the computer of a diffusion process. A computer, being a physical device, can only represent or compute with band-limited signals, and so in the light of the discussion of Section 2.2, a computer can only represent a physical process exactly, and not a diffusion process. Thus to simulate a given diffusion process, we must choose a physical process suitable for representation on a computer which models the essential statistical behaviour of the diffusion process.

Completing the circle, it is sometimes of interest to find a diffusion process which models the essential statistical behaviour of a given physical process. This approach is especially helpful when we are interested in the transient statistical behaviour of physical systems, for this transient behaviour cannot be simply obtained by analytic means for non-Markovian (i.e. physical) processes. In contrast, for Markovian (ie. diffusion)processes transient F.P. techniques are at our disposal, and particularly simple transient solutions are obtained if the process is linear (see [74]; an example is given in [57]).

Both these problems can be approached by studying the relation between physical and diffusion processes. This relation was introduced in Section 2.2 where we discussed the results of Clark, and has been discussed in different forms by Wong and Zakai [24, 25, 102] and Stratonovich [21]. In Section 4.1 we derive approximate

expressions for the second order statistics for the increments of a physical process, and choose an "equivalent" diffusion process by relating these statistics to the first two incremental moments of a diffusion process. Our argument, although basically similar to Stratonovich's, emphasises the method of characterising physical noise processes introduced by Clark. Unlike Clark and Wong and Zakai, we are not primarily concerned with the limiting properties of families of physical processes, but we explore the statistical relationship between a diffusion process and one given physical process (in this sense we follow Stratonovich). In this way, we show what diffusion process approximately shares the statistical properties of the given physical process. Our results duplicate Stratonovich's except that

(a) we include the non-stationary noise case;

(b) the manner in which we introduce the approximations gives a different, and perhaps clearer, insight into the conditions under which the statistical properties of the systems are approximately equivalent;

(c) owing to the structure of our formulation, our results are easier to apply to at least some physical systems found in practice. This is mainly due to our method of matching incremental moments (e.g. see [57]), but is also helped by our method of characterising physical noise processes.

In Section 4.2 we do consider the limiting properties of families of physical processes but not in the rigourous fashion of Clark and Wong and Zakai. In determining what diffusion process coincides with the limiting member of a family of physical processes as the upper frequency of the physical noise is extended to infinity, we consider convergence in distribution while Clark and Wong and Zakai consider the more confining convergence in mean square. While their use of convergence is pertinent to the individual sample paths of the processes, our use of convergence is pertinent to the statistical properties of ensembles of the processes, and as such is sufficient to show the convergence of simulation exercises. The value of Section 4.2 is to show the consistency of the approach of Section 4.1 as the upper frequency parameter of the physical noise is

refined, with particular reference to the method of characterising the physical noise process.

Using the characterisation we have chosen, we show that the limiting process is equivalent to replacing the physical noise by a special type of white noise which preserves the parameters of the original physical noise. At this stage we are faced with the choice of stochastic calculus we use to describe the process with white noise, and find that the use of the Stratonovich calculus in conjunction with the special type of white noise gives the stochastic equation which is closest to the original physical equation (and thus can be said to be more physically meaningful than other stochastic equations). We state the relation between this stochastic equation and the Stratonovich equation using the conventional white noise, and the Ito equation.

The results of Sections 4.1 and 4.2 are useful for choosing equivalent diffusion processes for physical processes found in practice, and in Section 4.3 we apply these results to the analysis of the topical example of linear systems with random coefficients. Our analysis of linear systems with random coefficients is an extension of the literature on the subject in the sense that we allow a more general form of physical noise process than previous authors.

In Section 4.4, we rephrase the results of Sections 4.1 and 4.2 so that we can choose a suitable physical process which is equivalent to a given diffusion process in order that the latter can be approximately simulated on a computer.

We also mention the problem of simulating a given physical process on a computer. If the noise of the physical process cannot be conveniently duplicated on the computer, a suitable noise must be chosen which then defines a second physical process. It turns out that the statistical equivalence of these physical processes can be discussed in terms of their equivalent diffusion processes, and this is done in Section 4.5.

4.1 Choosing an Equivalent Diffusion Process for a Physical Process by Matching Finite Incremental Statistics

Consider the diffusion process $x(t)$ described by the Ito stochastic differential equation (s.d.e):

$$dx(t) = f(x,t)\,dt + F(x,t)\,dw(t). \tag{4.1.1}$$

As $x(t)$ is a continuous Markov process, the statistical properties of $x(t)$ are completely specified by the first two incremental moments of the process (c.f. Section 2.1 or Appendix A)

$$b(x,t) = \underset{\delta t \downarrow o}{\text{limit}} \frac{1}{\delta t} E[\delta x \,|\, x, t] = f(x,t) \tag{4.1.2a}$$

and

$$a(x,t) = \underset{\delta t \downarrow o}{\text{limit}} \frac{1}{\delta t} E[\delta x\, \delta x^T \,|\, x,t] = FF^T(x,t). \tag{4.1.2b}$$

Now consider the physical process $X(t)$ described by the ordinary differential equation (o.d.e.)

$$\dot{X}(t) = g(X,t) + G(X,t)\, y(t). \tag{4.1.3}$$

As $X(t)$ is a non-Markovian process, its statistical properties are not completely specified by two incremental moments of the form (4.1.2). Indeed, the limiting operation involved in the definition of the second incremental moment (4.1.2b) results in a zero value for $a(X,t)$ when applied to physical processes, as $E[\delta X\, \delta X^T \,|\, X,t]$ is of order $O(\delta t^2)$ for physical processes for small δt.

This last point was expressed by Doob in 1942 [104] who showed that the classical Brownian motion process for which $E[\delta X\, \delta X^T \,|\, X,t]$ is of order $O(\delta t)$ is not physically realisable, and that the Brownian motion-type process occurring in practice is a smoothed version of the classical one. The smoothed version has second increments $E[\delta X\, \delta X^T \,|\, X,t]$ of order $O(\delta t^2)$, but if the time increment δt is made large enough, the second increment becomes proportional to δt rather than δt^2. For this to occur, δt must be somewhat larger than the memory time of the smoothing

device operating on the true Brownian noise. The second increment of the physical process (4.1.3) also has this property when δt is chosen somewhat larger than the significant memory time of the physical noise $y(t)$. This is a general property of continuous non-Markovian processes of the form (4.1.3).

A condition that a continuous process be a Markov process is that successive increments of the process $X(t_2) - X(t_1)$, $X(t_1) - X(t_o)$ etc. are independent of each other, even as the time increments $(t_2 - t_1)$, $(t_1 - t_o)$ are reduced to zero. Closely connected with the property of the preceding paragraph, the physical process (4.1.3) has successive increments which are approximately independent provided that the time increments are not reduced below δt, where δt is somewhat larger than the significant memory time of the physical noise $y(t)$. For time increments smaller than δt, the increments of $X(t)$ are no longer nearly independent and the process appears non-Markovian. However, if we are content to observe the physical process $X(t)$ only every δt time units, then it will appear to us to be a Markovian process. This is equivalent to saying we will only observe the process $X(t)$ through an instrument whose upper frequency of resolution is approximately δt^{-1}.

As the physical process $X(t)$ appears to be Markovian over time increments of the order of δt, and the diffusion process $x(t)$ is defined by the two incremental properties (4.1.2) which can be written as

$$E[\delta x \,|\, x,t] = f(x,t)\,\delta t + o(\delta t), \qquad (4.1.4a)$$

$$\text{and} \quad E[\delta x\, \delta x^T \,|\, x,t] = FF^T(x,t)\,\delta t + o(\delta t), \qquad (4.1.4b)$$

for a finite time increment δt (i.e. not taking the limit $\delta t \downarrow o$), it seems reasonable to choose a diffusion approximation to $X(t)$ by matching the quantities (4.1.4) of an arbitrary diffusion process to $E[\delta X \,|\, X,t]$ and $E[\delta X\, \delta X^T \,|\, X,t]$ of the given physical process, over time increments δt for which $X(t)$ appears Markovian, i.e. δt somewhat greater than the memory time of the noise $y(t)$.

This is the approach of this section, and the arguments of the two preceding paragraphs present a heuristic justification for the method, which is similar to that given by Stratonovich [21: Ch. 4, Sec. 7]. The justification will become clearer after the following derivation, when we discuss the conditions under which the modelling of physical processes by diffusion processes is valid, and state what properties of the physical process the diffusion process is expected to model.

4.1.1 Derivation of $E[\delta X | X,t]$ for a Physical Process

Consider the n-dimension physical process $X(t)$ described by the o.d.e.

$$\dot{X}(t) = G(X,t)\, y(t) \qquad (4.1.5)$$

where $y(t)$ is an m-dimension zero mean, non-stationary, physical noise process which is specified only by its matrix correlation function

$$R(t,\tau) = E[y(t)\ y(t-\tau)^T]. \qquad (4.1.5')$$

The noise $y(t)$ has a memory or correlation time τ_{cor}, which is defined so that $R(t,\tau)$ is essentially zero for $|\tau| > \tau_{cor}$ for all t in the time domain of interest. The quantity τ_{cor}^{-1} is of the order of the upper frequency f_u of the physical noise. We also need a quantity called the response time or relaxation time τ_{rel} of the system [21, p.99], It is analogous* to the time constant of a linear system, and for the system (4.1.5) it can be defined as

$$\tau_{rel} = E[G_X\ (2D)^{\frac{1}{2}}]^{-1} \qquad (4.1.5'')$$

where $2D$ is the intensity coefficient of the noise $y(t)$. Then τ_{rel}^{-1} is of the order of the upper frequency response of the

---

* i.e. of the order of

system, and the ratio of τ_{rel} to τ_{cor} gives the relative time scales of response of the system and noise.

Compared with equation (4.1.3) we consider the physical process with $g(X,t) = 0$ here, as this non-random term contributes no unusual properties to the increments derived below: the contribution to the first increment is $g(X,t)\,\delta t$ and to the second increment is $O(\delta t^2)$ - thus the term $g(X,t)$ goes over directly into the diffusion model. Compared with Stratonovich [21: Ch.4, Sec.9] we allow $y(t)$ to be non-stationary, and compared with Clark [22] we allow $y(t)$ to be non-stationary and non-Gaussian as well.

To obtain an expression for $E[\delta X \mid X,t]$ we write the state vector equation (4.1.5) in component form

$$\dot{X}_i(t) = \sum_k^m G_{ik}(X,t)\, y_k(t) \quad , \qquad i = 1, n, \tag{4.1.6}$$

and consider a typical component of δX,

$$\delta X_i(t) = X_i(t + \delta t) - X_i(t), \tag{4.1.7}$$

where $\delta \cdot$ is a forward difference operator operating over a time increment δt. Integrating (4.1.6) we have

$$\delta X_i(t) = \sum_k^m \int_t^{t+\delta t} G_{ik}(X,u)\, y_k(u)\, du. \tag{4.1.8}$$

We can see right away that if G is not a function of X, $E[\delta X \mid X,t] = 0$, as G is then independent of y, and y has a zero mean value. However, the interesting case is when G does depend on X, and we see below how this dependence affects $E[\delta X \mid X,t]$.

We can express the dependence of $G_{ik}(X,u)$ of (4.1.8) on the u parameter by the integral relation

$$G_{ik}(X,u) = G_{ik}(X,t) + \sum_j^n \int_t^u \frac{\partial G_{ik}(X,v)}{\partial X_j(v)}\, \dot{X}_j(v)\, dv$$

$$+ \int_t^u \frac{\partial G_{ik}(X,v)}{\partial v}\, dv, \qquad (4.1.9)$$

and then can write (4.1.8) as

$$\delta X_i(t) = \sum_k^m \int_t^{t+\delta t} G_{ik}(X,t)\, y_k(u)\, du$$

$$+ \sum_{j,k} \int_t^{t+\delta t} \int_t^u \frac{\partial G_{ik}(X,v)}{\partial X_j(v)} \cdot X_j(v)\, dv\, y_k(u)\, du$$

$$+ \sum_k^m \int_t^{t+\delta t} \int_t^u \frac{\partial G_{ik}(X,v)}{\partial v}\, dv\, y_k(u)\, du,$$

$$= T_1 + T_2 + T_3 . \qquad (4.1.10)$$

Assumption A1: We assume that the contribution of T_1 to the conditional expectation of equation (4.1.10) can be neglected. That is, we assume

$$E\Big[\sum_k^m \int_t^{t+\delta t} G_{ik}(X,t)\, y_k(u)\, du \;\Big|\; X,t\Big] = 0. \qquad (4.1.11)$$

The approximation involved in this assumption relies on the significant memory time of each component of $y(t)$ being small with respect to δt. To see this, suppose that we can bound each component of the correlation function of $y(t)$ by an exponential:

$$\left|R_{jk}(t,\tau)\right| \leqslant M e^{-|\tau|/N}, \quad j, k = 1, m, \qquad (4.1.12)$$

where M is finite and the maximum correlation time $\tau_{cor} = 5N \ll \delta t$. The left hand side of (4.1.11) is a random variable and we must show that it has a small mean and variance. For example, the mean value is

$$\sum_k^m \int_t^{t+\delta t} E[G_{ik}(X,t)\, y_k(u)]\, du. \qquad (4.1.13)$$

Now as $G_{ik}(X,t)$ does not depend explicitly on $y_k(u)$ for $u > t$ (or vice versa), we can express the correlation between $X(t)$ and $y(u)$ via their mutual dependence on the intermediate variable $y(t)$. This gives an explicit functional dependence of the integrand of (4.1.13) on u, by using the normalized correlation function

$$\rho_{jk}(t,\tau) = \frac{R_{jk}(t,\tau)}{R_{jk}(t,o)} \leqslant e^{-|\tau|/N} . \qquad (4.1.14)$$

Then we can write (4.1.13) as

$$\sum_{k}^{m} \int_{t}^{t+\delta t} \sum_{j}^{m} E[G_{ik}(X,t)y_j(t)]\, \rho_{jk}(t,\ t-u)\ du \qquad (4.1.15)$$

which, from (4.1.14), is less than or equal to

$$\sum_{k,j}^{m} E[G_{ik}(X,t)\ y_j(t)]\ N, \qquad (4.1.16)$$

provided $\tau_{cor} \ll \delta t$. Now $N = \tau_{cor}/5 = O(\tau_{cor})$, but we cannot explicitly evaluate $E[\cdot]$ in (4.1.16) as we cannot eliminate the functional dependence on X. However as $X(t)$ is an integral function of $y(t')$, $t' \leqslant t$, and the response time τ_{rel} of $X(t)$ is much greater than the memory time of $y(t)$, the correlation between $y(t)$ and $X(t)$ is very small. This correlation goes to zero as τ_{cor} goes to zero, so it seems plausible that

$$E[G_{ik}(X,t)\ y_j(t)] = O(\tau_{cor})$$

but we cannot show this. Thus we shall say that the mean value (4.1.13) is $O(\tau_{cor})$ times a small factor. A similar argument would show that the mean square of the left hand side of (4.1.11) is $O(\tau_{cor}^{2})$ with a small coefficient.

This completes our justification of introducing assumption A1. The expected error in the conditional increment $E[\delta X \mid X,t]$ is at most $O(\tau_{cor})$ provided $\tau_{cor} \ll \tau_{rel}$.

Returning to the main argument and using assumption A1, we substitute the value of $\dot{X}_j(v)$ from (4.1.6) into T_2 of (4.1.10). The equation for the increment in $X_i(t)$ then becomes*

$$\delta X_i(t) = \sum_{j,k} \int_t^{t+\delta t} \int_t^u \frac{\partial G_{ik}}{\partial X_j}(v) \left[\sum_l^m G_{jl}(v)\, y_l(v) \right] dv\, y_k(u)\, du$$

$$+ \sum_k^m \int_t^{t+\delta t} \int_t^u \frac{\partial G_{ik}}{\partial v}(v)\, dv\, y_k(u)\, du. \qquad (4.1.17)$$

Assumption A2: We assume that the variations of G, G_X and G_t are finite within the interval $(t, t+\delta t)$. Then we can change the time parameter of these functions from v to t and bring them outside the $\int_t^u . \, dv$ integral with an error of only $O(\delta t)$, as $(v - t) \leqslant \delta t$. However these terms are also arguments of the $\int_t^{t+\delta t} . \, du$ integral, and when this is evaluated, the contribution of A2 to the error of $\delta X_i(t)$ is $O(\delta t^2)$.

Now the second term of (4.1.17) is the same as T_1 of (4.1.10) with $G_{ik}(X,t)$ replaced by $\frac{\partial G_{ik}(X,t)}{\partial t}(u - t)$ where $(u - t)$ is $O(\delta t)$. Thus this term is approximately an order smaller than T_1 of (4.1.10) and is neglected. Equation (4.1.17) then becomes

$$\delta X_i(t) = \sum_{j,k,l} \frac{\partial G_{ik}}{\partial X_j}(t)\, G_{jl}(t) \int_t^{t+\delta t} \int_t^u y_k(u)\, y_l(v)\, dv\, du. \qquad (4.1.18)$$

As an heuristic explanation, the validity of the approximation involved in assumption A2 relies upon $G(X,v)$ and $G_X(X,v)$

---

* We continue to use equality signs in the subsequent equations, keeping in mind the assumptions and approximations made. We also write $G(X,v)$ as $G(v)$ when it is specifically the time parameter in which we are interested.

remaining constant for v in $[t, t+\delta t]$. This in turn depends on $X(v)$ remaining constant in this interval, and we know that significant changes in X occur in the order of the response or relaxation time τ_{rel} of the system. Thus we need $\delta t \ll \tau_{rel}$ for assumption A2 to be valid.

Taking the conditional expectation (given X,t) of both sides of (4.1.18), and remembering that G is a non-random function of X, we have

$$E[\delta X_i(t) \mid X,t] = \sum_{j,k,l} \frac{\partial G_{ik}}{\partial X_j}(t)\, G_{jl}(t) \int_t^{t+\delta t} \int_t^u R_{kl}(u,u-v)\, dv\, du \qquad (4.1.19)$$

where $R(v,v-u)$ is the non-stationary correlation function of $y(t)$ as defined in equation (4.1.5'). The use of the correlation function in (4.1.19) assumes that $y(u)$ and $y(v)$ are independent of $G_X(t)$ and $G(t)$, as we did in assumption A1.

Assumption A3: First note that by putting $u-v = \tau$, we can write

$$\int_t^u R_{kl}(u,u-v)\, dv \quad \text{as} \quad \int_0^{u-t} R_{kl}(u,\tau)\, d\tau \qquad (4.1.20)$$

in equation (4.1.19).

Then we define the quantity

$$A_{kl}(u) = \int_0^\infty R_{kl}(u,\tau)\, d\tau \,, \qquad k, l = 1, m, \qquad (4.1.21)$$

as the non-stationary characteristic matrix of the physical noise $y(t)$. As discussed after equation (2.2.4), this definition of $A(u)$ differs from the earlier definition (2.2.3), or Clark's, in that the limit of the upper frequency of the noise f_u going to infinity has not been taken here. In other words, the A matrix defined above is the characteristic matrix of a particular physical noise process $y(t)$ having the correlation function $R(t,\tau)$, whereas the earlier definition was of the characteristic matrix of the

limiting member of a family of physical processes. The definition (4.1.21) also allows for the non-stationarity of the physical noise $y(t)$, although in assumption A4 below we require that the change in $A(t)$ be bounded over time intervals of δt, and in Section 4.2, we assume that the change in $A(t)$ is negligible over time intervals of τ_{cor}. In the rest of the thesis, we shall assume that, unless otherwise stated, A and B (2.2.2c) are functions of time without necessarily writing the t parameter.

In assumption A3 we replace the quantity (4.1.20) in (4.1.19) by the quantity $A_{kl}(u)$ of (4.1.21). This assumption changes the upper limit of the integral (4.1.20) from $(u - t)$ to infinity, but we note that τ_{cor} is defined so that

$$\int_0^{\tau_{cor}} R(u,\tau)\, d\tau \doteq \int_0^{\infty} R(u,\tau)\, d\tau, \qquad (4.1.22)$$

and so the error involved in A3 is negligible for all u in the interval $[t, t+\delta t]$ except $u < t + \tau_{cor}$. Thus the error is negligible for all but the ratio $\tau_{cor}/\delta t$ of the u interval, which means that for a fixed δt, the error vanishes linearly with τ_{cor} (i.e. the error is $O(\tau_{cor})$ assuming the integral of the correlation function remains bounded). But even for $u < t + \tau_{cor}$ the error is not too large, for most of the value of the integral (4.1.22) is accumulated very near the origin, $\tau = 0$. In any case, the validity of the approximation of A3 relies upon $\tau_{cor} << \delta t$, and again we note that, like A1, the error involved goes to zero as τ_{cor} goes to zero.

Using assumption A3, the conditional increment (4.1.19) becomes

$$E[\, \delta X_i(t) \mid X,t \,] = \sum_{j,k,l} \frac{\partial G_{ik}}{\partial X_j}(t)\, G_{jl}(t) \int_t^{t+\delta t} A_{kl}(u)\, du. \qquad (4.1.23)$$

Assumption A4: We assume that the variation of $A(u)$ in the interval $[t, t+\delta t]$ is bounded so that

$$\int_t^{t+\delta t} A_{kl}(u)\, du = A_{kl}(t)\,\delta t + O(\delta t^2).$$

This is the same order of error as involved in assumption A2, and using A4 the conditional increment becomes

$$E[\delta X_i(t) \mid X,t] = \sum_{j,k,l} \frac{\partial G_{ik}}{\partial X_j}(t)\, G_{jl}(t)\, A_{kl}(t)\,\delta t,$$

$$= \sum_{k,l} (Q_{kl}(X,t))_i\, A_{kl}(t)\ \delta t, \qquad (4.1.24a)$$

or in vector form

$$E[\delta X(t) \mid X,t] = \sum_{k,l} Q_{kl}(X,t)\, A_{kl}(t)\ \delta t \qquad (4.1.24b)$$

where Q_{kl} is the n-vector introduced in (2.2.2a) as a shorthand way of writing the sum over j of the $G_x G$ term. Equation (4.1.24) is the main result we seek, as it gives an approximate value for the expected drift $\delta X(t)$ of the physical process $X(t)$ during time δt, where δt is much greater than the memory time τ_{cor} of the noise $y(t)$, but much less than the response time τ_{rel} of the process $X(t)$.

Comments on the Approximations

Essentially the difficulty in making precise statements about the magnitude or even order of the errors made in the approximations above stems from the non-Markovian nature of the physical process $X(t)$. We are evaluating the expected value of $\delta X(t)$ given $X(t)$, but as X is not a Markov process, the knowledge of X or its probability density at time t does not give us the future statistics

of X, as it would if X were Markov. The processes $X(t)$ and $y(t)$ may form a joint Markov process, but we are not given $y(t)$ or even the structure of the generator of $y(t)$, and hence cannot exactly find the future statistics of X.

Then the best we can do is average the contribution of $y(u)$ over the interval $[t, t+\delta t]$, but this does not make up for the lack of "initial condition" information of $y(t)$. However, despite the lack of precision in specifying the accuracy of the approximations made, the main point of the above analysis is to show that the contribution of the noise $y(u)$ to the expected drift $\delta X(t)$ is approximately given by the semi-infinite integral of the correlation function of $y(u)$, (4.1.21). We have shown that the expression (4.1.24) is a good $O(\delta t^2)$ approximation for values of τ_{cor} much less than δt but greater than zero, and in the next section, we use the fact that the approximation becomes exactly $O(\delta t^2)$ as τ_{cor} tends to zero (the errors involved in A1 and A3 then disappear).

4.1.2 Derivation of $E[\delta X \delta X^T | X,t]$ for a Physical Process

The derivation of the second order conditional increment follows that of the first order conditional increment given above, and involves assumptions equivalent to A1, A2, A3 and A4. The same remarks apply to the validity of the assumptions, and the derivation is outlined briefly below.

Using the expression (4.1.8) for the increment $\delta X_i(t)$ and a similar expression for $\delta X_j(t)$, we can write the product of the increments as

$$\delta X_i(t) \delta X_j(t) = \sum_{k,l} \int_t^{t+\delta t} \int G_{ik}(u)\, y_k(u)\, G_{jl}(v)\, y_l(v)\, dv\, du \,. \qquad (4.1.25)$$

We need not bother with the expression (4.1.9) which translates the time parameter from t to u and v, as there are now two $[t, t+\delta t]$ integrals and we can directly write

$$G_{ik}(u)\, G_{jl}(v) \;=\; G_{ik}(t)\, G_{jl}(t) \;+\; O(\delta t^2). \qquad (4.1.26)$$

Bringing the product (4.1.26) outside the integrals in (4.1.25) corresponds to approximation A2, and involves an $O(\delta t^2)$ error which is small if $\delta t << \tau_{rel}$.

Then following (4.1.19) we write the second order conditional increment as

$$E[\delta X_i(t)\, \delta X_j(t) \,|\, X,t] \;=\; \sum_{k,l} G_{ik}(t)\, G_{jl}(t) \int_t^{t+\delta t}\!\!\int R_{kl}(u,u-v)\,dv\; du, \qquad (4.1.27)$$

where an assumption equivalent to A1 assumes that $G(t)$ is independent of $y(u)$ and $y(v)$ for $u,v \geq t$. Now consider the integral

$$\int_t^{t+\delta t} R_{kl}(u,\, u-v)\; dv \;=\; \int_{u-t-\delta t}^{u-t} R_{kl}(u,\tau)\; d\tau. \qquad (4.1.28)$$

The range of τ in this integral always includes zero (for all u in $[t,\, t+\delta t]$), and assuming $\tau_{cor} << \delta t$, we can set (4.1.28) approximately equal to

$$\int_{-\tau_{cor}}^{\tau_{cor}} R_{kl}(u,\tau)\; d\tau \;\doteq\; \int_{-\infty}^{\infty} R_{kl}(u,\tau)\; d\tau \qquad (4.1.29)$$

We now define a new quantity

$$A^*_{kl}(u) \;=\; \int_{-\infty}^{o} R_{kl}(u,\tau)\; d\tau, \qquad k,\, l \;=\; 1,\, m, \qquad (4.1.30)$$

which can be regarded as a supplementary characteristic matrix, supplementary to $A(u)$ defined earlier (4.1.21). Note that

$$A^*_{kl}(u) \;=\; \int_{-\infty}^{o} E[y_k(u)\; y_l(u-\tau)]\; d\tau$$

and putting $\tau' = -\tau$, we have

$$A^*_{kl}(u) \;=\; \int_{o}^{\infty} E[y_l(u+\tau')\; y_k(u)]\; d\tau'. \qquad (4.1.31)$$

Now if $y(t)$ is sufficiently stationary over time intervals of τ_{cor} so that we can put

$$E[y_l(u+\tau')\, y_k(u)] = E[y_l(u)\, y_k(u-\tau')], \qquad \tau' \leq \tau_{cor}, \tag{4.1.32}$$

equation (4.1.31) becomes

$$A^*_{kl}(u) = \int_0^\infty R_{lk}(u, \tau')\, d\tau'$$

$$= A_{lk}(u). \tag{4.1.33}$$

Thus we have $A^*(u) = A^T(u)$ and the information in the supplementary characteristic matrix (4.1.30) is contained in the characteristic matrix (4.1.21). In the sequel, when we refer to the characteristic matrix of a physical noise $y(t)$, we implicitly refer to $A(t)$ and $A^*(t)$ if the noise is non-stationary ($2m^2$ parameters specified), and to only $A(t)$ if the noise is stationary in the sense of (4.1.32) (m^2 parameters specified).

Noting that

$$A(u) + A^*(u) = \int_{-\infty}^{\infty} R(u,\tau)\, d\tau, \tag{4.1.34}$$

and using the approximation (4.1.29), we replace (4.1.28) by $[A_{kl}(u) + A^*_{kl}(u)]$ in (4.1.27) and obtain

$$E[\delta X_i(t)\, \delta X_j(t) \mid X,t] = \sum_{k,l} G_{ik}(t)\, G_{jl}(t) \int_t^{t+\delta t} [A_{kl}(u) + A^*_{kl}(u)]du. \tag{4.1.35}$$

This approximation is similar to A3, and the error involved is small provided $\tau_{cor} \ll t$. Again, this error approaches zero as τ_{cor} approaches zero.

Following A4 (an $O(\delta t^2)$ error) we then obtain (analogous to (4.1.24a)) for the second order conditional increment of $X(t)$

$$E[\delta X_i(t)\delta X_j(t) \mid X,t] = \sum_{k,l} G_{ik}(t)\, G_{jl}(t)\, [A_{kl}(t) + A^*_{kl}(t)]\, \delta t, \tag{4.1.36a}$$

or in vector form

$$E[\delta X(t)\, \delta X(t)^T \mid X,t] = G(t)\,[A(t) + A^*(t)]\, G^T(t)\, \delta t. \qquad (4.1.36b)$$

If δt was not much greater than τ_{cor}, the assumption A3 leading to equation (4.1.35) would not be valid, and we would find that $E[\delta X\, \delta X^T \mid X,t]$ is proportional to δt^2. This is the property of physical processes noted after equation (4.1.3), and is an indication that the increments of $X(t)$ appear Markovian over time increments δt substantially larger than the significant memory time of the noise, τ_{cor}.

4.1.3 A Diffusion Model for the Physical Process

As the increments of the physical process $X(t)$ appear Markovian over time increments $\delta t >> \tau_{cor}$, a plausible method of choosing a diffusion process $x(t)$ whose statistical behaviour models that of $X(t)$ is to choose one whose increments $\delta x(t)$ have approximately the same first and second order conditional expectations as the physical process over the same time increment δt. Approximate expressions for these quantities are given in equation (4.1.4), and Doob [20, Ch. 6, Sect. 3] shows that the existence of the local properties (4.1.4) defines a diffusion process, in particular the process $x(t)$ given in (4.1.1). Thus it is proposed to choose a diffusion model for $X(t)$ by matching (4.1.4a) to (4.1.24) and (4.1.4b) to (4.1.36).

Matching the first order increments (4.1.4a) and (4.1.24) we choose $f(x,t)$ to be

$$f(x,t) = \sum_{k,l} Q_{kl}(X,t)\, A_{kl}(t), \qquad (4.1.37)$$

and matching the second order increments (4.1.4b) and (4.1.36),

we choose $FF^T(x,t)$ to be

$$FF^T(x,t) = G(X,t)\,[A(t) + A^*(t)]\,G^T(X,t). \qquad (4.1.38)$$

The quantities $f(x,t)$ and $FF^T(x,t)$ are sufficient to define the statistical behaviour of a diffusion process $x(t)$. If a specific structure of the diffusion process is desired, $A + A^*$ must be factored into B and B^T so that

$$B\,B^T = A + A^*, \qquad (4.1.39)$$

and the diffusion process $x(t)$ is defined by the Ito s.d.e. (4.1.1)

$$dx(t) = \sum_{k,l} Q_{kl}(x,t)\,A_{kl}(t)\,dt + G(x,t)\,B(t)\,dw(t). \qquad (4.1.40)$$

The probability density $P(x,t)$ of the diffusion process (4.1.40) is described by the Fokker-Planck differential equation

$$\frac{\partial P}{\partial t} = -\sum_{i}^{n} \frac{\partial}{\partial x_i}[f_i P] + \tfrac{1}{2}\sum_{i,j}^{n} \frac{\partial^2}{\partial x_i \partial x_j}[(FF^T)_{ij} P] \qquad (4.1.41)$$

with suitable initial conditions $P(x,t_o)$. If $P(x,t_o)$ is a delta function $\delta(x-x_o)$ then the FP equation describes the transition probabilities of the diffusion process $P(x,t \mid x_o,t_o)$.

The following conjecture is the main result of Section 4.1:

The diffusion process (4.1.40) is a "diffusion model of" or an "equivalent diffusion process for" the physical process $X(t)$ of (4.1.5) in the following sense. The solution of the FP equation (4.1.41) gives approximate values for the probability density $P(X,t)$ or the transition probability density $P(X,t \mid X_o,t_o)$ of the physical process $X(t)$, for values of t sufficiently far removed from the "initial condition time" t_o so that $(t - t_o) >> \tau_{cor}$. The approximation is valid as long as the response time of the physical process, τ_{rel}, is much greater than the memory time of the physical noise, τ_{cor}.

The validity of the conjecture comes from the assumptions made in the derivation of the conditional moments of the increments of

the physical process. We have shown that the first two conditional moments of the physical process and its diffusion model approximate to each other over time increments δt which are much greater than τ_{cor} but much less than τ_{rel}. Thus the inequality

$$\tau_{rel} >> \tau_{cor} \tag{4.1.42}$$

is a general necessary condition that the conjecture be valid. This is the same condition expressed by Stratonovich [21, eqn. 4.184]. We could also have shown the approximate equality of higher conditional moments of the increments, but only the first two were needed to define the diffusion model. It is the fact that the moments of the conditional increments of the processes $(x-x_o \mid x_o,t_o)$ approximate to each other that shows that the transition probability density $P(x, t_o+\delta t \mid x_o,t_o)$ of the diffusion process (given by the FP equation) gives an approximate expression for that of the physical process. This also implies that $P(x, t_o+t \mid x_o,t_o)$ is an approximate expression for that of the physical process for all t which greatly exceeds the correlation time τ_{cor}, for this transition probability over large time increments can be constructed from joint transition probabilities involving successive jumps over time increments of the order of δt (this can be done for Markov processes, and we know that the physical process is approximately Markov for increments in the order of δt). Finally we extend the conjecture to say that $P(x, t_o+t)$ of the diffusion process gives an approximate expression for that of the physical process for an arbitrary initial distribution $P(x,t_o)$, for all $t >> \tau_{cor}$, for this function can be constructed from a convolution of the transition probabilities over the same time interval t and the initial distribution.

This is the same conjecture as presented by Stratonovich [21, Ch.4, Sects. 7-9], and in fact he justifies it more directly by deriving a FP-type equation for the physical process which approximately equals the FP equation (4.1.41) when the inequality (4.1.42) holds. He gives an "error" term which distinguishes the FP-type

equation of the physical process from the FP equation (4.1.41) which is small when the inequality (4.1.42) holds, but this error term is not useful for determining the error in the transition density function.

In addition, Stratonovich's method depends on the time derivative of the cumulant function of the increment $(x(t) - x_o)$ becoming constant as $t - t_o$ greatly exceeds τ_{cor} [21: p.85], and this assumption precludes the consideration of non-stationary noise sources. Although it is possible that Stratonovich's method could be modified to accommodate non-stationary noises, our method accommodates them with virtually no increase in complexity over stationary noises.

Apart from the non-stationarity allowed, the main advantage of our derivation over Stratonovich's is that ours emphasises the manner in which the properties of the physical noise affect the drift of the physical process (c.f. Section 4.1.1). It is this drift which represents the difference between apparently similar physical and diffusion processes, and is fundamental to the recent interest in the relation between physical and diffusion processes. Our method clearly shows how the drift is affected by some unusual noise source properties, such as the property of asymmetry introduced by Clark (that of the matrix A (4.1.21) being asymmetrical).

4.1.4 Experimental Results

Because of the difficulty of treating non-Markovian systems analytically in the time domain, the analysis above could not be very specific about the error involved in the conjecture of the last section, and in particular we could not specifically relate the error to the ratio τ_{rel}/τ_{cor}. We have, however, suggested that this ratio is the most important parameter influencing the accuracy of the diffusion model, and an example will give us an indication of how large this ratio must be in order to achieve good modelling accuracy. As the modelling accuracy is very good (better than 1% in mean square for large τ_{rel}/τ_{cor}) the example

below constitutes an experimental confirmation of the validity of the conjecture and of the usefulness of the method of transient analysis of non-Markovian systems.

The example consists of a linear first order filter driven by a pseudo random binary sequence (PRBS) and is reported in detail in [57]. Although a deterministic signal, the PRBS has the properties of a physical noise process, $y(t)$, and has an effective maximum correlation time of $\tau_{cor} = \Delta$, the basic bit interval.* The filter, being linear, has a simple relaxation time $\tau_{rel} = T$, the filter's time constant, and this parameter is associated with the physical process, $X(t)$, the filter's output.

As the transient statistics of the PRBS depend on the integral of the PRBS, $S(t)$, this quantity is the "state" of the noise, and when forming a diffusion model for the output of the filter, $X(t)$, the extra state variable, $s(t)$, which models $S(t)$, must be added to the model. As, in this example, $S(t)$ incorporates all the statistical information of the noise process, the construction of a diffusion model for $X(t)$ can proceed by first choosing a diffusion model for $S(t)$, and then the addition of the state variable $x(t)$ modelling $X(t)$ is straightforward. Thus the diffusion model is found by matching the first and second order conditional increments of $S(t)$ to $s(t)$ by the method of this chapter, and the state variable $x(t)$ was added by simply appending the differential equation of the filter. This results from a special property of the

---

* As the PRBS is periodic, the correlation function is also periodic, with the same period $L\Delta$. However, the correlation time τ_{cor} must be chosen only with regard to the central period of the correlation function, $|\tau| < \frac{1}{2} L\Delta$. This is because the effect of the periodicity is to make the spectrum discrete, and does not change the envelope, or more specifically, the effective upper frequency f_u of the spectrum, which is determined by the central period of the correlation function. It is this latter parameter which influences the validity of the results of this chapter, and thus $\tau_{cor} = O(f_u^{-1})$ is the effective correlation time of the central period of the correlation function.

PRBS, and for more general noise processes, the choice of a diffusion model for $X(t)$ would proceed directly by matching the increments of $X(t)$ and $x(t)$.

Although we did not directly use the correlation function or characteristic matrix of the PRBS in evaluating the conditional increments, the relevance of the quantities τ_{cor} and τ_{rel} given above is not diminished, and the example clearly shows the effect of the ratio τ_{rel}/τ_{cor} on modelling accuracy. As mentioned in [57], the percentage error between the mean square of $X(t)$ and $x(t)$ is felt to be a representative error measure, and this is shown as a function of the filter time constant $T = \tau_{rel}$ in Figure 4.1.1 as

$$100\,\frac{E[x^2(t)] - E[X^2(t)]}{E[X^2(t)]}\,. \qquad (4.1.43)$$

This error function is almost exactly uniform in time, and so the time dependence is not shown in Figure 4.1.1.

The function $E[x^2(t)]$ is derived in [57]. In curve 1 of Figure 4.1.1, the function $E[X^2(t)]$ is evaluated at the PRBS switching points, $t = \Delta, 2\Delta, 3\Delta \ldots$, while in curve 2, $E[X^2(t)]$ is estimated from the continuous curve $X(t)$ by sampling $X(t)$ every $\Delta/40$ time units. These estimates were obtained to a high accuracy on a digital computer, and the $E[\cdot]$ operation was performed by averaging over every possible starting point of the PRBS ($L = 127$ in this case). Although the "continuous" error of curve 2 is more appropriate to the modelling of general random processes by diffusion processes, it is felt that the "sampled" error of curve 1 is a fairer basis of judging modelling accuracy in the present case, because of the special property of the PRBS that it only exhibits randomness at the discrete points $t = \Delta, 2\Delta, 3\Delta \ldots$ For this reason, the diffusion model was formed by matching increments over time increments Δ, and the diffusion model $s(t)$ was not expected to model the fine structure of the integrated PRBS between the sample points (remember that the integral of the PRBS, $S(t)$, is a ramp function between the sample points, while the

diffusion model $s(t)$ has the fine random structure of a Brownian motion between the sample points - the two are meant to be statistically equivalent only at the sample points). This point is a result of the inherently discrete nature of the PRBS and is not a general property of our method of choosing diffusion processes to model continuous non-Markovian processes.

With this qualification, we can judge the modelling accuracy by curve 1 of the Figure 4.1.1. It is noted that the error depends strongly on τ_{rel}/τ_{cor} until τ_{rel} exceeds about $3\tau_{cor}$ when the error curve flattens out and remains constant. Thus $\tau_{rel} > 3\tau_{cor}$ could be taken as a criterion for good modelling accuracy, but it is perhaps more realistic to look at the quantity

$$\tau'_{rel} = 4T = 4\tau_{rel} \qquad (4.1.44)$$

as this represents the maximum "memory" of the filter, and is thus more analogous to τ_{cor}, the maximum memory of the noise, than τ_{rel} is. Thus our criteria for good modelling accuracy is that

$$\tau'_{rel} > 12\,\tau_{cor}, \qquad (4.1.45)$$

which says roughly that the upper frequency of the physical noise should be about an order of magnitude higher than the upper pass frequency of the system.

Thus it seems that the PRBS example is reasonable evidence in support of the conjecture of Section 4.1.3, but as the PRBS has rather special properties, experimental work should be carried out with more general types of physical noise and systems in order to test the conjecture thoroughly. However, the facilities were not available for making accurate (e.g. 0.1%) estimates of generalised random functions, and the PRBS has the distinctive advantage that its statistical properties can be measured exactly with a finite number of trials (equal to the length of the code - $L = 127$ in the example shown in Figure 4.1.1).

4.2 A Limiting Form of a Physical Process

In Section 4.1, we have shown two main points. Firstly, we have derived approximate expressions for the incremental statistics of physical processes driven by band limited noise sources, and showed how a useful "equivalent" diffusion process could be obtained from these expressions. The derivation has suggested conditions under which the diffusion process does model the physical process, and we have presented an example which verified that the diffusion process modelled the statistics of the physical process quite accurately within the stated conditions. The example suggests that this method is a powerful method of analysing the transient statistics of non-Markovian systems.

The second point is that we have shown what properties of the physical noise source contribute to the statistics of the physical process. These properties are summarised in a compact form in the characteristic matrices $A(t)$ and $A^*(t)$, and as such these matrices form a sufficient characterisation of the physical noise source when we are interested in the statistical properties of systems driven by the noise.

In this section, we consider a limiting operation on a physical process, whereby the upper frequency of the physical noise is extended to infinity in such a way as to preserve the characteristic parameters of the physical noise. With the important assumption that the limiting physical process is a Markov process, we show that the physical process converges in distribution to a particular diffusion process, and this diffusion process is the same as the "equivalent" diffusion process of the previous section. Thus the limiting operation shows the consistency of choosing equivalent diffusion processes by the method of the last section.

Consider the physical process $X(t)$ defined by the ordinary differential equation

$$\dot{X}(t) = G(X,t)\, y(t), \tag{4.2.1}$$

where the m vector noise process $y(t)$ is defined only by its non-stationary correlation function

$$R(t,\tau,\tau_{cor}) = E[y(t)\, y^T(t-\tau)]. \tag{4.2.2}$$

The maximum correlation time parameter τ_{cor} is defined more rigidly than in the last section as being the maximum time shift $|\tau|$ in equation (4.2.2) for which $R(t,\tau,\tau_{cor})$ is non-zero. The characteristic matrices are then defined in terms of $R(t,\tau,\tau_{cor})$ as

$$A(t,\tau_{cor}) = \int_0^{\tau_{cor}} R(t,\tau,\tau_{cor})\, d\tau, \tag{4.2.3a}$$

and

$$A^*(t,\tau_{cor}) = \int_{\tau_{cor}}^{0} R(t,\tau,\tau_{cor})\, d\tau. \tag{4.2.3b}$$

As before, we assume that the functions $G(t)$, $G_X(t)$, $A(t)$ and $A^*(t)$ are of finite variation in the finite time range of interest.

Consider the following limiting operation on $X(t) = X(t,\tau_{cor})$. We extend the upper frequency of the noise source $y(t)$ to infinity in such a way that τ_{cor} tends to zero, and the noise characteristic parameters (4.2.3) are unaltered. The following theorem gives the statistical structure of the limiting physical process $X(t,o)$.

Theorem 4.2.1. If τ_{cor} tends to zero in such a way that $A(t,\tau_{cor})$ and $A^*(t,\tau_{cor})$ are independent of τ_{cor}, the physical process $X(t,\tau_{cor})$ of equation (4.2.1) converges in distribution to the diffusion process $x(t)$ defined by the Ito s.d.e.

$$dx(t) = \sum_{k,l} Q_{kl}(x,t)\, A_{kl}(t)\, dt + G(x,t)\, B(t)\, dw(t) \tag{4.2.4a}$$

where Q_{kl} is the n vector with i:th component

$$(Q_{kl})_i = \sum_j \frac{\partial G_{ik}(x,t)}{\partial x_j} G_{jl}(x,t), \qquad (4.2.4b)$$

$G(x,t)$ is the function appearing in (4.2.1),

$A(t) = A(t,\tau_{cor})$ of (4.2.3a),

and $BB^T(t) = A(t) + A^*(t)$ of (4.2.3),

with the assumption that the limiting process $X(t,o)$ is a continuous Markov process. The convergence is uniform in a compact set of X and t.

If $X(t,o)$ is a continuous Markov process, its probability density is given by the solution of a Fokker-Planck equation, and convergence in distribution [20, page 9] is assured if the density of $x(t)$ of (4.2.4) is given by the same FP equation. That the limiting physical process $X(t,o)$ is a Markov process is a subtle point of continuous stochastic process theory which is difficult to demonstrate rigorously. However, it is a common assumption that continuous systems driven by delta correlated noise are Markov processes, and as τ_{cor} tends to zero, $R(t,\tau,\tau_{cor})$ tends to a delta function (for example, Stratonovich [21, Ch.4] makes this assumption without comment).

Proof: Having assumed that $X(t,o)$ has a FP equation, we proceed to find this FP equation by deriving the first two incremental moments (4.1.2) of $X(t,o)$.

In Section 4.1.1, we derived $E[\delta X \mid X,t]$ for a particular physical process $X(t,\tau_{cor})$. We follow this derivation below, noting how the limiting operation, τ_{cor} tends to zero, affects the assumptions and approximations made in Section 4.1.1.

We begin with equation (4.1.10) which is the last exact equation in the derivation of Section 4.1.1. The error involved in assumption A1 tends to zero as τ_{cor} tends to zero, as the limiting $y_k(u)$ has a zero mean value and has no correlation with $G_{ik}(X,t)$ for

$u \geq t$. This is due to the delta correlation of $y(t)$, for while $y_k(t)$ has an instantaneous correlation with $\dot{X}_i(t)$, it has none with $X_i(t)$.

Then, allowing the assumption A2 which involves an $O(\delta t^2)$ error, we obtain equation (4.1.18)

$$\delta X_i(t) = \sum_{j,k,l} \frac{\partial G_{ik}}{\partial X_j}(t)\, G_{jl}(t) \int_t^{t+\delta t} \int_t^u y_k(u)\, y_l(v)\, dv\, du + O(\delta t^2), \qquad (4.2.5)$$

where, unlike equation (4.1.18), the equality sign in (4.2.5) is a proper equality when the limit of τ_{cor} is taken.

Taking the condition expectation of (4.2.5) and then referring to assumption A3, the $\int_t^u \cdot\, dv$ integral of (4.2.5) is exactly evaluated as $A_{kl}(u)$ when τ_{cor} equals zero, and then allowing the assumption A4 which involves an $O(\delta t^2)$ error, we obtain the equation

$$E[\delta X_i(t) \mid X,t] = \sum_{j,k,l} \frac{\partial G_{ik}}{\partial X_j}(t)\, G_{jl}(t)\, A_{kl}(t)\, \delta t + O(\delta t^2) \qquad (4.2.6)$$

which is an exact equation (to the given error term) when τ_{cor} equals zero. Thus we have shown the existence of the limit (in vector form)

$$\underset{\delta t \downarrow o}{\text{Limit}} \left(\frac{1}{\delta t} \underset{\tau_{cor} \downarrow o}{\text{Limit}} E[\delta X \mid X,t] \right) = \sum_{k,l} Q_{kl}(X,t)\, A_{kl}(t), \qquad (4.2.7)$$

where Q_{kl} is defined in (4.2.4b).

Now, the outside limit in equation (4.2.7) is a definition of the first incremental moment of a continuous Markov process, and the inside limit implies that we are operating on the limiting process $X(t,o)$, and so we recognize the right hand side of (4.2.7) as the first incremental moment of the Markov process $X(t,o)$.

In a similar fashion, we can follow the derivation of Section 4.1.2 and find that the second order conditional increment is exactly

$$E[\,\delta X_i(t)\delta X_j(t)\,|\,X,t] = \sum_{k,l} G_{ik}(t)\,G_{jl}(t)\,[A_{kl}(t) + A^*_{kl}(t)]\,\delta t + O(\delta t^2) \qquad (4.2.8)$$

when the limit of τ_{cor} is taken. Then the following limit exists (in vector form)

$$\lim_{\delta t \downarrow o}\left(\frac{1}{\delta t}\lim_{\tau_{cor}\downarrow o} E[\,\delta X\,\delta X^T\,|\,X,t]\right) = G(X,t)\,[A(t) + A^*(t)]G^T(X,t), \qquad (4.2.9)$$

which can be interpreted as the second incremental moment of the Markov process $X(t,o)$.

As $X(t,o)$ is a continuous process, higher incremental moments are zero [21, page 62], and the first order probability density function of $X(t,o)$ is given by the FP equation associated with the incremental moments (4.2.7, 9) (c.f. equations (1.2.1, 2)). But the diffusion process $x(t)$ (4.2.4) also has these incremental moments, as

$$[G\ B]\,[G\ B]^T = G[A + A^*]\,G^T, \qquad (4.2.10)$$

and so the first order probability density function of $x(t)$ is given by the same FP equation as $X(t,o)$. This means that given the same initial conditions, the probability densities $P[x(t)]$ and $P[X(t,o)]$ are identical for $t > t_o$. Thus the process $X(t,\tau_{cor})$ converges to $x(t)$ in distribution, uniform in X and t, and the proof of Theorem 4.2.1 is complete.

Comments on Theorem 4.2.1

(1) If the noise $y(t)$ is non-stationary, then it is not clear whether the limiting operation of letting τ_{cor} tend to zero while preserving the characteristic parameters $A(t)$ and $A^*(t)$ also preserves the correct time scale of the statistical variations involved in the non-stationarity of $y(t)$. The doubt exists because the time resolution of the delta correlated noise is different

from that of the original physical noise. However, for the case where the speed of the statistical variations of $y(t)$ is slow compared with the frequency content of the physical noise, we have $A^*(t) \doteq A^T(t)$, and the time scale of the non-stationarity is not affected by the limiting operation.

(2) It is likely that if we impose the further condition that

$$\underset{\tau_{cor} \downarrow 0}{\text{l.i.m.}} \int_0^t y(s)\, ds = \int_0^t B(s)\, dw(s), \qquad (4.2.11)$$

where $B(t)$ is as used in equation (4.2.4), then the physical process $X(t,\tau_{cor})$ converges in the mean to the diffusion process $x(t)$. Convergence in the mean is a much stronger concept of convergence than convergence in distribution (see footnote in fourth page of Appendix A), and implies that the sample paths of $X(t,\tau_{cor})$ converge to those of $x(t)$. Clark [22] proves this convergence for the case where $y(t)$ is stationary and Gaussian. It is when $y(t)$ is non-Gaussian that an approach such as Clark's becomes difficult to follow, for the r.h.s. of (4.2.11) is Gaussian, and it is not clear whether the signal

$$\int_0^t y(s)\, ds$$

becomes Gaussian when the limit of τ_{cor} is taken. Doob [20, page 98] suggests that the limiting signal is Gaussian, for he states that the Brownian motion process (such as the r.h.s. of (4.2.11)) is essentially the only continuous process with independent increments (the l.h.s. of (4.2.11) is continuous and has independent increments in the limit as τ_{cor} tends to zero). However, this property is difficult to demonstrate for an arbitrary signal $y(t)$ which is only specified by its characteristic matrices, and the convergence in the mean of $X(t,\tau_{cor})$ for an arbitrary $y(t)$ is a subject for future research.

