A New Approach to Linear Filtering and Prediction Problems

R. E. KALMAN

Abstract

The classical filtering and prediction problem is re-examined using the Bode-Shannon representation of random processes and the state transition method of analysis of dynamic systems. New results are:

(1) The formulation and methods of solution of the problem apply without modification to stationary and nonstationary statistics and to growing-memory and infinite-memory filters.

(2) A nonlinear difference (or differential) equation is derived for the covariance matrix of the optimal estimation error. From the solution of this equation the co-efficients of the difference (or differential) equation of the optimal linear filter are ob-tained without further calculations.

(3) The filtering problem is shown to be the dual of the noise-free regulator problem.

The new method developed here is applied to two well-known problems, confirming and extending earlier results.

The discussion is largely self-contained and proceeds from first principles; basic concepts of the theory of random processes are reviewed in the Appendix.

1 Introduction

An important class of theoretical and practical problems in communication and control is of a statistical nature. Such problems are: (i) Prediction of random signals; (ii) separation of random signals from random noise; (iii) detection of signals of known form (pulses, sinusoids) in the presence of random noise.

In his pioneering work, Wiener [1] showed that problems (i) and (ii) lead to the so-called Wiener-Hopf integral equation; he also gave a method (spectral factorization) for the solution of this integral equation in the practically important special case of stationary statistics and rational spectra.

Many extensions and generalizations followed Wieners basic work. Zadeh and Ragazzini solved the finite-memory case [2]. Concurrently and independently of Bode and Shannon [3], they also gave a simplified method [2] of solution. Booton discussed the nonstationary Wiener-Hopf

equation [4]. These results are now in standard texts [5-6]. A somewhat different approach along these main lines has been given recently by Darlington [7]. For extensions to sampled signals, see, e.g., Franklin [8], Lees [9]. Another approach based on the eigenfunctions of the Wiener-Hopf equation (which applies also to nonstationary problems whereas the preceding methods in general dont), has been pioneered by Davis [10] and applied by many others, e.g., Shinbrot [11], Blum [12], Pugachev [13], Solodovnikov [14].

In all these works, the objective is to obtain the specification of a linear dynamic system (Wiener filter) which accomplishes the prediction, separation, or detection of a random signal.

Present methods for solving the Wiener problem are subject to a number of limitations which seriously curtail their practical usefulness:

(1) The optimal filter is specified by its impulse response. It is not a simple task to synthesize the filter from such data.

(2) Numerical determination of the optimal impulse response is often quite involved and poorly suited to machine computation. The situation gets rapidly worse with increasing complexity of the problem.

(3) Important generalizations (e.g., growing-memory filters, nonstationary prediction) require new derivations, frequently of considerable difficulty to the nonspecialist.

(4) The mathematics of the derivations are not transparent. Fundamental assumptions and their consequences tend to be obscured.

This paper introduces a new look at this whole assemblage of problems, sidestepping the difficulties just mentioned. The following are the highlights of the paper:

(5) *Optimal Estimates and Orthogonal Projections.* The Wiener problem is approached from the point of view of conditional distributions and expectations. In this way, basic facts of the Wiener theory are quickly obtained; the scope of the results and the fundamental assumptions appear clearly. It is seen that all statistical calculations and results are based on first and second order averages; no other statistical data are needed. Thus difficulty (4) is eliminated. This method is well known in probability theory (see pp. 75–78 and 148–155 of Doob [15] and pp. 455–464 of Loève [16]) but has not yet been used extensively in engineering.

(6) *Models for Random Processes.* Following, in particular, Bode and Shannon [3], arbitrary random signals are represented (up to second order average statistical properties) as the output of a linear dynamic system excited by independent or uncorrelated random signals (white noise). This is a standard trick in the engineering applications of the Wiener theory [2–7]. The approach taken here differs from the conventional one only in the way in which linear dynamic systems are described. We shall emphasize the concepts of state and state transition; in other words, linear systems will be specified by systems of first-order difference (or differential) equations. This point of view is natural and also necessary in order to take advantage of the simplifications mentioned under (5).

(7) *Solution of the Wiener Problem.* With the state-transition method, a single derivation covers a large variety of problems: growing and infinite memory filters, stationary and nonstationary statistics, etc.; difficulty (3) disappears. Having guessed the state of the estimation (i.e., filtering or prediction) problem correctly, one is led to a nonlinear difference (or differential) equation for the covariance matrix of the optimal estimation error. This is vaguely analogous to the Wiener-Hopf equation. Solution of the equation for the covariance matrix starts at the time t_0 when the first observation is taken; at each later time t the solution of the equation represents the covariance of the optimal prediction error given observations in the interval (t_0, t). From the covariance matrix at time t we obtain at once, without further calculations, the coefficients (in general, time-varying) characterizing the optimal linear filter.

(8) *The Dual Problem.* The new formulation of the Wiener problem brings it into contact with the growing new theory of control systems based on the state point of view [17–24]. It turns out, *surprisingly*, that the Wiener problem is the dual of the noise-free optimal regulator problem, which has been solved previously by the author, using the state-transition method to great advantage [18, 23, 24]. The mathematical background of the two problems is identical–this has been suspected all along, but until now the analogies have never been made explicit.

(9) *Applications.* The power of the new method is most apparent in theoretical investigations and in numerical answers to complex practical problems. In the latter case, it is best to resort to machine computation. Examples of this type will be discussed later. To provide some feel for applications, two standard examples from nonstationary prediction are included; in these cases the solution of the nonlinear difference equation mentioned under (7) above can be obtained even in closed form.

For easy reference, the main results are displayed in the form of theorems. Only Theorems 3 and 4 are original. The next section and the Appendix serve mainly to review well-known material in a form suitable for the present purposes.

