PREFACE

Teaching stochastic processes to students whose primary interests are in applications has long been a problem. On one hand, the subject can quickly become highly technical and if mathematical concerns are allowed to dominate there may be no time available for exploring the many interesting areas of applications. On the other hand, the treatment of stochastic calculus in a cavalier fashion leaves the student with a feeling of great uncertainty when it comes to exploring new material. Moreover, the problem has become more acute as the power of the differential equation point of view has become more widely appreciated. In these notes, an attempt is made to resolve this dilemma with the needs of those interested in building models and designing algorithms for estimation and control in mind. The approach is to start with Poisson counters and to identity the Wiener process with a certain limiting form. We do not attempt to define the Wiener process per se. Instead, everything is done in terms of limits of jump processes. The Poisson counter and differential equations whose right-hand sides include the differential of Poisson counters are developed first. This leads to the construction of a sample path representations of a continuous time jump process using Poisson counters. This point of view leads to an efficient problem solving technique and permits a unified treatment of time varying and nonlinear problems. More importantly, it provides sound intuition for stochastic differential equations and their uses without allowing the technicalities to dominate. In treating estimation theory, the conditional density equation is given a central role. In addition to the standard additive white noise observation models, a number of other models are developed as well. For example, the wide spread interest in problems arising in speech recognition and computer vision has influenced the choice of topics in several places.

Contents

Chapter 1

Probability Spaces

We recall a few basic ideas from probability theory and, in the process, establish some of the notation and language we will use.

1.1 Sets and Probability Spaces

A set consisting of a finite collection of elements is said to be *finite*. In this case the *cardinality* of the set is the number of elements in the set. If a set is finite, or if its elements can be put into one-to-one correspondence with the positive integers, it is said to be *countable*. Sets that are not countable, such as the set of real numbers between zero and one, are said to be *non-denumerably infinite*. If A and B are subsets of a set S we use $A \cup B$ and $A \cap B$ to denote the union and intersection of A and B, respectively. We denote the empty set by ϕ. We use the notation $\{0,1\}^S$ to indicate the set of all subsets of S with ϕ and S included. This is sometimes called the *power set*. It is easy to see that if S is a finite set with cardinality s then the cardinality of $\{0,1\}^S$ is 2^s and hence $\{0,1\}^S$ is also finite. On the other hand, if S is countably infinite then, as was discussed by Cantor, $\{0,1\}^S$ is non-denumerably infinite. This is a strong suggestion that one must exercise care in attempting to reason about the set of all subsets of S when S is infinite.

In 1933 the mathematician A. N. Kolmogorov described a precise mathematical model for the subject of probability. This model has come to be widely used because it is both elegant and self contained. It is not, however, necessarily easy to relate the Kolmogorov axioms to real world phenomena, nor are the axioms he used the only ones that deserve consideration. For example, the use of probability amplitudes in quantum mechanics calls for a rather different set of ideas. Although we will not give the arguments in detail, everything in these notes is consistent with the Kolmogorov point of view. The idea behind the Kolmogorov formalism is that one associates to the set S a collection of subsets $\mathcal{P}$, called the *events*, such that each $P \in \mathcal{P}$ has a well defined probability, $\mu(P)$. Our intuitive ideas about probability are mostly based on the analysis of simple situations for which S is a finite set and $\mathcal{P}$ is the set of all subsets of S. In this case we ask that $\mu(S) = 1, \mu(\phi) = 0$ and that μ be additive on disjoint sets. This amounts to asking that

$$\mu(P_1 \cup P_2) = \mu(P_1) + \mu(P_2) - \mu(P_1 \cap P_2)$$

Sometimes such μ's are called additive set functions.

If the cardinality of S is infinite, and this is the situation that occurs most frequently is our applications, then it is usually impossible to develop a satisfactory theory in which $\mathcal{P}$ is the set

of all subsets of S. The problem stems from the fact that although the set of all subsets of a finite set is finite, the cardinality of the set of all subsets of an infinite set is not only infinite but is essentially larger than that of the original set in the sense that the elements of S cannot be placed in one-to-one correspondence with those of $\{0,1\}^S$. One needs to be content with a more restricted theory. In this case one asks that μ be additive on *countable* disjoint unions i.e., that

$$\mu\left(\bigcup_{i=1}^{\infty} P_i\right) = \sum_{i=1}^{\infty} \mu(P_i)$$

if the P_i are pair-wise disjoint but makes no claim about more general unions.

Collections of subsets, finite or infinite, that are closed under finite unions and intersections are said to form a *field of sets*. A collection $\mathcal{P}$ of subsets of a set S is said to be a *σ-field* if it is closed under countably infinite unions and intersections. The natural setting for probability theory is $(S, \mathcal{P}, \mu)$ with $\mathcal{P}$ being a σ-field and μ being additive on countable disjoint unions.

The basic construct in the Kolmogorov system is a triple $(S, \mathcal{P}, \mu)$, called a *probability space*. The elements have the following properties:

1. S = set
2. $\mathcal{P}$ = a collection of subsets of S that includes S and the empty set and is closed under complementation, countable unions and countable intersections.
3. μ = a map of $\mathcal{P}$ into $[0,1]$ such that if $\{A_i\}$ is a disjoint collection then

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i)$$

In order to describe a specific probability space it is necessary to specify $S, \mathcal{P}$ and μ. If S is a finite set then we can take $\mathcal{P}$ to be the set of all subsets of S and can specify μ by giving its value on each element of S with the understanding that its value on arbitrary subsets is just the sum of its values on the constituents of the subset.

Example: Let S be the set of n-tuples, $S = \{(a_1, a_2, \ldots, a_n)\}$, with the a_i taking on the value zero or the value one. This set has cardinality 2^n. We take $\mathcal{P}$ to be the set of all subsets of S, a set with cardinality 2^{2^n}. Let α be a real number between zero and one and let the probability of the occurrence of $(a_1, a_2, \ldots, a_n)$ be

$$p\big((a_1, a_2, \ldots, a_n)\big) = \alpha^{\Sigma a_i}(1-\alpha)^{n-\Sigma a_i}$$

In view of the identity $1 = (\alpha + (1-\alpha))^n$ we see that $\sum_{k=0}^{n} \binom{n}{k} \alpha^k (1-\alpha)^{n-k} = 1$. We extend the definition of p from S to P using $p(A \cup B) = p(A) + p(B)$ for $A \cap B = \phi$. Thus p defines a probability measure on the set of all binary n-tuples.

1.2 Probability Distributions on Vector Spaces

We often base our models of real world phenomena on variables whose measured value can be any real number or on vectors whose components are real numbers. This works well for phenomena described by differential equations, linear algebra, etc. It is, however, the source of some technical problems in probability theory because of the necessity of defining a suitable collection of subsets

of $\mathbb{R}$. More specifically, one needs to define a collection that is big enough to be useful but not so big as to cause logical difficulties. It may be of some comfort to consider the fact that the rational numbers, like the integers, are countable and the set of n-tuples with rational entries are countable as well.

The standard topology on $\mathbb{R}$ can be explained this way. We say that a subset $P \subset \mathbb{R}$ is *open* if for each $x_0 \in P$ there exists an $\epsilon > 0$ such that $\{x \big| |x - x_0| < \epsilon\} \subset P$. It is not hard to show that finite unions and finite intersections of open sets are open. The countably infinite cases are different. The infinite union of open sets is open, but it can happen that an infinite intersection of open sets is closed. For example

$$\bigcap_{n=1}^{\infty} \{x \big| (-1/n \leq x \leq 1 + 1/n)\} = [0, 1]$$

Thus the smallest σ-field that contains all the open intervals $(a/b, c/d)$ with a, b, c, d being integers (a countable family) contains all the closed intervals as well.

There is a rich collection of subsets of $\mathbb{R}^n$ called the *Borel sets* which are generated from the open subsets of $\mathbb{R}^n$ by taking countable intersections and countable unions, countable intersections of countable unions, etc. The development of measure theory, which we do not undertake here, focuses on Borel sets and, in some cases, a certain larger collection called Lebesgue measurable sets. For our purposes the Borel sets are adequate. It is often useful to consider the probability triple $(S, \mathcal{P}, \mu)$ where S is $\mathbb{R}^n$, $\mathcal{P}$ is the Borel sets and μ is the measure one gets by extending the ordinary idea of length to the Borel sets.

If $S = (-\infty, \infty)$ and if $\mathcal{P}$ is the Borel σ-field then we can specify μ by giving a nondecreasing function ϕ which maps $(-\infty, \infty)$ onto $[0, 1]$ with

$$\phi(\hat{x}) = \text{ probability } x \leq \hat{x}$$

In this case ϕ is said to be the *probability distribution function*. We see immediately that

$$\mu\big((a, b]\big) = \phi(b) - \phi(a)$$

and, using the axioms, we can extend the definition of μ to countable unions and countable intersections. If it happens that ϕ is a differentiable function of x then $d\phi/dx$ is called the *probability density* associated with $(S, \mathcal{P}, \mu)$.

Example 1: One-dimensional gaussian densities. The integral $I = \int_{-\infty}^{\infty} e^{-x^2} dx$ can be evaluated by noticing that $I^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-(x^2+y^2)} dx dy$ and then transferring to polar coordinates to get $I^2 = \int_0^{\infty} \int_0^{2\pi} r e^{-r^2} d\theta dr = \pi$. By rescaling x we see that

$$\int_{-\infty}^{\infty} e^{-x^2/2\sigma} dx = \sqrt{2\pi\sigma}$$

thus $(1/\sqrt{2\pi\sigma}) e^{-x^2/2\sigma}$ is a probability density on $(-\infty, \infty)$. Karl Fredrich Gauss (1777-1855), in addition to his many other claims to fame, made an extensive study of this density and for this reason it is usually named after him. Generalizing slightly so as to include a possible shift of origin the gaussians take the form

$$\rho(x) = \frac{1}{\sqrt{2\pi\sigma}} \, e^{-(x-m)^2/2\sigma}$$

with m being the mean and σ the variance. An integration by parts is all that is required to show that the moments of a zero mean Gaussian random variable satisfy

$$\mathcal{E}x^p = (2\pi\sigma)^{-1/2}\int_{-\infty}^{\infty} x^p e^{-x^2/2\sigma} = \sigma(p-1)\mathcal{E}x^{p-2}$$

and so, while for p odd $\mathcal{E}(x-m)^p = 0$, for p even we have

$$\mathcal{E}(x-m)^p = \sigma^{p/2}(p-1)(p-3)\cdots(3)(1)$$

thus

$$\mathcal{E}(x-m)^p = (\sigma/2)^{p/2}p!/(p/2)!$$

The Gaussian density is the only probability density with these moments.

Example 2: n-dimensional gaussian densities. Now let x be an n-vector. Because the integral over all space of $e^{-x_1^2/2\sigma_1}e^{-x_2^2/2\sigma_2}\cdots e^{-x_n^2/2\sigma_n}$ is just the product of the integrals of the factors we see that

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} e^{-x_1^2/2\sigma_1}e^{-x_2^2/2\sigma_2}\cdots e^{-x_n^2/2\sigma_n}dx_1dx_2\cdots dx_n$$
$$= \sqrt{(2\pi)^n\sigma_1\sigma_2\cdots\sigma_n}$$

We may say, therefore, that if D is diagonal and positive definite with eigenvalues $\{d_i\}$, then

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} e^{-x^T(2D)^{-1}x}dx_1dx_2\cdots dx_n = \sqrt{(2\pi)^n d_1d_2\cdots d_n}$$

Given any n by n matrix Q which is symmetric and positive definite one can find an orthogonal matrix Θ such that $\Theta^TQ\Theta$ is diagonal. Thus by letting $z = \Theta^Tx$ we see that the density

$$\rho(x) = \frac{1}{\sqrt{(2\pi)^n\det Q}}\, e^{-x^T(2Q)^{-1}x}$$

becomes

$$\rho(z) = \frac{1}{\sqrt{(2\pi)^n\det Q}}\, e^{-z^T\Theta^T(2Q)^{-1}\Theta z}$$

which is a nonnegative function such that

$$\int_{\mathbb{R}^n}\rho(x)dx = 1$$

That is, $\rho(x)$ is a probability density. If we make the further change of variables, $w = x + m$ we get the still normalized, but more general, form

$$\rho(w) = \frac{1}{\sqrt{(2\pi)^n\det Q}}e^{-(w-m)^T(2Q)^{-1}(w-m)}$$

called the *multidimensional Gaussian density* of variance Q and mean m. That is to say, we have the identities

$$m = \frac{1}{\sqrt{(2\pi)^n\det Q}}\int_{\mathbb{R}^n} xe^{-(x-m)^T(2Q)^{-1}(x-m)}dx$$

and

$$Q = \frac{1}{\sqrt{(2\pi)^n\det Q}}\int_{\mathbb{R}^n}(x-m)(x-m)^Te^{-(x-m)^T(2Q)^{-1}(x-m)}dx$$

valid for all $Q = Q^T > 0$. The latter of these equations can be verified by computing the derivative with respect to α of

$$\sqrt{(2\pi)^n \det \alpha Q} = \int_{\mathbb{R}^n} e^{-x^T(2\alpha Q)^{-1}x}\, dx$$

and then evaluating the result at $\alpha = 1$ to get

$$\frac{-n}{2\sqrt{(2\pi)^n \det Q}} = \int_{\mathbb{R}^n} -\frac{1}{2}x^T Q^{-1} x e^{x^T(2Q)^{-1}x}\, dx$$

1.3 Independence and Conditional Probability

Given a probability space $(S, \mathcal{P}, \mu)$ with P_1 and P_2 in $\mathcal{P}$, Consider the difference

$$d(P_1, P_2) = \mu(P_1 \cap P_2) - \mu(P_1) \cdot \mu(P_2)$$

In general, this need not be zero. However, if it is we will say that P_1 and P_2 are *independent*. Assuming independence often simplifies the analysis. On the other hand, when making measurements one hopes that the results of the measurements will reveal something about the state of the system being measured, i.e., that the measurements are not independent of the other variables present. This leads to the study of conditional probabilities. We use the notation $\mu(P_1|P_2)$ to denote the probability of P_1 given P_2.

Because estimation theory involves estimating a random variable on the basis of observations, we are often in the situation of computing conditional probabilities. A basic tool will be the rule of Bayes which expresses the conditional probability of A given B in terms of the conditional probability of B given A and the probabilities of A and B alone. This rule is

$$p(A|B) = p(B|A) \cdot p(A)/p(B)$$

The demonstration of this fact is completely elementary being based on the fact that one of the four possibilities, A and B, A and not B, not A and B, not A and not B, must occur. If they occur with probabilities $p_1,\ p_2,\ p_3$ and p_4 respectively, then

$$p(A|B) = \frac{p_1}{p_1 + p_3}$$

and

$$p(B|A) \cdot p(A)/p(B) = \frac{p_1}{p_1 + p_2} \cdot \frac{p_1 + p_2}{p_1 + p_3}$$

These are equal and nothing more is to be said.

If we are dealing with continuous random variables which have probability densities, then there is an infinitesimal version of the rule in which the probability p is replaced by the density ρ, i.e.,

$$\rho(a|b) = \rho(b|a) \cdot \rho(a)/\rho(b)$$

To see this, let a and b be real valued random variables with a joint probability density ψ, then the conditional probability of a given b is clearly

$$\rho(a|b) = \psi(a, b) / \int_{-\infty}^{\infty} \psi(\xi, b) d\xi$$

A similar formula holds for $\rho(b|a)$, the probability density of a is just

$$\bar{\rho}(a) = \int_{-\infty}^{\infty} \psi(a,\xi)d\xi$$

Thus we have a probability density version of Bayes' rule which is identical in form to the non-infinitesimal version with the probabilities being replaced by probability densities.

Example 1: If we observe $x+n=y$ and if we know that x, n and y are real valued random variables with densities ρ_1, ρ_2 and ρ_3, respectively, then, assuming that n and x are independent, the probability density for x, conditioned on the fact that $x+n=y$, is

$$\begin{aligned}\rho(x|y) &= \rho(y|x)\cdot\rho_1(x)\big/\rho_3(y) \\ &= \rho_2(y-x)\cdot\rho_1(x)\big/\int_{-\infty}^{\infty}\rho_2(y-\eta)\cdot\rho_1(\eta)d\eta\end{aligned}$$

Notice that we used a convolution formula to express the density of $y=x+n$ in terms of the densities of x and n

$$\rho_3(y) = \int_{-\infty}^{\infty}\rho_2(y-\eta)\rho_1(\eta)d\eta$$

Example 2: If in the previous example x and n have Gaussian distributions with means $\bar{x}$ and 0, respectively, and variances σ_x and σ_n, respectively, then having observed that $x+n=y$ the conditional density for x changes from its *a priori* form

$$\rho_1(x) = \frac{1}{\sqrt{2\pi\sigma_x}}\, e^{-(x-\bar{x})^2/2\sigma_x}$$

to

$$\begin{aligned}\rho(x|y) &= c\cdot e^{-(y-x)^2/2\sigma_n}\, e^{-(x-\bar{x})^2/2\sigma_x} \\ &= \frac{1}{\sqrt{2\pi\sigma_3}} e^{-(x-\alpha)^2/2\sigma_3}\end{aligned}$$

with $\sigma_3 = \sigma_x\sigma_n/(\sigma_x+\sigma_n)$ and α being $(\sigma_x y+\sigma_n\bar{x})/(\sigma_x+\sigma_n)$.

1.4 Statistical Measures

If x is a real valued random variable distributed with density ρ and if $\psi\colon \mathbb{R}\to\mathbb{R}$ is a map then one can investigate the existence of

$$\mathcal{E}\psi(x) = \int_{-\infty}^{\infty}\psi(x)\rho(x)dx$$

Of course such integrals can fail to exist in a variety of ways. If

$$\mathcal{E}x = \int_{-\infty}^{\infty} x\rho(x)dx$$

exists then we say that the density ρ has finite expectation. If

$$\mathcal{E}x^2 = \int_{-\infty}^{\infty} x^2\rho(x)dx$$

exists we will say that the density ρ has finite variance.

1.5 Transformation of Densities

If y is a random variable that takes on values in $\mathbb{R}^n$ and if $\phi\colon \mathbb{R}^n \to \mathbb{R}^m$ is a differentiable map, then $x = \phi(y)$ is also a random variable. Many problems involve making computations that characterize the way x is distributed, given the distribution of y. In the situation where y has a smooth density and $\phi\colon \mathbb{R}^1 \to \mathbb{R}^1$ the matter can be dealt with rather easily. Let $S \subset \mathbb{R}^1$ denote the set $S = \{x | x = \phi(y) \text{ has a solution}\}$. If x has a probability density $\tilde{p}$ it follows that $\tilde{p}$ is zero for any x not in S. On the other hand, at a point x_0 in S there is an inverse image y_0 such that $x_0 = \phi(y_0)$. If we linearize ϕ near y_0 then

$$\phi(y_0 + \delta) = x_0 + \left.\frac{\partial \phi}{\partial y}\right|_{y_0} \delta + \text{ higher order terms}$$

and the measure of the set of points $T = \{y| \ |y - y_0| \leq \epsilon\}$ is approximately $p(y_0) \cdot \epsilon$. Thus the measure of the set that T maps into is $p(y_0) \cdot \epsilon$. However, the length of this interval in x-space is the absolute value of $\left.\frac{\partial \phi}{\partial y}\right|_{y_0} \epsilon$. Thus, we see that $\tilde{p}(x) = \frac{1}{|\partial \phi / \partial y|} p\big(\phi^{-1}(x)\big)$. However a given x may have several inverse images. In such cases we need to express matters as

$$\tilde{p}(x) = \sum_{\{y | \phi(y) = x\}} \frac{1}{|\partial \phi / \partial y|} p(\phi^{-1}(x))$$

with the understanding that $\partial \phi / \partial y$ is to be evaluated at the appropriate inverse image of x. Suppose now that $m = n > 1$. If $\rho_1(\cdot)$ and $\rho(\cdot)$ are the probability densities of the random variables x and y with x and y taking on values in $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively, then the appropriate formula involves the determinant of the Jacobian

$$\rho_1(x) = \sum_{\{y | \phi(y) = x\}} \frac{1}{|\det(\partial \phi / \partial y)|} \rho(\phi^{-1}(y))$$

where the sum is to be taken over all inverse images of the given value of y. Of course this only works if $x = \phi(y)$ has a finite number of solutions.

Example 1: Suppose that y has gaussian density $\rho(y) = \frac{1}{\sqrt{2\pi\sigma}} e^{-y^2/2\sigma}$ and that $r = y^2$. Thus $dr = 2ydy$. Let $\tilde{\rho}$ denote the density with respect to r. Then $\tilde{\rho}$ is zero for negative arguments. There are two inverse images for $r > 0$ and the above formula yields

$$\tilde{\rho}(r) = \frac{1}{\sqrt{2\pi\sigma}} \left(\frac{1}{2\sqrt{r}} e^{-r/2\sigma} + \frac{1}{2\sqrt{r}} e^{-r/2\sigma} \right) = \frac{1}{\sqrt{2\pi\sigma}} \frac{1}{\sqrt{r}} e^{-r/2\sigma}$$

Example 2: Let x_1 and x_2 be gaussian with density

$$\rho(x_1, x_2) = \frac{1}{2\pi\sigma} e^{-(x_1^2 + x_2^2)/2\sigma}$$

with $r = \sqrt{x_1^2 + x_2^2}$ and let $\theta = \tan^{-1}(x_2/x_1)$. Find the density with respect to (r, θ). In this case r is non negative and each $r > 0$ has a circle of inverse images. The Jacobian of the map from (x_1, x_2) to (r, θ) is given by

$$\begin{bmatrix} \frac{\partial r}{\partial x_1} & \frac{\partial r}{\partial x_2} \\ \frac{\partial \theta}{\partial x_1} & \frac{\partial \theta}{\partial x_2} \end{bmatrix} = \begin{bmatrix} \frac{x_1}{\sqrt{x_1^2 + x_2^2}} & \frac{x_2}{\sqrt{x_1^2 + x_2^2}} \\ \frac{-x_2}{x_1^2 + y_2^2} & \frac{x_1}{x_1^2 + x_2^2} \end{bmatrix}$$

The determinant of the Jacobian is simply r^{-1}, and so

$$\tilde{\rho}(r, \theta) = \frac{r}{2\pi\sigma} e^{-r^2/2\sigma}$$

Remark: If x and y take on values in $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively, then the map $(x, y) \mapsto x$ gives rise to the change of density formula

$$\rho_1(x) = \int_{\mathbb{R}^m} \rho_0(x, y) dy$$

Example 3: Suppose that (x, y) is a gaussian random vector in $\mathbb{R}^2$, with zero mean and variance

$$\Sigma = \begin{bmatrix} a & b \\ b & a \end{bmatrix}$$

Show that the probability that the product xy is positive is given by

$$p = \frac{2}{\pi} \tan^{-1} \sqrt{\frac{a+b}{a-b}}$$

We begin by observing the identity

$$\begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} a & b \\ b & a \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} a+b & 0 \\ 0 & a-b \end{bmatrix}$$

This shows that the change of variables

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{a+b}} & 0 \\ 0 & \frac{1}{\sqrt{a-b}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

results in a pair of variables that are gaussian with zero mean and unity variance. Thus, key to the solution of the problem is determining which values of (u, v) correspond to a positive value of xy. To this end, observe that xy changes sign if and only if either x or y changes sign. However,

$$\sqrt{2}\sqrt{a+b}u = x + y$$

$$\sqrt{2}\sqrt{a-b}v = -x + y$$

Thus $x = 0$ implies that

$$\sqrt{2}\sqrt{a+b}u = \sqrt{2}\sqrt{a-b}v$$

and $y = 0$ implies that

$$\sqrt{2}\sqrt{a+b}u = -\sqrt{2}\sqrt{a-b}v$$

The straight lines defined by these equations pass through the origin and define an angle

$$\theta = tan^{-1} \frac{\sqrt{a+b}}{\sqrt{a-b}}$$

Because the denisty is circularly symmetric in this coordinate system, this angle, divided by π is the value of the integral of the density over the region where xy is positive.

Example 4: Suppose that x is a real valued random variable taking on values in the interval $[0, \infty)$. Suppose that the probability density for x is $\rho(x) = \lambda e^{-\lambda x}$. If $y = \lfloor x \rfloor$, i.e., if y is the largest integer less than or equal to x then what is the probability density for y and what is the expected value of y^p? Direct analysis shows that the probability p_n that $y = n$ is given by

$$p_n = \int_n^{n+1} \lambda e^{-\lambda x} dx = e^{-\lambda n}(1 - e^{-\lambda})$$

To evaluate the expected value of y^p we need to evaluate

$$\mathcal{E}y^p = \sum_{n=0}^{\infty} n^p e^{-\lambda n}(1 - e^{-\lambda})$$

However, if we observe that this is expressible as

$$\mathcal{E}y^p = (-1)^p(1 - e^{-\lambda})\frac{d^p}{d\lambda^p}\sum_{n=0}^{\infty} e^{-\lambda n}$$

then, summing the geometric series, we see that

$$\mathcal{E}y^p = (-1)^p(1 - e^{-\lambda})\frac{d^p}{d\lambda^p}\frac{1}{e^{\lambda} - 1}$$

and, in particular, that $\mathcal{E}y = 1/(e^{\lambda} - 1)$

Example 5: Let $H = H^T$ be an n by n matrix. We denote its ij^{th} entry by h_{ij} and of course $h_{ij} = h_{ji}$. We assume that the h_{ij} are random and are distributed according to a zero mean gaussian distributions. The h_{ij} are independent except for the requirement that $h_{ij} = h_{ji}$. Given this distribution on the entries of H we want to use the change of coordinates formula to find the distribution law for the eigenvalues. To begin with, we recall that it is possible to represent an arbitrary symmetric matrix H as $\Theta^T\Lambda\Theta$ with Λ being diagonal and Θ being orthogonal. In this case the diagonal entries of Λ are the eigenvalues. Given H, there is more than one orthogonal matrix such that $\Theta H\Theta^T$ is diagonal. However, if the eigenvalues of H are distinct then Θ is unique to within multiplication on the left by a diagonal matrix D whose diagonal entries are ± 1 and a further multiplication on the left by a permutation matrix P. Taking the derivative of $\Theta^T\Lambda\Theta = H$ we get $d\Omega H - Hd\Omega = dH$ with $d\Omega$ being skew symmetric. Adding in the change of Λ we get

$$Hd\Omega - d\Omega H + \Theta^T(d\Lambda)\Theta = dH$$

where $d\Omega$ is skew-symmetric and equal to $d\Theta\Theta^T$ Write $ad_H(d\Omega) = [H, d\Omega]$. We see that the inverse image in H space associated with the set of eigenvalues

$$|\lambda_1 - \lambda_a| \leq \epsilon, |\lambda_2 - \lambda_b| \leq \epsilon \ldots, |\lambda_n - \lambda_c| \leq \epsilon$$

is just $ad_H^{-1}(dH)$. We can use this to show that the eigenvalues of H are distributed according to the law

$$\rho(\lambda_1, \lambda_2, \ldots, \lambda_n) = Ne^{-\Sigma\lambda_1^2}\prod_{i,j}(\lambda_i - \lambda_j)$$

where N is the normalization factor needed to make the area under the function one.

1.6 Sampling

Given a random variable x whose density is unknown one can attempt to find the density by sampling x repeatedly and then determining ρ from various "averages" of the samples. For example, one can form the sample mean

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k$$

and the sample variance

$$\mathcal{E}(x - \bar{x})^2 = \frac{1}{n}\sum_{l=1}^{n}\left(x_l - \frac{1}{n}\sum_{k=1}^{n} x_k\right)^2$$

and show that under reasonable assumptions on the underlying distribution these sample statistics will approach the true mean and variance, as k approaches infinity.

1.7 Exercises 1

1. Consider a coin which comes up heads with probability p and tails with probability $q = 1-p$. If the coin is tossed n times and the tosses are independent show that the probability that the n tosses will yield exactly k heads is

$$C_{k,n} = \binom{n}{k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

2. Show that

$$\lim_{\epsilon \to \infty} \left(1 - \frac{t}{\epsilon}\right)^{\epsilon} = e^{-t}$$

Give this a probabilistic interpretation, restricting ϵ to integer values.

3. If ln denotes the natural logarithm then

$$\ln\, n! = \sum_{k=2}^{n} \ln\, k \approx \int_2^n \ln\, x\, dx = x\, \ln\, x\big|_2^n - \int_2^n 1\, dx = n\, \ln(n) - 2\ln(2) - (n-2)$$

and so

$$n! \approx n^n e^{-n}$$

Refine this idea to obtain the better approximation

$$n! \approx e\sqrt{n} n^n e^{-n}$$

If you are more ambitious you can try to get the still better Stirling's approximation

$$n! \approx \sqrt{2\pi n} n^n e^{-n}$$

Hint: For the latter analysis you may wish to consider the gamma function

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$$

which, as can be verified by repeated integration-by-parts, for positive integer values of x satisfies $\Gamma(x) = (x-1)!$.

4. If x_1 and x_2 are scalar Gaussian random variables with zero mean and variance σ and if $y_1 = \max\{x_1, x_2\}$, and $y_2 = \min\{x_1, x_2\}$, show that the probability density for y_1 is

$$\rho(y_1) = \frac{1}{\sqrt{2}\pi\sigma} \int_{-\infty}^{y} e^{-\eta^2/2\sigma} dy$$

5. Let H be an n by n symmetric matrix with real entries. Suppose that the entries are independent and identically distributed random variables, distributed according to a gaussian distribution with zero mean and variance σ. Show that the sum of the squares of the eigenvalues of H is distributed according to the Poisson density for $\beta = \Sigma\lambda^2$

$$\rho(\beta) = N e^{-\beta/n\sigma}$$

6. Let $\rho_1 \colon (-\infty, \infty) \to [0, \infty)$ and $\rho_2 \colon (-\infty, \infty) \to [0, \infty)$ be such that

$$\int_{-\infty}^{\infty} \rho_i(x) dx = 1; i = 1, 2$$

Show that if ρ_3 is defined by

$$\rho_3(x) = \int_{-\infty}^{\infty} \rho_1(x-\eta)\rho_2(\eta)d\eta$$

then

$$\int_{-\infty}^{\infty} \rho_3(x)dx = 1$$

7. Let ρ_1 and ρ_2 in problem 6 be equal and suppose that

$$\int_{-\infty}^{\infty} x\rho_1(x)dx = 0 \; ; \; \int_{-\infty}^{\infty} x\rho_2(x)dx = 0$$

and

$$\int_{-\infty}^{\infty} x^2\rho_1(x)dx = a \; ; \; \int_{-\infty}^{\infty} x^2\rho_2(x)dx = a$$

Show that

$$\int_{-\infty}^{\infty} x^2\rho_3(x)dx = 4a$$

8. If $\mu_1 \colon \mathcal{P} \to [0,1]$ and $\mu_2 \colon \mathcal{P} \to [0,1]$ are two different probability measures corresponding to the same choice of S and $\mathcal{P}$ then we can define the distance between μ_1 and μ_2 as

$$d(\mu_1, \mu_2) = \max_{P \in \mathcal{P}} |\mu_1(P) - \mu_2(P)|$$

If S is a finite set and $\mathcal{P}$ is the set of all subsets of S then show that

$$d(\mu_1, \mu_2) = \sum_{s_i \in S} |\mu_1(s_i) - \mu_2(s_i)|$$

show that $d(\mu_1, \mu_2) + d(\mu_2, \mu_3) \geq d(\mu_1, \mu_3)$.

9. Let S be a finite set and let $\mathcal{P}$ be the set of all subsets of S. Find the probability measure $\mu \colon \mathcal{P} \to [0,1]$ that maximizes the sum

$$H = \sum_{s_i \in S} \mu(s_i) \ln(\mu(s_i))$$

10. If $S = (-\infty, \infty)$ and $\mathcal{P}$ is set of all Borel subsets of $(-\infty, \infty)$ find the probability density function μ that maximizes

$$H = \int_{-\infty}^{\infty} \mu(x) \ln \mu(x) \, dx$$

subject to $\mathcal{E}x = 0$ and $\mathcal{E}x^2 = \sigma$.

11. Suppose that $a_1, a_2, \ldots, a_n$ are independent gaussian random variables with mean 0 and variance σ. Consider

$$x_n(t) = a_0 + \sqrt{2} \sum_{k=1}^{n} a_k \cos kt$$

(i) Show that $\mathcal{E}\, x_n(t) = 0$

(ii) Show that $\mathcal{E}\, x_n^2(t) \leq n$

(iii) Compute the probability density for $x(t)$

If we integrate this sum from $t = 0$ we get

$$y_n(t) = a_0 t + \sqrt{2} \sum_{k=1}^{n} \frac{a_k}{k} \sin kt$$

Repeat the above calculations for $y_n(t)$. What about taking the limit as n goes to infinity?

12. Let $Q = Q^T$ be positive definite and let x be a vector-valued Gaussian random variable with zero mean and variance Q Assume $R = R^T$. Show that

$$\mathcal{E} e^{x^T (2R)^{-1} x}$$

exists if $Q^{-1} - R^{-1}$ is positive definite and that in this case

$$\mathcal{E} e^{x^T (2R)^{-1} x} = \frac{\sqrt{\det(Q^{-1} - R^{-1})^{-1}}}{\sqrt{\det Q}}$$

13. If x is a random variable that takes on values in a finite set $x = \{x_1, x_2, \ldots, x_n\}$ then the possible ways to assign probabilities to this set can be identified with the set

$$S = \{(p_1, p_2, \ldots, p_n) | p_i \geq 0; \Sigma p_i = 1\}$$

S is a closed bounded subset of $\mathbb{R}^n$ called the *standard simplex*. (a) Show that if T is a linear transformation of $\mathbb{R}^n$ into $\mathbb{R}^n$ then T maps the standard simplex into itself if and only if the matrix representation of $T = (t_{ij})$ has nonnegative entries and

$$\sum_{j=1}^{n} t_{ij} = 1$$

14. consider a random vector in $\mathbb{R}^2$ with a gaussian distribution having mean 0 and variance Σ. Suppose that

$$\Sigma = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$$

Show that if $a = c$ then the probability that $x_1 x_2$ is positive is

$$p = \frac{1}{\pi} \tan^{-1} \frac{a+b}{a-b}$$

Hint: The change of variables $y_1 = (x_1 + x_2)/\sqrt{2}, y_2 = (x_1 - x_2)/\sqrt{2}$, followed by the change of variables $z_1 = \sqrt{(a+b)} y_1, z_2 = \sqrt{(a-b)} y_2$ makes the equiprobable contours circles. The positive quadrant is mapped into a cone defined by a pair of lines passing through the origin and separated by an angle

$$\theta = \tan^{-1} \frac{a+b}{a-b}$$

1.8 Notes and References

The basic elements of set theory are covered in many books introducing topology, measure theory, etc. In the last century Cantor showed that the treatment of infinite sets leads to nontrivial problems. With the formalization of various aspects of combinatorics it is now widely appreciated that there are decidedly nontrivial questions about finite sets as well. A standard reference is

1. F. Hausdorff, *Set Theory*, Chelsea, New York, 1962. (Translation of the original German edition)

Basic works on probability include:

1. A. N. Kolmogorov, *Foundations of Probability Theory*, Chelsea, New York, 1950. (Translation of the 1933 German edition)

2. W. Feller, *An Introduction to Probability Theory and Its Applications*, J. Wiley, New York, 1956.

Chapter 2

Poisson Counters and Differential Equations

A stochastic process is, from one point of view, just a random variable whose events are collections of functions of one or more independent variables. We are mostly interested in the cases where there is just one independent variable and we think of it as being time. It is not our intention to give a precise, general definition at this point. We prefer to discuss some specific cases instead. Indeed, we begin our discussion of a very flexible and interesting family of Markov processes based on a simple counting model of the type which describes a great many natural phenomena. We introduce stochastic differential equations, discuss statistical properties of their solutions, and derive characterizations of the corresponding probability laws.

2.1 Poisson Counters

By a *continuous time, countable state Markov process* we understand an ensemble of sample paths $\{x(\cdot)\}$ that are functions of time with the functions taking on values in a countable set X, subject to the condition that probability that $x(t+\tau) = x_i$ given the values of $x(\sigma)$ for all $\sigma \leq t$ is the same as the probability that $x(t+\tau) = x_i$ given only $x(t)$.

We let $\mathbb{Z}$ denote the integers $\{0, \pm 1, \pm 2, \dots\}$ and let $\mathbb{Z}^+$ denote the positive integers $\{1, 2, \dots\}$. Let x_t be a stochastic process which takes on values in the nonnegative integers $\mathbb{Z}^+ \cup \{0\}$, and let it be characterized by the transition rule.

$$\dot{p}_n(t) = -\lambda p_n(t) + \lambda p_{n-1}(t) \quad ; \quad p_0(0) = 1 \quad ; \quad p_i(0) = 0 \text{ for } i > 0$$

where λ is a constant called the *counting rate* and $p_i(t)$ is the probability that $x_t = i$. We call this the *Poisson counter* of rate λ. We can write these equations as an infinite system

$$\begin{bmatrix} \dot{p}_0 \\ \dot{p}_1 \\ \dot{p}_2 \\ \vdots \end{bmatrix} = \begin{bmatrix} -\lambda & 0 & 0 & \cdots \\ \lambda & -\lambda & 0 & \cdots \\ 0 & \lambda & -\lambda & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix} \begin{bmatrix} p_0 \\ p_1 \\ p_2 \\ \vdots \end{bmatrix} ; \quad p(0) = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \end{bmatrix}$$

and solve them one at a time, starting with the equation for p_0. This yields

$$\begin{aligned} p_0(t) &= e^{-\lambda t} \\ p_1(t) &= \lambda t e^{-\lambda t} \\ p_2(t) &= (\lambda^2 t^2/2!)e^{-\lambda t} \\ \cdots &= \cdots \\ p_n(t) &= (\lambda^n t^n/n!)e^{-\lambda t} \end{aligned}$$

Notice that the sum $p_0(t)+p_1(t)+p_2(t)+\cdots$ is one as it should be. If x is a real valued random variable we let $\mathcal{E}x$ denote its expected value. We justify our use of the term "counting rate" with the calculation

$$\begin{aligned} \mathcal{E}x_t &= \sum_{n=0}^{\infty} n p_n(t) = \sum_{n=0}^{\infty} (n(\lambda t)^n/n!)e^{-\lambda t} \\ &= e^{-\lambda t}(\lambda t e^{\lambda t}) \\ &= \lambda t \end{aligned}$$

Thus the expected value of x_t increases by one every $1/\lambda$ units of time. More generally, we can use this same method of computation to evaluate $\mathcal{E}x_t^p$ in terms of the lower order moments. That is,

$$\begin{aligned} \mathcal{E}x_t^p &= e^{-\lambda t}\sum_{n=0}^{\infty} n^p(\lambda t)^n/n! \\ &= e^{-\lambda t}\lambda t\sum_{n=0}^{\infty} (n)^p(\lambda t)^{n-1}/(n)! \\ &= e^{-\lambda t}\lambda t\sum_{n=0}^{\infty} (n)^{p-1}(\lambda t)^{n-1}/(n-1)! \\ &= e^{-\lambda t}\lambda t\sum_{r=0}^{\infty} (r+1)^{p-1}(\lambda t)^{r}/(r)! \\ &= e^{-\lambda t}\lambda t\sum_{r=0}^{\infty}\sum_{k=0}^{p-1} \binom{p-1}{k} r^k(\lambda t)^r/r! \\ &= \lambda t\sum_{r=0}^{p-1} \binom{p-1}{r}\mathcal{E}x_t^r \end{aligned}$$

Thus, for example, $\mathcal{E}x_t^3 = (\lambda t)(\mathcal{E}x_t^0 + 2\mathcal{E}x_t^1 + \mathcal{E}x_t^2)$

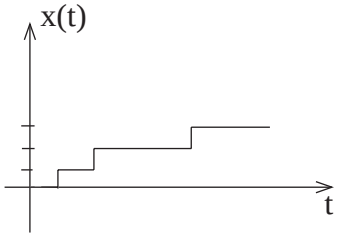

Figure 2.1. A sample path for a Poisson counter

We will also have occasion to use a *bidirectional Poisson counter* of rate λ. This is a process can be defined as the difference between two independent poisson counters of rate λ. It takes on

values in $\mathbb{Z}$ and starts at zero at $t = 0$. It's probability law is defined implicitly by the family of ordinary differential equations.

$$\dot{p}_i = +\lambda p_{i-1} - 2\lambda p_i + \lambda p_{i+1} \quad ; \quad i = 0, \pm 1, \pm 2, \ldots$$

subject to the initial conditions $p_0(0) = 1$ and $p_i(0) = 0$ for $i \neq 0$. One sees easily that the sum of the p's is always one. In order to develop the properties of the bidirectional counter it is useful to introduce the *generating function*

$$g(t, z) = \sum_{i=-\infty}^{\infty} z^i p_i(t)$$

Because the p_i are nonnegative and sum to one, this series converges uniformly for z on the unit circle, i.e., for $z = e^{i\theta}$ and θ real. Clearly $\partial g(t, z)/\partial t = \lambda(z - 2 + z^{-1})g(t, z)$ and so

$$g(t, z) = e^{-2\lambda t} e^{(z+z^{-1})\lambda t}$$

In terms of $z = e^{i\theta}$ we have

$$g(t, e^{i\theta}) = e^{(2cos\theta - 2)\lambda t}$$

Using the binomial formula to expand $e^{(z+z^{-1})\lambda t}$ we see that $g(t, z)$ can be expressed as

$$\begin{aligned} g(t, z) &= e^{-2\lambda t} \sum_{n=0}^{\infty} \sum_{k=0}^{n} \binom{n}{k} z^{n-2k} (\lambda t)^n / n! \\ &= e^{-2\lambda t} \sum_{n=0}^{\infty} \sum_{k=0}^{n} \frac{n!}{(n-k)!k!} z^{n-2k} (\lambda t)^n / n! \end{aligned}$$

or

$$p_n(t) = e^{-2\lambda t} \sum_{m=0}^{\infty} \frac{t^{2m}}{2^{2m} m!(n+m)!}$$

It is well known that this implies that p_n can be expressed in terms of the Bessel functions of the second kind and of integral order. (These are ordinary Bessel functions with purely imaginary argument.) In fact

$$I_n(t) = \sum_{m=0}^{\infty} \frac{t^{2m}}{2^{2m} m!(n+m)!}$$

is the series definition of the n^{th} Bessel function of the second kind. Thus

$$p_n(t) = e^{-2\lambda t} I_n(\lambda t)$$

It may also be shown that

$$\ddot{I}_n(t) + \frac{1}{t}\dot{I}_n(t) + \left(1 + \frac{n^2}{t^2}\right) I_n(t) = 0$$

with $I_0(0) = 1, \dot{I}_0(0) = -1$ and $I_n(0) = 0; n \neq 0$.