Other Forms of the Diffusion Process (4.2.4)

In Theorem 4.2.1, we have given the Ito s.d.e. for the diffusion process $x(t)$ which has the same distribution as the limiting form of the given physical process $X(t)$. It is interesting to see what other stochastic equations can describe the diffusion process $x(t)$.

By multiplying equation (A22) by the noise scaling factor $[A_{kl}(t) + A^*_{kl}(t)]$, the Ito s.d.e. (4.2.4) can be turned into a Stratonovich s.d.e. for $x(t)$ by subtracting the conversion term

$$\tfrac{1}{2} \sum_{k,l} Q_{kl}(x,t) \, [A_{kl}(t) + A^*_{kl}(t)] \, dt \qquad (4.2.12)$$

from the Ito equation. Thus the Stratonovich s.d.e. for the diffusion process $x(t)$ is

$$\bar{d}x(t) = \tfrac{1}{2} \sum_{k,l} Q_{kl}(x,t) \, [A_{kl}(t) - A^*_{kl}(t)] \, dt$$

$$+ \; G(x,t) \, B(t) \, \bar{d}w(t), \qquad (4.2.13)$$

where $\bar{d}\cdot$ represents a stochastic increment in the Stratonovich sense (see Appendix A).

As mentioned in Appendix A, the Stratonovich equation has the advantage that it can be manipulated by many of the ordinary rules of calculus. For example, we could calculate the statistics of the conditional increments of $x(t)$ by integrating equation (4.2.13) by the normal rules of calculus as in Sections 4.1.1, 2, where we must interpret $\dfrac{B(t) \, \bar{d}w(t)}{dt}$ of (4.2.13) as white noise with the correlation function

$$[A(t) + A^*(t)] \, \delta(\tau), \qquad (4.2.13')$$

$\delta(\cdot)$ being the usual (symmetrical) Dirac delta function. The characteristic matrix of this white noise is the matrix

$$\tfrac{1}{2}\,[A(t) \;+\; A^*(t)], \qquad (4.2.14)$$

which is symmetrical if the noise $y(t)$ is stationary. The first incremental moment of the Stratonovich s.d.e. (4.2.13) is found by substituting the characteristic matrix (4.2.14) for $A_{kl}(t)$ in (4.2.7) and adding on the drift term of (4.2.13). This gives

$$\begin{aligned} b(x,t) \;&=\; \overline{\sum_{k,l}}\; Q_{kl}\,\tfrac{1}{2}\,[A_{kl}(t) \;+\; A^*_{kl}(t)] \\ &\qquad +\; \tfrac{1}{2}\,\overline{\sum_{k,l}}\; Q_{kl}[A_{kl}(t) \;-\; A^*_{kl}(t)], \\ &=\; \overline{\sum_{k,l}}\; Q_{kl}\,A_{kl}(t), \end{aligned} \qquad (4.2.15)$$

which confirms that the Stratonovich s.d.e. (4.2.13) has the correct drift term (compare (4.2.15) with (4.2.7)). A similar analysis would show that it also has the correct second incremental moment.

If the noise $y(t)$ is stationary and has a symmetrical characteristic matrix A, or $A(t) = A^*(t)$ for non-stationary noise, then the Stratonovich s.d.e. (4.2.13) for the diffusion process $x(t)$ has no drift term, and the equation (4.2.13) is similar in form to the equation for the physical process (4.2.1). However, even if these conditions do not hold, we can construct a new type of stochastic equation to describe the diffusion process $x(t)$ which has no drift term, and thus is similar to equation (4.2.1). This possibility is suggested by the freedom of defining the stochastic integral in various ways as in equation (A9), but instead of varying θ in equation (A9), we keep $\theta = \frac{1}{2}$ (the Stratonovich integral) and modify the definition of white noise to give the diffusion process $x(t)$ the correct first incremental moment. We see below that this is, in fact, equivalent to varying θ and keeping the same noise.

Define a new white noise $z(t)$ to have the correlation function

$$R_z(t,\tau) \;=\; \delta_A(t,\tau)\;, \qquad (4.2.16)$$

where $\delta_A(t,\tau)$ is a modified delta function which has the properties

$$\int_{0}^{\mu} \delta_A(t,\tau)\, d\tau = A(t), \qquad (4.2.17a)$$

$$\int_{-\mu}^{0} \delta_A(t,\tau)\, d\tau = A^*(t), \qquad (4.2.17b)$$

and

$$\int_{-\mu}^{\mu} \delta_A(t,\tau)\, d\tau = A(t) + A^*(t), \qquad (4.2.17c)$$

for any arbitrarily small positive μ, where $A(t)$ and $A^*(t)$ are the characteristic matrices of the physical noise $y(t)$ of equation (4.2.1). Thus the new white noise $z(t)$ has the same characteristic matrix as the physical noise $y(t)$, and as far as the statistical properties of $X(t)$ or $x(t)$ are concerned, $z(t)$ has the same effect as a random forcing function of a differential equation as the limiting form of $y(t)$ as τ_{cor} tends to zero. This is as far as we can take the comparison, however, as $z(t)$ is Gaussian and the limiting form of $y(t)$ is non-Gaussian in general (c.f. comment (2) following Theorem 4.2.1).

The following note shows that the definition (4.2.16) is a proper, if unusual, definition of white noise.

Note on Definition of White Noise

The classical concept of white noise consists of a Gaussian signal with a continuous power spectral density which is uniform at all frequencies (the Gaussian assumption is common but not necessary - we use it here to ensure that the integral of white noise is the Wiener process). White noise then has the correlation function (5.1.3b)

$$R(t,\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} S(t,\omega)\, e^{j\omega\tau}\, d\omega = 2D(t)\, \delta(\tau), \qquad (4.2.18)$$

which is a delta function in τ, and $2D(t)$ is the intensity matrix of the noise. But this delta function need not be a symmetrical delta function, as the spectrum is defined in terms of the correla-

tion function as (5.1.3a)

$$S(t,\omega) = \int_{-\infty}^{\infty} R(t,\omega)\, e^{-j\omega\tau}\, d\tau,$$

and

$$\int_{-\infty}^{\infty} 2D(t)\, \delta(\tau)\, e^{-j\omega\tau}\, d\tau = \int_{-\infty}^{\infty} \delta_A(t,\tau)\, e^{-j\omega\tau}\, d\tau \qquad (4.2.19)$$

when $2D(t) = A(t) + A^*(t)$ and $\delta_A(t,\tau)$ is defined as the asymmetrical delta function (4.2.17). Thus a flat spectrum $S(t,\omega) = S(t)$ implies a delta correlated noise, but does not restrict the shape of the delta function. Indeed, the analysis above suggests that white noise which is used to replace physical noise when forming a diffusion model should in general have an asymmetric correlation function $\delta_A(t,\tau)$. In Section 4.3, we note that many authors have used symmetrical delta functions to characterise a white noise process, and so have not allowed their diffusion process model a wide class of physical processes.

Consider now the stochastic differential equation

$$\dot{x}(t) = G(x,t)\, z(t) \qquad (4.2.20)$$

which is to be integrated like a Stratonovich equation, bearing in mind that the white noise $z(t)$ has the property (4.2.16), which distinguishes it from the more usual white noise used by Stratonovich. We write equation (4.2.20) in the form given to facilitate comparison with the physical equation (4.2.1), but as both sides of (4.2.20) are always infinite, the equation should more properly be written as an equation involving stochastic differentials or integrals as in Appendix A.

Equation (4.2.20) can be integrated by the procedure of Section 4.1.1 even though the right hand side is always infinite, for $z(t)$ is replaced by its correlation function, which is then integrated correctly even though it has a singularity. The important integral is the one-sided integral of the correlation function of $z(t)$ in equation (4.1.19), and this is integrated exactly as $A(t)$ of (4.2.17a). Then it follows directly that the first incremental moment of the diffusion process $x(t)$ of equation (4.2.20) is equal to equation (4.2.7), which means that the new stochastic equation (4.2.20) describes the same process as the Ito equation (4.2.4) and the Stratonovich equation (4.2.13) (again one can also show the equivalence of the second incremental moments).

The new stochastic equation (4.2.20) corresponds to the generalised stochastic equation (A9) with

$$\theta = \theta(t) = 1 - \frac{A(t)}{A(t) + A^*(t)} \qquad (4.2.21)$$

for the scalar noise case, and in the vector case, θ, as well as being time varying, also varies with each component of the equation.

The purpose of introducing the three stochastic equations (4.2.4, 13, 20) for the diffusion process $x(t)$, which is statistically equivalent to the limiting form of the physically realisable process $X(t)$ of equation (4.2.1), is to trace the origin of some confusion which has arisen in the literature concerning the physical interpretation of stochastic equations.

For mathematical convenience, many authors have used stochastic equations to describe dynamic systems when deriving the control, filtering, stability, etc., of the systems. Until very recently, most authors have implied, or at least left it open for their readers to infer, that their stochastic equations are term-by-term equivalents of limiting forms of physically realisable processes. Early authors, for example Kozin and Bogdanoff [105], use Ito equations, but we see

that these equations are similar to physical equations only when the noise is additive (then $Q = 0$ in equation (4.2.4)), Later authors use Stratonovich equations, for example Leibowitz [106], Gray and Caughey [41], and Ariaratnam and Graefe [14: II, case 2], but even these equations are similar to physical equations only when the physical noise has the property $A(t) = A^*(t)$ - see equation (4.2.13).

However, in general we see that the stochastic equation which is similar to the associated physically realisable system is neither an Ito nor a Stratonovich equation, but the new equation (4.2.20). The new equation does not have the convenient properties of Ito or Stratonovich equations, and is introduced only to illustrate this point.

Much of the literature on the subject has concerned linear systems with random coefficients (where the noise is non-additive), and this example is discussed in more detail in the next section.

4.3 Applications to Linear Systems with Random Coefficients

Systems with random coefficients have received considerable attention in the literature in the last five years, since it was discovered that adding noise to a coefficient of a differential equation had a markedly different effect from adding noise as a simple additive term. This is because the coefficient noise gives a bias to the drift of the process (see equation (4.1.24)) while additive noise does not. (Leibowitz [106] was one of the first authors to demonstrate this.) Because of the analytical difficulties associated with non-linear systems, the literature has been confined to linear systems, and more recently their stability properties have been discussed, as adding noise to a coefficient can appreciably change the stability characteristics of a system (again, additive noise has little or no effect on stability).

These results are all derived in the stochastic calculus, and we are left with the problem of adapting them to physical situations. This problem has not been satisfactorily treated in the literature, although Gray and Caughey [41] and Kulman [103] (of the Stratonovich school) do make significant advances in this direction. Few authors suggest that physical equations should have different (bias) terms in them compared with the equivalent stochastic equations (Clark [22] is the best of the exceptions) and we are often left with the implication that physically realisable equations can be obtained by simply replacing the white noise in the stochastic equations by band-limited noise (and vice versa).

For the case of linear systems with random coefficients, this implication is always wrong when applied to Ito stochastic equations, for the noise in such systems is non-additive, and a bias term must always be used when converting from Ito to physical equations (see the QA term in (4.2.4) compared with (4.2.1)). A significant advantage of the Stratonovich stochastic calculus is that this implication is true when $A(t) = A^*(t)$ (compare equation (4.2.13) with (4.2.1)), and so the results derived in the Stratonovich calculus (e.g. [106], [14: II, case 2], and Stratonovich's many results) can be applied without modification to physical situations

provided the physical noise has the property $A(t) = A^*(t)$. Clark's main advance was in pointing out that many physical noise sources do not have this property, and in these cases, Stratonovich equations cannot be applied to physical situations without modification. He considers stationary noises where the condition $A(t) = A^*(t)$ is reduced to the symmetry of the A matrix, $A = A^T$. In Sections 4.1, 2, we extend his results under more general conditions (non-Gaussian noises allowed) but weaker implications (convergence in distribution only) to non-stationary noise sources where Clark's condition $A = A^T$ becomes $A(t) = A^*(t)$.

The conclusion is that, when relating stochastic equations to physical equations, or vice versa, we must in general use the results of Sections 4.1, 2, rather than the more specific results of previous authors. To further justify this, let us look again at our results, and in particular, at the limiting operation in Section 4.2.

We are given a particular physical process (4.2.1) and we wish to choose a diffusion process which retains the essential statistical characteristics of the physical process. To do this, in Section 4.2 we impose a limiting operation on the physical noise $y(t)$ such that the physical process $X(t)$ retains, to a good approximation, the expected value of its first and second order conditional increments. We argued that this was equivalent to retaining the first order transition probability density of $X(t)$ for transitions over a time interval greater than the correlation time of the noise. Looking at the limiting operation in the other direction, as τ_{cor} is increased, we noted that this statistical property was retained to a good approximation as long as the noise memory time was significantly less than the system memory time.

What we are doing in the limiting operation is forming an equivalence class of physical processes, each member of which approximately shares certain properties with the original physical process (4.2.1). To what extent the members of the equivalence class must retain the properties of $X(t)$ depends on the use we make of the members of the class (in particular, we are interested in the limiting member, $\tau_{cor} = 0$). In this chapter, we show that if we wish to maintain the first order distributional properties of $X(t)$,

then the members of the limiting class of noises must retain the properties $A(t)$ and $A(t) + A^*(t)$.

In connection with the latest literature on linear systems with random coefficients (e.g. [26], [41], [14]), our main point is this: these authors use Stratonovich stochastic equations with a symmetrically correlated white noise, and in the sense of this chapter, these authors' diffusion processes only represent the distributional properties of physical processes possessing the property $A(t) = A^*(t)$. Their results would apply without modification to all physical processes if our equivalence class of physical processes needed only to retain the property $A(t) + A^*(t)$ of the physical noises (we see Section 4.1.2 where we see that this property gives us the conditional expectation of the second order increment of $X(t)$ - a property analogous to the dispersion of the diffusion process $x(t)$). But we see in Section 4.1.1 that in order to maintain the distributional properties of $X(t)$, we must preserve $A(t)$ as well as $A(t) + A^*(t)$ of the physical noise (this then gives the processes the correct drift as well as the correct dispersion).

This repeats the conclusion at the end of the last section: if we are changing a physical process (4.2.1) into an "equivalent" diffusion process by simply replacing the physical noise $y(t)$ by white noise, then we must use a special type of white noise $z(t)$ which has the asymmetrical delta correlation function $\delta_A(t,\tau)$ - and obtain the diffusion process (4.2.20) which we pointed out could be written as the more conventional and convenient Ito equation (4.2.4) and Stratonovich equation (4.2.13). Previous authors who use the Stratonovich equation with symmetrically correlated white noise do not allow our generality, as will be illustrated by the example below. The distributional characteristics of a linear system with random coefficients will be specified by finding the FP equation of the limiting form of the given physical process.

Fokker-Planck Equation for a Linear System with Random Coefficients

We consider the physical process studied by Gray [15] described by the ordinary differential equation

$$\frac{d^{(n)}\,X(t)}{dt^n} + \sum_{i}^{n} [b_i + y_i(t)] \frac{d^{(i-1)}\,X(t)}{dt^{i-1}} = y_{n+1}(t) + c(t), \tag{4.3.1}$$

where $y(t)$ is an $(n+1)$ vector physical noise which Gray specifies by its limiting correlation function

$$E[y(t)\, y^T(t-\tau)] = 2D\,\delta(\tau) \tag{4.3.2}$$

involving the symmetrical delta function. His noise is stationary, and in our notation the right hand side of (4.3.2) equals $[A + A^T]\,\delta(\tau)$. However, our argument is that we should specify the noise by the limiting correlation function $\delta_A(\tau)$ which involves knowing A as opposed to $A + A^T$. ** The values b_i and function $c(t)$ are known constants and time varying functions.

Gray's example (4.3.1) is a special case of the physical process (4.1.3), and it is convenient to change to the state variable notation of (4.1.3). We put

$$\frac{d^{(i-1)}\,X(t)}{dt^{i-1}} = X_i(t), \qquad i = 1, n, \tag{4.3.3}$$

and find the coefficients $g(X,t)$ and $G(X,t)$ are

$$g_i(X,t) = X_{i+1}(t), \qquad i = 1, n-1,$$

$$g_n(X,t) = -\sum_{i}^{n} b_i\, X_i(t) + c(t),$$

** As Gray considers stationary noises in his example, we do also. The following analysis holds for non-stationary noises by replacing A^T by $A^*(t)$, which involves knowing $A^*(t)$ as well as $A(t)$.

and $$G_{ij}(X,t) = 0 \quad , \qquad i = 1, n-1, \text{ for all } j,$$

$$G_{nj}(X,t) = -X_j, \qquad j = 1, n$$

$$= 1 \quad , \qquad j = n+1. \tag{4.3.4}$$

We derive the FP equation which approximately gives the probability density of $X(t)$ by finding the diffusion process which is the limiting form of the physical process in the sense of Section 4.2. To do this we must find the bias term appearing in the Ito equation (4.2.4) or the Stratonovich equation (4.2.13).

The quantity $(Q_{kl})_i$ of (4.2.4b) is zero except when $i = k = n$ when we have

$$(Q_{nl})_n = X_l(t) , \qquad l = 1, n$$

$$= -1 \quad , \qquad l = n+1. \tag{4.3.5}$$

Then from (4.2.4), the bias term appropriate to the Ito form of the equivalent diffusion process is

$$\left[\sum_{l}^{n} X_l(t)\, A_{nl} - A_{n,n+1} \right] dt, \tag{4.3.6}$$

and from (4.2.13), the bias term appropriate to the Stratonovich form is

$$\tfrac{1}{2} \left[\sum_{l}^{n} X_l (A_{nl} - A_{nl}^T) - (A_{n,n+1} - A_{n,n+1}^T) \right] dt. \tag{4.3.7}$$

We can then write down the diffusion process equivalent to the physical process (4.3.4) as, in Ito form,

$$dx_i(t) = g_i(x,t)\, dt + \sum_{j}^{n+1} G_{ij}(x,t)(B\,dw(t))_j, \quad i = 1, n-1,$$

and

$$dx_n(t) = \left[g_n(x,t) + \sum_{l}^{n} x_l\, A_{nl} - A_{n,n+1} \right] dt \tag{4.3.8}$$

$$+ \sum_{j}^{n+1} G_{nj}(x,t)\ (B\ dw(t))_j,$$

where B is arbitrarily chosen to satisfy $BB^T = A + A^T$.

Similarly the Stratonovich form is

$$\bar{d}x_i(t) = g_i(x,t)dt + \sum_{j}^{n+1} G_{ij}(x,t)\ (B\ \bar{d}w(t))_j, \quad i = 1, n-1,$$

and

$$\bar{d}x_n(t) = \left[g_i(x,t) + \sum_{1}^{n} \tfrac{1}{2}\ x_1(A_{n1} - A_{n1}^T) - \tfrac{1}{2} A_{n,n+1} + \tfrac{1}{2} A_{n,n+1}^T\right] dt$$

$$+ \sum_{j}^{n+1} G_{nj}(x,t)\ (B\ \bar{d}w(t))_j. \qquad (4.3.9)$$

In the notation of Gray, this Stratonovich form becomes

$$\bar{d}\left[\frac{d^{(n-1)}\ x(t)}{dt^{n-1}}\right] + \sum_{i}^{n} \left[(b_i + \tfrac{1}{2} A_{in} - \tfrac{1}{2} A_{in}^T)dt + (B\bar{d}w(t))_i\right]\left[\frac{d^{(i-1)}\ x(t)}{dt^{i-1}}\right]$$

$$= (B\bar{d}w(t))_{n+1} + \left[\tfrac{1}{2} A_{n,n+1}^T - \tfrac{1}{2} A_{n,n+1} + c(t)\right]dt. \qquad (4.3.10)$$

Upon symbolic division of (4.3.10) by dt, it is interesting to compare the Stratonovich form of the equivalent diffusion system (4.3.10) to the original physical system (4.3.1). It is noted that if the characteristic matrix A of the noise vector $y(t)$ is symmetrical, then the terms involving A in (4.3.10) drop out, and the Stratonovich equation (4.3.10) can be obtained directly from physical process (4.3.1) by replacing the physical noise component $y_i(t)$ by the symmetrically delta correlated white noise $\frac{d}{dt}(B\ \bar{d}w(t))_i$. Note that the information of $A + A^T$ is retained by B in the form of BB^T.

This simple term-by-term correspondence shows the convenience of the Stratonovich calculus if the physical noise $y(t)$ is symmetrical. If the matrix A is not symmetrical, the terms

involving A in (4.3.10) must remain, or alternatively, as suggested in the previous section, the concept of white noise $\frac{dw(t)}{dt}$ could be altered to a noise with an asymmetric delta correlation function.

To write down the FP equation for the equivalent diffusion process, we use the Ito form (4.3.8) where the process' incremental moments are in explicit form. The drift term, or the first incremental moment, is straightforward, and the dispersion term, or the second incremental moment $a(x,t)$ is given by the $(n \times n)$ matrix

$$GB(GB)^T = GBB^TG^T = G(A + A^T)G^T = 2GDG^T \qquad (4.3.11)$$

which only has a lower right hand (n,n) element of

$$2\sum_{i,j}^{n} A_{ij}x_ix_j \sim 2\sum_{i}^{n} (A_{i,n+1} + A_{i,n+1}^T)x_i + 2A_{n+1,n+1}. \qquad (4.3.12)$$

Thus the FP equation of the equivalent diffusion process is given by

$$\frac{\partial P}{\partial t} = -\sum_{i}^{n-1} \frac{\partial}{\partial x_i}[x_{i+1}P] - \frac{\partial}{\partial x_n}\Big[\Big[-\sum_{i}^{n} b_ix_i + c(t) + \sum_{1}^{n} x_1A_{n1} - A_{n,n+1}$$

$$+ \frac{\partial^2}{\partial x_n^2}\Big[\Big[\sum_{i,j}^{n} A_{ij}x_ix_j - \sum_{i}^{n} (A_{i,n+1} + A_{i,n+1}^T)x_i + A_{n+1,n+1}\Big]P\Big]. \qquad (4.3.13)$$

The following simple example will show how this FP equation differs from that of Gray if the characteristic matrix A of the physical noise $y(t)$ is asymmetric.

Example

Consider the first order ordinary differential equation

$$\frac{dX(t)}{dt} = - X(t) - y_1(t)X(t) + y_2(t), \qquad (4.3.14)$$

which is a special case of Gray's system(4.3.3) with $n = 1$, $b_1 = 1$ and $c(t) = 0$. Let the two-dimensional physical noise process have the asymmetrical characteristic matrix (5.1.15, 16) derived as an example in the next chapter:

$$A = \begin{bmatrix} \frac{1}{2} & \frac{10}{11}\alpha \\ \frac{1}{11}\alpha & \frac{1}{2}(\alpha^2 + 1) \end{bmatrix} \qquad (4.3.15)$$

We see $g(x,t) = - X(t)$,

and $G(x,t) = [- X(t) \quad 1]$. $\qquad (4.3.16)$

From (4.3.8) we have the Ito form of the equivalent diffusion system as

$$dx(t) = [- x(t) + \tfrac{1}{2} x(t) - \frac{10}{11}\alpha]\, dt + G(x,t)\, B\, dw(t) \qquad (4.3.17)$$

where B is arbitrarily chosen to satisfy $BB^T = A + A^T$ of (4.2.15).

The FP equation of (4.3.17) is

$$\frac{\partial P}{\partial t} = - \frac{\partial}{\partial x}[(- \tfrac{1}{2} x - \frac{10}{11}\alpha)\, P] + \frac{\partial^2}{\partial x^2}[[\tfrac{1}{2} x^2 - \alpha x + \tfrac{1}{2}(\alpha^2 + 1)]P]. \qquad (4.3.18)$$

Now, treating this example in the fashion of Gray and Caughey and others, we only characterize the physical noise $y(t)$ by its intensity matrix $2D$ where

$$2D = A + A^T = 2 \begin{bmatrix} \frac{1}{2} & \frac{1}{2}\alpha \\ \frac{1}{2}\alpha & \frac{1}{2}(\alpha^2 + 1) \end{bmatrix}. \qquad (4.3.19)$$

This gives the FP equation (see Gray's equation (4))

$$\frac{\partial P}{\partial t} = -\frac{\partial}{\partial x}\left[(-\tfrac{1}{2}x - \tfrac{1}{2}\alpha)P\right]$$

$$+\frac{\partial^2}{\partial x^2}\left[(\tfrac{1}{2}x^2 - \alpha x + \tfrac{1}{2}(\alpha^2 + 1))P\right]. \qquad (4.3.20)$$

This equation agrees with our FP equation (4.3.18) only when $\alpha = 0$ which is when the characteristic matrix of the noise source is symmetrical (two independent noises in this case). This illustrates why the intensity matrix (4.3.19) does not sufficiently characterise the stationary physical noise vector $y(t)$ when the characteristic matrix A of $y(t)$ is asymmetrical. In the non-stationary noise case, the asymmetry condition $A \neq A^T$ changes to the condition $A(t) \neq A^*(t)$, and the same comments as to Gray's insufficient characterisation of the noise apply (i.e. we cannot fully characterise $y(t)$ by its intensity matrix $A(t) + A^*(t)$).

4.4 The Simulation of Diffusion Processes

In Sections 4.1 and 4.2 we have discussed the statistical relation between physical processes and diffusion processes by finding a diffusion process which statistically approximates to a given physical process (Section 4.1), or is statistically equivalent to the limiting form of a given physical process (Section 4.2). In some cases, we are primarily interested in the converse problem: given a particular diffusion process, how do we simulate it on an analogue computer? In other words, how do we choose a physical process (which can be represented to a arbitrary accuracy on a computer) which adequately represents the statistical behaviour of the diffusion process? The answer to this problem is implied in Clark's results or the results of this chapter, but there is no longer a unique solution for there is some freedom in the choice of physical noise source to use in the simulation.

Consider the diffusion process described by the Ito s.d.e.

$$dx(t) = f(x,t)\,dt + F(x,t)dw(t)\,, \qquad (4.4.1)$$

where $dw(t)$ is the stochastic increment of the m-vector unit parameter Wiener process. The components $dw_k(t)$, $k = 1, m$, are independent of each other, but the coefficient matrix $F(x,t)$ introduces cross-coupling between the components $dw_k(t)$ and $x_i(t)$, as well as introducing non-stationarity via the t parameter.

In choosing a physical process $X(t)$ to simulate the given diffusion process (4.4.1), we have some freedom at our disposal in the choice of physical noise source. The noise source must have sufficient degrees of independence to simulate the m degrees of independence of the Wiener process, but it can be a cross-correlated noise process provided the cross-coupling is sufficiently similar in form to that introduced by $F(x,t)$ of (4.4.1) so that the equality (4.4.5) can be satisfied. As $\dot{w}(t)$ is a stationary white noise, it is simplest to choose a stationary physical noise $y(t)$ for the simulation, provided the non-stationarity implied in $F(x,t)$

can be represented separately on the computer. A non-stationary noise can be used in the simulation, but this involves a matching operation (see below) with theoretical and practical difficulties, and so will not be considered.

Choice of Physical Process to Use in the Simulation

Consider the physical process $X(t)$ described by the ordinary differential equation

$$\dot{X}(t) = g(X,t) + G(X,t)\, y(t), \tag{4.4.2}$$

where $y(t)$ is an available m-vector stationary physical noise process of characteristic matrix A, which has a sufficiently high bandwidth that the inequality of Section 4.1, $\tau_{cor} \ll \tau_{rel}$ is satisfied by a reasonable margin.

From the argument of Section 4.1.2, the physical process (4.4.2) has the expected second order conditional increment which is approximately*

$$E[\delta X\, \delta X^T \,|\, X,t,\delta t] \doteq G(A + A^T)\, G^T\, \delta t, \tag{4.4.3}$$

whereas the diffusion process (4.4.1) has the second increment

$$E[\delta x\, \delta x^T \,|\, x,t,\delta t] = F\, F^T \delta t + o(\delta t). \tag{4.4.4}$$

Clearly, then, the first step in choosing an appropriate physical process to simulate is to choose the function G so that

$$G(A + A^T)\, G^T = F\, F^T, \tag{4.4.5}$$

---

* The term $g(X,t)$ of (4.4.2) is non-random in the sense that it does not depend directly on $y(t)$, in which case it contributes a negligible amount to the second-order conditional increment.

where the functions are evaluated at an arbitrary X (or x) and t. The equality (4.4.5) then insures that the processes $X(t)$ and $x(t)$ have approximately the same dispersion.

A simple procedure to achieve the equality (4.4.5) is to let G have the form

$$G = FC, \tag{4.4.6a}$$

where C is an $m \times m$ constant matrix. Then (4.4.5) becomes

$$FC(A + A^T)\, C^T F^T = F F^T,$$

which is equivalent to

$$C(A + A^T)\, C^T = I, \tag{4.4.6b}$$

where I is the $m \times m$ unit matrix. Since the matrix $(A + A^T)$ is symmetrical, it is congruent to I provided $(A + A^T)$ is of rank m, and a matrix C satisfying (4.4.6b) can be found.* If $(A + A^T)$ has rank lower than m, the noise source $y(t)$ is not sufficiently independent to represent the m degrees of independence of $\dot{w}(t)$. This can be remedied by making the components of $y(t)$ more independent of each other, or by increasing the dimension of $y(t)$. If $y(t)$ were non-stationary, the condition (4.4.6b) becomes $C(A + A^*)\, C^T = I$, and the matrix $C = C(t)$ cannot always be found.

Having chosen $G(X,t)$ to match the second order conditional increments of the diffusion process and the physical process, we must consider the first order conditional increments, and choose $g(X,t)$ accordingly. From the analysis of Section 4.1.1, the physical

---

* The square matrices A and B are <u>congruent</u> if and only if there exists a non-singular matrix C such that

$$C\, A\, C^T = B.$$

If A and B are both symmetrical and of the same rank, they are congruent and a matrix C can always be found.

process (4.4.2) has the expected conditional increment which is approximately

$$E[\delta X_i \mid X,t,\delta t] \doteq [g_i + \sum_{j,k,l} \frac{\partial G_{ik}}{\partial X_j} G_{jl} A_{kl}]\delta t, \qquad (4.4.7)$$

$$= [g_i + \sum_{k,l} (Q_{kl})_i A_{kl}]\delta t,$$

whereas the diffusion process (4.4.1) has the increment

$$E[\delta x_i \mid x,t,\delta t] = f_i \delta t + o(\delta t).$$

Thus the physical process (4.4.2) will have approximately the same drift as the diffusion process (4.4.1) if we put

$$g(X,t) = f(X,t) - \sum_{k,l} Q_{kl} A_{kl}, \qquad (4.4.8)$$

and the physical process which has the approximate incremental properties of the diffusion process (4.4.1) is then written as

$$\dot{X}(t) = f(X,t) - \sum_{k,l} Q_{kl}(X,t) A_{kl} + G(X,t)\, y(t). \qquad (4.4.9)$$

To sum up, we have chosen a physical process (4.4.9) which, in the fashion of Section 4.1, shares approximately the same statistical properties as the diffusion process (4.4.1). Bearing in mind the assumption on the upper frequency of the physical noise $y(t)$ and the related approximations of the analysis of Section 4.1, we can say that a computer realisation of the physical process (4.4.9) constitutes a simulation of the diffusion process (4.4.1).

The steps in choosing the physical process (4.4.9) are as follows:

(1) Choose a convenient noise source $y(t)$ which

(a) has m degrees of independence ($(A + A^T)$ has rank m);

(b) has a sufficiently high upper frequency so that $\tau_{cor} \ll \tau_{rel}$;

and (c) is suitable for representing accurately on the computer available.

(2) Evaluate the characteristic matrix A of the noise source, and choose $G(X,t)$ by the method of equation (4.4.6) so that the physical process (4.4.9) has the correct dispersion or variance. This is essentially an operation which gives the noise $y(t)$ the proper scaling factor C.

(3) Choose $g(X,t)$ according to equation (4.4.8) so that the physical process (4.4.9) has the correct drift.

4.5 The Simulation of Physical Processes

Consider the problem of simulating the physical process

$$\dot{X}(t) = g(X,t) + G(X,t)\, y(t), \qquad (4.5.1)$$

where $y(t)$ is a stationary physical noise vector of dimension m and characteristic matrix A. This problem is trivial if the noise $y(t)$ can be exactly reproduced on the computer, provided of course that the other terms do not produce realizability or stability problems. If $y(t)$ cannot be exactly reproduced, then an alternative noise $\underline{y}(t)$ must be chosen, which can be reproduced on the computer. The new noise $\underline{y}(t)$ must have the same number of degrees of independence as $y(t)$, and should have approximately the same frequency content.

The choice of a physical process $\underline{X}(t)$ involving $\underline{y}(t)$ which has approximately the same statistical behaviour as the process (4.5.1) can proceed by the method of Section 4.1:

(1) we scale the noise $\underline{y}(t)$ correctly by matching the approximate expressions for the second order conditional increments;

(2) we then choose the correct drift term for the new process by matching the approximate expressions for the first order conditional increments.

To do this, we only need the characteristic matrix $\underline{A}$ of the new noise $\underline{y}(t)$.

If you like, the choice of $\underline{X}(t)$ can be visualized in two parts:

(a) by the method of Section 4.4, we construct a physical process $\underline{X}(t)$ suitable for representation on the computer which is statistically "equivalent" to the diffusion process found in (b) below;

(b) By the method of Sections 4.1 or 4.2 we find the diffusion process which is statistically "equivalent" to the given physical process $X(t)$, equation (4.5.1).

Each of these operations involves choosing parameters to match second increments and then first increments, and the end result is the same.

Clark [22, Ch. 3] discusses the topic of this section in more detail and points out that the choice of $\underline{X}(t)$ is simplified if we can choose a noise $\underline{y}(t)$ whose characteristic matrix $\underline{A}$ is congruent to A, the characteristic matrix of $y(t)$. In this case, the ordinary differential equation governing the process $\underline{X}(t)$ is simply

$$\dot{\underline{X}}(t) = g(\underline{X},t) + G(\underline{X},t)\, C\, \underline{y}(t), \tag{4.5.2}$$

where C is the $m \times m$ constant noise scaling matrix satisfying

$$A = C\, \underline{A}\, C^T, \tag{4.5.3}$$

and hence satisfying also

$$A + A^T = C(\underline{A} + \underline{A}^T)\, C^T \tag{4.5.4}$$

By the method of Section 4.1, we can show that the equality (4.5.4) ensures that the processes $X(t)$ and $\underline{X}(t)$ have approximately the same expected second order conditional increment, and the equality (4.5.3) ensures that the contributions of $Gy(t)$ and $GC\underline{y}(t)$ to the first order increments are approximately the same, as $C\underline{y}(t)$ has the characteristic matrix $C\, \underline{A}\, C^T$. The matrix C can be considered as a noise scaling factor, and $C\underline{y}(t)$ can be considered as a scaled version of the noise $y(t)$.

If A is a symmetric matrix of rank m, then any noise $\underline{y}(t)$ which also has a symmetric characteristic matrix $\underline{A}$ of rank m allows a matrix C to be found. Indeed, if we can choose $\underline{y}(t)$ according to a non-singular transformation of the form

$$\underline{y}(t) = C^{-1}\, y(t) \tag{4.5.5}$$

then equations (4.5.1) and (4.5.2) are identical and the simulation is exact. This, of course, is not much different from using $y(t)$

itself on the computer, and if this is not possible then it is not likely that a noise of the form (4.5.5) could be represented on the computer.

If A is not symmetric, then it may not be possible to find a suitable noise source which has a characteristic matrix $\underline{A}$ congruent to A. For a computer noise $\underline{y}(t)$ with an arbitrary characteristic matrix $\underline{A}$, it is still convenient to choose the computer noise term as $G(X,t)\, C\underline{y}(t)$, for a C can always be found to ensure that the equality (4.5.4) holds, which scales the noise term properly. The noise term $GC\underline{y}$ contributes

$$\sum_{k,l} Q_{kl} (C\, \underline{A}\, C^T)_{kl}$$

to the drift of the process $\underline{X}(t)$, and the noise term Gy contributes

$$\sum_{k,l} Q_{kl}\, A_{kl}$$

to the drift of the process X(t), where Q_{kl} is used as before, e.g. in (4.4.7). Thus the physical process $\underline{X}(t)$ which has the same drift as the process X(t) is defined by the ordinary differential equation

$$\dot{\underline{X}}(t) = G(\underline{X},t) + \sum_{k,l} Q_{kl}(\underline{X},t)(A - C\underline{A}C^T)_{kl} + G(\underline{X},t)C\underline{y}(t). \quad (4.5.6)$$

Thus to simulate the physical process X(t) involving the noise y(t) on a computer, we replace the noise y(t) by a noise $C\,\underline{y}(t)$ which has as many degrees of independence as y(t) and is suitable for representing on the computer. The equation governing the system $\underline{X}(t)$ we simulate depends on the relation between the characteristic matrices A and $C\,\underline{A}\,C^T$ of the noises y(t) and $C\,\underline{y}(t)$. If these are equal, i.e. condition (4.5.3) holds, the equation (4.5.2) is programmed on to the computer, which is simply the original system equation (4.5.1) with the new noise term. If these are not equal, then a correction term must be added to correct

the drift of the process, and the equation (4.5.6) is the resultant equation to program on to the computer.

Note that we have only specified the noises $y(t)$ and $C\,\underline{y}(t)$ by their characteristic matrices, and find by the analysis of Sections 4.1, 4.2 that the processes $X(t)$ and $\underline{X}(t)$ have approximately the same probability density functions. If we are interested in a more detailed statistical comparison between $X(t)$ and $\underline{X}(t)$, then we should choose $\underline{y}(t)$ so that $y(t)$ and $C\,\underline{y}(t)$ have approximately the same correlation functions or spectral densities, for then the processes $X(t)$ and $\underline{X}(t)$ will have approximately the same spectral properties as well as amplitude density properties.

Again, as in Section 4.4, we have not discussed non-stationary processes as the matching of the parameters in (4.5.4) is not always possible, and the generation of non-stationary noises may not be convenient unless the non-stationarity has a very simple structure. An example would be a noise $y(t)$ with a time-varying gain, for then the non-stationarity could be factored out and, in the simulation, a stationary noise $\underline{y}(t)$ could be used with a time-varying scaling factor $C(t)$.

CHAPTER 5

ANALOGUE SIMULATION

Following the discussion in Section 4.4, this chapter will discuss the simulation of the diffusion process (4.4.1) on an analogue computer. In Section 4.5, it was stated that the problem of simulating a physical process could be rephrased as a problem of simulating a diffusion process by considering the diffusion process equivalent to the given physical process.

To simplify our discussion, we assume that the analogue computer at our disposal is an ideal one in the sense that its elements perform the various functions of summing, integrating and function generating exactly for frequencies up to a given upper frequency. We then assume that the physical noise $y(t)$ which we use in the simulation has no frequency components beyond this upper frequency so that the physical system, (4.4.9), can be represented exactly on the analogue computer. If this is not the case then a more complicated analysis would have to be applied using the true transfer functions of the analogue computer components.

If the analogue computer components were ideal over an infinite frequency range and an exact white noise $\dot{w}(t)$ could be generated, then the diffusion system (4.4.1) could be simulated directly when put in the Stratonovich s.d.e. form. Some authors in the past have implicitly assumed that such a system could be built (e.g. [49], p.352, Figure 1), but we have seen this is not possible from the discussion of Section 2.2.

With the assumption that the system (4.4.9) can be simulated exactly, the only point to be discussed is the choice of a suitable noise source $y(t)$, and the determination of its characteristic matrices $A(t)$ and $A^*(t)$. This chapter will discuss the choice of $y(t)$, and then give some computed examples which illustrate how the characteristic matrices and the bias term QA must be taken into account in the simulation of diffusion processes.

5.1 Some Useful Noise Sources and their Characteristic Matrices

5.1.1 Noises Generated by Linear Shaping Filters

In general, physically realizable noise sources have spectral densities which are approximately rational functions of frequency. By factoring these functions, linear multivariable filters can be derived which transform a noise vector with a given spectral density matrix into another noise vector with a desired spectral density matrix [81]. Generating noise sources with arbitrary spectral properties is conveniently done on an analogue computer by this method, particularly as a number of independent noise components can be obtained by non-linear operations on a single Gaussian noise source [82].

As an example of this sort of noise generator, consider the configuration of Figure 5.1.1, where $z(t)$ is an m-vector physical noise source with a known correlation function $R_{zz}(t,\tau)$, and $y(t)$ is the desired noise vector obtained from the output of the linear $m \times m$ shaping filter of impulse response $h(t,u)$. The following matrix relations hold for real stochastic processes:

Convolution Relation

$$y(t) = \int_{o}^{\infty} h(t,u)z(t-u)du \qquad (5.1.1)$$

Spectral Relation

$$S_{yy}(t,\omega) = H(t,j\omega)S_{zz}(t,\omega)H^{T}(t,-j\omega) \qquad (5.1.2)$$

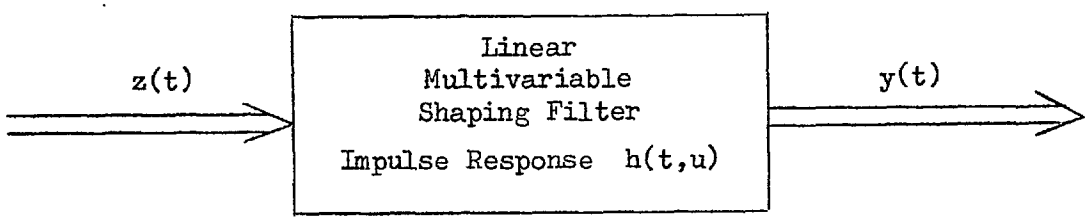
Fourier Transform Relations*

$$S..(t,\omega) = \int_{-\infty}^{\infty} R..(t,\tau)e^{-j\omega\tau}d\tau \qquad (5.1.3a)$$

$$R..(t,\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} S..(t,\omega)e^{j\omega\tau}d\omega. \qquad (5.1.3b)$$

---

* This Fourier transform relation is that suggested as a convention by Fuller [86], although our definition of $R(t,\tau)$ differs from Fuller's in the reversal of the time shift parameter τ. Our choice of $R(t,\tau)$ follows Clark [22] and is the transpose of Fuller's definition when the noise concerned is stationary.



Autocorrelation

$R_{zz}(t,\tau)$ $R_{yy}(t,\tau)$

Spectrum

$S_{zz}(t,\omega)$ $H(t, j\omega)$ $S_{yy}(t,j\omega)$

Figure 5.1.1 Generation of an Arbitrary Non-Stationary Noise Source

To obtain the characteristic matrix of $y(t)$, we need an expression for $R_{yy}(t,\tau)$. Postmultiplying (5.1.1) by $y^T(t-\tau)$ and ensemble averaging, we have

$$\begin{aligned} R_{yy}(t,\tau) &= E[y(t)\, y^T(t-\tau)] \\ &= E\Big[\int_0^\infty h(t,u)\, z(t-u)\, y^T(t-\tau)\, du\Big] \\ &= \int_0^\infty h(t,u)\, R_{zy}(t-u,\, \tau-u)\, du. \end{aligned} \tag{5.1.4}$$

Similarly,

$$\begin{aligned} R_{zy}(t-u,\, \tau-u) &= E[z(t-u)\, y^T(t-\tau)] \\ &= E\Big[\int_0^\infty z(t-u)\, z^T(t-\tau-v)\, h^T(t-\tau,v)dv\Big] \\ &= \int_0^\infty R_{zz}(t-u,\, \tau-u+v)\, h^T(t-\tau,v)\, dv. \end{aligned} \tag{5.1.5}$$

Substituting (5.1.5) into (5.1.4) we obtain

$$R_{yy}(t,\tau) = \int_0^\infty h(t,u) \int_0^\infty R_{zz}(t-u,\ \tau-u+v)\ h^T(t-\tau,\ v)\ dv\ du. \qquad (5.1.6)$$

The characteristic matrix $A(t)$ of $y(t)$ is then

$$A(t) = \int_0^\infty R_{yy}(t,\tau)\ d\tau$$

$$= \int_0^\infty \int_0^\infty h(t,u) \int_0^\infty R_{zz}(t-u,\ \tau-u+v)\ h^T(t-\tau,\ v)\ dv\ du\ d\tau, \qquad (5.1.7)$$

and the supplementary characteristic matrix is

$$A^*(t) = \int_{-\infty}^0 R_{yy}(t,\ \tau)\ d\tau$$

$$= \int_{-\infty}^0 \int_0^\infty h(t,u) \int_0^\infty R_{zz}(t-u,\ \tau-u+v)\ h^T(t-\tau,v)\ dv\ du\ d\tau. \qquad (5.1.8)$$

Suppose that the input noise $z(t)$ is broad-band stationary noise with an upper frequency much higher than the pass band of the filter $h(t,u)$. Then as the noise $z(t)$ is additive, we can replace $z(t)$ by white noise with the correlation function**

$$R_{zz}(t,\tau) = 2\ D_z(t)\ \delta(\tau), \qquad (5.1.9)$$

where $2\ D_z(t)$ is the intensity matrix of the noise $z(t)$.

Then as the integral $\int_0^\infty . \ d\tau$ of (5.1.7) only has a value when

---

** Remembering the discussion on the definition of white noise in Section 4.2, it is immaterial here what type of delta function is used in (5.1.9), as the noise $z(t)$ is additive.

$\tau = u - v$ is positive, that is when $u \geq v$, we have

$$A(t) = \int_0^\infty h(t,u)\ 2D_z(t-u) \int_0^u h^T(t-u+v\ ,\ v)\ dv\ du, \tag{5.1.10}$$

and similarly

$$A^*(t) = \int_0^\infty h(t,u)\ 2D_z(t-u) \int_u^\infty h^T(t-u+v\ ,\ v)\ dv\ du. \tag{5.1.11}$$

It is convenient to check the values of $A(t)$ and $A^*(t)$ so obtained by the zero frequency spectral relation

$$\begin{aligned} A(t) + A^*(t) &= \int_{-\infty}^{\infty} R_{yy}(t,\tau)\ d\tau, \\ &= S_{yy}(t,o), \\ &= H(t,o)\ S_{zz}(t,o)\ H^T(t,o), \\ &= 2\ D_y(t). \end{aligned} \tag{5.1.12}$$

For the case where the input noise $z(t)$ is white, equation (5.1.9), this relation becomes

$$A(t) + A^*(t) = 2\ H(t,o)\ D_z(t)\ H^T(t,o). \tag{5.1.13}$$

Example of a Noise Source with an Asymmetrical Characteristic Matrix

In the light of the results of Chapter 4, it is interesting to see that a simple noise source can have a characteristic matrix $A(t)$ which is not symmetrical.

Consider a two dimensional noise source $y(t)$ in which the component $y_1(t)$ is generated by a broad-band noise $z_1(t)$ passing

through a linear first order filter of time constant 10, and the component $y_2(t)$ is generated by a second broad band noise $z_2(t)$, independent of $z_1(t)$, passing through a linear first order filter of time constant 1. Such a noise source $y(t)$ has a diagonal, and hence symmetrical, characteristic matrix.

Now consider the case when a certain factor α of the channel 1 input noise $z_1(t)$ leaks over into the input of the channel 2 filter, as shown in Figure 5.1.2.

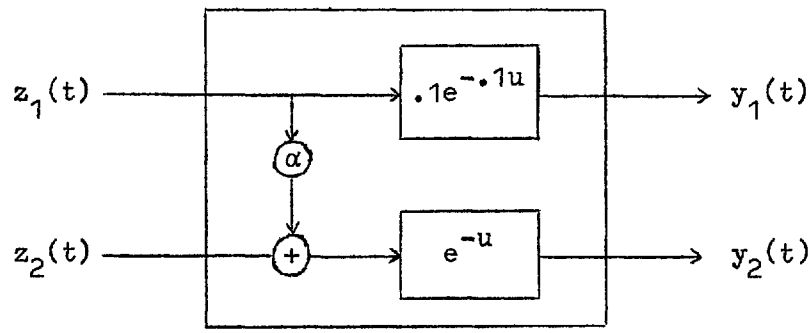


Figure 5.1.2 An Asymmetrical Noise Source

The impulse response of the shaping filter is now

$$h(t,u) = \begin{bmatrix} .1e^{-.1u} & 0 \\ \alpha e^{-u} & e^{-u} \end{bmatrix}, \quad u \geq 0, \qquad (5.1.14)$$

and we assume that the broad band input noise is normalised so that

$$R_{zz}(t,\tau) \doteq \delta(\tau)\, I,$$

or

$$D_z(t) = D_z = \tfrac{1}{2} I,$$

where I is the unit matrix.

Then by evaluating the integrals (5.1.10, 11) we find

$$A(t) = A = \begin{bmatrix} \frac{1}{2} & \frac{10}{11}\alpha \\ \frac{1}{11}\alpha & \frac{1}{2}(1+\alpha^2) \end{bmatrix}, \qquad (5.1.15)$$

and

$$A^*(t) = A^* = \begin{bmatrix} \frac{1}{2} & \frac{1}{11}\alpha \\ \frac{10}{11}\alpha & \frac{1}{2}(1+\alpha^2) \end{bmatrix} = A^T. \qquad (5.1.16)$$

It is noted that the cross-coupling introduced by a non-zero coefficient α in Figure 5.1.2 causes the vector noise source $y(t)$ to have an asymmetrical characteristic matrix. If the cross coupling coefficient α were time-varying, then the integrals (5.1.10, 11) would not have had the simple evaluations (5.1.15, 16) and $A^*(t)$ would not have equalled $A^T(t)$.

As a check on (5.1.15, 16), we note that

$$H(t,o) = \begin{bmatrix} 1 & 0 \\ \alpha & 1 \end{bmatrix},$$

and hence from (5.1.12) we have

$$A + A^* = \begin{bmatrix} 1 & \alpha \\ \alpha & 1+\alpha^2 \end{bmatrix} = 2\,D_y\,, \qquad (5.1.17)$$

which agrees with the sum of (5.1.15) and (5.1.16).

Checking A and A^* by calculating the intensity (5.1.17) of the noise $y(t)$ is not a complete check on A or A^*, as it does not bring out the asymmetry of the matrices A or A^*. This is to say that D_y does not contain all the information of A (or A^*)

and hence does not sufficiently characterize the vector noise $y(t)$ when the noise $y(t)$ is used in a non-additive situation. We illustrated this by an example in Section 4.3.