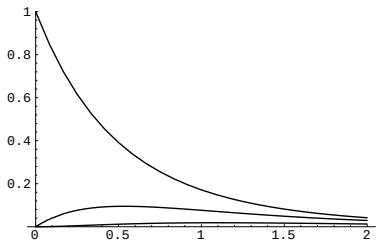

Figure 2.1. The plots of p_1, p_2 and p_3 for the bidirectional counter.

The utility of counters is greatly extended by combining them with some ideas from differential equations. Consider a differential equation in $\mathbb{R}^n$ written as

$$\dot{x}(t) = f(x(t), t)$$

or as

$$x(t) = x(0) + \int_0^t f(x(\sigma), \sigma)\, d\sigma$$

Suppose that $f(\cdot, \cdot)$ is continuous in both arguments and suppose that

$$||f(x, \sigma) - f(y, \sigma)|| \leq k||x - y||$$

for some k and all σ, x and y. It is known that such an equation has a unique solution corresponding to a given $x(0)$. We are interested in a stochastic version of this, i.e., stochastic differential equations. Consider

$$x(t) = x(0) + \int_0^t f\big(x(\sigma), \sigma\big)\, d\sigma + \int_0^t g\big(x(\sigma), \sigma\big)\, dN_\sigma \tag{2.1}$$

where N is a Poisson counter. How can we give meaning to such an object?

Definition: A function $x(\cdot)$ is a *solution of (2.1) in the Itô sense* if, on an interval where N is constant, x satisfies $\dot{x} = f(x, t)$ and if, N jumps at t_1, x behaves in a neighborhood of t according to the rule

$$\lim_{\substack{t \to t_1 \\ t > t_1}} x(t) = g(\lim_{\substack{t \to t_1 \\ t < t_1}} x(t), t_1) + \lim_{\substack{t \to t_1 \\ t < t_1}} x(t)$$

and $x(\cdot)$ is taken to be continuous from the left.

Notation: When this definition is in force it is common to rewrite equation (2.1) as

$$dx = f(x, t)dt + g(x, t)dN$$

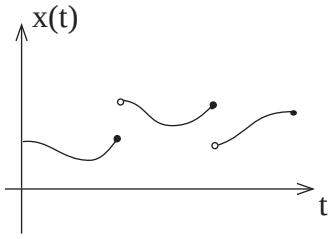

Figure 2.1. Showing a sample path of a solution.

Example: Consider

$$dx = xdt + xdN \;\; ; \;\; x(0) = 1$$

where N is a Poisson counter of rate λ. Then $x(t)$ will be of the form

$$x(t) = \begin{cases} e^t & 0 \le t \le t_1 \\ 2e^t & t_1 < t \le t_2 \\ 4e^t & t_2 < t \le t_3 \\ \cdots & \end{cases}$$

if $t_1, t_2, \ldots$ are the times at which the jumps of N occur.

Beware: An important part of the definition of the solution of an Itô equation is that the equation must be solved for dx. You cannot manipulate an Itô equation with dN's on the right-hand side before solving it. For example, one cannot replace the equation $dx = xdt + xdN$ by the equation

$$\frac{dx}{x} = dt + dN$$

and then integrate as one would do in the classical situation to get

$$x(t) = Me^{N(t)+t}$$

This is *not* the solution we defined above for the equation $dx = xdt + xdN$. **DO NOT FAIL TO GRASP THIS POINT**.

How can one simulate such a differential equation on a computer? Most implementations of the better known computer languages allow the program to call a random number generator that returns a "random" number that is, to the accuracy of the real numbers being used, uniformly distributed between zero and one. If r_i for $i = 0, 1, 2, \ldots$ are random variables that are independent and uniformly distributed on the interval [0,1] and if we chose a small constant h as a step size, we can simulate a Poisson counter via a difference equation approximation corresponding to step size h? Let

$$a((n+1)h) = a(nh) + m(n) \qquad ; \qquad n = 0, 1, 2, \ldots$$

with $m(n)$ being one if r_n is smaller than $h\lambda$ and zero otherwise. This means that in h units of time the probability of jump is $h\lambda$. A more accurate version would equate this probability to $1 - e^{-\lambda h}$; this could be achieved by equating $m(n)$ to one if r_n is smaller than $1 - e^{-\lambda h}$ and zero otherwise.

2.2 Finite-State, Continuous-Time Jump Processes

Consider a finite set $X = \{x_1, x_2, \ldots, x_n\}$ and consider a stochastic process x_t which takes on values in X. Let $p_i(t)$ be the probability that $x_t = x_i$ and suppose that

$$\dot{p}_i(t) = \sum_{j=1}^{n} a_{ij} p_j(t)$$

The matrix

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

is called the *intensity* matrix of the process. Conservation of probability, (i.e., the condition $\dot{p}_1 + \dot{p}_2 + \cdots + \dot{p}_n = 0$) requires that the entries in the columns of A sum to zero, i.e.,

$$\sum_{i=1}^{n} a_{ij} = 0 \quad ; \quad j = 1, 2, \ldots, n$$

Non-negativity of the p_i requires that we impose the condition

$$a_{ij} \geq 0 \quad ; \quad i \neq j$$

for otherwise we would get negative value for $p_i(t)$ if $p_j(0) = 1$ and a_{ij} were negative for some $i \neq j$. These two conditions and no others are required for A to be an intensity matrix. Of course these imply $a_{ii} \leq 0$. Because the columns sum to zero the set of all n by n intensity matrices is parameterized by $n(n-1)$ parameters which we may take to be the (a_{ij}) for $i \neq j$.

Example: Consider a continuous time jump process taking on values in the set $\{4, 2, -5\}$. Let p_1 be the probability that $x = 4$, p_2 be the probability that $x = 2$ and p_3 be the probability that $x = -5$. Let the probabilities evolve according to

$$\frac{d}{dt}\begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix} = \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ 1 & 0 & -1 \end{bmatrix}\begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix}$$

find the steady state probability distribution. Find the steady state mean of x. Find the steady state value of $\mathcal{E}(x - \bar{x})^2$. Make a quanitative statement about how fast p aproaches its steady state value.

Solution: If the steady state eigenvector has components $[a, b, c]$ then from the first component we see $a = b$ and from the second $b = c$ and so the steady state is $[1/3, 1/3, 1/3]$. Thus the expectation of x is 4/3+2/3-5/3=1/3. Similarly, $\mathcal{E}(x - \bar{x})^2 = 1/3(121/9 + 25/9 + 196/9) = 402/27 = 134/9$. The real parts of the two nonzero eigenvalues are -1.5 and the rate of approach is $e^{\lambda t}$.

We can give a sample path realization of such a process in terms of Poisson counters. Suppose we consider

$$dx = -2x dN \ ; \ x(0) = 1 \tag{2.2}$$

then x jumps first to -1, then back to +1, etc., so we have defined a finite state process. (If $x(0) = a$ then $x(t)$ takes on the values a and $-a$.) If we let p_1 be the probability that $x(t) = 1$ and let p_2 be the probability that $x(t) = -1$, then the corresponding description in terms of an intensity matrix is expressible in terms of the rate of the counter N as

$$\begin{bmatrix} \dot{p}_1 \\ \dot{p}_2 \end{bmatrix} = \begin{bmatrix} -\lambda & \lambda \\ \lambda & -\lambda \end{bmatrix}\begin{bmatrix} p_1 \\ p_2 \end{bmatrix} \tag{2.3}$$

To see this notice that the probability that there will be a jump in Δt seconds is $\lambda \Delta t +$ higher order terms in Δt. The following language will be used. We speak of (2.2) as being a *sample-path* description of the process and speak of (2.3) as being a *probabilistic description*.

Given a probabilistic description of an arbitrary finite state continuous-time jump process we can carry out an analogous construction to get a sample path description corresponding to it. One way to do this goes as follows. Code the states $\{x_1, x_2, \ldots, x_n\}$ of the finite-state, continuous-time jump process as distinct real numbers, say $z_1, z_2, \ldots, z_n$. Let ϕ_{ij} for $i, j = 1, 2, \ldots, n$ and $i \neq j$ be such that

$$\phi_{ij}(z) = \begin{cases} 0 & \text{if } z \neq z_j \\ z_i - z_j & \text{if } z = z_j \end{cases}$$

Let N_{ij} be a Poisson counter with rate λ_{ij}. Consider

$$dz = \sum_{i \neq j=1}^{n} \phi_{ij}(z) dN_{ij} \quad ; \quad z(0) \in \{z_1, z_2, \ldots, z_n\}$$

(a) If this process starts in the set $\{z_1, z_2, \ldots, z_n\}$ it remains in the set $\{z_1, z_2, \ldots, z_n\}$
(b) If $p_i(t)$ is the probability that $x(t) = z_i$, then

$$\dot{p}_i = \sum_{j=1}^{n} \lambda_{ij} p_j - \left(\sum_{j=1}^{n} \lambda_{ij} \right) p_i$$

(Notice there is no dependence on λ_{ii} because the terms cancel.) That is, if we wish to obtain a realization of $\dot{p} = Ap$ we may do so by choosing the rates $\lambda_{ij} = a_{ij}$ for $i \neq j$. This gives a sample path description of the continuous-time jump process. In this sense we may think of Poisson counters as having a certain "universal" property insofar as continuous time jump processes are concerned.

Example: Find functions $\phi_1(x), \phi_2(x)$ and $\phi_3(x)$ and counting rates $\lambda_1, \lambda_2, \lambda_3$ such that if $x(0) \in \{3, 7, 9\}$ then for

$$dx(t) = \phi_1\big(x(t)\big) dN_1 + \phi_2\big(x(t)\big) dN_2 + \phi_3\big(x(t)\big) dN_3$$

it happens that $x(t)$ belongs to {3,7,9} for all future time and

$$\begin{bmatrix} \dot{p}_1(t) \\ \dot{p}_2(t) \\ \dot{p}_3(t) \end{bmatrix} = \begin{bmatrix} -3 & 0 & 8 \\ 3 & -2 & 0 \\ 0 & 2 & -8 \end{bmatrix} \begin{bmatrix} p_1(t) \\ p_2(t) \\ p_3(t) \end{bmatrix}$$

with $p_1(t)$ being the probability that $x(t)$ is 3, $p_2(t)$ being the probability that $x(t) = 7$ and $p_3(t)$ being the probability that $x(t)$ is 9.

To get the appropriate ϕ's we see that

$$\phi_1(x) = (x - 9)(x - 7)/6$$

takes on the value 4 at $x = 3$ and is zero if $x = 7$ or $x = 9$. Thus it can be used with a counter of rate 3 to generate the transaction from 3 to 7. Likewise, we use

$$\phi_2(x) = (x - 3)(x - 9)/4$$

and

$$\phi_3(x) = -(x - 3)(x - 7)/2$$

taking the rates of N_1, N_2 and N_3 to be 3,2 and 8 completes the specification.

2.3 The Itô Rule for Jump Processes

If we have a stochastic differential equation

$$dx = f(x)dt + \sum_{i=1}^{m} g_i(x)dN_i \;\; ; \;\; x \in \mathbb{R}^n$$

and if $\psi : \mathbb{R}^n \to \mathbb{R}$ is a given function, then of course $\psi\big(x(t)\big)$ is a stochastic process and using the given definition we see that

$$d\psi(t,x) = \frac{\partial \psi}{\partial t}dt + \left\langle \frac{\partial \psi}{\partial x}\, , f(x) \right\rangle dt + \sum_{i=1}^{n} [\psi\big(t, x + g_i(x)\big) - \psi(t,x)]dN_i$$

This is the so-called *Itô rule* for jump processes. In writing it down we have used the fact that the probability that two counters will jump at the same time is zero and therefore such a possibility can be ignored. Notice that in this context the Itô rule is a completely trivial consequence of the definition of the solution of the differential equation.

Example: Given that

$$dx(t) = -x(t)dt + dN_1(t) - dN_2(t)$$

where N_1 and N_2 are Poisson counters of rates λ_1 and λ_2, respectively, find an equation for x^2.

Solution:

$$\begin{aligned} dx^2(t) &= -2x^2(t)dt + ((x(t)+1)^2 - x^2)dN_1 + ((x(t)-1)^2 - x^2(t))dN_2 \\ &= -2x^2(t)dt + (2x(t)+1)dN_1 + (1 - 2x(t))dN_2 \end{aligned}$$

2.4 Computing Expectations

If x satisfies the differential equation

$$dx = f(x,t)dt + \sum_{i=1}^{m} g_i(x,t)dN_i$$

how can we get an equation for the expectation of x? The key observations here are that

(a) $(\mathcal{E}N(t) - \lambda t) = 0$ (This was established in section 2.1) and

(b) the probability that $N(t)$ will jump in the interval $[t, t+\Delta]$ is independent of the value of $x(t)$.

Thus

$$\mathcal{E}\big(x(t+\Delta) - x(t)\big) = \mathcal{E}\int_t^{t+\Delta} f\big(x(\sigma),\sigma\big)\, d\sigma + \sum_{i=1}^{m} \mathcal{E}\int_t^{t+\Delta} g_i\big(x(\sigma),\sigma\big)dN_i(\sigma)$$

Expanding in a Taylor series and taking the limit as Δ goes to zero we get

$$\frac{d}{dt}\mathcal{E}x(t) = \mathcal{E}f\big(x(t),t\big) + \sum_{i=1}^{m}\Big(\mathcal{E}g_i\big(x(t),t\big)\Big)\lambda_i$$

Thus the general rule for computing expectations is to replace dN_i by $\lambda_i dt$, divide by dt and take expectations.

Example 1: Given

$$dx = -x(t)dt + dN_1 - dN_2$$

the above rule implies that

$$\frac{d}{dt}\mathcal{E}x = -\mathcal{E}x + \lambda_1 - \lambda_2$$

Using the Itô rule for x^2 and the calculating the expectation we get

$$\frac{d}{dt}\mathcal{E}x^2 = -2\mathcal{E}x^2 + \mathcal{E}(2x+1)\lambda_1 + \mathcal{E}(1-2x)\lambda_2$$

etc.

Example 2: Let n be a Poisson counter of rate λ and let x and z satisfy $z(t) \in \{-1, 1\}$

$$dx = -xdt + zdt$$

$$dz = -2zdN$$

We wish to find an equation for the variance of x. We begin by using the differentiation rule to write

$$dx^2 = -2x^2dt + 2xzdt$$

$$d(xz) = (-xz + z^2)dt - 2xzdN$$

Taking expectations we get

$$\frac{d}{dt}\mathcal{E}x^2 = -2\mathcal{E}x^2 + 2\mathcal{E}xz$$

$$\frac{d}{dt}\mathcal{E}xz = -(1+2\lambda)\mathcal{E}xz + 1$$

These linear equations can then be solved by standard means. Notice that steady-state value is

$$\mathcal{E}x^2 = 1/(1+2\lambda)$$

2.5 The Equation for the Density

In order to gain a more complete understanding of the class of problems introduced in 1.3 we will develop an equation for the probability law associated with the state. In general this is complicated because the solution of stochastic differential equations involving Poisson counters need not visit all points in the space. However, when it does, and does so often enough, there is a nice theory.

Let $\mathcal{A}$ denote the set of all subsets of $\mathbb{R}^n$ which are parallelpipeds; i.e., sets that are the intersections of n slices of the form $\{x | a_i \le x \le b_i\}$. Given an equation

$$dx = f\big(x(t)\big)dt + \sum_{i=1}^{m} g_i\big(x(t)\big)dN_i$$

suppose that there exists a differentiable function ρ such that for all sets $A \subset \mathcal{A}$

$$p(x(t) \in A) = \int_A \rho(t,x)dx$$

In this case it is of interest to find an equation for the evolution of ρ. The device to be used here will be used several times in these notes. First of all introduce a smooth "test function" ψ, a map of $\mathbb{R}^n$ into $\mathbb{R}$. Use the Itô rule to get

$$d\psi = \frac{\partial \psi}{\partial x} f dt + \sum_{i=1}^{m} \big(\psi\big(x + g_i(x)\big) - \psi(x)\big)\, dN_i$$

Next we compute the expected value of ψ using the results of section 1.4. This gives

$$\frac{d}{dt}\mathcal{E}\psi\big(x(t)\big) = \mathcal{E}\frac{\partial \psi}{\partial x} f + \mathcal{E}\sum_{i=1}^{m} \big(\psi\big(x(t) + g_i(x(t))\big) - \psi\big(x(t)\big)\big)\,\lambda_i$$

Now notice that if ρ exists then we can also compute the expectation of ψ by integrating against ρ. That is, $\mathcal{E}\psi = \int \psi(x)\rho(x)dx$. Differentiating this expression with respect to time gives

$$\frac{\partial}{\partial t}\int \psi(x)\rho(t,x)dx = \int \left[\frac{\partial \psi(x)}{\partial x} f(x) + \sum \big(\psi\big(x + g_i(x)\big) - \psi(x))\lambda_i\big)\right] \rho(t,x)dx$$

If ρ is smooth we may use integration-by-parts to get

$$\int \psi(x)\frac{\partial \rho(t,x)}{\partial t}dx = \int -\psi\frac{\partial}{\partial x}(f\rho) - \sum \lambda_i \psi\rho dx + \sum \int \lambda_i \psi(x + g_i(x))\rho(t,x)dx$$

provided $\psi(x) = 0$ for $|x|$ sufficiently large.

In order to go further it is necessary to make some assumptions about the function defined by

$$\tilde{g}_i(x) = x + g_i(x)$$

If, for example, x takes on values in $\mathbb{R}^1$ and $\tilde{g}$ defines a map of $\mathbb{R}^1$ onto $\mathbb{R}^1$ which is one to one, then letting

$$dz = |\det(I + \frac{\partial g_i}{\partial x})|\, dx$$

allows us to change variables in the last integral in the following way.

$$\int \psi(x + g_i(x))\rho(t,x)dx = \int \psi(z)\rho\big(t, \tilde{g}_i^{-1}(z)\big)\left|\det\left(I + \frac{\partial g_i}{\partial x}\right)\right|^{-1} dz$$

In this case we can argue that because ψ is arbitrary this integral equation can be replaced by the differential difference equation.

$$\frac{\partial \rho(t,x)}{\partial t} = -\frac{\partial}{\partial x}\left[f(x)\rho(t,x)\right] + \sum_{i=1}^{m} \lambda_i\Big(\rho\big(t, \tilde{g}_i^{-1}(x)\big)\Big|\Big(\det\big(I + \frac{\partial g_i}{\partial x}\big)\Big)\Big|^{-1}_{x+g_i(x)} - \rho(t,x)\Big)$$

This is, then, an evolution equation for the density, *provided* a smooth density exists. If $\tilde{g}$ is not one to one but yet has a finite number of solutions for each point in the range, this argument can be modified to yield a density equation.

Example 1: Consider the linear equation

$$dx = -xdt + dN_1 - dN_2$$

with N_1 and N_2 being standard Poisson counters of rate λ. It can be thought of as a stabilized version of the bidirectional counter. If a smooth density exists it satisfies

$$\frac{\partial \rho}{\partial t} = \frac{\partial}{\partial x} x\rho + \lambda\rho(t, x-1) - 2\lambda\rho(t,x) + \lambda\rho(t, x+1)$$

This is an example of a functional differential equation. Explicit solutions of this type of equation are virtually nonexistent in spite of their simple appearance. On the other hand, qualitative information about the solution can often be obtained. In this case there is a steady state solution satisfying

$$0 = \frac{\partial}{\partial x} x\rho + \lambda\rho(t, x-1) - 2\lambda\rho(t,x) + \lambda\rho(t, x+1)$$

It can be solved using Fourier transforms.

Example 2: Consider a model for a *queueing problem* in which customers arrive for service and service is provided as long as there is a customer to be served. Assume that the arrival of customers is modeled by a Poisson process of rate λ and that each customer requires μ amount of time to be served. (We could also let μ be random at the expense of some extra work.) Let $f_p(x)$ denote the function whose value is 1 if x is positive and zero if x is non positive. The variable x in the differential equation

$$dx = -\frac{1}{\mu} f_p(x)dt + dN$$

then describes the total effort that the server must provide to meet the demands of those presently in the queue, including the customer currently being served. The probability density satisfies

$$\frac{\partial \rho(t,x)}{\partial t} = \frac{\partial f_p(x)\rho(t,x)}{\mu \partial x} + \lambda\rho(t, x-1) - \lambda\rho(t,x)$$

We expect that there will be a steady state condition if the arrival rate is smaller than the service rate, i.e. if $\lambda/\mu < 1$. Of course there is a nonzero probability that the queue will be empty. Taking expectations of both sides of the sample path equation we see that in steady state $\mathcal{E} f_p(x) = \lambda\mu$. Thus in steady state the probability that the queue is empty is just $1 - \lambda\mu$. There is a steady state solution of the density equation having the form of a delta function plus a density

$$\rho_{ss}(x) = (1 - \lambda\mu)\delta(x) + \psi(x)$$

We can solve for ψ only by dividing the positive half-line in to unit length segments. In the segment $(0, 1]$ the $\rho(x-1)$ term is zero and thus we need only solve the homogenous equation

$$\frac{1}{\mu}\frac{d\rho}{dx} - \lambda\rho = 0$$

and so $\rho(x) = \beta^{\lambda\mu x}$. We observe that the drift term adds probability to the set $\{x|x = 0\}$ at the rate $\rho(0^+)/\mu = \beta/\mu$ and the counter removes probability at the rate $(1 - \lambda\mu)\lambda$. These must be equal in equilubrium and so $\beta = (1 - \lambda\mu)\lambda\mu$. On the interval $(1, 2]$ we have

$$\frac{1}{\mu}\frac{d\rho}{dx} - \lambda\rho = (1 - \lambda\mu)\lambda\mu e^{\lambda\mu(x-1)}$$

This is a linear differential equation with a confluence between the forcing term and the homogeneous response so the solution on $(1, 2]$ takes the form

$$\rho(x) = (ax + b)e^{\lambda\mu x}$$

With a and b being determined by the boundry condition at $x = 1$ and the differential equation itself. Continuing on, for the interval $(2, 3]$ we have the differential equation

$$\frac{1}{\mu}\frac{d\rho}{dx} - \lambda\rho = (a(x-1) + b)e^{\lambda\mu(x-1)}$$

which has a double confluence and hence a solution of the form of an exponential times a cubic polynomial, etc. It is of interest to note that while the exponential involved grows with x, the polynomials are such that ρ decays to zero.

Example 3. Consider the coupled equations of example 2 of the previous section. In this case it is convenient to write the equation for the density as a pair of coupled equations. Let $\rho_+(t,x)$ be the probability density associated with x, given that $z=+1$ and let $\rho_-(t,x)$ be the probability density associated with x given that $z=-1$. In terms of this notation we have

$$\frac{\partial \rho_+(t,x)}{\partial t} = \frac{\partial}{\partial x}(x-1)\rho_+(t,x) + \lambda(\rho_-(t,x) - \rho_+(t,x))$$
$$\frac{\partial \rho_-(t,x)}{\partial t} = \frac{\partial}{\partial x}(x+1)\rho_-(t,x) + \lambda(\rho_+(t,x) - \rho_-(t,x))$$

2.6 The Backwards Evolution Equation

Although there are many problems in which the forward evolution of the density, as described in the previous section, is the most natural evolution to focus on, there are other problems for which an evolution in a negative direction is important. For example, if it is known that $x(1)=x_1$ we may wish to know the probability density at $t=0$. More generally, given $\rho(1,x)$ what governs the evolution of ρ backwards in time? Because the definition of what one means by a solution of the differential equation does not treat t and $-t$ symmetrically, it is not to be expected that one can get the evolution of ρ backwards in time simply by changing the sign of the right-hand side of the equation.

To gain some intuition about this question it is useful to consider a finite state discrete-time process whose transition probabilities are such that

$$p(i+1) = Ap(i)$$

Given that $x(1)=x_k$, we see that the probability vector at time 0 is determined by the k^{th} row of A. In fact, if the entries in the k^{th} row of A sum to s_k then

$$p(0) = \frac{1}{s_k}A^k \;\; ; \;\; A^k = k^{th} \text{ row of } A$$

We can, therefore, express the backwards equation as

$$p(i-1) = A^T D p(i)$$

with T denoting the transpose of A and D being a diagonal matrix that normalizes the columns of A^T so that the sum of their entries is one.

The corresponding equation situation in continuous time relates the forward equation $\dot{p} = Ap$ to the backwards equation

$$\dot{p} = (A^T + D)p$$

with D being a diagonal matrix chosen so as to make the sum of the entries in a column of $A^T + D$ equal to zero.

Example: Consider a continuous-time jump process whose probability law satisfies

$$\begin{bmatrix} \dot{p}_1 \\ \dot{p}_2 \\ \dot{p}_3 \end{bmatrix} = \begin{bmatrix} -1 & 1 & 2 \\ 1 & -1 & 1 \\ 0 & 0 & -3 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix}$$

Given that $p_1(1)=1, p_2(1)=0, p_3(1)=0$, find $p(0)$. To get the backward propagation equation we observe that

$$A = \begin{bmatrix} -1 & 1 & 2 \\ 1 & -1 & 1 \\ 0 & 0 & -3 \end{bmatrix} \;\; , \;\; A^T = \begin{bmatrix} -1 & 1 & 0 \\ 1 & -1 & 0 \\ 2 & 1 & -3 \end{bmatrix}$$

$$\begin{bmatrix} \dot{p}_1 \\ \dot{p}_2 \\ \dot{p}_3 \end{bmatrix} = (A^T + D)p = \begin{bmatrix} -3 & 1 & 0 \\ 1 & -2 & 0 \\ 2 & 1 & 0 \end{bmatrix} p = Bp$$

$$D = \begin{bmatrix} -2 & & \\ & -1 & \\ & & 3 \end{bmatrix}$$

$$p(0) = e^B p(1)$$

Example: Consider a continuous-time jump process whose probability law satisfies

$$\begin{bmatrix} \dot{p}_1 \\ \dot{p}_2 \\ \dot{p}_3 \end{bmatrix} = \begin{bmatrix} -1 & 1 & 1 \\ 1 & -1 & 1 \\ 0 & 0 & -2 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix}$$

Given that $p_1(1) = 1, p_2(1) = 0, p_3(1) = 0$, find $p(0)$. (You will need to solve a system of three linear equations for a particular initial condition; the eigenvalues are 0, -1, -3.)

Solution: The relevant backwards time matrix, $B = A^T - D$ is

$$B = \begin{bmatrix} -2 & 1 & 0 \\ 1 & -2 & 0 \\ 1 & 1 & 0 \end{bmatrix}$$

The eigenvalues of B are $\lambda = 0, -1, -3$. Some work shows that

$$e^{Bt} = \begin{bmatrix} e^{-2t}\cosh t & e^{-2t}\sinh t & 0 \\ e^{-2t}\sinh t & e^{-2t}\cosh t & 0 \\ 1 - e^{-t} & b(t) & 1 \end{bmatrix}$$

The value of the probability at $t = 0$ is simply thie first column of this matrix evaluated at t=1.

2.7 Computing Temporal Correlations

Let x_t be a finite-state, continuous-time jump process which takes on values in the set $\{x_1, x_2, \ldots, x_n\} \subset \mathbb{R}$. To compute the expected value of $\mathcal{E}(x_t x_{t+\tau})$ we let $\psi_{ij}(t, \tau)$ denote the probability that $x(t) = x_i$ and $x(t + \tau) = x_j$ and use the formula

$$\mathcal{E} x_t x_{t+\tau} = \sum_{i,j} \ \psi_{ij}(t, \tau) \ x_i x_j$$

Now the probability that $x_t = x_i$ is just $p_i(t)$. The probability that $x_{t+\tau} = x_j$ given that $x_t = x_i$ is just the ji^{th} element of $e^{A\tau}$, provided $\tau \geq 0$. Thus we see that for $\tau \geq 0$.

$$\mathcal{E}(x_t x_{t+\tau}) = \sum_{i,j} \ p_i(t)\phi_{ji}(\tau) x_i x_j$$

where $\phi_{ij}(\tau)$ is the ij^{th} entry in $e^{A\tau}$. Of course if we are given $p_i(0)$, then this is

$$\mathcal{E}x_t x_{t+\tau} = \sum_{i,j,k} \phi_{ik}(t) p_k(0) \phi_{ji}(\tau) x_i x_j$$

If we want to write this in matrix notation, it is

$$\mathcal{E}x(t)x(t+\tau) = \quad [x_1(t), x_2(t), \ldots, x_n(t)] e^{A\tau} \quad \begin{bmatrix} p_1(t) & 0 & 0 \\ 0 & p_2(t) & 0 \\ 0 & 0 & \ddots \end{bmatrix} \quad \begin{bmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_n(t) \end{bmatrix}$$

An application of the same methodology yields formulas for more complicated statistical properties such as

$$\phi(t_1, t_2, \cdots, t_n) = \mathcal{E}x_{t_1} x_{t_2} \ldots x_{t_n}$$

There is a second approach which is sometimes more convenient if we have a sample path description. Consider

$$dx_t = f(x_t)dt + \sum g_i(x_t) dN_{it}$$

and

$$\frac{d}{dt}\mathcal{E}x_t x_\tau = \mathcal{E}x_t f(x_\tau) + \sum \mathcal{E}x_t g_i(x_\tau)\lambda_i$$

Now suppose that f is linear and g is constant. Then

$$\frac{d}{dt}\mathcal{E}x_t x_\tau = \mathcal{E}\alpha x_t x_\tau + \mathcal{E}x_t \beta_i \lambda_i$$

and we can obtain $\mathcal{E}x_t x_\tau$ from $\mathcal{E}x_t^2$ and $\mathcal{E}x_t$. Again the sample path approach is very effective.

Example: Consider the stochastic process y defined by

$$y(t) = \begin{cases} \alpha & \text{if } x(t) = 1 \\ \beta & \text{if } x(t) = -1 \end{cases}$$

where the transition probabilities for the x process are such that

$$\begin{bmatrix} \dot{p}_1 \\ \dot{p}_{-1} \end{bmatrix} = \begin{bmatrix} -a & b \\ a & -b \end{bmatrix} \begin{bmatrix} p_1 \\ p_{-1} \end{bmatrix}$$

(p_1 = probability $x = 1$; p_{-1} = probability x = -1).

(a) Find conditions such that

$$\lim_{t\to\infty} \mathcal{E}y(t) = 0$$

(b) Find conditions such that (a) holds and

$$\lim_{t\to\infty} \mathcal{E}y(t)y(t+\tau) = e^{-|\tau|}$$

Begin by finding a sample-path description for $y(t)$; i.e., by finding appropriate ϕ_{ij} as in section 2.2. Consider the specification

$$\phi_{\alpha\beta}(y) = \begin{cases} 0 & \text{if } y \neq \beta \\ \alpha - \beta & \text{if } y = \beta \end{cases}$$

for which $\phi_{\alpha\beta}(y) = \alpha - y$ is suitable. By the same token, we may choose $\phi_{\beta\alpha}(y) = \beta - y$. Now consider the process whose sample-path description is

$$dy = (\alpha - y)dN_{\alpha\beta} + (\beta - y)dN_{\beta\alpha},$$

with counters of rates $\lambda_{\alpha\beta}, \lambda_{\beta\alpha}$ respectively; if $y = \alpha$ this reduces to $dy = (\beta - \alpha)dN_{\beta\alpha}$, so that $\dot{p}_\beta = \lambda_{\beta\alpha}$ and hence $\dot{p}_\alpha = -\lambda_{\beta\alpha}$. Repeating this argument assuming $y = \beta$ and combining, we get

$$\begin{aligned} \dot{p}_\alpha &= -\lambda_{\beta\alpha}p_\alpha + \lambda_{\alpha\beta}p_\beta \\ \dot{p}_\beta &= \lambda_{\beta\alpha}p_\alpha - \lambda_{\alpha\beta}p_\beta. \end{aligned}$$

Now if we set $\lambda_{\alpha\beta} = b$ and $\lambda_{\beta\alpha} = a$ we recover the probabilistic description given in the problem and hence a sample- path description of $y(t)$:

$$dy = (\alpha - y)dN_b + (\beta - y)dN_a,$$

where the subscripts denote the counting rates.

a) Using the sample-path description we see immediately that

$$\frac{d}{dt}\mathcal{E}y(t) = a\beta + b\alpha - (a+b)\mathcal{E}y(t),$$

so that $\lim_{t\to\infty}\mathcal{E}y(t) = \frac{a\beta+b\alpha}{a+b}$. Thus the required condition is $a\beta + b\alpha = 0$.

b) To find the correlation $\mathcal{E}y(t)y(t+\tau)$ we first need to find $\mathcal{E}y^2(t)$. The Itô formula gives

$$\begin{aligned} dy^2(t) &= \Big(\big(y + (\alpha - y)\big)^2 - y^2\Big)dN_b + \Big(\big(y + (\beta - y)\big)^2 - y^2\Big)dN_a \\ \tfrac{d}{dt}\mathcal{E}[y^2(t)] &= b(\alpha^2 - \mathcal{E}[y^2(t)]) + a(\beta^2 - \mathcal{E}[y^2(t)]) \\ &= a\beta^2 + b\alpha^2 - (a+b)\mathcal{E}[y^2(t)] \end{aligned}$$

thus

$$\mathcal{E}[y^2(t)] = \tfrac{a\beta^2+b\alpha^2}{a+b} + ke^{-(a+b)t}$$

To compute the correlation we rewrite the original Itô equation using τ as the independent variable:

$$\begin{aligned} d_\tau y(\tau) &= \big(\alpha - y(\tau)\big)dN_b(\tau) + \big(\beta - y(\tau)\big)dN_a(\tau) \\ d_\tau y(t)y(\tau) &= \big(\alpha - y(\tau)\big)y(t)dN_b(\tau) + \big(\beta - y(\tau)\big)y(t)dN_a(\tau). \end{aligned}$$

Assuming $\tau > t$ we take the expectation, e.g., $\mathcal{E}[y(t)dN_a(\tau)] = a\mathcal{E}[y(t)]$

$$\tfrac{d}{d\tau}\mathcal{E}[y(t)y(\tau)] = -(a+b)\mathcal{E}[y(t)y(\tau)]$$

or

$$\mathcal{E}[y(t)y(\tau)] = \mathcal{E}[y^2(t)]e^{-(a+b)(\tau-t)}$$

for $\tau \geq t$. Using symmetry arguments, changing from τ to $t+\tau$ and substituting for the variance,

$$\mathcal{E}[y(t)y(t+\tau)] = \frac{a\beta^2 + b\alpha^2}{a+b}e^{-(a+b)|\tau|} + ke^{-(a+b)(t+|\tau|)}.$$

In the limit as $t \to \infty$ the second term vanishes; in order to have the first term be the desired quantity we must have $a+b=1$ and $a\beta^2 + b\alpha^2 = 1$. We can make this a little simpler by noting that $a\beta^2 + b\alpha^2 = (a\beta + b\alpha)(\alpha+\beta) - (a+b)\alpha\beta = 1$.

2.8 Linear Systems with Jump Process Coefficients

Recall one fact from the Lagrange interpolation. We may find a polynomial function $\phi : \mathbb{R} \to \mathbb{R}$ such that $\phi(z_i) = a_i$ for a given set of pairs $(z_1, a_1), (z_2, a_2) \ldots (z_n, a_n)$ (no z's repeated) and ϕ may be taken to be of degree n. In fact

$$\phi(z) = a_1\frac{(z-z_2)(z-z_3)\ldots(z-z_n)}{(z_1-z_2)(z_1-z_3)\ldots(z_1-z_n)} + \cdots + a_n\frac{(z-z_1)(z-z_2)\ldots(z-z_{n-1})}{(z_n-z_1)(z_n-z_2)\ldots(z_n-z_{n-1})}$$

Thus if we wish to find a sample path description of a finite state continuous time jump process we may take it to be of the form

$$dz = \sum_{i,j}\phi_{ij}(z)dN_{ij}$$

with ϕ_{ij} being polynomials of degree n where n is the number of states.

Now consider the linear stochastic differential equation

$$\dot{x} = A(z)x + b(z)$$

where z is a finite state continuous time jump process. We can use the elementary device of appending 1 to x to rewrite this as

$$\frac{d}{dt}\begin{bmatrix} 1 \\ x \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ b(z) & A(z) \end{bmatrix}\begin{bmatrix} 1 \\ x \end{bmatrix}$$

thus bringing it to a form in which there is no inhomogeneous term

$$\frac{d}{dt}\tilde{x} = \tilde{A}(z)\tilde{x}$$

(The reason for wanting to do this will appear shortly.)