5.1.2 Pseudo Random Sequences

Maximum-length pseudo random sequences generated by linear digital shift registers are a useful source of random noise, and are receiving much current attention. A survey of the principles of generation and the properties of such sequences is given by Davies [84]. The simplest sequences are binary sequences, where the noise takes on the values $+1$ and -1, and more complicated sequences such as ternary sequences only have advantages over binary sequences in certain applications not related to their use as a simulation noise source [85].

Briefly, the advantages of pseudo random binary sequences (PRBS) over conventional noise generators are as follows:

(a) they are easy to build from cheap digital binary logic circuitry (or simple to program on to a digital computer);

(b) their output is not subject to drift or other variations associated with non-digital electronic circuitry, and the noise records are easily reproducible;

(c) the bandwidth of the PRBS is easily adjusted by altering the clock rate of the shift registers;

(d) the PRBS is periodic, and marking pulses are conveniently obtained at any points in the period to control the modes of a repetitive analogue computer;

(e) the statistical properties (e.g. $R(t,\tau)$) of the PRBS are known exactly for a finite time record of the signal (equal to any multiple of the period of the signal), whereas the exact statistical properties of a conventional noise source are only known a priori over an infinite record length;

(f) uniformly distributed or binomial sequences are obtained from simple operations on the shift register;

(g) approximately Gaussian noise can be obtained by filtering the PRBS. This fact is useful in simulation exercises, as is discussed in more detail below.

A study by Roberts and Davis [87] has suggested that a filtered PRBS could be a useful approximation to Gaussian noise, and as such would be useful in simulation exercises. They consider a first order linear filter, and derive the autocorrelation function and power density spectrum of the filtered PRBS. Both these functions are similar in form to filtered white noise, provided the filter time constant is at least twice the bit interval Δ of the PRBS*.

However, Roberts and Davis also determine the first order probability density function of the filtered PRBS experimentally, and find that as the value of the filter time constant is increased, the nearness of the density function to the Gaussian improves up to a certain point and then deteriorates as the filter time constant is further increased. This tendency away from the Gaussian density function when the PRBS is "heavily smoothed" is rather unexpected as the general tendency for non-Gaussian signals is to become more Gaussian when they are smoothed. Indeed this is so for the filtered random telegraph signal [89], which has many outward similarities with the filtered PRBS.

Roberts and Davis offer no explanation for this unusual phenomena, but the theoretical study of Tausworthe [90] and the experimental results of White [91] suggest an explanation. Tausworthe shows that n-tuples of adjacent bits of the PRBS form a mutually uncorrelated set of binary digits, provided that the length of the set n is less than or equal to the length M of the shift register generating the PRBS. This is an expected result, for any n adjacent bits uniquely define the $(n+1)$st

---

* An error has been found in Roberts and Davis' derivation of the autocorrelation function of the filtered PRBS [88]. However, the final answer they give is accurate for values of filter time constant $T \leq 10\Delta$, and the error does not affect their conclusions regarding the approximation of the filtered PRBS to Gaussian noise.

bit in the sequence when n is greater or equal to M, but not if n is less than M. Illustrating this result, White [91] shows that the sum of these n adjacent bits in the sequence forms a good binomial distribution only if n is less than or equal to M.

The connection with Roberts and Davis' results is as follows. Filtering the PRBS is equivalent to taking a weighted sum of adjacent bits of the PRBS according to the shape and length of the memory (impulse response = $e^{-t/T}$) of the filter. When the memory of the filter extends significantly beyond $M\Delta$, where Δ is the bit length of the PRBS, the effect of the filter is to form a weighted sum of more than M adjacent bits of the PRBS. As these bits are no longer independent, we cannot expect the weighted sum to be Gaussian. This effect is shown in Figure 5 of [87] where, for the three binary sequences tested ($M = 6$, 7 and 9), the density functions of the filtered PRBS were nearest the Gaussian when the value of the filter time constant T equalled $\frac{M\Delta}{2.5}$. As $e^{-2.5} = 0.08$, this T could be considered as the value of time constant at which the filter memory begins to significantly extend beyond $M\Delta$, and the explanation given above is supported.

The implications of the effect of these non-Gaussian properties of the filtered PRBS on simulation exercises has not been fully assessed. It does seem that we must be careful in testing non-linear systems with such signals or even linear systems when their higher order statistical properties are desired.

A Diffusion Model for Pseudo Random Binary Sequence

A model [57]*was constructed for a PRBS for two reasons: (a) there was a desire to see how the deterministic properties of the PRBS (as opposed to the random properties of the random telegraph signal, for example) affected the statistics of systems driven by the PRBS. The deterministic properties are the properties of periodicity and the fact that the total number of +1 and -1 bits in each period of the PRBS is known. The mechanism of drawing balls

---

* Appendix D

from an urn without replacement was a convenient model for the latter property.

(b) As the PRBS is potentially a convenient noise source for a repetitive simulation, there was a desire to know if the PRBS had any unusual transient properties. Continuous diffusion processes have significant advantages over non-Markovian processes for transient analysis purposes [74], and the method of Section 4.1 conveniently gave a Markov diffusion model for the non-Markovian urn mechanism.

The construction of the diffusion model of the PRBS and the statistical analysis of the output of a linear first order filter driven by the PRBS has been reported in detail in [57]. The main conclusions are as follows:

(1) The mean square or variance of the filtered PRBS has a discontinuous time derivative at $t = L\Delta$ if the significant memory of the filter exceeds the period $L\Delta$ of the PRBS. Thus we must have the filter time constant $T < \frac{1}{4} L\Delta$ to ensure that this unusual transient effect is not present. In repetitive simulations, however, it is usually convenient to choose the period of the PRBS greater than the time of interest of the transient simulation (so that the computer can be reset before the beginning of the next period), and the unusual variance transient will not be experienced.

(2) The transient mean square or variance of the filtered PRBS is different from that of filtered white noise unless T is substantially less than $\frac{L\Delta}{4}$, and so we should choose T less than about $\frac{L\Delta}{25}$ if the PRBS is to simulate the transient properties of filtered white noise. (A choice of $T = \frac{L\Delta}{25}$ gave an 8% discrepancy in transient mean square.) If we are not concerned about simulating filtered white noise, any T can be chosen, as the transient statistics of the filtered PRBS are known for all T [57].

Characteristic Matrix of Filtered PRBS

As the PRBS is periodic, the PRBS or functions of it does not possess a characteristic matrix by the strict definition of Section 4.1, equation (4.1.21), as the semi-infinite integral of the correlation function of the PRBS is not convergent. However, if we are using the PRBS only over one period, $L\Delta$, we can forget that it is a periodic signal, and use the finite record correlation functions to obtain an effective characteristic matrix.

For example, the autocorrelation function of the filtered PRBS is derived in [88]. Provided $T \ll \frac{1}{8} L\Delta$, the expression given for the shift parameter τ in the range $[-\frac{1}{2}L\Delta, \frac{1}{2}L\Delta]$ can be used for the finite record autocorrelation function with the assumption that the function is zero outside this range. The validity of this assumption was confirmed by an experimental method of determining the characteristic matrix described in Section 5.2.3. The characteristic matrix of a vector noise source involving the PRBS could be calculated by methods similar to Section 5.1.1.

5.2 Experimental Results Illustrating the Differing Biases of Physical Processes and Diffusion Processes

As an experimental justification of the results of Chapter 4, the purpose of this section is to

(1) illustrate that a simulation of a diffusion process by the "naive" method of programming the Ito stochastic differential equation directly on to the analogue computer may give the simulation a wrong bias;

(2) predict what the proper bias should be by calculating the characteristic matrix of the physical noise used in the simulation;

(3) illustrate that the simulation of a diffusion process using the bias calculated in (2) has the correct statistical bias.

In sections 5.2.1-3 below, we give three examples: the first two involve scalar noise sources and show the need of the correction term, QA of equation (4.4.9). The third example uses a vector noise source and shows some of the consequences of using a noise source with an asymmetrical characteristic matrix in a simulation.

The first example, Section 5.2.1, illustrates a method of determining the characteristic matrix A of a scalar noise source experimentally. It is seen that this is essentially a problem of scaling the noise properly, and in the example, this scaling affects the drift of the process in a simple manner. It is then seen that our method of determining A is most relevant to ensuring that the simulated process has the correct drift.

The second example, Section 5.2.2, outlines the construction of a non-linear filter, and illustrates that some of the results of modern stochastic control theory (developed in the stochastic calculus) can be wrongly applied in practice unless the results of Chapter 4 are taken into consideration.

The third example, Section 5.2.3, illustrates a method of determining the matrix A experimentally for a vector noise source.

5.2.1 Example Illustrating the Scaling of an Independent Noise Source

We begin with a simple example for which the correct drift of the system is easily obtained. Consider the diffusion process given by the Ito s.d.e.

$$\begin{aligned} dx_1(t) &= \sigma\, dw(t) \\ dx_2(t) &= x_1(t)\, \sigma\, dw(t) \end{aligned} \qquad (5.2.1)$$

which is the example given as an Ito stochastic integral at the end of the first half of Appendix A. A naive method of simulating this process is to divide equation (5.2.1) symbolically by dt and programming the resultant equation

$$\begin{aligned} \dot{\underline{X}}_1(t) &= y(t) \\ \dot{\underline{X}}_2(t) &= \underline{X}_1(t)\, y(t) \end{aligned} \qquad (5.2.2)$$

directly on to the analogue computer using a suitable physical noise source $y(t)$ as a replacement for the white noise $\sigma \dot{w}(t)$. We do this below, showing that the simulation has an incorrect bias, and incidentally, showing that the simulation (5.2.2) is correct if (5.2.1) is interpreted as a Stratonovich s.d.e.

Suppose we have at our disposal a zero mean, high bandwidth, stationary noise generator. Basically we must determine what parameters of the generator output we must measure in order that the statistics of the simulation be known.

Scaling of a Scalar Noise Source

In this section, we present a simple method of measuring the generator output which sufficiently characterises the noise for the case where the noise is scalar, or vector with independent components. In these cases the noise characteristic matrix A is diagonal, $A = A^T$, and the zero frequency power spectral density of the noise,

$A + A^T = B B^T$, is the only parameter we must measure. In Section 5.2.3, a more elaborate method of measuring the generator output is presented, which characterises a vector noise source with cross-correlated components, so that the off-diagonal elements of A can be determined.

Consider the output $y(t)$ of a scalar noise generator (the components of a vector independent noise generator can be considered individually in this fashion). We must determine the zero frequency power spectral density $S_{yy}(0)$ of $y(t)$. Note that a signal with no delta functions in its continuous power spectrum $S_{yy}(\omega)$ has exactly zero power at any given frequency ω_o - thus our zero mean signal $y(t)$, which has zero d.c. power ($\omega_o = 0$), does have a non-zero zero frequency power spectral density. Also note that the meter commonly found on commercial noise generators which measures the r.m.s. value or (total) power of the output is not sufficient to determine $S_{yy}(0)$ unless the detailed shape of the power spectrum is known.

Essentially we have a scaling problem, for the amplitude of $y(t)$ must be adjusted so that $y(t)$ of (5.2.2) models $\sigma \dot{w}(t)$ of (5.2.1). But as we are interested in the statistics of $\underline{X}(t)$ which is an integral function of $y(t)$, it is more appropriate to consider the matching of

$$\int_o^t y(s)\, ds \quad \text{to} \quad \sigma w(t).$$

This is supported by the fact that Clark considers the convergence of the physical process (5.2.2) to the diffusion process (5.2.1) when the integral of $y(t)$ converges in the mean to $\sigma w(t)$ [see equation (2.2.2c)].

Thus we should integrate the output of the noise generator and scale $y(t)$ so that the variance of the integral equals the variance of $\sigma w(t)$, $\sigma^2 t$. ($\int_o^t y(s)\, ds$ has a zero mean and in most cases will be near Gaussian, so the variance is a sufficient matching parameter). This matching is a statistical operation, and so as many trials as conveniently possible should be carried out to

ensure an accurate matching. But note that non-overlapping increments of the form

$$\int_t^{t+\delta t} y(s)\, ds \doteq \sigma[w(t+\delta t) - w(t)] \qquad (5.2.3)$$

are essentially independent provided that δt is substantially greater than the significant memory time of the noise $y(t)$. This choice of δt entails the same considerations as in Section 4.1.

To illustrate the independence of the increments (5.2.3), consider $y(s)$ generated by a linear stationary filter $h(s-u)$ whose significant memory time is $\tau_{cor} = \tau$. Then

$$y(s) = \int_{-\infty}^{s} h(s-u)\, dw(u) \doteq \int_{s-\tau}^{s} h(s-u)\, dw(u),$$

and

$$\int_t^{t+\delta t} y(s)\, ds \doteq \int_t^{t+\delta t} \int_{s-\tau}^{s} h(s-u)\, dw(u)\, ds.$$

Now assuming that δt is much greater than τ, we can interchange the order of integration simply to give

$$\int_t^{t+\delta t} y(s)\, ds \doteq \int_t^{t+\delta t} dw(u) \int_u^{u+\tau} h(s-u)\, ds. \qquad (5.2.4)$$

But the latter integral of (5.2.4) is independent of u and letting

$$\int_u^{u+\tau} h(s-u)\, ds = K, \qquad (5.2.5)$$

we have

$$\int_t^{t+\delta t} y(s)\, ds \doteq K \int_t^{t+\delta t} dw(u) = K[w(t+\delta t) - w(t)]. \qquad (5.2.6)$$

But the right hand side of (5.2.6) represents independent increments for non-overlapping time segments $(t, t+\delta t)$, and so the left hand side of (5.2.6) approximately possesses this property.

From equations (5.1.10, 11)

$$A(t) + A^*(t) = \int_0^\infty h(t,u)\, 2\, D_z(t-u) \int_0^\infty h^T(t-u+v,v)\, dv\, du,$$

and for the stationary independent noises considered in this section, this expression reduces to

$$A = \tfrac{1}{2}\left[\int_0^\infty h(\cdot, u)\, du \right]^2,$$

where A has only diagonal elements and $2D = I$ for the unit parameter white noise $\dot{w}(t)$ used above. Then from (5.2.5) we have

$$A = \tfrac{1}{2} K^2$$

where $K^2 \delta t$ is the variance of the increments (5.2.6) of the integral of the physical noise (note that the non-stationary $h(t,u)$ and stationary $h(s-u)$ weighting functions have different convolution integrals).

Thus the characteristic matrix of linear, independent, stationary noise sources can be found by the simple graphical method given in this section. This method likely also works for non-Gaussian noises, as this method is just a means of relating the low frequency components of $y(t)$ to that of white noise $\dot{w}(t)$, but this point is difficult to demonstrate.

Experimental Results

In the simulations of this section and the next, an Advance optical disc low frequency Gaussian noise generator was used which had an upper frequency of 50 Hz (say $\tau_{cor} = .02$). The integral of the output signal was recorded on a pen recorder with a 20 Hz response, and the increments (5.2.6) were calculated from the recording

for 40 adjacent time intervals of $\delta t = 0.25$ seconds. The average variance of these 40 increments was found to be 0.0088, which, from (5.2.6), is a random variable with mean $K^2 \delta t$ and variance $\frac{2K^4 \delta t^2}{40}$ (standard deviation = $K^2 \delta t / 4.5$). Thus an estimate of K^2 is 0.035 ± 0.007, or an estimate of K is 0.187 ± 0.02. Then in the simulation of (5.2.2) the generator output must be scaled by the factor $\frac{\sigma}{0.187}$ in order that $y(t)$ of (5.2.2) be a valid approximation to $\sigma \hat{w}(t)$ of (5.2.1). In other words, the generator output noise has the characteristic matrix whose single element is

$$A = \tfrac{1}{2} K^2 = \tfrac{1}{2} (0.035),$$

while the unit parameter white noise $\hat{w}(t)$ has the characteristic matrix $A = \frac{1}{2}$.

This statistical estimation of the $S(o)$ parameter of the generator output was repeated for the same noise record using adjacent time increments of $\delta t = 0.5$. The estimated K was 0.198 and the agreement verified the independence of the successive increments of (5.2.3). In general, this seems to be an attractive method of measuring $S(o)$, as the method is simple and depends little on the upper frequency response of the recording apparatus.

Using $\sigma = 0.187$ for convenience in (5.2.1), the system (5.2.2) was simulated on a TR-48 analogue computer, and the curve of Figure 5.2.1 was recorded for $\underline{X}_2(t)$, using the same noise record that we used to estimate K above. The integral of this noise record is $\underline{X}_1(t)$ and is also shown.

Now, (5.2.2) is an ordinary differential equation and can be integrated by the normal rules of calculus. We have

$$\underline{X}_2(t) = \int_o^t \underline{X}_1(s)\, y(s)\, ds,$$

$$= \int_o^t \underline{X}_1(s)\, d\underline{X}_1(s),$$

$$= \tfrac{1}{2} \underline{X}_1^2(s) \Big|_0^t = \tfrac{1}{2} \underline{X}_1^2(t), \qquad (5.2.7)$$

assuming zero initial conditions.

But $\underline{X}_1(t)$ has the statistics of a Wiener process in the large, whose mean square equals $\sigma^2 t$. Thus

$$E[\underline{X}_2(t)] = \tfrac{1}{2} E[\underline{X}_1^2(t)] = \tfrac{1}{2}\sigma^2 t = 0.0175\, t\,. \qquad (5.2.8)$$

This mean value $0.0175\, t$ is drawn in Figure 5.2.1, and appears to be a good estimate of the mean value of the particular curve $\underline{X}_2(t)$ shown. This curve was typical of many obtained, but unfortunately it was not convenient to obtain statistical estimates on a large scale on the analogue computer to obtain a good estimate of $E[\underline{X}_2(t)]$. However, our point here is that $E[\underline{X}_2(t)]$ of (5.2.2) does not equal $E[x_2(t)]$ of (5.2.1), for $x_2(t)$ has a zero mean value (see (A6b)) while $\underline{X}_2(t) = \tfrac{1}{2}\underline{X}_1^2(t)$ is always positive and so has a non-zero mean value.

Thus in order to simulate the diffusion process (5.2.1) correctly, we must modify the physical process (5.2.2) so that $E[\underline{X}_2(t)] = E[x_2(t)] = 0$.* But from (5.2.7), $\underline{X}_2(t) = \tfrac{1}{2}\underline{X}_1^2(t)$ provided the analogue computer operations are performed accurately (the multiplication of $\underline{X}_1(t)$ and $y(t)$ in (5.2.2) likely introduces the largest error), and $\underline{X}_1(t)$ is a good approximation to a Wiener process with $E[\underline{X}_1^2(t)] = \sigma^2 t$. This latter point illustrates the importance of scaling $y(t)$ as we did, for we took care to ensure that

$$\int_t^{t+\delta t} y(s)\, ds = \underline{X}_1(t + \delta t) - \underline{X}_1(t)$$

has the correct mean square increments.

---

* We are already ensured that $E[\underline{X}_1(t)] = E[x_1(t)] = 0$ as $y(t)$ has a zero mean value, and as discussed in Section 4.1, a discrepancy in mean square between the physical and diffusion process does not exist provided $y(t)$ is scaled correctly.

Thus $E[\underline{X}_2(t)] = \frac{1}{2}\sigma^2 t$ in (5.2.2), and it is clear that we can modify the system (5.2.2) to make $E[X_2(t)] = 0$ by putting $X_2(t) = \underline{X}_2(t) - \frac{1}{2}\sigma^2 t$. Then with $\underline{X}_1(t)$ $(= X_1(t))$ unaltered, the ordinary differential equation generating $X(t)$ is

$$\dot{X}_1(t) = y(t)$$

$$\dot{X}_2(t) = -\tfrac{1}{2}\sigma^2 + X_1(t)\, y(t), \qquad (5.2.9)$$

and a simulation of (5.2.9) on an analogue computer should provide an unbiased estimate of the diffusion process (5.2.1).

We see that this result agrees with the general result of Section 4.4. Relating the diffusion process (4.4.1) to the present example (5.2.1), we note

$$f(x,t) = 0,$$

$$F_{11}(x,t) = \sigma,$$

and $$F_{21}(x,t) = \sigma x_1(t). \qquad (5.2.10)$$

As we have scaled $y(t)$ so $E[\int_0^t y(s)\, ds] = \sigma^{-2}t$, we have $A = \frac{1}{2}\sigma^2$, and from (4.4.6b), $C = \sigma^{-1}$. Then from (4.4.6a),

$$G_{11}(X,t) = 1,$$

and $$G_{21}(X,t) = X_1(t). \qquad (5.2.11)$$

Also from (4.4.7),

$$(Q)_1 = 0$$

and $$(Q)_2 = 1,$$

which leads to the bias term $$-QA = \begin{bmatrix} 0 \\ -\frac{1}{2}\sigma^2 \end{bmatrix} \qquad (5.2.12)$$

in (4.4.9). Then (5.2.11) and (5.2.12) result in the equivalence of the physical processes (4.4.9) and (5.2.9), which, from the results of Chapter 4, indicate that (5.2.9) is the correct physical process to simulate the statistics of the diffusion process (5.2.1.).

This is illustrated by the experimental run of Figure 5.2.1, for $X_2(t)$ of equation (5.2.9) equals $\underline{X}_2(t)$ of equation (5.2.2) (the curve is so indicated in Figure 5.2.1) minus the ramp $\frac{1}{2}\sigma^2 t$ (also shown in Figure 5.2.1 as $E[\underline{X}_2(t)]$). Subtracting these two curves appears to give a good zero mean signal and so this limited experimental evidence indicates that the physical process (5.2.9) is the correct process to simulate the diffusion process (5.2.1). Further experimentation on a digital computer confirmed that $X_2(t)$ of (5.2.9) is a random signal chosen from a distribution of mean zero and variance $\frac{3}{4}(\sigma^2 t)^2$ (the correct variance).

5.2.2 Construction of a Non-Linear Filter

The mathematical convenience of the stochastic calculus has led to its use in the formulation and solution of a number of problems of filtering and optimal control of continuous stochastic systems (see, for example [92-99], [49]). Essentially we have the problem (mentioned only by Wonham of the above authors) that the physical situation that the formulation treats must be described in the stochastic calculus and the solution obtained in the stochastic calculus must be translated into the ordinary calculus so that such a controller or filter can be built. This problem has been discussed theoretically by Clark in the context of filtering problems [22, Chapter 6], and in this section we illustrate how such a filter should be built. An example of a non-linear filter derived by Wonham [49], is convenient for our purposes.

Consider a random telegraph signal $x(t) = \pm 1$ with zero mean value and an expected number of transitions per unit time of v. The signal $x(t)$ is observed by an instrument with additive white noise

$$\bar{x}(t) = x(t) + \beta \dot{w}(t) , \qquad (5.2.13)$$

and the filtering problem is to obtain the best estimate of $x(t)$ given all past measurements, $\bar{x}(s)$, $s \leq t$. This estimate is conveniently summarised in the "sufficient statistic"

$$p(t) = \text{Prob}[x(t) = +1 \mid \bar{x}(s)] - \text{Prob}[x(t) = -1 \mid \bar{x}(s)]. \qquad (5.2.14)$$

Assuming the transitions to be Poisson distributed, $x(t)$ is a Markov process (continuous in time but not in state), and the integral of the measurements $\bar{z}(t) = \int^t \bar{x}(s)ds$ is a continuous Markov process given by the Ito s.d.e.

$$d\bar{z}(t) = x(t)\, dt + \beta\, dw(t). \qquad (5.2.15)$$

Then $\bar{z}(t)$ is the "measurement state".

A physical process which $x(t)$ approximates is one in which the transitions are Poisson distributed except that a lower limit is placed on the interval between successive transitions (this puts an effective upper frequency limit on the physical process corresponding to $x(t)$). However, as the measurement noise $\beta \dot{w}(t)$ of (5.2.13, 15) is additive, there is no problem in interpreting

$$\dot{\bar{z}}(t) = x(t) + \beta \dot{w}(t) \qquad (5.2.16)$$

as an ordinary differential equation governing the physical instrument when $\dot{w}(t)$ is not strictly white.

Although $x(t)$ is not a diffusion process and thus cannot be described by an Ito s.d.e., it is a non-linear process in the sense that the density of $x(t)$ is non-Gaussian. Thus the optimal

filter generating $p(t)$ of (5.2.14) is non-linear, and Wonham gives the Ito s.d.e. governing the filter as

$$dp(t) = -2v\,p\,dt - \beta^{-2}p(1-p^2)\,dt + \beta^{-2}(1-p^2)\,d\bar{z}(t) \qquad (5.2.17)$$

where $d\bar{z}(t)$ is the measurement (input to the filter) in the form (5.2.15).

The filter is non-linear as evidenced by the $(1-p^2)$ term in (5.2.17), but what is more important from our point of view is that the noise in (5.2.17) is non-additive. This is seen by combining (5.2.17) and (5.2.15) to give

$$dp(t) = -2\,v\,p\,dt - \beta^{-2}p(1-p^2)\,dt + \beta^{-2}(1-p^2)\,x\,dt$$

$$+ \beta^{-2}(1-p^2)\,\beta\,dw(t), \qquad (5.2.18)$$

where the last term of (5.2.18) represents non-additive noise. Thus care must be taken in converting the Ito s.d.e. (5.2.17) for the non-linear filter to an ordinary differential equation so that the filter can be simulated or constructed. This difficulty does not arise in the construction of linear filters (Kalman filters) where the noise is additive in the filter equations [100]. The same remarks apply to linear vs. non-linear stochastic control problems, except that the control (and filtering) of linear systems with stochastic coefficients [94] entails the same considerations as non-linear control problems in the context of this chapter, as the noise is non-additive.

We now use the results of Section 4.4 to show what physical system we must construct in order to obtain the correct filter as represented by the Ito s.d.e. (5.2.17). In the notation of equation (4.4.1), we have from (5.2.18),

$$f(p,t) = -2\,v\,p - \beta^{-2}p(1-p^2) + \beta^{-2}(1-p^2)x \qquad (5.2.19)$$

and $$F(p,t) = \beta^{-1}(1-p^2). \qquad (5.2.20)$$

As the noise in (5.2.18) is scalar, we can simply put $G(p,t) = F(p,t)$, and we only have to be careful to scale the physical noise $y(t)$ properly so that it approximates to the unit parameter white noise $\dot{w}(t)$. From the last section we see that this scaling makes the characteristic matrix of $y(t)$, $A = \frac{1}{2}$.

Then from (4.4.7),

$$QA = -\beta^{-2} P(1-P^2), \tag{5.2.21}$$

and the physical process (4.4.9) corresponding to the diffusion process (the filter) (5.2.18) is

$$\dot{P}(t) = -2 v P + \beta^{-2}(1-P^2)\, x + \beta^{-2}(1-P^2)\, \beta\, y(t). \tag{5.2.22}$$

But, as mentioned earlier, the "physical" observation $x(t) + \beta y(t)$ (5.2.16) can be considered equivalent to the "white noise" observation (5.2.13) and so the ordinary d.e. for the filter is

$$\dot{P}(t) = -2 v P(t) + \beta^{-2}(1 - P^2(t))\, \bar{x}(t) \tag{5.2.23}$$

where $\bar{x}(t)$ is the observation driving the filter. This equation can be compared with the ordinary d.e. obtained by using the terms of the Ito s.d.e. (5.2.17) directly:

$$\dot{\underline{P}}(t) = -2 v \underline{P}(t) - \beta^{-2} \underline{P}(t)\, (1 - \underline{P}^2(t)) + \beta^{-2}(1 - \underline{P}^2(t))\, \bar{x}(t). \tag{5.2.24}$$

From the results of Chapter 4, equation (5.2.23) should give the correct, and equation (5.2.24) should give an incorrect physical approximation to the non-linear filter (5.2.17).

The filters of equations (5.2.23) and (5.2.24) were simulated on a TR-48 analogue computer. The random telegraph signal $x(t)$ was obtained from a bistable ($\pm$ 1) circuit driven from the zero crossings of a zero-mean low frequency Gaussian signal [101, Chapter 10], and the mean switching rate was $v = 0.35$ switches per second. The noisy measurement $\bar{x}(t)$ was obtained by adding the noise of the last section to $x(t)$. The scaling factor β corresponds to K or σ of the last section, and in this section, the generator output was doubled, giving $\beta = 0.374$.

Figures 5.2.2, and 5.2.3, show simulations of equations (5.2.23) and (5.2.24) respectively, using the same noise record and random telegraph signal. When $P(t)$ is close to $\pm$ 1, the filter is quite confident that it has the right answer. Comparing the two filters qualitatively, we see that the filter of equation (5.2.24) is less sure of itself in the "steady state" than the filter of equation (5.2.23), although each filter seems equally quick in recognising that a switch has occurred.

More quantitatively, the mean square estimation error suggested by Wonham [49, equation (28)] was evaluated for each filter. For the record shown, the filter of equation (5.2.23) had a mean square estimation error of 0.26, while the filter of equation (5.2.24) had an error of 0.38. Wonham derives the theoretical mean square estimation error of the proper filter described by the Ito s.d.e. (5.2.17) and gives it as 0.27 for the parameters of the present example [49, Figure 2]. Wonham also gives the mean square estimation error of the best linear (Wiener) filter as 0.35. Thus we see that the mean square estimation error of the filter of equation (5.2.23) is very close to the theoretical error, but the error of the filter of equation (5.2.24) is considerably higher, and is even worse than the best linear filter.

Although it is difficult to place estimates of accuracy on the filter performance data given above, we can at least be confident from Figures (5.2.2) and (5.2.3) that the mean square estimation error of the equation (5.2.23) filter is <u>lower</u> than that of the equation (5.2.24) filter. But by definition, the optimal filter (5.2.17) must give the <u>lowest</u> mean square estimation error, and so

of the two filters simulated, equation (5.2.23) is the better approximation to the correct filter. To this extent, the experimental evidence of the example of this section substantiates the theoretical results of Chapter 4. Digital computer simulations confirm this with a higher confidence.

In [49] Wonham gives a block diagram for an analogue implementation of his optimal non-linear filter (5.2.17). He programs it directly from the Ito equation (5.2.17), that is, the filter of equation (5.2.24), and states that this is to be an ideal analogue device for generating p(t). Although the meaning of "ideal analogue device" is open to various interpretations, if it is to be interpreted as an analogue device which obeys the normal rules of calculus (i.e. integration) over an infinite bandwidth, then it would solve Stratonovich s.d.e.'s correctly but not Ito s.d.e.'s. But the limiting form of equation (5.2.23) (as the noise bandwidth is extended to infinity) is the Stratonovich s.d.e. for the optimal filter, and the limiting form of (5.2.24) is the Ito s.d.e. for the optimal filter. With this assumption, the conjecture stated by Wonham that an ideal analogue device should be programmed according to the Ito equation is wrong. Bypassing the speculation on the properties of an ideal analogue device, the results of Chapter 4 and this section indicate that a practical analogue device should be programmed according to the Stratonovich-like equation (5.2.23) and not according to the Ito-like equation (5.2.24).

Recently Wonham [96] has reversed his earlier conjecture, and suggests that an analogue device should be programmed according to the Stratonovich equation for the filter. This agrees with our present results. He quotes (private communication, June 1965) experimental work of Ternan, which, although unreported, is likely coincidental with, and contemporary with (February-April 1965), the present authors' experimental work. Later, Wong and Zakai [102] and Kulman [103] also support the new conjecture. Other authors such as Ariaratnam and Graefe [14], Gray and Caughey [41], Caughey and Dienes [13], and Leibowitz [106], who do not specifically mention physical processes (in our sense), support the conjecture insofar

as they use Stratonovich equations as opposed to Ito equations to describe systems driven by ideal white noise (the inference being that we should program a computer directly from their equations).

None of these authors, however, consider the simulation of a stochastic system using a vector physical noise source with an asymmetrical characteristic matrix. The next section will illustrate how simulation results can be biased if care is not taken over the possible asymmetry of a vector physical noise source.

5.2.3 Example using an Asymmetrical Noise Source

In the last two sections we considered the simulation of examples with scalar noise sources. We found that if we wished to simulate a diffusion process described by an Ito s.d.e., the naive method of programming the Ito equation directly on to the computer gave results with a wrong bias, while programming from the associated Stratonovich equation gave correct results.

Having learned this lesson (along with the authors mentioned at the end of the last section), let us consider an example of a diffusion process with a vector noise source and see what happens when we try and simulate it by programming the Stratonovich equation directly.

Consider a special case of the example discussed by Astrom [26]: the diffusion process described by the Ito s.d.e.

$$dx(t) = x(t)\, dw_1(t) + dw_2(t), \tag{5.2.25}$$

where $\dot{w}(t)$ is a two-dimensional white noise with the intensity matrix

$$2D = \frac{1}{15}\begin{bmatrix} 1 & .2 \\ .2 & 1 \end{bmatrix}, \tag{5.2.26}$$

(this means that the noises $\dot{w}_1(t)$ and $\dot{w}_2(t)$ have the same co-spectral density and have a small cross-correlation). Then from

(A22) (where the matrix $2D$ is the identity matrix) we can change the Ito equation (5.2.25) into a Stratonovich equation by subtracting the bias term

$$\sum_{k,l}^{2} Q_{kl} D_{kl} \, dt = (D_{11}x + D_{12}) \, dt \qquad (5.2.27)$$

from the Ito equation. The resultant Stratonovich s.d.e. for the diffusion process $x(t)$ is

$$\bar{d}x(t) = -(D_{11} x(t) + D_{12})dt + x(t) \, \bar{d}w_1(t) + \bar{d}w_2(t), \qquad (5.2.28)$$

and we shall simulate the diffusion process $x(t)$ by programming the ordinary differential equation

$$\dot{X}(t) = -D_{11} X(t) - D_{12} + X(t) \, y_1(t) + y_2(t) \qquad (5.2.29)$$

directly on to the analogue computer, using a two-dimensional noise source $y(t)$ which appears to have the property (5.2.26).

A Filtered PRBS Noise Source

Following the discussion of Section 5.1.2, we use a filtered PRBS as a primary noise source. We use a 15 stage shift register which generates an $L = 32{,}767$ bit code, driven at a clock frequency of 2.16 kHz giving a bit period of $\Delta = 0.463$ msec (then $L\Delta = 15.2$ secs.). The filter we use has a time constant of $T = 4$ msec, which equals $8.7\,\Delta$ (this is close to the value 6Δ which gives the best Gaussian signal), giving the noise a correlation time, $\tau_{cor} = 20$ msec.

As a second noise source, we take a delayed version of the filtered PRBS. We take $y_1(t)$ off the first stage of the shift register and $y_2(t)$ off the last (15th) stage, both channels being filtered as above. Then

$$y_2(t) = y_1(t - \tau') \qquad (5.2.30)$$

where $\tau' = 14\Delta = 6.5$ msecs. Then from [88], the correlation coefficient between $y_1(t)$ and $y_2(t)$ is approximately

$$e^{-6.5/4} = 0.198.$$

This was confirmed experimentally by evaluating integrals of the form

$$\frac{1}{L\Delta} \int_0^{L\Delta} y_i(t)\, y_j(t)\, dt \quad , \quad i, j = 1, 2 \qquad (5.2.31)$$

on the analogue computer, where the value 0.185 was obtained for the normalised cross-correlation.

Now, on the face of it, it would seem that this two-dimensional noise source has the right correlation properties to simulate the white noise with the intensity (5.2.26). It is not suggested that this noise source is a sensible one to use for this example, as the fact that $y_2(t)$ is merely a delayed version of $y_1(t)$ hints that this two-dimensional noise source has unusual properties. However, it is a convenient noise source to illustrate the following point: we cannot characterise a physical noise source by its intensity, or even by its zero-shift correlation function, as we might be tempted to do from the following argument.

Authors, e.g. [14, 41], who give Stratonovich s.d.e.'s for diffusion processes define the noise in such equations, e.g. $\dot{w}(t)$ in (5.2.28), by its correlation property

$$R(\tau) = E[\dot{w}(t)\, \dot{w}^T(t - \tau)] = 2D\delta(\tau) , \qquad (5.2.32)$$

where $\delta(\cdot)$ is the Dirac delta function. They state, or imply, that when their equation describes a physical situation, the relation (5.2.32) is approximately true (see also Astrom [26. p. 318], but he does not use Stratonovich equations). That is, the high bandwidth physical noise $y(t)$ (which $\dot{w}(t)$ models in the Stratonovich equation) appears to have a delta correlation function when viewed at the relative time scale of the physical system.

Now consider the filtered PRBS noise source $y(t)$ which we propose to use in the simulation of (5.2.25 or 28) by computing with the equation (5.2.29). According to the implications of the current literature, we must ensure that $y(t)$ approximately has the property (5.2.32), where D is specified in (5.2.26). The important point is this: the implication is that the correlation function $R_y(\tau)$ of $y(t)$ should have an area of $2D$, concentrated in a small region about the origin. Thus we should measure the zero frequency spectral density of $y(t)$, and the noise can then be scaled properly. We show in this section that this is an insufficient characterisation of the physical noise.

It is possible to go one step further in error. It is sometimes not convenient to measure the spectral density of a physical noise. The correlation function is computed instead, and the noise source is scaled to have the property (5.2.32) via its correlation properties. Now our primary noise source has a correlation time of 20 msec and $y_2(t)$ is shifted by 6.5 msec, giving a joint maximum correlation time of 26.5 msec. We will likely be interested in the system $X(t)$'s behaviour over intervals in the order of hundreds of milliseconds to seconds, and so the correlation equipment at our disposal likely only has a time shift resolution of say 100 msec. Using this equipment, the noise correlation function appears to be a delta function with $R_y(0)$ proportional to

$$\begin{bmatrix} 1 & .185 \\ .185 & 1 \end{bmatrix} .$$

The temptation now is to apply an arbitrary scaling technique (such as applying the technique of Section 5.2.1 to one component of $y(t)$), and decide that our noise source, after scaling, has the intensity matrix

$$2D_y = \frac{1}{15} \begin{bmatrix} 1 & .185 \\ .185 & 1 \end{bmatrix} . \qquad (5.2.33)$$

The Simulation of Equation (5.2.29)

Using this noise source, the equation (5.2.29) was programmed on to the TR-48 analogue computer, and curve 1 of Figure 5.2.4 was obtained. Now, according to our method of characterising the noise source, the components $y_1(t)$ and $y_2(t)$ have the same statistical properties and so the statistics of the simulation should not be affected by interchanging $y_1(t)$ and $y_2(t)$. This was done, and curve 2 of Figure 5.2.4 was obtained for $X(t)$ using the same portion of the PRBS noise record and the same initial conditions.

It is noticed that curves 1 and 2 have a distinctively different drift, as their separation increases linearly with time. In fact, as we have used the same noise record, the separation is deterministic, and the curves shown are typical of many obtained. The simulations in the curve 1 group had a mean value near zero, while the simulations in the curve 2 group had a mean value close to the ramp function $0.05t$, which is curve 3 shown in Figure 5.2.4.

Thus the simulations using the original noise source, and using this noise source with reversed components, have different drifts, and so both of the arrangements cannot be correct simulations of the given diffusion process $x(t)$. Our point here is not to show experimentally which arrangement gives the correct simulation, but to show that curves 1 and 2 have different drifts, which is not predicted by our method of characterising the noise source. To trace the origin of these differing drifts, let us characterise the noise $y(t)$ by the method proposed in Chapter 4.

Experimental Determination of the Characteristic Matrix of a Vector Correlated Noise Source

The ij:th element of the characteristic matrix A is given as (see equation (4.1.21) but here $y(t)$ is stationary)

$$A_{ij} = \int_0^\infty E[y_i(t)\, y_j(t-\tau)]\, d\tau. \qquad (5.2.34)$$

Putting $s = t - \tau$ and interchanging the integral and the expecta-

tion in (5.2.34) we have

$$A_{ij} = E[y_i(t) \int_{-\infty}^{t} y_j(s)\, ds] ,$$

$$= E[y_i(t)\, Y_j(t)] , \qquad (5.2.35)$$

where $$Y_j(t) = \int_{-\infty}^{t} y_j(s)\, ds. \qquad (5.2.36)$$

Now, $y(t)$ is a stationary signal, but $Y(t)$ is not a stationary signal (it behaves like a Wiener process in the large), and so for the purpose of computing the statistical average (5.2.35) it is convenient to replace $Y(t)$ by a stationary signal $\bar{Y}(t)$ so that

$$E[y(t)\, Y(t)] \doteq E[y(t)\, \bar{Y}(t)] . \qquad (5.2.37)$$

Writing (5.2.35) as

$$E[y_i(t) \int_{-\infty}^{t} [1 - e^{-a(t-s)} + e^{-a(t-s)}] y_j(s)\, ds] = E_1 + E_2 \qquad (5.2.38)$$

where

$$E_1 = E[y_i(t) \int_{-\infty}^{t} [1 - e^{-a(t-s)}]\, y_j(s)\, ds], \qquad (5.2.39)$$

and $$E_2 = E[y_i(t) \int_{-\infty}^{t} e^{-a(t-s)}\, y_j(s)\, ds], \qquad (5.2.40)$$

we see that the term E_1 is effectively zero as long as $e^{-a(t-s)} \doteq 1$ for $t - s$ less than τ_{cor} (for $t - s > \tau_{cor}$, $E[y(t)\, y(s)] = 0$). Thus if we choose a so that $e^{-a\tau_{cor}} = .98$, (i.e. $a\tau_{cor} = .02$ or $a = 1$), the contribution of E_1 to (5.2.38) can be neglected and from (5.2.40) we find

$$A_{ij} \doteq E_2 = E[y_i(t) \int_{-\infty}^{t} e^{-a(t-s)}\, y_j(s)\, ds] \qquad (5.2.41)$$

$$= E[y_i(t)\ \bar{Y}_j(t)]$$

where $$\bar{Y}_j(t) = \int_{-\infty}^{t} e^{-a(t-s)}\ y_j(s)\ ds. \qquad (5.2.42)$$

The signal $\bar{Y}_j(t)$ is recognized as the output of a linear first order filter of time constant a^{-1} whose input is the signal $y_j(t)$. Thus in making the replacement $Y(t) = \bar{Y}(t)$ of (5.2.37) we replace the integral of $y(t)$ by a filtered $y(t)$. The filter behaves like an integrator for past inputs up to τ_{cor} time units ago, and then beyond that, tends to "forget the past". Thus $\bar{Y}(t)$ is a stationary signal which behaves like $Y(t)$ as far as evaluating the statistical average (5.2.35) is concerned, and being stationary, the average can be conveniently performed by integrating:

$$A_{ij} = \frac{1}{T} \int_{o}^{T} y_i(t)\ \bar{Y}_j(t)\ dt. \qquad (5.2.42)$$

The integral in (5.2.42) was performed on the analogue computer and is shown in Figure 5.2.5 for the evaluation of A_{21}, A_{11}, and A_{12}, which are estimated as the slopes of the average lines drawn (averaged over 10 seconds). The curve for A_{22} is indistinguishable from the A_{11} curve and is not shown. The fact that the curves are never far from their average slopes attests to the stationarity of the product in the integral (5.2.42), and the fluctuations near the end of the curves (t = 8 to 10) indicate the confidence interval in estimating the slopes. The vertical scale of the graph is arranged so that the estimate of A_{ij} is read directly as the value of the ramp at $t = 10$, and the estimates are

$$A_{21} = 0.061$$

$$A_{11} = A_{22} = .033$$

$$A_{12} = .0055$$

with a confidence interval of about $\pm$ .001 on each estimate. Writing these values in a form suitable for comparison with D_y

of (5.2.33), we have

$$A = \frac{1}{2} \cdot \frac{1}{15} \begin{bmatrix} 1.0 & .17 \\ 1.83 & 1.0 \end{bmatrix} \qquad (5.2.43)$$

The accuracy of the relative magnitudes of the components of this matrix can be checked from the theoretical correlation function of y(t) [88]. A typical component $R_{21}(\tau)$ is shown in Figure 5.2.6, and other components are shifted or transposed versions of the curve shown.

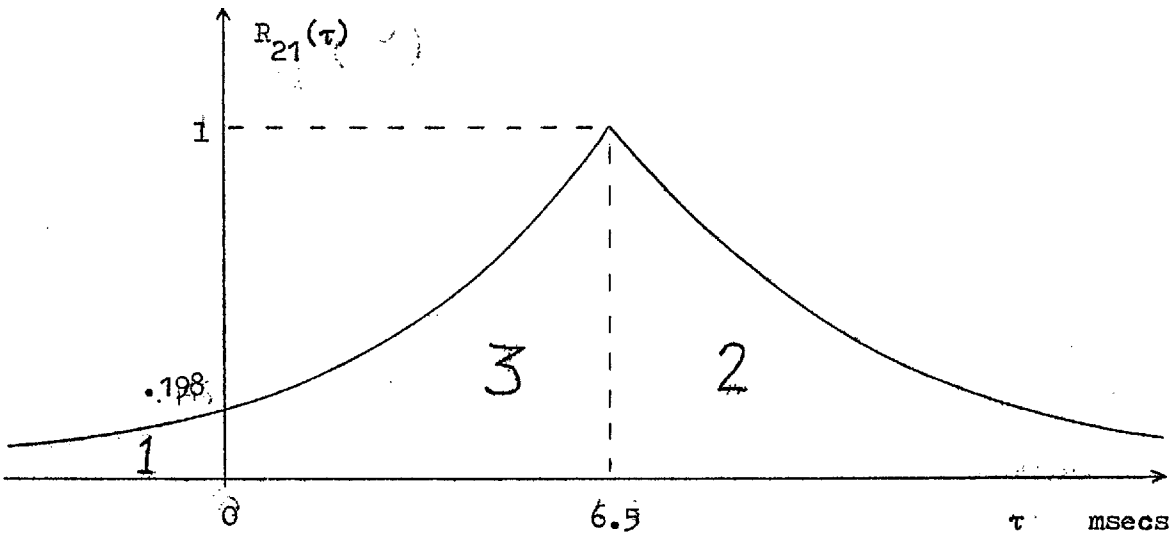


Figure 5.2.6 Normalised Correlation Function of Two Dimensional Noise Source

The curve is very nearly made up of exponential segments whose time constant is that of the filter, 4 msecs., and the relative sizes of A_{12}, A_{11}, and A_{21} are found by the relative areas marked 1, 2, and 2 plus 3. These are 0.198, 1.000, and 1.802. As these values are within the confidence limits we had assigned to the component of the matrix in (5.2.43), the validity of our experimental method of evaluating the components of the characteristic matrix A is verified (apart from a possible scaling factor error . To check

this we would have to know the exact values of many of the analogue computer and PRBS generator components. However, in this section we wish to illustrate the effect of having A_{12} and A_{21} different from D_{12} and D_{21}, and so we will not go into this point).

The Statistics of the Physical Process (5.2.29)

Having found the characteristic matrix of the physical noise $y(t)$, we can evaluate the statistical behaviour of the physical process $X(t)$ of equation (5.2.29). In particular, we are interested in the drift of the process (i.e. $E[\delta X \mid X,t]$), and how it is affected by interchanging $y_1(t)$ and $y_2(t)$.

The noise terms of (5.2.29) contribute a positive drift given by equation (5.2.27) with A_{kl} replacing D_{kl} (c.f. equation (4.1.24), and adding this to the non-random drift in (5.2.29) we get

$$E[\delta X \mid X,t] = [(A_{11} - D_{11})\, X(t) + A_{12} - D_{12}]\, \delta t \qquad (5.2.44)$$

where A is given in (5.2.43) and D in (5.2.26). Thus we have

$$E[\delta X \mid X,t] = -\frac{.03}{30}\, \delta t = -0.001\, \delta t\,. \qquad (5.2.45)$$

Now, if the noises $y_1(t)$ and $y_2(t)$ are interchanged, the system drift becomes

$$E[\delta X \mid X,t] = [(A_{22} - D_{22})\, X(t) + A_{21} - D_{21}\,]\, \delta t$$

$$= \frac{1.63}{30}\, \delta t = 0.054\, \delta t, \qquad (5.2.46)$$

which is quite different from the essentially zero drift of (5.2.45). These figures agree well with the observed drifts in Figure 5.2.4, where the drift of curve 1 is near zero but slightly negative, and the drift of curve 2 was estimated as $0.050\, t$ (curve 3).

The diffusion process $x(t)$ we are trying to simulate has a zero drift, and in fact we see from our method of calculating A

that it was rather fortuitous that the first arrangement of noises gave a near zero drift (A_{12} was only near D_{12} because of the exponential shape of the correlation function, Figure 5.2.6). The only point we wish to make here, is that the characteristic matrix A must, in general, be evaluated for a physical noise process, for unless this matrix is symmetrical, the information in the intensity matrix D is not sufficient to specify the statistics of the simulation. In fact, our primitive method* of calculating D (leading to (5.2.33)) did not even evaluate the intensity correctly, for from (5.2.43),

$$2D = A + A^T = \frac{1}{15}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}. \qquad (5.2.47)$$

The matrix (5.2.47) only has a rank of one, which means that our two-dimensional noise source does not have sufficient degrees of freedom to simulate the two-dimensional white noise $\dot{w}(t)$ of the diffusion process (this confirms our earlier suspicion that our noise source is not a very suitable two-dimensional noise source).

To sum up, if we have a noise source at our disposal which has a symmetric characteristic matrix, then a diffusion process can be simulated by programming the Stratonovich s.d.e. for the diffusion process directly on to the analogue computer, using a properly scaled version of the noise source, which only involves the intensity of the noise, $2D$. If the noise source has an asymmetrical characteristic matrix, correction terms have to be added to the drift of the Stratonovich equation which involves the characteristic matrix of the noise, A. In this case, the Stratonovich equation has no advantage, and in Section 4.4 we give the correction terms pertinent to the Ito form of the equation of the diffusion process.

---

* This method of calculating the intensity $2D$ by evaluating $R(0)$ would only have worked correctly if each component of the correlation function had the same shape.

Thus unless it is known that a noise source has a symmetrical characteristic matrix (e.g. it may be known that the components are all independent of each other), the characteristic matrix A must be evaluated. If the correlation function $R(\tau)$ of the noise is known, then A can be found from the integral definition (4.1.21). If this is not the case, then the characteristic matrix A can be evaluated experimentally by the simple, and seemingly accurate, method given in this Section.

CHAPTER 6

Digital Simulation

Many of the interesting points concerning the relation between ordinary differential equations and stochastic differential equations can be demonstrated on an analogue (or hybrid) computer, and some of these points (particularly those concerning the role of the noise characteristic matrix) have been illustrated by examples in the previous chapter. On a digital computer, we usually use random numbers as a physical noise source when simulating high frequency random phenomena, and so this chapter will not dwell on the choice of noise source or the evaluation of characteristic matrices. Our main interest lies in the choice of digital formulae for solving ordinary differential equations involving the random number noise. We choose two suitable formulae, and discuss their efficiency and convergence rates.

Although many of the techniques used in the data reduction part of digital simulation were conventional, one interesting technique was developed, and is presented in Section 6.3. This was a smoothing technique used in estimating functional solutions from random data, and takes advantage of the efficient representation of functional information afforded by orthogonal expansions.

6.1 Digital Noise and the Form of the Ordinary Differential Equation

Consider the problem of simulating the diffusion process $x(t)$ described by the Ito stochastic differential equation

$$\frac{dx(t)}{dt} = f(x,t) + F(x,t)\,\frac{dw(t)}{dt} \qquad (6.1.1a)$$

on a digital computer. Here $\dot{w}(t)$ is the formal derivative of a

unit parameter, independent component, m-vector Wiener process, and we use this notation as opposed to the more accepted dx, dw, notation to facilitate comparison of equation (6.1.1) with ordinary differential equations. The diffusion process $x(t)$ is also described by the Stratonovich stochastic differential equation

$$\frac{\bar{d}x(t)}{dt} = f(x,t) - \frac{1}{2}\sum_{k,l}^{m} Q_{kl}(x,t) + F(x,t)\frac{\bar{d}w(t)}{dt} \qquad (6.1.1b)$$

where $Q_{kl}(x,t)$ is the n-vector given in equation (A11'b).

The noise $\dot{w}(t)$ in equation (6.1.1) is theoretical white noise with an infinite amplitude and a flat power density spectrum of unity extending to infinity. As mentioned before, such a noise process is not physically realisable, and must be replaced by a physical noise process which is suitable for computing purposes. This turns the stochastic differential equation into an ordinary differential equation.

On a digital computer, approximate solutions of ordinary differential equations are obtained at sample points $n\Delta t$, $n = 0, 1, 2...$ where Δt is a discretization time or step length associated with the numerical formula used to solve the ordinary differential equation. The mechanics of obtaining the solution do not specify the value of the solution between the sample points, nor (except in special cases) is the value of the derivative used between the sample points.

In Chapter 4, we showed that the first order probability density of $x(t)$ was well approximated by an appropriate physical process $X(t)$ provided $\tau_{cor} << \tau_{rel}$ and the physical noise was characterized by the matrix A. We also showed that the transition probability density (or second order probability density) of $x(t)$ was well approximated by that of $X(t)$ for transitions over time increments δt somewhat higher than but approaching τ_{cor}. Connected with this is the fact that the power density spectrum of $x(t)$ is well approximated by that of $X(t)$ for frequencies up to approximately τ_{cor}^{-1} (again provided $\tau_{cor} << \tau_{rel}$, in which case the spectrum will be small at $f = \tau_{cor}^{-1}$ anyway).

Thus for accuracy and spectral resolution considerations, τ_{cor} of the physical noise should be chosen as small as possible, and in any case much less than τ_{rel}. But we noted above that the noise $y(t)$, being a term on the right hand side of the differential equation, is sampled every Δt time units, and so for a given Δt* the physical noise is given a maximum frequency content (or minimum τ_{cor}) by making the samples $y(n\Delta t)$ independent of each other. As $\dot{w}(t)$ is a zero mean Gaussian noise, it is convenient to use pseudo random zero mean Gaussian numbers for the values $y(n\Delta t)$, generated by a computer algorithm such as that mentioned in Section 3.2.6. We describe below how the variance of the Gaussian random numbers is chosen.

We have specified the structure of the physical noise $y(t)$ at the sample points $t = n\Delta t$, but in order to know what ordinary differential equation we are solving, we must know the structure of $y(t)$ between the sample points. The specification of this inter-sample point structure is arbitrary, and we shall assume that the noise $y(t)$ is constant between the sample points. That is

$$y(t) = y(n\Delta t) \quad , \quad n\Delta t \leq t < (n+1)\Delta t. \qquad (6.1.2)$$

The following argument suggests that this choice of structure is a natural one. As the noise $y(t)$ is piecewise constant, the integral of $y(t)$ is piecewise linear. Now, being interested in the properties of $X(t)$ which is an integral function of $y(t)$, we must match the integral of $y(t)$ to the integral of the noise $\dot{w}(t)$ it is replacing, i.e. to the Wiener process $w(t)$. If the Wiener process is specified at $t = n\Delta t$, and no information is given about its behavior between these points, it seems natural to approximate the Wiener process by joining the sampled values $w(n\Delta t)$ by straight lines. The integral of the piecewise constant $y(t)$ does this, and to this extent the choice of the piecewise constant structure is a natural one.

---

* factors concerning the choice of Δt are discussed later.

This matching can be used to choose the variance V of the random numbers $y(n\Delta t)$. The integral of $y(t)$ has increments over time Δt of zero mean value and variance $V\,\Delta t^2$. Then, as the Wiener process $w(t)$ has increments of zero mean and variance Δt, we must choose V so that

$$V = E[y(n\Delta t)^2] = \Delta t^{-1}. \qquad (6.1.3)$$

The analysis of Chapter 4 can be used to show that the ordinary differential equation using the piecewise constant structure of $y(t)$ with the variance (6.1.3) has the same incremental statistics (to the accuracy of Section 4.1) as the diffusion process described by the Stratonovich s.d.e. having the same terms as the o.d.e. For example, consider the simulation of the linear diffusion process given by the Stratonovich s.d.e.

$$\frac{\bar{d}x(t)}{dt} = a\,x(t) + b\,x(t)\,\frac{\bar{d}w(t)}{dt}\ , \qquad (6.1.4a)$$

or by the Ito s.d.e.

$$\frac{dx(t)}{dt} = (a + \tfrac{1}{2}\,b^2)\,x(t) + b\,x(t)\,\frac{dw(t)}{dt}\ . \qquad (6.1.4b)$$

We propose to simulate $x(t)$ by using the o.d.e.

$$\dot{X}(t) = a\,X(t) + b\,X(t)\,y(n\Delta t) \qquad (6.1.5)$$

obtained directly from the Stratonovich equation (6.1.4a). The noise $y(n\Delta t)$ has the piecewise constant structure introduced above, and so for $n\Delta t \leq t < (n+1)\Delta t$ we can write (6.1.5) as

$$\dot{X}(t) = c\,X(t), \qquad (6.1.6)$$

where c is a Gaussian random number of mean a and variance $b^2\,\Delta t^{-1}$. Integrating (6.1.6) over Δt we find

$$\Delta X(n\Delta t) = X((n+1)\,\Delta t) - X(n\Delta t)$$

$$= X(n\Delta t)\,[e^{c\Delta t} - 1]$$

$$= X(n\Delta t)\,[c\Delta t + \tfrac{1}{2}\,c^2\,\Delta t^2 + \ldots\,]. \qquad (6.1.7)$$

Then

$$E[\Delta X(n\Delta t) \mid X(n\Delta t)] = X(n\Delta t)\,[a\Delta t + \tfrac{1}{2}\,b^2\Delta t] + o(\Delta t), \qquad (6.1.8)$$

which, to the indicated error magnitude, is the correct first increment of the diffusion process (6.1.4). Also, from (6.1.7),

$$\Delta X^2(n\Delta t) = X^2(n\Delta t)\,[c^2\,\Delta t^2 + \ldots\,],$$

and

$$E[\Delta X^2(n\Delta t) \mid X(n\Delta t)] = X^2(n\Delta t)\,[b^2\,\Delta t] + o(\Delta t), \qquad (6.1.9)$$

which again, to the indicated error magnitude, is the correct property of the diffusion process (6.1.4).

Thus by arguing along the lines of Section 4.1 we have shown that the ordinary differential equation (6.1.5) is a correct simulation of the diffusion process (6.1.4), an argument which is also demonstrated below for general (non-linear) processes. Thus if a physical process is formed by replacing white noise by piecewise constant noise, the equation we must work from is the Stratonovich s.d.e. of the diffusion process. This point is confirmed by Wong and Zakai [24, 102] who show that such a physical process converges in mean square to the diffusion process given by the term-by-term equivalent Stratonovich s.d.e., as Δt goes to zero, when the increments of the integral of $y(t)$ are related sample pathwise to the increments of the Wiener process (we have only related these increments statistically, and as a result only obtained convergence in distribution - see Section 4.2). If the physical noise $y(t)$ has some other inter-sample point structure, such as a linear interpolation, the equivalent o.d.e. can still be obtained directly from the Stratonovich equation (as the independent white noise source is symmetrical) but the simple scaling method

we have used may no longer be convenient, and the following more general method can be used.

To investigate the scaling of the noise $y(n\Delta t)$ further, let us calculate the characteristic matrix of the noise. The correlation function of the piecewise constant noise has the triangular form shown in Figure 6.1.1, where V is the variance of $y(n\Delta t)$ and Δt is the inter-sample time. For a vector independent piecewise constant noise source, the cross-correlations are all zero, and each component of the noise has the autocorrelation function shown. Then the characteristic matrix is $A = \frac{1}{2} V\Delta t\ I$ (6.1.10)
and $A + A^T = V\ \Delta t\ I.$ (6.1.11)

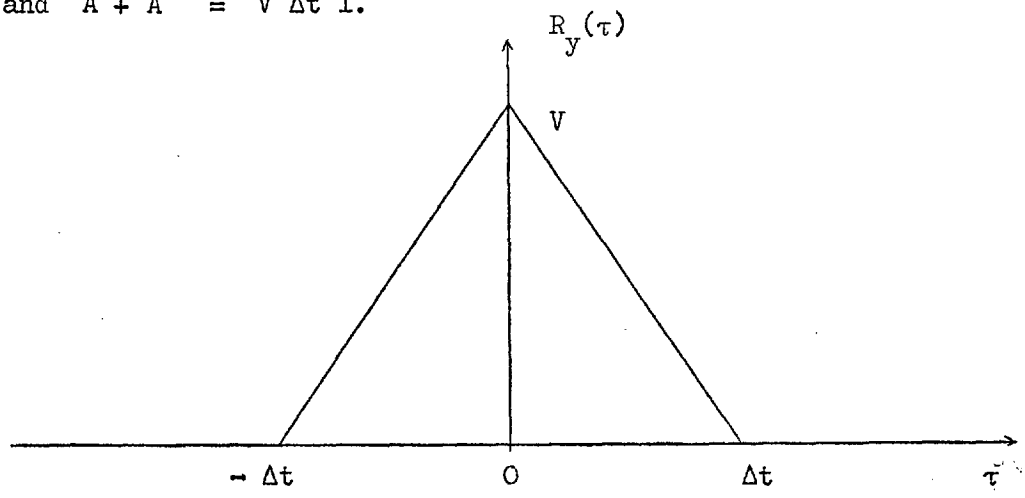


Figure 6.1.1 Correlation Function of Piecewise Constant Noise

Now consider the simulation of the diffusion process (6.1.1) by computing with the o.d.e. obtained from the Stratonovich s.d.e. (6.1.1b):

$$\dot{X}(t) = f(X,t) - \tfrac{1}{2} \sum_{k,l}^{m} Q_{kl}(X,t) + F(X,t)\ y(t), \quad (6.1.12)$$

using the piecewise constant noise $y(t)$ discussed above. To scale the noise correctly (i.e. choose its variance V), we evaluate the second order conditional increment of the physical process (6.1.12). From Section 4.1.2,

$$E[\Delta X^2 \mid X,t] \doteq F\ (A + A^T)\ F^T\ \Delta t. \quad (6.1.13)$$

But from equation (6.1.1) the diffusion process $x(t)$ has

$$E[\Delta x^2 \mid x,t] = F F^T \Delta t + o(\Delta t), \qquad (6.1.14)$$

and so from (6.1.11) we must specify the noise variance as

$$V = \Delta t^{-1} \qquad (6.1.15)$$

With this choice of noise variance, we see from Section 4.1.1 that the approximate first order conditional increment of the physical process (6.1.12) is

$$E[\Delta X \mid X,t] \doteq f(X,t)\,\Delta t \qquad (6.1.16)$$

which agrees with that of the diffusion process (6.1.1). Thus to simulate the diffusion process (6.1.1) on a digital computer using a step length of Δt time units, we solve the ordinary differential equation (6.1.12) which is obtained directly from the Stratonovich s.d.e. of the diffusion process by replacing the white noise $\frac{\bar{d}w(t)}{dt}$ by a physical noise $y(t)$ which consists of independent piecewise constant segments Δt long, each with a variance of Δt^{-1}.

6.2 Digital Solution of the Ordinary Differential Equation and Convergence to the S.D.E.

Having chosen a physical process (o.d.e.) to model the diffusion process (s.d.e), we must consider the second problem of choosing a suitable algorithm for obtaining an approximate solution to the o.d.e. on the digital computer. As this second operation also involves an error (which depends on Δt), there are two steps in the convergence of the digital computer simulation to the diffusion process, and these are shown diagrammatically in Figure 6.2.1.

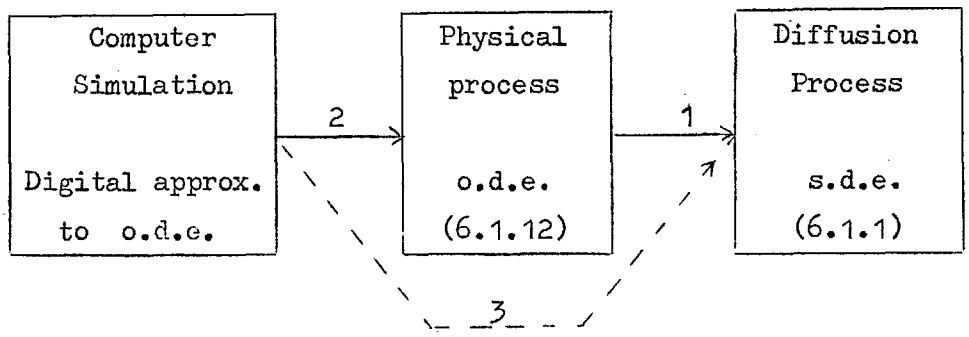


Figure 6.2.1: Steps in the Convergence of a Digital Computer Simulation to a Diffusion Process

In the analogue computing in Chapter 5, we had assumed that the physical process was simulated exactly, and so the question of convergence along step 2 in Figure 6.2.1 did not arise. Then we were only concerned with step 1, and in Section 4.1 we presented an analysis which showed what parameters affected the convergence in distribution along step 1. In this section, we review the considerations affecting the convergence along step 1, and show how the choice of digital computer algorithm affects the convergence along path 2. Using a simple formula for simulating s.d.e.'s directly, the formulation of the physical process in Figure 6.2.1 can be bypassed, and convergence along path 3 can be considered.

6.2.1 Convergence of the Physical Process to the Diffusion Process (Path 1)

In Section 4.1 we evaulated approximate expressions for the first and second conditional increments of a physical process over a time increment δt, and showed how these parameters specified an "equivalent" diffusion process whose statistical behaviour was approximately the same as that of the physical process. In particular,

we showed that, for a fixed δt, the error in evaluating the expression $E[\delta X \mid X,t]$ was proportional to τ_{cor} (c.f. A1 and A3), and that, discounting this error, the expression was the same as that of the equivalent diffusion process (to the stated δt accuracy). This implies that the expected value of the increment in $X(t)$ converges to that of $x(t)$ with an error term proportional to τ_{cor}. As the increment is proportional to δt, then the statement above implies that the linear error between the expected values of $X(t)$ and $x(t)$ (from a common starting point) is also $O(\tau_{cor})$.

This point can be confirmed by following an analysis of Franklin [107]. He considers a sampled data model of a diffusion process with additive noise and considers the convergence of expected values of the sampled data model to those of the diffusion process and finds the convergence to be $O(\Delta t)$, i.e. $O(\tau_{cor})$. That is, he shows

$$E[\phi(X(t))] = E[\phi(x,t))] + O(\tau_{cor}) \text{ as } \tau_{cor} \longrightarrow 0, \quad (6.2.1)$$

for a sufficiently smooth function $\phi(\cdot)$. His proof can be extended to non-additive noise, although the transformations he uses in the proof then become time-varying, with a considerable loss of simplicity.

Franklin's digital model is, in effect, a direct simulation of the diffusion process without specifying the particular physical process that we have in Section 6.1, and so his convergence is along path 3 in Figure 6.2.1. Nevertheless, his digital model is a physical process and can be considered to use the piecewise constant noise we use, and so the convergence he proves is closely related to the convergence along path 1 in Figure 6.2.1. Indeed we see later that his digital model is related to our physical process by being a suitably biased forward difference method of solving the o.d.e. of our physical process, and so the error in simulation must be at least as large as that of our physical model. By this token, the convergence along path 1 of Figure 6.2.1 for expected values of functions of $X(t)$ to those of $x(t)$ can be no worse than $O(\tau_{cor})$.

The convergence (6.2.1) is difficult to demonstrate experimentally, for it involves forming a statistical estimate of a statistical parameter (the l.h.s. of (6.2.1)) which is itself subject to sampling error. Instead it is more convenient to investigate sample path convergence on the computer, which is the kind of convergence considered by Clark [22].

As our physical noise is Gaussian and stationary, Clark's results can be applied to the convergence along path 1 in Figure 6.2.1 (in [22, Chapter 4] he considers such a piecewise constant noise process). He shows that the mean square error in sample paths is of order $O(\tau_{cor})$, that is

$$E[(X(t) - x(t))^2] = O(\tau_{cor}). \qquad (6.2.2)$$

But the diffusion process $x(t)$ (and the limiting $X(t)$ process) is a process whose increments $x(t+\Delta t) - x(t)$ are of the same order as the square of the increments. Thus it seems plausible that we can infer from (6.2.2) that the linear error in sample path is also $O(\tau_{cor})$

$$E[\,|X(t) - x(t)|\,] = O(\tau_{cor}), \qquad (6.2.3)$$

and not $O(\tau_{cor}^{\frac{1}{2}})$ as might be suggested by taking the square root of (6.2.2). The following experimental evidence supports the convergence rate (6.2.3).

Experimental Test of Sample Path Convergence

Consider the simulation of the diffusion process given by the Ito s.d.e.

$$dx(t) = -x(t)\,dt + \alpha(1 + x(t))\,dw(t), \qquad (6.2.4a)$$

which is another example studied by Astrom [26, ex. 2]. From (6.1.1b) the Stratonovich s.d.e. for this process is

$$\bar{d}x(t) = -[x(t) + \tfrac{1}{2}\alpha^2(1 + x(t))]\,dt + \alpha(1 + x(t))\,\bar{d}w(t), \qquad (6.2.4b)$$

and from (6.1.12), the o.d.e. using piecewise constant noise for the physical process whose statistics approximate those of the diffusion process (6.2.4) is

$$\dot{X}(t) = -[X(t) + \tfrac{1}{2}\alpha^2(1 + X(t))] + \alpha(1 + X(t))\,y(t). \qquad (6.2.5)$$

As $y(t)$ is constant in the interval $t = [n\Delta t, (n+1)\Delta t)$, equation (6.2.5) can be rearranged to give the following first order linear constant coefficient ordinary differential equation

$$\dot{X}(t) = -[1 + \tfrac{1}{2}\alpha^2 - \alpha\, y(n\Delta t)]\,X(t) - \tfrac{1}{2}\alpha^2 + \alpha\, y(n\Delta t)$$

$$= -\mu_1(n\Delta t)\,X(t) + \mu_2(n\Delta t).$$

This equation can be solved explicitly in the Δt interval to give

$$X((n+1)\Delta t) = \frac{\mu_1(n\Delta t)}{\mu_2(n\Delta t)} + \left[X(n\Delta t) - \frac{\mu_2(n\Delta t)}{\mu_1(n\Delta t)}\right] e^{-\mu_1(n\Delta t)\Delta t}. \qquad (6.2.6)$$

Now to test the convergence of the sample paths of the physical process (6.2.5) to those of the diffusion process (6.2.4) as a function of Δt, we must consider a particular realisation of the Wiener process $w(t)$. Then for a given Δt, we choose $y(n\Delta t)$ as

$$y(n\Delta t) = \Delta t^{-1}\,[w((n+1)\Delta t) - w(n\Delta t)] \qquad (6.2.7)$$

so that the integral of $y(t)$ equals $w(t)$ at the sample points, and the continuous curve converges to $w(t)$ in the mean as Δt goes to zero. That is, as we refine Δt, the integral of $y(t)$ more and more closely resembles the particular realisation of $w(t)$ we have chosen, and as a result, we are in a position to compare the convergence of the sample paths of $x(t)$ to $X(t)$. The piecewise constant $y(t)$ (6.2.7) is the same as that we had chosen earlier (6.1.2), except that before we had only matched the integral

of $y(t)$ to the Wiener process statistically.

Clark [22] and Wong and Zakai [24, 102] show that the sample paths of $X(t)$, (6.2.5), converge in the mean to the sample paths of $x(t)$, (6.2.4), for the choice of physical noise (6.2.7). Having accepted this, our interest lies in testing the rate of convergence, and this can be done by computing trajectories using the formula (6.2.6) for successive refinements of Δt.

This was done for a particular Wiener process $w(t)$ record 10 seconds long. The values $X(n)$, $n = 1, 2, \ldots 10$, were computed using (6.2.6, 7) for Δt successively equal to 1, 0.1, 0.01 and 0.001. The trajectory with the finest subdivision, $\Delta t = 0.001$, was taken as $x(t)$ and the other trajectories were compared with it at the integral values of time, and the error function on the l.h.s. of (6.2.3) was evaluated. Table 6.2.1 gives these error values for three different noise standard deviations.

Each entry in the table is an average of 10 points and so the standard deviation of each entry is approximately $10^{-\frac{1}{2}}$ or 0.33 times the entry. Within these limits, the entries in the table (scanning horizontally) are proportional to Δt for each of the three noise ratios tested.

This is clearly seen from the linear relationship of the log-log plot of Figure 6.2.2, where each line drawn has a slope of 1, indicating that errors are proportional to Δt^1. Then as $\tau_{cor} = \Delta t$ for the piecewise constant noise $y(n\Delta t)$, we have experimentally verified the convergence rate $O(\tau_{cor})$ of (6.2.3).

Remarks

The contents of this section (6.2.1) are an early look at a new and rather complex field - the convergence of a particular ordinary differential equation to a stochastic differential equation. For this reason, the comments we have made must be regarded as being somewhat speculative in nature, and much more work remains to be done on both theoretical and experimental lines to establish meaningful and practical norms of convergence.

As a problem in numerical analysis, the difficulty is that we are dealing with a random equation, but above all, an equation which has an unusual incremental property when the "Δt tends to zero" limiting properties are studied. This is the property that the increments Δx and the square of the increments $(\Delta x)^2$ are both of the same order of magnitude, $O(\Delta t)$, when conditional expectations are taken and Δt approaches zero (the property $E[\Delta x \mid x,t] = O(\Delta t)$ is due to the non-random part of the process, and the property $E[(\Delta x)^2 \mid x,t] = O(\Delta t)$ is due to the random part of the process). As the equation for $x(t)$ is random, error terms which result when we try and simulate the equation only have a meaning when expectations are taken, but because of the unusual order of magnitude property of the increments of such processes, the usual concepts of numerical analysis cannot be applied to stochastic equations. New outlooks, such as that provided by Franklin [107], are needed. Our analysis of Section 4.1, and that of Stratonovich [21], is similar in philosophy to Franklin's, and all three approaches illustrate the difficulty of making quantitative statements about the error of diffusion approximations.

Thus the error analysis of diffusion approximations is still in an early development stage, and although we have made some progress, much work remains to be done. At this stage, numerical experimentation must be relied upon to determine the reliability of solution, and in this vein, the theoretical and experimental results of the next section comparing the order and efficiency of digital integration formulae are the most concrete results of this chapter.

6.2.2 Discrete Approximation to the Ordinary Differential Equation (Path 2)

We now turn our attention to the problem of obtaining an approximate solution to the ordinary differential equation (6.1.12) on a digital computer. We are now interested in the convergence along path 2 in Figure 6.2.1. This is a more conventional problem

in numerical analysis than that of the previous section, yet the particular o.d.e. (6.1.12) with piecewise constant noise has some special properties connected with the lack of smoothness of the r.h.s. which bear special consideration.

The Order of Digital Approximations to O.D.E.'s

Consider the problem of obtaining an approximation on a digital computer to the solution of the o.d.e.

$$\dot{X}(t) = G(X)\, y(t), \tag{6.2.8}$$

which is the same as equation (6.1.12) with the non-random terms removed. On the computer, approximate solutions to (6.2.8) are obtained over discrete time steps Δt long. Over a single time increment, the increment in the true solution of (6.2.8) can be written as the Taylor series

$$\begin{aligned} \Delta X(t) &= X(t + \Delta t) - X(t) \\ &= \dot{X}(t)\,\Delta t + \tfrac{1}{2}\ddot{X}(t)\,\Delta t^2 + \tfrac{1}{3!}\dddot{X}(t)\,\Delta t^3 + \ldots \end{aligned} \tag{6.2.9}$$

provided all the higher derivatives on the r.h.s. exist. These derivatives are all evaluated at t, the beginning of the Δt interval.

To obtain an approximation to $X(t + \Delta t)$ given $X(t)$ which is accurate to an error term proportional to Δt^{n+1}, we must use a digital formula for $\Delta X(t)$ which agrees with (6.2.9) for all terms on the r.h.s. of (6.2.9) up to and including the Δt^n term. Then provided the $(n+1)$:th derivative of $X(t)$ is bounded in the interval $[t, t+\Delta t)$, the error in $\Delta X(t)$ will be of order $O(\Delta t^{n+1})$. Such a formula is called an n:th order formula.

The simplest formula uses only the first term on the r.h.s. of (6.2.9), and is the first order Euler (or forward difference) formula:

$$\Delta X(t) = \dot{X}(t)\,\Delta t = G(X(t))\,y(t)\,\Delta t. \qquad (6.2.10)$$

The formula is particularly simple to program, as the r.h.s., Gy, is known explicitly. However, the accuracy of the Euler formula is often too limited for a given amount of computing, and so higher order formula are usually used.

Multi-step Formulae

Higher order formulae which directly use the terms on the r.h.s. of (6.2.9) are sometimes inconvenient, as the higher derivatives $\ddot{X}(t)$, $\dddot{X}(t)$, etc., may not be explicitly available. They then have to be obtained by differencing methods, and an efficient way of doing this is to express these higher derivatives as functions of $X(t')$ and $\dot{X}(t')$, where $t' = t - \Delta t$, $t - 2\Delta t$, etc. These quantities are readily available as they have already been evaluated at previous time steps. Such formulae are called multi-step formulae, as they utilize the value of X and $\dot{X}$ at points other than in the current time step $[t, t + \Delta t)$.

For example, we can write $\ddot{X}(t)$ as

$$\ddot{X}(t) \doteq [\dot{X}(t) - \dot{X}(t - \Delta t)]\,\Delta t^{-1}, \qquad (6.2.11)$$

from which we get the second order formula

$$\Delta X(t) = [\tfrac{3}{2}\,\dot{X}(t) - \tfrac{1}{2}\,\dot{X}(t - \Delta t)]\,\Delta t. \qquad (6.2.12)$$

Now for such a formula to have second order accuracy, we do not require that $\dddot{X}(t)$ be bounded in $[\Delta t, t + \Delta t)$ as we would for a second order formula based on the series (6.2.9). Instead we require that $\ddot{X}(t)$ be bounded over the <u>double</u> interval $[t - \Delta t, t + \Delta t)$ so that the substitution (6.2.11) is correct up to second order terms in (6.2.12). But our particular o.d.e. (6.2.8) does not have this property, as $y(t)$ is only continuous over single time steps, and so the formula (6.2.12) is not a second order formula when applied to the o.d.e. (6.2.8).

This is true in general for multi-step formulae. The accuracy of multi-step formulae depend on the smoothness of the derivative $\dot{X}(t)$ over more than one time step Δt. But equations such as (6.2.8) which involve a piecewise constant noise $y(t)$ have derivatives $\dot{X}(t)$ which are discontinuous at the sample points $t = n\,\Delta t$. Then the usual error analysis does not apply to multi-step formulae applied to such equations, and the error is indeterminate. Thus multi-step formulae cannot be used to obtain high order solutions to equations involving piecewise constant noise.

From another viewpoint, the random numbers $y(n\Delta t)$, $n = 0, 1, 2 \ldots$ are independent of each other, and so values of $\dot{X}(t)$ of (6.2.8) over successive time intervals are uncorrelated with each other. But multi-step formulae using expansions such as (6.2.11) use the relation between $\dot{X}(t)$ at previous successive time steps to estimate higher derivatives needed in the higher order formulae. This approach is invalid in the present example, for as successive values of $\dot{X}(n\Delta t)$ are uncorrelated, they contain no information on the higher derivatives in their neighbourhood.

This criticism does not apply to single step formulae, as they do not use any past information of $\dot{X}(t)$ when obtaining the solution in a particular time step. In this sense, single step formulae are conceptually similar to the Markov Processes we are simulating, for the statistics of the future solution step $\Delta X(t)$ are given entirely by the present state $X(t)$ and not by any past state information.

Single Step Formulae

The Order of Single Step Formulae when Applied to the O.D.E. with Piecewise Constant Noise Terms

With this justification, we restrict our attention to single step formulae for solving o.d.e.'s such as (6.2.8) involving piecewise constant noise terms on the r.h.s. We have already given the simplest single step formula, the Euler first order formula (6.2.10),

and of the various higher order single step formulae, those based on the Runge-Kutta method are the most commonly used [108, p.211; 55, p.195].* In this section we consider the order of convergence of single step formulae when applied to the o.d.e. (6.2.8).

Considering equation (6.2.8), we replace the piecewise constant noise $y(t)$ by a Gaussian $(0, \Delta t^{-1})$ random number c in the interval $[t, t+\Delta t)$. Then in equation (6.2.9) we have the following values for the derivatives:

$$\dot{X}(t) = c\, G(X)$$

$$\ddot{X}(t) = c\, G_x(X)\, \dot{X} = c^2 G_2(X)$$

$$\dddot{X}(t) = c^2 [G_2(X)]_x \dot{X} = c^3 G_3(X),$$

and in general

$$X^{(n)}(t) = c^n G_n(X), \qquad (6.2.13)$$

where $G_n(X)$ is a function depending on $G(X)$ and its higher derivatives, but <u>not</u> depending on c. The we can write (6.2.9) as

$$\Delta X(t) = G(X)\, c\Delta t + \tfrac{1}{2} G_2(X) c^2 \Delta t^2 + \frac{1}{3!} G_3(X) c^3 \Delta t^3 + \ldots \qquad (6.2.14)$$

and we know from our earlier discussion that for a particular constant c, an n:th order formula computes $\Delta X(t)$ correctly up to terms involving Δt^n in equation (6.2.14).

Now let us look at the order of convergence of such formulae. In normal concepts of convergence, the parameters of the o.d.e. such as c and the functions $G_n(X)$ in (6.2.14) are kept constant, and we evaluate how the error in $\Delta X(t)$ behaves when Δt is

---

* The discussion in the rest of this section pertaining to the Runge-Kutta formula apply equally well to other higher order single step formulae, except that the efficiencies of each formula will be different.

refined. However, in our context, the o.d.e. (6.2.8) is meant to simulate a diffusion process, and as reasoned in Section 6.1, we should make the noise $y(t)$ as high frequency as possible, and this is done by assigning a new random value to c every Δt.* Then for this special equation, our concept of convergence must be altered to allow c to depend on Δt, as Δt is refined.

Then the dependence of the various terms in equation (6.2.14) on Δt is not the same as the Δt^n factor which explicitly appears in (6.2.14). However, the digital formula still computes $\Delta X(t)$ to an accuracy depending on those factors Δt^n which do explicitly appear in (6.2.14), as the formula is derived from the Taylor series expansion on which (6.2.14) is based.

To be more precise, we look at two kinds of convergence: sample path convergence and statistical convergence. In sample path convergence we are interested in the magnitude of the error in computing $\Delta X(t)$. But $\Delta X(t)$ and its errors are statistical quantities, and so we must take the expected value of the magnitude of the error in order to obtain a useful error norm.

If the Taylor series expansion (6.2.14) is truncated after the Δt^n term, the resultant error in evaluating $\Delta X(t)$ can be expressed as

$$\text{Error}_n = \frac{1}{(n+1)!} G_{n+1}(X') \, c^{n+1} \, \Delta t^{n+1}, \qquad (6.2.15)$$

where $X' = X(t')$ for some t' in the interval $[t, t+\Delta t]$, provided that $G_{n+1}(\cdot)$ is bounded in the interval (unless the contrary is specifically mentioned, we always assume our Taylor series expansions are validated by this sort of bound). Then the error of a general n:th order formula is proportional to (6.2.15), and we may write it as

$$\text{Error}_n = \bar{G}_{n+1}(X') \, c^{n+1} \, \Delta t^{n+1}, \qquad (6.2.16)$$

where $\bar{G}_{n+1}(X')$ is again some bounded function independent of c.

The expected magnitude of the sample path error is then $E\,|\text{Err}_n|$.

---

* This means that as Δt is refined, the o.d.e. in the middle box of Fig. 6.2.1 varies. In this sense, the convergence along path 2 is not a normal concept of convergence.

To evaluate this we note that c is a zero mean Gaussian random variable, independent of the other quantities in (6.2.16), with the following properties:

$$E\left|c\right| = (2/\pi)^{\frac{1}{2}} \Delta t^{-\frac{1}{2}},$$

$$E\left|c^2\right| = \Delta t^{-1},$$

$$E\left|c^3\right| = (8/\pi)^{\frac{1}{2}} \Delta t^{-1.5}$$

$$E\left|c^4\right| = 3\,\Delta t^{-2},$$

and in general

$$E\left|c^n\right| = \text{ISP} = \Delta t^{-n/2} . \quad * \qquad (6.2.17)$$

Then from (6.2.16) and (6.2.17) we have

$$E\left|\text{Err}_n\right| = \text{ISP} = \Delta t^{\frac{n+1}{2}} . \qquad (6.2.18)$$

Now for digital formulae applied to smooth o.d.e.'s we noted that the order of the formula was defined to be one less than the power of Δt appearing in the truncation error. By analogy, we define the order of sample path convergence of formulae applied to o.d.e.'s involving piecewise constant noise as being one less than the power of Δt in (6.2.18). This new order of convergence of digital formulæ is shown in the last column of Table 6.2.2., where the first two columns are the normal definition of order of the formula and the power of Δt in (6.2.18).

We now look at the <u>statistical convergence</u> of the digital formulae. In simulation exercises, we are interested in evaluating functions such as $E[\phi(X)]$ which we estimate by simulating many trajectories of $X(t)$ and forming the appropriate average. For

---

* = ISP = denotes "is proportional to"

certain functions $\phi(\cdot)$ we can tolerate an error in $X(t)$ [or $\Delta X(t)$] as long as it is unbiased, and so for statistical convergence (or convergence in distribution) we take as an error norm

$$\left| E\ (Err_n) \right| \quad , \tag{6.2.19}$$

which is the magnitude of the ensemble average of the error in $\Delta X(t)$. Thus in contrast to the sample path error, we are now interested in the size of the expected error instead of the expected value of the size of the error.

We now remove the absolute value signs in (6.2.17) and find that

$$E(c^n) = 0 , \qquad n \text{ odd} ,$$

$$\text{and} \qquad E(c^n) = \text{ISP} = \Delta t^{-n/2}, \qquad n \text{ even} . \tag{6.2.20}$$

If n is odd ($n+1$ is even) we then obtain from (6.2.16) the error norm (6.2.19) as

$$\left| E\ (Err_n) \right| = \text{ISP} = \Delta t^{\frac{n+1}{2}} , \quad n \text{ odd}, \tag{6.2.21a}$$

but if n is even, the expression becomes zero. But the error term (6.2.15) of the truncated Taylor series is a composite error term which includes the error of all the higher order terms which have been truncated, and if the Δt^{n+1} term has a zero mean value because n is even, then the Δt^{n+2} term does not have a zero mean value. Then the error can be written as

$$\left| E\ (Err_n) \right| = \text{ISP} = c^{n+2}\, \Delta t^{n+2}, \quad n \text{ even},$$

and from (6.2.20), this becomes

$$\left| E\ (Err_n) \right| = \text{ISP} = \Delta t^{\frac{n+2}{2}} , \quad n \text{ even}. \tag{6.2.21b}$$

The resultant order of the formula in the sense of statistical convergence is given in Table 6.2.3. This order is higher than that

of sample path convergence (Table 6.2.2) by one half, for digital formulae which are normally of even order (such as the 4th order Runge-Kutta (RK)).

In summary, we see that when solving o.d.e.'s such as (6.2.8) which have terms which depend on Δt (the variance of $y(t)$ equals Δt^{-1}), the order of the truncated terms in the Taylor series (6.2.9) is different from the power of Δt which appears in (6.2.9). The result of this is that digital formulae which are "order n" when applied to smooth o.d.e.'s become less than n:th order when applied to o.d.e.'s involving piecewise constant noise. The new order depends on whether we are considering sample path error or statistical (ensemble average) error, and is given in Tables 6.2.2, 3.

The Euler Formula

From Tables 6.2.1 , 2, we see that the Euler formula when applied to (6.2.8) has an error which is proportional to Δt - that is, it is a zeroth order formula. By normal o.d.e. standards, this means that the formula is not consistent. That is, as Δt goes to zero, the solution of the discrete formula does not tend to the solution of the continuous equation (6.2.8). In that case, the Euler formula cannot be used as a digital approximation to the o.d.e. along path 2 of Figure 6.2.1, when simulating diffusion processes. Let us look at this point further, and see if we can alter the Euler formula so that it can be used to simulate diffusion processes.

Consider the example of simulating the diffusion process (6.1.4) by applying the Euler formula to the o.d.e. (6.1.5). Then applying (6.2.10) to (6.1.6), we have

$$\Delta X(t) = X(t)\, c\, \Delta t, \tag{6.2.22}$$

where c is a Gaussian $(a, b^2\Delta t^{-1})$ random number. Taking the conditional expectation of (6.2.22) we have

$$E\left[\Delta X \mid X(t)\right] = X(t)\left[a\, \Delta t\right] \tag{6.2.23}$$

which is not the correct property (6.1.8) of the diffusion process (6.1.4). This illustrates that the Euler approximation (6.2.22) does not converge to the solution of the o.d.e. (6.1.5) which does have the property (6.1.8). This is because the Euler formula has omitted the Δt term $\frac{1}{2} b^2 \Delta t\, X(t)$ in (6.1.8) which came from the Δt^2 term of (6.1.7).

Thus in order that the Euler formula can be used to simulate diffusion processes, the conditional expectation of $\Delta X(t)$ must be obtained correctly. We note that the diffusion process $x(t)$ described by the Ito s.d.e. (6.1.1 a)

$$\frac{dx(t)}{dt} = f(x,t) + F(x,t)\, \dot{w}(t)$$

has the increment

$$E\left[\Delta x \,|\, x(t)\right] = f(x,t)\, \Delta t + o(\Delta t),$$

and that the Euler approximation to the solution of the o.d.e.

$$\dot{X}(t) = f(X,t) + F(X,t)\, y(t) \qquad (6.2.24)$$

has the increment

$$E\left[\Delta X \,|\, X(t)\right] = f(X,t)\, \Delta t + o(\Delta t), \qquad (6.2.25)$$

when $y(t)$ is a zero mean signal independent of $X(t)$. Also if $y(t)$ is a piecewise constant process with a variance of Δt^{-1} then the Euler approximation to the o.d.e. (6.2.24) has the second conditional increment

$$E\left[\Delta X^2 \,|\, X(t)\right] = F\, F^T (X,t)\, \Delta t + o(\Delta t) \qquad (6.2.26)$$

which is the correct property of the diffusion process.

We thus conclude that if we wish to simulate the diffusion process (6.1.1) on a digital computer using the Euler formula,

we must apply the Euler formula to the o.d.e. (6.2.24) which is obtained directly from the Ito s.d.e. (6.1.1a) by replacing the unit parameter white noise $\dot{w}(t)$ by a piecewise constant noise $y(t)$ which has a zero mean value and a variance of Δt^{-1}. From the consideration of the first two incremental properties (6.2.25, 26) we do not require that $y(t)$ be Gaussian, but as the signal $\dot{w}(t)$ it is replacing is Gaussian, better accuracy will be obtained if $y(t)$ is Gaussian (particularly with regard to the accuracy of the higher moments).

Remarks

Referring to Figure 6.2.1, the simulation of a diffusion process by the Euler method does not consist of the paths of convergence shown as 1 and 2. This is because the physical process (6.2.24) does not converge to the diffusion process (6.1.1) (in the sense of Section 4.2), and also because the Euler solution does not converge to the solution of the physical process (6.2.24). Thus the Euler simulation method can be considered as a direct simulation of the diffusion process, as shown along path 3 of Figure 6.2.1.

However, the accuracy of the Euler method (in obtaining the property (6.2.25), for example) depends upon the smallness of the truncation error of the Euler formula applied the o.d.e. (6.2.24), and depends upon the inequality $\tau_{cor} << \tau_{rel}$ being satisfied (or then the properties (6.2.25, 26) would not ensure accurate simulation). Thus the convergence along path 3 of Figure 6.2.1 entails the same considerations as the convergence along paths 2 and 1.

The use of the Euler formula to simulate stochastic differential equations has been known for some time, and is sometimes called the Maruyama approximation to the s.d.e. (see [102], or [109] for Maruyama's paper). It is the method used by Astrom [26] and Franklin [107], and as far as the author knows, no other method for the digital simulation of diffusion processes has been reported.

Higher Order Formulae

In [26], Astrom simulates a diffusion process using the Euler method in which he has to use a very small step length Δt. The impetus for the work of this section came from this paper, with the hope of showing that stochastic differential equations could be simulated on a digital computer more efficiently using higher order formulae.

Referring to Figure 6.2.1, we know that the exact solution of the o.d.e. (6.1.12) is a proper simulation of the diffusion process (6.1.1) as long as $\tau_{cor} << \tau_{rel}$ (path 1). The question now is: how accurately must we solve the o.d.e. (6.1.12) on the digital computer in order that the approximate solution so obtained is also a proper simulation of the diffusion process (6.1.1) (path 2 is now also involved)?

The answer is in two parts. Returning to our example (6.1.4), we see that to obtain the correct first incremental property (6.1.8), the o.d.e. (6.1.6) must be solved correctly up to and including second order terms in the expansion (6.1.7). That is, our digital computing formula must use the $\frac{1}{2} c^2 \Delta t^2$ term in (6.1.7) so that the bias $\frac{1}{2} b^2 \Delta t$ appears in (6.1.8). It is easily verified that this is a general rule: in order for the convergence along path 2 of Figure 6.2.1 to be consistent, we must use a digital formula which is at least of second order when referred to smooth o.d.e.'s. Referring to Tables 6.2.2, 3 we see that this ensures that the order of the formula when applied to the o.d.e. with piecewise constant noise is greater than zero (see third column), which ensures its consistency.

Secondly, we must choose a step length Δt which is low enough so that the incremental properties (6.2.25, 26) are obtained with sufficient accuracy. This refers to the errors in path 2 of Figure 6.2.1, and is a point which is best investigated experimentally. An example is given below.

To sum up, if we wish to simulate the diffusion process (6.1.1) on a digital computer, there are two possibilities. Firstly, we may use simple forward differences (the Euler formula), working directly from the Ito s.d.e. (6.1.1a) and replacing $\dot{w}(t)$ by Gaussian $(0, \Delta t^{-1})$ random numbers. Secondly, we may form an o.d.e. from the Stratonovich s.d.e. (6.1.1b) replacing $\dot{w}(t)$ by piecewise constant noise with a random, Gaussian $(0, \Delta t^{-1})$ amplitude, and solve the resulting o.d.e. on the computer using a single step formula which is at least a second order formula. The following experimental results illustrate that for some diffusion processes, the first method is more efficient in respect of computing time, and for others the second approach is better.

Euler vs. Runge-Kutta Formulae

Let us compare the efficiency of the Euler method and the o.d.e. method (using a fourth order Runge-Kutta formula) of simulating the diffusion process (6.1.1). Two points are noticed:

(a) If $F(x,t) = 0$ in (6.1.1), the diffusion process reduces to a deterministic process, whose simulation is a classical numerical analysis problem. Then the Euler method has an $O(\Delta t^2)$ error and the Runge-Kutta (RK) method has an $O(\Delta t^5)$ error, and it is well accepted that the RK method is more efficient of computing time for all but the most trivial examples, even though the RK formula requires about four times as much computation as the Euler formula for a given Δt.

(b) If $f(x,t) = 0$ and $F(x,t) =$ constant, the diffusion process reduces to a Wiener process, whose statistics are represented exactly (at $t = n\Delta t$) by adding Gaussian random numbers. That is, we use the o.d.e.

$$\dot{X}(t) = K\, y(n\,\Delta t)$$

which both the Euler and RK formulae compute exactly for any Δt. Then as the Euler formula is simpler, it is more efficient.

However, general diffusion processes fall in between these two extremes of a purely deterministic process and a Wiener process, and either method may be more efficient.

Some computational experience has indicated that the choice of methods generally depends on the relative sizes of the drift term $f(x,t)$ and the random term $F(x,t)\,\dot{w}(t)$ in equation (6.1.1). For low noise values, the RK method was more efficient, and as the relative size of the random term was increased, there came a "break-even" point, beyond which the Euler method was more efficient. However, this effect depended on the particular form of each equation studied, and no general rule could be stated. Numerical experimentation has to be relied upon to determine the more efficient method (and best Δt) for each particular example. Below, we discuss an example of Astrom ([26], Figure 6, Table 2; or our equation (6.2.4) with $\alpha = 2^{\frac{1}{2}}$) and show that the RK method combined with a particular choice of Δt is the more efficient method for this example.

To test the efficiency of computing formula, we shall begin by duplicating the results of Astrom. We simulate the o.d.e.

$$\dot{X}(t) = -X(t) + 2^{\frac{1}{2}}\,(1 + X(t))\,y(t) \qquad (6.2.27)$$

(obtained directly from the Ito s.d.e.) using forward differences over a step length of $\Delta t = 0.002$, and estimate the steady state density $P(x)$ by averaging over 50,000 adjacent solution points. The density $P(x)$ was estimated in the interval $|x| < 1$ by quantising x in 0.05 steps, and the root mean square error was found over these 41 estimates (Astrom [22, eqn. 77] gives the true density of the diffusion process $x(t)$).

Next we use the alternative method of simulating the o.d.e.

$$\dot{X}(t) = -2X(t) - 1 + 2^{\frac{1}{2}}(1 + X(t))y(t) \qquad (6.2.28)$$

(obtained directly from the Stratonovich s.d.e.) using a fourth-order Runge-Kutta formula. We begin with $\Delta t = 0.002$, but realise that the RK method allows a much larger Δt to be used before any significant error arises in integrating (6.2.28) over Δt. As a result, we compare the Euler and RK methods for a variety of Δt values, each time keeping the number of samples equal to 50,000 to keep the computational effort constant.

The RMS error in the estimated density function was evaluated in each case, and is plotted in Figure 6.2.3 against a logarithmic Δt scale. Up until $\Delta t = 0.002$, the errors in the Euler and the RK methods are equal, as the truncation error in each formula is negligible*. Beyond $\Delta t = 0.002$, however, the error in the Euler method becomes higher than that of the RK method, as the truncation errors of the first order Euler formula begin to become significant. The truncation errors of the RK formula do not become significant until beyond $\Delta t = 0.2$ (see Figure 6.2.4), hence the error using the RK method continues to decrease up to this point.

The conclusions are as follows. For a fixed number of samples, it is best to use as large a Δt as possible so that the maximum amount of independent statistical information is generated. The memory time of the system, τ'_{rel}, is about 2 time units, and so from this point of view, we should take $\Delta t = 2$. However, the statistical error between the physical process (6.2.28) and the diffusion process (6.2.4) is appreciable unless the condition $\tau_{cor} << \tau'_{rel}$ is satisfied (c.f. path 1 in Figure 6.2.1). We learned from the PRBS example in Section 4.1.4 that the error involved in path 1 of Figure 6.2.1 became small when τ_{cor} was less than about $0.1\ \tau'_{rel}$.

Combining these two considerations, we decide that $\Delta t = 0.2$ is the best choice of Δt. Now the main point is this: errors along

---

* The error in this region is almost all due to the finite record length, for, as τ_{cor} is much less than τ'_{rel} ($\doteq 2$), the error involved in path 1 is very small. Thus as Δt increases in this region, a larger record length is used (record length = 50,000 Δt), and more independent information is used in the statistical estimate of the density function.

path 2 in Figure 6.2.1 are small for this value of Δt using the RK method, while for the Euler method, a much smaller Δt has to be used to keep the truncation errors small. An indication of the truncation errors involved in integrating equations (6.2.27) and (6.2.28) by the Euler and RK methods respectively is given in Figure 6.2.4, where the number of correct decimal places (a logarithmic error scale) is plotted against log Δt.

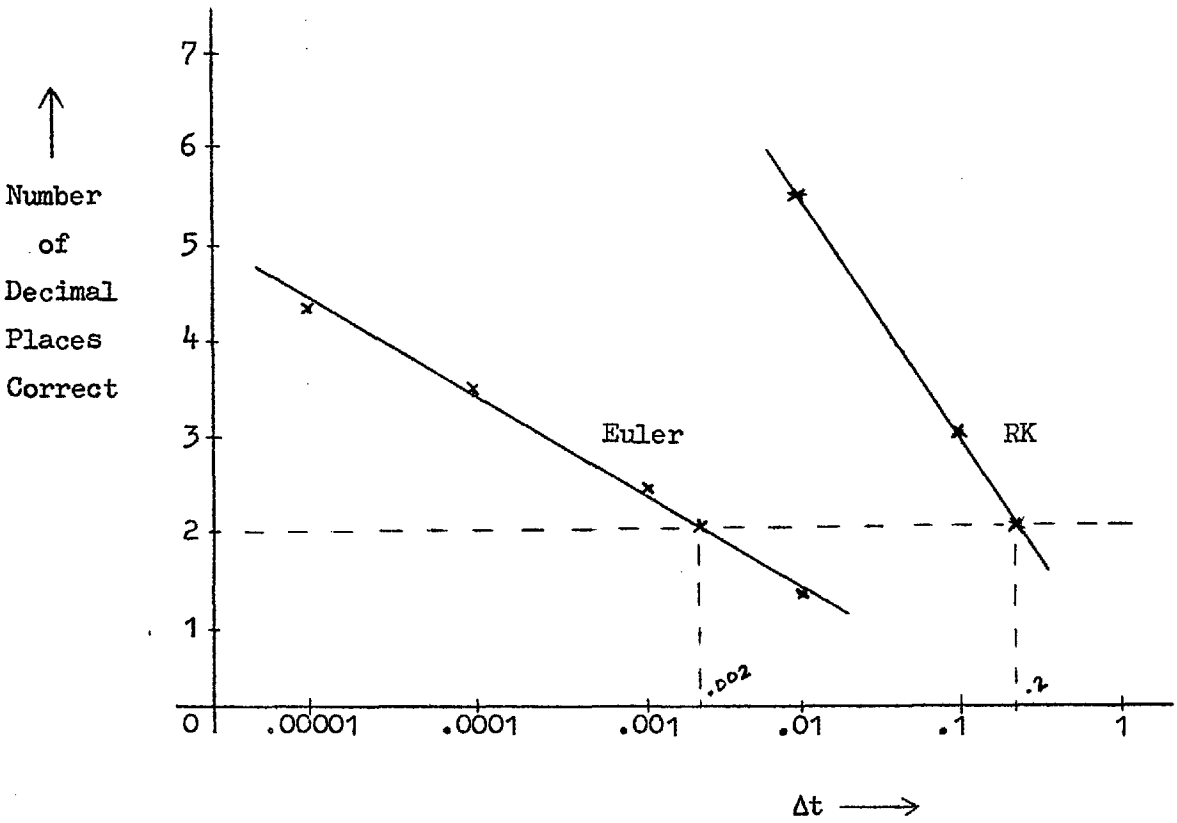


Figure 6.2.4 Sample Path Accuracy of Euler and Runge-Kutta Formulae Applied to Stochastic Equations

If the ordinates of Figure 6.2.4 were truncation error, then the slopes of the lines shown would be close to 1 and $2\frac{1}{2}$ for the Euler and RK methods, indicating convergence rates of $O(\Delta t)$ and $O(\Delta t^{2\frac{1}{2}})$ respectively. This agrees with our theory, as summarised in Table 6.2.2.