If z is a FSCT jump process, then model it as above with the ϕ_{ij} polynomial. Then $A(z)$ can be expressed as

$$A(z) = A_0 + zA_1 + \dots z^n A_n$$

In fact, the Lagrange formula does the job for us again

$$A(z) = A(z_1) \cdot \frac{(z-z_2)(z-z_3)\dots(z-z_n)}{(z_1-z_2)(z_1-z_3)\dots(z_1-z_n)} + \dots + A(z_n)\frac{(z-z_1)(z-z_2)\dots(z-z_{n-1})}{(z_n-z_1)(z_n-z_2)\dots(z_n-z_{n-1})}$$

Thus we can write, dropping the $\tilde{}$ over x,

$$\begin{aligned} dz &= \sum \phi_{ij}(z) dN_{ij} \\ \dot{x} &= (A_0 + zA_1 + \dots + z^{n-1}A_{n-1})x \end{aligned}$$

We now take the major step which allows us to put these into a form which permits us to compute the statistical properties of the solutions. It goes like this. *Consider, along with the given x equation, an equation for zx, z^2x, ... , $z^{n-1}x$.* Since z takes on the values of $z_1, z_2, \dots, z_n$ and no others, we see that

$$p(z) = (z-z_1)(z-z_2)\dots(z-z_n) = 0$$

and thus z^n can be expressed as $z^n = p_0 + p_1 z + \dots + p_{n-1}z^{n-1}$ for some choice of p_i. Using the Itô rule we then obtain (with A's above)

$$\begin{aligned} d\begin{bmatrix} x \\ zx \\ \vdots \\ z^{n-1}x \end{bmatrix} &= \begin{bmatrix} A_0 & A_1 & \cdots & A_{n-1} \\ \tilde{A}_{n-1} & \tilde{A}_0 & \cdots & \tilde{A}_{n-2} \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix} \begin{bmatrix} x \\ zx \\ \vdots \\ z^{n-1}x \end{bmatrix} dt \\ &\quad + \sum \begin{bmatrix} 0 & 0 & & 0 \\ \phi_{ij}^{21} I & \phi_{ij}^{22} I & \dots & \phi_{ij}^{2n} I \\ \cdots & \cdots & \cdots & \cdots \\ \phi_{ij}^{n1} I & \phi_{ij}^{n2} I & \dots & \phi_{ij}^{nn} I \end{bmatrix} \begin{bmatrix} x \\ zx \\ \vdots \\ z^{n-1}x \end{bmatrix} dN_{ij} \end{aligned}$$

where ϕ_{ij}^{kl} are such that

$$dz^{k-1} = \sum_{l,j,i} z^{l-1} \phi_{ij}^{kl} dN_{ij}$$

What we have accomplished with this device is the reduction of the original problem to one of the form (new notation)

$$dx = Axdt + \sum B_i x dN_i$$

The advantage of this is that we can compute statistical properties of this equation using the calculus developed earlier.

Example: Consider the linear equation with an additive jump process term

$$dx = Axdt + bz\,dt$$

where

$$dz = -2z dN \quad ; \quad z(0) \in \{-1, 1\}$$

This set of equations is linear as it stands.

$$\begin{aligned} d\begin{bmatrix} x \\ z \end{bmatrix} &= \begin{bmatrix} A & b \\ 0 & 0 \end{bmatrix}\begin{bmatrix} x \\ z \end{bmatrix}dt + \begin{bmatrix} 0 & 0 \\ 0 & -2 \end{bmatrix}\begin{bmatrix} x \\ z \end{bmatrix}dN \\ \frac{d}{dt}\mathcal{E}\begin{bmatrix} x(t) \\ z(t) \end{bmatrix} &= \begin{bmatrix} A & b \\ 0 & -2\lambda \end{bmatrix}\mathcal{E}\begin{bmatrix} x \\ z \end{bmatrix} \\ \frac{d}{dt}\mathcal{E}x(t) &= A\mathcal{E}x + b\mathcal{E}z \\ \frac{d}{dt}\mathcal{E}z(t) &= -2\lambda\mathcal{E}z(t) \end{aligned}$$

To compute the variance notice that x is continuous so that the Itô rule in this case simplifies to

$$\begin{aligned} dxx^T &= (Axx^T + xx^TA^T)dt + (bx^T + xb^T)zdt \\ dzx &= zdx + xdz \\ &= (Axz + bz^2)dt - 2zxdN \\ dz^2 &= 0 \end{aligned}$$

Solving the second of these we get (note $\mathcal{E}z^2 = 1$)

$$\frac{d}{dt}\mathcal{E}zx = A(\mathcal{E}zx) + b - 2\mathcal{E}\lambda zx$$

so

$$\begin{aligned} \mathcal{E}zx &= e^{(A-2\lambda I)t}a \\ \frac{d}{dt}\mathcal{E}(xx^T) &= A(\mathcal{E}xx^T) + (\mathcal{E}xx^T)A^T + ba^Te^{(A-2\lambda I)^t} + e^{(A-2\lambda I)^Tt}ab^T \end{aligned}$$

2.9 Exercises 2

1. Suppose that x satisfies the linear differential equation

   $$\dot{x} = Ax + bu \quad ; \quad x(0) = 0$$

   with u being a finite state continuous time jump process taking on the values $1, 0$, and -1 with transition probabilities

   $$\frac{d}{dt}\begin{bmatrix} p_{+1} \\ p_0 \\ p_{-1} \end{bmatrix} = \begin{bmatrix} -1 & 1 & 0 \\ 1 & -2 & 1 \\ 0 & 1 & -1 \end{bmatrix}\begin{bmatrix} p_{+1} \\ p_0 \\ p_{-1} \end{bmatrix}$$

   Find a differential equation for the expected value of xx^T.

2. Consider the stochastic differential equation

   $$dx = -\tfrac{1}{2}x(1-x)dN_1 - \tfrac{1}{2}x(x+1)dN_2 + (x^2-1)dN_3 - (x^2-1)dN_4$$

   Suppose that $x(0) \in \{1, 0, -1\}$. Suppose that the N_i are Poisson counters with

   $$\mathcal{E}N_i = \lambda_i t$$

   If $p_1(t)$ is the probability that $x(t) = 1, p_2(t)$ is the probability that $x(t) = 0$ and $p_3(t)$ is the probability that $x(t) = -1$, find A such that

   $$\frac{d}{dt}\begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{21} & a_{31} \\ a_{12} & a_{22} & a_{32} \\ a_{13} & a_{23} & a_{33} \end{bmatrix}\begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix}$$

3. In a certain mixture of radioactive isotopes there are two main reactants which decay according to the stochastic equations

$$dx = -100xdN_1$$
$$dy = xdN_1 - ydN_2$$

where N_1 and N_2 are Poisson counters of rate 1. If $x(0) = 10$ and $y(0) = 2$ how long will it be before we can assert that for all future time

$$Ey(t) \leq 1$$

4. Consider the Itô equation

$$dx = adN_1 - bdN_2$$

with N_1 and N_2 being Poisson counting process of rate λ and μ, respectively.

   (a) Compute $\mathcal{E}x(t)$ in terms of a, b, λ and μ. Suppose that we chose a, b, λ and μ so that $\mathcal{E}x(t) \equiv 0$.

   (b) Under this hypothesis compute $\mathcal{E}x^2(t)$.

   (c) Under this hypothesis and under certain further conditions it will happen that we can let the rates λ and μ go to infinity and the amplitudes a and b go to zero and get a limiting value for $\mathcal{E}x^2(t)$. What is the most general condition under which this is true?

5. An n by n matrix A is said to be a *circulant matrix* if $a_{ij} = b_{j-i} \mathrm{mod} n$. For example, three by three matrices are of the circulant if they take the form

$$A = \begin{bmatrix} b_0 & b_1 & b_2 \\ b_2 & b_0 & b_1 \\ b_1 & b_2 & b_0 \end{bmatrix}$$

   (a) Show that the sum of the two circulant matrices is circulant and that the product of two circulant matrices is circulant.

   (b) If A and F are circulant show that $AF = FA$.

   (c) Show that there exists a unitary matrix U that diagonalizes circulant matrices.

   (d) If A is circulant we can associate with it a polynomial

$$\hat{A}(z) = \sum_{i=1}^{n-1} b_i z^i$$

   which can be thought of as being a finite z-transform of A.

   (e) Show that if A and F are circulant with $T = AF$ then $\hat{T}(z) = \hat{A}(z)\hat{F}(z)$ if we agree to interpret $e^{\hat{A}(z)} \mathrm{mod}(z^{\dim A} - 1)$.

6. Show that there exists n by n infinitesimally stochastic matrices whose eigenvalues are $\{(-1 + e_i)\}_{i=1}^n$ where $\{e_i\}_{i=1}^n$ are the nth roots of 1.

7. Consider the unidirectional Poisson counter N of rate λ. As we have seen, the expected value of $N(t)$ is λt and

$$\mathcal{E}(N(t) - \lambda t)^2 = \lambda t$$

Thus if we fix t, say $t = 1$ then the process $(N(t)/\lambda - t)$ has a variance $1/\lambda$ and an expected value of zero. Show that

$$\lim_{\lambda \to \infty} \mathrm{prob}\{N(t)/\lambda > 1\} = \begin{cases} 1 & t > 1 \\ 0 & t < 1 \end{cases}$$

This implies that we can use Poisson counters to model events that are arbitrarily close to certain events.

8. This is a continuation of the previous problem. Consider the function $f\colon [0,\infty] \to [0,1]$ defined by

$$f(x) = \begin{cases} 1\,; & 0 \le x < 1 \\ 0\,; & 1 \le x < \infty \end{cases}$$

If we would like to find a state transition associated with a finite state continuous time jump process that has this function as the distribution of its transition time we need to find an approximation in terms of real exponentials. The function $g\colon [0,\infty) \to [0,1]$ defined by

$$g(x) = 1 - e^{-x}$$

is a candidate. Find a better approximation by making use of n by n transition matrices of the circulant form

$$A_n = \begin{bmatrix} -1 & 1 & 0 & 0 & \cdots \\ 0 & -1 & 1 & 0 & \cdots \\ 0 & 0 & -1 & 1 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & 0 & 0 & 0 & \cdots \end{bmatrix}$$

Hint: The eigenvalues of A_n are $-1 + r_n$ with r_n being an n^{th} root of unity.

9. Consider an Itô equation that evolves in the finite set S

$$dx = \Sigma \phi_i(x)\, dN_i \quad ; \quad x(0) \in S$$

Suppose that $\psi\colon S \to \{0,1\}$ and suppose that $\dot{p} = Ap$ governs the probability that $x(t) = x_i$. Show that the probability that $\psi = 1$ is $\sum_{\iota \in I} p_i(t)$ where I is the set of states that ψ maps into 1.

10. A matrix is said to be doubly stochastic if it is stochastic and its rows sum to one.

    (a) Show that every permutation matrix is doubly stochastic.

    (b) Show that the set of all doubly stochastic matrices is a convex set characterized by $(n-1)^2$ parameters.

    (c) Show that if Θ is a real n by n orthogonal matrix then the matrix Ψ whose ij^{th} entry is $\psi_{ij} = (\theta_{ij})^2$ is doubly stochastic. Such matrices are said to be orthostochastic.

    (d) Show that not every doubly stochastic matrix is orthostochastic.

    (e) Show that every permutation matrix is orthostochastic.

11. Suppose that $x(t)$ is Markovian continuous time jump process taking values in the set $X = (x_1, x_2, \ldots, x_n)$. Suppose $Y = \{y_1, y_2, \ldots, y_m\}$ with $m < n$ is a second finite set and let $f : X \to Y$ be any mapping. In this case $y(t) = f\big(x(t)\big)$ is called a *hidden Markov model*. Compute the probability distributions for the time that the y processes spends in state j before jumping to state i. Show that it is a sum of exponentials

$$q(i,j,t) = \sum_{k=1}^{r} p_{ijk} e^{-\lambda_k\, t}$$

12. Let $\{T_1, T_2, \ldots, T_k\}$ be a sequence of transition times associated with a two state continuous time jump process with intensity matrix

$$A = \begin{bmatrix} -p & q \\ p & -q \end{bmatrix}$$

Suppose that $\epsilon_1, \epsilon_2, \ldots, \epsilon_k$ are small and sum to zero. Compute to first order in ϵ the probability that the process will have transition times $\tilde{T}_1, \tilde{T}_2, \ldots, \tilde{T}_n$ with $T_1 - \epsilon \le \tilde{T}_1 \le T_1 + \epsilon, \quad T_2 - \epsilon \le \tilde{T}_2 \le T_2 + \epsilon, \ldots, T_k - \epsilon \le \tilde{T}_k \le T_k + \epsilon$?

13. In the notes we have taken the counting rates to be constant. If the counting rates are state dependent everything is more or less the same. Consider

$$\dot{x} = -x + z$$
$$dz = -2z\, dN$$

with the rate of dN being equal to $|x|$. Show that the Itô rule is still valid with λ's being state dependent and that

$$\frac{d}{dt}\mathcal{E}\, x = -\mathcal{E}\, x + \mathcal{E}\, |x|$$

Suppose that $x(0)$ is positive with probability one. What is the probability that x stays positive forever.

14. Let N be a Poisson Counter of rate λ. Derive the Fokker-Planck equation for

$$dx = -2x\, dt - .5x dN$$

Construct a set of ordinary (deterministic) differential equations (some in t and some in τ) whose solution will determine $\mathcal{E}x(t)x(t+\tau)$.

15. Let N be a Poisson counter with rate λ. Compute the steady state covariance for

$$\begin{aligned} dx_1 &= -2x_1 dN \quad ; \quad x_1(0) = 1 \\ dx_2 &= -x_2 dt + x_1 dt \quad ; \quad x_2(0) = 0 \end{aligned}$$

i.e., compute

$$M(\tau) = \lim_{t\to\infty} \mathcal{E} \begin{bmatrix} x_1(t)x_1(t+\tau) & x_1(t)x_2(t+\tau) \\ x_2(t)x_1(t+\tau) & x_2(t)x_2(t+\tau) \end{bmatrix}$$

16. Let N_1 and N_2 be Poisson counters of rate λ. For

$$dx = -xdt + dN_1 - dN_2$$

the probability density, if it exists, satisfies

$$\frac{\partial \rho}{\partial t}(t,x) = \frac{\partial}{\partial x} x\rho(t,x) + \big(\rho(t,x+1) + \rho(t,x-1) - 2\rho(t,x)\big)\lambda$$

(a) When will $\rho(t,x)$ be continuous? (Discuss at an intuitive level.)

(b) Describe the qualitative behavior of $\rho(t,x)$ as t becomes large - make a sketch.

(c) Back up your analysis by computing the differential equation for

$$\mathcal{E}x^n(t)$$

and using it to solve for $\lim_{t\to\infty} \mathcal{E}x^2(t)$.

17. A *Markov renewal process* is a generalization of the continuous time jump processes considered here. In this case one has a finite or countable state space X and the process jumps from one state to another as in the continuous time jump process but in this case there is a distribution $q(i,j,t)$ which is the probability distribution for the time spent in state j before jumping to state i. Clearly

$$\lim_{t\to\infty} q(i,j,t) = p_{ij}$$

must define a matrix with nonnegative entries whose columns sum to one. Show that if $q(i,j,t)$ takes the special form

$$q(i,j,t) = p_{ij}(1 - e^{-\lambda_{ij} t})$$

with p_{ij} constant then the Markov renewal process is a continuous time jump process.

18. Consider a two dimensional situation in which $x(t,\tau)$ takes on the values {0,1,2, ... } with $x(0,0) = 0$ and x being a Poisson process of rate λ in t along lines of constant τ and a Poisson process of rate μ in τ along lines of constant t. Does it follow that

$$x(t,\tau) = \phi(t) \cdot \psi(\tau)$$

19. Let $\alpha, \beta, \gamma, \delta$ be four real numbers with $0 < \alpha < \beta < \gamma < \delta$. Let N be a standard Poisson counter of rate λ. Show that the probability that $N(\alpha) = 0, N(\beta) = 1, N(\gamma) = 1, N(\delta) = 2$ is given by

$$p = \lambda^2(\beta - \alpha)(\delta - \gamma)e^{-\lambda\delta}$$

20. The discrete time markov process whose transition law is

$$\begin{bmatrix} p_1(k+1) \\ p_2(k+1) \\ p_3(k+1) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} p_1(k) \\ p_2(k) \\ p_3(k) \end{bmatrix}$$

is actually deterministic in the sense that a knowledge of the state of process at any one moment in time determines the state (not just its probability distribution) at any other point in time. The only continuous time analog of such "non-mixing" processes is the trivial process $x(t) =$ constant. If $p(t)$ is a 3-vector, and if

$$\dot{p}(t) = Ap(t)$$

with A being infintessimally stochastic, and if $p(0) = e_1$, the standard basis vector, show that

$$\min_{t>0} \|e^{At}e_1 - e_2\| + \min_{t>0} \|e^{At}e_1 - e_3\| \geq \sqrt{2}$$

where $\|p\| = \sqrt{p_1^2 + p_2^2 + p_3^2}$. (This describes a sense in which a continuous time finite state jump process can not be too close to a deterministic process.)

NOTES AND REFERENCES

1. Poisson Counters are discussed in most introductory books on probability. One classic is

   (a) W. Feller, *Probability Theory*, Vol. I, John Wiley, New York, 1957 and Volume II, 1966

   Because we will make heavy use of systems of ordinary (deterministic) differential equations, the reader may want a reference. Everything we need and much more can be found in:

   (a) E. A. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, McGraw Hill, New York, 1955

2. Continuous time jump process are discussed in many books. An introductory treatment can be found in

   (a) M. Rosenblatt, *Random Processes*, Oxford University Press, New York, 1968.

3. This is an almost trivial special case of the general form of Itô's rule. The origin of the idea is Itô's much cited paper

   (a) K. Itô, "Stochastic Integral", *Proc. Imperial Acad. Tokyo*, Vol. 20, 1944, pp. 519-524.

4. Computing an expectation is often the key step in getting a practical result. We have arranged the presentation so that the reader quickly arrives at a point where some interesting calculations can be done. One advantage of the differential equations approach to stochastic process is that such computations are usually rather transparent.

5. The computation of two point statistics will play an important role in the stochastic realization work to be discussed in chapter V.

6. Jump processes and linear systems are sometimes used as models for systems with failure modes.

7. These equations are of theoretical interest but they are very seldom soluble in closed form. In some cases the steady state value can be computed even though the time dependent solution is not known. In the case of finite state continuous time jump process the Peron-Frobenius theory applies and in the non finite state cases these are conditions under which one can guarantee the existence of an invariant measure as well.

Chapter 3

Wiener Processes and Differential Equations

In this chapter we introduce a second basic stochastic process, the Wiener process. The Itô calculus will be extended to include stochastic equations involving Wiener processes and the computation of expectations, the equation for the probability law, etc. will be developed.

3.1 Gaussian Distributions

From one point of view, the special significance of the Gaussian density derives, in part, from its role in solving a very basic differential equation variously called the *diffusion equation* or the *heat equation*. In one dimension, this is the partial differential equation of evolutionary type

$$\frac{\partial \rho(t,x)}{\partial t} = \frac{1}{2}\frac{\partial^2 \rho(t,x)}{\partial x^2}$$

which governs diffusion. A short calculation shows that the function $\rho(t,x) = \frac{1}{\sqrt{2\pi t}}e^{-x^2/2t}$ satisfies this equation for all $t > 0$. In fact, the general solution of this equation for twice differentiable initial data $\rho(0,x)$ is, for $t > 0$, given by

$$\rho(t,x) = \int_{\mathbb{R}^n} \frac{1}{\sqrt{2\pi t}} e^{-(x-y)^2/2t}\rho(0,y)dy$$

For this reason $\frac{1}{\sqrt{2\pi t}}e^{-x^2/2t}$ is often called the *heat kernel*.

There is a simple n-dimensional generalization. If $Q = Q^T$ is positive definite, and if we let q_{ij} denote the ij^{th} entry of Q then the partial differential equation

$$\frac{\partial \rho(t,x)}{\partial t} = \frac{1}{2}\sum_{i,j=1}^{n} q_{ij}\frac{\partial}{\partial x_i}\frac{\partial}{\partial x_j}\rho(t,x)$$

can be thought of as an n-dimensional generalization of the one-dimensional equation given above. It has the fundamental solution

$$\rho(t,x) \;=\; \frac{1}{\sqrt{\det Q(2\pi t)^n}} e^{-x^T(2Qt)^{-1}x}$$

The solution for smooth initial data $\rho(0,x)$ is

$$\rho(t,x) = \int_{\mathbb{R}^n} \frac{1}{\sqrt{\det Q}(2\pi t)^n} e^{-(x-y)^T(2Qt)^{-1}(x-y)} \rho(0,y)dy$$

The Gaussian distribution arises as a limiting form of the binominal distribution in the following way. Notice that

$$\begin{aligned}\lim_{n\to\infty}\left(1-\frac{\alpha^2}{n}\right)^n &= \lim_{n\to\infty} e^{n\ln(1-\alpha^2/n)} \\ &= e^{-\alpha^2}\end{aligned}$$

3.2 Brownian Motion

Robert Brown (1773–1858) was a Scottish botanist who became interested in stochastic processes after looking under a microscope at grains of pollen suspended in water. He published influential papers on this subject in 1827. Norbert Wiener undertook the mathematical study of stochastic processes in the 1920's after important work earlier in the century by Bachelier, Einstein, Smoluchowski, *et al.* There are several ways of approaching the mathematics of this subject. We chose one that continues from our work on jump processes.

Let $y(t)$ be a bidirectional Poisson counter of rate $\lambda/2$ and let $x_\lambda(t) = y(t)/\sqrt{\lambda}$. We may realize this process as

$$dx_\lambda = \frac{1}{\sqrt{\lambda}}(dN_1 - dN_2) \quad ; \quad x(0) = 0$$

where N_1 and N_2 are ordinary Poisson counters of rate $\lambda/2$. A quick calculation based on the Itô rule of chapter 2 shows that

$$dx_\lambda^p = \big((x_\lambda + \frac{1}{\sqrt{\lambda}})^p - x_\lambda^p\big)dN_1 + \big((x_\lambda - \frac{1}{\sqrt{\lambda}})^p - x_\lambda^p\big)dN_2$$

For reasons of symmetry, if p is an odd positive integer $\mathcal{E}x_\lambda^p = 0$. If p is an even positive integer then the expectation rule of chapter 2 yields

$$\frac{d}{dt}\mathcal{E}x_\lambda^p = \frac{1}{2}\mathcal{E}\Big((x_\lambda + \frac{1}{\sqrt{\lambda}})^p + (x_\lambda - \frac{1}{\sqrt{\lambda}})^p - 2x_\lambda^p\Big)$$

$$= \binom{p}{2}\mathcal{E}x_\lambda^{(p-2)} + \frac{1}{\lambda}\binom{p}{4}\mathcal{E}x_\lambda^{(p-4)} + \cdots$$

where the terms not written involve powers of λ^{-1} greater than one. Because of this, for t fixed there exists a high counting rate limit for the expected value of each moment

$$\lim_{\lambda\to\infty}\mathcal{E}x_\lambda^p(t) = \frac{1}{2}\int_0^t p(p-1)\lim_{\lambda\to\infty}\mathcal{E}x_\lambda^{p-2}(\sigma)d\sigma$$

Solving this set of equations beginning at $p = 2$ and continuing to larger, even values of p we see

that the limiting values of the moments are

$$\begin{aligned}
\lim_{\lambda\to\infty} \quad \mathcal{E}x_\lambda^2(t) &= t \\
\lim_{\lambda\to\infty} \quad \mathcal{E}x_\lambda^4(t) &= 3t^2 \\
\lim_{\lambda\to\infty} \quad \mathcal{E}x_\lambda^6(t) &= 5\cdot 3\cdot t^3 \\
\cdots &= \cdots \\
\lim_{\lambda\to\infty} \quad \mathcal{E}x^p(t) &= (p-1)(p-3)\ldots t^{p/2} \\
&= \frac{p!}{(p/2)!}\cdots\left(\frac{t}{2}\right)^{p/2}
\end{aligned}$$

Thus in the limit as λ goes to infinity we get the moments associated with a Gaussian density having zero mean and variance t. Likewise, we see that for $\tau \geq t$ we have the differential equation

$$\begin{aligned}
\frac{d}{d\tau}\mathcal{E}x(t)x(\tau) &= \mathcal{E}x(t)\frac{1}{\sqrt{\lambda}}(dN_1(\tau) - dN_2(\tau)) \\
&= 0
\end{aligned}$$

and so $\mathcal{E}x(t)x(t+\tau) = \mathcal{E}x^2(t)$ for $\tau \geq t$.

We remark that $x_\lambda(t)$ has three properties which are more or less obvious

1. $x_\lambda(0) = 0$.
2. $x_\lambda(t) - x_\lambda(\tau)$ is a random variable whose distribution is dependent only on $|t-\tau|$ and if $[t,\tau]\cap[s,\sigma]$ is empty, the random variables $x_\lambda(t)-x_\lambda(\tau)$ and $x_\lambda(s)-x_\lambda(\sigma)$ are independent. (Because N on the interval $[s,\sigma]$ is independent of N on the interval $[t,\tau]$)
3. The limit as $\lambda\to\infty$ of $\mathcal{E}\big(x_\lambda(t)-x_\lambda(\tau)\big)^2 = \mathcal{E}x_\lambda^2(t) - 2\mathcal{E}x_\lambda(t)x_\lambda(\tau) + x_\lambda^2(\tau)$ exists and is just $|t-\tau|$.

From the results of section 2.5 we see that if there exists a probability density for x_λ then it is given by

$$\frac{\partial\rho(t,x)}{\partial t} = \frac{\lambda}{2}\left(\rho\Big(t, x+\frac{1}{\sqrt{\lambda}}\Big) - 2\rho(t,x) + \rho\Big(t, x-\frac{1}{\sqrt{\lambda}}\Big)\right)$$

Notice that in the limit as λ goes to infinity, we get (formally) the diffusion equation

$$\frac{\partial\rho(t,x)}{\partial t} = \frac{1}{2}\frac{\partial^2\rho(t,x)}{\partial x^2}$$

This analysis shows that if we use the Itô rule for jump processes, then certain limits as λ goes to infinity exist. There is, in fact, a stochastic process known as a standard Wiener process, which has the properties we have derived for the above limiting forms and, in addition, is continuous with probability one. Continuity is, of course, not surprising in view of the fact that $\mathcal{E}(x(t)-x(\tau))^2 = |t-\tau|$. The analysis of limits given here does not, in and of itself, establish the existence of a stochastic process having the limiting properties. We will not, in these notes, prove its existence. We will, instead, appeal to limiting processes of the above type whenever we need to explore its properties. For convenience we will refer to the limiting form as being *Brownian motion*.

3.3 Stochastic Differential Equations

This same limiting process can be applied to give meaning to a general class of equations which we write as

$$dx = f(x)dt + \sum_{i=1}^{n} g_i(x)dw_i$$

By a solution of this equation we understand the limit as λ goes to infinity of the solution of

$$dx = f(x)dt + \sum_{i=1}^{n} g_i(x)(dN_i - dN_{-i})/\sqrt{\lambda}$$

where N_i and N_{-i} are independent Poisson counters with rate $\lambda/2$.

Example 1: (The Ornstein-Uhlenbeck Process) Consider the process

$$dx = vdt$$

$$dv = -\alpha v dt + dw$$

The process x is obtained by integrating the v process. The v process would be Brownian motion if α were zero. If α positive, and this is the situation which is of interest in applications, we will show that the probability distribution of v approaches a steady-state value as t goes to infinity. We will give more details in later sections.

3.4 The Itô Rule

Let ψ be a twice differentiable function of x. Consider the evaluation of $d\psi$ for

$$dx = f(x)dt + \sum_{i=1}^{n} g_i(x)(dN_i - dN_{-i})/\sqrt{\lambda}$$

where N_i are Poisson counters of rate $\lambda/2$. Using the Itô rule of chapter 2 we get

$$\begin{aligned} d\psi &= \left\langle \frac{\partial \psi}{\partial x}\ ,\ f(x) \right\rangle dt + \sum_{i=1}^{n} \left[\psi\big(x + \frac{1}{\sqrt{\lambda}} g_i(x)\big) - \psi(x) \right] dN_i \\ &\quad + \sum_{i=1}^{n} \left[\psi\big(x - \frac{1}{\sqrt{\lambda}} g_i(x)\big) - \psi(x) \right] dN_{-i} \end{aligned}$$

In order to explore the limit as λ goes to infinity, we expand ψ in a Taylor series about x. The result is

$$\begin{aligned} d\psi &= \left\langle \frac{\partial \psi}{\partial x}\ ,\ f(x) \right\rangle dt + \sum_{i=1}^{n} \left\langle \frac{\partial \psi}{\partial x}\ ,\ g_i(x) \right\rangle (dN_i - dN_{-i})/\sqrt{\lambda} \\ &\quad + \sum_{i=1}^{n} \frac{1}{2} \langle g_i(x)\ ,\ \frac{\partial^2 \psi}{\partial x^2} g_i(x) \rangle (dN_i + dN_{-i})/\lambda + O(1/\lambda^{3/2}) dN_i \end{aligned}$$

Consider now

$$dz_\lambda = (dN_1 + dN_{-1})/\lambda$$

where N_1 and N_{-1} are independent Poisson counters of rate $\lambda/2$. Using the formula for expectation from chapter 2, we see that

$$\mathcal{E}z_\lambda(t) = t + \mathcal{E}z(0)$$

Let $m_\lambda = z_\lambda^2(t)$. Then from the Itô rule for jump processes we see that

$$dm_\lambda = \left(\left(z_\lambda + \frac{1}{\lambda}\right)^2 - z_\lambda^2\right)(dN_1 + dN_2) = \frac{2z_\lambda}{\lambda}(dN_1 + dN_2) + \frac{1}{\lambda^2}(dN_1 + dN_2)$$

A short calculation, using $\mathcal{E}dN_i = (\lambda/2)\,dt$, yields

$$\frac{d}{dt}\mathcal{E}m_\lambda(t) = \mathcal{E}(2z\lambda + \frac{1}{\lambda}$$

Thus if $\mathcal{E}z(0) = 0$ we have

$$\mathcal{E}m_\lambda(t) = t^2 + t/\lambda$$

It is remarkable that the variance of the process defined by z, namely

$$\sigma(t) = \mathcal{E}\big(z_\lambda(t) - \mathcal{E}z_\lambda(t)\big)^2$$

which goes along with the initial condition $z(0) = 0$ is just

$$\begin{aligned}\sigma(t) &= t^2 + t/\lambda - 2t^2 + t^2 \\ &= t/\lambda\end{aligned}$$

and hence goes to zero as λ goes to infinity. As λ becomes larger the uncertainty associated with z decreases and, in a sense, z tends to the deterministic process defined by $z(t) = t$. As a result, the limiting form of the Itô rule takes a simple form. We may make the replacement, as λ goes to infinity, $(dN_i + dN_{-i})/\lambda = dt$. Thus for

$$dx = f(x)dt + \sum_{i=1}^{m} g_i(x)dw_i$$

we have

$$d\psi = \left\langle \frac{\partial\psi}{\partial x}\ ,\ f(x)\right\rangle dt + \sum_{i=1}^{m}\left\langle \frac{\partial\psi}{\partial x}\ ,\ g_i(x)\right\rangle dw_i + \frac{1}{2}\sum_{i=1}^{m}\left\langle g_i(x)\ ,\ \frac{\partial^2\psi}{\partial x^2}g_i(x)\right\rangle dt$$

where we have abbreviated $(dN_i - dN_{-i})/\sqrt{\lambda}$ as dw_i. This is known as the *Itô rule for Brownian motion.*

Example 1: Consider

$$dx = -xdt + xdw$$

Suppose we want to find an equation for $z = x^2$. Using the Itô rule we have

$$dx^2 = 2x(-xdt + xdw) + x^2dt$$

The last term results from an evaluation of $1/2(\partial^2\psi/\partial x^2)\cdot(g(x))^2dt$. The net result is

$$\begin{aligned}dz &= -2zdt + 2zdw + zdt \\ &= -zdt + 2zdw\end{aligned}$$

Example 2: Consider the pair of Itô equations

$$dx = dw \;\; ; \;\; dy = d\nu$$

with w and ν being independent Wiener processes. If we describe matters in polar coordinates, $x = r\cos\theta$, $y = r\sin\theta$, then of course $r = \sqrt{x^2+y^2}$ and $\theta = \tan^{-1}(y/x)$. To obtain the differential equations for r and θ we can use the Itô rule as follows. First of all,

$$\frac{\partial r}{\partial x} = \frac{x}{\sqrt{x^2+y^2}} \;\; ; \;\; \frac{\partial r}{\partial y} = \frac{y}{\sqrt{x^2+y^2}} \;\; ; \;\; \frac{\partial \theta}{\partial x} = \frac{-y}{x^2+y^2} \;\; ; \;\; \frac{\partial \theta}{\partial y} = \frac{x}{x^2+y^2}$$

The matrix of second partial derivatives is

$$\begin{bmatrix} \frac{\partial^2 r}{\partial x^2} & \frac{\partial^2 r}{\partial x \partial y} \\ \frac{\partial^2 r}{\partial x \partial y} & \frac{\partial^2 r}{\partial y^2} \end{bmatrix} = \begin{bmatrix} \frac{y^2}{\sqrt{x^2+y^2}^3} & \frac{-xy}{\sqrt{x^2+y^2}^3} \\ \frac{-xy}{\sqrt{x^2+y^2}^3} & \frac{x^2}{\sqrt{x^2+y^2}^3} \end{bmatrix} \;\; ; \;\; \begin{bmatrix} \frac{\partial^2 \theta}{\partial x^2} & \frac{\partial^2 \theta}{\partial x \partial y} \\ \frac{\partial^2 \theta}{\partial x \partial y} & \frac{\partial^2 \theta}{\partial y^2} \end{bmatrix} = \begin{bmatrix} \frac{xy}{(x^2+y^2)^2} & \frac{-1}{x^2+y^2} \\ \frac{-1}{x^2+y^2} & \frac{xy}{(x^2+y^2)^2} \end{bmatrix}$$

Using this in the Itô rule we have

$$dr = \frac{xdw + yd\nu}{\sqrt{x^2+y^2}} + \frac{dt}{2\sqrt{x^2+y^2}}$$

A similar calculation shows that θ satisfies

$$d\theta = \frac{xd\nu - ydw}{x^2+y^2} + \frac{xydt}{(x^2+y^2)^2}$$

Expressing the right-hand sides of these equations in terms of r and θ we get

$$dr = \frac{1}{2r}dt + \sin\theta d\nu + \cos\theta dw$$

$$d\theta = \frac{\cos\theta d\nu - \sin\theta dw}{r} + \frac{\sin\theta\cos\theta dt}{r^2}$$

3.5 Expectations

Adopting the notation of the previous section we now consider stochastic equations of the form

$$dx = f(x)dt + \sum_{i=1}^{m} g_i(x)dw_i$$

with w_i being a Wiener process. What is the rule for computing expectations for such an equation? If we replace dw_i by

$$dw_{i\lambda} = \frac{1}{\sqrt{\lambda}}(dN_i - dN_{-i})$$

where N_i and N_{-i} are Poisson counters of rate $\lambda/2$, then it is clear that in the limit as λ goes to infinity we get simply

$$\frac{d}{dt}\mathcal{E}x = \mathcal{E}f(x)$$

Thus the rule for computing expectation is simply that $\mathcal{E}g(x)dw = 0$. This is a consequence of the fact that we defined the solutions of our jump-process equations to be continuous from the left and have defined the solution after the jump in terms of the left limit.

The result of this method of defining solutions is to destroy the usual time reversal symmetry of the time axis in a way that has no analogy in the study of ordinary differential equations.

Example: To illustrate, consider the model

$$dx = xdt + \alpha x dw$$

This might represent, for example, the growth of capital when the interest rate is a random process. We see that

$$\mathcal{E}x(t) = e^t \mathcal{E}x(0)$$

The expectation of $x^2(t)$ can be computed by first using the Itô rule to get

$$dx^2 = 2x^2 dt + 2\alpha x^2 dw + \alpha^2 x^2 dt$$

and then taking expectations to get

$$\frac{d}{dt}\mathcal{E}x^2(t) = (2 + \alpha^2)\mathcal{E}[x^2(t)]$$

Example 2: Consider system

$$dx = vdt \ ; \ x(0) = 0$$

$$dv = adt \ ; \ v(0) = 0$$

$$da = dw \ ; \ a(0) = 0$$

To compute the covariance of $(x(t), v(t), a(t))$ we write this in matrix notation

$$d\begin{bmatrix} x \\ v \\ a \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}\begin{bmatrix} x \\ v \\ a \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} dw$$

and take the expectation to get we get

$$\frac{d}{dt}\mathcal{E}\begin{bmatrix} x \\ v \\ a \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}\mathcal{E}\begin{bmatrix} x \\ v \\ a \end{bmatrix}$$

In terms of the notation

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \ ; \ B = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

we have

$$\mathcal{E}\begin{bmatrix} x(t) \\ v(t) \\ a(t) \end{bmatrix} = e^{At}\mathcal{E}\begin{bmatrix} x(0) \\ v(0) \\ a(0) \end{bmatrix} = 0$$

If we let Σ denote the covariance then

$$\dot{\Sigma} = A\Sigma + \Sigma A^T + BB^T$$

$$\Sigma(t) = \int_0^t e^{A\sigma} BB^T e^{A^T\sigma} d\sigma$$

$$e^{At} = I + At + \frac{1}{2}A^2t^2 = \begin{bmatrix} 1 & t & \frac{t^2}{2} \\ 0 & 1 & t \\ 0 & 0 & 1 \end{bmatrix}$$

This yields

$$\int_0^T \begin{bmatrix} 1 & t & \frac{t^2}{2} \\ 0 & 1 & t \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ t & 1 & 0 \\ \frac{t^2}{2} & t & 1 \end{bmatrix} dt$$

$$= \int_0^T \begin{bmatrix} 0 & 0 & \frac{t^2}{2} \\ 0 & 0 & t \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ t & 1 & 0 \\ \frac{t^2}{2} & t & 1 \end{bmatrix} dt$$

$$= \int_0^T \begin{bmatrix} \frac{t^4}{4} & \frac{t^3}{2} & \frac{t^2}{2} \\ \frac{t^3}{2} & t^2 & t \\ \frac{t^2}{2} & t & 1 \end{bmatrix} dt$$

$$= \begin{bmatrix} \frac{1}{20}T^5 & \frac{1}{8}T^4 & \frac{1}{6}T^3 \\ \frac{1}{8}T^4 & \frac{1}{3}T^3 & \frac{1}{2}T^2 \\ \frac{1}{6}T^3 & \frac{1}{2}T^2 & T \end{bmatrix}$$

3.6 A Digression on Stratonovic Calculus

An important aspect of the Wiener process is the formula for the quadratic variation

$$\lim_{\Delta_i \to 0} \sum |w(t_{i+1}) - w(t_i)|^2 = |a - b|$$

where $\Delta_i = t_{i+1} - t_i$ and the sum is over subdivisions of the interval $[a, b]$. This fact underlies the useful calculation rule, $dw \cdot dw = dt$. In order to define what is meant by the stochastic differential equation

$$dx = f(x)dt + g(x)dw$$

one integrates to get

$$x(t) - x(0) = \int_0^t f(x(\sigma))d\sigma + \int_0^t g(x(\sigma))dw$$

and transfers the difficulty to that of interpreting the integral. In the figure below we illustrate part of the construction of the (non anticipatory) Itô integral

$$\int_a^b g(x(t))dw(t) = \lim_{\Delta_i \to 0} \sum g(x(t_i))(w(t_i + \Delta_i) - w(t_i))$$

On the other hand, in the (partially anticipatory) Stratonovic calculus one uses a different definition of the integral

$$\int_a^b g(x(t))đw(t) = \lim_{\Delta_i \to 0} \sum g(x(t_i + \Delta_i/2))(w(t_i + \Delta_i) - w(t_i))$$

Using the notation and setup of the figure shown below, we explore the relationship between the Itô and Stratonovic integrals.

The Itô integral is based on the interpretation of

$$I = \int_a^b g(x(t))\, dw_t$$

in which $g(x(t))$ is evaluated at the left end point of each subdivision of $[a, b]$ and dw_t is treated as occurring "in advance" of any change in $x(t)$. That is to say, we define the integral as the limit of sums of the type

$$I = \sum_i g(x(t_i))(w(t_i + \Delta) - w(t_i))$$

On the other hand, we could have used instead

$$I = \sum g\left(x(t_i + \Delta/2)\right)\left(w(t_i + \Delta) - w(t_i)\right)$$

evaluating g in the middle of the subdivision. Unlike the situation one has in Riemannian integration where such a change makes no difference, in this case one gets a different integral. Since this "central difference" interpretation will be useful later, we will now discuss the precise way in which it differs from the Itô interpretation. These remarks lead to a different, but for our purposes equally expressive, version of the stochastic calculus. It is also widely used. The key observation is that from Taylor series we get

$$g(x(t + \Delta/2)) = g(x(t)) + \left.\frac{\partial g}{\partial x}\right|_{x(t)} \Delta x/2$$

If we evaluate Δx as

$$\Delta x = f(x + \Delta x/2)\Delta t + g(x + \Delta x/2)\Delta w$$

we see that

$$g(x(t) + \Delta x/2) \approx g(x(t)) + (\partial g/\partial x)g(x)\Delta w + \text{ terms of order } \Delta t, (\Delta w)^2 \text{ and higher}$$

Recall that when we introduced the Wiener process as the limit of a certain construction involving Poisson counters, we found that

$$\mathcal{E}(w(t) - w(\tau))^2 = |t - \tau|$$

and that as a consequence of this it was necessary to treat the differential calculus with additional care. In particular, it happened that $(dw)^2$ is, in a certain sense, first order and equal to the

deterministic differential dt. The change of variables and expectation formula associated with the Itô calculus is summarized by

$$\begin{aligned} dx &= f(x)\,dt + g(x)\,dw \\ d\psi &= \left\langle \frac{\partial \psi}{\partial x}\ ,\ f(x)\,dt + g(x)\,dw \right\rangle + \frac{1}{2} g^T(x) \frac{\partial^2 \psi}{\partial x^2} g(x) dt \\ d\mathcal{E}\psi/dt &= \left\langle \frac{\partial \psi}{\partial x}\ ,\ f(x) \right\rangle + \frac{1}{2} g^T(x) \frac{\partial^2 \psi}{\partial x^2} g(x) \end{aligned}$$

When we switch and represent dynamics by differential equations which are to be interpreted as central differences, these equations change. The change can be best appreciated by noting that a Taylor series expansion yields

$$g(x(t + \triangle/2)) = g(x(t)) + \frac{1}{2}\frac{\partial g}{\partial x}(x(t))\triangle + \frac{1}{2}(\triangle/2)^2 \frac{\partial^2 g}{\partial x^2}(x)$$

But Δx is $f(x)\Delta t + g(x)\Delta w$ and since $(\Delta w)(\Delta t)$, but not $(\Delta w)^2$, is of order higher than Δt, we see that

$$\int_{\text{Itô}} g(x(t))\,dw = \int_{central} g(x_t)\,dw - \frac{1}{2}\int (\partial g/\partial x) g\,dt$$

Thus we have a different calculus; different from, but equally expressive as, the Itô calculus. Using barred d's to indicate differentials in the Stratonovic sense, we can summarize via

$$\begin{aligned} đx &= f(x)\,dt + g(x)\,đw \\ đ\psi &= \left\langle \frac{\partial \psi}{\partial x}\ ,\ f(x)\,dt + g(x)\,đw \right\rangle \\ d(\mathcal{E}x)/dt &= \mathcal{E}\left(f(x) + \frac{1}{2}\left(\frac{\partial g}{\partial x}\right) g(x) \right) \end{aligned}$$

If x satisfies the Itô equation

$$dx = f(x)dt + \sum_{i=1}^{m} g_i(x) dw_i$$

then it satisfies the Stratonovic equation

$$đx = (f(x) - \tfrac{1}{2}\sum_{i=1}^{m} \frac{\partial g}{\partial x} g_i)dt + \sum_{i=1}^{m} g_i(x) đw_i$$

Figure 3.1. Illustrating the increment $w(t + \Delta t) - w(t)$ and $g(x(t))$.

3.7 The Fokker-Planck Equation

Hand in hand with the Itô differentiation formula goes the evolution equation for the probability density. Consider in $\mathbb{R}^n$ the Itô equation

$$dx = f(x)dt + \sum_{i=1}^{m} g_i(x)dw_i$$

and let ψ be a smooth function having compact support. Then

$$d\psi = \left\langle \frac{\partial \psi}{\partial x} \ , \ f(x)dt + \sum_{i}^{m} g(x)dw_i \right\rangle + \frac{1}{2}\sum_{i=1}^{m} \left\langle \frac{\partial^2 \psi}{\partial x^2} g_i(x) \ , \ g_i(x) \right\rangle dt$$

If we are interested in $\frac{d}{dt}\mathcal{E}\psi$, we have on one hand

$$\begin{aligned} \frac{d}{dt}\mathcal{E}\psi &= \frac{d}{dt}\int_{\mathbb{R}^n} \psi(x)\rho(t,x)dx \\ &= \int_{\mathbb{R}^n} \psi(x)\frac{\partial \rho(t,x)}{\partial t}dx \end{aligned}$$

On the other hand,

$$\begin{aligned} \frac{d}{dt}\mathcal{E}\psi &= \mathcal{E}\left\langle \frac{\partial \psi}{\partial x} \ , \ f(x) \right\rangle + \frac{1}{2}\sum_{i=1}^{m} \mathcal{E}\left\langle \frac{\partial^2 \psi}{\partial x^2} g_i(x) \ , \ g_i(x) \right\rangle \\ &= \int_{\mathbb{R}^m} \left(\left\langle \frac{\partial \psi}{\partial x} \ , \ f(x) \right\rangle + \frac{1}{2}\sum_{i=1}^{n} \left\langle \frac{\partial^2 \psi}{\partial x^2} g_i(x) \ , \ g_i(x) \right\rangle \right) \rho(t,x)dx \end{aligned}$$

Integrating this by parts we get

$$\frac{d}{dt}\mathcal{E}\psi = \int_{\mathbb{R}^n} -\psi \left\langle \frac{\partial}{\partial x} \ , \ \rho(t,x)f(x) \right\rangle + \psi \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{n} (\frac{\partial}{\partial x_j}\frac{\partial}{\partial x_k} g_j^i(x) g_k^i(x))\rho(t,x)dx$$

Putting these two formulas for $\frac{d}{dt}\mathcal{E}\psi$ together we have

$$0 = \int_{\mathbb{R}^n} \psi \left(\frac{\partial \rho}{\partial t} + \left\langle \frac{\partial}{\partial x} \ , \ \rho(t,x)f(x) \right\rangle - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{n} \frac{\partial^2}{\partial x_j \partial x_k} g_j^i(x) g_k^i(x)\rho(t,x) \right) dx$$

The only way this can hold true for all ψ is if

$$\frac{\partial \rho(t,x)}{\partial t} = -\left\langle \frac{\partial}{\partial x} \ , \ \rho(t,x)f(x) \right\rangle + \frac{1}{2}\sum \frac{\partial^2}{\partial x_i \partial x_j} g_i(x) g_j(x)\rho(t,x)$$

This is the Fokker-Planck equation; the evolution equation for the probability law.

Example 1: Consider the Itô equation

$$dx = -xdt + dw$$

The corresponding Fokker-Planck equation is

$$\frac{\partial \rho(t,x)}{\partial t} = \frac{\partial x\rho(t,x)}{\partial x} + \tfrac{1}{2}\frac{\partial^2}{\partial x}\rho(t,x)$$

If $\rho(0,x)$ is given then one can verify that

$$\rho(t,x) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma(t)}} e^{-(x-e^{-t}\eta)^2/2\sigma(t)} \rho(0,\eta) d\eta$$

with $\sigma(t) = \frac{1}{2}(1-e^{-2t})$

Example 2: Seeking an interesting topic for a term paper, a novice in finance asks the following question. If I can solve for the joint probability distribution for the pair of Itô equations

$$dx = -axdt + bxdw_1$$

$$dy = -cydt + fydw_2$$

with w_1 and w_2 independent, can I also solve for the joint probability density when

$$\begin{bmatrix} dx \\ dy \end{bmatrix} = \begin{bmatrix} -a & 0 \\ 0 & -c \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} dt + \begin{bmatrix} -b & h \\ g & -f \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} dw_1 + \begin{bmatrix} -b' & h' \\ g' & -f' \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} dw_2$$

Assuming suitable initial conditions, write down a joint density $\rho(t,x,y)$ in the first case using the fact that the logrthims of x and y are Gaussian for all time if they are at $t=0$. (Do not forget that we are dealing with Itô equations here.) Should the student expect to be successful in the more general situation? Explain your answer carefully, including the role of a term of the form $AB-BA$.

Solution: If we have $dx = -axdt + bxdw$ then we can apply the Itô rule to get the differential of $\psi(x) = \ln x$ as

$$d\psi(x) = -adt + bdw - \frac{1}{2}\frac{b^2x^2}{x^2}dt = (-a - \frac{1}{2}b^2)dt + bdw$$

This can be solved for a Gaussian density of mean $(c_1 - a - b^2/2)t$ and variance $c_2 + tb^2$. The same remarks apply to the second equation. The matrix system can not be solved in closed form because the linear system is time varying and does not decouple.

Example 3: Consider the stochastic equations

$$\begin{aligned} dx &= -(10+z)xdt + dw \\ dz &= -2zdN \end{aligned}$$

with $z(0) \in \{-1,1\}$ and N a Poisson counter of rate λ. Instead of writing $\rho(t,x,z)$ we write $\rho(t,x,1) = \rho_+(t,x)$ and $\rho(t,x,-1) = \rho_-(t,x)$. In this notation one has the following pair of equations for the density

$$\frac{\partial \rho_+(t,x)}{\partial t} = \lambda(\rho_-(t,x) - \rho_+(t,x)) + \frac{\partial}{\partial x} 11x\rho_+(t,x) + \frac{1}{2}\frac{\partial^2}{\partial x^2}\rho_+(t,x)$$

$$\frac{\partial \rho_-(t,x)}{\partial t} = \lambda(\rho_+(t,x) - \rho_-(t,x)) + \frac{\partial}{\partial x} 9x\rho_-(t,x) + \frac{1}{2}\frac{\partial^2}{\partial x^2}\rho_-(t,x)$$

Example 4: Consider the equation

$$\begin{bmatrix} dx \\ dy \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} dt + \begin{bmatrix} dw \\ d\nu \end{bmatrix} + \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} (dN_1 - dN_2)$$

with N_1 and N_2 being Poisson counters of rate λ and w and ν being standard Wiener processes. Find a differential equation for the mean and the variance of (x,y). Find the Fokker-Planck

equation for $\rho(t,x,y)$. Discuss the influence of the size of λ on the existence of a steady-state solution. (Assume throughout that w, ν, N_1, and N_2 are all independent, and that θ is constant) Working in terms of matrices, and using Σ to denote the expectation of xx^T, we have

$$\dot{\Sigma} = -2\Sigma + I + \mathcal{E}((x+\Theta x)(x+\Theta x)^T - xx^T)\lambda + ((x-\Theta x)(x-\Theta x)^T - xx^T)\lambda$$

This, in turn, simplifies to

$$\dot{\Sigma} = -2\Sigma + I + 2\Theta\Sigma\Theta^T\lambda$$

Clearly, we see that if $|\lambda|$ is larger than 1 this system might be unstable, depending on the eigenvalues of the operator $L(\cdot) = \Theta(\cdot)\Theta^T$.

In terms of components everything is much more messy.

$$d\mathcal{E}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}\mathcal{E}\begin{bmatrix} x \\ y \end{bmatrix}dt + \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}\mathcal{E}\begin{bmatrix} x \\ y \end{bmatrix}(\lambda dt - \lambda dt)$$

and

$$\mathcal{E}\begin{bmatrix} x \\ y \end{bmatrix} = e^{\begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}t}\mathcal{E}\begin{bmatrix} x(0) \\ y(0) \end{bmatrix}$$

$$\lim_{t\to 0}\mathcal{E}\begin{bmatrix} x \\ y \end{bmatrix} = 0$$

$$dx^2 = 2x(-x)dt + 2xdw + dt + [(x + x\cos\theta + y\sin\theta)^2 - x^2]dN_1 + [(x - x\cos\theta - y\sin\theta)^2 - x)^2]dN_2$$

From this we can take expectations to get

$$\frac{d}{dt}\mathcal{E}x^2 = -2\mathcal{E}x^2 + (\mathcal{E}(x\cos\theta + y\sin\theta)^2)\lambda + 1$$

$$\frac{d}{dt}\mathcal{E}x^2 = -2\mathcal{E}x^2 + 2\lambda\sin 2\theta\mathcal{E}xy + \lambda\mathcal{E}x^2 + \lambda\mathcal{E}y^2 + 1$$

We omit further details.

For the Fokker-Planck equation we have

$$\frac{\partial\rho(t,x,y)}{\partial t} = 2(\frac{\partial}{\partial x}x + \frac{\partial}{\partial y}y)\rho(t,x,y) + \frac{1}{2}(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2})\rho(t,x,y) + f\lambda(\rho(t,x_-,y_-) - 2\rho(t,x,y) + \rho(t,x_+,y_+))$$

where

$$x_\pm = x \mp (x\cos\theta + y\sin\theta)$$
$$y_\pm = x \mp (y\cos\theta - x\sin\theta)$$

and

$$f^{-1} = \det\begin{bmatrix} 1+\cos\theta & \sin\theta \\ -\sin\theta & 1+\cos\theta \end{bmatrix}$$

Example 5: Consider the piecewise linear stochastic equation of the Itô type

$$dx = f(x)dt + \sqrt{a}dw$$

where

$$f(x) = \begin{cases} -x+2 & x \geq 1 \\ x & |x| < 1 \\ -x-2 & x \leq -1 \end{cases}$$

Write down the Fokker-Planck equation. Find a steady state solution of the Fokker Planck equation by piecing together, in a continuous way, three Gaussian solutions and normalizing appropriately. (It will happen that one of "Gaussians " has a positive quadratic term in the exponent. This should not cause you difficulty however.) Show that as a approaches zero the steady state density tends, in a weak sense, to the sum of two delta functions and determine where the delta functions are centered.

Solution: The solution of the variance equation in the central region $|x| \leq 1$ is

$$\rho = be^{x^2/a}$$

In the side regions for which $|x| \geq 1$ we have

$$\rho = ce^{-(x-2)^2/a}$$

and

$$\rho = de^{-(x+2)^2/a}$$

By symmetry, $c = d$. Continuity at ± 1 gives an overall solution of the form

$$\rho_{ss}(x) = \begin{cases} ce^{-(x-2)^2/a} & x \geq 1 \\ ce^{-2/a}e^{x^2/a} & |x| < 1 \\ ce^{-(x+2)^2/a} & x \leq -1 \end{cases}$$

We must chose c so that the area under ρ_{ss} is one and so

$$c^{-1} = \int_{-\infty}^{-1} e^{-(x-2)^2/a}dx + \int_{-1}^{1} e^{-2/a}e^{x^2/a}dx + \int_{1}^{\infty} e^{-(x+2)^2/a}dx$$

Example 6: Suppose that x and y satisfy the Itô equation

$$\begin{bmatrix} dx \\ dy \end{bmatrix} = \begin{bmatrix} -adt & dw \\ -dw & -adt \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} b & 0 \\ 0 & b \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} dN$$

with $a = 1/2$ and

$$b = -1 \pm \frac{1}{2}\sqrt{4 + \frac{1}{x^2+y^2}}$$

Assuming that $x^2(0) + y^2(0) = 1$, describe the set of points that sample trajectories will visit. Hint: Use the Itô rule to compute $d(x^2 + y^2)$.

Solution: Applying the Itô rule to the function $\psi(x, y) = x^2 + y^2$ we get

$$d\psi(x, y) = 0dt + 0dw + [(x + bx)^2 - x^2 + (y + by)^2 - y^2]dN$$

which evaluates to

$$d\psi(x, y) = (x^2(2b + b^2) + y^2(2b + b^2))dN = dN$$

For the given definition of b this is of the form $d\psi = dN$. The solutions will only visit those circles centered at the origin that have radii which are positive integers.

3.8 Stochastic Approximation

There are various senses in which a sequence of real-valued random variables might be said to converge. It is useful to keep in mind that if $\{x_1, x_2, ...x_k, ...\}$ is a sequence of random variables then there is associated with each element of the sequence a different probability distribution, say $P_1, P_2, ...P_k, ...$ Suppose, that these distributiojns have densities, $p_1, p_2, ...p_k, ...$ The densities are L_1 functions mapping the real line into the positive half-line. If the sequence of random variables is to converge then the sequence of densities must in some, as yet to be described sense, "converge" to a delta function. Such an analysis requires a creful choice of topologies. On the otherhand, if we are content to consider convergence in the sense that

$$\lim_{k\to\infty} \int_R (x - x_0)^2 p_k(x) dx = 0$$

then convergence is much easier to deal with. The literature on stochastic approximation is usually concerned with discrete time models. there is usually a more or less obvious way to pass from a discrete-time description to a continuous-time discription. Here we only consider the latter.

Consider the following result on the asymptotic properties of the determninistic equation.

Lemma: Let a and b be integrable functions with b nonnegative. The solutions of

$$\dot{x}(t) = a(t)x(t) + b(t)$$

go to zero as t goes to infinity if

$$\lim_{t\to\infty} \int_0^t a(\sigma) d\sigma \to -\infty$$

and

$$\lim_{t\to\infty} \int_0^t b(\sigma) d\sigma < \infty$$

Proof: The solution of this equation is given by

$$x(t) = e^{\int_0^t a(\tau)d\tau} x(0) + \int_0^t e^{\int_\eta^t a(\tau)d\tau} b(\eta) d\eta$$

The asymptotic properties of a imply that the first term goes to zero. Because of the convergence of the integral of b, given any $\epsilon > 0$ there exists a time T such that

$$\int_T^\infty b(\tau) d\tau \leq \epsilon$$

Fot $t \geq T$ we have

$$x(t) = e^{\int_0^t a(\tau)d\tau} x(0) + \int_0^T e^{\int_\eta^t a(\tau)d\tau} b(\eta) d\eta + \int_T^t e^{\int_\eta^t a(\tau)d\tau} b(\eta) d\eta$$

The first two terms on the right go to zero and the third can be made arbitrarly small by choice of T.

Now consider the scalar stochastic equation

$$dw = a(t)x(t)dt + dw$$

It has an associated variance equation

$$\dot{\sigma}(t) = -2a(t)\sigma(t) + a^2(t)$$

Applying the lemma we see that if

$$\lim_{t\to\infty} \int_0^t 2a(\sigma)d\sigma \to -\infty$$

and

$$\lim_{t\to\infty} \int_0^t a^2(\sigma)d\sigma < \infty$$

then the variance goes to zero and we can assert that the random variable $x(t)$ converges to zero as t goes to infinity in mean square sense.

The main fact here is that a must go to zero at just the right rate if x is to go to zero. This sam analysis can be used much is different if x is replaced by $f(x)$ in the right-hand side of this equation, as long as $f(\cdot)$ is a first and third quadrent, monotone increasing function.

3.9 Exit Times

Consider the problem of determining the probability that the solution of

$$dx = f(x)dt + g(x)dw; x(0) \in S_1$$

leaves an open connected set $S \supset S_1$ before time t. Such problems arise in the analysis life expectancy of a machine, the time to finicancial ruin, etc. One way to formulate this is to consider a modified process which satisfies the given equation as long as $x \in S$ and satisfies $dx = 0$ once x reaches the boundary of S. The corresponding Fokker-Planck equation in the set S is

$$\frac{\partial \rho(t,x)}{\partial t} = -\sum_{i=1}^{n} \frac{\partial}{\partial x_i} f_i(x)\rho(t,x) + \sum_{i=1}^{n}\sum_{j=1}^{n} \tfrac{1}{2} \frac{\partial^2}{\partial x_i \partial x_j} g_i(x)g_j(x)\rho(t,x)$$

In terms of a physical picture we can imagine that the boundary of S absorbs the process and so, insofar as motion inside S is concerned, the appropriate boundary condition is $\rho(t,x) = 0$ for x on the boundary of S. To answer the original question then, we would need to solve for ρ subject to this boundary condition. The probability that x does not leave S on the internal $[0,t]$ is then just

$$p = \int_S \rho(t,x)dx$$

with the integral extending over the open set S.

Example: Suppose that we have a Gauss-Markov process

$$dx = -xdt + dw; x(0) = 0$$

and want to know the probability that $x(t)$ has not left the interval $[-\pi, \pi]$ over the period $0 \le t \le 1$. The Fokker-Planck equation is

$$\frac{\partial \rho(t,x)}{\partial t} = \frac{\partial}{\partial x} x\rho(t,x) + \tfrac{1}{2}\frac{\partial^2}{\partial x^2}\rho \quad ; \quad \rho(t,-1) = \rho(t,1) = 0$$

The nature of the boundary conditions suggest that we seek an even solution using an expansion in terms of trigonometric functions.

$$\rho(t,x) = \sum_{n=0}^{\infty} p_n(t) \cos nx$$

Standard seperation of variable techniques yield an equation for the individual p_n,

$$\dot{p}_n = p_n - n^2 p_n - \frac{1}{n} p_n$$

Because the probability that x has not left $[-\pi, \pi]$ in the interval $[0,t]$ is just $\rho_0(t)$ we see that

$$\text{prob } = e^{-t}$$

3.10 Computing Temporal Correlations

We can use the same multivariable method for computing temporal correlations that were discussed in section 2.7. If x satisfies the Itô equation

$$dx = -f(x)dt + g(x)dw$$

then for $\tau > t$, $x(t)x(\tau)$ satisfies

$$d_\tau x(t)x^T(\tau) = x(t)f^T x(\tau)d\tau + x(t)g^T(x(\tau))dw_\tau$$

Thus we can take expectation to get

$$\frac{d}{d\tau}\mathcal{E}x(t)x^T(\tau) = \mathcal{E}x(t)f^T(x(\tau))$$

This relationship can often be used with a formula for $\mathcal{E}x(t)x^T(t)$ to actually compute correlations.

3.11 Linear Equations

The combination of the Itô rule and the results of 3.5 on computing expectations gives an effective method for computing the statistical properties of solutions of stochastic equations of the form

$$dx = Axdt + \sum_{i=1}^{m} B_i x dw_i + \sum_{i=1}^{k} b_i dv_i$$

where w_i and v_i are independent Wiener processes. The special cases

$$dx = Axdt + \sum_{i=1}^{k} b_i dv_i$$

which involve only additive white noise terms have been used extensively as models for physical and economic systems. The observation that in terms of an enlarged vector $(1, x^T)^T$ we can write

$$d\begin{bmatrix} 1 \\ x \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & A \end{bmatrix}\begin{bmatrix} 1 \\ x \end{bmatrix}dt + \sum \begin{bmatrix} 0 & 0 \\ 0 & B_i \end{bmatrix}\begin{bmatrix} 1 \\ x \end{bmatrix}dw_i + \sum \begin{bmatrix} 0 & 0 \\ b_i & 0 \end{bmatrix}\begin{bmatrix} 1 \\ x \end{bmatrix}dv_i$$

means that it is never necessary to display the inhomogeneous term explicitly. For this reason we begin by developing the basic properties of

$$dx = Axdt + \sum_{i=1}^{m} B_i x dw_i$$

Fact 1: If x satisfies a linear sample path equation then

$$\frac{d}{dt}\mathcal{E}x(t) = A\mathcal{E}x(t)$$

Proof: Take expectations of both sides of (1). Use the fact that $\mathcal{E}g(x)dw = 0$.

Fact 2: If x satisfies a linear sample path equation then for each positive integer p the set of moments of order p satisfy a linear homogeneous differential equation.

Proof: We introduce the notation $x^{[p]}$, where x is an n-vector and p is a positive integer, to denote the $\binom{n+p-1}{p}$-component vector whose entries are the independent monomials homogeneous of degree p. That is,

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \;\; ; \;\; x^{[p]} = \begin{bmatrix} x_1^p \\ x_1^{p-1}x_2 \\ \vdots \\ x_n^p \end{bmatrix}$$

If $\dot{x} = Ax$, then it is easy to see that $x^{[p]}$ also satisfies a linear equation so that we have, for $A_{[p]}$ suitably defined,

$$\frac{d}{dt}x^{[p]} = A_{[p]}x^{[p]}$$

We take this as a definition of $A_{[p]}$. In applying the Itô rule to (3.1) we get

$$dx^{[p]} = A_{[p]}x^{[p]}dt + \sum B_{i[p]}x^{[p]}dw_i + \text{“Itô term”}$$

the Itô term denoting the contributions due to

$$\left\langle \frac{\partial^2 x^{[p]}}{\partial x_i \partial x_j} g_i \;,\; g_i \right\rangle dt$$

in using the Itô rule. Because the entries of $x^{[p]}$ are homogeneous of degree p, their second derivatives are homogeneous of degree $p-2$. Since $B_i x dw_i$ is linear in x, we see that the Itô term is homogeneous of degree p in $(x_1, x_2, \ldots, x_n)$ or, what is the same, linear in $x^{[p]}$. Thus after taking expectation, we get

$$\frac{d}{dt}\mathcal{E}x^{[p]} = L_{[p]}\mathcal{E}x^{[p]}$$

for some $L_{[p]}$. One can show that $L_{[p]}$ is given by

$$L_{[p]} = \left(A - \frac{1}{2}\sum_{i=1}^{m} B_i^2\right)_{[p]} + \frac{1}{2}\sum_{i=1}^{m}(B_{i[p]})^2$$

Fact 3: If x satisfies a linear sample path equation then for $\tau \geq 0$

$$\mathcal{E}x(t)x^T(t+\tau) = \mathcal{E}x(t)x^T(t)e^{A^T\tau}$$

Proof: It is easy to see that

$$\frac{d}{d\tau}\mathcal{E}x(t)x^T(t+\tau) = \mathcal{E}x(t)x^T(t+\tau)A^T$$

which is solved with initial condition $\mathcal{E}x(t)x^T(t)$.

3.12 Asymptotic Behavior

Under suitable hypothesis the probability density $\rho(t,x)$ associated with an Itô equation will approach a limit as t goes to infinity. This means that the process is approaching a "steady state" in the sense that certain statistical properties are approaching constant values. In many cases it is only this steady state which is experimentally observable and a great deal of the study of statistical physics and classical communication theory is based on steady state analysis.

For a one dimensional linear system the situation is as follows. If we have

$$dx = axdt + dw$$

then the 2nd moment equation is

$$\frac{d}{dt}\mathcal{E}x^2 = 2a\mathcal{E}x^2 + 1$$

and so

$$\mathcal{E}x^2(t) = e^{2at}\mathcal{E}\left(x^2(0) + \frac{1}{2a}\right) - \frac{1}{2a}$$

Thus we see that if a is positive the variance goes to infinity with increasing t whereas if a is negative it goes to $-1/2a$ with increasing t. In fact, since the response to $\rho(0,x) = \delta(x - x_0)$ is

$$\rho(t,x) = \frac{1}{\sqrt{2\pi\sigma(t)}}\, e^{-a(x-x_0e^{at})^2/2\sigma(t)}$$

we see that, regardless of the form of the initial distribution, ρ goes to

$$\rho(x) = \sqrt{-a/\pi}\, e^{ax^2}$$

as t goes to infinity provided that a is negative. In this case the situation may be summarized as follows: *If the nonrandom part of the stochastic equation is asymptotically stable then the density approaches a steady state as t goes to infinity.*

This same conclusion is valid in many situations. In particular, for linear systems of the form

$$dx = Axdt + \sum_{i=1}^{m} b_i dw_i$$

we have a differential equation for the second moment $\Sigma(t) = \mathcal{E}x(t)x^T(t)$ taking the form

$$\dot{\Sigma}(t) = A\Sigma(t) + \Sigma(t)A^T + b_1 b^T{}_1 + \cdots + b_m b_m^T$$

(To see this, apply the Itô rule taking advantage of the fact that the w_i are independent and the fact that the second derivative $\frac{\partial^2 x_n x_m}{\partial x_i \partial x_j}$ takes a simple form) This has a solution which may be expressed as

$$\Sigma(t) = \int_0^t e^{A(t-\sigma)}BB^T\, e^{A^T(t-\sigma)}d\sigma + e^{At}\Sigma(0)e^{A^Tt}$$

where $B = (b_1, b_2, \ldots, b_n)$. If the eigenvalues of A have negative real parts, then $\Sigma(t)$ approaches a limit as t goes to infinity. The density ρ also approaches a limit in this case

$$\lim_{t\to\infty} \rho(t,x) = \frac{1}{\sqrt{(2\pi)^n \det \Sigma_\infty}}\, e^{-\frac{1}{2}x^T \Sigma_\infty^{-1} x}$$

where Σ_∞ satisfies

$$A\Sigma_\infty + \Sigma_\infty A^T + BB^T = 0$$

or

$$\Sigma_\infty = \int_0^\infty e^{At} BB^T e^{A^T t} dt$$

Example: Solve the Fokker-Planck equation for

$$\begin{aligned} dx_1 &= x_2 dt \\ dx_2 &= -x_1 dt + dw \end{aligned}$$

In this case

$$\begin{aligned} \mathcal{E}x(t) &= e^{At}\mathcal{E}x(0) \\ &= \bar{x}(t) \end{aligned}$$

and $\Sigma(t)$ is given by

$$\Sigma(t) = \int_o^t \begin{bmatrix} \cos(t-\sigma) & \sin(t-\sigma) \\ -\sin(t-\sigma) & \cos(t-\sigma) \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \cos(t-\sigma) & -\sin t(t-\sigma) \\ \sin(t-\sigma) & \cos(t-\sigma) \end{bmatrix} d\sigma$$

$$\Sigma(t) = \int_o^t \begin{bmatrix} \sin^2(t-\sigma) & \sin(t-\sigma)\cos(t-\sigma) \\ \sin(t-\sigma)\cos(t-\sigma) & \cos^2(t-\sigma) \end{bmatrix} d\sigma$$

With these definitions $\rho(t,x) = \frac{1}{\sqrt{(2\pi)^2 \det \Sigma(t)}}\, e^{-\frac{1}{2}\left(x-\bar{x}(t)\right)^T \Sigma^{-1}(t)(x-\bar{x}(t))}$

3.13 Finite Difference Approximations

For reasons having to do with numerical simulation as well as theoretical questions we discuss here the connection between stochastic differential equations and difference equation approximations.

Consider the difference equation

$$x((k+1)T) = x(kT) + Tf(x(kT)) + Tg(x(kT))n(kT)$$

If we think of this as an approximation to a differential equation, the question of refining the mesh size becomes of interest. We begin by comparing

$$x(T) = x(0) + T\, f\big(x(0)\big) + T\, g\big(x(0)\big)n(0)$$

with the value of $x(T)$ that results if we replace T by $T/2$, and iterate. In the latter situation we get at the first step

$$x(T/2) = x(0) + (T/2)f\big(x(0)\big) + (T/2)g\big(x(0)\big)n(0)$$

and, repeating this,

$$\begin{aligned} x(T) \quad = \quad & x(0) + (T/2)f\big(x(0)\big) + (T/2)g\big(x(0)\big)n(0) \\ & + (T/2)f\big(x(T/2)\big) + (T/2)g\big(x(T/2)\big)n(T/2) \end{aligned}$$

If we expand f and g in their Taylor series we can express this in terms of $x(0)$

$$\begin{aligned} x(T) &= x(0) + Tf\big(x(0)\big) + T/2g\big(x(0)\big)n(0) + T/2g\big(x(0)\big)n(T/2) \\ &\quad +(T/2)^2 g'\big(x(0)\big)g\big(x(0)\big)n(0)n(T/2) + \dots \end{aligned}$$

It is useful to focus on the variance associated with $x(T)$ in these two cases. Before refinement, we have

$$\mathcal{E}\big(x(T)-x(0)\big)\big(x(T)-x(0)\big)^T = T^2 g\big(x(0)\big)g^T\big(x(0)\big)\mathcal{E}n^2(0)$$

and after refinement

$$\mathcal{E}\big(x(T)-x(0)\big)\big(x(T)-x(0)\big)^T = T^2/4 g\big(x(0)\big)g^T\big(x(0)\big)\mathcal{E}n^2(0) + n^2\big(T/2\big)\big)$$

Thus, if we wish the variance of the solution to stay the same when we divide the interval in half, we must double the variance of the noise. For this reason it is a good idea to adopt the notation

$$x\big((k+1)T\big) = x(kT) + Tf\big(x(kT)\big) + \sqrt{T}\, g\big(x(kT)\big)n(kT)$$

In such a notation the variance of the noise term $n(kT)$ can remain constant as T is reduced.

We remark that if $\psi(\cdot)$ is a twice differentiable function of x, then

$$\psi\Big(x\big((k+1)T\big)\Big) = \psi\Big(\big(x(kT)\big) + Tf\big(x(kT)\big) + \sqrt{T}\, g\big(x(kT)\big)n(kT)\Big)$$

which yields

$$\begin{aligned} \psi\Big(x\big((k+1)T\big)\Big) &= \psi\big(x(kT)\big) + T\Big\langle \frac{\partial \psi}{\partial x}\ ,\ f\big(x(kT)\big)\Big\rangle + \sqrt{T}\,\Big\langle \frac{\partial \psi}{\partial x}\ ,\ g\big(x(kT)\big)\Big\rangle n(kT) \\ &\quad + \frac{T}{2}\Big\langle \frac{\partial^2 \psi}{\partial x^2} g\ ,\ g\Big\rangle n^2(kT) + \text{h.o.t.} \end{aligned}$$

This is fully consistent with the Itô differential rule.

3.14 Exercises 3

1. Consider the scalar Itô equation

$$dx = -xdt - \alpha x dw + \beta x d\nu \quad ; \quad x(0) = 0$$

where w and ν are independent Brownian motions. Derive an equation for the p^{th} moment of x and give conditions on α and β for the p^{th} moment to have a finite limit as t goes to infinity.

2. If x_1 and x_2 satisfy the Itô equations

$$\begin{aligned} dx_1 &= dw_1 \\ dx_2 &= dw_2 \end{aligned}$$

and if we make a change of variables $\theta = \tan^{-1}(x_2/x_1)$, $r = \big((x_1)^2 + (x_2)^2\big)^{\frac{1}{2}}$, then find an expression for the probability density $\rho(t,r,\theta)$ assuming that the initial density is rotationally symmetric. If $x_1(0) = x_2(0) = 0$ find the probability that $x_1^2(t) + x_2^2(t) \le a$.

3. Consider the stochastic differential equation

$$\begin{bmatrix} dx_1 \\ dx_2 \end{bmatrix} = \begin{bmatrix} -\frac{1}{2}dt & dw \\ -dw & -\frac{1}{2}dt \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

Show that $x_1^2(t) + x_2^2(t) =$ constant. Show that if $\theta = \tan^{-1}(x_1/x_2)$, then

$$d\theta = dw$$

Show that in terms of θ the probability density satisfies

$$\frac{\partial \rho(t,\theta)}{\partial t} = \frac{\alpha}{2}\frac{\partial^2 \rho(t,\theta)}{\partial \theta^2}$$

for a suitable value of α. Solve this equation for ρ on $-\pi \leq \theta \leq \pi$ given that $\theta(0) = 0$.

4. If w is a Wiener process defined for $0 \leq t \leq 1$ and if $dx = dw$ with $x(0) = 0$ evaluate

$$m(t) = \mathcal{E}e^{x^2(t)}$$

5. Consider, again, the polar coordinate representation of two-dimensional Wiener process as developed in example two of section 3.4.

$$dr = \frac{1}{r}dt + \sin\theta d\nu + \cos\theta dw$$

$$d\theta = \frac{\cos\theta d\nu - \sin\theta dw}{r} + \frac{\sin\theta\cos\theta}{r^2}dt$$

The corresponding Fokker-Planck equation makes apparent the rotational symmetry present.

$$\frac{\partial \rho(t,r,\theta)}{\partial t} = \frac{1}{2}\left(\frac{1}{r}\frac{\partial}{\partial r} + \frac{\partial^2}{\partial r^2} + \frac{1}{r^2}\frac{\partial^2}{\partial \theta^2}\right)\rho(t,r,\theta)$$

The term in parentheses on the right is just the Laplacian expressed in polar coordinates. Show that this equation admits the rotationally symmetric solution

$$\rho(t,r,\theta) = \frac{r}{2\pi t}e^{-r^2/2t}$$

6. Consider the Itô equation

$$\begin{aligned} dx &= dw_1 \\ dy &= dw_2 \\ dz &= x dw_2 - y dw_1 \end{aligned}$$

Show that

$$\frac{d}{dt}\mathcal{E}z^2(t) = \mathcal{E}(x^2(t) + y^2(t)) = 2t$$

and that

$$\frac{d}{dt}\mathcal{E}z^2(t)(x^2(t) + y^2(t)) = \mathcal{E}(x^2(t) + y^2(t))^2 + \mathcal{E}z^2(t)$$

7. Consider the Itô equations for x, y, z in the previous problem. Show that the corresponding Fokker-Planck Equation is

$$\frac{\partial \rho(x,t)}{\partial t} = \frac{1}{2}\left(\frac{\partial}{\partial x} - y\frac{\partial}{\partial z}\right)^2\rho(t,x) + \frac{1}{2}\left(\frac{\partial}{\partial y} - x\frac{\partial}{\partial z}\right)^2\rho(t,x)$$

Show that for each real value of λ the expression $\rho(t,x,y,z) = e^{-2\lambda t}e^{-\lambda(x^2+y^2)}\cos 2\lambda z$ satisfies this equation. It does not have a constant integral with respect to the variables x, y, z. Discuss the possibility of forming a weighted combination of solutions of this type to obtain a solution which remains normalized with growing values of t.

8. Consider the Itô equations

$$dx = xdt + (zx/\sqrt{x^2+y^2})dw + zy\sqrt{x^2+y^2}\,dt$$
$$dy = -ydt - (zy/\sqrt{x^2+y^2})dw - zx\sqrt{x^2+y^2}\,dt$$
$$dz = -z\sqrt{x^2+y^2}\,dt$$

Show that as t goes to infinity

$$\mathcal{E}\,x^2 \to 0$$
$$\mathcal{E}\,y^2 \to 0$$

and

$$z(t) \to 0$$

9. In the theory of linear electrical circuits one models a resistor-inductor circuit (following Nyquist and Johnson) via an Itô equation $L\,di = -Ri\,dt + \sqrt{kRT}\,dw$ where T is the temperature with respect to an absolute scale and k is a suitable constant (Boltzman's constant). Determine the expected value of the energy stored in the inductor in steady state.

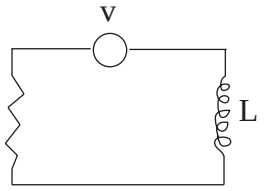

Figure 2.1. An electrical circuit with a Nyquist-Johnson resistor.

10. Consider the Wiener process x

$$dx = dw \quad ; \quad x(0) = 0$$

Let t be a positive number and suppose that it is known that $x(t) = 0$. For $0 \le \tau \le t$ find the probability density for $x(\tau)$ conditioned on the fact that $x(t) = 0$. What is the probability density for

$$n = \int_0^t x(\sigma)d\sigma$$

conditioned on the fact that $x(t) = 0$?

11. Suppose that x_1 and x_2 satisfy the differential equations

$$dx_1 = x_2 dt$$
$$dx_2 = (-x_1 - x_2)dt + dw$$

Suppose that the process has reached steady state. Show that the probability that $x(t)x(t+\tau)$ is positive is

$$p(\tau) = \frac{1}{\pi}\tan^{-1}\alpha$$

with α being derived from the autocorrelation function.

12. Given that

$$dx = ydt$$
$$dy = \alpha x dt + \beta y dt + d\omega$$

for what values of α and β will $\rho(t,x,y)$ have a limiting value as t goes to infinity? Suppose that $\beta = -3$ and $\alpha = -2$, find

$$\lim_{t\to\infty} \mathcal{E}y(t)y(t+\tau)$$

13. Compute the probability that the solution of

$$\begin{aligned} dx_1 &= dw_1 \quad ; \quad x_1(0) = 0 \\ dx_2 &= dw_2 \quad ; \quad x_2(0) = 0 \end{aligned}$$

lies in the disk $\{x_1, x_2 | x_1^2 + x_2^2 \leq 1\}$ at $t = 1$. Evaluate the expected value of

$$m(t) = e^{x_1(t)x_2(t)}$$

14. Consider the modified diffusion equation

$$\frac{\partial \rho}{\partial t} = \frac{1}{2}\frac{\partial^2 \rho}{\partial x^2} - \rho - x\frac{\partial \rho}{\partial x}$$

Find an Itô equation of the form

$$dx = f(x)dt + g(x)dw$$

such that the above diffusion equation describes the probability density. Solve for the density given $x(0) = 0$

15. Consider the equations

$$\begin{aligned} dx &= xdt + dw - dt \quad ; \quad x(0) = (0) \\ dy &= xydt \quad ; \quad y(0) = 1 \end{aligned}$$

Find the probability density $\rho(t, x, y)$?

16. Evaluate $\mathcal{E}x^2(t)$ for x satisfying

$$\begin{aligned} dz &= -2zdN \\ dx &= (\alpha + z)xdt + dw \end{aligned}$$

Here N is a Poisson counter of rate λ and w is a standard Wiener process.