From the point Δt where the graphs of Figure 6.2.3 separate (i.e. the truncation errors of the Euler method become significant above $\Delta t = 0.002$), we draw the dotted lines in Figure 6.2.4 and deduce that two decimal place accuracy is a minimum standard to ensure that errors in distribution due to truncation error are small. We also see that this accuracy is obtained in the RK method by a $\Delta t = 0.2$. A factor of 100 separates these two Δt's, and so the RK method is much more efficient even though the RK formula requires four times as much computation per time step as the Euler formula.

Another point in favour of the RK method is that $\Delta t = 0.2$ is the "break point" governing acceptable accuracy in both the convergences in paths 1 and 2 of Figure 6.2.1, and in this sense, the RK method is well matched to our method of approximating diffusion processes by physical processes. In fact, for a particular form of o.d.e., the Δt for a given truncation error is proportional to τ'_{rel}, and in this case for the acceptable truncation error we have $\Delta t = 0.1\ \tau'_{rel}$ for the RK formula, but $\Delta t = 0.001\ \tau'_{rel}$ for the Euler formula. This illustrates the suitability of the RK method for our example, but these ratios will depend on the particular form of the equation as well as the relative size of the noise term. In the author's experience with other examples and other noise ratios, it was found that the Runge-Kutta method was more efficient than the Euler method except when the system was heavily dominated by the noise (when the average size of the random term F was greater than about five times the size of the non-random term f in equation (6.1.1)).

On the digital computer, we can confirm the effects noted in Chapter 5 with a high degree of statistical confidence.* Essentially we were trying to show that ordinary differential equations to be programmed on to the analogue computer must not be obtained directly from the Ito s.d.e. of the diffusion process (we called this the naive method), but must be obtained from the s.d.e. by the methods of Chapter 4 (this reduced to using the Stratonovich equation directly if the noise was scalar or vector symmetrical). On the digital computer, the RK method parallels the analogue computer

---

* But at the expense of long computer runs.

method, and we can confirm the results of Chapter 5 using our present example by computing with the o.d.e. (6.2.27) obtained from the Ito s.d.e., and with the o.d.e. (6.2.28) obtained from the Stratonovich s.d.e., in each case using the RK method.

The latter of these two possibilities is what we have already done, and the resultant error in distribution is shown in Figure 6.2.3. Keeping $N = 50,000$, and using the best value of $\Delta t = 0.2$, we apply the RK method to equation (6.2.27), and find the RMS error in distribution to be 0.21. As the resultant histogram was statistically smooth, we deduce that this large error cannot come from random fluctuations (the error of 0.02 when simulating equation (6.2.28) under these conditions gives an idea of the size of the error due to random fluctuations in the sample), and conclude that equation (6.2.27) is the wrong o.d.e. to simulate.

Comparing equations (6.2.27) and (6.2.28), we see that equation (6.2.27) has a substantial positive drift compared with equation (6.2.28) for all values of X greater than -1. The effect of this error in the drift term is clearly seen in Figure 6.2.5, where the simulation of equation (6.2.27) by the RK method gives a distribution which is heavily biased to the right of the true density function, while the simulation of equation (6.2.28) gives a distribution which seems to have an unbiased error. In each case, the curves shown are sketched through the histogram points.

The non-linear filter of Section 5.2.2 was simulated on the digital computer using the RK method, and the same effect was noted. When equation (5.2.23) was used (the Stratonovich equation), the theoretical performance of the optimal filter was achieved with a high statistical confidence (mean square estimation error = 0.265 $\pm$.01, theoretical = 0.27). In contrast, when equation (5.2.24) was used (the Ito equation), the filter performance was substantially lower (mean square estimation error = 0.365 $\pm$ .01).

Thus using a digital computer, which is a convenient tool for collecting and analysing statistical data, we have confirmed the results of Chapter 4. The analogue computing of Chapter 5 was interesting, as the analogue can represent a variety of physical

noise sources exactly, and is thus a more realistic simulation tool than the digital computer. However, the analogue itself cannot conveniently collect and analyse statistical data, and so a hybrid computer would be the best tool to continue experimental work related to the results of Chapter 4. More examples should be tried, particularly using noise sources with asymmetrical characteristic matrices.

6.3 Digital Data Smoothing by Orthogonal Functions

In the course of studying Monte Carlo or simulation methods on the digital computer, a useful data smoothing technique was developed. The technique is applicable to function type solutions such as the probability density function $P(x)$ and advantage is taken of the orthogonality property of certain polynomial expansions. The idea came from the Hermite coefficient representation of density functions discussed in Section 2.4.

Consider the problem of forming an estimate of the probability density function $P(x)$ from statistically generated data. The usual technique is to quantise the x space into a convenient set of finite cells, and form the histogram of the quantised data. The main difficulty is that we must choose a compromise between smoothness and resolution, given a particular amount of data. For example, consider the estimation of the height $\bar{P}$ of an element of the histogram of width Δx. If there are N trials in the simulation, and n of these fall in the cell of width Δx, then a suitable normalised estimate $\bar{P}$ is

$$\bar{P} = \frac{n}{N \Delta x} . \tag{6.3.1}$$

Assuming the trials to be independent, the number n is from a binomial distribution and

$$E[n] = N P \Delta x,$$

and

$$\mathrm{Var}[n] = N P \Delta x (1 - P \Delta x),$$

where P is the true solution somewhere in the cell. Then we have

$$E[\bar{P}] = P, \tag{6.3.2a}$$

and

$$\mathrm{Var}[\bar{P}] = \frac{P(1 - P \Delta x)}{N \Delta x} . \tag{6.3.2b}$$

Thus for solution resolution* we need a small Δx and for solution smoothness (low Var $[\bar{P}]$) we want a large Δx. Our point here is that we may not be able to make the solution as smooth as we would like because of resolution or restrictions on the amount of data available.

Then we must find other ways of smoothing the solution, which gives a smooth estimated function $\bar{P}(x)$ but retains the essential features of the statistical data. The following method based on orthogonal polynomial expansions smooths the function $P(x)$ while leaving the first n moments of x unaltered.

6.3.1 Orthogonal Polynomial Expansions

Let $P(x)$ be expressed by the infinite functional series expansion

$$P(x) = \sum_{r=0}^{\infty} k_r J_r(x) g(x), \qquad (6.3.3)$$

where k_r is a constant, $J_r(x)$ is a polynomial of degree r, and $g(x)$ is a non-negative weighting function closely related to the expected shape of the function $P(x)$. The polynomials $J_r(x)$ can always be found by the Gram-Schmidt orthonormalisation process so that the following orthonormality relations hold:

$$\int_{-\infty}^{\infty} J_r(x) J_s(x) g(x)\, dx = \begin{cases} 1 & r = s \\ 0 & r \neq s \end{cases} \qquad (6.3.4)$$

A convenient numerical procedure for finding the coefficients of the

---

* One way of expressing solution resolution is that any functions $E[F(x)]$ which are to be estimated from the unquantised data, become $E[F(x + O(\Delta x)]$ when estimated from the histogram. For example, the second moment of x is estimated too high by

$$\frac{1}{12} \Delta x^2$$

from the histogram, if the quantisation intervals are equally spaced.

polynomials $J_r(x)$ is given in [100] where the basis functions f_i are taken as the linearly independent set of polynomials $1, x, x^2, x^3, \ldots$ In Section 2.4, we used the Gaussian curve as the weighting function $g(x)$ and found the polynomials $J_r(x)$ to be $(r!)^{\frac{1}{2}} H_r(x)$, where $H_r(x)$ were the Hermite polynomials.

From (6.3.3, 4), we find that the coefficients k_r are given by

$$k_r = \int_{-\infty}^{\infty} P(x)\, J_r(x)\, dx , \qquad r = 0, 1, 2, \ldots, \quad (6.3.5)$$

but as $P(x)$ is the density function of x, equation (6.3.5) can be written as

$$k_r = E\,[J_r(x)]. \quad (6.3.6)$$

This latter expression is particularly convenient for estimating k_r from data x.

Our smoothing method is based on the following assumptions:
(a) the expansion (6.3.3) is convergent, and the reconstructed function

$$P_n(x) = \sum_{r=0}^{n} k_r\, J_r(x)\, g(x) \quad (6.3.7)$$

is a good approximation to $P(x)$ for low values of n.
(b) the "high frequency" information in $P(x)$ is contained in the higher expansion functions $J_r(x)\, g(x)$, $r > n$, so that $P_n(x)$ is a smoother function than $P(x)$.

In section 2.4, we discussed the validity of these assumptions when $P(x)$ was near to the Gaussian in shape and the expansion was the Hermite polynomial expansion. Although not much is known about the expansion (6.3.3) for arbitrary weighting functions $g(x)$, it seems reasonable that if $g(x)$ is close to $P(x)$ in shape, then the expansion (6.3.3) is quickly convergent. Cramer [51] verifies this for certain orthogonal polynomial expansions, and for an arbitrary weighting function $g(x)$, the assumption can quickly be checked

computationally such as we have done in Section 2.4.2.

In general, the functions $J_r(x)\, g(x)$ contain more points of inflection (or maxima and minima) the larger r is made, as $J_r(x)$ is a polynomial of degree r. For example, for the Hermite polynomials $H_r(x)$ and the Gaussian curve $G(x)$, the functions $H_r(x)\, G(x)$ had r zero crossings. In this sense, the functions $J_r(x)\, g(x)$ are less smooth for increasing r, and the reconstructed function $P_n(x)$ of (6.3.7) is more smooth the lower we choose n.

Finally we show that the first n moments of x are correctly contained in the reconstructed density $P_n(x)$. Writing the m:th moment of x as

$$E\,[x^m] = \int_{-\infty}^{\infty} P(x)\, x^m\, dx, \tag{6.3.8}$$

which from (6.3.3) can be written as

$$E\,[x^m] = \sum_{r=0}^{\infty} k_r \int_{-\infty}^{\infty} x^m\, J_r(x)\, g(x)\, dx. \tag{6.3.9}$$

But as the polynomials $J_r(x)$ are linearly independent, we can express x^m as

$$x^m = \sum_{s=0}^{m} c_s\, J_s(x), \tag{6.3.10}$$

where the c_s are uniquely defined coefficients. Then (6.3.9) becomes

$$E\,[x^m] = \sum_{r=0}^{\infty} k_r \int_{-\infty}^{\infty} \sum_{s=0}^{m} c_s\, J_s(x)\, J_r(x)\, g(x)\, dx. \tag{6.3.11}$$

But from the orthogonality relations (6.3.4), the integral (6.3.11) only has a value when $r = s$, and then (6.3.11) can be written as

$$E[x^m] = \sum_{r=0}^{m} k_r\, c_r \tag{6.3.12}$$

That is, the coefficients k_r, $r > m$, are not used to calculate $E\,[x^m]$, and we could work back through equations (6.3.11) to (6.3.8) to show that

$$E[x^m] = \sum_{r=0}^{n} k_r \int_{-\infty}^{\infty} x^m J_r(x)\, g(x)\, dx$$

$$= \int_{-\infty}^{\infty} P_n(x)\, x^m\, dx \qquad (6.3.13)$$

provided $n \geq m$. Of course, we note that the coefficients k_r as found by (6.3.5) are independent of how high we truncate the series (6.3.3) to get the finite series (6.3.7), which is another result of the orthogonality relations.

6.3.2 Data Smoothing by Finite Expansions

Consider the problem of estimating the probability density function of a system $x(t)$ by simulating the system and obtaining N samples of x at time t. As discussed in Section 3.2.3, each sample $x_\alpha(t)$ contributes a delta function $\phi N^{-1} \delta(x - x_\alpha(t))$ to the estimate of $P(x,t)$. This, of course, gives a pathologically unsmooth solution estimate $\bar{P}(x,t)$, and so we usually average the delta functions in a region to obtain point (3.2.50) or histogram solutions (6.3.1).

An alternative smoothing procedure which does not involve the type of x quantisation as the point solutions is one which expands the estimate $\bar{P}(x,t)$ in the orthogonal polynomial series (6.3.3). As $\phi = 1$ for probability density functions (c.f. (3.2.22)), we write $\bar{P}(x,t)$ as

$$\bar{P}(x,t) = \sum_{\alpha}^{N} N^{-1} \delta(x - x_\alpha(t)). \qquad (6.3.14)$$

Then to find the coefficients $k_r(t)$ of the expansion, we substitute (6.3.14) for $P(x)$ in (6.3.5), and obtain

$$k_r(t) = \int_{-\infty}^{\infty} \sum_{\alpha}^{N} N^{-1} \delta(x - x_\alpha(t))\, J_r(x)\, dx,$$

$$= N^{-1} \sum_{\alpha}^{N} J_r(x_\alpha(t)), \qquad r = 0,1,2,\ldots, \qquad (6.3.15)$$

which is recognized as the expected value of $J_r(x)$ (6.3.6) at time t.

The coefficients k_r, $r = 0, 1, 2, \ldots$ can be evaluated simply by forming the expected value (6.3.6) using each sample $x_\alpha(t)$ as in (6.3.15). However, if N is very large, this may be a time-consuming operation, and quantising methods can be used to estimate $P(x,t)$ at suitable points and then the integral (6.3.5) is evaluated numerically. The quantisation here can be very fine compared to that discussed earlier (6.3.1), for we are not concerned about solution smoothness at this stage, and can choose Δx quite small to keep the $O(\Delta x)$ errors small. As discussed in Section 2.4.2, the numerical integration of (6.3.5) can be done very efficiently if the point estimates $\bar{P}(x_r, t)$ are obtained at the zeroes of the Gaussian quadrature formula associated with the weighting function $g(x)$ and the orthogonal polynomials $J_r(x)$ of (6.3.3).

The main advantage of the orthogonal polynomial representation (6.3.7), is that the information in each successive expansion function $J_r(x)\, g(x)$ is independent information, which means that the coefficients k_r, $r = 0, n$, are not a function of n, the truncation point. This means that we can compute k_r by the formula (6.3.15) without knowing beforehand where we will truncate the series. However, as the data is random, the coefficient series $k_0, k_1, k_2, \ldots$, computed from (6.3.15) will not have the nice convergence properties that the series evaluated from the true (smooth) density function $P(x)$ (see (6.3.5) - we have chosen the weighting function $g(x)$ so that the series for $P(x)$ is quickly convergent. As in Section 2.4.4, it may be necessary to normalise the data $x_\alpha(t)$ to assure this convergence.).

This is the basis of our filtering method: the true density $P(x)$ is represented by a few coefficients which are nicely converging. But we notice that the coefficient series evaluated from the random data do not converge, and after the first few coefficients, they appear to take on random values. We conclude then that the first few coefficients obtained from the random data contain all the information we want in $\bar{P}(x,t)$, and that the higher coefficients contain information which is essentially random error resulting from

the finite sample size. With this justification, we truncate the series for $\bar{P}(x,t)$ to obtain $\bar{P}_n(x,t)$, which is then a smoother estimate of $P(x,t)$ than $\bar{P}(x,t)$.

Thus the series is truncated using a priori knowledge or assumptions on the convergence rate of the true density function $P(x,t)$. In view of this, it is useful to look into the origin of the random components of the coefficients, and in particular, to see how their size is affected by the sample size, and how their size affects the accuracy of the reconstructed distribution.

Expansion of Samples from the Gaussian Distribution

To investigate this point, it is convenient to choose an expansion whose coefficients are purely random: the expansion of $\bar{P}(x,t)$ by the Hermite polynomial series (2.4.6) when $x(t)$ is a Gaussian (0, 1) random number. Then $k_o = 1$ and all higher coefficients should be zero, but in fact have a zero mean random component due to the finite sample size N.

The coefficients k_1 to k_{10} were computed for samples of N = 500 to 10,000 Gaussian random numbers. The root mean square values of these coefficients are plotted in Figure 6.3.1 as a function of N to a log-log scale.

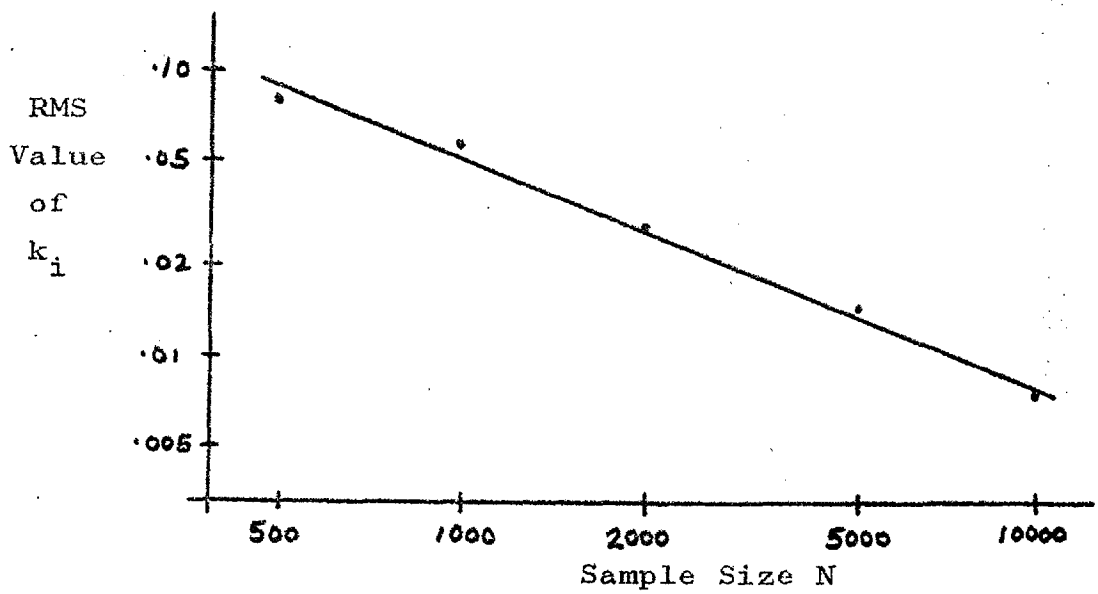


Fig. 6.3.1 RMS Value of Expansion Coefficients.

The slope of the line drawn through the points is -0.65, which indicates that the variance of the error in the coefficients is proportional to $N^{-1.3}$. As $J_r(x)$ is a function with a fixed variance, the theory of sampling statistics suggests that the variance of k_r should be proportional to N^{-1}. The convergence of the coefficient error with respect to sample size N that we have observed is somewhat higher than the theoretical figure, but it is not an inconsistent estimate considering the number of points in Figure 6.3.1.

We also observed that the variance of the coefficients k_r was independent of r, which is because the variance of $J_r(x)$ is independent of r. This is a result of the scaling of the form of Hermite polynomial expansion we have used (see the note following equation (2.4.8) - the functions are normalised as in (6.3.4)), which also means that the error in the reconstructed distribution $\bar{P}_n(x,t)$ caused by an error in k_r, $r \leq n$, is independent of r.

With this assumption, we proceed to look at the error in the reconstructed distribution as a function of the sample size N. The density $P(x,t)$ was estimated at the abscissae $x = -4.8(0.3)4.8$ by forming the histogram of cell width 0.3 and by filtering with the Hermite polynomial series, keeping coefficients up to k_{10}.

A typical estimate of the Gaussian density is shown in Table 6.3.1 where $N = 2000$ data are used. The true density is also given in the table and an error measure, consisting of the absolute error in the ordinates averaged over the 33 abscissae shown, was evaluated. For this example, the error in the histogram was .058 and the error in $\bar{P}_{10}(x,t)$ was .020. In forming $\bar{P}_{10}(x,t)$ we first evaluated the mean and variance of the data as

$$\begin{aligned} \text{Mean} &= -0.013 \\ \text{Variance} &= 1.012 \end{aligned}$$

Then the data were normalised to zero mean and unit variance, and the Hermite expansion coefficients were found for the normalised data as follows:

k_0	k_1	k_2	k_3	k_4	k_5	k_6	k_7	k_8	k_9	k_{10}
1.00000	-.00005	-.008	-.024	.009	-.054	.044	-.017	.022	-.004	-.030

The type of smoothing afforded by the reconstructed distribution $\bar{P}_{10}(x,t)$ can be judged by the entries in Table 6.3.1. The density $\bar{P}_{10}(x,t)$ has exactly the same first 10 moments as the data, and so smoothing is not achieved by lowering the variance of the estimates of these moments. The higher moments, however, are altered, by limiting the maximum frequency content of $\bar{P}_{10}(x,t)$. The highest reconstruction function $J_{10}(x)\ g(x)$ of (6.3.7) has 11 maxima or minima, of which only 7 are of appreciable magnitude, and so the function $\bar{P}_{10}(x,t)$ cannot follow the 33 degrees of freedom of the histogram. In this sense, adjacent values of the histogram are smoothed*, but in a way which does not involve any discretization of the x variable.

The error measure was evaluated for the histogram and for $\bar{P}_{10}(x,t)$ for various numbers of data, $N = 500$ to 10,000, and is shown in Figure 6.3.2 as a function of the RMS value of the expansion coefficients. It is noted that a linear relationship exists between the size of the expansion coefficients and the error in the estimated density $\bar{P}_{10}(x,t)$, a result which is expected from the linear expression (6.3.7) for $\bar{P}_{10}(x,t)$. The same linear relationship holds for the error in the histogram which shows that the error in the histogram and the error in $\bar{P}_{10}(x,t)$ are linearly related, by a factor of about 2.

Another effect observed was that for a fixed N, the error in $\bar{P}_n(x,t)$ was proportional to $n^{\frac{1}{2}}$. This is a consequence of the independence of the coefficients k_r and the normalization of the expansion functions. Thus adding another term k_n in the expansion adds another function which has the same variance as, and is independent

---

* The actual smoothing process is more complicated than this, as seen from the form of the orthogonal polynomial expansion. It is difficult, however, to make a more precise illustrative statement than this.

of, the previous expansion functions.

Thus from our exercise of smoothing the density of Gaussian random numbers by our orthogonal polynomial series, we have gained a feeling for the errors caused in the estimated density by introducing random components into the coefficients of the expansion series. It is clear that the series k_r should be truncated as soon as the mean value of the coefficients reaches zero. This point can be determined from an a priori knowledge of the true density $P(x,t)$, or by repeating the experiments and noting the statistical properties of each coefficient. Also, via Figures 6.3.1, 2 we have related the error in the estimated density to the number N of data used in the experiment, for a fixed number of expansion coefficients. Further, this error depended on the square root of the number of random expansion coefficients used, and as long as enough coefficients are used to accurately represent the true density, the error should be independent of the shape of the true density. These considerations help us to choose the number of coefficients n in the expansion, and the number N of data to collect in the statistical trials.

Example of Chapter 2

To illustrate these methods, we study the example of Chapter 2. From Section 2.4.4, we recall that the expansion coefficients up to k_8 were needed to give a good representation of the true solution $P(x,t)$, and to be safe we shall choose k_{10} as our upper limit. We decide on a reasonable error limit, say .02, and see from Figures 6.3.1, 2 that $N = 5000$ should estimate the density function to this accuracy.

Figure 6.3.3 shows the results of a typical run, using the initial conditions of the example of Figure 2.3.5. The equations were simulated by the Runge-Kutta method with a step length of $\Delta t = 0.1$, and Figure 6.3.3. gives the estimated solution at $t = 2.0$. The finite difference solution of Figure 2.3.5 is accurate to better than 1%, and is taken as the true solution $P(x,t)$. Also shown in

Figure 6.3.3 is the histogram obtained from the 5000 data and $\Delta x = 0.3$, and the estimate $\bar{P}_{10}(x,t)$ of the Hermite polynomial expansion.

It is clearly seen that the estimate $\bar{P}_{10}(x,t)$ is much smoother and at most points closer to the true solution $P(x,t)$ than the histogram. Averaged over the 21 points $x = -3.0(0.3)3.0$, the histogram had an average absolute error of 0.031, while the smoothing afforded by the polynomial expansion reduced the error to 0.014. These error figures are consistent with those of Figure 6.3.2, and once again we see that the polynomial smoothing cuts the errors in the estimated density down by a factor of 2.

The mean and the variance of the 5000 data are

Mean = 0.001

Variance = 1.249

and the expansion coefficients of the normalised data are

k_0	k_1	k_2	k_3	k_4	k_5	k_6	k_7	k_8	k_9	k_{10}
1.00000	.0006	.00007	.0007	-.105	.007	.066	-.013	-.023	.017	-.001

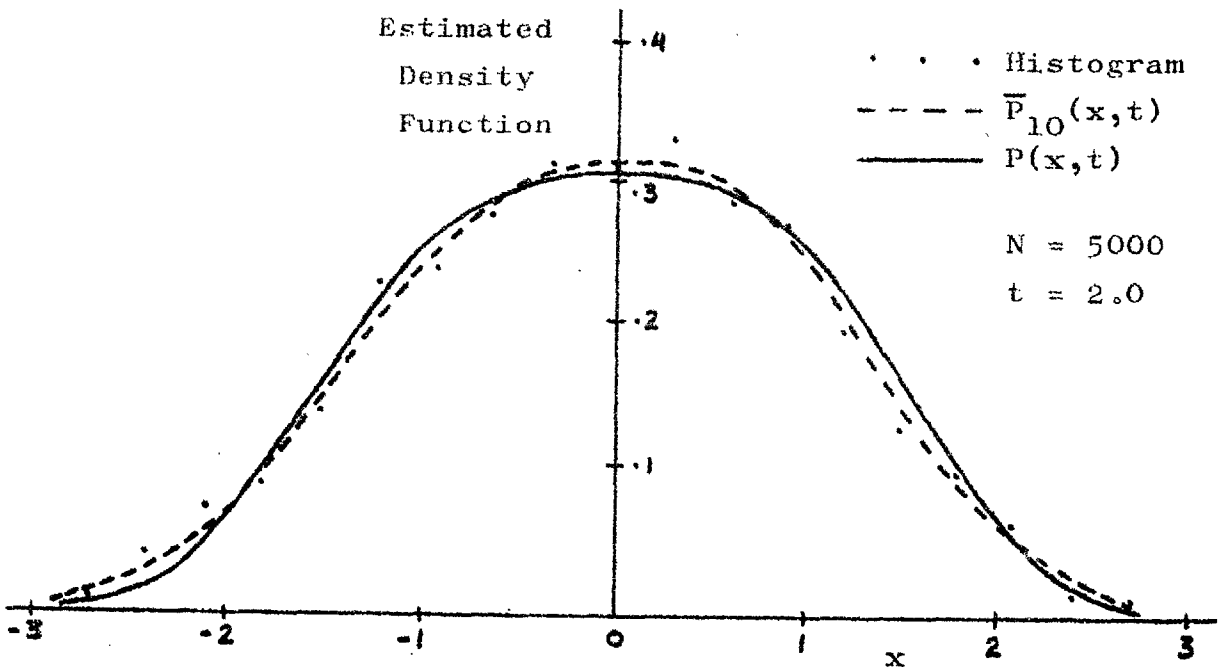


Fig. 6.3.3 Monte Carlo Solution of Fig. 2.3.5 Example.

We, of course, know that the true solution has a zero mean and is symmetrical, and so we could introduce further smoothing by setting the odd coefficients equal to zero. This cannot be done for general polynomial expansions (6.3.3), however, as the orthogonal polynomials $J_r(x)$ for a general weighting function $g(x)$ do not separate into odd and even functions of x as the Hermite polynomials do (c.f. equation (2.4.4)).

We conclude that the smoothing of estimates of density functions obtained from simulated data can be usefully achieved by expanding the estimated density in orthogonal polynomial expansions, and truncating the series. The method relies upon having an a priori estimate of the shape (not mean or variance) of the true density function so that a quickly convergent orthogonal polynomial expansion can be chosen. This representation of the estimated density function has the advantage of using few parameters and the smoothing procedure does not alter the lower order moments of the data. The price we pay is in extra computing time to evaluate the expansion coefficients, and the success of the method may depend on the importance of computing time (from (6.3.15) we note that the time taken to evaluate the first n expansion coefficients is roughly that needed to estimate the first n moments of the data).

CHAPTER 7

CONCLUSIONS

We have already given summaries at the ends of most sections, and in some cases we have pointed out the limitations of the present work and indicated the needs for future investigation. In this concluding chapter we review the highlights of the present work and indicate the most promising avenues for future work.

Our work begins with the prediction problem and we attempt to obtain numerical estimates of the future statistics of non-linear stochastic systems. The direct approach is to use the Fokker-Planck equation: a parabolic partial differential equation whose initial conditions are the present probability density of the stochastic system and whose subsequent time solution gives the future probability density of the system.

In Chapter 2, numerical methods are presented for solving the FP equation on a digital computer. The first method was the classical numerical method of finite differences and the second was an integral transform method of some novelty using orthogonal polynomial expansions. Although the finite difference method was the better of the two methods for general problems, both methods suffered from the following limitations:

(a) The class of stochastic system whose statistics could be predicted by these methods is quite restricted. The most severe restriction is on system dimensionality, and the solution of problems of even second dimension is not a routine matter. The highest dimension problem that the finite difference method has been known to solve is a three dimension one, and the integral transform method is felt to be impractical for problems of higher than one dimension. In addition, the form of the non-linearity affected the solution, and in general, the smoother the non-linearity is, the better the accuracy is that can be obtained.

(b) In connection with the dimensionality aspect, we can easily reach the limit of computer size and speed. For even two dimensional problems, a computer of large storage capacity and high computing speed is needed to obtain useful solution accuracies. Although larger and faster computers are becoming more available, this does not speak well for the generality of the method.

(c) The background knowledge which must be gained to use these methods can be quite considerable. To someone whose main speciality is not numerical analysis (or applied mathematics), several months of study are needed to learn the basic principles of partial differential equation solution, and even then the newer methods for two and three dimension problems are difficult to follow. Thus an engineer having to estimate statistical behaviour of a system faces a large diversionary study before he can obtain useful results. Again, this is a major drawback of a method which hopes to have wide appeal.

Apart from the research problems in numerical analysis, a useful research topic would be to investigate the effect of the reduction of system dimensionality on the accuracy of statistical descriptions of the system. In general, as the dimension of a system is increased, less useful information is contained in the behaviour of the each additional component of the state vector added, and one intuitively feels that all the pertinent information could be summarised in a few components. The problem of dimensionality reduction has already been studied in the context of multivariable control problems, for the problems of choosing control algorithms (particularly optimal control ones) are also plagued by the curse of dimensionality. It has been found that the deterministic behaviour of some large systems can be well approximated by that of low dimension systems, and we must see if the same is true for the statistical behaviour of noisy systems. The analysis of Chapter 4 is one step in this direction, for there we approximate an n dimension physical process by an n dimension diffusion process. But the physical process can sometimes be represented as a diffusion process of dimension higher than n (c.f. Section 2.2), and the n dimension diffusion process can be considered as a

reduced-dimension approximation to the high dimension system. The analysis of Section 4.1 then pertains to the statistical accuracy of such an approximation, and the analysis could be extended to pursue this matter.

In the hopes of relieving some of the limitations described above, simulation methods are investigated for studying the statistical behaviour of noisy systems. This entails two main theoretical considerations, that of the relation between diffusion processes and the solution of parabolic equations, and that of the relation between diffusion processes and physically realisable processes.

In Chapter 3, the first of these questions is considered. We detail the connection between the simulation of a diffusion process and the solution of the process' FP equation, and point out that the simulation constitutes a Monte Carlo solution of the FP equation. We show that the derivation of the FP equation requires that the trajectories of the simulated diffusion process are continuous or are conserved in the sense that they do not begin or end within the time interval of the simulation. We deduce that any parabolic equation which can be written in the form of a FP equation has a diffusion process associated with it, so that the parabolic equation is the FP equation of the diffusion process. Then a Monte Carlo solution for such a parabolic equation can be obtained by simulating the diffusion process in a simulation with conserved trajectories.

In general, parabolic equations do not have the form of FP equations, but we show that for any given parabolic equation, a FP equation can be chosen which is nearest to the given equation. The diffusion process associated with the nearest FP equation is called the underlying diffusion process of the given parabolic equation, and we show that a Monte Carlo solution of the parabolic equation can be obtained by a modified simulation of the underlying diffusion process. The modification consists of forcing the density field of simulated trajectories to grow and decay by allowing the trajectories to terminate or new ones to begin within the time interval of the simulation. This violation of the principle of conservation of trajectories which

we had observed when solving FP equations accounts for the differences between the general parabolic equation and its nearest FP equation.

In this way we could solve a general parabolic equation by simulation techniques. We show under what conditions the solution of parabolic equations remains positive, and we see that the mechanics of the simulation assured positive solutions under these conditions. We point out that parabolic equations usually arise in physical situations in which a random element of a diffusive character is present, and what we are doing in our Monte Carlo solution procedure is simulating the diffusion process which is inherent in the situation which the parabolic equation describes. We pursue this analogy for the example of heat conduction in a solid, and show that the simulated trajectories are a specific form of heat energy. We see that the conservation of trajectories in the simulation agrees with the conservation of thermal energy in the interior of the region where the parabolic equation is defined, and we can use the principle of the conservation of heat and the specification of thermal flux to specify the behaviour of the simulated trajectories at boundaries where the parabolic equation is not defined.

We present numerical results which give no indication that the solution method has any bias in the steady state, and argue that the solution method does not introduce any transient errors into the solution, apart from the Δt quantisation necessary on the digital computer. We show that the estimated solutions have accuracies agreeing with the laws of sampling statistics, and that these laws present the major obstacle of the Monte Carlo method. This is that the accuracy of the estimated solutions depends on the square root of the number of simulated trajectories or trials, and so the limitation of computer size and speed restricts the accuracy of solution that can be attained. To this extent one of the limitations discussed in connection with the direct solution of the FP equation is still present, but the other two limitations concerning system complexity and mathematical background necessary are largely removed. We also note that modern hybrid computers to a great extent relieve the remaining limitation.

The future work on this topic should proceed initially along experimental lines, preferably with a suitable tool as a hybrid computer. We have presented a general theory concerning the Monte

Carlo solution of parabolic equations, but have not established the practicality of the method. Other investigators, however, have established the practicality of Monte Carlo methods, and the points we must investigate is where the mechanics of our simulation differ from theirs. The difference is in the implementation of boundary conditions and the non-conserved trajectories. It is these points which allow our Monte Carlo method to tackle certain parabolic equations that the previous Monte Carlo methods could not.

In Chapter 4, we discuss the relation between diffusion processes and processes which are physically realisable (physical processes). The distinction between these two types of process is necessary, as diffusion processes are random processes involving the theoretical concept of white noise, and as such cannot be exactly duplicated in practice. However, as diffusion processes are continuous Markov processes (physical processes are not), they have advantages of mathematical convenience which make it desirable to approximate physical processes by diffusion processes for analysis purposes. For example, we can write down parabolic equations to describe the statistics of diffusion processes and extend this concept into a general connection between parabolic equations and diffusion processes. Also, in order to perform the simulations of diffusion processes required in Chapter 3, we have to convert them to physical processes so that they can be represented on a computer which is a physical device.

In Section 4.1, we investigate the relation between physical and diffusion processes by evaluating their incremental statistics. Diffusion processes have increments which are Gaussian in the small, and thus their statistics are described by two incremental moments. These incremental moments are precisely defined quantities when the time increment δt is set to zero, but for a non-zero δt these quantities are specified to a first order accuracy (i.e. they are given as $b(x,t)\,\delta t + o(\delta t)$). We find that these quantities can be evaluated for physical processes to the same accuracy provided the upper frequency of the noise generating the physical process is substantially higher than the maximum frequency content of the output

of the process. In effect, this criterion ensures that the physical process has properties which are nearly Markovian, justifying the approximation of the physical process by a diffusion process (or vice versa).

Our approach in Section 4.1 is conceptually similar to Stratonovich [21], except that we use a different form of equation to describe the physical process which conveniently separates the physical noise into a separate factor. In this way, we can characterise the noise process independently of the physical process equations, which is a convenience when evaluating the properties of alternative noise source choices. The characterisation we use is that introduced by Clark, and our analysis clearly shows that this characterisation contains exactly the minimum number of parameters needed to specify the statistics of the physical process. This characterisation and the separation of the noise factor allow us to consider non-stationary noise sources which Stratonovich does not consider, and our analysis shows the need of the criterion on the noise and process upper frequencies more clearly than Stratonovich's.

In particular, we emphasise the convenience of choosing approximating diffusion processes for given physical processes by evaluating the approximate incremental moments of the physical process and constructing the diffusion which has the same moments. The effectiveness of the method is illustrated in Appendix D, where we analyse the transient statistical properties of a filtered pseudo random binary sequence (PRBS). This example illustrates the ease of choosing the diffusion approximation by matching the finite increments, and illustrates the power of analysis (particularly transient) afforded by the Markov property of the diffusion process (here we use the stochastic calculus, outlined in Appendix A). The example illustrates the accuracy of using the statistics of the diffusion process to approximate those of the physical process, particularly as a function of the ratio of the upper frequencies of the noise (the PRBS) and the process (the output of the filter).

In Section 4.1, we considered a physical process generated by a physical noise source which was characterised by a set of parameters, and that provided the upper frequency of the noise was sufficiently

high, the physical process could be approximated by an equivalent diffusion process. In Section 4.2, we consider a limiting operation on the physical process whereby the upper frequency of the physical noise is extended to infinity in such a way as to preserve the characteristic parameters of the noise. We then find that the limiting physical process, assuming it is a Markov process, converges in distribution to a diffusion process which is the same diffusion process as we had chosen in Section 4.1 to approximate the statistics of the physical process. This means that the method of choosing approximating diffusion processes of Section 4.1 is consistent in the sense that the error in approximating the statistics of the physical process goes to zero as the main error parameter (the ratio of the upper frequencies of the noise and the process) reaches its limit (infinity).

We show that there are various ways of writing equations for diffusion processes which depend on the definition of white noise and on the interpretation of the stochastic integral. We point out that these various interpretations of stochastic equations has led to ambiguities and contradictions in the recent literature, particularly when writing Fokker-Planck equations for processes involving white noise. The ambiguity has usually centred around whether the white noise equations given are to be interpreted as Ito stochastic equations, or limiting forms of physically realisable equations (for example, Stratonovich equations). In Section 4.3, we show what form this ambiguity takes when dealing with linear systems with random coefficients, and show that the ambiguity can be removed by specifying the (limiting) physical noise by its complete set of characteristic parameters as well as specifying the type of stochastic equation we use.

In Sections 4.1, 2, we discuss the approximate equivalence of diffusion and physical processes in terms of their statistical properties. A major application of this theory is in simulation problems, where we hope to duplicate on a computer the statistical properties of a particular random system. But a computer is a physical device and can only duplicate physical processes, and so if we wish to simulate a diffusion process on a computer (for example, to apply the Monte Carlo methods of Chapter 3), we must first convert it to a physical

process with equivalent statistical properties. In Section 4.4, we apply the results of Section 4.1 to this problem and show how, given a particular diffusion process to simulate, we can choose a physical process which is suitable for implementing on an analogue or digital computer. We see that this is essentially a problem of choosing a suitable noise source, scaling it properly, and then adding an appropriate bias term to assure the correct statistical properties. To do this, we need only evaluate the characteristic parameters of the noise source we have chosen.

In Section 4.5, we show how a given physical process can be simulated on a computer. If the given physical process cannot be represented directly on the computer itself, the results of Section 4.1 can be used to choose another physical process which has approximately the same statistical properties as the given process. Again, we see that we must only know the characteristic parameters of the noise sources concerned.

Future work based on Chapter 4 should aim towards cleaning up the analysis of Section 4.1 in the hope of making more precise statements on the statistical error involved in approximating one process by another. Indeed, even the concepts of statistical error and statistical convergence are not well established, and it is not clear what norms or criterion we should apply when stating that "one process statistically approximates another". The error analysis of Section 4.1 is complicated by the fact that the derivation which matches the processes' statistics manipulates the incremental statistics of the process, while we are usually interested in the error in the statistics of the process itself, as opposed to its increments. Stratonovich's analysis suffers from this same problem, but Clark's analysis [22] is more direct as it does discuss the error in the process itself (even his analysis, though, does not bound the error, but only gives an order of convergence). A useful addition to Clark's work would be to extend it to the non-stationary and non-Gaussian physical processes which are considered in Chapter 4 of this thesis.

In Chapters 5 and 6, we present some of the practical aspects involved in simulating diffusion processes on analogue and digital computers. This is equivalent to the problem of building a physical system which has been derived and specified in the stochastic calculus, and we show how to build an optimal non-linear filter which is an example of some of the interesting results of modern stochastic control theory.

Chapter 5 concentrates on analogue computing, and we see that the main problem is solved when we choose an appropriate noise source and evaluate its characteristic parameters. Two common noise sources are discussed, and we give practical methods of evaluating their characteristic parameters. In addition, some experimental results illustrate that significant errors can occur in simulations if the results of Chapter 4 are not followed. In particular, we show that the performance of the non-linear filter can be considerably below the optimum if the filter is not built correctly.

Chapter 6 pertains to digital computing, and concentrates on the choice of numerical formula used to solve the ordinary differential equation (o.d.e.) describing the physical process which approximates the given diffusion process. We find that the particular o.d.e. used to simulate diffusion processes has a Taylor series incremental expansion which is unusual in its dependence on Δt, the digital time increment. This alters the convergence rates of the various numerical formulae used to solve o.d.e.'s, and we explicitly give the new convergence rates. We reason that single step formulae must be used to simulate diffusion processes, and some experimental results with the Euler and Runge-Kutta formulae confirm the convergence rates mentioned above.

Previous authors have only used the Euler formula to simulate diffusion processes, and we use the results of Chapter 4 to show that higher order formulae can be used as well. We give experimental evidence to show that the higher order formulae can simulate diffusion processes more efficiently than the Euler formula, except when the diffusion process is heavily dominated by the random terms.

The convergence rate discussed above refers to the convergence of the digital formulae to the true solution of the o.d.e., as opposed to the convergence of the statistical approximation presented in Chapter 4. In Section 6.2.1, we do discuss this latter convergence, but once again, we cannot make statements of any certainty. We give various arguments which suggest that the convergence in distribution is proportional to the inverse of the upper frequency of the physical noise, and give limited experimental evidence to confirm this. As mentioned earlier, much more theoretical as well as experimental work is needed to investigate this convergence.

In general, the experimental results of Chapters 5 and 6 have confirmed the theory of Chapter 4 with a certain degree of confidence (the most certain statement we can make is that the results introduce no statistical evidence that the theory of Chapter 4 is not correct). This experimental work should be continued, particulary with other examples, to establish the theory with a higher degree of confidence. A statistical analysis computer such as a special purpose hybrid computer would be particularly useful in this respect.

APPENDIX A

THE STOCHASTIC CALCULUS

A1. Stochastic Integrals and Integral Equations

The normal rules of calculus, based on the Riemann or Stieltjes concept of the integral, are defined in such a way that they can treat functions which satisfy certain smoothness or boundedness conditions. When we wish to treat functions which are not smooth, that is, are not differentiable or are of unbounded variation, then we must use rules of calculus based on different definitions. This new calculus has been called the stochastic calculus, and has been discussed by (among others) Bernstein, Doob, Ito and Wiener. The most complete accounts of the stochastic calculus appear in the books by Doob [20] and Skorokhod [73], although their arguments are usually confined to scalar examples.

The rules of the stochastic calculus, and a comparison with the ordinary calculus, will be discussed in terms of the integral or integral equation. Other rules, such as differentiation and the manipulation of differential, come directly from the concept of the integral, and will be mentioned later. The rules of the stochastic calculus, as discussed by the authors above, will be referred to as the Ito calculus. Some new rules for treating stochastic equations are due to Stratonovich [50]. As the latter rules have some interesting advantages in connection with the topics of this thesis, they will also be discussed below, and will be referred to as the Stratonovich calculus.

A.1.1 The Ito Stochastic Integral

Consider the ordinary integral (for the time being, in the scalar case only)

$$X(t) = X(o) + \int_o^t F(s)\, y(s)\, ds, \qquad (A1)$$

in which the integrand $F(t)\, y(t)$ is continuous with a bounded derivative (that is, the integrand satisfies normal Lipschitz conditions. The following argument can be extended to functions $F(t)\, y(t)$ with a finite (or even countable) number of discontinuities, as long as the functions remain bounded.) We will associate the function $y(t)$ with a band limited smooth physical noise process, as discussed in Section 2.2.

The Riemann definition of the definite integral of such smooth functions is as follows [42, p.101]:

Let the interval (o, t) be divided into n intervals by inserting the points of subdivision t_i in such a way that

$$0 < t_1 < t_2 \ldots t_{n-2} < t_{n-1} < t. \qquad (A2)$$

Let θ_i be any point in the interval of length $\Delta t_i = t_i - t_{i-1}$. Then

$$X(t) = X(o) + \underset{n \to \infty}{\text{limit}} \sum_{i=1}^{n} F(\theta_i)\, y(\theta_i)\, \Delta t_i \qquad (A3)$$

provided that the limit assures that the maximum Δt_i tends to zero. Clearly, setting θ_i equal to t_{i-1} or t_i makes the sum in (A3) analogous to the simple forward or backward difference formulae for solving differential equations, and the Riemann definition shows that forward and backward difference methods give the same results in the limit for arguments which are smooth functions.

Now consider the replacement of the physical noise $y(t)$ of (A1) by white noise. White noise is defined here* as the formal derivative of the Wiener process $w(t)$, and as such does not exist physically as it is everywhere infinite (or, $w(t)$ is of unbounded variation). As $y(t) = \frac{dw(t)}{dt}$ is now always infinite, the concept of the Riemann integral is not valid, and we must define a new integral to deal with such functions. This definition of the stochastic integral has been given by Ito [43] and a presentation of its properties is given in Doob [20, Ch. 9]. We will give some pertinent properties below.

Let $w(t)$ be the unit parameter Wiener process such that $w(o) = 0$, and for any $t > s$,

$$E\left[w(t) - w(s) \right] = 0, \tag{A4}$$

$$E\left[(w(t) - w(s))^2 \right] = t - s, \tag{A5}$$

---

* In Section 4.2 we shall allow a more general definition

and the expected value of the product of any non-overlapping increments is zero (that is, the Wiener process has infinitely divisible independent increments). The Wiener process is continuous with probability one, but is not differentiable, and so we will prefer to write the stochastic integral analogous to (A1) in Stieltjes form. Let the interval (o, t) have the partition (A2). Then the Ito stochastic integral is defined as (with $x(o) = 0$)

$$x(t) = \int_o^t F(s)\, dw(s) = \text{l.i.m.} \sum_i^n F(t_{i-1}) \left[w(t_i) - w(t_{i-1})\right] \qquad (A6)$$

where,
as n tends to infinity the maximum $\Delta t_i \downarrow 0$.* Note that the summation formula corresponds to that of the forward differences mentioned earlier in connection with the Riemann integral. The d· of (A6) is to be interpreted as a stochastic increment in the Ito sense (see the discussion on differential equations later). The stochastic integral (A6) has the mean and mean square

$$E\left[x(t)\right] = E\left[\int_o^t F(s)\, dw(s)\right] = 0, \qquad (A7)$$

---

* l.i.m. or "limit in the mean" or "mean square (m.s.) convergence" is one of the strongest of the concepts of stochastic convergence [44, p.136 or 45, p.598]. If x_n is a stochastic process and $\underset{n\to\infty}{\text{l.i.m.}}\ x_n = x$ then

$$\lim_{n\to\infty} E\left[(x_n - x)^2\right] = 0 \quad \text{and}$$

$$E\left[(x_n - x_m)^2\right] < \mu \quad \text{for all } n, m > n_o,$$

where $n_o < \infty$ exists for any arbitrarily small but positive μ.

$$E\,[x^2(t)] \;=\; E\,\Big[\int_0^t F(s)\,dw(s)\,\Big]^2 \;=\; \int_0^t F^2(s)\,ds. \qquad \text{(A8)}$$

The integral on the far right hand side of (A8) is a normal Riemann integral, and its existence is a necessary and sufficient condition for the existence of the stochastic integral (A6).

The stochastic integral (A6) is usually presented in the more general form where $F(t)$ is allowed to be a random function $F(w, t)$. Then the definition (A6) becomes

$$x(t) = \int_0^t F(w,s)dw(s) = \text{l.i.m.} \sum_i^n F(w(t_{i-1}),t_{i-1})[w(t_i) - w(t_{i-1})] \qquad \text{(A6a)}$$

The mean value (A7) is unchanged, but the mean square (A8) becomes

$$E\,[x^2(t)] \;=\; \int_0^t E\,[F^2(w,\,s)]\,ds. \qquad \text{(A8a)}$$

The properties (A7, A8, A8a) are a direct consequence of the forward differences used in the definition (A6, A6a) involving a limiting operation on a finite difference representation of the integrand. From (A6a), a typical term in the limiting sum for $E\,[x(t)]$ is $E\,[F(w(t_{i-1}),\,t_{i-1})[w(t_i) - w(t_{i-1})]]$. But $F(w(t_{i-1}),\,t_{i-1})$ depends only on the past $w(s)$, $s \leqslant t_{i-1}$, and so is independent of the forward increment $[w(t_i) - w(t_{i-1})]$. Further, as $w(t)$ is an infinitely divisible random process [20], then this independence is preserved even as the limit of the maximum $\Delta t_i \downarrow 0$ is taken. Then as the $\Delta w(t_i)$ increments (A4) have zero mean, (A7) follows.

Similarly, a typical term in the product sum for $E[x^2(t)]$ is

$$E\left[F(w(t_{i-1}),\, t_{i-1})[w(t_i) - w(t_{i-1})]F(w(t_{j-1}),t_{j-1})[w(t_j) - w(t_{j-1})]\right].$$

Again, owing to the independence of increments of $w(t)$, this term only has a non-zero contribution for $i = j$, in which case, from (A5), the squared $\Delta w(t_i)$ increment can be replaced by its mean value Δt_i. Then the limiting sum becomes the normal Riemann integral (A8) or (A8a).