17. Consider a system of the form

$$\begin{aligned} dx &= (-2zx - x)dt + dw \\ dz &= -2zdN \; ; \; z(0) \in \{\pm 1\} \end{aligned}$$

with N a Poisson Counter of rate λ. The x equation is "unstable" with growth e^t when $z = -1$ but is stable with decay e^{-3t} when $z = +1$. of course $z = +1$ and $z = -1$ are equally likely and using the methods of section 3.5 we see that

$$\begin{aligned} \frac{d}{dt}\mathcal{E}x^2 &= -4\mathcal{E}zx^2 - 2\mathcal{E}x^2 + 1 \\ \frac{d}{dt}\mathcal{E}zx^2 &= -4\mathcal{E}x^2 - 2\mathcal{E}zx^2 - 2\lambda\mathcal{E}zx^2 + \mathcal{E}z \end{aligned}$$

Show that the solutions of these equations approach a constant regardless of the initial conditions.

18. Consider the pair of equations

$$\begin{aligned} dx_1 &= -2x_1 dN \quad ; \quad x_1(0) \in \{-1, 1\} \\ dx_2 &= -10x_2 dt + x_1 dt + dw \end{aligned}$$

with N being a Poisson counter of rate λ and w being a Weiner process. Compute

$$\phi(t, \tau) = \mathcal{E}x_2(t)x_2(t + \tau) \quad ; \quad \tau \geq 0$$

Does there exist a limiting value for $\mathcal{E}x_1^2(t)$ and $\mathcal{E}x_2^2(t)$ as t goes to infinity?

19. Compute the autocorrelation function for the steady state of $x(t)$ where

$$\begin{aligned} dx(t) &= -x(t)dt + z(t)x(t)dt + dw \\ dz &= -2zdN \quad ; \quad z(0) = 1 \end{aligned}$$

with N being a Poisson counter of rate λ.

20. Consider the pair of stochastic differential equations

$$\begin{aligned} dx &= zxdt + dw \\ dz &= -2z\, dN \quad ; \quad z(0) \in \{-1, +1\} \end{aligned}$$

where N is a poisson counter of rate λ. The Itô rule gives

$$d\psi(x,z) = \frac{\partial \psi}{\partial x}\, zx\, dt + [\psi(x,-z) - \psi(x,z)]\; dN + \frac{1}{2}\frac{\partial^2 \psi}{\partial x^2}\, dt + \frac{\partial \psi}{\partial x} dw$$

Assume that the initial density of x exists and that a density exists for all positive time, say $\rho\colon \mathbb{R}^+ \times \mathbb{R} \times \{1,-1\} \to \mathbb{R}$.

Writing $\rho(t,x,1)$ as $\rho_+(t,x)$ and $\rho(t,x,-1)$ as $\rho_-(t,x)$ show that

$$\begin{aligned} \frac{\partial \rho_+(t,x)}{\partial t} &= \frac{1}{2}\frac{\partial^2 \rho_+(t,x)}{\partial x^2} - \frac{\partial x \rho_+(t,x)}{\partial x} + \lambda\big(\rho_-(t,x) - \rho_+(t,x)\big) \\ \frac{\partial \rho_-(t,x)}{\partial t} &= \frac{1}{2}\frac{\partial^2 \rho_-(t,x)}{\partial x^2} + \frac{\partial x \rho_-(t,x)}{\partial x} + \lambda\big(\rho_+(t,x) - \rho_-(t,x)\big) \end{aligned}$$

21. Consider the pair of Itô equations

$$\begin{aligned} dx &= z(t)xdt + dw \quad ; \quad x(0) \in \mathbb{R}^1 \\ dz &= -2z(t)dN \quad ; \quad z(0) \in \{\pm 1\} \end{aligned}$$

where w is a standard Wiener process and N is a Poisson counter of rate λ. Find the Fokker-Planck equation for this system and express it in terms of two coupled diffusion equations. Does there exist a limiting value as $t \to \infty$ for $\mathcal{E}x^2(t)$?

22. (Heat baths and equipartition of energy) Let x take on values in $\mathbb{R}^n$, let S be a real skew-symmetric matrix and let ϵ be a positive number. Consider the Itô equation

$$dx = (S - \epsilon GG^T)xdt + \sqrt{\epsilon} G dw$$

Assume that $(G, SG, \ldots, S^{n-1}G)$ has rank n. In steady state the variance satisfies

$$\Sigma(S - \epsilon GG^T)^T + (S - \epsilon GG^T)\Sigma = -\epsilon GG^T$$

Show that $\Sigma = (1/2)I$ is the unique solution of this equation for all $\epsilon > 0$. Now consider

$$\begin{bmatrix} dx \\ dz \end{bmatrix} = \begin{bmatrix} S - GG^T & B \\ -B^T & \Omega \end{bmatrix} \begin{bmatrix} x \\ z \end{bmatrix} dt + \begin{bmatrix} Gdw \\ 0 \end{bmatrix}$$

Evaluate the steady state variance assuming $\Omega = -\Omega^T$ and rank $(B, \Omega B, \ldots, \Omega^k B) = k$, where $k =$dim $x+$dim z.

NOTES AND REFERENCES

[1] Gaussian distributions are also called normal distributions. Most applied mathematical treatments of partial differential equations do not solve the heat equations in quite enough generality for our purposes, but the basic ideas are widely discussed. For example, see

1. I. Stakgold, *Partial Differential Equations*, Van Nostrand, New York, 1991.

[2–3] This kind of limiting approach starting from Poisson process was worked out by Paul Levy around 1940. A more recent and highly readable account can be found in:

1. H. P. McKean, *Stochastic Integrals*, McGraw Hill, New York, 1969.

Alternative points of view are explored in:

1. E. Wong, *Stochastic Process in Dynamical Systems and Information Theory*, McGraw Hill, New York, 1969.

[4] Notice that we can keep the same point of view as was used in Chapter 2.

[5] The Fokker-Planck equation plays a fundamental role in many physical problems. It is also the point of departure for the study of the very important conditional density equations to be taken up in Chapter 6.

[6] The original papers on stochastic approximations are H. Robbins and S. Monro, "A Stochastic Approximation Method,", *Annals of Mathematical Statistics*, vol. 22, pp. 400-407, 1951. and J. Kiefer and J. Wolfowitz, "Stochastic Estimation of the Minimum of a Regression Function,", *Annals of Mathematical Statistics*, vol. 23, pp. 462-466, 1952. An early applicationin control is given in H. J. Kushner, "Hill Climbing Methods for the Optimization of Multi-Parameter Noise Disturbbed Systems," 1962 JACC.

A recent reference dealing with the stability of stochastic differential equations is: L. Arnold

[7] A classic text on stationary processes, including material on the expected number of zero-crossings is: Harald Cramér and M. R. Leadbetter, *Stationary and Related Stochastic Processes* John Wiley, New York, 1966.

[8] The exercises involving coupled diffusion equations are roughly analogous to problems in quantum mechanics involving spin. See, for example,

1. Gordon Baym, Lectures on Quantum Mechanics, University of Illinois, W. A. Benjamin, Inc., New York, 1969.
2. Eugene Gutts, "Brownian Motion and Indeterminacy Relations", in *Stochastic Processes in Chemical Physics*, (K.E. Shuker, Ed.), Interscience, New York, 1969.

Chapter 4

Pseudorandom Processes

4.1 Pseudorandom Number Generators

A deterministic process, such as the calculation of the decimal expansion of π, may produce a sequence of digits that appears to be random if one checks the relative frequency of occurrence of a particular digit or pair of digits even though there may be sophisticated tests powerful enough to reveal the deterministic origin of the sequence. Do there exist "universal" tests for randomness and what tests would one use to attempt to use to discover if a sequence is truly random? Are all points in the sample space represented with the right frequency? Are there correlations between successive points? Any finite list of tests could be extended by considering more complex tests. The sequence of digits of π would fail certain highly nonlinear tests for randomness but in many applications these tests would be irrelevant.

The generation by computer of "random" numbers and sequences of random numbers is obviously quite important for simulation and even for some types of computation. Most high level computer languages make available random number generators that produce an empirical distribution appearing to be consistent with the uniform distribution on [0,1]. Of course the real situation is a bit different. The random numbers produced by computer are all rational and, far from being independent. Successive elements in a sequence are related by a deterministic rule.

Consider the equation

$$x(k+1) = (kx(k) + b) \bmod 1$$

If k is large, say about 10^5, then multiplication by k can be thought of as shifting the decimal point of x five places to the right. The addition of b further rearranges matters and reducing modulo one serves to make everything to the left of the decimal point irrelevant. Thus successive terms in the sequence, while deterministically related, are related in such a highly nonlinear way that insofar as linear correlation is concerned they may appear to be independent, depending on the exact values of $x(0), k$ and b.

4.2 Uniform versus Gaussian Distributions

According to the central limit theorem, under rather weak hypothesis, the sum of n independent, identically distributed zero mean, variance σ, random variables has a density which when suitably scaled tends to a Gaussian with increasing n. That is, the density for the random variable y_N

$$y_N = \sum_{i=1}^{N} \frac{x_i}{\sigma\sqrt{n}}$$

tends to a gaussian of unity variance as N goes to ∞. Expressed in terms of convolution, we have for a unity variance distribution

$$g(y_2) = f(x) * f(x)$$

as the density of

$$y_2 = (x_1 + x_2)$$

Thus the density for y_n, in the limit as n goes to infinity, is

$$\lim_{n\to\infty} \underbrace{f(\frac{x}{\sqrt{n}}) * f(\frac{x}{\sqrt{n}}) * \cdots * f(\frac{x}{\sqrt{n}})}_{n \text{ factors}} = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$$

One can investigate how this works using the density $f(x) = 1$ for $|x| \leq \frac{1}{2}$ and zero otherwise. (In this case the variance is $\frac{1}{12}$.) The number of terms required to get a good approximation depends on the circumstances. Naturally the probability that $(x_1 + x_2 + \cdots + x_n)$ exceeds $n/2$ is zero. Thus this remains of limited validity even for n large. However, near the origin, say $|x| \leq 3$ the match is already quite good for $n = 20$.

4.3 The Results of Jacobi and Bohl, Sierpinskii and Weyl

In this section we sketch two results on deterministic processes.

Theorem 1: Let x be a scalar, $0 \leq x < 1$, and ω a real number. If $f(x) = (x + \omega) \bmod 1$ then the orbits of $x(k+1) = f(x(k))$ are dense in $[0, 1]$, if and only if ω is irrational.

Proof: If $\omega = \frac{q}{p}$, then $x(p) = x(0) + \frac{pq}{p}) \bmod (1) = x(0)$ and so $x(p) = x(0)$. Thus all motions are periodic. If w is irrational then $f^n(x) \neq f^m(x)$ for m and n distinct integers; if not

$$x + n\omega = x + m\omega$$

which means $(n - m)\omega$ is an integer. Thus each orbit contains an infinity of points in $[0, 1]$. There must be at least one limit point. That is, given $1/2 > \epsilon > 0$ there exists $m \neq n$ such that $|f^n(x) - f^m(x)| \leq \epsilon$. This means

$$|f^{(n-m)}(x)| \leq \epsilon$$

because $|f(x) - f(y)| = |x - y|$. (Recall that 0 and 1 are to be thought of as being the same point.)

Let $n-m=q$. If $f^q(x)-x$ is positive, then $f^{2q}(x)-x$ is also. We observe that the sequence $x, f^q(x), f^{2q}(x), \ldots$ satisfies

$$|f^{rq}(x) - f^{sq}(x)| > \epsilon$$

for r and s integers between 0 and $\frac{1}{\epsilon}$. The same is true of $f^q(x)-x$ negative. But ϵ is an arbitrary positive number so the sequence $f^r(x)$ is dense in $[0,1]$.

Vector Version: Suppose $x(k+1)=x(k)+\omega$ with $x(k)\in\mathbb{R}^n$. In this case $x(k)$ winds densely if $k\cdot\omega\in\mathbb{Z}$ and $k\in\mathbb{Z}^n$ implies $k=0$.

The following theorem is often credited to Bohl, Sierpinskii and Weyl.(See [4].) It shows that the process described in Theorem 1 generates a uniform distribution if ω is irrational. It is formulated in terms of a rotation on the circle S^1.

Theorem 2: If $f: S^1 \to S^1$ is a rotation of the circle through an angle ω, if ω is incommensurate with 2π, and if ψ is a Riemann integrable function, then

$$\lim_{N\to\infty} \frac{1}{N}\sum_{k=1}^{N} \psi(f(x)) = \int_{S^1} \psi(x)\ dx$$

Proof: (This is only a partial proof.) For the sake of making the steps in the proof more transparent, we transform the problem to the unit circle. Write $z=e^{2\pi ix}$ and $\theta = e^{2\pi i\omega}$. We study the difference equation $z(k+1)=e^{2\pi i\omega}z(k)$. Now

$$\begin{aligned} \tfrac{1}{p}\textstyle\sum_{k=0}^{p-1}\left(e^{2\pi ik}z(0)\right)^k &= \tfrac{1}{p}\textstyle\sum_{k=0}^{p}\theta^k z(0)^k \\ &= \begin{cases} 1 & p=0 \\ \\ \frac{1}{p}z^p & \frac{\theta^{np}-1}{\theta^p-1} \end{cases} \end{aligned}$$

If $\psi(z)=\sum_{i=1}^{k} a_i z^i$ then we just add up the above. This gives the result when ψ is a polynomial in $e^{2\pi ix_0}$. We omit the demonstration that any Riemann integrable function can be approximated by such polynomials with sufficient precision so as to complete the proof.

4.4 Other Difference Equations

There are completely deterministic procedures for determining the decimal expansion of π. Yet if we considered a discrete time stochastic process $x(t)$ taking on values in the set {0,1, ... ,9} with uniform distribution and independent trials, we would have great difficulty in designing a test which would distinguish the above expansion from a sample function of the stochastic process. This is but one of many procedures for generating what we might call pseudorandom processes. Recently it has become a very popular game to invent deterministic schemes for generating processes whose behavior looks stochastic. This may be done using difference equations or differential equations. Because rather less is known about the differential equation case we begin with a brief discussion of difference equations.

Perhaps the best known pseudorandom difference equation is the so-called "logistic" equation

$$x(i+1) = 1 - k\big(x(i) - \frac{1}{2}\big)^2$$

considered on the interval $0 \leq x \leq 1$ and for $0 \leq k \leq 1$. In this range of values we may distinguish between several types of behavior according to

(a) $0 < k < \frac{1}{4}$
(b) $\frac{1}{4} < k <\sim .91$
(c) $\sim .91 < k < 1$

In the first case all solutions approach an equilibrium point. In the second region there are many periodic solutions – more as k gets larger. In the last case the motion appears to be quite chaotic indeed.

If $f\colon [0,1] \to [0,1]$ is such that for each value of y there is a finite number of values x such that $f(x) = y$, and if f is differentiable with a derivative that is nonzero except for a finite number of points, then we can consider the equation

$$\rho(x) = \sum_{\substack{\text{inverse} \\ \text{images}}} \frac{1}{|f'(x)|} \rho(f^{-1}(x))$$

This is the functional equation that a steady state density for x would need to satisfy if it were to exist.

Example 1: Let $f(x) = 4x - 4x^2$. In this case $f(f(\dots f(x) \dots))$ is a polynomial of degree 2^n where n is the number of compositions. Clearly this has at most 2^n distinct periodic solutions of period n because a polynomial of degree 2^n has, at most, 2^n real roots. It is easy to see that if all the roots are real and distinct then there are 2^n distinct periodic solutions of period n.

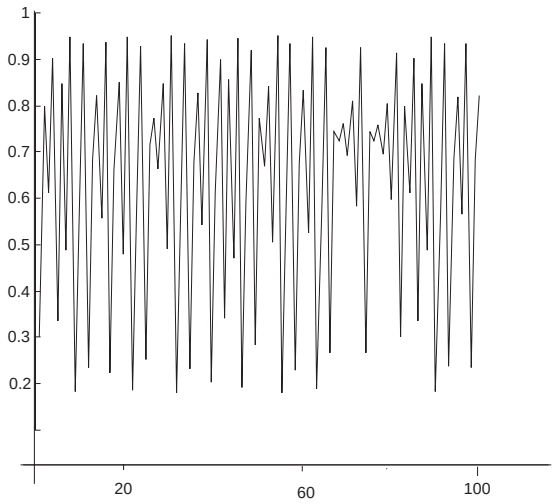

Figure 4.3.Trajectory of the logistic equation $x(k+1) = 4x(k) - 4x^2(k)$.

The functional equations that the steady state invariant density would need to satisfy can be found as follows. From

$$y = 4x - 4x^2$$

we see that

$$(y-1) = -(2x-1)^2$$

or

$$x = \tfrac{1}{2}(1 \pm \sqrt{1-y})$$

thus

$$\rho(y) = \frac{1}{\sqrt{1-y}}\big(\rho(\tfrac{1}{2}+\tfrac{1}{2}\sqrt{1-y}) + \rho(\tfrac{1}{2}-\tfrac{1}{2}\sqrt{1-y})\big)$$

Example 2: Let f be defined by

$$f(x) = \begin{cases} 2x: & 0 \le x \le \frac{1}{2} \\ 2-2x: & \frac{1}{2} \le x \le 1 \end{cases}$$

then the difference equation

$$x(k+1) = f(x(x))$$

gives rise to the invariant measure equation

$$\rho(x) = \tfrac{1}{2}\rho(x/2) + \tfrac{1}{2}\rho\big((1-x)/2\big)$$

Clearly this admits the solution $\rho(x) \equiv 1$. Value $x(0) = a$ gives rise to a periodic solution of period n if

$$f(f(\dots f(a)\dots)) = a$$

If $n = 2$ and $a < 1/2$ this is

$$2 - 4a = a$$

4.5 Differential Equations

We have seen that simple difference equations of the form $x(k+1) = f(x(k))$ can have solutions that are highly unstable but yet are not unbounded. Differential equations can exhibit the same type of behavior but unless time enters explicitly they must be of dimension 3 or more.

Consider the third order equation

$$x^{(3)} + x^{(2)} + 1.25x^{(1)} - 1.8\sin(x) = 0$$

Figure 4.3 shows a trajectory in $(x, \dot{x})$-space.It is confined to a relatively small region of the space but yet it does not appear to be periodic.

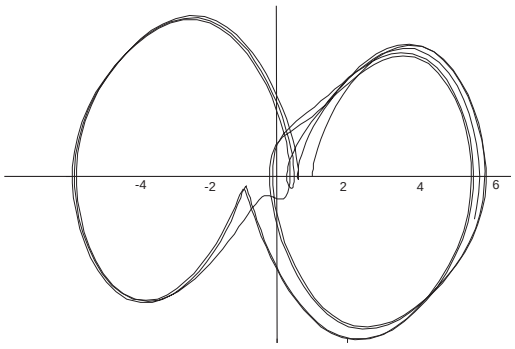

Figure 4.3.Trajectory of the third order equation.

4.6 Exercises 4

1. Any given a number in $[0, 1)$ has a binary expansion $x = .a_1a_2a_3\cdots$. Of course a_1 is one if $x \geq 1/2$, a_2 is one if $x - a_1 \geq 1/4, \ldots$ etc. Describe the process of generating the a_i as a pair of scalar difference equations by introducing $b_i = x - a_1, a_2, \ldots, a_{i-1}$ then finding g_1 and g_2 so that

$$a_{i+1} = g_1(a_i, b_i)$$
$$b_{i+1} = g_2(a_i, b_i)$$

4.7 Notes and References

1. S. K. Park and K. W. Miller, "Random Number Generators: Good ones are hard to find," *Comm. A.C.M.*, vol. 32, (1988), pp. 1192-1201.

2. Gérard Iooss and Daniel D. Joseph, *Elementary Stability and Bifurcation Theory*, Springer Verlag, New York, 1980.

3. A. J. Lictenberg and M. J. Lieberman, *Regular and Stochastic Motion*, Springer-Verlag, New York, 1983.

4. J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems and Bifurcation of Vector Fields*, Springer-Verlag, New York, 1983.

5. V.I. Arnold and A. Alvez, *Ergolic Problems of Classical Mechanics*, Benjamin, Reading, MA, 1969.

6. Edward Ott, *Chaos in Dynamical Systems*, Cambridge University Press, Cambridge, England, 1993

Chapter 5

System Concepts

This chapter marks the transition from pure stochastic processes to the subject of stochastic control. We assemble here a few control and system theoretic tools that will be used in the later chapters. We are interested in modeling systems that have control terms (inputs), stochastic terms, and explicitly observable terms (outputs). One of the basic problems we will discuss is that of finding a stochastic differential equation whose solutions have certain specified statistical properties. A second basic problem is that of modeling energy flow in systems with inputs and outputs.

5.1 Deterministic Linear Systems

A deterministic, continuous time, differential equation based, input-output model takes the form

$$\dot{x} = f(x, u, t); \quad y = h(x, t)$$

A stochastic, differential equation based, input-output model takes the form

$$dx = f(x, u, t)dt + \sum_{i=1}^{k} g_i(x, u)dw_i + \sum_{j=1}^{l} \tilde{g}_i(x, u)dN_i \ ; \ dy = h(x)dt + i(x)d\nu$$

Mathematical models of the form

$$\dot{x} = Ax + Bu \ ; \ y = Cx$$

with x, u, and y vector valued are called *linear systems.* If A, B, and C are constant then it is said to be a *time invariant linear system.* If $\Phi(t, \sigma)$ satisfies the equations

$$\frac{d}{dt}\Phi(t, \sigma) = A(t)\Phi(t, \sigma) : \Phi(\sigma, \sigma) = I$$

then Φ is said to be a fundamental solution of $\dot{x} = Ax$. If A is constant then $\Phi(t, \sigma)$ is just $e^{A(t-\sigma)} = I + A(t - \sigma) + A^2(t - \sigma)^2/2 + \cdots$. The so-called *variation of constants formula* gives y in terms of $x(\sigma)$ and u

$$y(t) = C(t)\Phi(t, \sigma)x(\sigma) + \int_{\sigma}^{t} C(t)\Phi(t, \sigma)B(\sigma)u(\sigma)d\sigma$$

The function T defined by $T(t,\sigma) = C(t)\Phi(t,\sigma)B(\sigma)$ is sometimes called the *weighting pattern*. If A, B, and C are constant then it takes the form

$$T(t,\sigma) = Ce^{A(t-\sigma)}B$$

with e^{At} being defined by the (convergent) series

$$e^{At} = I + At + A^2t^2/2! + \cdots$$

A linear system is said to be *controllable* if for any given value of $x(\sigma)$ and any given $t > \sigma$ there is a control $u(\cdot)$ defined on $[\sigma, t]$ such that $u(\cdot)$ drives the system to the state zero at time t. The matrix

$$W(\sigma,t) = \int_\sigma^t \Phi(t,\sigma)B(\sigma)B^T(\sigma)\Phi^T(t,\sigma)d\sigma$$

is called the *controllability gramian* for the interval $[\sigma, t]$. Notice that $W(\sigma,t)$ is automatically nonnegative definite. An application of the control $u(t) = B^T(t)\Phi^T(t,\sigma)W^{-1}(t,\sigma)x(\sigma)$ shows that if $W^{-1}(t,\sigma)$ exists then the system is controllable. If it is positive definite then the system is controllable. If A and B are constant then $W(\sigma,t)$ is positive definite if and only if the matrix $(B, AB, \ldots, A^{n-1}B)$ (commas denote column partition) has rank equal to $n = \dim x$.

The problem of observability relates to the possibility of determining the value of $x(t)$ from a knowledge of $y(\cdot) = cx(\cdot)$ over some interval $[t_0, t_1]$. The answer to this type of question can also be reduced to the investigation of the definiteness of a symmetric matrix. Consider the expression

$$\Phi^T(t_1,t_0)c^T(t)y(t) = \int_{t_0}^t \Phi^T(t_1,t_0)c^T(t)c(t)\Phi(t_1,t_0)dtx_0 + \int_{t_0}^t \Phi^T(t,\sigma)c^T(t)\int_0^t \Phi(t,\sigma)u(\sigma)d\sigma dt$$

the matrix

$$M(t_1,t_0) = \int_{t_0}^{t_1} \Phi^T(t_1,t_0)c^T(t)c(t)\Phi(t_1,t_0)dt$$

is called the *observability gramian*.

If A, B, and C are constant matrices then one can associate to the system

$$\dot{x} = Ax + Bu; y = Cx$$

a matrix of rational functions

$$G(s) = C(Is - A)^{-1}B$$

called the *transfer function*. It has a number of properties that make it useful in doing systems analysis. In the first place, if we recall that if the real parts of the eigenvalues of A are all less than σ then $x(t)e^{-\sigma t}$ approaches zero as t goes to infinity and we may define the the Laplace transform for $\mathcal{R}e\ s > \sigma$ by the integral

$$L\big(x(\cdot)\big)(s) = \int_0^\infty e^{-st}x(t)dt$$

It is not too hard to see that the Laplace transform of the matrix exponential e^{At} is given by

$$\int_0^\infty e^{-st}e^{At}dt = (Is - A)^{-1}$$

Thus we see that the transfer function is the Laplace transform of the weighting pattern.

It is of importance to observe that if the real parts of the eigenvalues of A are negative then not only do all the unforced solutions of $\dot{x} = Ax + Bu$ go to zero as t goes to infinity, but in addition,

the solution corresponding to $u = u_0 \cos wt$ approaches the form $x(t) = x_1 \cos \omega t + x_2 \sin \omega t$ and we can compute vectors x_1 and x_2 from u_0 and the transfer function. In fact, if we write

$$C(Ii\omega - A)^{-1}B = \mathcal{R}eG(i\omega) + i\mathcal{I}mG(i\omega)$$

then

$$x_1 = \mathcal{I}m(i\omega)u_0; x_0 = \mathcal{R}eG(i\omega)u_0$$

This justifies the use of the term "frequency response".

If a linear model is to describe certain types of physical situations then additional symmetries and structure maybe present. For example, if

$$\dot{x} = Ax + Bu \ ; \ y = Cx$$

describes a passive electrical network with u being a vector of currents associated with a collection of terminal pairs and y being the corresponding voltages, then $\langle u, y\rangle$ is the rate of flow of energy into the system. If the stored electrical energy is given by $x^T Q^T x$ then of course

$$\begin{aligned}\frac{d}{dt}x^T Qx &= x^T(QA + A^T Q)x + x^T QBu + u^T B^T Qx \\ &= x^T(QA + A^T Q)x + 2\langle y, u\rangle\end{aligned}$$

thus we must have $B = C^T$ and $QA + A^T Q \leq 0$.

A linear stationary system that is controllable and observable and symmetric in the sense that CA^iB for $i = 1, 2, \ldots$ is a symmetric matrix is said to be *passive* if there exists a symmetric positive definite matrix Σ such that

$$\frac{d}{dt}x^T(t)\Sigma x(t) \leq y^T(t)u(t)$$

Example: Consider the linear Itô equation

$$dx = ydt$$

$$dy = zdt + adw$$

$$dz = -xdt - 2ydt - zdt + dw$$

Display the equation for the steady state variance as a function of the parameter a. Is there any value of a such that the steady state variance is not positive definite? You can answer this by solving for the variance if you like but there are other ways to get the answer.

Solution: Writing this in vector matrix notation we have

$$\begin{bmatrix} dx \\ dy \\ dz \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & -2 & -1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} dt + \begin{bmatrix} 0 \\ a \\ 1 \end{bmatrix} dw$$

The steady state variance equation is

$$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & -2 & -1 \end{bmatrix} \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix} + \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix} \begin{bmatrix} 0 & 0 & -1 \\ 1 & 0 & -2 \\ 0 & 1 & -1 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & a^2 & a \\ 0 & a & 1 \end{bmatrix} = 0$$

It can be solved one component at a time. From the first row, $\sigma_{12} = 0$, $\sigma_{22} = -\sigma_{13}$ and $\sigma_{23} = \sigma_{11} + \sigma_{13}$. Turning to the second row, we see that $\sigma_{23} = -a^2/2$ and that $\sigma_{33} - \sigma_{23} - 2\sigma_{22} =$

$-a$. Thus $\sigma_{11} + \sigma_{13} = -a^2/2$ and $\sigma_{33} - 2\sigma_{22} = -a - a^2/2$. Finally, the last equation is $\sigma_{13} + 2\sigma_{23} + \sigma_{33} = 1/2$. Thus, in addition to $\sigma_{12} = 0$ and $\sigma_{23} = -a^2/2$ we have

$$\sigma_{11} + \sigma_{13} = -a^2/2$$

$$\sigma_{22} = -\sigma_{13}$$

$$\sigma_{33} - \sigma_{22} = -a^2$$

$$\sigma_{13} + \sigma_{33} = a^2$$

This leads to $\sigma_{33} - 2\sigma_{22} = a + a^2$, and $\sigma_{22} + \sigma_{33} = 1/2 - a^2$ which implies that $\sigma_{33} = 1/4 + a/2$ and $\sigma_{22} = 1/4 - a/2 - a^2$, and

Thus

$$\begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix} = \begin{bmatrix} 1/2 + a + a^2 & 0 & -1/2 - a - 3a^2/2 \\ 0 & 1/2 + a + 3a^2/2 & -a^2/2 \\ -1/2 - a - 3a^2/2 & -a^2/2 & \frac{1}{2} + a + 5a^2/2 \end{bmatrix}$$

The determinant vanishes, signaling a lack of positive definiteness, if $1 - a - a^2 = 0$.

An easier way to do the problem is to observe that the eigenvalues of A have negative real parts and thus there is a steady state solution. The steady state solution will be positive definite if the syistem is controllable. The controllability matrix is

$$[b, Ab, A^2b] = \begin{bmatrix} 0 & a & 1 \\ a & 1 & -2a - 1 \\ 1 & -2a - 1 & a - 1 \end{bmatrix}$$

and its determinant is $-a^3 - 3a^2 - 2a - 1$. The covariance matrix fails to be positive definite when the system fails to be controllable.

5.2 Covariance and the Power Spectrum

We now turn to a more general situation which illustrates the much broader scope of the previous somewhat special reasoning. For a more detailed account of the background ideas involved here see the literature.

A real valued stochastic process is said to be *wide-sense stationary* if the statistical properties $\mathcal{E}y(t)$ and $\mathcal{E}y(t)y(\tau)$ are, respectively, independent of t and dependent only on $|t-\tau|$. A stationary process is said to be *ergodic* if all statistical properties can be computed as time averages. For example, if y is ergodic, then

$$\mathcal{E}y(t)y(t+\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} y(t)y(t+\tau)dt$$

with y, as it appears on the right, being any particular sample path.

In applications one is often interested in processes of this type and usually measures the power spectrum rather than the autocorrelation function. For a sine wave, $y(t) = a \sin \omega t$ the power is concentrated at ω and the power density (power per unit bandwidth) at that point is a

delta function of strength a^2. In general, the power spectrum, defined for zero mean, wide-sense stationary processes by

$$\Phi(\omega) = \lim_{T\to\infty} \frac{1}{2T} \left| \int_{-T}^{T} e^{-i\omega t} y(t) dt \right|^2$$

is a measure of the "power" at ω in the process y.

For stationary, ergodic processes, the power spectrum and the autocorrelation function are related by the well known Wiener-Khinchin theorem which states that the power spectrum is simply the Fourier transform of the autocorrelation function. We sketch how this goes. By definition

$$\Phi(\omega) = \lim_{T\to\infty} \frac{1}{2T} \left| \int_{-T}^{T} e^{-i\omega t} \phi(t) dt \right|^2$$

which we can write as

$$\begin{aligned} \Phi(\omega) &= \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} \int_{-T}^{T} e^{-i\omega t} y(t) e^{i\omega\tau} y(\tau) dt d\tau \\ &= \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} \int_{-T}^{T} e^{-i\omega(t-\tau)} y(t) y(\tau) dt d\tau \end{aligned}$$

Let $\sigma = t - \tau$ and use the ergodic property to get

$$\Phi(\omega) = \lim_{T\to\infty} \int_{-T}^{T} e^{-i\omega\sigma} \mathcal{E} y(t) y(t-\sigma) d\sigma$$

which is the desired result.

Since the power spectrum of a real valued stochastic process is clearly nonnegative at each frequency, we see that there must be some nontrivial conditions on a function ϕ in order for it to be an autocorrelation function. These are best brought out by means of the *Parseval relation* which states that for complex valued functions which are square integrable on the interval $(-\infty, \infty)$ and related by

$$\Phi(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-i\omega t} y(t) dt$$

we have

$$\int_{-\infty}^{\infty} |y(t)|^2 dt = \int_{-\infty}^{\infty} |\Phi(\omega)|^2 d\omega$$

Example: Let $y(t) = e^{-|t|}$ defined on $(-\infty, \infty)$. Its Fourier transform is

$$\tilde{y}(w) = \frac{1}{\sqrt{2\pi}} \left(\int_0^{\infty} e^{-iwt} e^{-t} dt + \int_{-\infty}^{0} e^{-iwt} e^{t} dt \right) = \frac{1/\sqrt{2\pi}}{1+w^2}$$

If $\Phi(\omega)$ is nonnegative on the real axis, then of course for any $u(\cdot)$

$$0 \le \int_{-\infty}^{\infty} \Phi(\omega) |\tilde{u}(\omega)|^2 d\omega$$

Since multiplication goes over into convolution after Fourier transformation, we have

$$0 \le \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \phi(t-\tau) u(t) u(\tau) dt d\tau$$

Functions ϕ which satisfy this inequality are said to be *positive definite* (in the sense of Bochner). The autocorrelation function of a stationary process is positive definite.

One can gain some appreciation for what it means for a function to be positive definite by letting u be a zero except for a string of narrow pulses centered at $t_1, t_2, \ldots, t_n$. In this case define

$$\Phi(u) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \phi(t-\tau)u(t)u(\tau)\, dtd\tau$$

Then the integral is approximated by

$$[\alpha_1, \alpha_2, \ldots, \alpha_n] \begin{bmatrix} \phi(t_1-t_1) & \phi(t_1-t_2) & \cdots & \phi(t_1-t_n) \\ \phi(t_2-t_1) & \phi(t_2-t_2) & \cdots & \phi(t_2-t_n) \\ \cdots & \cdots & \cdots & \cdots \\ \phi(t_n-t_1) & \phi(t_n-t_2) & \cdots & \phi(t_n-t_n) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix}$$

where α_i is the area under the i^{th} pulse. Clearly we require that the symmetric matrix on the right should be positive definite.

5.3 The Zero-Crossings Problem

We consider briefly here a a problem that comes up frequently in applications. We wish to evaluate the expected number of times a gauss-markov process will cross a given level in a given period of time. The essential calculation will be given in detail but some of the details of the argument will only be treated superficially.

Consider the Itô equation

$$dx = (-x+v)dt \;\; ; \;\; x(0) = 0$$

$$dv = (-v-x)dt + dw \;\; ; \;\; v(0) = 0$$

Compute the steady state value of the covariance matrix

$$\Sigma(\tau) = \mathcal{E} \begin{bmatrix} x(t) \\ x(t+\tau) \end{bmatrix} \begin{bmatrix} x(t) & x(t+\tau) \end{bmatrix}$$

$$dx = (-x+v)dt \;\; ; \;\; x(0) = 0$$

$$dv = (-v-x)dt + dw \;\; ; \;\; v(0) = 0$$

The steady state value of the covariance matrix

$$\Sigma(\tau) = \mathcal{E} \begin{bmatrix} x(t) \\ x(t+\tau) \end{bmatrix} \begin{bmatrix} x(t) & x(t+\tau) \end{bmatrix}$$

can be obtained from the steady state solution of the variance equation associated with (x, v),

$$0 = \begin{bmatrix} -1 & 1 \\ -1 & -1 \end{bmatrix} \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} + \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} \begin{bmatrix} -1 & -1 \\ 1 & -1 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

by solving

$$\Sigma(\tau) = \Sigma_\infty e^{A^T\tau}$$

Note that

$$e^{A^T\tau} = \begin{bmatrix} e^{-\tau}\cos\tau & e^{-\tau}\sin\tau \\ -e^{-\tau}\sin\tau & e^{-\tau}\cos\tau \end{bmatrix}$$

The steady state value of Σ is

$$\Sigma_\infty = \begin{bmatrix} \frac{1}{8} & \frac{1}{8} \\ \frac{1}{8} & \frac{3}{8} \end{bmatrix}$$

5.4 Stochastic Realization Problem

The basic problem in stochastic realization is that of finding a dynamical model, say of the form

$$dx = f(x)dt + \Sigma g_i(x)dw_i + \Sigma\tilde{g}_i(x)dN_i$$

and a map $h(\cdot)$ such that $y(t) = h(x(t))$ has certain, preassigned, statistical properties. Stated in this way, the stochastic realization question is a very general one and one for which a general answer is not likely to be too useful. In control theory and communication theory one is often interested in specifying just the mean and covariance; most commonly

$$\begin{aligned} &\mathcal{E}y(t) = 0 \\ &\mathcal{E}y(t)y(t+\tau) = \phi(\tau) \end{aligned}$$

and then asking for a linear model of the form

$$\begin{aligned} dx &= Axdt + Bd\omega \\ y &= cx \end{aligned}$$

such that y has the desired statistical properties. We will see that this particular version of the stochastic realization problem has a solution which is essentially unique.

Before discussing the mathematical treatment of the problem of stochastic realizations we give a few examples.

Blackbody Realization

The most striking example of a stochastic realization problem comes from physics and concerns the so-called "blackbody radiation problem". In the late 1890's experimenters in Berlin were measuring the power spectra of the radiant energy coming from the sun. The existing theories as developed by Wien and Rayleigh-Jeans did not agree with each other and did not explain the experimental results in a uniform way. Letting ω denote the frequency of the electromagnetic radiation, Max Planck noticed that for appropriate values of α and β the function

$$\Phi(\omega) = \frac{\beta|\omega|^3}{e^{\alpha|\omega|} - 1}$$

provided an excellent fit to the observed data, and more importantly, suggested his quantum hypothesis in order to explain this function. Thus it is fair to say that insofar as Planck's contributions are concerned, the "old quantum mechanics" grew out of an attempt to solve a particular stochastic realization problem.

Speech Synthesis

No two people pronounce a given word the same way. This does not cause much difficulty for humans but it is the source of great difficulty for researchers trying to build artificial speech recognition systems. It means that there is not a single acoustical pressure wave that corresponds to the word "dizziness" but rather, a class of signals. Conversely, it is not appropriate to model "dizziness" by a single signal but rather it should be thought of as a family of possible realizations. Consider a finite state process of the form

$$dx = \sum_{i=1} \phi_i(x) dN_i$$
$$y(t) = \sum_{i=1}^{r} \phi_i(x) b_i(t)$$

where the $b_i(t)$ are fixed deterministic or stochastic processes. According to this point of view spoken English corresponds to a set of stochastic models.

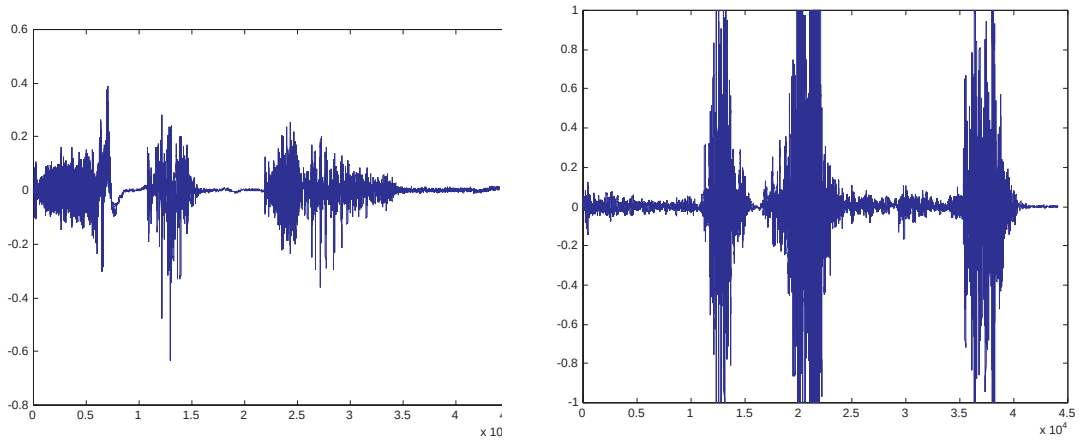

Figure 5.1. A waveform of the author saying the word "spectral" (left) and "factor" (right).