In a similar fashion, we can define the Ito stochastic integral equation

$$x(t) = \int_0^t F(x(s),s)dw(s) = \text{l.i.m.} \sum_i^n F(x(t_{i-1}),t_{i-1})[w(t_i) - w(t_{i-1})], \qquad \text{(A6b)}$$

where the sum on the right hand side is the limiting forward difference sequence as before, and the integral equation can be solved by the method of Picard iterations.* This process $x(t)$ is still a martingale, that is $E[x(t)] = 0$, but the expression for the mean square

$$E[x^2(t)] = \int_0^t E[\cdot^2(x,\, s)]\, ds \qquad \text{(A8b)}$$

is itself an integral equation and cannot be readily evaluated in general.

---

* See [20, p.279]

A.1.2. The Generalization of the Stochastic Integral - The Stratonovich Form

Noting that the definition of the stochastic integral using the forward difference sum (A6) is rather arbitrary, it is possible to consider other definitions. Astrom [46] and Gray and Caughey [41] have suggested a generalized stochastic integral defined by

$$x(t) = \int_o^t F(w(s), s)\, dw(s)$$

$$= \text{l.i.m.} \sum_i^n F(\theta w(t_{i-1}) + [1 - \theta]w(t_i), t_{i-1})[w(t_i) - w(t_{i-1})], \tag{A9}$$

where $\theta = [0, 1]$. The case $\theta = 1$ is the previous forward difference definition (A6) of the Ito stochastic integral, and the case $\theta = 0$ is a backward difference definition, with other values of θ giving linear combinations of these extreme definitions. We will see that, unlike the Riemann integral, the value of the stochastic integral (A9) depends on the value of θ.

Of these definitions, the most useful is that of the central difference $\theta = \frac{1}{2}$, for it gives the integral the same expected value as the corresponding Riemann integral. This definition has been discussed by Stratonovich [50] and thus we will call

$$x(t) = \int_o^t F(w(s), s)\, \bar{d}w(s)$$

$$= \text{l.i.m.} \sum_i^n F\left(\frac{w(t_{i-1}) + w(t_i)}{2}, t_i\right)[w(t_i) - w(t_{i-1})] \tag{A10}$$

the Stratonovich stochastic integral, where $\bar{d}\cdot$ of (A10) is to be distinguished from $d\cdot$ of (A6) and is a stochastic increment in the Stratonovich sense.

The Stratonovich integral has the same mean square value (A8a) as the Ito integral, but it is no longer a martingale as

$$E[x(t)] = E[\int_0^t F(w(s),s)\bar{d}w(s)] = \tfrac{1}{2}\int_0^t E[\frac{\partial F}{\partial w}(w(s),s)]ds. \qquad (A11)$$

This result can be seen by expanding a typical term in the sum (A10) about $w(t_{i-1})$, and we have

$$F(\frac{w(t_{i-1}) + w(t_i)}{2}, t_i)\,\Delta w(t_i)$$

$$= F(w(t_{i-1}),t_i)\Delta w(t_i)+\frac{\partial F}{\partial w}(w(t_{i-1}),t_{i-1})[w(t_{i-1}+\frac{\Delta t_i}{2})-w(t_{i-1})]\Delta w(t_i)+o(\Delta t_i).$$

The first term on the right hand side is the same term as in the Ito definition and has a zero expected value, but the second term has an expected value

$$E\,[\frac{\partial F}{\partial w}(w(t_{i-1}),\ t_{i-1})\,]\ \tfrac{1}{2}\,\Delta t_i,$$

which leads to the Riemann integral on the right hand side of (A11).

Analogous to (A6b) we can define the Stratonovich integral equation

$$x(t) = \int_0^t F(x(s),\ s)\ \bar{d}w(s), \qquad (A10a)$$

which has the same mean square value (A8b) as the Ito integral

equation, but a mean value of

$$E[x(t)] = \tfrac{1}{2} \int_0^t E[F \frac{\partial F}{\partial x} (x(s), s)] \, ds. \qquad (A11a)$$

A.1.3 Vector Stochastic Integral Equations

The definition of the Ito and Stratonovich integral equations (A6b) and (A10a) as limits of forward and central differences respectively extend directly to the vector case where

$x(t)$ is an n-vector,

$F(x(t), t)$ is an n x m coefficient matrix

and $w(t)$ is the m-vector unit parameter independent Wiener process, with $w(o) = 0$ and for $t > s$, $t' > s'$,

$$E[w_i(t) - w_j(s)] = 0, \quad \text{all } i, j, \qquad (A4')$$

$$E[(w_i(t) - w_i(s))(w_j(t') - w_j(s'))]$$

$$= \text{Minimum } [(t' - s), (t - s')]^* , \quad i = j,$$

$$= 0, \quad i \neq j. \qquad (A5')$$

The Ito integral equation for $x(t)$

$$x(t) = \int_0^t F(x(s), s) \, dw(s),$$

has a zero mean value (that is, $x(t)$ is a martingale) and a

---

* Either of these arguments is set to zero if it is negative.

mean square value

$$E[x(t)x(t)^T] = \int_o^t E[FF^T(x(s), s)]\, ds. \qquad (A8'b)$$

The analogous Stratonovich stochastic integral equation has this same mean square value, but $x(t)$ no longer a martingale, as it has the mean value

$$E[x(t)] = E[\int_o^t F(x(s),s)\, \tilde{d}w(s)]$$

$$= \tfrac{1}{2} \int_o^t E[\sum_{k,l}^m Q_{kl}(s)]\, ds, \qquad (A11'a)$$

where $Q_{kl}(s)$ is the n-vector with the i:th component

$$(Q_{kl})_i = \sum_j^n \frac{\partial F_{ik}(x(s),s)}{\partial x_j} F_{jl}(x(s),s) . \qquad (A11'b)$$

Thus the normal integral

$$\tfrac{1}{2} \int_o^t \sum_{k,l}^m Q_{kl}(s)\, ds$$

is the difference in bias between apparently similar Ito and Stratonovich integral equations, or from another viewpoint, this integral will be the difference between Ito and Stratonovich integral equations for the same diffusion process $x(t)$. This integral must be subtracted from the Ito integral equation to turn it into a Stratonovich one, or added to a Stratonovich integral equation to turn it into an Ito equation for the same process.

A.1.4 Example

A simple example will illustrate that the rules of the Stratonovich stochastic calculus resemble those of the ordinary calculus, but those of the Ito stochastic calculus do not.

Consider the Ito stochastic integral

$$x(t) = \int_0^t w(s)\, dw(s),$$

which, from (A7) has a zero mean value. In comparison, consider the Stratonovich stochastic integral

$$x(t) = \int_0^t w(s)\, \bar{d}w(s),$$

which, from (A11) has a mean value of $\frac{1}{2}t$.

Now if $w(s)$ were a smooth function so that we could interpret either of these integrals as a normal integral (in Stieltjes form), then we would have

$$x(t) = \tfrac{1}{2} w^2(t),$$

which from the property (A5) of the Wiener process has the mean value $\frac{1}{2}t$. The Stratonovich form of the stochastic integral has this same mean value, but the Ito form does not.

A2. Stochastic Differentials and Differential Equations

[20, Ch.6, Sect. 3; 49, Appendix I]

A.2.1 The Ito Stochastic Differential Equation

If we allow $F(s)$ in (A1) to be also a function of $X(s)$, then differentiation of the resulting integral equation leads to the ordinary differential equation

$$\dot{X}(t) = F(X(t), t)\, y(t). \tag{A12}$$

As before, the same equation with $y(t)$ replaced by white noise is not meaningful, as $y(t)$, and hence $\dot{X}(t)$, would be everywhere infinite. However, we gave the resulting equation a meaningful interpretation as a stochastic integral equation (A6b). As in most cases, though, it is more convenient to consider differential instead of integral equations, we will write the stochastic equation (A6b) in symbolic form as

$$dx(t) = F(x(t), t)\, dw(t). \tag{A13}$$

We will discuss the vector case, where $w(t)$ is an m vector of independent unit parameter Wiener processes (A4', A5'), $x(t)$ is an n vector diffusion process, and $F(x(t), t)$ is an $n \times m$ coefficient matrix. We will include a drift term $f(x(t), t)dt$ in (A13), although this term contributes no unusual properties. Then (A13) becomes the Ito stochastic differential equation [47, 48],

$$dx(t) = f(x(t), t)dt + F(x(t), t)\, dw(t). \qquad (A14)$$

the $d\cdot$ operator in equation (A14) is an Ito stochastic differential, and is different from the operator with the same symbol in the ordinary equation (A12). The equation (A14) cannot be divided by dt as $dw(t)$ is $O(dt^{\frac{1}{2}})$, and thus the equation does not specify the value of the time derivative of $x(t)$. Loosely speaking, equation (A14) gives the size of the deterministic and random contributions to $x(t + dt) - x(t)$ during a small time increment dt.

The equation (A14) has a unique solution $x(t)$ which is a continuous diffusion process provided $f(x(t), t)$ and $F(x(t), t)$ are bounded functions of t and satisfy Lipschitz conditions with respect to $x(t)$.

The diffusion process $x(t)$ has the local properties

$$E\,[\Delta x(t) \mid x(t)] = \int_t^{t+\Delta t} f(x(s), s)ds + O(\Delta t^{3/2}), \qquad (A15)$$

$$\text{and} \quad E\,[\Delta x(t)\, \Delta x(t)^T \mid x(t)] = \int_t^{t+\Delta t} F\, F^T(x(s),s)ds + O(\Delta t^2), \qquad (A16)$$

where Δ is a forward difference operator.

These local properties are more commonly expressed in their limiting form as the first and second incremental moments of the diffusion process $x(t)$:

$$b(x, t) = \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t}\, E\,[\Delta x(t) \mid x(t)] = f(x(t), t), \qquad (A17)$$

$$a(x,t) = \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} E[\Delta x(t)\Delta x(t)^T \mid x(t)] = F F^T(x(t),t). \quad (A18)$$

These first and second incremental moments form a unique description of the diffusion process $x(t)$. That higher incremental moments are zero assures the continuity of the diffusion process, and that $a(x, t)$ of (A18) is not zero assures that $x(t)$ is a Markov process.

A.2.2 The Stratonovich Stochastic Differential Equation

In a similar fashion, the Stratonovich stochastic integral equation (A10a) can be written in symbolic form as the Stratonovich stochastic differential equation

$$\bar{d}x(t) = g(x(t),t)dt + F(x(t),t)\bar{d}w(t). \quad (A19)$$

We have included the drift term $g(x(t),t)dt$ and will consider the vector case as in (A14). As we have kept the same noise coefficient matrix $F(x(t),t)$ as in (A14), the diffusion process $x(t)$ described by the Stratonovich equation (A19) shares the same second incremental moment (A18) with the Ito equation (A14), but has the different first incremental moment

$$b(x, t) = g(x(t), t) + \tfrac{1}{2} \sum_{k,l}^{m} Q_{kl}(t), \quad (A20)$$

where $Q_{kl}(t)$ is the n vector given earlier (A11'b). Thus if

the Ito equation (A14) and the Stratonovich equation (A19) are to define the same diffusion process, we must have

$$f(x(t), t) = g(x(t), t) + \tfrac{1}{2} \sum_{k,l}^{m} Q_{kl}(t). \qquad (A21)$$

Thus the quantity

$$\tfrac{1}{2} \sum_{k,l}^{m} Q_{kl}(t)\, dt \qquad (A22)$$

can be regarded as a conversion term which can be added to the Stratonovich equation (A19) to turn it into an Ito equation for the same diffusion process, or subtracted from the Ito equation (A14) to turn it into a Stratonovich equation for the same process. In this thesis, the quantity (A22) is often called the bias term, as it represents the relative bias in the drift terms of equivalent Ito and Stratonovich equations.

A.2.3 Example

The following scalar example [49] will illustrate the differing rules of the Ito and Stratonovich calculus in differential form.

Consider the diffusion process $x(t)$ given by

$$x(t) = e^{w(t)}, \qquad (A23)$$

where $w(t)$ is a scalar unit parameter Wiener process (A4, A5). What is the Ito stochastic differential equation for $x(t)$? Remembering from (A6) that the Ito calculus is based on the forward difference operator, we can interpret the Ito stochastic increment $dx(t)$ as

$$dx(t) = e^{w(t + dt)} - e^{w(t)},$$

where $$e^{w(t + dt)} = e^{w(t) + dw(t)}.$$

Then $$dx(t) = e^{w(t)}[e^{dw(t)} - 1]$$

$$= e^{w(t)}[dw(t) + \frac{(dw(t))^2}{2} + \frac{(dw(t))^3}{6} \ldots.]. \quad (A24)$$

Taking the conditional expectation of both sides we have

$$E[dx(t) \mid x(t)] = x(t)[\tfrac{1}{2} dt + O(dt^2)],$$

and so the first incremental moment (A17) of $x(t)$ is

$$b(x, t) = \tfrac{1}{2} x(t). \quad (A25)$$

Similarly from (A24),

$$(dx(t)^2 = x^2(t)[(dw(t))^2 + (dw(t))^3 + \ldots.],$$

and $$E[(dx(t))^2 \mid x(t)] = x^2(t)[dt + O(dt^2)].$$

Thus the second incremental moment (A18) of $x(t)$ is

$$a(x, t) = x^2(t). \quad (A26)$$

Then by comparison with (A14), (A17) and (A18), we can write down the Ito stochastic differential equation for $x(t)$ as

$$dx(t) = \tfrac{1}{2}\, x(t)\, dt + x(t)\, dw(t). \qquad \text{(A27)}$$

Paralleling this derivation, we see from (A10) that the Stratonovich increment $\bar{d}x(t)$ can be interpreted as a central difference increment, and so from (A23) we have

$$\bar{d}x(t) = e^{w(t + \frac{1}{2}dt)} - e^{w(t - \frac{1}{2}dt)},$$

and as

$$\bar{d}w(t) = w(t + \tfrac{1}{2}dt) - w(t - \tfrac{1}{2}dt),$$

we have

$$\bar{d}x(t) = e^{w(t - \frac{1}{2}dt) + \bar{d}w(t)} - e^{w(t - \frac{1}{2}dt)}$$

$$= e^{w(t - \frac{1}{2}dt)}\left[\bar{d}w(t) + \frac{(\bar{d}w(t))^2}{2} + \frac{(\bar{d}w(t))^3}{6} \ldots\ \right].$$

Now we can write

$$e^{w(t - \frac{1}{2}dt)} = e^{w(t)} - \tfrac{1}{2}\, e^{w(t)}\, \bar{d}w(t) + O(\bar{d}w(t))^2,$$

and so we have

$$\bar{d}x(t) = \left[e^{w(t)} - \tfrac{1}{2}\, e^{w(t)}\, \bar{d}w(t)\right]\left[\bar{d}w(t) + \frac{(\bar{d}w(t))^2}{2}\right] + O(\bar{d}w(t))^3,$$

$$= e^{w(t)}\, \bar{d}w(t) + O(\bar{d}w(t))^3.$$

Then, once again passing to the limit via the system's incremental moments (A20), (A18), we arrive at the Stratonovich stochastic differential equation for $x(t)$

$$\bar{d}x(t) = x(t)\,\bar{d}w(t). \tag{A28}$$

As $F(x(t), t) = x(t)$ in this case, the scalar $a(t) = x(t)$, and the process $x(t)$ of equation (A28) has the first incremental moment (A20)

$$b(x, t) = \tfrac{1}{2}\,x(t), \tag{A29}$$

and from (A18), the second incremental moment is

$$a(x, t) = x^2(t). \tag{A30}$$

Then by comparing incremental moments (A29, 30) with (A25, 26), we see that equations (A27) and (A28) represent the same diffusion process. We could, of course, have written down (A28) from (A27) by simple reference to the conversion term (A21, 22).

Note that if (A27) and (A28) were to be considered as ordinary differential equations (after dividing by dt) and integrated by the normal rules of calculus, the Ito equation (A27) would lead to the wrong solution

$$x(t) = e^{\frac{1}{2}t + w(t)},$$

while the Stratonovich equation (A28) leads to the correct solution

$$x(t) = e^{w(t)}.$$

This example again illustrates that Stratonovich equations can be manipulated according to the rules of the calculus of smooth functions, while Ito equations cannot be handled by these rules [50]. However, the Ito form has the advantage that a system's incremental moments are expressed explicitly in terms of the coefficients of the differential or integral equation, and so we will usually prefer to use the Ito form of stochastic equations.

APPENDIX B

THE NORMALIZED FP EQUATION

(see Section 2.4.4)

Consider the scalar stochastic differential equation

$$dx(t) = b(x, t)dt + F(x, t)\, dw(t) \tag{B1}$$

where $a(x, t) = F^2(x, t)$, and $b(x, t)$ and $a(x, t)$ are the first and second incremental moments of the diffusion process (B1).

$$\text{Let } m(t) = E\,[x(t)] \tag{b2}$$

$$\text{then } dm(t) = d\,E\,[x(t)] = E\,[dx(t)]$$

$$= E\,[b(x, t)]\, dt. \tag{B3}$$

Thus the differential equation for $m(t)$ is

$$\dot{m}(t) = E\,[b(x, t)]. \tag{B4}$$

$$\text{Let } v(t) = E\,[x^2(t)]. \tag{B5}$$

$$\text{Then } dv(t) = d\,E\,[x^2(t)],$$

$$= E\,[d(x^2(t))], \tag{B6}$$

$$= E\left[\frac{\partial}{\partial x}(x^2(t))\, b(x, t)\right] dt$$

$$+ \tfrac{1}{2} E\left[\frac{\partial^2}{\partial x^2}(x^2(t))\, a(x, t)\right] dt, \tag{B7}$$

$$= 2\,E\,[x\, b(x, t)]\, dt + E\,[a(x, t)]\, dt. \tag{B8}$$

Thus the differential equation for $v(t)$ is

$$\dot{v}(t) = 2\,E\,[x\,b(x,\,t)] + E\,[a(x,\,t)]. \tag{B9}$$

For a discussion of the rules of the Ito stochastic calculus which lead to the formulae (B3) and (B7), the reader is referred to Appendix A, or more specifically to [74].

Let $\sigma(t)$, the standard deviation of $x(t)$, be defined as

$$\sigma(t) = [v(t) - m^2(t)]^{\frac{1}{2}}. \tag{B10}$$

If we assume the variance $\sigma^2(t)$ is never zero, we can define the normalized or standardized variable $y(t)$ by the transformation

$$y(t) = \frac{x(t) - m(t)}{\sigma(t)}\,. \tag{B11}$$

We will derive the stochastic differential equation for $y(t)$. Noting

$$x(t) = \sigma(t)\,y(t) + m(t), \tag{B12}$$

we have (dropping the t parameter)

$$dx = \sigma\,dy + y\,d\sigma + dm, \tag{B13}$$

as $\sigma(t)$ has no dispersion. Thus

$$dy = \sigma^{-1}\,[dx - y\,d\sigma - dm]. \tag{B14}$$

From (B10) we have

$$d\sigma = \tfrac{1}{2}\sigma^{-1}\,[dv - 2\,m\,dm], \tag{B15}$$

and so

$$dy = \sigma^{-1} [b(x, t)\, dt + F(x, t)\, dw - y(\dot{v} - 2\, m\, \dot{m})\, \tfrac{1}{2}\sigma^{-1}\, dt - \dot{m}\, dt]. \tag{B16}$$

Now if we assume m, v, σ, $\dot{m}$, $\dot{v}$ are known functions of time (in reality they must be obtained simultaneously with the statistics of $y(t)$) then we can write the first and second incremental moments of the diffusion process (B16) as

$$\underline{b}(y, t) = q_1\, b(\sigma y + m, t) + q_2\, y + q_3, \tag{B17}$$

$$\underline{a}(y, t) = q_1^2 a(\sigma y + m, t), \tag{B18}$$

where $b(x, t) = b(\sigma y + m, t)$ is the first incremental moment of the $x(t)$ process (B1), and $a(x, t) = a(\sigma y + m, t)$ is the second. The quantities q_1, q_2 and q_3 are the time functions

$$q_1(t) = \sigma^{-1}(t), \tag{B19a}$$

$$q_2(t) = -\tfrac{1}{2}\, \sigma^{-2}(t)\, [\dot{v}(t) - 2\, m(t)\, \dot{m}(t)], \tag{B19b}$$

and $$q_3(t) = -\sigma^{-1}(t)\, \dot{m}(t). \tag{B19c}$$

It is noted that $\underline{b}$ and $\underline{a}$ are polynomials in y if b and a were polynomials in x (and of equal degree in y).

Let $Q(y, t)$ be the density function of the normalized variable $y(t)$. As $y(t)$ is normalized, $Q(y, t)$ will have a concise representation in the Hermite polynomial expansion

$$Q(y, t) = \sum_r c_r(t)\, H_r(y)\, G(y). \tag{B20}$$

$Q(y, t)$ obeys the FP equation (2.4.41) associated with the incremental moments (B17, 18), and it will be convenient to obtain a numerical solution for $Q(y, t)$ by the Hermite transform method of Section 2.4.3.

Using the parameters values $b = b_3 x^3$ and $a = a_o$ and noting that

$$\frac{\partial b}{\partial y} = 3 q_1 b_3 (\sigma y + m)^2 + q_2,$$

and $$\frac{\partial a}{\partial y} = \frac{\partial^2 a}{\partial y^2} = 0,$$

we have from (2.4.29)

$$\begin{aligned}
\sum_r \dot{c}_r H_r G &= -[3 q_1 b_3 (\sigma y + m)^2 + q_2] \sum_r c_r H_r G \\
&+ [q_1 b_3 (\sigma y + m)^3 + q_2 y + q_3] \sum_r c_r H_{r+1} G \\
&+ \tfrac{1}{2} q_1^2 a_o \sum_r c_r H_{r+2} G. \qquad \text{(B21)} \\
&= [-3 b_3 \sigma^2 y^2 - 6 b_3 \sigma m y - 3 b_3 m^2 - q_2] \sum_r c_r H_r G \\
&+ [b_3 \sigma^2 y^3 + 3 b_3 \sigma m y^2 + (3 b_3 m^2 + q_2) y + q_1 b_3 m^3 + q_3] \sum_r c_r H_{r+1} G \\
&+ \tfrac{1}{2} q_1^2 a_o \sum_r c_r H_{r+2} G.
\end{aligned}$$

Using the relations (2.4.26j) for $y^s H_r(y)$ and collecting terms of equal H index together, we have

$$\sum_r \dot{c}_r H_r G = [-3 b_3 \sigma^2 r(r-1) + (r+1) r (r-1) b_3 \sigma^2] \sum_r c_r H_{r-2} G$$

$$+ [-6mb_3\sigma r + (r+1)r3b_3m\sigma] \sum_r c_r H_{r-1} G$$

$$+ [-3b_3\sigma^2(2r+1) - 3b_3m^2 - q_2 + 3(r+1)^2 b_3\sigma^2 + (r+1)(3b_3m^2 + q_2)] \sum_r c_r H_r G$$

$$+ [-6mb_3\sigma + (2r+3)3b_3m\sigma + q_1 b_3 m^3 + q_3] \sum_r c_r H_{r+1} G$$

$$+ [-3b_3\sigma^2 + 3(r+2)b_3\sigma^2 + 3b_3m^2 + q_2 + \tfrac{1}{2}q_1^2 a_o] \sum_r c_r H_{r+2} G$$

$$+ [3b_3m\sigma] \sum_r c_r H_{r+3} G$$

$$+ [b_3\sigma^2] \sum_r c_r H_{r+4} G. \quad \text{(B22)}$$

Multiplying by $(s!)^{-1} H_s(y)$ and integrating over the infinite y range, we use the orthogonality relations (2.4.2) to eliminate H and G from (B22), and obtain the simultaneous non-linear equations for the coefficients $c_s(t)$,

$$\dot{c}_s = [-3(s+2)(s+1)b_3\sigma^2 + (s+3)(s+2)(s+1)b_3\sigma^2] c_{s+2}$$

$$+ [-6(s+1)b_3m\sigma + 3(s+2)(s+1) b_3m\sigma] c_{s+1}$$

$$+ [-3(2s+1)b_3\sigma^2 - 3b_3m^2 - q_2 + 3(s+1)^2 b_3\sigma^2 + (s+1)(3b_3m^2 + q_2)] c_s$$

$$+ [-6b_3m\sigma + 3(2s+1)b_3m\sigma + q_1 b_3 m^3 + q_3] c_{s-1}$$

$$+ [-3b_3\sigma^2 + 3sb_3\sigma^2 + 3b_3m^2 + q_2 + \tfrac{1}{2} q_1^2 a_o] c_{s-2}$$

$$+ [3b_3m\sigma] c_{s-3} + [b_3\sigma^2] c_{s-4} \quad , \quad s = 0, 1, 2 \ldots, \quad \text{(B23)}$$

remembering that these c_i whose indices are negative or above the truncation point are set to zero. These equations must be solved simultaneously with those for $m(t)$ and $v(t)$, (B4) and (B9). The latter equations can be solved numerically by using the expansion for $P(x, t)$ and Hermite quadrature. From (B4) we have

$$\dot{m}(t) = \int_{-\infty}^{\infty} b(x, t)\, P(x, t)\, dx. \qquad \text{(B24)}$$

Now $P(x, t)$ can be written as

$$P(x, t) = \sigma^{-1}(t)\, Q\left(\frac{x - m}{\sigma}, t\right), \qquad \text{(B25)}$$

$$= \sigma^{-1}(t) \sum_r c_r H_r\left(\frac{x - m}{\sigma}\right) G\left(\frac{x - m}{\sigma}\right). \qquad \text{(B26)}$$

Thus
$$\dot{m}(t) = \sum_r \int_{-\infty}^{\infty} b(x, t)\sigma^{-1}(t)c_r(t)H_r\left(\frac{x - m}{\sigma}\right)G\left(\frac{x - m}{\sigma}\right) dx,$$

$$= \sum_r \int_{-\infty}^{\infty} b(\sigma y + m, t)c_r(t)H_r(y)G(y)\, dy, \qquad \text{(B27)}$$

as $dx = \sigma\, dy$. ($\sigma(t)$ and $m(t)$ are constant during the integral.)

Now (B27) is in the form (2.4.20) for Hermite quadrature and so

$$\dot{m}(t) = (2\pi)^{\frac{1}{2}} \sum_r c_r(t) \sum_{s=1}^{N} h_s[b(\sigma y_s + m, t)H_r(y_s)]. \qquad \text{(B28)}$$

From (B9) we have

$$\dot{v}(t) = 2 \int_{-\infty}^{\infty} x\, b(x, t)P(x, t)\, dx + \int_{-\infty}^{\infty} a(x, t)P(x, t)\, dx,$$

$$= \sum_r c_r(t) \int_{-\infty}^{\infty} [2xb(x, t) + a(x, t)]H_r\left(\frac{x-m}{\sigma}\right) G\left(\frac{x-m}{\sigma}\right)\sigma^{-1}\, dx,$$

$$= \sum_r c_r(t) \int_{-\infty}^{\infty} [2(\sigma y + m)b(\sigma y + m, t) + a(\sigma y + m, t)]H_r(y)G(y)\, dy,$$

$$= (2\pi)^{\frac{1}{2}} \sum_r c_r(t) \sum_{s=1}^{N} h_s[2(\sigma y_s+m)b(\sigma y_s+m,t) + a(\sigma y_s+m,t)]H_r(y_s) \tag{B29}$$

Now, substituting in the parameters of our example, $b(x, t) = b_3x^3$ and $a(x, t) = a_o$, we have the equations

$$\dot{m}(t) = (2\pi)^{\frac{1}{2}} \sum_r c_r(t) \sum_{s=1}^{N} h_s b_3(\sigma(t)y_s + m(t))^3 H_r(y_s),$$

$$\dot{v}(t) = (2\pi)^{\frac{1}{2}} \sum_r c_r(t) \sum_{s=1}^{N} h_s[2b_3(\sigma(t)y_s+m(t))^4+a_o]H_r(y_s),$$

and

$$\sigma(t) = [v(t) - m^2(t)]^{\frac{1}{2}}. \tag{B30}$$

APPENDIX C

Calculation of Flux Hitting Boundary in Heat Conduction Problem [Section 3.2.5]

Consider a simulation of the heat conduction equation with the parameters of Case III, equation (3.2.72). The dynamics of the particles in the simulation are given by

$$dx_i(t) = (2K_i)^{\frac{1}{2}} dw_i(t) \quad , \quad i = 1, 3 \tag{C1}$$

where K_2 and K_3 are constants and K_1 is constant except across the discontinuity in Figure 3.2.2. The dynamics in each coordinate axis are independent of each other, and so the motion of the particle in the x_1 direction is given only by

$$dx_1(t) = (2K_1)^{\frac{1}{2}} \, dw_1(t). \tag{C2}$$

Let us calculate the number of trajectories which hit the boundary in Figure 3.2.2. from the left in Δ seconds. From the properties (A7) and (A8) of the Ito s.d.e. (C2), in time Δ the change in x_1 is a Gaussian random variable with zero mean and variance

$$\sigma^2 = 2 K_1 \Delta. \tag{C3}$$

Consider a particle to the left of and a distance y from the boundary. Then the probability that it will cross the boundary in the next Δ seconds is

$$\text{Prob}[\Delta x_1 \geq y] = \int_y^{\infty} \frac{1}{(2\pi)^{\frac{1}{2}}\sigma} e^{-\frac{z^2}{2\sigma^2}} dz \, ,$$

$$= \text{erf}(\sigma, y). \tag{C4}$$

Now if $D(y)$ is the density of particles at a distance y from the boundary, then the number of particles hitting the boundary per cm^2 in Δ seconds is

$$\Omega^- = \int_0^\infty D(y)\ \mathrm{erf}\ (\sigma^-, y)\ dy. \tag{C5}$$

Now from Figure 3.2.2 we can set $D(y) = D - y\ D^-_{x_1}$. (C6)

Then

$$\Omega^- = \int_0^\infty D\ \mathrm{erf}(\sigma,y)dy \ - \int_0^\infty y\ D^-_{x_1}\ \mathrm{erf}(\sigma,y)dy,$$

and integrating once by parts we obtain

$$\Omega^- = -\int_0^\infty D\ y\ d(\mathrm{erf}(\sigma,y)) \ + \ \tfrac{1}{2}\int_0^\infty y^2\ D^-_{x_1}\ d(\mathrm{erf}(\sigma,y)).$$

But $d(\mathrm{erf}(\sigma,y)) = -G(\sigma,y)\ dy$, where $G(\sigma,y)$ is a zero mean, σ^2 variance, distribution of the Gaussian random variable y. Then

$$\Omega^- = \int_0^\infty D\ y\ G(\sigma,y)\ dy \ - \ \tfrac{1}{2}\int_0^\infty y^2\ D^-_{x_1}\ G(\sigma,y)\ dy$$

$$= D\ \sigma\ (2\pi)^{-\frac{1}{2}} \ - \ \frac{1}{4}\ D^-_{x_1}\ \sigma^2$$

from the properties of the Gaussian distribution. Substituting in the value of $\sigma^-_{,\Delta}$ (c3), we find the number of particles hitting the boundary from the left per cm^2 in Δ seconds is

$$\Omega^- = \left(\frac{K_1^- \Delta}{\pi}\right)^{\frac{1}{2}} D \ - \ \tfrac{1}{2}\ K_1^-\ \Delta\ D^-_{x_1}\ . \tag{C7}$$

The first term of (C7) is the number of particles hitting the boundary due to the density D at the boundary, and the second term is the number due to the density gradient $D^-_{x_1}$ at the left hand side of the boundary.

Following through a similar argument for the particles to the right of the boundary in Figure 3.2.2, we find the number of particles hitting the boundary from the right per cm^2 in Δ seconds is

$$\Omega^+ = \left(\frac{K_1^+ \Delta}{\pi}\right)^{\frac{1}{2}} D + \tfrac{1}{2} K_1^+ \Delta D_{x_1}^+ , \qquad (C8)$$

as $D(y)$ is now $D + y D_{x_1}^+$.

APPENDIX D

A White Noise Model of a Pseudo Random Binary Sequence

NOTE

This Appendix was initially prepared as a separate report, and so it is self-contained in its use of symbols, references, and numbering of equations, pages, sections and appendices. In this respect, continuity is lost with the rest of the thesis.

A WHITE NOISE MODEL OF A PSEUDO RANDOM BINARY SEQUENCE

by I. G. CUMMING

Centre for Computing and Automation,
Imperial College,
London, S.W.7.

MARCH 1967

Abstract

A maximal-length pseudo random binary sequence (PRBS) is modelled by a modified white noise, and the model is useful for obtaining the transient statistics of integral functions of the PRBS, such as the output of a dynamic system with a PRBS input. As an example, the transient mean and mean square of a filtered PRBS are derived, and an exact simulation verifies that these quantities are determined accurately by our method. This example is of current interest because of the possibility of using a filtered PRBS to approximate a low frequency Gaussian noise source in simulation exercises, and the transient statistics derived in this paper help to compare the filtered PRBS with Gaussian noise when they are used in transient situations such as in a repetitive mode simulation.

The method of transient analysis used here arises from a new method of approximating continuous non-Markovian processes by continuous Markov processes. The method appears to be a useful way of deriving the transient statistics of non-Markovian systems.

GLOSSARY OF SYMBOLS

A	amplitude of PRBS
a	inverse of filter time constant T
e	the exponential constant
E[.]	expectation operator
f,F,G	arbitrary scalar functions of a random variable x and time in Appendix A
G(t)	time function of a PRBS realisation
L	number of bits in maximum length PRBS = $2^M - 1$
M	order of PRBS = number of shift register stages needed to generate PRBS
m.(t)	mean value (e.g. of y(t))
n,N	non-negative integers
O(.)	"of order equal to"
o(.)	"of order higher than"
P	a probability
p	$= \frac{d}{dt}$, used as 1/p in Figure 1 to denote an integrator
r,u,v	dummy integration parameters in Appendices
r(t)	white noise model of PRBS
S(t)	an integral used in urn model of PRBS
s(t)	diffusion model of S(t)
T	time constant of first order linear filter
t	time
t*	t (modulo LΔ)
v.(t)	mean square value (e.g. of y(t))
w(t)	unit parameter Wiener process

$x(t)$	a diffusion process
$y(t)$	model of filtered PRBS
$z(t)$	ideal white noise
β	a step function of time, defined in equation (B7')
$\delta\cdot$	an increment operator
Δ	duration of one bit of PRBS
τ	time shift
PRBS	a maximum-length pseudo random binary sequence

INDEX

1. Introduction

In a recent paper, Roberts and Davis (1966) show how a filtered maximal-length pseudo random binary sequence (PRBS) can be a useful approximation to band limited Gaussian noise. Thus a filtered PRBS is a convenient noise source for system test signals or simulation exercises, and one must relate the variance of the filtered PRBS to the noise source it is simulating (which could be considered as the same filter being driven by white noise). In a later note, Roberts (1966) shows that if the PRBS is used to replace white noise directly (with equal low frequency power spectral density) an error in the steady state variance of the filtered noise will occur, particularly if the filter has a long time constant T compared with the PRBS bit interval Δ.

One of the most attractive features of using the PRBS for simulation purposes is its repeatability, as a typical simulation exercise might involve repeated trials on an iterative system using the same noise record. As well as the digital computer implementation mentioned by Roberts (1966), this type of repetitive simulation is conveniently accomplished on an analogue computer, as the reset-integrate modes can be easily slaved to a reference bit of the cyclic PRBS.

With this application in mind, it will be useful to know the transient statistics of the filtered PRBS. As a continuous signal, the PRBS is non-Markovian, and as such it is difficult to obtain equations for the transient statistics of functions of the PRBS. Using a new method, we construct a continuous Markov process (a diffusion process, or a process driven by ideal white noise) which accurately models integral functions of the PRBS.

A continuous Markov process has advantages for transient analysis, as differential equations are readily obtained for the statistics of the process.

Thus our method of choosing a diffusion process which approximates a non-Markovian process (physical process) constitutes a useful tool of analysis for non-Markovian processes.

We use the diffusion model to derive the transient mean and mean square (and thus variance) of the filtered PRBS. We find that these transient statistics have interesting properties if the significant memory of the filter is as long as the period of the PRBS. An exact simulation of the filtered PRBS is used to show the accuracy of the diffusion model. The model also provides a comparison between the PRBS and white noise.

2. Construction of Model of the PRBS

Urn mechanism

The pseudo random binary sequence is a deterministic sequence which has some of the properties of a proper random sequence. We begin by constructing a discrete random process which preserves the most striking deterministic property of maximum length pseudo random sequences - the property that the total number of positive and negative bits in one period of the sequence is known. A random process which preserves this property is the mechanism of drawing labelled balls from an urn without replacement.

We consider a maximum length PRBS generated by an M stage shift register, with bits of amplitude + A and - A held during the bit interval of Δ time

units. The sequence contains $L = 2^M - 1$ bits and is periodic with a period of $L\Delta$ time units. In each period there are 2^{M-1} positive bits and $2^{M-1} - 1$ negative bits, which gives the PRBS a mean value of $\frac{A}{L}$. It is convenient to assume for the moment that the sequence has an equal number of positive and negative bits, and thus has a zero mean value. This assumption, which is removed later without any ensuing error, considerably simplifies the equations to follow.

Consider an urn which initially contains L balls, half of which are labelled + A and half labelled - A. We draw the balls randomly from the urn every Δ time units and do not replace them. The random sequence of length $L\Delta$ so produced has considerably more degrees of freedom in terms of possible outcomes than the pseudo random sequence (e.g. the PRBS has a maximum run of M bits while the urn mechanism may have a run of $\frac{1}{2}$ L bits), but it turns out that the statistical properties of such a sequence closely model those of the PRBS.

The fundamental property of the urn mechanism is that at any given time the probability of the next ball (bit) being + A depends on the ratio of positive and negative balls left in the urn at that time, which is also characterised by the sum of positive and negative balls already drawn. At time $n\Delta$, define the "sum" $S(n\Delta)$ as

$$S(n\Delta) = \int_0^{n\Delta} G(t)\, dt \quad , \qquad n = 0, 1, 2, \ldots . L \quad , \qquad (1)$$

where $G(t) = \pm A$ represents the ball drawn at each sampling instant $n\Delta$ and is constant over the next Δ time units. Then at time $n\Delta$ there are $L - n$ balls remaining in the urn, of which there are $\frac{S(n\Delta)}{A\Delta}$ more - A balls than + A balls. Thus we have

$$P(n\Delta) = \text{Prob}\,[G(t) = +A \mid S(n\Delta)] \quad , \quad t = (n\Delta , n\Delta + \Delta) ,$$
$$= \frac{1}{2} - \frac{S(n\Delta)}{2A\Delta(L - n)} \quad . \qquad (2)$$

We note the following points about the urn mechanism:

1. The urn mechanism is a discrete Markov process as its probabilistic dynamics are given by the present state of the process $S(n\Delta)$ in equation (2). Later an analogous quantity $s(t)$ will be the state of the continuous Markov process which approximates to this discrete process.

2. $S(L\Delta) = 0$ with probability one.

3. The effect of the urn mechanism is to randomise the starting point of the PRBS. That is, in the transient situations which we consider in the sequel, the starting point of the PRBS is not known a priori.

Diffusion Model

Although $S(n\Delta)$ is a discrete Markov process when considered at the sample points $t = n\Delta$, $n = 1, 2 \ldots$, when considered as a continuous function of time $S(t)$, it has the characteristics of a non-Markovian process (e.g. it is differentiable almost everywhere). For the purposes of transient analysis, we wish to model $S(t)$ by a continuous Markov or diffusion process $s(t)$.

A diffusion process is a random process obtained by exciting a continuous dynamic system with ideal white noise. As white noise is always infinite in amplitude, integral functions of white noise such as diffusion processes are never differentiable and so cannot be described by ordinary differential equations. Instead we use the stochastic calculus, and we describe diffusion processes by Ito stochastic differential equations (s.d.e.), whose relevant properties are given in Appendix A.

In particular, a diffusion process is completely specified by two infinitesimal incremental properties (A5) and (A6), which are often expressed as the first two incremental moments (A5') and (A6'). A continuous non-Markovian

process does not have a second incremental moment (A6') as the right hand side of equation (A6) is $O(\delta t^2)$ in that case. However, quantities analogous to the incremental moments of a diffusion process can be defined for a continuous non-Markovian process by considering the increments (A5) and (A6) over a finite time increment and not taking the limits indicated in (A5') and (A6'), Then we can construct a diffusion process to model the continuous non-Markovian process by matching the finite increments (A5) and (A6) of the two processes. This technique of forming diffusion models for physical processes is thought to be new, and is described in a more general context by Cumming 1967 c.

In the present example, the diffusion process $s(t)$ is matched to the non-Markovian process $S(t)$ by matching the finite increments (A5) and (A6) over the time increment Δ, the PRBS bit interval. This time increment is chosen as this is the minimum time increment over which the PRBS (or urn mechanism) exhibits random properties. This is because the discrete urn mechanism only undergoes random changes at the time points $t = n\Delta$, and we see later that the diffusion model approximates the integral properties of the PRBS most accurately at the sample points $t = n\Delta$.

Define the finite increment of $S(t)$ over a time increment Δ from time $t = n\Delta$ as

$$\delta S(n\Delta) = S((n+1)\Delta) - S(n\Delta). \qquad (3)$$

Then from equation (1), $\delta S(n\Delta) = \pm A\Delta$ with the associated probabilities given by $P(n\Delta)$ of equation (2) and $1 - P(n\Delta)$. Then corresponding to the finite increments (A5) and (A6), we have

$$E[\delta S(n\Delta) \mid S(n\Delta)] = A\Delta(2P(n\Delta) - 1)$$

$$\simeq -\frac{S(n\Delta)}{L\Delta - n\Delta}\Delta , \qquad (4)$$

and $$E\left[(\delta S(n\Delta))^2 \,\middle|\, S(n\Delta)\right] = (A^2\Delta)\,\Delta\,. \qquad (5)$$

Then ignoring the error terms given in (A5) and (A6), the diffusion process $s(t)$ with the increments (4) and (5) has the coefficients

$$f(s,t) = -\frac{s(t)}{L\Delta - t}\,, \qquad (6)$$

and $$F^2(s,t) = A^2\Delta\,. \qquad (7)$$

Then, referring to (A1), the diffusion process $s(t)$ is described by the Ito s.d.e.

$$ds(t) = -\frac{1}{L\Delta - t}\,s(t)\,dt + A\,\Delta^{\frac{1}{2}}\,dw(t)\,,$$
$$s(o) = o\,,\quad t = [o, L\Delta]\,, \qquad (8)$$

where $w(t)$ is a unit parameter Wiener process, and it is convenient to consider $\dot{w}(t) = \frac{dw(t)}{dt}$ as white noise (with unit power spectral density in [power per cycle per unit time] units). The stochastic integral (Doob, 1953) equivalent to equation (8) is

$$s(t) = A\Delta^{\frac{1}{2}}(t - L\Delta)\int_o^t (u - L\Delta)^{-1}\,dw(u)\,,\quad t = [o, L\Delta]\,, \qquad (9)$$

which illustrates that the diffusion process $s(t)$ which models the integral of the zero mean PRBS is an integral function of white noise (replace $dw(u)$ by $\dot{w}(u)\,du$ in equation (9)).

White Noise Model of the PRBS

The model of the zero mean PRBS is the derivative of $s(t)$ which can be obtained by formally dividing the Ito s.d.e. (8) by dt. At this stage we add the positive bias $\frac{A}{L}$ which we had neglected in constructing the urn

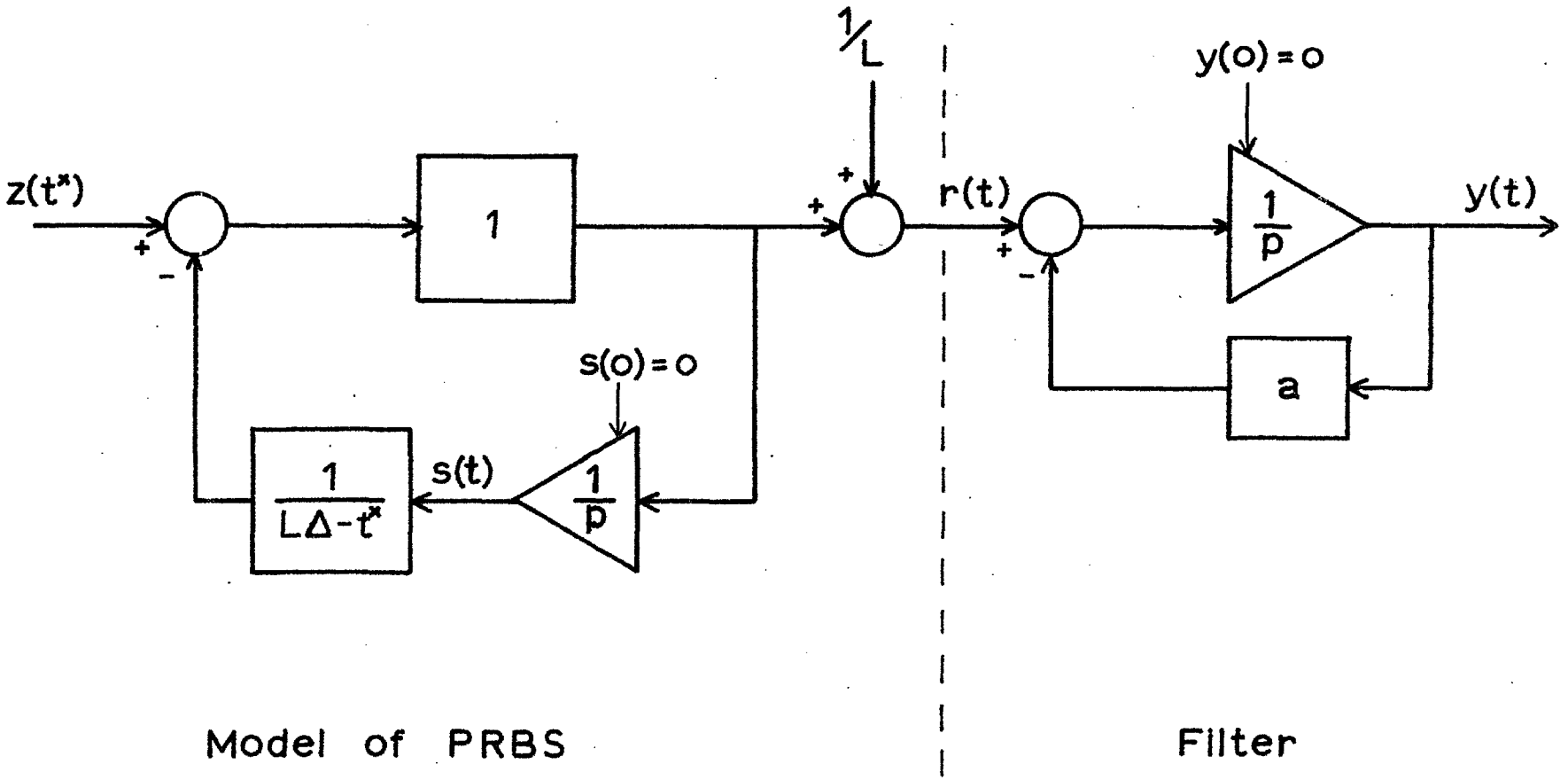


Fig 1. Continuous Model of a Filtered PRBS

mechanism, and we make the model valid for all positive time by imposing the periodicity property of the PRBS. This is done by introducing the cyclic time parameter

$$t^* = t \text{ (modulo } L\Delta \text{)},$$

and we shall see later that $s(t) = s(t^*)$.

The PRBS is then modelled by the modified white noise

$$r(t) = -\frac{s(t)}{L\Delta - t^*} + \frac{A}{L} + z(t^*), \quad t \geq 0 \qquad (10)$$

where $s(t)$ is defined by equations (8) or (9) and is the integral of $r(t) - \frac{A}{L}$, and $z(t^*) = A\,\Delta^{\frac{1}{2}}\,\dot{w}(t^*)$ is a periodic white noise with a flat power spectral density of $A^2\,\Delta$. An ideal circuit generating $r(t)$ is given in Figure 1, where the value $A = 1$ is used.

Diffusion Model of the Filtered PRBS

The model of the PRBS, $r(t)$, is a modified white noise, and integral functions of $r(t)$ are diffusion processes. Because of the method of constructing the model by forming a discrete urn model $S(t)$ for the integrated PRBS and then forming a diffusion approximation $s(t)$ for $S(t)$, the white noise model $r(t)$ is not meant to model the fine structural properties of the PRBS, but integral functions of $r(t)$ are meant to model integral functions of the PRBS. Such a function is the output of a filter driven by a PRBS.

Consider the output of a linear first order filter of time constant $T = a^{-1}$ whose input is a PRBS. This situation is the same as that considered by Roberts and Davis (1966) except that our filter has a gain of T over their filter. The diffusion model $y(t)$ of the filter output is obtained by passing $r(t)$ of equation (10) through the filter $\frac{1}{p+a}$ as shown in Figure 1. The Ito

s.d.e. for the diffusion process $y(t)$ is

$$dy(t) = \left(- a\, y(t) - \frac{s(t)}{L\Delta - t^*} + \frac{A}{L} \right) dt + A\Delta^{\frac{1}{2}}\, dw(t^*) ,$$

$$t \geq o \quad , \quad y(o) = y_o \, , \tag{11}$$

which must be solved simultaneously with equation (8) remembering that $s(t) = s(t^*)$.

The equivalent stochastic integral for $y(t)$ is $y(t) = \int_o^t e^{-a(t-u)}\, r(u)\, du$,

$$= \int_o^t e^{-a(t-u)} \left[- \frac{s(u)}{L\Delta - u^*} + \frac{A}{L} \right] du$$

$$+ \int_o^t e^{-a(t-u)}\, A\Delta^{\frac{1}{2}}\, dw(u^*) \, , \quad t \geq o \, , \tag{12}$$

where $u^* = u$ (modulo $L\Delta$). Equation (12) illustrates that $y(t)$ is an integral function of $r(t)$, or that the filter output is an integral function of the PRBS.

3. Statistics of the Integrated and Filtered PRBS

By the methods discussed in Appendix A, the statistics of $s(t)$ and $y(t)$ can be obtained from their stochastic equations, (8) and (11) or (9) and (12). As $s(t)$ and $y(t)$ are linear functions of white noise, they are Gaussian and only their first two moments need be evaluated.

Statistics of Integral of PRBS

For convenience we consider the statistics of $s(t)$ of equations (8) or (9), which is the integral of the PRBS less the bias $\frac{A}{L}$.

Let $m_s(t) = E\left[s(t) \,|\, s(o) = 0 \right]$.

Then from equation (A12), dropping the t parameter,

$$\hat{m}_s = - \frac{m_s}{L\Delta - t} \quad , \quad m_s(o) = o \, , \quad t = [o, L\Delta] \, .$$

from which we have $m_s(t) = o \, , \quad t = [o, L\Delta] \, .$ (13)

This result is also easily deduced by applying equation (A7) to (9), as $E[f(s, u)] = o$ in the stochastic integral (9). The mean value of the integral of the model of the PRBS with bias $\frac{A}{L}$ equals $\frac{A}{L} t$.

Now let $v_s(t) = E[s^2(t) \mid s(o) = o]$.

Then from equation (A13) we have

$$\hat{v}_s = \frac{- 2 v_s}{L\Delta - t} + A^2 \Delta \, , \quad v_s(o) = o, \quad t = [o, L\Delta] \, ,$$

which has the solution

$$v_s(t) = A^2\Delta \left(1 - \frac{t}{L\Delta}\right) t \quad , \quad t = [o, L\Delta] \, . \qquad (14)$$

Equation (14) is also the variance of the integral of the zero mean PRBS or of the PRBS with bias $\frac{A}{L}$. The curve $v_s(t)$ is a parabola with a maximum value of $\frac{A^2\Delta^2 L}{4}$ occurring at the half period $t = \frac{1}{2} L\Delta$. It can be compared with the variance $(A^2\Delta) t$ of the Wiener process or Brownian motion obtained by deleting the urn model property represented by the feedback $- \frac{s(t)}{L\Delta - t}$ in equation (8). Thus the integral of the PRBS is only a good approximation to a Brownian motion or random walk for values of time t much less than the period $L\Delta$.

From equation (14) we note that $v_s(L\Delta) = 0$ which means that $s(L\Delta) = 0$ with probability one, as we should expect from the cyclic probability of the urn mechanism. Thus as the input noise $\hat{w}(t)$ of equation (8) is periodic, it is easy to deduce that $s(t)$ is periodic, that is $s(t) = s(t^*)$. Then equations (8), (9), (13) and (14) can be made valid for $t > L\Delta$ by replacing

t and u on the right hand sides by t^* and u^*.

It is noted that the PRBS amplitude A is simply a scaling factor appearing linearly in the forcing functions of the equations for $s(t)$ and $r(t)$, equations (8), (9), (11) and (12), and will appear as a factor A^2 in all variance or mean square equations such as (14). With this in mind, it is convenient to set $A = 1$ in all subsequent equations.

Statistics of the Filtered PRBS

In Appendix B we have derived the mean and mean square of $y(t)$, the model of the filtered PRBS, assuming $y_o = 0$. A similar analysis could be carried out for any dynamic system driven by the PRBS, but as we see from the comments following equation (A13), a direct solution cannot be obtained for the moment equations if the system is non-linear.

The main results are the expressions for the mean and mean square of the filtered PRBS derived from the continuous diffusion model. The expression for the mean $m_y(t)$ is given as

$$m_y(t) = \frac{1}{aL}\,(1 - e^{-at}) \quad , \quad t \geq 0, \qquad \ldots (B2)^\dagger$$

and is due solely to the bias $\frac{1}{L}$ entering the first order filter. No unusual effects are caused by the periodic or urn mechanism properties of the PRBS.

The expression for the mean square $v_y(t)$ is

$$v_y(t) = \frac{2}{a^2L}\,[(\tfrac{1}{L} - 1)(1 - e^{-at}) + (\tfrac{1}{2} - \frac{1}{2L} + \frac{aL\Delta}{4})(1 - e^{-2at})] \quad , \quad t = [0, L\Delta], \quad (B4)$$

---

† The equation numbers (B2), (B4) etc. refer to Appendix B.

and for $t > L\Delta$, $v_y(t)$ involves functions of this initial cycle curve, and is given by

$$v_y(t) = v_y(t^*)\,[1 + 2\,\beta(N)\,e^{-aL\Delta}] + v_y(I\Delta)\,\beta^2(N)\,e^{-2at^*}$$

$$+ \frac{2}{a^2 L}\,\beta(N)\,\left(\frac{1}{L} - 1\right)\,(1 - e^{-at^*})\,(e^{-at^*} - e^{-aL\Delta}), \qquad \ldots\ (B10)$$

where $t = \left[NL\Delta,\ (N+1)L\Delta\right)$, $t^* = t$ (Modulo $L\Delta$),

and

$$\beta(N) = \frac{1 - e^{-aNL\Delta}}{1 - e^{-aL\Delta}}, \quad N = 0, 1, 2 \ldots \qquad \ldots\ (B7')$$

This expression is plotted in Figure 2 for $L = 127$, $\Delta = 1$ and various values of the filter time constant. It is convenient to parameterise the time constant as a multiple of $L\Delta$, for the unusual properties of $v_y(t)$ are only in evidence if the memory of the filter exceeds $L\Delta$. This is seen in our example (the memory of the filter being given by its weighting function e^{-at}), for if $e^{-aL\Delta} \doteq 0$ then $\beta(N = 1) \doteq 1$ and the second $L\Delta$ cycle of $v_y(t)$ will be equal to the constant steady state value $\beta^2 v_y(L\Delta) = v_y(L\Delta)$ of (B12). Furthermore, if $e^{-aL\Delta} \doteq 0$, then the transient part of $v_y(t)$ of (B4) will have died away by $t = L\Delta$, and equation (B4) can be considered to hold for all time. For a first order filter, the significant memory extends to about 4 time constants or $4/a$. Thus for curves 4, 5 and 6 of Figure 2, the filter memory is shorter than $L\Delta$, and $v_y(t)$ is essentially smooth for all time.

If the filter memory exceeds $L\Delta$, as in curves 1, 2 and 3 of Figure 2, $v_y(t)$ will have successive discontinuous derivatives at $t = L\Delta, 2L\Delta, \ldots NL\Delta$ until $\beta(N)$ reaches the constant β of (B8). Also, $v_y(t)$ loses its dependence on t^* at a rate depending on the convergence of $\beta(N)$, and so when $\beta(N) = \beta$, the mean square of $y(t)$ is equal to its steady state value of

$$v_y \text{ (steady state)} = \beta^2 v_y(L\Delta). \qquad \ldots\ldots \text{(B12)}$$

As $\beta(N)$ converges to a constant as $e^{-aNL\Delta}$ converges to zero, and $e^{-aNL\Delta}$ is a discrete form of e^{-at}, the filter's memory, we are left with the expected result that we must wait for a period equal to the significant memory of the filter for the filter's initial transients to die away before a stationary signal $y(t)$ is obtained. However, unless $y(0)$ is randomised according to its steady state distribution (i.e. $y(0)$ is set equal to $\beta y(L\Delta) = y(N_{large} L\Delta)$), this transient will exist, and we have used our analysis to show the form of the transient for a particular initial value, $y(0) = 0$. Further, by evaluating the function $E[y(t^*)\, y(t^* + \tau)]$ in a manner similar to Appendix C, we could derive the entire transient second order properties of $y(t)$, that is, its transient correlation function. This function would become stationary in the same way as $v_y(t)$ above, and would be a periodic function of τ, with period $L\Delta$.

We can illustrate that the unusual transient properties of the statistics of $y(t)$ are due to the periodic and urn model properties of the PRBS by comparing our continuous model of the filtered PRBS with the same model with these periodic and urn model properties removed, which is essentially the same filter with white noise input. This is done in section 5, and we see that the filter with white noise input does not exhibit these unusual transient effects.

4. Accuracy of the Diffusion Model

As the diffusion model $y(t)$ of the filtered PRBS is Gaussian the error in mean square is considered to be a reasonable and sufficient measure of

accuracy of the model in distribution.* As the model is primarily meant to represent the filtered PRBS at the sample points $t = n\Delta$, $n = 0, 1, 2 \ldots$, the main error check of the transient results was carried out at these points.

The accuracy of equations (B4), and (B10) for the mean square of $y(t)$ was checked by simulating the PRBS and filter output exactly on a digital computer at the sample points $t = n\Delta$, and forming $v_y(n\Delta)$ by ensemble averaging over all possible starting points of the PRBS. The $M = 7$, $L = 127$ PRBS was used and Δ was arbitrarily set to unity.

The percentage error between $v_y(t)$ and the true mean square at the sample points is shown as curve 1 in Figure 3, as a function of filter time constant T. The error over the transient curve was almost exactly uniform in time and so the time dependence of error is not given. For values of T beyond 2Δ, the absolute percentage error remains at less than 1% and the error finally settles down at -0.80% for all T beyond the scale of the graph including the integrator case, $v_s(t)$.

Two comments are relevant to this error curve. Firstly, the modelling accuracy is good only for values of filter time constant greater than 2Δ. This agrees with the conditions given by Cumming (1967 c) for good modelling accuracy — that the significant memory time of the system (equal to about 4 T in the present example) must be substantially greater than the significant

---

* At least this is so when the filtered PRBS itself is near Gaussian (see section 5 or Roberts and Davis, 1966), but when T is large and the filter output non-Gaussian, other error measures may have to be checked, depending on the use to which the model will be put. The expression (I2) for the mean value of the filtered PRBS is exact.

correlation time of the input noise (equal to Δ for the PRBS). Thus the inequality $T > 2\Delta$ satisfies this criterion by a reasonable margin.

Secondly, by changing L and Δ it was confirmed that the error for large T is very nearly constant at a factor of $-\frac{1}{L}$ ($-\frac{1}{L}$ equals - 0.79% in the example illustrated). But from Cumming (1967 b, equation (29)) we see the PRBS has a low frequency power of $\frac{1}{L}(1 + \frac{1}{L})$ concentrated at discrete frequencies separated by $\frac{1}{L\Delta}$ cycles/unit time, which averages out to the equivalent of a continuous power density spectrum of $(1 + \frac{1}{L})\Delta$. Thus it is reasonable that the white noise which replaces the PRBS in our diffusion model should have this power density spectrum instead of the value Δ used in equation (8) which we obtained via the second increment of equations (5) and (7). Increasing the power of the model noise from Δ to $(1 + \frac{1}{L})\Delta$ would largely cancel the error in mean square quoted above, but this would be difficult to justify from the method we have used to choose a diffusion model. In any case, if greater model accuracy is desired, all noise terms (i.e. $dw(t)$ in equations such as (8) and (11)) can be multiplied by the factor $(1 + \frac{1}{L})^{\frac{1}{2}}$. This would not effect the $m(t)$ equations, and would multiply the $v(t)$ equations by a factor of $(1 + \frac{1}{L})$ in the same way as the noise scaling factor A we have dropped.

The accuracy of the mean square of the diffusion model of the filtered PRBS was also checked against the <u>continuous</u> mean square of the true filtered PRBS, and is shown as curve 2 in Figure 3. The continuous mean square was estimated by sampling an exact simulation of the filtered PRBS every $\frac{1}{40}\Delta$ time units and averaging over all starting points of the PRBS. Again the error was uniform in time, and in the steady state ($t > 4T$) the error of the estimate in the simulation was checked using the known steady state mean square (see Cumming, 1967 b, equation (20)), and was equal to 0.02%, uniform in T.

As seen in Figure 3, the diffusion model does not represent the mean square of the continuous filter output as well as it does the sampled filter output, except for $T > 50\,\Delta$ (beyond range of graph) when curve 2 falls just below 1%. This is because the white noise model of the PRBS does not represent the fine structure of the PRBS, but as seen from the method of constructing the model, the diffusion model is meant to represent integral functions of the PRBS at the sample points $t = n\Delta$, $n = 1, 2, \ldots$ (remember that the integral of the PRBS is a ramp function between the sample points, while the integral of the modified white noise has the fine random structure of Brownian motion - the two are meant to be statistically equivalent only at the sample points). This point is a result of the inherently discrete nature of the PRBS and is not a general property of our method of choosing diffusion processes to model continuous non-Markovian processes.

5. Comparison of PRBS with White Noise

Through equation (10) or Figure 1, the PRBS can be explicitly compared with white noise $z(t)$ of power per cycle per unit time of Δ (or more accurately $\Delta\,(1 + \frac{1}{L})$). The following points are noted:

1. The PRBS has a small positive bias equal to $\frac{1}{L}$.

2. The PRBS is periodic with period equal to $L\,\Delta$ as represented by the periodic time parameter t^*.

3. The PRBS has the urn model property represented by the first term of equation (5) or the feedback loop in Figure 1.

The effect of one or both of the last two points is to reduce the effective power of the PRBS compared with white noise when used as a test

signal at the input of a system. This effect is dependent on the bandwidth of the system and has been noted earlier by Roberts (1966) who compares the steady state variance of filtered white noise with that of the filtered PRBS. Using the present analysis we can compare the transient variance of filtered white noise with that of the filtered PRBS.

This comparison is made in Appendix B, where equation (B14') gives the variance of white noise of power density Δ passed through a first order filter of time constant $T = a^{-1}$. Equation (B4') gives the analogous expression for the filtered PRBS for $t = [0, L\Delta]$, and for $t > L\Delta$ the expression for the variance can be deduced from equation (B10) for the mean square.

Comparing these equations, we see that the variances of the filtered PRBS and filtered white noise are essentially equal when $\frac{aL\Delta}{4}$ is substantially greater than one. For $L = 127$ and $T = a^{-1} = 5\Delta$, the difference in the transient variances is a maximum of 8%, but for larger values of filter time constant, this difference increases substantially. Furthermore, for $T > 25\Delta$, the unusual discontinuity effects of Figure 2 are noticed in the variance of the filtered PRBS which are not present in the filtered white noise.

In the steady state, this disparity of variances can be cancelled by applying a scaling factor to the PRBS. However, owing to the extra transient term in (B4') compared with (B14') and the discontinuity effects noticed when $T > 25\Delta$, the disparity in the transients of the variances cannot be simply cancelled. Thus when using the PRBS to simulate white noise in a situation where the transient statistics are of interest, we must approach the results with caution if the effective time constants of the system under test are greater than 5Δ (or $> 0.039\, L\, \Delta$ for an arbitrary value of L).

Another property of the filtered PRBS which distinguishes it from filtered white noise has been brought out by Roberts and Davis (1966). This property is that beyond a certain value of filter time constant T, further increases in T make the first order probability density function of the filter output less and less like the Gaussian distribution. The distribution of the PRBS itself is concentrated at + 1 and - 1, and as the PRBS is more heavily filtered, we should expect the distribution of the filtered PRBS to tend uniformly towards the Gaussian in much the same way as the smoothed random telegraph signal does (Wonham and Fuller, 1958). Roberts and Davis show experimentally that as the value of the time constant T is increased, the proximity of the distribution of the filtered PRBS to the Gaussian increases up to a certain value of $T = T'$, and then decreases for higher values of T.

An explanation of this property is suggested by a theoretical result of Tausworthe (1965) who shows that sets of m adjacent bits of the PRBS can be considered as a set of statistically independent variables only for m less than or equal to M, the number of stages in the generating shift register. Near-Gaussian distributions are formed by summing independent variables, and when the effective memory time of the filter extends beyond $M\Delta$, then the filter (which is forming a weighted sum of the input) is summing variables which are no longer independent, and the resultant output distribution can no longer be expected to be near Gaussian. Thus the distribution of the filtered PRBS should be nearest the Gaussian at $T' \doteq 0.4\, M\Delta$. This is in close agreement with the experimental findings of Roberts and Davis. This effect was first pointed out by White (1966) in connection with summing adjacent bits of the PRBS to form a binomial distribution.

The diffusion model we propose for the PRBS (and the linear filter output) is Gaussian and so it does not possess the non-Gaussian property discussed above. Indeed, the model accuracy is at a maximum for large values of T when the true signal tends away from the Gaussian.

6. Conclusions

As a random noise source, a maximal length pseudo random binary sequence has two unusual properties. These are the non random properties of a) being periodic, and b) having a fixed (and almost equal) number of positive and negative bits in each period. These properties give the PRBS some advantages for use as an input signal in testing dynamic systems, but at the same time we must appreciate how these properties can introduce unexpected effects into the system's statistics. In this paper we present a continuous diffusion model which is useful for determining the transient statistics of systems which are driven by a PRBS.

The non-random properties of the PRBS are incorporated into a discrete urn mechanism, and then by a new method, a continuous Markov (diffusion) process is chosen which models the urn mechanism, which is non-Markovian when considered as a continuous process. As diffusion processes are more convenient to analyse, particularly for transient properties, than continuous non-Markovian processes, the construction of diffusion models for non-Markovian processes is a convenient tool of analysis of physical random phenomena.

The analysis was applied to a PRBS put through a linear first order filter. A PRBS of 127 bit length was used as an example, and the model

represented the transient mean square statistics of the filtered PRBS at the switching points of the PRBS to better than 1% accuracy, provided the filter time constant was not too short compared with the PRBS bit interval. For short values of the time constant, the upper frequency of the input noise (the PRBS) is not sufficiently higher than the pass band of the system (the filter) for modelling accuracy to be maintained. It was also noted that the diffusion model represented the mean square of the continuous filter output less accurately than the discrete output, but perhaps still sufficiently accurately for some purposes.

The white noise model of the PRBS provides a comparison between the PRBS and white noise. We see that the transient statistics of the filtered PRBS can be quite different from that of filtered white noise, a factor which could be important in repetitive-type simulations.

Acknowledgements

The author is indebted to J.M.C. Clark for many discussions and to Professor J.H. Westcott for his encouragement. The financial support of a NATO scholarship administered by the National Research Council of Canada is gratefully acknowledged.

References

Cumming, I.G., 1967 a, "Derivation of the Moments of a Continuous Stochastic System", to appear, Int. J. Control.

1967 b, "The Autocorrelation Function and Spectrum of a Filtered Pseudo Random Binary Sequence", to appear, Proc.IEE.

1967 c, "Computing Aspects of Problems in Non-linear Prediction and Filtering", Ph.D. thesis, Centre for Computing and Automation, Imperial College, University of London.

Doob, J.L., 1953, "Stochastic Processes", Chapters VI and IX. John Wiley & Sons, New York.

Kushner, H.J., 1964, "On the Differential Equations Satisfied by Conditional Probability Densities of Markov Processes, with Applications", J. S.I.A.M. (Control), 2, 1, pp. 106 - 119.

Roberts, P.D. & Davis, R.H., 1966, "Statistical Properties of Smoothed Maximal-length Linear Binary Sequences", Proc. IEE, 113, 1, pp 190 - 196, January 1966.

Roberts, P.D., 1966, "Use of Pseudorandom Binary Sequences in the Digital Simulation of Control Systems Subjected to Random Input Signals", Electronics Letters, 2, 3, pp 105 - 106, March 1966.

Tausworthe, R.C., 1965, "Random Numbers Generated by Linear Recurrence Modulo Two", Math. of Computation, 19, pp 201 - 209, April 1965.

White, R.C., 1966, "Experiments with Digital Computer Simulations of Pseudo-Random Noise Generators", ACL Memo No. 128, Dept. of Electrical Eng., University of Arizona, October, 1966.

Wonham, W.M., and Fuller, A.T., 1958, "Probability Densities of the Smoothed Random Telegraph Signal", J. Elect. & Control, 4, 6, pp 567-76, June 1958.

Wonham, W.M., 1965, "Some Applications of Stochastic Differential Equations to Optimal Non-linear Filtering", Appendix 1., J. S.I.A.M. (Control), 2, 3, pp 347-369.

Appendix A

Stochastic Equations

A diffusion process $x(t) = x(t, w)$ which depends on a unit parameter Wiener process $w(t)$ can be defined by the Ito stochastic differential equation (s.d.e.) (see, for example, Wonham, 1965, Appendix 1)

$$dx(t) = f(x,t)\, dt + F(x,t)\, dw(t) \quad , \qquad x(o) = o. \tag{A1}$$

This equation does not have the properties of an ordinary differential equation (e.g. it does not specify the value of the derivative of $x(t)$), and some authors prefer to regard it as a symbolic way of writing the better known stochastic integral equation (Doob, 1953)

$$x(t) = \int_o^t f(x,u)\, du + \int_o^t F(x,u)\, dw(u). \tag{A2}$$

Interpreting $\delta.$ as a finite forward difference increment over the time increment δt, the diffusion process $x(t)$ has the following local properties

$$\delta x(t) = f(x,t)\, \delta t + F(x,t)\, \delta w(t) + o(\delta t), \tag{A3}$$

and

$$(\delta x(t))^2 = F^2(x,t)\, (\delta w(t))^2 + o(\delta t). \tag{A4}$$

As δx and $(\delta x)^2$ are random variables, equations (A4) and (A5) are formal, and in particular, the error terms in these equations only have a precise meaning when conditional expectations are taken. Then we have,

$$E\,[\delta x \mid x, t] = f(x,t)\, \delta t + o(\delta t), \tag{A5}$$

and

$$E\,[(\delta x)^2 \mid x, t] = F^2(x, t)\, \delta t + o(\delta t), \tag{A6}$$

as $(\delta w)^2$ is a random variable with mean δt and standard deviation $2^{\frac{1}{2}}\,\delta t$. Even though the standard deviation of $(\delta w)^2$ is the same order as the mean, we can neglect the random part of $(\delta w)^2$ if we are only interested in integral functions of δx (that is, if we are only interested in $x(t)$), as the contribution of the random part of $(\delta w)^2$ to the integral becomes negligible by the central limit theorem.† It is the fact that the first term on the right hand side of (A6) is $O(\delta t)$ that distinguishes the stochastic calculus from the ordinary calculus, for the ordinary calculus applies to smooth functions for which $E\,[(\delta x)^2 \,|\, x,\, t\,] = O\,(\delta t^2)$.

The incremental properties (A5) and (A6) of the diffusion process (A1) or (A2) are more commonly given as the first and second incremental moments

$$\underset{\delta t \downarrow o}{\text{Limit}} \quad \frac{1}{\delta t} \quad E\,[\,\delta x \,|\, x,t\,] \;=\; f\,(x,t), \tag{A5'}$$

and

$$\underset{\delta t \downarrow o}{\text{Limit}} \quad \frac{1}{\delta t} \quad E\,[(\delta x)^2 \,|\, x,\, t] = F^2\,(x,\,t). \tag{A6'}$$

Equations for other properties can be written down directly from the stochastic integral equation (Doob, 1953); for example

$$E\,[x(t)] \;=\; \int_o^t E\,[f(x,u)\,]\,du\,, \tag{A7}$$

and

$$E\,[x^2\,(t)\,] \;=\; \Big[\int_o^t E\,[f(x,\,u)]\;\;du\,\Big]^2 \;+\; \int_o^t E\,[F^2(x,u)\,]\,du. \tag{A8}$$

Equations for the Moments of $x(t)$

Consider a function of x, $G(x)$. By Taylor's series, a forward increment in $G(x)$ is written as

---

† For a more detailed explanation justifying this point, see Kushner (1964, Appendix 1.)

$$\delta G = G_x \delta x + \tfrac{1}{2} G_{xx} (\delta x)^2 + o((\delta x)^2) , \qquad (A9)$$

where the subscripts denote partial derivatives.

Taking the conditional expectation of both sides of (A9) and using the properties (A5) and (A6) of the conditional x increments, we have

$$E[\delta G \mid x, t] = G_x f(x,t) \delta t + \tfrac{1}{2} G_{xx} F^2(x,t) \delta t + o(\delta t), \qquad (A10)$$

as $(\delta x)^2$ and δt are the same order for stochastic increments. Taking the expected value of both sides of (A10), interchanging the $E[\cdot]$ and $\delta \cdot$ operators on the left hand side, dividing both sides by δt and passing to the limit $\delta t \downarrow 0$, we have

$$\frac{d E[G]}{dt} = E[G_x f(x,t)] + \tfrac{1}{2} E[G_{xx} F^2(x,t)] , \qquad (A11)$$

as $E[E[\delta G \mid x,t]] = E[\delta G] = \delta E[G]$. Equation (A11) is an ordinary differential equation which can be used to find the moments of $x(t)$. Setting $G(x)$ successively equal to x and x^2, we have

$$\frac{d E[x]}{dt} = E[f(x,t)] , \qquad (A12)$$

and

$$\frac{d E[x^2]}{dt} = 2 E[x f(x,t)] + E[F^2(x,t)] . \qquad (A13)$$

These differential equations are usually easier to handle than the corresponding integral equations (A7) and (A8). However, they are only easy to evaluate if $f(x,t)$ and $F(x,t)$ are constants or linear functions of x, in which case the right hand sides of (A12) and (A13) involve only the unknowns $E[x]$ and $E[x^2]$. Otherwise these equations will be imbedded in an infinite set of simultaneous differential equations for all the moments of x.

A more detailed derivation of these equations, including the differential equations for the moments of a vector diffusion process is given by Cumming (1967a).

APPENDIX B

Derivation of Mean and Mean Square of Filtered Signal $y(t)$

Consider the joint diffusion process $[s(t), y(t)]$ in the interval $[0, L\Delta]$

$$ds(t) = \frac{-s(t)}{L\Delta - t}\,dt + \Delta^{\frac{1}{2}}\,dw(t), \qquad s(0) = 0; \qquad \ldots (B1)$$

$$dy(t) = \left(-a\,y(t) - \frac{s(t)}{L\Delta - t} + \frac{1}{L}\right)dt + \Delta^{\frac{1}{2}}dw(t), \quad a \neq 0, \; y(0) = 0.$$

From equations (13) and (14) we know $m_s(t) = 0$

$$v_s(t) = \Delta\left[1 - \frac{t}{L\Delta}\right]t.$$

Let $m_y(t) = E[y(t) \mid y(0) = 0]$.

Then by the methods of Appendix A, we have

$$\frac{d\,m_y}{dt} = -a\,m_y + \frac{1}{L}, \qquad m_y(0) = 0$$

whence

$$m_y(t) = \frac{1}{aL}\left(1 - e^{-at}\right). \qquad \ldots\ldots (B2)$$

Let $v_{sy}(t) = E[s(t)\ y(t) \mid s(0) = 0;\ y(0) = 0]$.

Then by a formula analogous to (A9) for the differential of a product

$$\delta(sy) = s\delta y + y\,\delta s + \delta s\,\delta y + o(\delta s^2) + o(\delta y^2)$$

we obtain the differential equation for $v_{sy}(t)$ (see Cumming, 1967 a, for more details)

$$\frac{dv_{sy}}{dt} = \left(-a - \frac{1}{L\Delta - t}\right) v_{sy} + \Delta - \frac{1}{L}t\ ,\quad v_{sy}(o) = 0.$$

Solving we have $v_{sy}(t) = \frac{L\Delta - t}{aL}(1 - e^{-at})$. ... (B3)

It is easy to deduce that $v_{sy}(t)$ is periodic in $L\Delta$ with $v_{sy}(nI\Delta) = 0$, $n = 0, 1, 2 \ldots$ i.e. we can replace t by t^*, and also that (B2) holds for $t > I\Delta$ without replacing t by t^*.

Let $v_y(t) = E[y^2(t) \mid y(0) = 0,\ s(0) = 0]$.

Then

$$\frac{dv_y}{dt} = -2a\,v_y - \frac{2v_{sy}}{L\Delta - t} + \frac{2}{L}m_y + \Delta\ ,$$

or

$$\frac{dv_y}{dt} + 2a\,v_y = -\frac{2}{aL}\left(1 - \frac{1}{L}\right)(1 - e^{-at}) + \Delta\ ,\ v_y(o) = 0.$$

Solving we have

$$v_y(t) = \frac{2}{a^2L}\left[\left(\frac{1}{L} - 1\right)(1 - e^{-at}) + \left(\frac{1}{2} - \frac{1}{2L} + \frac{aL\Delta}{4}\right)(1 - e^{-2at})\right],$$

$$o \leqslant t \leqslant L\Delta\ . \qquad \ldots\ldots \quad (B4)$$

By subtracting the square of the mean (B2), we obtain a simpler expression for the variance

$$\mathrm{Var}_y(t) = \frac{2}{a^2 L}\left[-(1 - e^{-at}) + \left(\frac{1}{2} + \frac{aL\Delta}{4}\right)(1 - e^{-2at}) \right]. \quad \ldots (B4')$$

Solution for $v_y(t)$ for $t > L\Delta$

To solve for $y(t)$ and the statistics of $y(t)$ for $t > L\Delta$, we will use the periodic property of $Z(t^*)$ or $d\omega(t^*)$.

Writing $y(t)$ of equation (6) as a stochastic integral

$$y(t) = \int_0^t e^{-a(t-u)} \left[-\frac{S(u^*)}{L\Delta - u^*} + \frac{1}{L} \right] du + \Delta^{\frac{1}{2}} \int_0^t e^{-a(t-u)}\, d\omega(u^*) , \quad \ldots (B5)$$

where $u^* = u$ (modulo $L\Delta$).

Now consider t in the semi-open interval $[NL\Delta, (N + 1)L\Delta)$. Then

$$t = NL\Delta + t^*$$

$$u = nL\Delta + u^* \quad , \quad n = 0, 1, 2 \ldots N.$$

Breaking the integrals of (B5) into $L\Delta$ time segments, we have

$$y(t) = \sum_{n=0}^{N-1} \int_{o}^{L\Delta} e^{-a(NL\Delta - nL\Delta + t^* - u^*)} h(u^*)\, du^* + \int_{o}^{t^*} e^{-a(t^* - u^*)} h(u^*)\, du^* + \Delta^{\frac{1}{2}} \sum_{n=0}^{N-1} \int_{o}^{L\Delta} e^{-a(NL\Delta - nL\Delta + t^* - u^*)}\, d\omega(u^*) + \Delta^{\frac{1}{2}} \int_{o}^{t^*} e^{-a(t^* - u^*)}\, d\omega(u^*) , \qquad \ldots (B6)$$

where $h(u^*) = \frac{-S(u^*)}{L\Delta - u^*} + \frac{1}{L}$.

Noting that the second and fourth terms of (B6) sum to $y(t^*)$, and on bringing $e^{-at^*}e^{-a(N-n-1)L\Delta}$ out of the remaining integrals, these latter integrals become the integrals for $y(L\Delta)$. Thus

$$y(t) = y(t^*) + \sum_{n=0}^{N-1} e^{-at^*}e^{-a(N-n-1)L\Delta}\, y(L\Delta) ,$$

or

$$y(t) = y(t^*) + \beta(N)\, e^{-at^*}\, y(L\Delta) , \qquad \ldots (B7)$$

where

$$\beta(N) = e^{-a(N-1)L\Delta} \sum_{n=0}^{N-1} e^{anL\Delta} = \sum_{n=0}^{N-1} e^{-anL\Delta} = \frac{1 - e^{-anL\Delta}}{1 - e^{-aL\Delta}} \qquad \ldots (B7')$$

In the steady state, $\beta(N)$ approaches the constant $\beta = \frac{1}{1 - e^{-aL\Delta}}$ (B8)

Taking the expected value of (B7) we have

$$m_y(t) = m_y(t^*) + \beta(N)\, e^{-at^*}\, m_y(L\Delta)\,, \qquad \ldots \text{(B9)}$$

which, upon substitution of $\beta(N)$ of (B7'), does not depend on t^*, verifying that (B2) holds for $t > L\Delta$.

Squaring (B7) and taking the expected value we have

$$v_y(t) = v_y(t^*) + \beta^2(N)\, e^{-2at^*}\, v_y(L\Delta) + 2\beta(N)\, e^{-at^*}\, E\,[\, y(t^*)\, y(L\Delta)\,],$$

where $E\,[\, y(t^*)\, y(L\Delta)\,]$ has been evaluated in Appendix C, equation (C2). Then

$$v_y(t) = v_y(t^*)\,[1 + 2\beta(N)\, e^{-aL\Delta}] + v_y(L\Delta)\, \beta^2(N)\, e^{-2at^*} + \frac{2}{a^2 L}\, \beta(N)\left(\frac{1}{L} - 1\right)(1 - e^{-at^*})(e^{-at^*} - e^{-aL\Delta}), \qquad \ldots \text{(B10)}$$

where $v_y(t^*)$ and $v_y(L\Delta)$ are taken from (B4).

A simpler expression is obtained for $v_y(t)$ at the cycle points $t = NL\Delta$, where (with $t^* = 0$)

$$v_y(NL\Delta) = \beta^2(N)\, v_y(L\Delta), \qquad \ldots \text{(B11)}$$

which approaches a constant as $\beta(N)$ approaches β of equation (B8).

Steady State Solution of $v_y(t)$, with $\beta(N) = \beta$

From (B10), (B4) and (B8) we have

$$v_y(t_{large}) = \frac{2}{a^2L}\left[\left(\frac{1}{L} - 1\right)\left(1 - e^{-at^*}\right) + \left(\frac{1}{2} - \frac{1}{2L} + \frac{aL\Delta}{4}\right)\left(1 - e^{-2at^*}\right)\right] \cdot$$

$$\cdot\left[1 + \frac{2e^{-aL\Delta}}{1 - e^{-aL\Delta}}\right] + \frac{2}{a^2L}\left[\left(\frac{1}{L} - 1\right)\left(1 - e^{-aL\Delta}\right) + \left(\frac{1}{2} - \frac{1}{2L} + \frac{aL\Delta}{4}\right)\left(1 - e^{-2aL\Delta}\right)\right] \cdot$$

$$\cdot\left[\frac{e^{-2at^*}}{\left(1 - e^{-aL\Delta}\right)^2}\right] + \frac{2}{a^2L}\left[\left(\frac{1}{L} - 1\right)\left(1 - e^{-at^*}\right)\left(e^{-at^*} - e^{-aL\Delta}\right)\left(\frac{1}{1 - e^{-aL\Delta}}\right)\right],$$

$$= \frac{2}{a^2L}\beta^2 V(t^*),$$

where $V(t^*) = \left(\frac{1}{L} - 1\right)\left(1 - e^{-at^*}\right)\left(1 - e^{-2aL\Delta}\right)$

$$+ \left(\frac{1}{2} - \frac{1}{2L} + \frac{aL\Delta}{4}\right)\left(1 - e^{-2at^*}\right)\left(1 - e^{-2aL\Delta}\right)$$

$$+ \left(\frac{1}{L} - 1\right)\left(1 - e^{-aL\Delta}\right)e^{-2at^*} + \left(\frac{1}{2} - \frac{1}{2L} + \frac{aL\Delta}{4}\right)\left(1 - e^{-2aL\Delta}\right)e^{-2at^*}$$

$$+ \left(\frac{1}{L} - 1\right)\left(1 - e^{-aL\Delta}\right)\left(e^{-at^*} - e^{-aL\Delta} - e^{-2at^*} + e^{-a(t^*+L\Delta)}\right),$$

$$= \left(\frac{1}{L} - 1\right)\left(1 - e^{-at^*}\right)\left(1 - e^{-aL\Delta}\right)\left(1 + e^{-aL\Delta}\right)$$

$$+ \left(\frac{1}{2} - \frac{1}{2L} + \frac{aL\Delta}{4}\right)(1 - e^{-2aL\Delta})$$

$$+ \left(\frac{1}{L} - 1\right)(1 - e^{-aL\Delta})e^{-at^*}(1 + e^{-aL\Delta})$$

$$- \left(\frac{1}{L} - 1\right)(1 - e^{-aL\Delta})e^{-aL\Delta} \quad ,$$

$$= \left(\frac{1}{L} - 1\right)(1 - e^{-aL\Delta}) + \left(\frac{1}{2} - \frac{1}{2L} + \frac{aL\Delta}{4}\right)(1 - e^{-2aL\Delta}) \quad ,$$

$$= \frac{a^2L}{2} v_y(L\Delta).$$

Thus $$v_y(t_{large}) = \beta^2 v_y(L\Delta) \qquad \ldots (B12)$$

Comparison of $y(t)$ with Filtered White Noise

Consider the output $\underline{y}(t)$ of a filter of weighting function e^{-at} and input $Z(t) + \frac{1}{L}$. This situation is the same as the PRBS model with the periodicity and urn model properties removed. Then analogous to (B5) we have

$$\underline{y}(t) = \frac{1}{L} \int_0^t e^{-a(t-u)}\, du + \Delta^{\frac{1}{2}} \int_0^t e^{-a(t-u)}\, d\omega(u).$$

Then from (A10) and (A11) we have

$$m_{\underline{y}}(t) = \frac{1}{aL}(1 - e^{-at}) \quad , \qquad \ldots \text{(B13)}$$

which is identical to $m_y(t)$, and

$$v_{\underline{y}}(t) = \left[\frac{1}{aL}(1 - e^{-at})\right]^2 + \Delta \int_0^t e^{-2a(t-u)}\, du \quad ,$$

$$= \frac{2}{a^2L}\left[\frac{1}{L}(1 - e^{-at}) + \left(\frac{aL\Delta}{4} - \frac{1}{2L}\right)(1 - e^{-2at})\right]. \qquad \ldots \text{(B14)}$$

This expression can be compared with (B4) or, more conveniently, the expression for the variance of $\underline{y}(t)$

$$\mathrm{Var}_{\underline{y}}(t) = \frac{2}{a^2L}\left[\frac{aL\Delta}{4}(1 - e^{-2at})\right] \qquad \ldots \text{(B14')}$$

can be compared with (B4').

APPENDIX C

Evaluation of $E\,[y(t^*)\, y(L\Delta)\,]$

We write $y(\cdot)$ in stochastic integral form (B5), and break the range of integration for $y(L\Delta)$ into two parts about t^*. Then we have

$$y(L\Delta) = \int_0^{t^*} e^{-a(L\Delta-t^*)} e^{-a(t^*-u)} h(u)\,du + \int_{t^*}^{L\Delta} e^{-a(L\Delta-u)} h(u)\,du$$

$$+ \Delta^{\frac{1}{2}} \int_0^{t^*} e^{-a(L\Delta-t^*)} e^{-a(t^*-u)}\, d\omega(u) + \Delta^{\frac{1}{2}} \int_{t^*}^{L\Delta} e^{-a(L\Delta-u)}\, d\omega(u)\,,$$

$$= e^{-a(L\Delta-t^*)} y(t^*) + \int_{t^*}^{L\Delta} e^{-a(L\Delta-u)} h(u)\,du + \Delta^{\frac{1}{2}} \int_{t^*}^{L\Delta} e^{-a(L\Delta-u)}\, d\omega(u). \qquad \ldots (C1)$$

Then $E\,[y(t^*)\, y(L\Delta)\,]$

$$= e^{-a(L\Delta-t^*)} v_y(t^*) + E\left[y(t^*) \int_{t^*}^{L\Delta} e^{-a(L\Delta-u)} h(u)\,du\right]$$

as the last integral of (C1) has zero mean (see (A10)) and no correlation with $y(t^*)$ as the lower limit of integration is t^*. Then

$$E[y(t^*)y(L\Delta)] = e^{-a(L\Delta-t^*)} v_y(t^*) + \int_{t^*}^{L\Delta} e^{-a(L\Delta-u)} \frac{m_y(t^*)}{L}\,du$$

$$- \int_{t^*}^{L\Delta} e^{-a(L\Delta-u)} \frac{E[y(t^*)S(u)]}{L\Delta - u}\,du\,,$$

where $E[y(t^*)S(u)]$ has been evaluated later, equation (C3).

Then

$$E[y(t^*)y(L\Delta)] = e^{-a(L\Delta-t^*)}v_y(t^*) + \int_{t^*}^{L\Delta} e^{-a(L\Delta-u)}\left[\frac{m_y(t^*)}{L} - \frac{1}{aL}(1 - e^{-at^*})\right]du$$

$$= e^{a(t^*-L\Delta)}v_y(t^*) + \frac{1}{a^2L}\left(\frac{1}{L} - 1\right)(1 - e^{-at^*})(1 - e^{a(t^*-L\Delta)}). \qquad \ldots (C2)$$

Evaluation of $E[y(t^*)\, S(u)]$, $\quad t^* \leqslant u \leqslant L\Delta$

Breaking the range of integration for $S(u)$ in two parts at t^*, we have for $y(t^*)$ and $S(u)$

$$y(t^*) = \int_0^{t^*} e^{-a(t^*-v)}h(v)dv + \Delta^{\frac{1}{2}} \int_0^{t^*} e^{-a(t^*-r)}d\omega(r),$$

$$= \underline{1} + \underline{2},$$

$$S(u) = \Delta^{\frac{1}{2}}(u - L\Delta) \int_0^{t^*} \frac{1}{r - L\Delta}\, d\omega(r) + \Delta^{\frac{1}{2}}(u - L\Delta) \int_{t^*}^{u} \frac{1}{r - L\Delta}\, d\omega(r),$$

$$= \underline{3} + \underline{4}.$$

Then $E[y(t^*)\, S(u)] = E[\underline{1}\cdot\underline{3} + \underline{1}\cdot\underline{4} + \underline{2}\cdot\underline{3} + \underline{2}\cdot\underline{4}]$.

The terms $E[\underline{1}\cdot\underline{4} + \underline{2}\cdot\underline{4}]$ drop out as $\underline{4}$ has zero mean value (A10) and no correlation with random variables depending on $\omega(r)$ for $r \leqslant t^*$.

The term $E[\underline{2}.\ \underline{3}]$ is the expected value of the product of two stochastic integrals, and is given by a generalization of (A11), or refer to Doob[3], p. 429, eqn. (2.2) where $dF(r) = dr$, and we have

$$E[\underline{2}\ .\ \underline{3}] = \Delta\, e^{-at^*}(u - L\Delta) \int_o^{t^*} \frac{e^{ar}}{r - L\Delta}\, dr\ ,$$

$$= \Delta\, e^{-at^*}(u - L\Delta)e^{aL\Delta} \int_o^{t^*} \frac{e^{a(r - L\Delta)}}{r - L\Delta}\, d(r - L\Delta)\ .$$

Now $E[\underline{1}\ .\ \underline{3}] = E\Big[\int_o^{t^*} e^{-a(t^*-v)}\Big[\Delta^{\frac{1}{2}} \int_o^{v} \frac{1}{r - L\Delta}\, d\omega(r) + \frac{1}{L}\Big]dv\ .$

$$.\ \Delta^{\frac{1}{2}}\ (u - L\Delta) \int_o^{t^*} \frac{1}{r - L\Delta}\, d\omega(r)\ \Big].$$

The part of the first integral involving $\frac{1}{L}$ is a determinate integral and the expected value of its product with the latter zero mean stochastic integral is zero. We can change the order of integration of the remainder of the first integral by setting

$$\int_o^{t^*} \int_o^{v} d\omega(r)\, dv = \int_o^{t^*} \int_r^{t^*} dv\, d\omega(r)$$

and we have

$$E[\underline{1}\ .\ \underline{3}] = E\Big[\Delta e^{-at^*}(u - L\Delta) \int_o^{t^*} \frac{1}{r - L\Delta} \int_r^{t^*} e^{av} dv\, d\omega(r)\ .\ \int_o^{t^*} \frac{1}{r - L\Delta}\, d\omega(r)\Big]$$

$$= E\Big[\Delta e^{-at^*}(u - L\Delta) \int_o^{t^*} \frac{1}{r - L\Delta} \Big[\frac{1}{a}\, e^{at^*} - \frac{1}{a}\, e^{ar}\Big] d\omega(r)\ .\ \int_o^{t^*} \frac{1}{r - L\Delta}\, d\omega(r)\Big]$$

Using the formula for the expected value of the product of two stochastic integrals again, we have

$$E[\underline{1} \cdot \underline{3}] = \Delta e^{-at^*}(u - L\Delta) \int_o^{t^*} \frac{1}{(r - L\Delta)^2} \left[\frac{1}{a} e^{at^*} - \frac{1}{a} e^{ar}\right] dr \ ,$$

$$= \Delta e^{-at^*}(u - L\Delta) \left[\frac{1}{a} e^{at^*}\left(-\frac{1}{L\Delta} - \frac{1}{t^* - L\Delta}\right) - \frac{1}{a} \int_o^{t^*} \frac{e^{ar}}{(r - L\Delta)^2} dr\right].$$

Now

$$\frac{1}{a} \int_o^{t^*} \frac{e^{ar}}{(r - L\Delta)^2} dr = \frac{1}{a} e^{aL\Delta} \int_o^{t^*} \frac{e^{a(r-L\Delta)}}{(r - L\Delta)^2} d(r - L\Delta)$$

$$= \frac{1}{a} e^{aL\Delta} \left[-\frac{e^{a(r-L\Delta)}}{r - L\Delta}\right]_o^{t^*} + e^{aL\Delta} \int_o^{t^*} \frac{e^{a(r-L\Delta)}}{r - L\Delta} d(r - L\Delta) .$$

The integral of the last term cancels out the contribution of $E[\underline{2} \cdot \underline{3}]$, and we obtain

$$E\left[y(t^*)\, S(u)\right] =$$

$$\Delta e^{-at^*}(u - L\Delta) \left[\frac{1}{a} e^{at^*}\left(-\frac{1}{L\Delta} - \frac{1}{t^* - L\Delta}\right) - \frac{1}{a} e^{aL\Delta}\left[-\frac{e^{a(t^*-L\Delta)}}{t^* - L\Delta} + \frac{e^{-aL\Delta}}{-L\Delta}\right]\right]$$

$$= \frac{\Delta(u - L\Delta)}{a} \left[-\frac{1}{L\Delta} - \frac{1}{t^* - L\Delta} + \frac{1}{t^* - L\Delta} + \frac{e^{-at^*}}{L\Delta}\right] ,$$

$$= \frac{(L\Delta - u)}{aL} (1 - e^{-at^*}). \qquad \ldots (C3)$$

REFERENCES

[1] Merklinger, K. J.: "Control Systems Disturbed by Gaussian Processes". Ph.D. Thesis, Cambridge 1963 + IFAC 1963 Paper

[2] Buhr, R.: "Computational Analysis of Non-Linear Stochastic Control Systems", Ph.D. Thesis, Cambridge 1966

[3] Chuang, K. and Kazda, L. F.: "A Study of Non-Linear Systems with Random Inputs". AIEE Trans. Appl. & Industry, 42, May 1959

[4] Sawaragi, Y, Sunahara, Y. & Soeda, T.: "The Discrepancy from the Normal Distribution of the Probability of the Response of Non-Linear Control Systems Subjected to a Gaussian Random Input". Tech. Rep., Engng. Res. Inst., Kyoto Univ. (Japan), 11, 2, pp. 19-38, (Rep. 79), March 1961

[5] Viterbi, A. J.: "Phase-Locked Loop Dynamics in the Presence of Noise by Fokker-Planck Techniques". Proc. IEEE, Dec. 1963, pp. 1737-53

[6] Barrett, J. F.: "Application of Kolmogorov's Equations to Randomly Disturbed Automatic Control Systems". Proc. IFAC Congress, Moscow 1960, pp. 724-33

[7] Ariaratnam, S. T.: "Random Vibrations of Non-Linear Suspensions". J. Mech. Eng. Science, 2, 3, 1960, pp. 195-201

[8] Khazen, E. M.: "Evaluation of the One-dimensional Probability Densities and Moments of a Random Process in the Output of an Essentially Non-Linear System". Prob. Th. and Appl., 6, 1961, pp. 117-23

[9] Fuller, A. T.: "Notes on Fokker-Planck Equations". Internal Communication, Engineering Dept., Cambridge University, October 1962

[10] Butchart, R. L.: "An Explicit Solution to the Fokker-Planck Equation for an Ordinary Differential Equation". Int. J. Control, 1, 3, March 1965, pp. 201-208

[11] Wax, N. (Editor): "Selected Papers on Noise and Stochastic Processes", Dover, New York, 1954

[12] Bogdanoff, J. L. & Kozin, F.: "Moments of the Output of Linear Random Systems". J. Acous. Soc. Amer. 34, 8, August 1962, pp. 1063-6

[13] Caughey, T. K. & Dienes, J. K.: "The Behaviour of Linear Systems with Random Parametric Excitation". J. Maths. & Phys., 41, 4, December 1962, pp. 300-318

[14] Ariaratnam, S. T. & Graefe, P. W. U.: "Linear Systems with Stochastic Coefficients, I, II, III". Inter. J. Control I: 1, 3, pp. 239-50; II: 2, 2, pp. 161-169; III: 2, 3, pp. 205-210 (1965)

[15] Gray, A. H.: "Behaviour of Linear Systems with Random Parametric Excitation". J. Acous. Soc. Amer., 37, 2, Feb. 1965, pp. 235-39

[16] Chan, S. Y. & Chuang, K.: "A Study of Linear Time-Varying Systems Subject to Stochastic Disturbances". Automatica 4, pp. 31-48, 1966

[17] Caughey, T. K. & Dienes, J. K.: "Analysis of a Non-Linear First Order System with a White Noise Input". J. Appl. Phys. 32, 11, Nov. 1961, pp. 2476-79

[18] Astrom, K. J.: "Analysis of a First Order Non-Linear System with a White Noise Forcing Function". TN. 18.057, IBM Nordiska Laboratorier, Sweden, September 1961

[19] Khazen, E. M.: "Estimating the Density of the Probability Distribution for Random Processes in Systems with Non-Linearities of Piecewise-Linear Type". Prob. Th. & Appln. 6, 1961, pp. 214-19

[20] Doob, J. L.: "Stochastic Processes". John Wiley, New York, 1953

[21] Stratonovich, R. L.: "Topics in the Theory of Random Noise, Volume 1". Gordon and Breach, 1963

[22] Clark, J. M. C.: "The Representation of Non-Linear Stochastic Systems with Applications to Filtering". Ph.D. Thesis, University of London, 1966

[23] Khazen, E. M.: Discussion of Barrett's 1960 IFAC Paper [6]

[24] Wong, E., & Zakai, M.: "On the Relation Between Ordinary and Stochastic Differential Equations". Int. J. Engng. Sci., 3, pp. 213-229, 1965

[25] Wong, E. and Zakai, M.:"On the Convergence of Ordinary Integrals to Stochastic Integrals". Ann. Math. Stat., 36, pp. 1560-1564, October 1965

[26] Åström, K. J.: "On a First Order Stochastic Differential Equation". Int. J. Control, 1, 4, April 1965

[27] Young, D.: "The Numerical Solution of Elliptic and Parabolic Partial Differential Equations", which is Chapter 15 or "Modern Mathematics for the Engineer", Second Series, edited by E. F. Beckenbach, McGraw-Hill, N.Y., 1961.