Turbulence

Other problems of this type which are not yet solved include an explanation of the power spectra of certain fluid mechanics problems. In this area the nature of the underlying model is still very much up in the air. In the 1940's Kolmogorov applied physical principles to reason that the power spectrum associated with turbulence should, for large value of w behave like $\pi(\omega) = k\omega^{-5/2}$. The status of this theory is still uncertain.

5.5 Linear Stationary Realizations

As we have seen in section 2.5, the covariance associated with

$$dx = Axdt + \sum_{i=1}^{m} b_i dw_i \quad ; \quad x(t) \in \mathbb{R}^n$$

is given by

$$\mathcal{E}x(t)x^T(t+\tau) = \Sigma(t)e^{A^T\tau} \quad ; \quad \tau \geq 0$$

where Σ satisfies

$$\dot{\Sigma} = A\Sigma + \Sigma A^T + \sum_{i=1}^{m} b_i b_i^T$$

This means that in the limit as t goes to infinity, we obtain an autocorrelation function which depends on τ only provided that the eigenvalues of A have negative real parts. Under this hypothesis we have a unique solution Σ to

$$A\Sigma + \Sigma A^T + \sum_{i=1}^{m} b_i b_i^T = 0$$

Moreover, this solution is nonnegative definite and can be expressed as

$$\Sigma = \sum_{i=1}^{m} \int_0^\infty e^{At} b_i b_i^T e^{A^T t} dt$$

It is not too hard to see that Σ is actually positive definite if the system

$$\dot{x} = Ax + \sum_{i=1}^{m} b_i u_i$$

is controllable. That is to say, Σ is positive definite if the set $\{A^i b_j\}$ spans $\mathbb{R}^n$.

In any case, assuming only that $\mathcal{R}e\ \lambda(A) < 0$ we have for $\tau > 0$

$$\lim_{t\to\infty} \mathcal{E}x(t)x(t+\tau) = \left(\sum_{i=1}^{m} \int_0^\infty e^{At} b_i b_i^T e^{A^T t} dt \right) e^{A^T \tau}$$

If $y = cx$ is any linear functional on x, then

$$\lim_{t\to\infty} \mathcal{E}\left[y(t)y^T(t+\tau)\right] = c\Sigma e^{A^T \tau} c^T$$

where Σ is as above.

5.6 Spectral Factorization

Given a linear stochastic equation of the form

$$dx = Axdt + Bdw \quad ; \quad x(t) \in \mathbb{R}^n$$

with (A, B) a controllable pair ($(B, AB, \ldots, A^{n-1}B)$ has rank n), we know that if the eigenvalues of A have negative real parts, then

$$\begin{aligned} \lim_{t\to\infty} \mathcal{E}x(t)x^T(t+\tau) &= \Sigma e^{A^T \tau} \ ; \ \tau \geq 0 \\ \lim_{t\to\infty} \mathcal{E}x(t)x^T(t+\tau) &= e^{-A\tau}\Sigma \ ; \ \tau \leq 0 \end{aligned}$$

Using the definition of the Fourier transform we obtain

$$\begin{aligned} \Phi(\omega) &= \int_{-\infty}^{\infty} e^{-i\omega\tau} \phi(\tau) d\tau \\ &= \int_0^\infty e^{-i\omega\tau} \Sigma e^{A^T \tau} d\tau + \int_{-\infty}^{0} e^{-i\omega\tau} e^{-A\tau} \Sigma d\tau \\ &= -\Sigma(-i\omega I + A^T)^{-1} - (i\omega I - A)^{-1}\Sigma \end{aligned}$$

We can use the formula $\Sigma A^T + A\Sigma + BB^T = 0$ to reexpress this. Adding and subtracting $i\omega\Sigma$ we get

$$(A - Ii\omega)\Sigma + \Sigma(Ii\omega + A^T) = -BB^T$$

Pre and post multiplication by $-(Ii\omega - A)^{-1}$ and $(Ii\omega + A^T)^{-1}$ yields

$$\Sigma(-Ii\omega - A^T)^{-1} + (Ii\omega - A)^{-1}\Sigma = (Ii\omega - A)^{-1}BB^T(-Ii\omega - A^T)^{-1}$$

Thus

$$\Phi(\omega) = (-Ii\omega - A)^{-1}BB^T(Ii\omega - A^T)^{-1}$$

This shows that the Fourier transform of the autocorrelation function is hermitian and nonnegative. Looking at $y = cx$ we get for the Fourier transform of $\mathcal{E}cx(t)cx(t+\tau) = \phi(\tau)$

$$\mathcal{F}(\phi(\cdot)) = c(-Ii\omega - A)^{-1}bb^T(Ii\omega - A^T)^{-1}c^T$$

The solution of the present version of the stochastic realization problem is based on this identity and the following lemma. (*The spectral factorization lemma.*)

Lemma: Given an even function $\psi(s) = \psi(-s)$ which is real, rational, and nonnegative on $s = i\omega$ for all ω, there exists $r(s)$, real, rational, and having no poles in the half plane $\mathcal{R}e\ s > 0$ such that

$$\psi(s) = r(s)r(-s)$$

Moreover, $r(s)$ is analytic wherever ψ is.

Proof: See Finite Dimensional Linear Systems.

Suppose we are given a bounded, continuous real valued function $\phi(\cdot) = \phi(-\cdot)$ which for $t \geq 0$ is of the form

$$\phi(t) = \sum \alpha_{ij}\ t^i\ e^{\lambda_j\ t}\ ;\ (\lambda_j \text{ may be complex})$$

positive definite in the sense of Bochner. Its Fourier transform Φ is nonnegative and hence can be expressed as

$$\Phi(\omega) = r(i\omega)r(-i\omega)$$

with $r(s)$ a real rational function with its poles in $\mathcal{R}e\ s > 0$. (Poles on $\mathcal{R}e\ s = 0$ are ruled out by the fact that they lead to infinite total power in the power spectrum.) In view of these remarks we can express the inverse Fourier transform as

$$(\mathcal{F}^{-1}r)(t) = ce^{At}b$$

Tracing backwards the above formulae yield the identity

$$\lim_{t\to\infty} \mathcal{E}y(t)y(t+\tau) = \phi(\tau)$$

for

$$\begin{aligned} dx &= Axdt + bd\omega \\ y &= cx \end{aligned}$$

Thus to solve the problem of realizing a given stationary covariance $\phi(\cdot)$ with a Wiener process model it is enough to

(i) Transform $\phi(\cdot)$
(ii) Factor the transform using spectral factorization
(iii) Express $r(s)$ as $c(Is - A)^{-1}b$
(iv) Construct
$$dx = Axdt + bd\omega$$
$$dy = cx$$

Example: Let the power spectrum be

$$\begin{aligned}\Phi(w) &= \frac{1}{1+w^2} + \frac{1}{9+w^2} \\ &= \frac{10+2w^2}{(1+w^2)(9+w^2)} \\ &= \sqrt{2}\frac{(\sqrt{5}+w)}{(1+iw)(3+iw)} \cdot \sqrt{2}\frac{(\sqrt{5}-iw)}{(1-iw)(3-iw)}\end{aligned}$$

A corresponding realization of the covariance is provided by

$$\begin{bmatrix} dx_1 \\ dx_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -3 & -4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} dt + \begin{bmatrix} 0 \\ 1 \end{bmatrix} dw \;\; ; \;\; dy = [\sqrt{10}\,,\, \sqrt{2}] \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} dt$$

5.7 The Gauss-Markov Heat Bath

This section describes an application of the above ideas to a basic problem in statistical mechanics. The stochastic realization problems which appears here is distinguished from those just considered by a need to realize jointly a certain "coupling behavior" (e.g., an impedance) together with a power spectrum. This idea is illustrated very nicely by the Nyquist-Johnson model for a resistor. In fact the model proposed by them, when used with the theory of linear systems, leads to a rather nice packaging of a certain circle of ideas in statistical mechanics. The advantage of the system theory formalization is that it allows one to discuss the interaction between systems in a very precise way. We will illustrate this with a discussion of heat baths. In the process we will need to clairfy the following ideas.

1. Linear passive systems and stochastic dynamical systems
2. The fluctuation dissipation equality and temperature.
3. The heat bath as a stochastic dynamical system.

By a linear, finite dimensional, gaussian system (FDLGS) we understand a pair

$$dx = Axdt + Budt + Gdw$$
$$y = Cx + Du$$

Here dw is a vector valued Wiener process and x, u, y are all vector valued. This system is said to be *minimal* if (A, B) is a controllable pair and (A, C) is an observable pair.

We refer to
$$G(s) = C(Is - A)^{-1}B + D$$
as the *transfer function* associated with the system and we call

$$\phi(s) = (D + C(-Is - A)^{-1}G)^T(D + C(Is - A)^{-1}G)$$

the *power spectrum* of the system. We say that the system is *externally reciprocal* if $G(s) = G^T(s)$. The system is said to be *deterministically passive* if the eigenvalues of A have real parts that are negative and if $G(i\omega) + G^T(-i\omega) \geq 0$. A deterministically passive system of this form is said to be *monotemperaturic* if there exists a proportionality factor $\beta \geq 0$ such that the power spectrum of y and the frequency response are related by

$$\big(C(-Is-A)^{-1}G+D\big)^T\big(C(Is-A)^{-1}G+D\big) = \beta[C(-Is+A)^{-1}B^T+D^T+C(Is-A)^{-1}B+D] \quad (*)$$

In the language of thermodynamics, this equation expresses a frequency independent fluctuation-dissipation proportionality. In the language of system theory it expresses a proportionality between the power spectrum and the parahermetian part of the transfer function. If equality (*) holds we call β the *temperature* of the system. Several justifications for this definition can be given. One is that (*) expresses a property of electrical networks constructed from linear constant inductors, capacitors and resistors in Nyquist-Johnson form and all of the same temperature. It is best thought of as the fluctuation-dissipation equality in the context of linear systems excited by white noise.

One of the beautiful facts about the linear theory of equilibrium thermodynamics is the equipartition of energy theorem which states that the expected value of the energy of each mode of a system in equilibrium is the same. For example, a balloon in still air can be expected to have as much kinetic energy as an O_2 molecule. This idea finds its expression here in terms of the interconnection of lossless systems with monotemperaturic systems.

We will say that a linear stochastic system has the equipartition property if there exists a positive number β such that whenever it is connected to a conservative system with impedance $Z(s)$ the resulting system has a unique invariant measure and for this measure

$$\mathcal{E}yy^T = \beta G_0 \; ; \;\; G(s) = G_0 s^{-1} + G_1 s^{-2} + \dots$$
$$\mathcal{E}uu^T = \beta Z_0 \; ; \;\; Z(s) = Z_0 s^{-1} + Z_1 s^{-2} + \dots$$

regardless of the impedance $Z(s)$ of the conservative system.

The explanation of this definition is that if one chooses normal coordinates for the lossless system then it appears as

$$\dot{x} = \Omega x + Bu; \quad y = B^T x$$

Equipartition means that $\mathcal{E}xx^T = \beta I$ for some β and thus $\mathcal{E}yy^T = \beta B^T B$. However $Z(s) = B^T B s^{-1} + B^T A B s^{-2} + \dots$. Similar but more involved calculations apply to $\mathcal{E}uu^T$. The following theorem is easily verified.

Theorem 1: A linear stochastic system has the equipartition property if and only if it is monotemperaturic.

Monotemperaturic linear stochastic systems have the following canonical form, relating to the Darlington normal form for passive electrical networks [5].

Theorem 2: If S is a minimal, externally reciprocal, monotemperaturic system then we can make a linear change of coordinates on x such that it takes the form ($\Omega = -\Omega^T$)

$$dx = (\Omega - \frac{1}{2\beta}GG^T)xdt + Budt + Gdw$$
$$dy = B\dot{x}dt + D\dot{u}dt + \sqrt{2\beta D}df$$

Conversely, any system of this form is monotemperaturic.

Proof: That systems of this form are monotemperaturic is just a calculation.

Consider the deterministic system

$$\dot{x} = (A - GG^T)x + Bu_1 + Gu_2 \quad ; \quad y_1 = Bu; y_2 = G^T x$$

Adopt the notation

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \end{bmatrix} = \begin{bmatrix} G_{11}(s) & G_{12}(s) \\ G_{21}(s) & G_{22}(s) \end{bmatrix} \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \end{bmatrix}$$

The characterization of monotemperaturic implies that this system is convervative and hence it can be realized with $A - \frac{1}{2\beta}GG^T$ skew symmetric.

The classical equipartition of energy theorem is quite elegantly expressed by saying that a linear stochastic differential equation of the form

$$dx = (S - BB^T)xdt + Cudt + Bdw; \quad y = Cx$$

$(S = -S^T)$ has the property that its steady state variance is $\frac{1}{2}\mathrm{I}$ and if we interconnect it with a lossless system

$$z = \Omega x + Gu; \quad v = G^T z : \quad \Omega = \Omega^T$$

then the random process (x, z) has variance $\frac{1}{2}\mathrm{I}$ as well.

We can use these ideas to explain the Rayleigh-Jeans law in the following way. If we have a one parameter family of monotemperaturic systems of the form

$$dx = (S - \epsilon^2 BB^T)xdt + \epsilon Bdw; \quad y = cx; \quad S = -S^T$$

Then in steady state

$$\mathcal{E}y(t)y(t+\tau) = \frac{1}{2}ce^{A_\epsilon^T \tau}c^T$$

where

$$A_\epsilon = S - \epsilon^2 BB^T$$

There is, of course, a limiting value for A_ϵ as ϵ goes to zero and the resulting power spectrum is pure line spectrum. The lines in the line spectrum correspond to the eigenvalues of S. Thus the number of eigenvalues of S in an interval $\omega_0 < \omega < \omega_0 + \delta$ determine the amount of power in the interval provided we select the linear functional c in such a way as to weight all modes equally.

One feature of monotemperaturic systems is that they form an "interconnectable set" in that if we interconnect two systems which satisfy this form of the fluctuation-dissipation equality then the resulting system does also. Within the realm of linear gaussian systems there is a generalization of this idea which is significant, namely the fact that if $Re\ g(i\omega) = \chi(\omega)\psi(i\omega)$ for a system S_1 and if the same is true for a system S_2 then it remains true for the interconnection of S_1 and S_2. Let us denote by $P(\chi)$ the set of passive systems with this property.

Theorem 3: Suppose that

$$dx = (Ax + bu)dt + Bdw; \quad y = cx$$

is given with (A, b, c) a minimal triple and (A, B) a controllable pair. Suppose further that the eigenvalues of A have negative real parts and that $Re\ g(i\omega) \geq 0$ for all ω. If q and p satisfy

$$\begin{bmatrix} \dot{q} \\ \dot{p} \end{bmatrix} = \begin{bmatrix} 0 & \omega \\ -\omega & 0 \end{bmatrix} \begin{bmatrix} q \\ p \end{bmatrix} + \begin{bmatrix} \epsilon \\ 0 \end{bmatrix} u^T \ ; \ y^T = [\epsilon, 0] \begin{bmatrix} q \\ p \end{bmatrix}$$

We define an interconnected system by

$$d\begin{bmatrix} x \\ q \\ p \end{bmatrix} = \begin{bmatrix} A & b\epsilon & 0 \\ -c & 0 & \omega \\ 0 & -\omega & 0 \end{bmatrix} \begin{bmatrix} x \\ q \\ p \end{bmatrix} dt + \begin{bmatrix} Bd\omega \\ 0 \\ 0 \end{bmatrix}$$

If $G(s) = c(Is - A)^{-1}b$ and if $AQ + QA^T = BB^T$ with $\phi(s) = c(Is - A)^{-1}Q(-Is + A^T)c^T$ then

$$\Sigma_0 = \begin{bmatrix} Q & 0 \\ 0 & \alpha I \end{bmatrix} \text{ with } \alpha = \frac{\psi(i\omega)}{\mathcal{R}eg(i\omega)}$$

is the weak coupling limit of the steady state variance limiting value as ϵ goes to zero.

Proof: The proof of this theorem involves solving for the limiting value of the solution of a system equation of the form $(A + \epsilon B)x = c$ with A singular. In the first place, for $\mathcal{R}e\, G(i\omega) > 0$ we see that the eigenvalues of

$$\bar{A} = \begin{bmatrix} A & \epsilon b & Q \\ -c & 0 & \omega \\ 0 & -\omega & 0 \end{bmatrix}$$

have negative real parts for $\epsilon \neq 0$. thus for $\epsilon \neq 0$ there is a unique positive definite variance $\Sigma(\infty)$ which is the solution of

$$\bar{A}(\epsilon)\Sigma(\infty) + \Sigma(\infty)\bar{A}^T = \begin{bmatrix} BB^T & 0 \\ 0 & 0 \end{bmatrix} = M$$

To solve this we make use of the representation

$$\Sigma(\infty) = \int_0^\infty e^{\bar{A}^T(\epsilon)t} M e^{A(\epsilon)t} dt$$

transformed by Parseval's theorem to

$$\Sigma(\infty) = \frac{1}{2\pi i} \int_{-i\infty}^{i\infty} (Is - \bar{A}^T(\epsilon))^{-1} M (Is - \bar{A}(\epsilon))^{-1} ds$$

this formula then readily yields the desired result.

We may then think of any system

$$dx = Axdt + Budt + d\omega; \quad y = cx$$

with $c(Is - A)^{-1}b$ positive real as defining a *heat bath* in thesense that if we couple it loosely an oscillator and allow it to come to equilibrium the oscillator will possess a certain expected energy, independent of the details of the coupling or the oscillator.

Now it is well known in a three dimensional cavity the number of eigenfrequencies for the wave equation between ω_0 and $\omega_0 + \delta$ is approximately $\delta\omega_0^2$ for ω_0 not too small. Thus if the coupling of each oscillator to y (as represented by c) is of the same strength, then the power spectral density is proportional to ω^2 as indicated above.

In the classical theory one takes $\chi(\omega)$ to be independent of the frequency and identifies its value, via physical reasoning, with 1/2 the temperature times Boltzmann's constant. In a quantum theory, in order to be consistent with the black body radiation curve, it would be necessary to take $\chi(\omega)$ to be a member of the one-parameter-family

$$\chi(\omega) = \frac{h\omega/T}{e^{k\omega/k^T} - 1}$$

We make some additional remarks here. In the first place, for an oscillator with weak coupling to a heat bath we find that the steady state value of the expected value of the energy depends on the natural frequency ω of the oscillator and is simply $\chi(\omega)$. If we connect with weak coupling a lossless first order system – a capacatance or an inductance – we get $\chi(0)$ or $\chi(\infty)$ respectively.

5.8 Reducibility

Let A be the infinitesimal generator of a finite state, continuous time Markov process whose state space is a subset of the real line. If $\{x_1, x_2, \ldots, x_n\}$ are the values taken on by this process, then

$$y(t) = x(t)$$

is a real valued stochastic process and its autocorrelation function

$$\mathcal{E}y(t)y(t+\tau) = \phi(t,\tau)$$

can be computed. In particular we can ask if there exists a limit

$$\lim_{t\to\infty} \mathcal{E}y(t)y(t+\tau)$$

As might be expected, the existence and uniqueness of such a limit is dependent on the properties of A.

One says that a finite dimensional generator A is *irreducible* if there is no permutation matrix P such that

$$PAP^{-1} = \begin{bmatrix} A_{11} & 0 \\ A_{12} & A_{22} \end{bmatrix}$$

with A_{11} and A_{22} both square. It is clear that an infinitesimal generator A is always singular. If, in fact, there are two distinct probability vectors p_1 and p_2 such that $Ap_1 = Ap_2 = 0$, then of course

$$A(\alpha p_1 + (1-\alpha)p_2) = 0$$

for some value of α it will happen that $\alpha p_1 + (1-\alpha)p_2$ has nonnegative entries with some entries zero. By a permutation we can arrange these to be at the top. Thus

$$\begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{2n} & \ldots & a_{nn} \end{bmatrix} \begin{bmatrix} 0 \\ x_l \end{bmatrix} = 0$$

Since x_l has positive entries and since A is nonnegative off the diagonal, we see that the upper right-hand block of A must be zero and that A must be reducible. Thus we have shown the following.

Lemma: If there is not a unique steady state probability distribution for $\dot{p} = Ap$, then A is reducible.

5.9 Covariance Generation with Finite State Processes

We now turn our attention to the question of generating a covariance using a finite state continuous time Markov process. This represents a considerable departure from the usual Gauss-Markov theory discussed in the previous section. The main question here is this. Given a stationary covariance $\phi(\tau)$ under what circumstances does there exist a finite state continuous time Markov process $x(\cdot)$ taking on values in a finite set X and a function $f : X \to \mathbb{R}$ (the real numbers) such that $y(t) = f[x(t)]$ has a specified mean and covariance?

Because we have shown that any finite state continuous time Markov process is equivalent to one which can be expressed as

$$dx(t) = \sum_{i=1}^{\delta'} B_i x(t) dN_i(t) \quad ; \quad x(t) \in \mathbb{R}^n$$

with $N_i(t)$ a standard Poisson counting process with rate λ_i, this work is a natural complement to the Gauss-Markov covariance generation problem.

We find it convenient to associate with each state of an n state Markov process a point in $\mathbb{R}^n$. This lets us visualize the process as jumping between points in a vector space and allows us to use certain familiar formulas from linear system theory. Let e_i be the i^{th} standard basic element in $\mathbb{R}^n$

$$e_i = [0, 0, \ldots, 0, 1, 0, \ldots, 0]^T$$
$$\qquad\qquad\qquad\qquad\qquad\qquad i^{\text{th}} \text{ coordinate}$$

and let x be a process which takes on values in the set $\{e_1, e_2, \ldots, e_n\}$. If $p_i(t)$ is the probability that $x(t) = e_i$, then our assumptions imply that there exists a constant matrix A such that

$$\begin{bmatrix} \dot{p}_1(t) \\ \dot{p}_2(t) \\ \vdots \\ \dot{p}_n(t) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{2n} & \ldots & a_{nn} \end{bmatrix} \begin{bmatrix} p_1(t) \\ p_2(t) \\ \vdots \\ p_n(t) \end{bmatrix}$$

Notice that because of the way we have embedded the states in $\mathbb{R}^n$ we have

$$p(t) = \mathcal{E}x(t)$$

As we discussed in chapter 2, the entries of A satisfy the conditions

(i) $a_{ij} \begin{cases} \geq 0 & \text{if } i \neq j \\ \leq 0 & \text{if } i = j \end{cases}$

(ii) $\sum_{i=1}^{n} a_{ij} = 0$

Such matrices are called infinitesimally stochastic. Of course Peron-Frobenius theory implies that A has a nontrivial null space. If we ask that the null space be one dimensional, then we are assured that there is a unique steady state probability distribution. For these reasons we will assume

(iii) the kernel of A is one dimensional.

Such processes are called irreducible.

As an immediate consequence of the definitions we see that

$$\mathcal{E}x(t) = e^{At}x(0)$$

Because the i^{th} and j^{th} components of x are never simultaneously nonzero and because the components take on only the values zero and one

$$\Sigma(t) = \mathcal{E}x(t)x^T(t) = \begin{bmatrix} p_1(t) & 0 & \ldots & 0 \\ 0 & p_2(t) & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & p_n(t) \end{bmatrix}$$

Moreover an elementary application of Bayes' rule yields

$$\mathcal{E}x(t)x^T(\tau) = \Sigma(t)e^{A^T(\tau-t)} \quad ; \quad \tau \geq t$$

The equations (1) and (2) together with one more remark will yield a restatement of our question in linear algebraic form. The additional remark is this. Any map of the state space of the Markov process into $\mathbb{R}$ is of the form $c^T x$ for some c in $\mathbb{R}^n$. Simply let the i^{th} component of c be $f(e_i)$ where $f : X \to \mathbb{R}$. Bringing these remarks together we have the following theorem.

Theorem 1: A stationary covariance $\phi(t-\tau) = \hat{\phi}(t,\tau)$ is realizable by a zero mean process which is a real valued function of a finite state, continuous time irreducible Markov process if and only if

$$\phi(t) = c^T \Sigma e^{At} c \quad ; \quad t > 0$$

for some pair (A, c) where A is an n by n matrix satisfying (i), (ii), (iii); c is an n vector such that $c^T x = 0$ for all x in the kernel of A, and

$$\Sigma = \begin{bmatrix} p_1 & 0 & \dots & 0 \\ 0 & p_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & p_n \end{bmatrix}$$

with $(p_1, p_2, \dots, p_n)$ being the probability vector in the kernel of A.

The determination of what functions can be expressed in the form required by equation (4.3) is made difficult by the requirement that A be infinitesimally stochastic. To get around the awkwardness of this constraint we focus attention on a special class of matrices. By a circulant matrix we understand a square matrix of the form

$$M = \begin{bmatrix} m_0 & m_1 & m_2 & \dots & m_{n-1} \\ m_{n-1} & m_0 & m_1 & \dots & m_{n-2} \\ & \ddots & \ddots & \ddots & \\ m_2 & m_3 & m_4 & \dots & m_1 \\ m_1 & m_2 & m_3 & \dots & m_0 \end{bmatrix}$$

Associated with each such M there is a polynomial $\hat{m}(z) = m_0 + m_1 z + \cdots + m_{n-1} z^{n-1}$. The eigenvalues of a circulant matrix are simply the values

$$\lambda_k = \hat{m}(e^{ik\theta}) \quad ; \quad \theta = \frac{2\pi}{n} \quad ; \quad k = 0, 1, \dots, n-1$$

Thus M meets condition (i), (ii), (iii) if and only if

(i′) the coefficients of $\hat{m}(z)$ are all nonnegative except for the constant term.
(ii′) $\hat{m}(1) = 0$
(iii′) $\hat{m}(z)$ does not vanish for z an n^{th} root of unity unequal to one.

Under these assumptions one sees easily that the solution of

$$\dot{p} = Mp$$

for $p(0)$ a probability vector, tends to

$$p_\infty = \frac{1}{n} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

Thus using such an M we see that the matrix Σ in equation (4.3) becomes $\frac{1}{n}I$ and we have

$$\tilde{\phi}(t) = \frac{1}{n} c^T e^{Mt} c \quad ; \quad t > 0$$

The condition that the mean should vanish, $c^T p_\infty = 0$, is also easily interpretable. In fact if

$$c^T = (c_0, c_1, \ldots, c_{n-1})$$

we introduce

$$\hat{c}(z) = c_0 + c_1 z + \cdots + c_{n-1} z^{n-1}$$

In this notation $c^T p_\infty = 0$ becomes

$$\hat{c}(1) = 0$$

We can also express $\tilde{\phi}$ succinctly in terms of $m(z)$ and $c(z)$ (see [2]).

$$\tilde{\phi}(t) = \sum_{z:z^n=1} \hat{c}(z)\hat{c}(z^{-1}) e^{\hat{m}(z)t} \quad ; \quad t \geq 0$$

The following lemma gives a somewhat more satisfactory form of this.

Lemma 1: The set of finite state continuous time realizable covariances include those expressible as

$$\tilde{\phi}(t) = \sum_{k=1}^{n-1} r_k e^{\hat{m}(e^{2\pi i k/n})t} \quad ; \quad t \geq 0$$

with m satisfying (i′), (ii′) and (iii′) and the r_k real and nonnegative. In particular

$$\psi(t) = re^{-\sigma t} \cos \omega t \quad ; \quad t \geq 0$$

is realizable if r and σ are real and positive and ω is real.

Proof: Of course $\hat{c}(z)\hat{c}(z^{-1})$ is, for z on the unit circle, real and nonnegative. Since $\hat{c}(1) = 0$, we must have $\hat{c}(z)\hat{c}(z^{-1})$ vanishing at $z = 1$ but otherwise we may pick the coefficients so that $\hat{c}(z)$ has arbitrary complex values at the roots of unity consistent with $c(z) = \overline{c(\overline{z})}$. Adding up the contribution from $\rho = e^{2\pi i k/n}$ we get

$$\tilde{\phi}_k(t) = 2|c(\rho)|^2 e^{Re\, \hat{m}(\rho)t} \cos \mathcal{I}m\, \hat{m}(\rho)t$$

But since $\hat{c}(\rho)$ is arbitrary, we see that $|\hat{c}(\rho)|^2$ can be any real nonnegative number. The general form given in the lemma then follows.

To show that the specific ψ given in the lemma is expressible in this way we make a particular choice of n and $\hat{m}(z)$. Let

$$\hat{m}(z) = a(1 - \alpha z - (1 - \alpha)z^2) \quad ; \quad a, \alpha > 0$$

and let $\sigma > 0$ and ω be given. At $z = e^{i2\pi/n} = \cos(2\pi/n) + i \sin(2\pi/n)$ the ratio of the real to the imaginary parts of $\hat{m}$ is

$$\gamma = \frac{1 - \alpha \cos(2\pi/n) - (1 - \alpha)\cos(4\pi/n)}{1 - \alpha \sin(2\pi/n) - (1 - \alpha)\sin(4\pi/n)}$$

Inspection of this equation shows that for any negative γ we can choose an integer n large enough so as to have a solution for α. (As γ approaches zero, $n(\gamma)$ goes to infinity.) Thus we can, with this choice of $m(z)$ adjust the magnitude and argument of $m(\exp(2\pi i/n)$ as needed to get the function $\psi(\cdot)$ of the lemma. Of course we pick c in such a way as to vanish on all n^{th} roots of unity except the two which enter in this discussion.

Lemma 1 makes it clear how to realize nonnegative linear combinations of the basic terms labeled there ψ; take the direct sum of realizations of the type constructed in its proof. However these realizations

(a) will not, in general, satisfy the irreducibility condition,
and

(b) give no suggestions as to how to realize covariances such as $\phi(t) = 10e^{-|t|} - e^{-5|t|}$ which are differences of positive definite functions but still positive definite.

We now establish the results necessary to get around these difficulties.

If $\phi(t)$ is a continuous, even, positive definite function, then according to the well known representation theorem of Bochner it can be expressed as

$$\phi(t) = \int_{[0,\infty)} \cos \omega t d\mu$$

for some nonnegative measure μ. Of course if μ is absolutely continuous with respect to Lebesgue measure, then we can write

$$\phi(t) = \int_0^\infty \cos(\omega t)\Phi(\omega)d\omega \quad ; \quad \Phi(\omega) \geq 0$$

displaying the power spectrum explicitly.

However, if we assume that ϕ is not only positive definite but in addition it is *strictly positive definite* in the sense that

$$\phi^\epsilon(t) = \phi(t)e^{2\epsilon|t|}$$

is for, some $\epsilon > 0$, also square integrable and positive definite, then we can express ϕ as

$$\phi(t) = \int_0^\infty e^{-\epsilon|t|} \cos(\omega t)\Phi(\omega)d\omega$$

with ϕ analytic. (This follows from Payley-Wiener theory; the Fourier transform of $\phi(t)e^{-\epsilon|t|}$ analytic in a strip of width 2ϵ centered on the ω-axis and Φ is its Fourier transform.) Let $\{t_i\}_{i=1}^r$ be any finite set of real numbers. We can approximate simultaneously the integrals $\phi(t_i)$ by Riemann sums and thus obtain

$$\omega(t) = \sum_{i=1}^m e^{-\epsilon t} \cos \omega_i t \phi(\omega_i) s_i + \epsilon(t)$$

with $|\epsilon(t_i)|$ less than any preassigned positive number. Both ψ and the approximation go to zero as $|t| \to \infty$; in view of the continuity of ϕ we see that it can be uniformly approximated by a linear combination of ψ-like terms *with positive coefficients*. The following theorem summarizes.

Theorem 2: Any continuous, strictly positive definite function is the uniform limit of a sequence ϕ_n of the form

$$\phi_n(t) = \sum_{k=1}^{m(n)} \alpha_k e^{-\epsilon t} \cos \omega_k(t) \quad ; \quad \alpha_k > 0$$

We now address the second problem mentioned at the start of this section.

Lemma 2: Stationary covariances of the form appearing in theorem 2 can be realized by a pair (A, c) satisfying conditions (i), (ii) and (iii).

Proof: Let (A_k, c_k) be a realization of $\alpha_k e^{-\epsilon t/2}\cos\omega_k t$ of the form given in the proof of lemma 1; i.e., of the circulant form. (Note we have $\epsilon/2$ as the decay factor.) Then for

$$A = \begin{bmatrix} A_1 & 0 & \dots & 0 \\ 0 & A_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & A_m \end{bmatrix} \quad ; \quad c = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_m \end{bmatrix}$$

it follows that

$$c^T e^{At} c = \sum_{k=1}^{m} \alpha_k e^{-\epsilon t/2}\cos\omega_k t$$

Now subtract from A the infinitesimally stochastic matrix

$$F = \frac{\epsilon}{2} I - \frac{\epsilon}{2n}\begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 1 \end{bmatrix}$$

where n is the sum of the dimensions of the A_i. Clearly $A - F$ is infinitesimally stochastic and irreducible. Notice that

$$c^T e^{(A-I/2)t} c = \sum_{k=1}^{m} \alpha_k e^{-\epsilon t}\cos\omega_k t$$

(This time the decay factor agrees with theorem 2.) Finally, because of the null spaces of each of the A_i we see that

$$A\begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 1 \end{bmatrix} = 0 \quad ;$$

and

$$c^T\begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 1 \end{bmatrix} = 0$$

Together these imply $c^T e^{(A-F)t} c = c^T e^{(A-\epsilon I/2)t} c$ and so $(A - F, c)$ meets all requirements.

5.10 Exercises 5

1. Find a realization for a stationary, zero-mean process with autocorrelation function equal to $10e^{-3|\tau|}$.

2. Under what circumstances is the process

$$y(t) = (\sin t)c^T x + (\cos t)d^T x$$

second order (wide sense) stationary if

$$dx = Axdt + Bdw$$

and the x process is in steady state?

3. Describe all the 2-dimensional Wiener process realizations of the power spectrum

$$\Phi(\omega) = \frac{1}{1+\omega^2} + \frac{4}{1+4\omega^2}$$

4. Find a finite state continuous time jump process realization of the power spectrum

$$\Phi(\omega) = \frac{1}{\omega^4 + \omega^2 + 1}$$

5. An n by n matrix P is said to be *orthostochastic* if $p_{ij} = \theta_{ij}^2$ with θ_{ij} being the ij^{th} element of an orthogonal matrix. Show that under these circumstances $p(k+1) = Pp(k)$ defines a Markov chain. Show that an orthostochastic matrix has nonnegative entries whose columns sum to one. How many parameters are needed to specify an n by n orthostochastic matrix and how many are needed to specify a stochastic matrix (i.e., -a matrix with nonnegative entries and columns which sum to one)?

6. Find a constant matrix A and vectors b and c, such that for

$$dx = Axdt + bdw \quad ; \quad y = cx$$

the power spectrum of y is given by

$$\phi(w) = \frac{1+w^2}{(1-7w^2)^2 + 1}$$

Describe the sense in which your answer is unique.

7. Consider an intensity matrix for a four-state continuous time jump process

$$A = \begin{bmatrix} -a & 0 & d & d \\ 0 & -a & e & e \\ b & b & -f & 0 \\ c & c & 0 & -f \end{bmatrix}$$

If p_i is the probability that $x = x_i$ and if $(x_1, x_2, x_3, x_4) = (m, n, m, n)$ then compute the expectation of $x(t)x(t+\tau)$. Hint: Show that the eigenvalues of the matrix A satisfy

$$(s+a)(s+f) - af = 0$$

8. Let c be an n dimensional row vector whose entries are chosen from $\{0, 1\}$. Let A be an infintessimal generator of a continuous time jump process and let $p(0)$ be a probability vector. Consider

$$\Psi(t) = ce^{At}p(0)$$

Show that if A is a circulant matrix corresponding to $-1 + z$ then

$$\Psi(t) = \sum_{k=0}^{n} cke^{-t}e^{2\pi ikt}$$

9. Find the values of a, b, λ_1 and λ_2 such that the stochastic differential equations driven by Poisson counters

$$dx = -2x dN_1 \quad ; \quad x(0) \in \{-1, 1\} \quad ; \quad \mathcal{E}N_1(t) = \lambda_1 t$$

$$dy = -2y dN_2 \quad ; \quad y(0) \in \{-1, 1\} \quad ; \quad \mathcal{E}N_2(t) = \lambda_2 t$$

with

$$z(t) = ax(t) + by(t)$$

generate an autocorrelation function for z which is, in steady state,

$$\mathcal{E}z(t)\, z(t+\tau) = 13e^{-|\tau|} + 2e^{-2|\tau|}$$

NOTES AND REFERENCES

1. A standard reference here is:
H. Bode and C. E. Shannon, A Simplified Derivation of Linear Least Squares Smoothing and Prediction Theory, Proceedings of the I.R.E., Vol. xx, (1950) pp. 417–425.

2. R. W. Brockett, "Stochastic Realization Theory and the Planck Law for Black Body Radiation," *Ricerche di Automatica*, Vol. 10 (1979) pp. 344-362.

3. B.D.O. Anderson and T. Kailath, "Forwards and Backwards Models for Finite-State Markov Processes", *Advances in Applied Probability*, Vol. 11, pp 118-133, (1979).

4. L. R. Rabiner, A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, *Proceedings of the IEEE*, Vol. 77, No. 2, (Feb. 1989), pp. 257–285.

Chapter 6

Estimation Theory

Given a noisy observation of a stochastic process and given some a priori statistical characterization of the process, how can one extract the best estimate of the process from its noisy version? This type of problem is ubiquitous in science and engineering. In this chapter we develop some of the basic results.

6.1 Preliminaries

We begin with the following simple situation. Let x be Markov process which evolves in discrete time, taking on values in a finite set $X = \{x_1, x_2, ..., x_N\}$. Let $p_i(k)$ be the probability that $x(k) = x_i$ and suppose that $p(k+1) = Ap(k)$. Let y be a second discrete time stochastic process taking on values in a finite set $Y = \{y_1, y_2, ..., y_N\}$. We postulate that $y(k)$ depends on $x(k)$ through a probabilistic law. The law is expressed by the conditional probability statement

$$p\big(y(k) = y_i | x(k) = x_j\big) = d_{ji}(y)$$

If we are given the values $y(0), y(1), \dots, y(k)$, what can we infer about $x(l)$? This problem, modified in various ways, is the subject of this chapter. The application of Bayes' rule will allow us to do the calculations we need. In order to see how this will work, consider the following basic calculation. Given $y(1) = y_i$, the probability that $x(1) = x_j$ is given by

$$p(x = x_j | y = y_i) = p(y = y_i | x = x_j) \cdot p(x = x_j)/p(y = y_j)$$

In vector notation

$$p(x|y) = p(y|x)p(x)/N$$

Where N is a normalization. The power of this approach comes from the fact that this can be used repeatedly together with the probability law for the evolution of x to solve a much more general class of problems. Letting $D = (d_{ji})$ as given above, and using $p(k+1) = Ap(k)$ we get

$$p(x(k+1) = x | y(k) = y_i) = D(y)Ap(x(k) = x)/N$$

Now suppose that the values $x(k)$ assumes are real numbers, $X = \{x_1, x_2, \dots, x_n\} \in \mathbb{R}^n$, and let x be a continuous time Markov process taking on values in X. Let w be a standard Wiener process. Suppose that $\tilde{p}_i$ is the probability that $x = x_i$ and suppose that

$$\frac{d}{dt}\tilde{p}(t) = A\tilde{p}(t)$$

Define y, a noisy observation of x by $dy = xdt + dw$. In 1965, Wonham [1] gave a derivation of the equation satisfied by the conditional probability $\rho(t, x|Y_t)$ where Y_t indicates the y process on the interval $[0, t]$. In terms of the notation

$$D = \begin{bmatrix} x_1 & 0 & \cdots & 0 \\ 0 & x_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & x_n \end{bmatrix} \quad ; \quad b = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Wonham's equation, expressed in Itô form, is

$$d\rho = A\rho + (D - \langle b, \rho\rangle I)\rho(dy - \langle b, \rho\rangle dt)$$

It is a straightforward exercise in the Itô calculus to verify that if $\Sigma\rho_i(0) = 1$ then $\Sigma\rho_i(t) = 1$ for all $t > 0$ and that if p satisfies the related Itô equation

$$dp = Apdt + Dpdy \tag{6.1}$$

then the elements of p_i remain nonnegative for $t > 0$ if they are initially nonnegative. Moreover, the equality

$$\rho(t) = p(t)/\Sigma p_i(t)$$

holds for all $t > 0$ if it holds initially. We call (6.1) the unnormalized conditional density equation. The purpose of this section is to use Bayes' rule to derive the conditional density equation.