[28] Richtmyer, R. D.: "Difference Methods for Initial Value Problems". Interscience, New York, 1957

[29] Cumming, I. G.: "Numerical Techniques for the Non-Linear Prediction Problem". Automatica, 3, 3-4, pp. 257-273, January 1966

[30] Crank, J. & Nicholson, P.: "A Practical Method for Numerical Integration of Solutions of Partial Differential Equations of Heat-conduction Type". Proc. Cambridge Philos. Soc., 43, p. 50, 1947

[31] Young, D.: "The Numerical Solution of Elliptic and Parabolic Partial Differential Equations", which is Chapter 11 of "Survey of Numerical Analysis", edited by J. Todd, McGraw-Hill, New York, 1962

[32] Booton, R. C.: "The Analysis on Non-Linear Control Systems with Random Inputs", Proc. Symp. Non-Linear Circuit Analysis, Polytechnical Inst. of Brooklyn, New York, 1953, pp. 161-173

[33] Lee, Y. W. : "Statistical Theory of Communication". John Wiley & Sons, New York, 1960

[34] Blackman, R. B. & Tukey, J. W.: "The Measurement of Power Spectra", Dover Publications, New York, 1959

[35] Private Discussion with Professor R. D. Richtmyer and Dr. H. W. Morton at the Culham Laboratories, Atomic Energy Establishment, 28th July, 1965

[36] Peaceman, D. W. & Rachford, H. H.: "The Numerical Solution of Parabolic and Elliptic Differential Equations". J. Soc. Indust. Appl. Math., 3, 1, pp. 28-41, March 1955

[37] Douglas, J.: "On the Numerical Integration of $U_{xx} + U_{yy} = U_t$ by Implicit Methods", J. Soc. Indust. Appl. Math., 3, 1, pp. 42-65, March 1955

[38] Douglas, J. & Rachford, H. H.: "On the Numerical Solution of Heat Conduction Problems in Two and Three Space Variables", Trans. Amer. Math. Soc, 82, pp. 421-439, 1956

[39] This was confirmed in private discussion with J. Douglas at the Conference on Functional Analysis, Ravello, July 1st, 1965

[40] Karplus, W. J.: "A Hybrid Computer Technique for Treating Non-Linear Partial Differential Equations". IEEE Trans., EC-13, 5, p. 597-605, October 1964

[41] Gray, A. H. & Caughey, T. K.: "A Controversy in Problems Involving Random Parametric Excitation". J. Maths. & Phys., 44, 3, September 1965, pp. 288-296

[42] Sokolnikoff, I. S.; "Advanced Calculus". McGraw-Hill, New York, 1939

[43] Ito, Kiyosi: "Stochastic Integrals", Proc. Imp. Acad. Tokyo, 20, pp. 519-524, 1944

[44] Bartlett, M. S.: "An Introduction to Stochastic Processes", Cambridge University Press, Cambridge, 1962

[45] Feller, W.: "An Introduction to Probability Theory and its Applications, Volume 2", John Wiley, New York, 1966

[46] Aström, K. J.: "On Stochastic Integrals", unpublished note

[47] Ito, Kiyosi: "Stochastic Differential Equations on a Differential Manifold", Nagoya Math. J., 1, pp. 35-47, 1950

[48] Ito, Kiyosi: "On Stochastic Differential Equations", Mem. Amer. Math. Soc., 4, 1951

[49] Wonham, W. M.: "Some Applications of Stochastic Differential Equations to Optimal Non-Linear Filtering". J. Soc. Indus. Appl. Maths., Ser. A. 2, 3, pp. 347-369, 1965.
Also helpful is his "Advances in Non-Linear Filtering", Lecture Notes, M.I.T., Summer Section, 1965 [96]

[50] Stratonovich, R. L.: "A New Form of Representation of Stochastic Integrals and Equations", Vestnik of Moscow University, Series 1, Math. & Mech., 1, January 1964. It has recently appeared in English: SIAM J. Control, 4, p. 362

[51] Cramer, H.: "Mathematical Methods of Statistics", Princeton University Press, Princeton, 1946

[52] Grad, H.; "Note on N-Dimensional Hermite Polynomials", Comm. Pure Appl. Math., 2, pp. 325-330, 1949

[53] Kuznetsov, P. I., Stratonovich, R. L., & Tikhonov, V. I.: "Quasi-Moment Functions in the Theory of Random Processes", Th. Prob. and its Applic., 5, 1, pp. 80-96, 1960

[54] Lanczos, C.: "Applied Analysis", Pitman, London, 1957

[55] Kopal, Z.: "Numerical Analysis", Chapman and Hall, London, 1955

[56] Barrett, J. F.: "Notes from a Seminar on Non-Linear Filtering," Imperial College. 3rd November, 1964

[57] Cumming, I. G.: "A White Noise Model of a Pseudo Random Binary Sequence", Imperial College Research Report ICST/EE/ACS - 3/66, August, 1966. Reprinted in Appendix D of this thesis.

[58] Kolosov, G. E. & Stratonovich, R. L.: "One Asymptotic Method for Solving Problems on Synthesis of Optimal Regulators", Automation and Remote Control, 25, 12, pp. 1483-1498, December 1964

[59] Davison, E. J. A.: "Automatic Control of High Order Systems", Ph.D. Thesis, Cambridge, 1964

[60] Ternan, J. G.: "The Estimation of Parameters and Co-ordinates of Markov Processes". Ph.D. Thesis, University of Cambridge, December 1965

[61] Metropolis, N. & Ulam, S.: "The Monte Carlo Method". J. Amer. Stat. Assoc., 44, p. 335, 1949

[62] Korn, G. A.: "Hybrid Computer Monte Carlo Techniques". Memo. No. 109, Analog-Hybrid Computer Laboratory, University of Arizona, 15th February, 1965. This has now appeared in his book "Random Process Simulation and Measurement". McGraw-Hill, 1966

[63] Chuang, K., Kazda, L. F. & Windenecht, T.: "A Stochastic Method of Solving Partial Differential Equations using an Electronic Analogue Computer". Project Michigan Report 2900-91-T, Willow Run Labs., University of Michigan, June 1960

[64] Little, W. D.: "Hybrid Computer Solutions of Partial Differential Equations by Monte Carlo Methods". Ph.D. Thesis, University of British Comulbia, October 1965

[65] Gel'fand, I. M. & Yaglom, A. M.: "Integration in Functional Spaces and its Applications in Quantum Physics". J. Math'l. Phys., 1, 1, pp. 48-69, January 1960

[66] Ito, K.: "Wiener Integral and Feynman Integral". Proc. 4th Berkeley Symposium, Volume 2, pp. 227-238, University of California Press, 1961

[67] Dynkin, E. B.: "Markov Processes", Springer-Verlag, 1965

[68] Feller, W.: "On Positivity Preserving Semi-Groups of Transformations on $C[r_1, r_2]$." Ann. Soc. Polon. Math, 25, pp. 85-94, 1952

[69] Edelbaum, T. N.: "Theory of Maxima and Minima", in "Optimization Techniques", edited by G. Leitmann, Academic Press, New York, 1962

[70] Bellman, R.: "Introduction to Matrix Analysis", McGraw-Hill, New York, 1960

[71] Schur, I.: "Bemerkungen zur Theorie der beschränkten Bilinearformen mit unendlichen vielen Veranderlichen," J. Reine Angew. Math., 140, pp. 1-28, 1911

[72] King, G. W.: "Monte Carlo Method for Solving Diffusion Problems", Indust. Eng. Chem., 43, pp. 2475-6, 1951

[73] Skorokhod, A. V.: "Studies in the Theory of Random Processes", Addison-Wesley, Reading, Mass.,1965

[74] Cumming, I. G.: "Derivation of the Moments of a Continuous Stochastic System", Research note 5-66, Automatic Control Systems, Elec. Eng. Dept., Imperial College, London, Oct. 1966 - to be published in Int. J. Control

[75] Kwakernaak, H.: "On-Line Iterative Optimisation of Stochastic Control Systems", Automatica, 2, 3, p. 195, January 1965

[76] Mayne, D. Q.: "A Gradient Method for Determining Optimal Control of Non-Linear Stochastic Systems", IFAC Symposium on Self-Adaptive Control Systems, Teddington, September, 1965

[77] Carslaw, H. S. & Jaeger, J. C.: "Conduction of Heat in Solids", 2nd Edition, Clarendon Press, Oxford, 1959

[78] Box, G. E. P. & Mueller, M. E.: "A Note on the Generation of Random Normal Deviates", Annals of Mathematics and Statistics, 29, pp. 610-611, 1958

[79] Mood, A. M. & Graybill, F. A.: "Introduction to the Theory of Statistics", 2nd Edition, McGraw-Hill, 1963

[80] Lee, R. C. K.: "Optimal Estimation, Identification and Control", M.I.T. Press, Cambridge, Mass., 1964

[81] Kavanagh, R. J.: "A Note on Optimum Linear Multivariable Filters", Proc. IEE, 108C, pp. 412-417, April 1961

[82] Kavanagh, R. J. & Nolan, R.C.: "Generation of Random Signals with a Specified Spectral Density Matrix", IEEE Trans., AC-9, 3, pp. 300-301, July, 1964

[83] Handler, H.: "High-speed Monte Carlo Technique for Hybrid-computer Solution of Partial Differential Equations", Ph.D. Thesis, Department of Electrical Engineering, University of Arizona, 1967

[84] Davies, W. D. T.: "Generation and Properties of Maximum-length Sequences", Control, 10, 96, p. 302; 97, p. 364; 98, p. 431, June-August 1966

[85] Clarke, D. W. & Godfrey, K. R.: "Three-level m-sequences and their Application in On-Line Hill-climbing", IEE Colloquium on Pseudo Random Sequences Applied to Control Systems, January 27th, 1967, IEE, Savoy Place, London, W.C.2.

[86] Fuller, A. T.: "Notes on the Random Telegraph Signal as an Approximation to Gaussian White Noise", J. Elect. & Control, 14, 6, pp. 669-73, June 1963

[87] Roberts, P. D. & Davies, R. H.: "Statistical Properties of Smoothed Maximal-length Linear Binary Sequences", Proc. IEE, 113, 1, pp. 190-6, January, 1966

[88] Cumming, I. G.: "The Autocorrelation Function and Spectrum of a Filtered Pseudo Random Binary Sequence", Research Note 2-67, Automatic Control Section, Centre for Computing and Automation, Imperial College, London, S.W.7. January, 1967

[89] Wonham, W. M. & Fuller, A. T.: "Probability Densities of the Smoothed Random Telegraph Signal", J. Elect. & Control, 4, 6, pp. 567-76, June 1958

[90] Tausworthe, R. C.: "Random Numbers Generated by Linear Recurrence Modulo Two", Math. of Computation, 19, pp. 201-9, April 1965

[91] White, R. C.: "Experiments with Digital Computer Simulations of Pseudo-Random Noise Generators", ACL Memo No. 128, Dept. of Electrical Engineering, University of Arizona, October 1966

[92] Florentin, J. J.: "Optimal Control of Continuous Time, Markov, Stochastic Systems", J. Elect. & Control, 10, 6, pp. 473-488, June 1961

[93] Dreyfus, S. E.: Some Types of Optimal Control of Stochastic Systems", J. SIAM Control, 2, 1, pp. 120-134, 1964

[94] Sancho, N. G. F.: "Optimal Control of Markov Stochastic Systems which have Random Variation of Gain of Plant", Int. J. Control, 3, 5, pp. 487-496, 1966

[95] Kushner, H. J.: "On the Differential Equations Satisfied by Conditional Probability Densities of Markov Processes, with Applications", J. SIAM Control, 2, 1, pp. 106-119, 1964

[96] Wonham, W. M.: "Advances in Non-Linear Filtering", Lecture at M.I.T. Summer Session, 1965, Boston, Mass.

[97] Bucy, R. S.: "Non-Linear Filtering Theory", IEEE Trans. on Auto. Control, AC-10, 2, p. 198, April 1965

[98] Sridhar, R. & Detchmendy, D. M.: "Sequential Estimation of States and Parameters in Noisy Non-Linear Dynamical Systems", Preprints, 1965 JACC, Rensselaer Polytechnic Inst., Troy, New York, pp. 56-63

[99] Jazwinski, A. H.: "Filtering for Non-Linear Dynamical Systems", IEEE Trans. on Auto. Control, AC-11, 4, pp. 765-766, October 1966

[100] Kalman, R. E.: "New Methods in Wiener Filtering Theory", Proc. First Symposium on Eng. Appl. of Random Function Theory & Probability, Wiley, 1963, pp. 270-388

[101] Bendat, J. S.: "Principles and Applications of Random Noise Theory", Wiley, New York, 1958

[102] Wong, E. & Zakai, M.: "On the Relation Between Ordinary and Stochastic Differential Equations and Applications to Stochastic Problems in Control Theory", Preprints, Third IFAC Congress, London, June 1966, Paper 3B

[103] Kulman, N. K.: "A Note on the Differential Equations of Conditional Probability Density Functions", J. Math. Anal. Appl., 14, pp. 301-308, 1966

[104] Doob, J. L.: "The Brownian Movement and Stochastic Equations", Ann. Maths, 43, 2, pp. 351-369, April, 1942; reprinted in [11]

[105] Kozin, F. & Bogdanoff, J. L.: "A Comment on 'The Behaviour of Linear Systems with Random Parametric Excitation'", J. Math. & Phys., 42, 4, pp. 336-337, December 1963

[106] Leibowitz, M. A.: "Statistical Behaviour of Linear Systems with Randomly Varying Parameters", J. Mathl. Phys., 4, 6, pp. 852-8, June 1963

[107] Franklin, J. N.: "Difference Methods for Stochastic Differential Equations", Maths. of Comp., 19, 92, pp. 552-561, October 1965

[108] Hamming, R. W.: "Numerical Methods for Scientists and Engineers", McGraw-Hill International Student Edition (Tokyo), 1962

[109] Maruyama, G.: "Continuous Markov Processes and Stochastic Equations", Rend. Circ. Mat. Palermo, Ser. 2, 4, pp. 48-90, 1955

[110] Davis, P. J.: "Orthonormalising Codes in Numerical Analysis", which is Chapter 10 of "Survey of Numerical Analysis", edited by J. Todd, McGraw-Hill, New York, 1962

Alphabetical List of Referenced Authors

my [74] reference

INT. J. CONTROL, 1967, VOL. 000, NO. 000, 000–000

CUMMING (I.G.) Ph.D 1967

Derivation of the Moments of a Continuous Stochastic System†

By I. G. CUMMING

Centre for Computing and Automation, Imperial College, London, S.W.7

[Received January 24, 1967]

ABSTRACT

The properties of the Ito stochastic differential equation give a simple derivation of differential equations for expected values of arbitrary functions of stochastic systems. In particular, differential equations for the moments of the system are derived. It is cautioned that care must be taken when applying these results to noisy systems occurring in practice.

§ 1. INTRODUCTION

CONSIDER a continuous-time, continuous-state stochastic system with state vector $x(t)$. Assuming the system state $x(t)$ constitutes a Markov process, then the independence of successive increments of the noise process generating $x(t)$ allows us to derive differential equations for the expected value of arbitrary functions of $x(t)$ and time t. The most common application is in the derivation of differential equations for the moments of $x(t)$.

A stochastic system which is a continuous Markov process (a diffusion process) is described by an Ito stochastic differential equation (s.d.e.). The properties of Ito equations are discussed in detail by Doob (1953, Chaps. 6 and 9), and the properties we need are given in eqns. (4) and (5) below. In this note we shall consider the system $x(t)$ defined by the Ito s.d.e.:

$$dx(t) = f(x,t)\,dt + F(x,t)\,dw(t), \tag{1}$$

or in component form:

$$dx_i(t) = f_i(x,t)\,dt + \sum_k^m F_{ik}(x,t)\,dw_k(t), \quad i = 1,n, \tag{1'}$$

where

$x(t)$ is the n-dimension state vector of the system,

$x(0)$, or its probability density function $P(x,0)$, is given,

$f(x,t)$, $F(x,t)$ are the known system dynamics, with components or elements f_i and F_{ij},

$d.$ is a stochastic increment in the Ito sense (Doob 1953, p. 273),

all sums, e.g. $\sum_i^n$, have a lower limit of 1, and the indicated upper limit,

† Communicated by Dr. A. T. Fuller

and

$w(t)$ is an m-dimension Wiener process with the following incremental properties:

$$E[dw_i(t)] = 0, \qquad i = 1, m,$$

$$E[dw_i(t)\, dw_j(t)] = 2D_{ij}(t)\, dt, \quad i,j = 1, m,$$

or

$$E[dw(t)\, (dw(t))^{\mathrm{T}}] = 2D(t)\, dt,$$

where $(.)^{\mathrm{T}}$ denotes the transpose of the vector or matrix argument. The Ito s.d.e. (1) usually assumes $w(t)$ is a unit parameter Wiener process, in which case $2D(t) = I$, the identity matrix. The present choice of Wiener process allows arbitrary scaling and cross-correlation of the noise sources in a convenient form. The present form is equivalent to one in which the $2D(t)$ factor is absorbed into the F matrix, and $w(t)$ *is* a unit parameter Wiener process. The formal derivative of $w(t)$ is the common concept of white noise, and in the present notation, $2D(t)$ is the intensity matrix or the uniform power spectral density matrix of the white noise (in units of (noise)^2 per cycle per unit time (Fuller 1963)).

Some authors prefer the symmetrical form of stochastic equation introduced by Stratonovich (1966):

$$\bar{d}x_i(t) = \bar{f}_i(x,t)\, dt + \sum_{k}^{m} F_{ik}(x,t)\, \bar{d}w_k(t), \quad i = 1, n, \qquad (1'')$$

where $\bar{d}.$ is a stochastic increment in the Stratonovich sense. The relation between the Stratonovich s.d.e. (1″) and the Ito s.d.e. (1′) when they define the same diffusion process $x(t)$ is that

$$f_i(x,t) = \bar{f}_i(x,t) + \sum_{j}^{n} \sum_{k,l}^{m} \frac{\partial F_{ik}(x,t)}{\partial x_j} F_{jl}(x,t)\, D_{kl}(t), \quad i = 1, n.$$

The probability density $P(x,t)$ of the system (1) satisfies the Fokker–Planck differential equation†:

$$\frac{\partial P}{\partial t} = -\sum_{i}^{n} \frac{\partial}{\partial x_i} [f_i P] + \sum_{i,j}^{n} \frac{\partial^2}{\partial x_i\, \partial x_j} [(FDF^{\mathrm{T}})_{ij} P], \qquad (2)$$

with suitable initial conditions $P(x, 0)$. The notation $(.)_{ij}$ denotes the ijth component of the matrix argument.

Consider an arbitrary function $G(x,t)$ whose partial derivatives $G_{x_i x_j}$ and G_t are jointly continuous and bounded on any finite interval of x and t. The expected value of G over x space, $E[G]$, is given by the integral of GP over all x. $E[G]$ is then a function of time, and a differential equation for $E[G]$ can be obtained by multiplying both sides of (2) by G and integrating by parts to eliminate the x dependence. This technique seems to have been

† In the sequel we will drop the (x, t) parameter dependence, with the understanding that all functions of $f(x, t)$, $F(x, t)$, $D(t)$, $P(x, t)$ and $G(x, t)$ are to be evaluated at these points.

first mentioned by Bogdanoff and Kozin (1962) in connection with the moments of linear systems, and has been used by several authors since—for example, Ariaratnam and Graefe (1965) and Sancho (1965).

The purpose of this note is to present an alternative method of deriving a differential equation for $E[G]$. The method given below uses the incremental properties of the Ito s.d.e. (1) directly, and avoids the procedure of integrating the Fokker–Planck eqn. (2) by parts. Although the method we present below is not a difficult result of continuous stochastic process theory (for example, see the related result of Skorokhod (1965, p. 96)), we introduce it here as a simpler method than that in current use in the engineering literature.

§ 2. The General Result

Consider an arbitrary function of $x(t)$ and t, $G(x,t)$, whose partial derivatives have the finite properties mentioned above. In this section we derive a differential equation for the expected value of $G(x,t)$, $E[G]$. Interpreting $\delta.$ as a finite forward increment operator over the time increment δt, we expand $\delta G(x,t)$ by Taylor series and obtain:

$$\delta G = \sum_{i}^{n} G_{x_i}\delta x_i + \frac{1}{2}\sum_{i,j}^{n} G_{x_i x_j}\delta x_i\,\delta x_j + G_t\,\delta t + o(\delta x\,\delta x^{\mathrm{T}}) + o(\delta t), \tag{3}$$

where the subscripts x_i and t denote partial derivatives, and $o(.)$ denotes 'of order higher than $(.)$'†.

To take the expected value of (3) we will need the following properties of the Ito s.d.e. (1):

$$E[\delta x\,|\,x] = f\,\delta t + o(\delta t) \tag{4}$$

and

$$E[\delta x\,\delta x^{\mathrm{T}}\,|\,x] = 2FDF^{\mathrm{T}}\,\delta t + o(\delta t). \tag{5}$$

From the property (5) we see that $\delta x\,\delta x^{\mathrm{T}}$ and δt are of the same order for the diffusion process (1), and so we will absorb the error term $o(\delta x\,\delta x^{\mathrm{T}})$ of (3) into $o(\delta t)$. Then taking the conditional expectation of (3) given x, we have:

$$E[\delta G\,|\,x] = \sum_{i}^{n} G_{x_i} f_i\,\delta t + \sum_{i,j}^{n} G_{x_i x_j}(FDF^{\mathrm{T}})_{ij}\,\delta t + G_t\,\delta t + o(\delta t), \tag{6}$$

as G is a non-random function of the (random) variable x. Then taking the expected value of (6) we have, as $E[E[\delta G\,|\,x]] = E[\delta G]$:

$$E[\delta G] = \sum_{i}^{n} E[G_{x_i} f_i]\,\delta t + \sum_{i,j}^{n} E[G_{x_i x_j}(FDF^{\mathrm{T}})_{ij}]\,\delta t + E[G_t]\,\delta t + o(\delta t), \tag{7}$$

provided the distribution of $x(t)$ is such that the indicated expected values exist. Now (7) is a non-random equation, and so when we divide by δt and pass to the limit $\delta t \to 0$, and interchange the linear $E[.]$ and d.

† For example, $o(\delta t)$ denotes terms with the limiting property:

$$\lim_{\delta t \to 0} \frac{o(\delta t)}{\delta t} = 0.$$

operators on the left-hand side, we obtain the ordinary differential equation:

$$\frac{dE[G]}{dt} = \sum_i^n E[G_{x_i} f_i] + \sum_{i,j}^n E[G_{x_i x_j} (FDF^{\mathrm{T}})_{ij}] + E[G_t] \tag{8}$$

for the expected value of the function $G(x,t)$, where $d.$ is now the usual differential operator of ordinary differential equations.

§ 3. Equations for the System Moments

The formula (8) is useful for obtaining differential equations for the moments $m_{c_1,c_2,\ldots,c_n}(t)$ of order N of the system (1), where

$$m_{c_1,c_2,\ldots,c_n}(t) = E[G(x(t))], \tag{9}$$

$$G(x(t)) = x_1^{c_1}(t)\, x_2^{c_2}(t) \ldots x_n^{c_n}(t), \tag{9'}$$

$c_1, c_2, \ldots, c_n$ are non-negative integers satisfying

$$c_1 + c_2 + \ldots + c_n = N, \quad N = 1, 2, \ldots,$$

and

G is no longer an explicit function of t (i.e. $G_t = 0$ in (8)).

In particular, the equations for the first few moments are:

First moments $G(x) = x_i$

$$\dot{m}_{c_i=1}(t) = E[f_i]; \tag{10}$$

Second moments $G(x) = x_i^2$ or $x_i x_j$

$$\dot{m}_{c_i=2}(t) = 2E[x_i f_i] + 2E[(FDF^{\mathrm{T}})_{ii}], \tag{11a}$$

$$\dot{m}_{c_i=1,c_j=1}(t) = E[x_i f_j + x_j f_i] + 2E[(FDF^{\mathrm{T}})_{ij}]; \tag{11b}$$

Third moments $G(x) = x_i^3$ or $x_i^2 x_j$ or $x_i x_j x_k$

$$\dot{m}_{c_i=3}(t) = 3E[x_i^2 f_i] + 6E[x_i (FDF^{\mathrm{T}})_{ii}], \tag{12a}$$

$$\dot{m}_{c_i=2,c_j=1}(t) = E[2x_i x_j f_i + x_i^2 f_j] + 2E[x_j (FDF^{\mathrm{T}})_{ii} + 2x_i (FDF^{\mathrm{T}})_{ij}], \tag{12b}$$

$$\dot{m}_{c_i=1,c_j=1,c_k=1}(t) = E[x_j x_k f_i + x_i x_k f_j + x_i x_j f_k] + 2E[x_k (FDF^{\mathrm{T}})_{ij} + x_j (FDF^{\mathrm{T}})_{ik} + x_i (FDF^{\mathrm{T}})_{jk}]. \tag{12c}$$

In each case, suitable initial conditions are obtained from $x(0)$ or $P(x,0)$.

The higher moments follow by substituting the appropriate function (9') of $G(x)$ into (8). On the assumption that $f(x,t)$ and $F(x,t)$ can be expressed as polynomials in x with time-varying coefficients, the expected values on the right-hand side of the moment eqns. (10)–(12) reduce to linear functions of the moments (9). Then the eqns. (10) to (12) and all higher order ones form an infinite set of simultaneous linear first-order differential equations for the moments.

These are not very useful equations, however, unless the simplification of system linearity applies. The system (1) is said to be linear if $f(x,t)$ *and* $F(x,t)$ are linear functions of x (or trivially independent of x). Only in this case does the equation for the Nth-order moment involve moments

of order N or less on the right-hand side, and eqns. (10), (11), ... can be solved explicitly for increasing N. Furthermore, if $f(x,t)$ is linear in x and $F(x,t)$ independent of x, then $x(t)$ will be Gaussian and only the first two moments will be needed†.

§ 4. Example: Moments of a Second-order Linear System

As an example, consider the second-order linear system with stochastic coefficients discussed by Ariaratnam and Graefe (1965, p. 247, eqn. (24)), where, in terms of the coefficients of eqn. (1):

$$\begin{aligned} f_1 &= x_2, \\ f_2 &= -a_1 x_1 - a_2 x_2, \\ F_{1i} &= 0, \quad i = 1,3, \\ F_{21} &= 1, \\ F_{22} &= -x_1, \\ F_{23} &= -x_2. \end{aligned}$$

Then from (10) we have the first-moment equations:

$$\begin{aligned} \dot{m}_{1,0}(t) &= E[f_1] = m_{0,1}(t), \\ \dot{m}_{0,1}(t) &= E[f_2] = -a_1 m_{1,0}(t) - a_2 m_{0,1}(t), \end{aligned}$$

and from (11) the second-moment equations:

$$\begin{aligned} \dot{m}_{2,0}(t) &= 2m_{1,1}(t), \\ \dot{m}_{1,1}(t) &= m_{0,2}(t) - a_1 m_{2,0}(t) - a_2 m_{1,1}(t), \\ \dot{m}_{0,2}(t) &= 2D_{22} m_{2,0}(t) + 2(2D_{23} - a_1) m_{1,1}(t) + 2(D_{33} - a_2) m_{0,2}(t) \\ &\quad + 2D_{11} - 4D_{12} m_{1,0}(t) - 4D_{13} m_{0,1}(t). \end{aligned} \tag{13}$$

The last equation points out misprints in a result of Ariaratnam and Graefe (1965, p. 247, eqn. (27)) and the same result of Sancho (1965, p. 524). The factor 2 immediately preceding D_{23} in eqn. (13) is missing in each of these papers.

§ 5. Conclusions

This note derives a differential equation for the expected value of an arbitrary function $G(x,t)$ of a continuous stochastic system $x(t)$ and time. It is assumed that the stochastic system is Markov and is defined by an Ito stochastic differential eqn. (1). The differential equation for $E[G]$ is obtained simply by substituting the coefficients of the stochastic system (1) into the general result (8).

By choosing functions $G(x,t)$ according to eqn. (9), differential equations are obtained for the moments of the stochastic system $x(t)$. Equations for the first three moments have been written down in eqns. (10) to (12) and

† Non-Gaussian initial conditions will make $x(t)$ temporarily non-Gaussian, and more moments may be needed.

equations for the higher-order moments follow directly. For a given stochastic system, the moment equations are obtained from the coefficients of the Ito s.d.e. defining the system, by substituting the coefficients into the appropriate eqns. (10) to (12). This procedure, and the derivation which led to it, is simpler than the earlier method of Bogdanoff and Kozin (1962) which involved integrating the Fokker–Planck equation by parts.

It should be emphasized that the results of this note (and the method of Bogdanoff and Kozin) come from the properties of continuous Markov processes, as represented by the Ito s.d.e. (1). However, the system $x(t)$ of eqn. (1) has properties which preclude the system from being physically realizable (for example, $x(t)$ is nowhere differentiable). The system $x(t)$ exists only as a mathematical concept, convenient for analysis purposes, and any continuous stochastic process arising in physical or engineering situations has certain smoothness properties which prevent it from being a continuous Markov process. A good account of the differences between continuous Markov (diffusion) processes and physical processes has been given by Gray and Caughey (1965) and Kulman (1966), and recent research has indicated that under certain conditions a physical process can be approximated by a diffusion process. The nature of this approximation has been studied in detail by Clark (1966), and some extensions and examples are given by Cumming (1967). Using such an approximation, an approximate expression for the moments of a physical random process can be obtained by the methods of this note.

Acknowledgments

The author is deeply indebted to J. M. C. Clark for an introduction to the properties of the stochastic calculus. The work was done under the direction of Professor J. H. Westcott of Imperial College, and financial support was obtained from a NATO science scholarship.

References

Ariaratnam, S. T., and Graefe, P. W. U., 1965, *Int. J. Control*, **1**, 239.

Bogdanoff, J. L., and Kozin, F., 1962, *J. acoust. Soc. Am.*, **34**, 1063.

Clark, J. M. C., 1966, Ph.D. Thesis, Electrical Engineering Department, Imperial College, London, S.W.7.

Cumming, I. G., 1967, Ph.D. Thesis, Centre for Computing and Automation, Imperial College, London, S.W.7.

Doob, J. L., 1953, *Stochastic Processes* (New York: John Wiley).

Fuller, A. T., 1963, *J. Elect. Control*, 14, 669.

Gray, A. H., and Caughey, T. K., 1965, *J. Math. Phys.*, **44**, 288.

Kulman, N. K., 1966, *J. Math. Anal. Appl.*, **14**, 301.

Sancho, N. G. F., 1965, *Int. J. Control.*, **2**, 509.

Skorokhod, A. V., 1965, *Studies in the Theory of Random Processes* (Reading, Mass.: Addison-Wesley).

Stratonovich, R. L., 1966, *Siam J. Control*, **4**, 362.

Reference [88]

CUMMING (I.G.)
PhD. 1967

The Autocorrelation Function and Spectrum of a Filtered Pseudo Random Binary Sequence

I. G. Cumming

Abstract

This note derives the output autocorrelation function and power density spectrum of a maximum length pseudo random binary sequence passed through a linear first order filter. This derivation points out an error in an earlier paper by Roberts and Davis.

January 1967.

Research Note 2 - 67
Automatic Control Section,
Centre for Computing
and Automation,
Imperial College,
London, S.W.7.

1. Introduction

We consider a maximum length pseudo random binary sequence (PRBS) $x(t)$ obtained from an N stage shift register with a digit period of δ seconds. The PRBS has an amplitude of +1 or -1 and a period $L\delta = (2^N - 1)\delta$ seconds.

We consider the output $y(t)$ of a linear first order filter of impulse response $g(\tau) = ae^{-a\tau}$ where $T = a^{-1}$ seconds is the time constant (interchange) of the filter. We derive the cross-correlation $\phi_{yx}(s)$, the output autocorrelation $\phi_y(s)$, and the output power density spectrum $\overline{\phi}_y(f)$ when the input to the filter is the PRBS $x(t)$.

This has been done previously by Roberts and Davis [1], but an omission has restricted the range of values of filter time constant over which their result is accurate. Sections 2 and 3 below will parallel the derivation in reference [1].

2. Cross-correlation and Output Autocorrelation Function

From equation (1) of reference [1], the output autocorrelation function $\phi_y(s)$ is given by a double convolution of the filter impulse response and the input autocorrelation function $\phi_x(\tau)$:

$$\phi_y(s) = \int_s^\infty a\, e^{-a(u-s)} \left[\int_0^\infty \phi_x(u - \tau_1)\, a\, e^{-a\tau_1}\, d\tau_1 \right] du. \qquad (1)$$

As $\phi_y(s)$ is an even function and periodic with period $L\delta$, it must only be evaluated over the interval $0 \leq S \leq \frac{1}{2}L\delta$. The interior integral in brackets in equation (1) is the cross-correlation function

$$\phi_{yx}(u) = E\left[y(t)\, x(t - u) \right] \qquad (2)$$

and will be evaluated first. The autocorrelation of the input PRBS is

$$\phi_x(\tau) = -\frac{1}{L} + \frac{L + 1}{L\delta} \sum_{k=-\infty}^{\infty} f(\tau - kL\delta), \qquad (3)$$

where $f(\tau) = u_{-2}(\tau + \delta) - 2u_{-2}(\tau) + u_{-2}(\tau - \delta)$,

$u_{-2}(\tau)$ being the unit ramp function. The last term of equation (3) consists of an infinite series of triangular spikes centred at $\tau = kL\delta$, $k = -\infty, \ldots -1, 0, 1, 2 \ldots \infty$. For example, the central spike ($k = 0$) is

$$f(\tau) = \delta - |\tau|, \qquad -\delta \leq \tau \leq \delta,$$

$$= 0 \qquad \text{elsewhere.}$$

Cross-correlation function

The cross-correlation function (2) must be obtained for all positive u, but it is also a periodic function (with period $L\delta$) and so must only be evaluated for $0 \leq u \leq L\delta$. It is convenient to evaluate this function seperately in 3 segments:

Segment 1, $\delta \leq u \leq L\delta - \delta$.

The integral for the cross-correlation function

$$\phi_{yx}(u) = \int_0^\infty \phi_x(u - \tau_1)\ a\ e^{-a\tau_1}\ d\tau_1 \qquad (4)$$

consists of a constant $-\frac{1}{L}$ coming from the same term of equation (3) plus a contribution obtained from the convolution of the filter weighting function $a\ e^{-a\tau_1}$ and all those triangular spikes appearing in (3) for which $k \leq 0$. The central spike ($k = 0$) contributes

$$\frac{L + 1}{L}\ \frac{2}{a\delta}\ (\cosh(a\delta) - 1)\ e^{-au} \qquad (5)$$

to the integral (4), and each of the other spikes ($k = -1, -2, \ldots -\infty$) contribute $e^{akL\delta}$ times the quantity (5) to the integral (4). But the contribution of all these spikes can be grouped together by the infinite series property

$$\sum_{k=0}^{\infty} e^{-akL\delta} = \frac{1}{1 - e^{-aL\delta}}\ .$$

Thus the sum of all contributions to the integral (4) is

$$\phi_{yx}(u) = -\frac{1}{L} + \frac{L+1}{L}\,\frac{2}{a\delta}\,\frac{\cosh(a\delta) - 1}{1 - e^{-aL\delta}}\,e^{-au}, \qquad \delta \leq u \leq L\delta - \delta. \tag{6}$$

This is essentially the same expression obtained by Roberts and Davis, [1] equation (2). However, they do not evaluate $\phi_{yx}(u)$ over the following two segments.

Segment 2, $0 \leq u \leq \delta$.

A portion ($\delta - u$ long) of the $k = 0$ spike of $\phi_x(u - \tau_1)$ is now to the left of $\tau_1 = 0$ in the integral (4). In this case, it is convenient to rewrite (4) as

$$\phi_{yx}(u) = \int_{u-\delta}^{\infty} \phi_x(u - \tau_1)\, a\, e^{-a\tau_1}\, d\tau_1 - \int_{u-\delta}^{0} \phi_x(u - \tau_1)\, a\, e^{-a\tau_1}\, d\tau_1 . \tag{4'}$$

The first integral of (4') now includes the whole of the $k = 0$ spike along with all negative k spikes as before, and its evaluation is given by equation (6) as the e^{-au} factor of (6) automatically allows for the change in the lower integration limit of (4'). The second integral of (4') is *

$$\frac{L+1}{L}\,\frac{1}{\delta}\int_{u-\delta}^{0} (\delta - u + \tau_1)\, a\, e^{-a\tau_1}\, d\tau_1$$

$$= \frac{L+1}{L}\,\frac{1}{\delta}\left[u - \delta - \frac{1}{a}\left(1 - e^{-a(u-\delta)}\right)\right], \tag{7}$$

and subtracting (7) from (6) we have

---

* We omit the $-\frac{1}{L}$ term of $\phi_x(u)$ as its complete contribution has been included in equation (6).

$$\phi_{yx}(u) = -\frac{1}{L} + \frac{L+1}{L}\,\frac{2}{a\delta}\,\frac{\cosh(a\delta) - 1}{1 - e^{-aL\delta}}\,e^{-au}$$

$$-\frac{L+1}{L}\,\frac{1}{a\delta}\left[\,a(u-\delta) - 1 + e^{-a(u-\delta)}\right], \qquad (8)$$

$$0 \leq u \leq \delta.$$

Segment 3, $L\delta - \delta \leq u \leq L\delta$.

A portion ($u - L\delta + \delta$ long) of the $k = +1$ spike of $\phi_x(u - \tau_1)$ is now to the right of $\tau_1 = 0$ in the integral (4), but in evaluating (4) to obtain the result (6) we had not included any contribution of the $k = +1$ spike. Thus to equation (6) we must add this contribution of part of the $k = +1$ spike, which is

$$\frac{L+1}{L}\,\frac{1}{\delta}\int_0^{u - L\delta + \delta} (u - L\delta + \delta - \tau_1)\,a\,e^{-a\tau_1}\,d\tau_1$$

$$= \frac{L+1}{L}\,\frac{1}{\delta}\left[\,u - L\delta + \delta - \frac{1}{a}\left(1 - e^{-a(u - L\delta + \delta)}\right)\right]. \qquad (9)$$

Then adding (9) to (6) we obtain

$$\phi_{yx}(u) = -\frac{1}{L} + \frac{L+1}{L}\,\frac{2}{a\delta}\,\frac{\cosh(a\delta) - 1}{1 - e^{-aL\delta}}\,e^{-au}$$

$$+\frac{L+1}{L}\,\frac{1}{a\delta}\left[a(u - L\delta + \delta) - 1 + e^{-a(u - L\delta + \delta)}\right],$$

$$L\delta - \delta \leq u \leq L\delta. \qquad (10)$$

Output autocorrelation function

To find the output autocorrelation function we must evaluate the integral (1)

$$\phi_y(s) = \int_s^\infty a\, e^{-a(u-s)}\, \phi_{yx}(u)\, du \qquad (11)$$

over the range $0 \leq s \leq \frac{1}{2}L\delta$, taking care to use the appropriate segment of the periodic function $\phi_{yx}(u)$ over the {infinite range of the du} integral. The integral (11) can be broken into 3 parts

$$\phi_y(s) = I_1 + I_2 + I_3 \qquad (12)$$

where I_j is the contribution to (11) of the j:th segment of the function $\phi_{yx}(u)$, remembering that each segment is infinitely repeated along the u axis at intervals of $L\delta$.

It is convenient to evaluate the integral (11) for 3 cases of the shift parameter s.

Case 1, $s = 0$

The integral I_1 of (12) is obtained by integrating (11) over the infinite series of segments of the u axis

$$\delta + kL\delta \leq u \leq L\delta - \delta + kL\delta, \qquad k = 0, 1, 2 \ldots\ldots \infty .$$

The first term of this series (i.e. k = 0) is

$$I_{1,0} = \int_{\delta}^{L\delta - \delta} a\, e^{-au}\, \phi_{yx}(u)\, du,$$

where $\phi_{yx}(u)$ is taken from equation (6). The result is

$$I_{1,0} = \frac{1}{L}\left[e^{-a(L\delta-\delta)} - e^{-a\delta}\right] + \frac{L+1}{L}\,\frac{1}{a\delta}\,\frac{\cosh(a\delta) - 1}{1 - e^{-aL\delta}}\left[e^{-2a\delta} - e^{-2a(L\delta - \delta)}\right]. \tag{13}$$

Subsequent terms in this series are

$$I_{1,k} = \int_{\delta + kL\delta}^{L\delta - \delta + kL\delta} a\, e^{-au}\, \phi_{yx}(u)\, du, \qquad k = 1, 2, \ldots\ldots\infty,$$

$$= I_{1,0}\; e^{-akL\delta}, \tag{14}$$

as $\phi_{yx}(u) = \phi_{yx}(u + kL\delta)$. As before, the infinite series has the evaluation

$$I_1 = \sum_{k=0}^{\infty} I_{1,k} = \frac{I_{1,0}}{1 - e^{-aL\delta}}. \tag{15}$$

Paralleling this development for I_1, I_2 is found by first evaluating

$$I_{2,0} = \int_{0}^{\delta} a\, e^{-au}\, \phi_{yx}(u)\, du = \frac{1}{L}\left[e^{-a\delta} - 1\right] + \frac{L+1}{L}\,\frac{1}{a\delta}\,\frac{\cosh(a\delta) - 1}{1 - e^{-aL\delta}}\left[1 - e^{-2a\delta}\right]$$

$$- \frac{L+1}{L}\left[-1 + \frac{e^{a\delta}}{2a\delta}\,(1 - e^{-2a\delta})\right], \tag{16}$$

where $\phi_{yx}(u)$ comes from equation (8) and the rest of the series sums

to give

$$I_2 = \sum_{k=0}^{\infty} I_{2,k} = \frac{I_{2,0}}{1 - e^{-aL\delta}} \cdot \qquad (17)$$

Similarly,

$$I_{3,0} = \int_{L\delta - \delta}^{L\delta} a\, e^{-au}\, \phi_{yx}(u)\, du \,,$$

$$= \frac{1}{L}\left[e^{-aL\delta} - e^{-a(L\delta - \delta)} \right]$$

$$+ \frac{L+1}{L} \frac{1}{a\delta} \frac{\cosh(a\delta) - 1}{1 - e^{-aL\delta}} \left[e^{-2a(L\delta - \delta)} - e^{-2aL\delta} \right]$$

$$+ \frac{L+1}{L} \left[e^{-aL\delta} + \frac{1}{2a\delta} \left(e^{-a(L\delta - \delta)} - e^{-a(L\delta + \delta)} \right) \right] , \qquad (18)$$

where $\phi_{yx}(u)$ comes from equation (10) and the rest of the series sums to give

$$I_3 = \sum_{k=0}^{\infty} I_{3,k} = \frac{I_{3,0}}{1 - e^{-aL\delta}} \cdot \qquad (19)$$

Summing the 3 segments (12) using the results (13) to (19) we obtain the mean square value of the filtered PRBS y(t)

$$\phi_y(0) = \frac{L+1}{L} \frac{1}{a\delta} \frac{\cosh(a\delta) - 1}{1 - e^{-aL\delta}} \left(1 + e^{-aL\delta}\right)$$

$$- \frac{L+1}{L} \left(\frac{\sinh(a\delta)}{a\delta} - 1 \right) - \frac{1}{L} \cdot \qquad (20)$$

Comparing this result with equation (4) of reference [1], the factor $(1 + e^{-aL\delta})$ in the first term of equation (20) is missing in [1]. This means the results are equivalent only for sufficiently large values of $aL\delta$. In the example discussed in [1], the value $L = 127$ is used, and the result given is only accurate for filter time constants $T = a^{-1} \leq 10\delta$.

Our result has been checked computationally by simulating the filtered PRBS exactly on a digital computer (i.e. to 8 digit word-length accuracy), and estimating $\phi_y(0)$ by sampling $y(t)$ every $\frac{1}{40}\delta$ seconds. The result (20) was checked for $L = 127$ and values of filter time constant in the range $\delta \leq T \leq 200\delta$, and the accuracy of $\phi_y(0)$ was better than 0.1%, uniformly in T, where 0.1% was the order of accuracy of estimating $\phi_y(0)$ in the simulation.

The result (20) for the mean square value of the filtered PRBS has also been checked independently in [2] where an approximate expression for the transient statistics of a filtered PRBS is developed (in contrast, no approximations are made in the analysis of this note). The agreement with equation (20) is better than 1% except for low values of the filter time constant ($T \leq 25\delta$) when the relevance of the results of [2] to equation (20) diminishes (the results of [2] approximate the mean square of the sampled filter output, $y(n\delta)$, $n = 0, 1, 2 \ldots$, while the present results represent the mean square of the continuous filter output, $y(t)$).

Case 2, $0 \leq s \leq \delta$.

The integral (11) can be broken up into 2 parts,

$$\phi_y(s) = \int_0^{\infty} a\, e^{-a(u-s)}\, \phi_{yx}(u)\, du - \int_0^{s} a\, e^{-a(u-s)}\, \phi_{yx}(u)\, du. \qquad (21)$$

The first integral of (21) is recognised as $e^{as}\, \phi_y(0)$, and the second integral is evaluated using $\phi_{yx}(u)$ from (8). This is similar to equation (16) and has the result

$$-\int_0^{s} a\, e^{-a(u-s)}\, \phi_{yx}(u)\, du = \frac{L+1}{L}\,\frac{1}{a\delta}\,\frac{\cosh(a\delta)-1}{1-e^{-aL\delta}}\left[e^{-as} - e^{as}\right]$$

$$+\frac{L+1}{L}\,e^{as}\left[-(1-e^{-as}) + \frac{e^{a\delta}}{2a\delta}(1-e^{-2as}) - \frac{s}{\delta}\,e^{-as}\right] - \frac{1}{L}(1-e^{as}). \qquad (22)$$

Adding e^{as} times $\phi_y(0)$ of equation (20) to equation (22) we obtain

$$\phi_y(s) = \frac{L+1}{L}\,\frac{1}{a\delta}\,\frac{\cosh(a\delta)-1}{1-e^{-aL\delta}}\left(1 + e^{-a(L\delta - 2s)}\right) e^{-as}$$

$$-\frac{L+1}{L}\left[\frac{\sinh(a\delta - as)}{a\delta} - \left(1 - \frac{s}{\delta}\right)\right] - \frac{1}{L}, \qquad 0 \le s \le \delta. \qquad (23)$$

Case 3, $\delta \le s \le \frac{1}{2}L\delta$.

The integral (11) can be broken up into 3 parts,

$$\phi_y(s) = \int_0^{\infty} a\, e^{-a(u-s)}\, \phi_{yx}(u)\, du - \int_0^{\delta} a\, e^{-a(u-s)}\, \phi_{yx}(u)\, du$$

$$-\int_{\delta}^{s} a\, e^{-a(u-s)}\, \phi_{yx}(u)\, du. \qquad (24)$$

The first integral of (24) is $e^{as}\,\phi_y(0)$, and the second integral is the same as equation (22) with an upper integration limit of δ instead of s. This contribution to (24) equals

$$\frac{L+1}{L}\,\frac{1}{a\delta}\,\frac{\cosh(a\delta)-1}{1-e^{-aL\delta}}\left[e^{-2a\delta}-1\right]e^{as}-\frac{L+1}{L}\,e^{as}\left[1-\frac{e^{a\delta}}{2a\delta}(1-e^{-2a\delta})\right]$$

$$-\frac{1}{L}\left[e^{-a\delta}-1\right]e^{as} \qquad (25)$$

The third integral of (24) is similar to equation (13) and its contribution to (24) equals

$$\frac{L+1}{L}\,\frac{1}{a\delta}\,\frac{\cosh(a\delta)-1}{1-e^{-aL\delta}}\left[e^{-2as}-e^{-2a\delta}\right]e^{as}-\frac{1}{L}\left[e^{-as}-e^{-a\delta}\right]e^{as}. \qquad (26)$$

Summing e^{as} times $\phi_y(0)$ of equation (20) to equations (25) and (26) we obtain

$$\phi_y(s)=\frac{L+1}{L}\,\frac{1}{a\delta}\,\frac{\cosh(a\delta)-1}{1-e^{-aL\delta}}\left(1+e^{-a(L\delta-2s)}\right)e^{-as}-\frac{1}{L},$$

$$\delta \leq s \leq \tfrac{1}{2}L\delta. \qquad (27)$$

This result is the same as that for Case 2, equation (23), except that the middle term of (23) is not present in (27).

Equations (23) and (27) for the autocorrelation function of the filtered PRBS can be compared with equations (3b) and (3a) respectively of reference [1]. It is noted that the factor $(1+e^{-a(L\delta-2s)})$ in the first terms of equations (23) and (27) does not appear in reference [1]. Again this restricts the range of values of filter time constant T over which the expressions (3a) and (3b) of reference [1] are accurate. Equations (23) and (27) have been checked computationally for

$L = 127$, $\delta \leq T \leq 200\delta$ and $0 \leq s \leq 5\delta$, and were found to be correct to within the accuracy of the simulation (0.1%).

It should be mentioned that the examples quoted by Roberts and Davis [1] involves values of filter time constant in the range $0.9\delta \leq T \leq 6.8\delta$, and as their results are accurate over this range, the results of the present note do not affect the discussion and conclusions given in [1].

3. Output Power Spectrum

The power density spectrum $\Phi_x(f)$ of the PRBS x(t) can be found from the Fourier transform of the periodic autocorrelation function $\phi_x(\tau)$, equation (3). The transform relation for the discrete spectra with fundamental frequency $\frac{1}{L\delta}$ Hz is [3, page 12]

$$\Phi_x\left(\frac{r}{L\delta}\right) = \frac{2}{L\delta} \int_0^{\frac{L\delta}{2}} \phi_x(\tau) \cos\left(\frac{2\pi r\tau}{L\delta}\right) d\tau, \qquad (28)$$

where r is any integer. Evaluating this expression gives

$$\Phi_x(f) = \frac{1}{L^2}\, u_o(f) + \sum_{\substack{r=-\infty \\ r\neq 0}}^{\infty} \frac{L+1}{L^2}\left(\frac{\sin\frac{r\pi}{L}}{\frac{r\pi}{L}}\right)^2 u_o\left(f - \frac{r}{L\delta}\right), \qquad (29)$$

where $u_o\left(f - \frac{r}{L\delta}\right)$ is a unit impulse function centred at $f = \frac{r}{L\delta}$.

The power density spectrum $\Phi_y(f)$ of the filtered PRBS y(t) can be obtained by multiplying the input spectrum $\Phi_x(f)$ of equation (29) by the square of the modulus of the complex system function $\frac{1}{1 + j2\pi fT}$ of the filter (as in reference [1]), or by Fourier transforming the output autocorrelation function $\phi_y(s)$ of equations (23) and (27). Either method gives the result

$$\Phi_y(f) = \frac{1}{L^2}\, u_o(f) + \sum_{\substack{r=-\infty \\ r\neq 0}}^{\infty} \frac{L+1}{L^2}\left(\frac{\sin\frac{r\pi}{L}}{\frac{r\pi}{L}}\right)^2 \frac{1}{1 + \left(\frac{2\pi r}{L}\right)^2 \left(\frac{1}{a\delta}\right)^2} u_o\left(f - \frac{r}{L\delta}\right). \qquad (30)$$

The input and output power spectra of equations (29) and (30) indicate misprints in the similar equations given by Roberts and Davis [1]. In equations (5) and (7) of [1], the denominator immediately following the summation sign is L instead of L^2.

The expression (30) for the output power spectrum was checked by summing the series of equation (30) for $|r| \leq 250$ for the range of filter time constants $\delta \leq T \leq 200\delta$. The resulting value of signal power obtained agreed with the mean square value of y(t), equation (20), to within 0.1%, uniformly in T.

References

1. Roberts, P. D., and Davis, R. H.: 'Statistical Properties of Smoothed Maximal-length Linear Binary Sequences', Proc. IEE, Jan. 1966, 113, (1), pp.190-6.

2. Cumming, I. G.: 'A White Noise Model of a Pseudo Random Binary Sequence', to appear, Int. J. Control.

3. Lee, Y. W.: 'Statistical Theory of Communication', (Wiley 1960).