To begin with, we look at a discrete time case. Consider

$$\begin{aligned} p(k+1) &= Ap(k) \\ y(k) &= x(k) + n(k) \end{aligned}$$

with the elements of random sequence $\{n(k)\}_{k=0}^{\infty}$ being independent for distinct values of k but distributed according to the same density, $f(n)$. According to Bayes' rule, if the probability density for x before the observation is p, then after an observation of value y we have (recall that $n = y - x$)

$$\begin{aligned} p(x|y) &= \bar{p}(y|x) \cdot p(x)/\bar{p}(y) \\ &= f(y - x) \cdot p(x)/\bar{p}(y) \end{aligned}$$

with $\bar{p}$ being the probability density of y. This can be written in vector notation as

$$\begin{bmatrix} p(x_1|y) \\ p(x_2|y) \\ \vdots \\ p(x_n|y) \end{bmatrix} = (1/\bar{p}(y)) \begin{bmatrix} f(y - x_1) & 0 & \cdots & 0 \\ 0 & f(y - x_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & f(y - x_n) \end{bmatrix} \begin{bmatrix} p(x_1) \\ p(x_2) \\ \vdots \\ p(x_n) \end{bmatrix}$$

If we insert $p(x(k))$ in place of $p(x)$ and then use $p(k) = Ap(x(k-1))$, we see that a propagation by A followed by an observation gives

$$p(x|y \text{ values up to } k) = F(y)\, A\, p(x|y \text{ values up to } k\text{-}1)/(1/\bar{p}(y))$$

with $F(y)$ being the diagonal matrix appearing in the previous equation. This is a key equation. In terms of an abbreviated notation we can express the evolution for the conditional density as

$$\tilde{p}(k+1) = (1/\bar{p}(y))F(y(k))A\tilde{p}(k)$$

Notice that it is nonlinear in $\tilde{p}$ because $\bar{p}(y)$ is a nonlinear function of $\tilde{p}$. Nonetheless, it is, as we will see, nonlinear in a trivial way.

Example: If x takes on the values ± 1 and if n is a Gaussian random variable with zero mean and variance σ, this equation takes the form

$$\begin{bmatrix} p_1(k+1) \\ p_2(k+1) \end{bmatrix} = r \begin{bmatrix} e^{-(y-1)^2/2\sigma} & 0 \\ 0 & e^{-(y+1)^2/2\sigma} \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} p_1(k) \\ p_2(k) \end{bmatrix}$$

with r being the scalar which normalizes the sum of the entries of the vector on the right-hand side to one.

We now examine what happens when we attempt to capture continuous time phenomena as a limit of a set of discrete time equations. In the first place, if we allowed ourselves more and more unbiased independent measurements of a quantity, we would, by the law of large numbers, reduce the error variance to zero. More precisely, if we have measurements $y_i = x + n_i$ for $i = 1, 2, \ldots, m$ with the n_i being independent, zero mean, Gaussian random variables, each of variance σ, then

$$\mathcal{E} \frac{1}{m} \sum_{i=1}^{m} y_i = x$$

and

$$\begin{aligned} \mathcal{E} \left(\frac{1}{m} \sum_{i=1}^{m} y_i - x \right)^2 &= \frac{1}{m} \sum_{i=1}^{m} (n_i)^2 \\ &= \frac{\sigma}{m} \end{aligned}$$

On the other hand, if every time we double the number of measurements we multiply the variance by 2, then we do not change the variance of the error. The former model, according to which, we simply make more measurements, is therefore unsuitable as a way to model a noisy continuous time observation process whereas the latter could be appropriate.

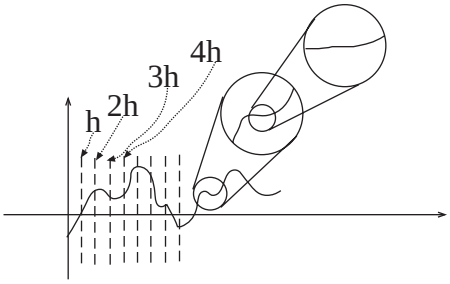

Figure 6.1. Illustrating the effect of dense sampling.

Consider observing $y = x + n$ with n being zero mean, Gaussian, and with variance σ. We have, as derived previously,

$$\begin{bmatrix} p_1(nh+h) \\ p_2(nh+h) \\ \vdots \\ p_n(nh+h) \end{bmatrix} = r \begin{bmatrix} e^{-(y-x_1)^2/2\sigma} & 0 & \cdots & 0 \\ 0 & e^{-(y-x_2)^2/2\sigma} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e^{-(y-x_n)^2/2\sigma} \end{bmatrix} A \begin{bmatrix} p_1(nh) \\ p_2(nh) \\ \vdots \\ p_n(nh) \end{bmatrix}$$

Consider replacing h by $h/2$ and σ by 2σ. In fact, consider doing it q times. After subtracting $p(nh)$ from both sides, we have as an approximation

$$\frac{h}{2^q}\frac{d}{dt}\,p(t) = \begin{bmatrix} (y-x_1)^2/\sigma(q) & 0 & \cdots & 0 \\ 0 & (y-x_2)^2/\sigma(q) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & (y-x_n)^2/\sigma(q) \end{bmatrix} A^{1/2q}\,p(t)$$

with $\sigma(q) = \sigma 2^q$. A suitable limiting process, letting $h\sigma = 1$, yields

$$\dot{p} = (\tilde{A} + yD - \tfrac{1}{2}D^2)p$$

where $\tilde{A}$ is the logarithm of A.

6.2 The Conditional Density Equation - I

Let us now consider the problem

$$\begin{aligned} \dot{p}(t) &= Ap(t) \\ p_i(t) &= \text{the probability that } x(t) = x_i \\ dy &= x\,dt + d\nu \end{aligned}$$

where ν is a standard Wiener process. From above we see that by using a central difference approximation for the time derivative we get for the conditional density (in unnormalized form)

$$đ\rho = (A - \tfrac{1}{2}D^2)\rho(t)dt + D\rho đy$$

with A being as above and D being the diagonal matrix

$$D = \begin{bmatrix} x_1 & 0 & \cdots & 0 \\ 0 & x_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & x_n \end{bmatrix}$$

Example: Consider the estimation of the values of a random telegraph wave

$$\begin{aligned} dx &= -2x dN \;\;;\;\; x(0) = +1 \\ dy &= x\,dt + d\nu \end{aligned}$$

The conditional density equation in unnormalized form is

$$\begin{bmatrix} dp_1 \\ dp_2 \end{bmatrix} = \begin{bmatrix} -\lambda dt - 1/2dt + dy & \lambda dt \\ \lambda dt & -\lambda dt - 1/2dt - dy \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \end{bmatrix}$$

6.3 The Conditional Density Equation - II

Now consider the model

$$\begin{aligned} dx &= f(x)dt + g(x)dw \\ dy &= h(x)dt + d\nu \end{aligned}$$

What is the corresponding conditional density equation? Without becoming too involved in the details, we note that in the previous section it was seen that the unnormalized conditional density equation involves two operators; A, which governs the evolution of the probability of the unobserved process, and D, which is a diagonal operator which multiplies the probability associated with x by the value that the observation takes on at x.

Reasoning via this analogy we take A to be the Fokker-Planck operator

$$A = -\frac{\partial}{\partial x} f(x) + \frac{1}{2} \sum \frac{\partial}{\partial x_i} \frac{\partial}{\partial x_j} g_i(x) g_j(x)$$

and let D be

$$D = \text{multiplication by } h(x)$$

Putting these ideas together we see that if we denote the Fokker-Planck operator by L then

$$\frac{\partial \rho(t,x)}{\partial t} = (L - \tfrac{1}{2} h^2(x))\rho + \frac{dy}{dt} h(x) \rho$$

is the central difference (Stratonovic) version of the conditional density equation for a diffusion process.

Example: Consider the simplest problem

$$\begin{aligned} dx &= dw \\ dy &= x\, dt + d\nu \end{aligned}$$

The conditional density equation in unnormalized form is

$$\frac{\partial \rho}{\partial t} = \left(\frac{1}{2} \frac{\partial^2}{\partial x^2} - \frac{1}{2} x^2 \right) \rho(t,x) + \frac{đy}{dt} x \rho(t,x)$$

If we assume a solution of the form

$$\rho(t,x) = e^{a(t)x^2 + b(t)x + c(t)}$$

we can derive equations for a, b, and c.

6.4 An Exponential Representation

There exist representations of solutions of differential equations that will let us establish a connection between the unnormalized conditional density equation and a certain Lie algebra. This material is most explicit in Wei-Norman, an earlier paper by Chen covers similar ground, and the basic ideas could probably be traced back at least to Lie and Cartan.

To begin with, consider the finite dimensional linear equation:

$$\dot{x} = (uA + vB)x$$

with u and v functions from $\mathbb{R}^1$ to $\mathbb{R}^1$ and A and B constant $n \times n$ matrices. Naively one might expect to find that the fundamental solution $\Phi(\cdot)$ is

$$\Phi(t) = e^{\left(\int_0^t u(\sigma) d\sigma\right) A + \left(\int_0^t v(\sigma) d\sigma\right) B} =$$

$$I + \left(\int_0^t u(\sigma)d\sigma\right) A + \left(\int_0^t v(\sigma)d\sigma\right) B + \frac{1}{2}\left(\left(\int_0^t u(\sigma)d\sigma\right) A + \left(\int_0^t v(\sigma)d\sigma\right) B\right)^2 + \dots$$

However,

$$\dot{\Phi}(t) = (uA + vB) + \frac{1}{2}(uA + vB) \cdot \int_0^t (uA + vB)d\sigma + \frac{1}{2}\left(\int_0^t (uA + vB)d\sigma\right)(uA + vB)d\tau + \dots$$

and in general it is not possible to factor out $(uA + uB)$ from this expression because A and B do not necessarily commute. Thus the above expression for Φ does not work. However, we can use the matrix identity

$$\begin{aligned} e^{-A}Be^{A} &= (1 + A + \tfrac{1}{2!}A^2 + \dots)B(1 - A + \tfrac{1}{2}A^2 \dots) \\ &= B + AB - BA + \tfrac{1}{2}(A^2B - 2ABA + BA^2) + \dots \\ &= B + [A, B] + \tfrac{1}{2}[A, [A, B]] + \dots \end{aligned}$$

This is sometimes called the Baker-Cambell-Hausdorff formula. Introduce the notation

$$ad_A^k B = [A, [A, [A, \dots [A, B]] \dots]; \quad k \geq 1$$
$$k \text{ times}$$
$$ad_A^0 B = B$$

For each choice of A, $ad_A^k(B)$ is a linear operator acting on B. It is common to express the above relationship as

$$\begin{aligned} \exp ad_A B &= ad_A^0 B + ad_A^1 B + \tfrac{1}{2!}ad_A^2 B + \dots \\ &= B + [A, B] + \tfrac{1}{2!}[A, [A, B]] + \dots \end{aligned}$$

and write

$$e^{A}Be^{-A} = \exp ad_A B$$

Wei and Norman investigated the differential equation

$$\dot{x} = (\sum_{i=1}^{k} u_i B_i)x$$

by looking for a solution $\Phi(t)x_0$ which can be represented as a product of exponentials

$$\Phi(t)x_0 = e^{g_1 A_1} e^{g_2 A_2} \dots e^{g_m A_m} x_0$$

in which $g_1, g_2, \dots, g_m$ are real valued functions of time and the A_i are somehow generated by the B_i. Without being specific about the latter process, we may differentiate to get

$$\frac{d}{dt}(e^{g_1 A_1} \dots e^{g_m A_m}) = \dot{g}_1 A_1 e^{g_1 A_1} \dots + e^{g_m A_m} + e^{g_1 A_1} \dot{g}_2 A_2 e^{g_2 A_2} \dots e^{g_m A_m} + \dots$$

$$e^{g_1 A_1} \dots e^{g_{m-1} A_{m-1}} \dot{g}_m A_m e^{g_m A_m}$$

Inserting exponentials and their inverses we can transform this into an expression in which all terms have a common factor $e^{g_1 A_1} \dots e^{g_m A_m}$ on the right. That is,

$$\frac{d}{dt}(e^{g_1 A_1} \dots e^{g_m A_m}) = \left(\dot{g}_1 A_1 + e^{g_1 A_1} \dot{g}_2 a_2 e^{-g_1 A_1} + \dots\right)\left(e^{g_1 A_1} \dots e^{g_m A_m}\right)$$

Applying the Baker-Campbell-Hausdorff-formula this can be written as:

$$\dot{g}_1 A_1 + \dot{g}_2(A_2 + g_1[A_1, A_2] + \frac{1}{2}g_1^2[A_1, [A_1, A_2]] + \dots) + \dots + \dot{g}_m \psi$$

with ψ being an expression containing the matrices $A_1, A_2, ...A_k$ and various products. This results in a set of differential equations for the g_i provided that the A_i are chosen so that any B_i is a linear combination of the A_i, any $[B_i, B_j]$ is a linear combination of the A_i, any $[B_i, [B_j, B_k]]$ is a linear combination of the A_i, etc. The smallest set of A's with this property forms a basis for a matrix Lie algebra and will be referred to as the *Lie algebra* generated by the A_i.

If the B_i are, themselves, linearly independent as elements of the vector space consisting of all real matrices of a certain size, then we may take $A_1 = B_1$, $A_2 = B_2$ etc for the first k elements of set A_i. Matters being so, and assuming that the A_i are a basis for the Lie algebra generated by the B_i, we have

$$[A_i, A_j] = \sum_k \gamma_{ijk} A_k$$

with certain coefficients γ_{ijk}. These are the so-called *structure constants* of the Lie algebra. If $\Phi(t)x_0$ satisfies the differential equation we must have

$$\dot{g}_1 A_1 + \dot{g}_2(A_2 + g_1[A_1, A_2] + \frac{1}{2}g_1^2[A_1, [A_1, A_2]] + \dots) + \dots$$

$$+ \dots + \dot{g}_m(\dots) = u_1 A_1 + \dots u_m A_m$$

and because the A_i are independent as vectors in $R^{n\times n}$ we get, on equating coefficients a set of equations of the form

$$\begin{aligned} \dot{g}_1 &= f_1(g_1, \dots, g_m, \quad \dot{g}_2, \dots, \dot{g}_m) + u_1 \\ \dot{g}_2 &= f_2(g_1, \dots g_m, \quad \dot{g}_2, \dots, g_m) + u_2 \\ &\vdots \\ \dot{g}_m &= f_m(g_1, \dots, g_m, \quad \dot{g}_2, \dots, g_m) + u_m \end{aligned}$$

We will refer to these as the Wei-Norman equations. They depend on the choice of a particular ordering of the exponential factors. Because $\Phi(0) = I$ we have initial conditions $g_1(0) = g_2(0) = \dots = g_m(0)$. An analysis shows that the Wei-Norman equations can always be solved on some interval $|t| \leq \epsilon$ however in most cases the solution cannot be continued for all time. A significant point is that the functions $f_1, \dots, f_m$ only depend on the structure constants $Upsilon_{ijk}$. That is, regardless of the representation of the Lie algebra we get the same Wei-Norman equations. We have here a situation such that by solving one set of nonlinear differential equations we simultaneously solve a whole family of linear evolution equations.

Example:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} a(t) & c(t) \\ 0 & b(t) \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \left[a(t) \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + b(t) \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} + c(t) \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \right] \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

Let $\mathcal{L}$ be the Lie algebra generated by

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \text{ and } \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$

We choose an ordered basis for $\mathcal{L}$:

$$A_1 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, A_2 = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \text{ and } A_3 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

The differential equations can now be written

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = (A_1\eta_1 + A_2\eta_2 + A_3\eta_3) \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

with $\eta_1 = a - b, \eta_2 = c, \eta_3 = b$. If we look for a fundamental solution having the form

$$\Phi(t) = e^{A_1 g_1} e^{A_2 g_2} e^{I g_3}$$

Then

$$\dot{\Phi} = (A_1 \dot{g}_1 + e^{A_1 g_1} A_2 \dot{g}_2 e^{-A_1 g_1} + \dot{g}_3 I) e^{A_1 g_1} e^{A_2 g_2} e^{I g_3}$$

Because we have a basis with $A_3 = I$ the expression for $\dot{\Phi}$ contains the term $\dot{g}_3 \Phi$ instead of a term

$$e^{A_1 g_1} e^{A_2 g_2} \dot{g}_3 A_3 e^{-g_2 A_2} e^{-g_1 A_1}$$

that is

$$\dot{\Phi} = (A_1 \dot{g}_1 + \dot{g}_2 (A_2 + g_1 [A_1, A_2] + \frac{1}{2} g_1^2 [A_1, [A_1, A_2]] + \dots) + \dot{g}_3 I) e^{A_1 g_1} e^{A_2 g_2} e^{I g_3}$$

Now $[A_1, A_2] = A_2$ so that

$$\begin{aligned} \dot{\Phi} &= (A_1 \dot{g}_1 + \dot{g}_2 (A_2 + g_1 A_2 + \tfrac{1}{2} g_1^2 A_2 + \dots) + \dot{g}_3 I) e^{g_1 A_1} e^{g_2 A_2} e^{g_3 I} \\ &= (A_1 \dot{g}_1 + \dot{g}_2 e^{g_1} A_2 + \dot{g}_3 I) e^{g_1 A_1} e^{g_2 A_2} e^{g_3 I} \end{aligned}$$

So

$$A_1 \dot{g}_1 + \dot{g}_2 e^{g_1} A_2 + \dot{g}_3 I = \eta_1 A + \eta_2 A_2 + \eta_3 I$$

The Wei-Norman equations become

$$\begin{aligned} \dot{g}_1 &= \eta_1 = a - b \\ \dot{g}_2 &= e^{-g_1} \eta_2 = c e^{-g_1} \\ \dot{g}_3 &= \eta_3 = b \end{aligned}$$

The differential equations can be solved directly. By solving them we do not only find a fundamental solution of the particular set of equations but also a fundamental solution associated with any family of operators that commute according to the same commutation relations.

6.5 Conditional Jump Processes

Let x be a stochastic process taking on values in the set $X = \{x_1, x_2, ..., x_n\}$ and let

$$\dot{p} = Ap$$

describe the evolution of the probability law for x. That is, $p_i(t)$ is the probability that the process $x(t)$ is in the state x_i. Consider a second continuous time jump process $y(t)$ taking on values in the set $Y = \{y_1, y_2, ..., y_m\}$ with a probability law that satisfies the equation

$$\dot{q} = B(x)q$$

That is, $q_i(t)$ is the probability that $y(t)$ is in state y_i. The problem to be solved is this. Given $p(0), A$, and $B(\cdot)$, find the conditional probability of $x(t)$ given $y(\sigma)$ for $0 \leq \sigma \leq t$.

We are observing y. To infer something about x, from a knowledge of y we can use Bayes' rule. We look at the behavior of y over a small interval of time, the interval $[0, \Delta]$. If over this interval y is constant, say $y(t) = y_i$, then

$$p(x|y) = p(y|x) \cdot p(x)/p(y)$$

Thus we may write

$$p(x(\Delta)|y) = \frac{1}{N} e^{(\mathrm{Diag}(B_{yy}(x_1)\dots B_{yy}(x_n))+A)\Delta} p(0)$$

Where $B_{yy}(x_i)$ is the diagonal element of $B(x)$ corresponding to the value that the y process takes on over the interval $[0, \Delta]$ and n is a normalization. On the other hand, if on $[0, \Delta]$ the y process jumps from y_i to y_j then, to within a normalization,

$$p(x|\text{ jump from } i \text{ to } j) \approx p\,(\text{jump from } i \text{ to } j|x) \cdot p(x)$$

At the moment of the jump, p changes according to

$$p(x|\text{ jump from } i \text{ to } j) = \mathrm{Diag}(B_{y_+y_-}(x_1)\dots B_{y_+y_-}(x_n)) \cdot p(x)$$

In order to incorporate this into a stochastic differential equation, it is convenient to assign y_i to a point in a vector space. For convenience we identify y_i with e_i, the i^{th} standard basis element in $\mathbb{R}^m$. Thus $\langle e_j, y\rangle$ jumps from 1 to zero if y jumps away from state j and jumps from zero to one if y jumps to state j. We seek a differential equation of the form

$$dp = (A + \mathrm{Diag}B_{yy}(x))pdt + \sum_{i\neq j} D_{ij} p\phi_j(y)\langle e_i, dy\rangle$$

for the evolution of p. Of course at a jump from i to j the quantity $\langle e_i, y\rangle$ decreases by one and $\langle e_j, y\rangle$ increases by one. Thus p goes to

$$p \mapsto \frac{1}{N}(p + \sum_{i\neq j} D_{ij}\phi_j(y)\langle e_i, dy\rangle)$$

By taking the D_{ij} to be the appropriate diagonal matrix and $\phi_j(y) = \langle e_j, y\rangle$, we alter p by

$$p \mapsto p + D_{ij}p$$

when y goes from y_j to y_i. Thus the conditional density is

$$dp = (A + B_{yy})pdt + \sum_{i\neq j}(B_{ij} - I)\langle e_j, y\rangle e_i dy$$

If x is a diffusion process then this equation takes a similar form.

Example: Consider a continuous time jump process that takes on the values ± 1 and is generated by the Itô equation

$$dx = -2xdN$$

Let q satisfy

$$\frac{d}{dt}\begin{bmatrix} q_1 \\ q_2 \\ q_3 \end{bmatrix} = \begin{bmatrix} -3+x & 1 & 2 \\ 1-x & -3 & 1+x \\ 2 & 2 & -3-x \end{bmatrix}\begin{bmatrix} q_1 \\ q_2 \\ q_3 \end{bmatrix} dt$$

If we assume that $y(t)$ takes on the values in the set $Y = \{e_1, e_2, e_3\}$ then we can construct the conditional density equation as above.

6.6 Extrapolation and Smoothing

It often happens that it is desireable to delay making an estimate of a process until further data has been received. This leads to a problem of the smoothing type. In its basic form,

the smoothing problem is that of determining the conditional probability distribution for $x(t_s)$, given $y(\cdot)$ over $[t_0, t_f]$ with $t_0 < t_s < t_f$. In this section we formulate problems of this type mathematically and derive the appropriate evolution equation for the conditional density.

Consider first a discrete time situation. Suppose that the probability that $x(0) = x_i$ is $p_i(0)$ and that in the absence of observations $p(k+1) = Ap(k)$. We have seen that at some future time the conditional probability can be obtained by normalizing

$$p(k) = D(y(k))AD(y(k))A...D(y(1))Ap(0)$$

with D being ther diagonal matrix defined in section 6.1.

In chapter two we discussed the problem of evolving the probability vector backwards in time and introduced the operator $A_B = A^T S$ with S being a diagonal matrix chosen so that $A^T S$ has columns that sum to one. Now suppose that we know that $x(t_f) = x_i$. In that case the probability distribution for x_{f-1} is $A_B e_i$. After we incorporate the observation at $t = t_{f-1}$ the probability changes to

$$p(k-1) = \frac{1}{N}(D(y(k-1))A_B e_i$$

Continuing in this way, we see that

$$\tilde{p}(t_s) = \frac{1}{N}(D(y(t_s))A_B...D(y(k-1))A_B e_i$$

We can now put these two calculations together using Bayes' rule. The first calculation gives the probability distribution of $x(t_f)$. The second gives that of $x(t_s)$, assuming that we know $x(t_f)$). Multiplying these together and normalizing, we have the desired conditional probability. We may express matters as

$$p(t_s|y(t),\ t_0 < t < t_f) = \frac{1}{N}\tilde{p}(t_s) \odot D(y(k))AD(y(k))A...D(y(1))Ap(0)$$

or

$$p(t_s|y(t),\ t_0 < t < t_f) = \frac{1}{N}\tilde{p}(t_s) \odot p(t_f)$$

We can adapt this analysis to continuous time and infinite state spaces. For example, in the case of continuous time jump processes the backwards equation for $\tilde{p}$ is simply

$$d\tilde{p} = (A_B - \frac{1}{2}D^2)\tilde{p}dt + D\tilde{p}dy$$

6.7 The Orthogonality Principle

In the study of inner products on functions spaces one often uses the definition

$$\langle f(\cdot), g(\cdot)\rangle = \int f(x)g(x)d\mu$$

In those cases where $d\mu$ is a weighted version of Borel measure one often writes

$$\langle f(\cdot), g(\cdot)\rangle = \int f(x)g(x)w(x)dx$$

Thus far we have concentrated on the problem of propagating the conditional density and have postponed all questions of what to do with the conditional density once we have it. In some

cases the goal of the analysis is to estimate the value of some function of the state, say $h(x)$. Naturally

$$\mathcal{E}h(x) = \int h(x)\rho(t,x)dx$$

where ρ is the conditional density. The following fact is basic.

Theorem: The estimate e that minimizes the expected value of the square of the eror between $h(x)$ and its estimate, is the conditional expectation,

$$e = \int h(x)\rho(t,x)dx$$

Proof: In order to prove this we make the following observations. If we observe y and if z is a casual functional of y and if e is y-measurable then so is $e + \epsilon z$ for all ϵ. If e minimizes

$$\mathcal{E}(e - h(x))^2$$

then we see that

$$\mathcal{E}(e + \epsilon z - h(x))^2 \geq \mathcal{E}(e - h(x))^2$$

¿From an analysis of this inequality in a neighborhood of zero we see that

$$\int z(e - h(x))\rho(t,x) = 0$$

That is to say, the error is orthogonal to any past measurable function of the observations.

6.8 Linear Estimation Theory

In some parts of the literature on stochastic differential equations one sees a notation in which a Stratonovic equation is divided by dt and $đw/dt$ is written as $\dot{w}$. This allows for simpler typography and we will use it in this section.

Consider a stochastic process y which is generated from a unity variance Wiener process w according to the equations

$$\dot{x}(t) = A(t)x(t) + B(t)\dot{w}(t) \;\; ; \;\; \dot{y}(t) = C(t)x(t)$$

Now suppose that we observe not $\dot{y}(t)$ itself but $\dot{y}(t) + \dot{v}(t)$ where v is a zero mean, unity variance, Wiener process. The question arises as to how, if at all, we can remove the effects of the noise v.

Lemma: A stochastic equation of the form

$$\dot{z}(t) = F(t)z(t) + H(t)(\dot{y}(t) + \dot{v}(t))$$

has a solution z which has the same expectation as x if and only if $\mathcal{E}z(0) = \mathcal{E}x(0)$ and the matrices F and H are chosen so that the equation takes the form

$$\dot{z}(t) = A(t)z(t) - G(t)[C(t)(x(t) - z(t)) + \dot{v}(t)]$$

Proof: This is a direct application of the rule for computing expectations.

If we seek an estimate which can be generated by a linear stochastic equation in the form described in the lemma, and if we want an unbiased estimate we must constrain the equations

for z in this way. Now from this point on there remains only the question of picking the best choice of G. Notice that if we introduce $e = x - z$, then

$$\dot{e}(t) = A(t)e(t) + G(t)[C(t)e(t) + \dot{v}(t)] + B(t)\dot{w}(t)$$

Writing down the variance equation (c.f. section 2.7) we have

$$\begin{aligned}\dot{\Sigma}_{ee}(t) &= A(t)\Sigma_{ee}(t) + \Sigma_{ee}(t)A^T(t) + G(t)C(t)\Sigma_{ee} \\ &\quad + \Sigma_{ee}(t)C^T(t)G^T(t) + G(t)G^T(t) + B(t)B^T(t)\end{aligned}$$

Now compare this with the solution of the equation

$$\dot{S}(t) = A(t)S(t) + S(t)A^T(t) - S(t)C^T(t)C(t)S(t) + B(t)B^T(t)$$

A little manipulation gives

$$\begin{aligned}\dot{\Sigma}_{ee}(t) - \dot{S}(t) &= (A(t) + G(t)C(t))(\Sigma_{ee}(t) - S(t)) + (\Sigma_{ee}(t) - S(t))(A(t) + G(t)C(t))^T \\ &\quad + G(t)G^T(t) + G(t)C(t)S(t) + S(t)C^T(t)G^T(t) + S(t)C^T(t)C(t)S(t)\end{aligned}$$

If we set $\Sigma_{ee}(t_0) = S(t_0)$, then we have

$$\Sigma_{ee}(t) - S(t) = \int_{t_0}^{t} \Phi_{A+GC^T}(t,\sigma)[G(\sigma) + S(\sigma)C^T(\sigma)][G(\sigma) + S(\sigma)C^T(\sigma)]^T \Phi^T_{A+GC^T}(t,\sigma)d\sigma$$

This shows that $\Sigma_{ee}(t) - S(t)$ is nonnegative definite and is zero if and only if $G(t) = -S(t)C^T(t)$. Thus we see that the minimum variance estimate of x is generated by

$$d\hat{x} = A\hat{x}dt + \Sigma_{ee}(t)C^T(dy - C\hat{x}dt)$$

Theorem: If z satisfies the equations

$$\dot{z}(t) = A(t)z(t) + S(t)C^T(t)(y(t) - C(t)z(t) + \dot{v}(t))$$

with $S(t)$ satisfying

$$\dot{S}(t) = A(t)S(t) + S(t)A^T(t) - S(t)C^T(t)C(t)S(t) + B(t)B^T(t) \;\; ; \;\; S(t_0) = \Sigma_{ee}(t_0)$$

then z is the minimum variance estimate of x.

There is one additional property of the optimal solution which we need later in the discussion of the separation theorem.

Theorem: The optimum error variance and the optimal estimate variance satisfy

$$\Sigma_{xx}(t) = \Sigma_{ee}(t) + \Sigma_{\hat{x}\hat{x}}(t)$$

or equivalently

$$\mathcal{E}\hat{x}(t)e^T(t) = 0$$

where $\Sigma_{xx} = \mathcal{E}xx^T$, $\Sigma_{\hat{x}\hat{x}} = \mathcal{E}\hat{x}\hat{x}^T$.

To see that the second condition is equivalent to the first observe that

$$\begin{aligned}x(t)x^T(t) &= \mathcal{E}(\hat{x}(t) + e(t))(\hat{x}(t) + e(t))^T \\ &= \mathcal{E}\hat{x}(t)\hat{x}(t) + \mathcal{E}e(t)e^T(t) + \mathcal{E}(\hat{x}(t)e^T(t) + e(t)\hat{x}^T(t))\end{aligned}$$

and that the third term is zero if and only if $\mathcal{E}\hat{x}e^T$ is zero. To see that $\mathcal{E}\hat{x}e^T = 0$ we observe that

$$\begin{bmatrix} \dot{\hat{x}}(t) \\ \dot{e}(t) \end{bmatrix} = \begin{bmatrix} A(T) & \Sigma_{ee}(t)C^T(t)C(t) \\ 0 & A(t) - \Sigma_{ee}(t)C^T(t)C(t) \end{bmatrix} \begin{bmatrix} \hat{x}(t) \\ e(t) \end{bmatrix} + \begin{bmatrix} \Sigma_{ee}(t)C^T(t) & 0 \\ \Sigma_{ee}(t)C^T(t) & B(t) \end{bmatrix} \begin{bmatrix} \dot{v} \\ \dot{w} \end{bmatrix}$$

Now the resulting variance equation is

$$\begin{aligned} \frac{d}{dt}\mathcal{E}\begin{bmatrix} \hat{x}\hat{x}^T & \hat{x}e^T \\ e\hat{x}^T & ee^T \end{bmatrix} &= \begin{bmatrix} A(t) & \Sigma_{ee}(t)C^T(t)C(t) \\ 0 & A(t) - \Sigma_{ee}(t)C^T(t)C(t) \end{bmatrix} \mathcal{E}\begin{bmatrix} \hat{x}\hat{x}^T & \hat{x}e^T \\ e\hat{x}^T & ee^T \end{bmatrix} \\ &+\mathcal{E}\begin{bmatrix} \hat{x}\hat{x}^T & \hat{x}e^T \\ e\hat{x}^T & ee^T \end{bmatrix} \begin{bmatrix} A(t) & \Sigma_{ee}(t)C^T(t)C(t) \\ 0 & A(t) - \Sigma_{ee}(t)C^T(t)C(t) \end{bmatrix}^T \\ &+ \begin{bmatrix} \Sigma_{ee}(t)C^T(t)C(t)\Sigma_{ee}(t) & \Sigma_{ee}(t)C^T(t)C(t)\Sigma_{ee}(t) \\ \Sigma_{ee}(t)C^T(t)C(t)\Sigma_{ee}(t) & B(t)B^T(t) + \Sigma_{ee}(t)C^T(t)C(t)\Sigma_{ee}(t) \end{bmatrix} \end{aligned}$$

A short calculation will verify that the solution of this is given by

$$\begin{bmatrix} H(t) & 0 \\ 0 & \Sigma_{ee}(t) \end{bmatrix}$$

where $H(t)$ is $\mathcal{E}\hat{x}(t)\hat{x}^T(t)$. This calculation depends on verifying that we may take

$$H(t) = \mathcal{E}x(t)x^T(t) - \mathcal{E}e(t)e^T(t)$$

where, of course, $\mathcal{E}ee^T = \Sigma_{ee}$ and

$$\begin{aligned} \frac{d}{dt}\mathcal{E}xx^T &= A\mathcal{E}xx^T + \mathcal{E}xx^TA^T + BB^T \\ \frac{d}{dt}\Sigma_{ee} &= A\Sigma_{ee} + \Sigma_{ee}A^T - \Sigma_{ee}C^TC\Sigma_{ee} + BB^T \end{aligned}$$

Example: Consider the two different models for signal and observation given below. Evaluate the variance of the steady-state estimation error in each case.

$$dx = -2xdt + dw \;\; ; \;\; dy = xdt + d\nu$$

$$dx = -3xdt + 1.5dw \;\; ; \;\; dy = xdt + d\nu$$

In which case is the estimation error larger and what is a qualitative explanation in terms of the power spectrum? To answer this, begin with a calculation of the covariance equation associated with the first model

$$\dot{\sigma}_1 = -4\sigma_1 + 1 - \sigma_1^2$$

and for the second model

$$\dot{\sigma}_2 = -6\sigma_2 + 2.25 - \sigma_2^2$$

We can calculate the steady state by setting the derivative equal to zero.

$$\sigma_1 = \frac{-4 + \sqrt{5}}{2} = \sqrt{5} - 2$$

$$\sigma_2 = \frac{-6 + \sqrt{45}}{2} = \frac{3}{2}(\sqrt{5} - 2)$$

Comparing these we have

$$\sigma_2 - \sigma_1 = \frac{1}{2}(\sqrt{5} - 2)$$

Because both systems 1 and 2 are stable there is a steady state for the unobserved process. The steady state variance of the first system when unobserved is 1/4; that of the second system is 2.25/6 which is larger. The signal to noise ratio of the observation process is the same in both cases so we may expect that the estimation error will be larger in the second case.

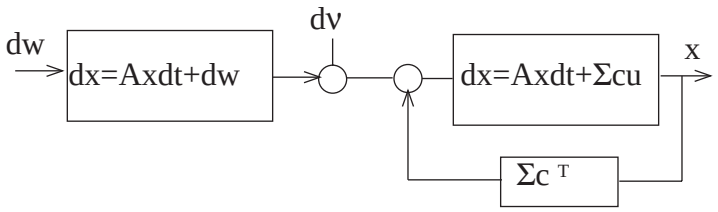

Figure 6.2.Block Diagram of Optimal Filter.

Smoothing: Consider the problem

$$dx = -xdt + dw \ ; \ dy = xdt + d\nu$$

Suppose that x is distributed with a gaussian density with $\mathcal{E}x(0) = 1$ and $\mathcal{E}(x(0) - 1)^2 = 2$.

(a) Describe a process for generating the conditional expectation of $x(1)$ given $y(t)$ for $0 \leq t \leq 1$.

(b) Describe a process for generating the conditional expectation of $x(\alpha)$ given $y(t)$ for $0 \leq t \leq 1$ and $0 \leq \alpha \leq 1$. In both cases "describe a process" means provide the appropriate differential equations.

The Kalman-Bucy filter will generate the conditional mean and the variance.

$$d\hat{x}(t) = -\hat{x}(t) + p(t)(dy(t) - \hat{x}(t)dt) \ ; \ \hat{x}(0) = 1$$

$$\dot{p} = -2p + 1 - p^2 \ ; \ p(0) = 2$$

If we use these equations on a forward sweep to generate the conditional density at $t = 1$ then we can run the related equation

$$d\hat{s}(t) = -\hat{s}(t) + q(t)(dy(1-t) - \hat{s}(t)dt)$$

together with

$$\dot{q} = -2q + 1 - q^2 \ ; \ q(1) = p(1)$$

Example: Consider the problem of generating the conditional density for x when x satisfies

$$dx_1 = x_2 dt + dw$$

$$dx_2 = 3x_1 dt - x_2 dt$$

and the observation equation is

$$dy = x_1 dt + ax_2 dt + d\nu$$

Assuming Gaussian initial data, display the differential equation for the error variance. Will the the solution fail to have a steady state value for any choice of a? Explain this.

Solution: the right-hand side of the error covariance equation is

$$\begin{bmatrix} 0 & 1 \\ 3 & -1 \end{bmatrix} \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} + \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} \begin{bmatrix} 0 & -3 \\ 1 & -1 \end{bmatrix} - \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} \begin{bmatrix} 1 & a \\ a & a^2 \end{bmatrix} \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

The A matrix has an eigenvalue with a positive real part. Thus if the system fails to be observable for any value of a, the error covariance will go to infinity if the unstable mode is unobservable. Otherwise it will have a (finite!) steady state value. The system fails to be observable when

$$[c; cA] = \begin{bmatrix} 1 & a \\ 3a & 1-3a \end{bmatrix}$$

is singular and this happens when $1 - a - 3a^2 = 0$ One of these roots results in a value of a for which the unobservable mode is stable and in this case the variance has a limit. The other root results in the unstable mode being unobservable and in this case no limit exists.

6.9 Identification of Linear Systems

Control theory techniques generally require the *a priori* knowledge of an accurate model for the dynamical system which is to be controlled. In some situations this is hypothesis is invalid and it is necessary to develop algorithms to identify the system before, or while, control is being exercised. This is a major problem. In this section we address it in the special case of linear systems with constant coefficients. This special case is a good example of applying the tools we have developed.

The approach to system identification is based on the conditional density equation. A simple example should help to fix ideas. Consider the equation

$$dx = axdt + dw_1 \quad ; \quad dy = xdt + dw_2$$

together with the stochastic equation

$$da = 0$$

What does this mean? If we had set da equal to cdw_3 for c nonzero, then this pair of equations would have described the evolution of the pair (a, x). The observations would have enabled us to learn something about the probability distribution of the pair. In the special case where c is zero, a is a constant but that does not mean that a is known! What this models is the situation where a is a random variable as opposed to being a full-fledged stochastic process. The initial distribution for (a, x) expresses our *a priori* knowledge and the observation of y allows us to improve on this.

The conditional density equation gives us a method to compute the joint conditional probability $\rho(t, a, x)$. In order to obtain the conditional probability for a alone, it is necessary to integrate with respect to x. Applying the general results of the previous chapter we have

$$\frac{\partial \rho(t,a,x)}{\partial t} = \left(-\frac{\partial}{\partial x} ax + \frac{1}{2}\frac{\partial^2}{\partial x^2} - \frac{1}{2}x^2 \right) \rho(t,a,x) + \frac{dy}{dt} x \rho(t,a,x)$$

According to the general philosophy we have been espousing here, the first thing to do is to compute the Lie algebra generated by the pair of operators.

In carrying out this computation there is one point which requires attention. Suppose that the algebra contains an operator L and also contains aL. Are these linearly independent elements of the Lie algebra? Unfortunately they are. Even though a is a constant and the operator "multiplication by a" remains just that and must not be confused with multiplication by a constant. Proceeding then, we see that the relevant Lie algebra here contains all of the following

$$a \ , \ \ -ax + \frac{\partial}{\partial x} \ , \ \ a^2 x - ax + \frac{\partial}{\partial x} \ , \ \ldots$$

and, in fact, is infinite dimensional.

Does this mean that there is nothing one can do with this estimation problem? Recall that the family of "reachable" conditional densities can be low dimensional either because the Lie algebra is low dimensional or because the initial condition is chosen in a suitable way. It is a general feature of conditional density equations which involve unknown parameters which do not change in time that the support of the conditional density does not increase as time evolves. If one knows *a priori* that a is an element of the set $S \subset R^1$, then the same is true for all time, regardless of the observation history. In particular, if we know that a is either -1 or -1.5, then the conditional density $\rho(t, a, x)$ can be computed rather easily.

6.10 Exercises 6

1. By a "Wiener filter" one usually means the steady state filter (state estimator) associated with a Gauss-Markov process observed with an additive noise term which is also a Gauss-Markov process. In terms of state variable models, we have

$$\begin{aligned} dx &= Axdt + bdw_1 \\ dn &= Fndt + gdw_2 \\ dy &= cxdt + hndt + dw_3 \end{aligned}$$

   set up the appropriate Riccati equation and estimation equation for the conditional mean (also least squares) estimate. Show that under appropriate assumptions about controllability, the relevant Riccati equation approaches a steady state as t goes to infinity.

2. Show that if $x(0)$ is distributed according to an arbitrary smooth density and satisfies the stochastic equation

$$dx = Axdt + bdw_1$$

   then provided that the eigenvalues of A have negative real parts, the probability density approaches a Gaussian density as t goes to infinity and this Gaussian density is independent of the initial density. Establish a similar result for the conditional density that goes along with the observation equation

$$dy = xdt + dw_2$$

3. Consider $2n$ by $2n$ dimensional matrices of the form

$$\begin{bmatrix} A & Q \\ R & -A^T \end{bmatrix}$$

   with Q and R symmetric. Show that these form a Lie algebra.

4. At $t = 0$ one knows that $x = \dot{x} = 0$ and over the interval $0 \leq t \leq t_0$ we observe $x(t) + \dot{v}(t)$ where $v(t)$ is a unity-variance Wiener process. Suppose that

$$\ddot{x}(t) + x(t) = \dot{\omega}(t)$$

   where ω is also a unity variance Wiener process. Find a filter which computes an unbiased, minimum variance estimate of x and $\dot{x}$, based on the observation $x(t) + \dot{v}(t)$. Examine the limiting case $t_0 \to \infty$; $t \to t_0$.

5. Suppose that $A \in R^{n\times m}$ is of rank m. Suppose that x is a zero mean Gaussian random variable. Let y be a given n-vector. Find v such that

$$\eta = \mathcal{E}\|Av - y + x\|^2$$

is minimized. Find the minimizing value of η.

6. Given that x and y are uncorrelated Gaussian random variables with mean $\bar{x}$ and $\bar{y}$ and variance Σ_{xx} and Σ_{yy}, find the mean and variance for $x + y$.

7. Consider the process x generated by

$$dx = Axdt + Bdw + \xi(t)dt \quad ; \quad x(0) = \text{Gaussian}$$

Where w is a standard m-dimensional Wiener process and ξ is a known function of time. If we observe

$$dy = cxdt + d\nu$$

With ν an n-dimensional Wiener process independent of w, write out in full detail a set of stochastic differential equations for the conditional mean given the observations. (These will look like the Kalman-Bucy equations with a suitable modification.)

8. Suppose that we have

$$dx = dw \quad ; \quad dy = csdt + dw + d\nu$$

where w and ν are independent Wiener processes. Derive the values of the coefficients of the unbiased, minimum variance estimator assuming the optimal filter to be one dimensional and linear.

9. Consider the system

$$dx = \alpha xdt + dw + budt \quad ; \quad dy = xdt + d\nu$$

where α is either -1 or -2 and $x(0)$ is Gaussian, independent of α. Suppose that the *a priori* probability is 1/2 that $\alpha = -1$ and 1/2 that $\alpha = -2$. If u is zero, find a finite set of equations which accept input dy and propagate the conditional probability of α.

   (a) Find equations for the conditional mean of x given that $u(t)$ is zero.

   (b) Consider letting $u(t) = k(t)z(t)$ where

$$dz(t) = -\beta(t)zdt + \gamma(t)dy$$

   Write an equation for the conditional mean of α.

   (c) Discuss qualitatively which choices of k, β and γ make it easiest to determine α.

10. Write down the differential equations for the conditional mean and the conditional variance for

$$dx = Axdt + \alpha Bdw \quad ; \quad dy = \beta Cxdt + dv \quad ; \quad \alpha_9\beta > 0$$

This is standard except for the appearance of the real parameters α and β. Assume that rank $(B, AB, ...A_{n-1}B)$ and rank $(C; CA; ...CA_{n-1})$ are $n = \dim x$. Show that the steady state conditional variance $\Sigma(\infty)$ is a monotone increasing function of α and a monotone decreasing functions of β where $\Sigma_1 \geq \Sigma_2$ means that the symmetric matrix $\Sigma_1 - \Sigma_2$ is nonnegative definite. (Please avoid heuristic arguments.)

11. Let $x(t)$ be a Markov process which takes on values in the set of {1,2,3,4} and let $p_i(t)$ be the probability that $x(t) = i$; assume that $\dot{p} = Ap$ describes the evolution of the probabilities. Suppose that at t = 1,2,3,... one makes an observation of x and through this observation learns if x is larger than 2.5 or smaller than 2.5 (there is no uncertainty about the fact). Give a rule for propagating the conditional probability in this case.

12. Consider the setup

$$\begin{bmatrix} dx_1 \\ dx_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} dt + \begin{bmatrix} 0 \\ u \end{bmatrix} dt + \begin{bmatrix} 0 \\ dw \end{bmatrix}$$
$$dy = x_1 dt + dv$$

Suppose that $x(0)$ is gaussian. Find the control law which minimizes

$$\eta = \mathcal{E} \int_0^1 x_1^2(t) + u^2(t) dt$$

Give all the details.

13. Show that the Lie algebra generated by the operators

$$A = \begin{bmatrix} \frac{\partial^2}{\partial x^2} - x^2 - \lambda + \frac{\partial}{\partial x} x & \lambda \\ \lambda & \frac{\partial^2}{\partial x^2} - x^2 - \lambda + \frac{\partial}{\partial x} x \end{bmatrix}$$
$$B = \begin{bmatrix} x & 0 \\ 0 & x \end{bmatrix}$$

is infinite dimensional by showing that it contains the two operators

$$C = \begin{bmatrix} \frac{\partial}{\partial x} + x & 0 \\ 0 & \frac{\partial}{\partial x} - x \end{bmatrix}$$
$$D = \begin{bmatrix} x + \frac{\partial}{\partial x} & 0 \\ 0 & x - \frac{\partial}{\partial x} \end{bmatrix}$$

14. The Pauli equation provides a quantum mechanical description of a spinning particle. It takes the form

$$\frac{ih}{2\pi} \begin{bmatrix} \frac{\partial \psi_+}{\partial t} \\ \frac{\partial \psi_-}{\partial t} \end{bmatrix} = \frac{1}{2m} \begin{bmatrix} \frac{\partial^2 \psi_+}{\partial r^2} \\ \frac{\partial^2 \psi_-}{\partial r^2} \end{bmatrix} - \begin{bmatrix} H_z & H_x - iH_y \\ H_x - iH_y & H_z \end{bmatrix} \begin{bmatrix} \psi_+ \\ \psi_- \end{bmatrix}$$

Find a solution corresponding to the right-hand equal to zero.

15. Let e be a random variable taking on values in $(-\infty, \infty)$. Let $\psi(e)$ be the probability density of e. Let x be a random variable taking on values in the set $\{-2, -1, 0, 1, 2\}$. Suppose that the probability that $x = -2$ is $p_{-2}, x = -1$ is $p_{-1}, x = 0$ is p_0 and $x = 1$ is p_1 and $x = 2$ is p_2. Added to this *aprori* information is the statement $x + e = b$. Find the probability distribution for x conditioned on this additional information. In particular, explain which aspects of $\psi(\cdot)$ matter and which are irrelevant.

16. If $\rho(t, x)$ satisfies the unnormalized conditional density equation

$$\frac{\partial \rho}{\partial t} = \frac{1}{2} \frac{\partial^2 \rho}{\partial x^2} - \frac{1}{2} x^2 \rho + \frac{dy}{dt} x \rho$$

What equation does $\psi(x)\rho(t, x) = \eta(t, x)$ satisfy? Explain the relevance for the propagation of the conditional density for

$$dx = f(x)dt + d\omega$$

with observation

$$dy = xdt + d\nu$$

17. Consider the pair of equations

$$\begin{aligned} dz &= -2z dN \quad ; \quad z(0) \in \{1, -1\} \\ dx &= -(2+z)x dt + dw \end{aligned}$$

and the observation

$$dy = x dt + d\nu$$

where N is a Poisson counter of rate λ and ν and w are standard Wiener processes. If λ is very small, then the coefficient in the equation for x is unlikely to change very frequently. Thus it is likely to be either

$$dx = -3x dt + dw$$

or

$$dx = -1x dt + dw$$

for long periods of time. If we knew that z never switched between these values and just took on one or the other of them, then we would use two Kalman-Bucy filters to decide which it is. What would be the equations of these filters?

18. Consider the stochastic process y generated by

$$\begin{aligned} dx &= f(x)dt + g(x)dw \\ dy &= h(x)dt + d\nu \end{aligned}$$

Suppose that an inspired engineer has found that the equation

$$dz = a(z)dt + b(z)dy$$

is such that

$$m(t) = \eta(z(t))$$

is a good estimate of $h(x)$. In fact, the engineer claims that it is the least squares estimate of x in the sense that for any other past measurable function of y, say n,

$$\mathcal{E}(h(x(t)) - n(t))^2 \geq \mathcal{E}(h(x(t)) - m(t))^2$$

Assuming the claim is correct, show that

$$\mathcal{E}(h(x(t)) - m(t)) \cdot r(t) = 0$$

if $r(t)$ satisfies any equation of the form

$$dr = \gamma(r)dt + \delta(r)dy$$

19. Verify that there exists constants a, b, c and f such that the fundamental solution for

$$\frac{\partial \rho(t,x)}{\partial t} = \frac{1}{2}\left(\frac{\partial^2}{\partial x^2} - x^2\right)\rho(t,x)$$

is given by

$$\rho(t,x) = \frac{a}{\sqrt{\sinh\, bt}}\, e^{\frac{-cx^2}{\tanh ft}}$$

20. If $y\colon [a,b] \to \mathbb{R}$ is given them in some sense

$$F(\omega) = \int_a^b e^{-i\omega t} y(t) dt$$

is an "approximate Fourier Transform" of y. If, on the other hand, we wish to construct a periodic extension of y we can evaluate

$$b_n = \int_a^b y(t) \cos\Big(\frac{d\pi t}{b-a}\Big) dt$$

$$c_n = \int_a^b y(t) \sin\Big(\frac{2\pi t}{b-a}\Big) dt$$

If y is a stochastic process generated by

$$dx = Axdt + bd\omega \quad ; \quad y = cx$$

then we may wish to define the short term spectral content by

$$F(\omega) = \int_{-\infty}^t e^{-t+\sigma} e^{i\omega\pi} d\sigma$$

6.11 Notes and References

Our basic approach follows the ideas described in

1 W.M. Wonham, "Some Applications of Stochastic Differential Equations to Optimal Nonlinear Filtering", *J. SIAM, Series A, Control*, Vol. 2, No. 3, (1965).

This path leading to the association of a Lie algebra with an estimation problem is described in

2 R.W. Brockett and J.M. C. Clark, Geometry of the Conditional Density Equations, *Analysis and Optimization of Stochastic Systems* (O.L.R. Jacobs et.al. Eds.), Academic Press, 1980.

3 R.W. Brockett, "Remarks on Finite Dimensional Nonlinear Estimation", *Astérisque* (75–76 Analyse des Systems), Société Mathématique de France, 1979.

The identification of an explicit class of solvable nonlinear filters is due to Benes

4 V.E. Benes, "Exact Finite-Dimensional Filters for Certain Diffusions with Nonlinear Drift", *Stochastics*, Vol. 5, (1981) pp. 65–92.

The smoothing solution was given by Mayne in

5 D. Q. Mayne, "A Solution of the Smoothing Problem for Linear Dynamic Systems", *Automatica*, Vol. 4, (1966), pp. 73–92.

Section 6

There is an alternative approach to estimation theory which is based on the idea that an estimate can be optimal in the least squares sense only if the difference between it and the true signal, i.e., the estimation error is orthogonal to any signal which can be generated from the observable quantities. This point of view is explored in

6 M. H. A. Davis, *Linear Estimation and Control*, Chapman-Hall, London, 1977.

Chapter 7

Stochastic Control

7.1 Stochastic Control with Perfect Observations

Suppose that x evolves according to an equation of the form

$$dx = f(x)dt + b(x)udt + g(x)dw$$

with u being a control and suppose we wish to minimize a loss function of the form

$$\eta = \mathcal{E}\int_{t_0}^{t_1} L(x(t), u(t))dt + \mathcal{E}\phi(x(t_1))$$

If we observe the state x we can attempt to reduce the expected value of a loss function by suitable selection of u as a function of x and t. Because u enters the the Fokker-Planck equation as in

$$\frac{\partial \rho(t,x)}{\partial t} = L\rho(t,x) + \sum \frac{\partial b_i(x)u(t,x)\rho(t,x)}{\partial x_i}$$

with L being the Fokker Planck operator with $u = 0$, and because we can express η as

$$\eta = \int_X \int_{t_0}^{t_1} L(x,t)\rho(t,x)dtdx + \int_X \phi(x(t_1))\rho(t_1,x)dx$$

the optimal control problem can be stated as a problem about controlling a deterministic partial differential equation with no reference to the sample path equation. In general, problems of this type are not easy to solve and so other approaches are adopted.

Consider the scalar equation with control,

$$dx = -xdt + udt + dw$$

Suppose that our goal is to minimize the steady state value of

$$\eta = \mathcal{E}x^2 + u^2$$

by means of a choice of feedback control law $u(\cdot)$. If there is a steady state density $\rho(x)$ then it satisfies

$$0 = -\frac{\partial}{\partial x}(-x + u(x))\rho(x) + \frac{1}{2}\frac{\partial^2 \rho(x)}{\partial x^2}$$

Thus in steady state

$$\frac{\partial \rho(x)}{\partial x} = 2(-x + u(x))\rho(x)$$

and we wish to minimize

$$\eta = \int_{-\infty}^{\infty} (x^2 + u^2)\rho(x)dx$$

If we introduce explicitly the constraint that the integral of ρ is one, we can treat this as an optimization problem of the optimal control type, with x being the independent variable. Introduce the hamiltonian h

$$h = 2p_1(-x + u)\rho(x) + p_2(x^2 + u^2)\rho(x) + p_3\rho(x)$$

and let $p_3 = 1$. From the maximum principle formalism we see that

$$\frac{dp_1}{dx} = -2(-x + u)p_1 + x^2 + u^2 + p_3$$

The optimal value of u minimizes $2p_1u + u^2$. If u has no constraints then we have $u = -p_1$ and

$$\frac{dp_1}{dx} = 2p_1 + 2p_1^2 + x^2 + p_1^2 + p_3$$

$$\frac{dp_3}{dx} = 2$$

One sees without too much effort that $u(x) = -(\sqrt{2} - 1)x$ satisfies this system.

Example 1: Consider the control problem

$$dx = -xdt + udt + dw \ \ ; \ \ x(0) = 0$$

and the performance measure

$$\eta = \mathcal{E} \int_0^T x^2 + u^2 dt$$

Suppose that $u = k(t, x)$. In this case the Fokker-Planck equation for the density $\rho(t, x)$ can be expressed in terms of k and takes the form

$$\frac{\partial \rho}{\partial t} = (\frac{\partial}{\partial x}(x - k(t, x)) + \frac{1}{2}\frac{\partial^2}{\partial x^2})\rho$$

If we express the performance measure in terms of ρ we get

$$\eta = \int_0^T \int_{-\infty}^{\infty} \rho(t, x)(x^2 + k^2(t, x))dxdt$$

We see that the optimal k satisfies As we will see in Section 7.2, this integral is minimized by letting $\hat{k}$ satisfy the equation

$$\dot{\hat{k}} = -2\hat{k} - \hat{k}^2 + 1 \ \ ; \ \ \hat{k}(T) = 0$$

and letting $k(t, x) = -\hat{k}(t)x$.

Example 2: The situation with respect to steady-state control is somewhat simpler. Consider the equation

$$dx = -f(x)dt + b(x)udt + g(x)dw \ \ ; \ \ x(0) = 0$$

and the performance measure

$$\eta = \lim_{t\to\infty} \mathcal{E}\phi(x(t), u(t))$$

Suppose that $u = k(t, x)$ so that the the Fokker-Planck equation for the density $\rho(t, x)$ takes the form

$$\frac{\partial \rho}{\partial t} = (\frac{\partial}{\partial x}(f(x) - b(x)k(t, x)) + \frac{1}{2}\frac{\partial^2}{\partial x^2})g^2(x)\rho$$

If we express the performance measure in terms of ρ we get

$$\eta = \int_{-\infty}^{\infty} \rho(t, x)(x^2 + k^2(t, x))dx$$

The minimization of this integral can be approached using the calculus of variations. Setting the time derivative of ρ equal to zero gives

$$(\frac{\partial}{\partial x}(f(x) - b(x)k(t, x)) + \frac{1}{2}\frac{\partial^2}{\partial x^2})g^2(x)\rho = 0$$

and t can be eliminated from the problem.

Example 3: Consider the control problem

$$dx = -xdt + udt + dw \ ; \ x(0) = 0$$

and the performance measure

$$\eta = \lim_{t\to\infty} \mathcal{E}x^2 + u^2$$

Suppose that $u = k(x)$ is only permitted to take on three values, a, b, c. The Fokker-Planck equation for the density $\rho(t, x)$ can be expressed in terms of k and takes the form

$$\frac{\partial \rho}{\partial t} = (\frac{\partial}{\partial x}(x - k(t, x)) + \frac{1}{2}\frac{\partial^2}{\partial x^2})\rho$$

If we express the performance measure in terms of ρ we get

$$\eta = \int_{-\infty}^{\infty} \rho(t, x)(x^2 + k^2(x))dx$$

If we introduce the parameters a, b, c we can parametrize the solution test and optimize. Symmetry suggests that the choice $a = a, b = 0, c = -a$ deserves special consideration. The critical question remains, however, as to how many times u should switch between the three possible values. The overall structure of the problem can be investigated using the maximum principle of Pontryagin, using x as the independent variable.

7.2 Stochastic Control with Noisy Observations

By far the most tractable class of stochastic control problems involve linear systems, Wiener processes and the minimization of the expectation of a quadratic form. Consider

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) + G(t)\dot{w}(t)$$

where $\dot{w}$ is m-dimensional, unity variance, white noise and where $x(t) \in \mathbb{R}^n$. We assume that we are given

$$\mathcal{E}x(t_0) = \bar{x}(t_0)$$
$$\mathcal{E}(x(t_0) - \bar{x}(t_0))(x(t_0) - \bar{x}(t_0))^T = \Sigma_{ee}(t_0)$$

and a performance measure of the form

$$\eta = \mathcal{E}\int_{t_0}^{t_1} x^T(t)L(t)x(t) + u^T(t)u(t)dt + \mathcal{E}x^T(t_1)Qx(t_1)$$

Our goal is to find a control law to minimize η. Making full use of the structure given we can find the optimal linear feedback control law for this system as follows.

Let $\Pi(t, Q, t)$ be the solution of the Riccati equation

$$\dot{K}(t) = -A^T(t)K(t) - K(t)A(t) + K(t)B(t)B^T(t)K(t) - L(t) \quad ; \quad K(t_1) = Q$$

If G were zero and if we knew x exactly, then the optimal control law would be $u = -B^T Kx$. In the present setting we can express η, through a completion of the square argument. The first step is to observe that

$$0 = \mathcal{E}x^T(t)K(t)x(t) - \mathcal{E}x^T(0)K(0)x(0) - \mathcal{E}\int_0^t \frac{d}{d\sigma}x^T(\sigma)K(\sigma)x(\sigma)d\sigma$$

Subtracting this from the right-hand side of the equation

$$\eta = \mathcal{E}\int_{t_0}^{t_1} x^T(t)L(t)x(t) + u^T(t)u(t)dt + \mathcal{E}x^T(t_1)Qx(t_1)$$

allows us to complete the square to get

$$\begin{aligned}\eta &= \mathcal{E}\int_{t_0}^{t_1} \|u(t) + B^T(t)\Pi(t, Q, t_1)x(t)\|^2 + tr(Q\Sigma(t_1)) \\ &\quad - \int_{t_0}^{t_1} tr[\dot{\Pi}(t, Q, t_1)\Sigma(t)] + tr[\Pi(t, Q, t_1)\dot{\Sigma}(t)]dt \\ &\quad + \int_{t_0}^{t_1} tr\Pi(t, Q, t_1)G(t)G^T(t)dt\end{aligned}$$

Integrating $\mathrm{tr}(\dot{\Pi}\Sigma + \Pi\dot{\Sigma})$ gives $\Pi\Sigma$. So we have

$$\begin{aligned}\eta = \mathcal{E}\int_{t_0}^{t_1} \|u(t) + B^T(t)\Pi(t, Q, t_0)x(t)\|^2 dt + tr\Pi(t_0, Q, t_1)\Sigma(t_0) \\ + \int_{t_0}^{t} tr\Pi(t, Q, t_1)G(t)G^T(t)dt\end{aligned}$$

Since all quantities are of fixed value except the first integral, we see that the optimal control law is the same as in the deterministic case

$$u(t) = -B^T(t)\Pi(t, Q, t_1)x(t)$$

and the minimum return is

$$\eta = tr[\Pi(t_0, Q, t_1)\Sigma(t_0)] + \int_{t_0}^{t_1} tr[\Pi(t, Q, t_1)G(t)G^T(t)]dt$$

Now consider a more difficult problem where the state is not measured exactly. We have the formula above, which is valid for all linear control laws applied to

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) + G(t)\dot{w}(t)$$

If we cannot observe x itself but only $cx+\dot{v}$, we write $x=\hat{x}+e$ where $\hat{x}$ is the minimum variance unbiased estimate. Then

$$\begin{aligned}\eta &= \mathcal{E}\int_{t_0}^{t_1}\|u(t)+B^T(t)\Pi(t,Q,t_1)\hat{x}(t)\|^2dt\\ &+\mathcal{E}\int_{t_0}^{t_1}u^T(t)B^T(t)\Pi(t,Q,t_1)[x(t)-\hat{x}(t)]dt\\ &+\mathcal{E}\int_{t_0}^{t_1}[\hat{x}(t)-x(t)]^T\Pi(t,Q,t_1)B(t)B^T(t)\Pi(t,Q,t_1)[\hat{x}(t)-x(t)]dt\\ &+\int_{t_0}^{t_1}tr[\Pi(t,Q,t_1)G(t)G^T(t)]dt+tr[\Pi(t_0,Q,t_1)\Sigma(t_0)]\end{aligned}$$

The control choice $u=F\hat{x}$ gives

$$\begin{aligned}\eta &= \mathcal{E}\int_{t_0}^{t_1}\|u(t)+B^T(t)\Pi(t,Q,t_1)x(t)\|^2dt\\ &+\mathcal{E}\int_{t_0}^{t_1}x^T(t)F^T(t)B^T(t)\Pi(t,Q,t_1)[x(t)-\hat{x}(t)]dt\\ &+\int_{t_0}^{t_1}tr[\Pi(t,Q,t_1)B(t)B^T(t)\Pi(t,Q,t_1)\Sigma(t)]dt\\ &+\int_{t_0}^{t_1}tr[\Pi(t,Q,t_1)G(t)G^T(t)]dt\end{aligned}$$

But on using the fact that $\mathcal{E}[x(t)-\hat{x}(t)]\hat{x}^T(t)=0$ we can drop out the second integral. Thus the best choice of u is

$$u(t)=-B^T(t)\Pi(t,Q,t_1)\hat{x}(t)$$

The figure below shows the implementation of the separation principle for $\eta=\mathcal{E}\int_{t_0}^{t_1}x^TLx+u^Tudt+x^T(t_1)Qx(t_1)$. The defining equations are

$$\begin{aligned}\dot{\Sigma}&=A\Sigma+\Sigma A^T-\Sigma C^TC\Sigma+GG^T \quad ; \quad \Sigma(t_0)=\Sigma_{00}\\ \dot{K}&=-A^TK-KA+KBB^TK-L \quad ; \quad K(t_1)=Q\end{aligned}$$

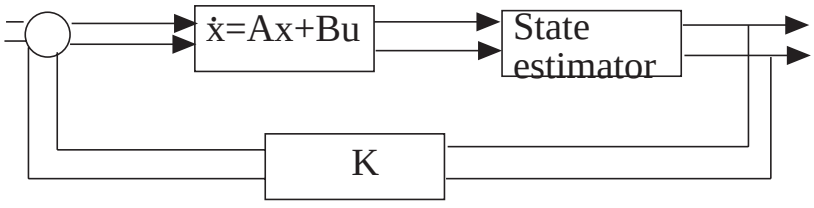

Figure 7.1: The implementation of the separation principle.

Example: Consider the scalar system

$$dx=(-x+u)dt+dw \quad ; \quad dy=xdt+dv$$

with $x(0)$ being distributed according to a Gaussian law with mean $\bar{x}$ and variance $\sigma(0) = \sigma_0$. Suppose we wish to minimize

$$\eta = \mathcal{E} \int_0^{t_1} (3x)^2(\tau) + \mu^2(\tau) d\tau + \mathcal{E} x^2(t_1)$$

According to the separation theorem we need to compute both the optimal filter to recover the best estimate of x and the optimal control assuming we know x. The latter involves solving

$$\dot{k} = 2k + k^2 - 9 \quad ; \quad k(1) = 1$$

We denote the solution of this equation by π.

We also need to solve for the optimal estimate of $x_0 - x_f$, where $x_f = \int_0^t e^{-t+\tau} u(\tau) d\tau$. If we denote this by $\hat{x}$ then for

$$\dot{\sigma} = -2\sigma + 1 - \sigma^2 \quad ; \quad \sigma(0) = \sigma_0$$

we would have

$$d\hat{x} = -2\hat{x} + \sigma(t)(dy - \hat{x}dt)$$

if u were 0.

Putting this together we have

$$dx = -xdt - \pi(t)\hat{x}(t)dt + dw$$

$$d\hat{x} = (-2 - \sigma(t))\hat{x}dt + \sigma(t)dy$$

$$u(t, x) = -\pi(t)\hat{x}$$

$$\dot{\pi}(t) = 2\pi(t) + \pi^2(t) - 9 \quad ; \; \pi(1) = 1$$

$$\dot{\sigma} = -2\sigma + 1 - \sigma^2 \quad \sigma(0) = \sigma_0$$

It may be enlightening to see this as a system

$$d \left| \begin{array}{c} x \\ \hat{x} \end{array} \right| = \begin{bmatrix} -1 & -\pi \\ \sigma c & -2 - \sigma \end{bmatrix} \begin{bmatrix} x \\ \hat{x} \end{bmatrix} dt + \begin{bmatrix} dw \\ k(t)dv \end{bmatrix}$$

$$\dot{\pi} = 2\pi + \pi^2 - 9 \quad \pi(1) = 1 \qquad \text{Final cond.}$$

$$\dot{\sigma} = -2\sigma + 1 - \sigma^2 \quad ; \quad \sigma(0) = \sigma_0 \qquad \text{Initial Cond.}$$

If we work with $\hat{x}$ and $e = x - \hat{x}$ then

$$d \begin{bmatrix} e \\ \hat{x} \end{bmatrix} = \begin{bmatrix} 2 - k & 0 \\ k & 2 - k \end{bmatrix} \begin{bmatrix} e \\ \hat{x} \end{bmatrix} dt + \begin{bmatrix} dw \\ kdv \end{bmatrix}$$

7.3 Stochastic Control with No Observations

If we are given a system of the form

$$dx = f(x,u)dt + \Sigma g_i(x,u)d\omega_i$$

with no observations of x available, what choice of u, possibly dependent on t, will minimize a performance measure of the expected value form? That is to say, how can we choose u so as to minimize

$$\mathcal{E}\phi(x(t)) = \int \rho(x,t)\phi(x)dx$$

Of course it may turn out that the best choice for u is some constant value. For example, if we have

$$dx = -xdt + udt + d\omega$$

Then

$$\frac{d}{dt}\mathcal{E}x = -\mathcal{E}x + u$$

and

$$\frac{d}{dt}\mathcal{E}(x - \mathcal{E}x)^2 = -2\mathcal{E}(x - \mathcal{E}x)^2$$

In this case u has no effect on the variance of x. Unless one knows the sign of $\mathcal{E}x(0)$ it is impossible to choose u so as to reduce $\mathcal{E}x(t)$.

This situation is not universal. There are classes of models for which open loop control produces interesting effects. One such class is modeled on the study of the thermodynamics of heat engines. If we combine the equation of a Nyquist-Johnson resistor with that of a linear capacitor having the capacitance c, the result is an Itô equation describing the voltage across the capacitance,

$$dcv = -gvdt + \sqrt{2gT}\, dw$$

The steady state value of $\mathcal{E}(v^2)$ is easily seen to be

$$\mathcal{E}v^2 = T/c$$

and so, in steady state, the expected value of the energy stored in the capacitor, $\mathcal{E}cv^2/2$, is just $T/2$. Notice that it does not depend on the values of g and c. This is, a very special form of the equipartition of energy theorem discussed in chapter 5.

We now investigate a stochastic control analog of the problem of extracting mechanical work from a heat bath. We do so by setting up a thermodynamic cycle based on a variable capacitance interacting with resistors at different temperatures.

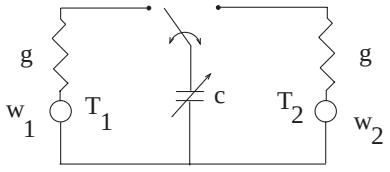

Figure 7.2: Extraction of work from an electrical circuit with noisy resistors and a variable capicator.

We are interested in analyzing the possibility of extracting energy from the system using control laws which depend on average values only and not properties of sample paths. One way

to get mechanical energy (i.e., work) out of such a system is by changing the capacitance when a charge is present. Because the energy stored in a capacitor is $cv^2/2 = q^2/2c$ where v is the voltage on the capacitor and q is the charge, we see that when we change the capacitance we change the energy stored in the capacitor. The nature of the change depends on the electrical connection in effect while the change is being made. We have the alternative expressions

$$\dot{E} = \frac{d}{dt}\frac{cv^2}{2} = \frac{d}{dt}\frac{q^2}{2c}$$

Thus we see that the effect is very different in the case of changing c at constant charge verses changing c at constant voltage. The force required to change the value of the capacitance, is expressible as the negative of the derivative of the potential energy, ($f = -\partial V/\partial x$)

$$f = -\frac{d}{dx}\frac{q^2}{2c} = \frac{c_x}{c^2}\frac{q^2}{2} \;\; ; \;\; c_x = \frac{dc}{dx}$$

The work done on the capacitor when it changes from an initial value of c_i to a final value of c_f at constant charge can be computed by integrating fdx but it more direct to observe that it is just the difference in the potential energy at the end points

$$W = \frac{q^2}{2c_f} - \frac{q^2}{2c_i}$$

From the stochastic equation for the circuit we get a variance equation

$$\dot{\sigma} = -2(\dot{c}/c)\sigma - 2\frac{g}{c}\sigma + 2g_iT_i/c^2 \;\; ; \;\; \sigma = \mathcal{E}v^2$$

with T_i being the temperature of the resistor connected to the capacitor. This can also be written as

$$\frac{d}{dt}c\sigma = -\frac{\dot{c}}{c}c\sigma - 2\frac{g}{c}c\sigma + 2g_iT_i/c$$

This equation clearly expresses the flow of energy. Along a path on which $\mathcal{E}v^2c$ is constant, the integral of the left-hand side is zero, the integral of the first term on the right is the work while the last two terms describe the flow of the heat supplied to the resistor. In view of this, we may say that the expected value of the work done when a capacitor changes slowly at an equilibrium condition associated with a resistor at temperature T is

$$\mathcal{E}W = \int_a^b \frac{c_x}{2c}Tdx = \frac{T}{2}\ln\frac{c(b)}{c(a)}$$

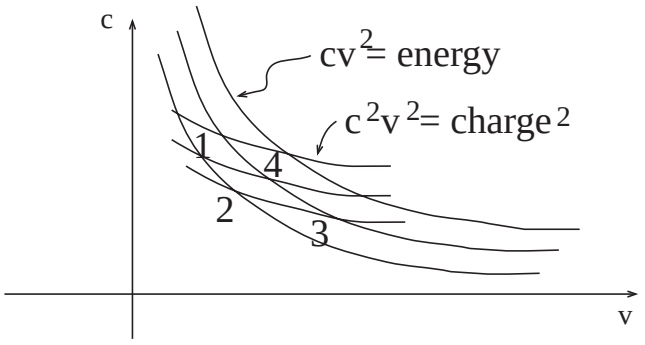

Figure 7.3: Constant charge paths and constant energy paths.

We now discuss the Carnot cycle using figure 7.3 as a guide. In passing from 1 to 2 the capacitor is connected to the resistor at temperature $T_1 = T_l$. Along this path work is done by

the capacitor in the amount

$$\mathcal{E}W_{12} = \frac{T_l}{2}\ln\frac{c_2}{c_1} \quad \text{(negative number)}$$

The next step is to follow the path from point 2 to point 3 which is a constant charge path. The capacitor is not connected to anything and work is done on the capacitor as required to change c at constant charge.

$$\mathcal{E}W_{23} = \frac{q^2}{2c_3} - \frac{q^2}{2c_2} = T_3 - T_2$$

where we have used $T = \mathcal{E}q^2/2c$. In passing from point 3 to point 4, the capacitor is attached to the resistor with temperature T_h along this path the capacitor does work in the amount

$$\mathcal{E}W_{34} = \frac{T_4}{2}\ln\frac{c_4}{c_3} \quad \text{(positive number)}$$

Finally, the capacitor is disconnected again and follows the constant charge path from point 4 to point 1. Along this path the capacitor does work on the environment in the amount

$$\mathcal{E}W_{41} = \frac{q^2}{2c_1} - \frac{q^2}{2c_4} = T_1 - T_4$$

Notice that because $T_1 = T_2$ and $T_3 = T_4$, the magnitude of the work done along the path from 2 to 3 is equal to the magnidude of the work done along the path from 4 to 1. They are of opposite sign and therefore cancel in the overall accounting of the work. To evaluate the performance of this system we examine the heat supplied at temperature $T_h = T_4 = T_3$ and compare it to the heat rejected at temperature $T_l = T_1 = T_2$. Defining the thermodynamic efficiency as $\eta = W/Q$ we see that because the magnitudes of the heat supplied along the paths 1-2 or 3-4 equals the magnitudes of the work done, we have

$$\eta = \frac{W}{Q_h} = \frac{W_{34} + W_{12}}{W_{34}} = \frac{\frac{T_h}{2}\ln\frac{c_4}{c_3} - \frac{T_l}{2}\ln\frac{c_1}{c_2}}{\frac{T_h}{2}\ln\frac{c_4}{c_3}}$$

we proceed to evaluate η. The key step in evaluating η is to establish the fact that along the closed path we have defined

$$\frac{c_1}{c_2} = \frac{c_4}{c_3}$$

To do this we combind the four equalities

$$c_1\mathcal{E}v_1^2 = c_2\mathcal{E}v_2^2 \; ; \; \mathcal{E}c_2^2v_2^2 = \mathcal{E}c_3^2v_3^2 \; ; \; \mathcal{E}c_3v_3^2 = \mathcal{E}c_4v_4^2 \; ; \; \mathcal{E}c_4^2v_4^2 = \mathcal{E}c_1^2v_1^2$$

These imply that

$$1 = \frac{c_1}{c_2}\frac{c_2}{c_3}\frac{c_3}{c_4}\frac{c_4}{c_1} = \frac{\mathcal{E}v_2^2}{\mathcal{E}v_1^2}\sqrt{\frac{\mathcal{E}v_3^2}{\mathcal{E}v_2^2}\frac{\mathcal{E}v_4^2}{\mathcal{E}v_3^2}}\sqrt{\frac{\mathcal{E}v_1^2}{\mathcal{E}v_4^2}} = \sqrt{\frac{\mathcal{E}v_2^2\mathcal{E}v_4^2}{\mathcal{E}v_1^2\mathcal{E}v_3^2}}$$

Again, using the four equations, we see that

$$1 = \sqrt{\frac{\mathcal{E}v_2^2\mathcal{E}v_4^2}{\mathcal{E}v_1^2\mathcal{E}v_3^2}} = \frac{\sqrt{c_2c_4}}{\sqrt{c_1c_3}}$$

which establishes the desired fact. Now, returning to the efficiency, we see that

$$\eta = \frac{\frac{T_h}{2}\ln\frac{c_4}{c_3} - \frac{T_l}{2}\ln\frac{c_1}{c_2}}{\frac{T_h}{2}\ln\frac{c_4}{c_3}} = \frac{T_h - T_l}{T_h}$$

According to the Kelvin-Planck statement of the second law, it is impossible to remove work from a single heat bath using a thermodynamic cycle. In our situation we see that we get no work out if $T_h = T_l$.

7.4 Exercises 7

1. Consider the stochastic differential equation

$$dx = -xdt + dw + udt; x(t) = 0$$

Suppose we set $u = -kx$. What value of k minimizes

$$\lim_{t\to\infty} E(x^2(t) + u^2(t))$$

2. Suppose that x satisfies the equation

$$dx = -xdt + dw + udt$$

and suppose that one observes

$$dy = xdt + d\nu$$

with w and ν being standard, independent, Weiner processes. If $x(0)$, the initial data for x, is distributed according to the probability density

$$\rho(x(0)) = \frac{1}{\sqrt{2\pi}}\, e^{-x^2(0)/2}$$

evaluate

$$\lim_{t\to\infty} Ex^2(t)$$

for $udt = kdy$. (Your answer will, of course, depend on k.)

3. Consider the second order equation

$$\begin{bmatrix} dx_1 \\ dx_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} dt + \begin{bmatrix} 0 \\ dw \end{bmatrix} + \begin{bmatrix} 0 \\ u \end{bmatrix} dt$$

Find $u(t) = a(t)x_1 + b(t)x_2$ such that

$$\eta = \mathcal{E} \int_0^1 u^2(t) + x_1^2(t)dt$$

is as small as possible.

4. Let x be an n-vector and let w be an n-dimensional Wiener process. Consider the Itô equations

$$\begin{array}{lll} dx = dw & ; & x(0) = 0 \\ dQ = xx^T\, dt & ; & Q(0) = 0 \end{array}$$

Compute $\mathcal{E}\, Q(t)$. Consider the deterministic control problem

$$\begin{array}{lll} \dot{x} = u & ; & x(0) = 0 \\ \dot{Q} = xx^T & ; & Q(0) = 0 \end{array}$$

Suppose that one wants to find the optimal control for achieving $Q(T) = M$, $x(T) = 0$ while minimizing

$$\eta = \int_0^T u^T u\, dt$$

Of course the optimal value of η depends on M and so we write $\eta(M)$. Show that $\eta(T)$ depends on M only through the eigenvalues of M. How does η depend on T? Write the Hamilton-Jacobi equation for this problem.

5. Write the Fokker-Planck equation for the stochastic differential equations appearing in the previous problem.

6. Consider the solution ρ of the Fokker Planck equation for $x(0) = 0$, $Q(0) = 0$. If $t > 0$, do you expect that $\rho(t, x, Q)$ will be positive for all x and $Q = Q^T$? Will it be positive for all x, Q such that $Q = Q^T > 0$?

7. Find the best open-loop u and v for the problem

$$dx = -xdt + udt + dw_1$$

$$dy = -ydt + vdt + dw_2$$

$$dz = (xv - yu)zdt + dw_3$$

with

$$\eta = \mathcal{E}(x^2 + y^2 + |z| + u^2 + v^2)$$

Index