2 Notation Coventions

Throughout the paper, we shall deal mainly with *discrete* (or *sampled*) dynamic systems; in other words, signals will be observed at equally spaced points in time (*sampling instants*). By suitable choice of the time scale, the constant intervals between successive sampling instants (*sampling periods*) may be chosen as unity. Thus variables referring to time, such as t, t_0, τ, T will always be integers. The restriction to discrete dynamic systems is not at all essential (at least from the engineering point of view); by using the discreteness, however, we can keep the mathematics rigorous and yet elementary. Vectors will be denoted by small bold-face letters: $\mathbf{a}$, $\mathbf{b}$, ..., $\mathbf{u}$, $\mathbf{x}$, $\mathbf{y}$, ... A vector or more precisely an n-vector is a set of

n numbers $x_1, \ldots x_n$; the x_i are the *co-ordinates* or components of the vector $\mathbf{x}$.

Matrices will be denoted by capital bold-face letters: $\mathbf{A}$, $\mathbf{B}$, $\mathbf{Q}$, $\mathbf{\Phi}$, $\mathbf{\Psi}$, ...; they are $m \times n$ arrays of elements a_{ij}, b_{ij}, q_{ij}, ... The *transpose* (interchanging rows and columns) of a matrix will be denoted by the prime. In manipulating formulas, it will be convenient to regard a vector as a matrix with a single column.

Using the conventional definition of matrix multiplication, we write the *scalar product* of two n-vector $\mathbf{x}$, $\mathbf{y}$ as

$$\mathbf{x}'\mathbf{y} = \sum_{i=1}^{n} x_i y_i = \mathbf{y}'\mathbf{x}$$

The scalar product is clearly a scalar, i.e., not a vector, quantity. Similarly, the quadratic form associated with the $n \times n$ matrix $\mathbf{Q}$ is,

$$\mathbf{x}'\mathbf{Q}\mathbf{x} = \sum_{i,j=1}^{n} x_i q_{ij} x_j$$

We define the expression $\mathbf{x}\mathbf{y}'$ where $\mathbf{x}'$ is an m-vector and $\mathbf{y}$ is an n-vector to be the $m \times n$ matrix with elements $x_i y_j$.

We write $E(\mathbf{x}) = E\mathbf{x}$ for the expected value of the random vector $\mathbf{x}$ (see Appendix). It is usually convenient to omit the brackets after E. This does not result in confusion in simple cases since constants and the operator E commute. Thus $E\mathbf{x}\mathbf{y}'$ = matrix with elements $E(x_i y_j)$; $E\mathbf{x}E\mathbf{y}'$ = matrix with elements $E(x_i)E(y_j)$.

For ease of reference, a list of the principal symbols used is given below.

3 Optimal Estimates

To have a concrete description or the type of problems to be studied, consider the following situation. We are given signal $x_1(t)$ and noise $x_2(t)$. Only the sum $y(t) = x_1(t) + x_2(t)$ can be observed. Suppose we have observed and know exactly the values of $y(t0)$, ..., $y(t)$. What can we infer from this knowledge in regard to the (unobservable) value of the signal at $t = t_1$, where t_1 may be less than, equal to, or greater than t? If $t_1 < t$, this is a *data-smoothing (interpolation)* problem. If $t_1 = t$, this is called *filtering*. If $t_1 > t$, we have a *prediction* problem. Since our treatment will be general enough to include these and similar problems, we shall use hereafter the collective term *estimalion*.

As was pointed out by Wiener [1], the natural setting of the estimation problem belongs to the realm of probability theory and statistics. Thus signal, noise, and their sum will be random variables, and consequently they may be regarded as random processes. From the probabilistic description of the random processes we can determine the probability with which a particular sample of the signal and noise will occur. For any given set of measured values $\eta(t_0)$, ..., $\eta(t)$ of the random variable $y(t)$ one can then also determine, in principle, the probability of simultaneous

Optimal Estimates

t	time in general, present time.
t_0	time at which observations start.
$x_1(t), x_2(t)$	basic random variables.
$y(t)$	observed random variable.
$x^*(t_1\|t)$	optimal estimate of $x_1(t_1)$ given $y(t_0), \ldots, y(t)$
L	loss function (non random function of its argument).
ϵ	estimation error (random variable).

Orthogonal Projections

$\mathcal{Y}(t)$	linear manifold generated by the random variables $y(t_0), \ldots, y(t)$.
$\bar{x}(t_1\|t)$	orthogonal projection of $x(t_1)$ on $\mathcal{Y}(t)$.
$\tilde{x}(t_1\|t)$	component of $x(t_1)$ orthogonal to $\mathcal{Y}(t)$.

Models for Random Processes

$\mathbf{\Phi}(t+1;t)$	transition matrix
$\mathbf{Q}(t)$	covariance of random excitation

Solution of the Wiener Problem

$\mathbf{x}(t)$	basic random variable.
$\mathbf{y}(t)$	observed random variable.
$\mathcal{Y}(t)$	linear manifold generated by $\mathbf{y}(t_0), \ldots, \mathbf{y}(t)$.
$\mathcal{Z}(t)$	linear manifold generated by $\tilde{\mathbf{y}}(t\|t-1)$.
$\mathbf{x}^*(t_1\|t)$	optimal estimate of $\mathbf{x}(t_1)$ given $\mathcal{Y}(t)$.
$\tilde{\mathbf{x}}(t_1\|t)$	error in optimal estimate of $\mathbf{x}(t_1)$ given $\mathcal{Y}(t)$.

occurrence of various values $\xi_1(t)$ of the random variable $x_1(t_1)$. This is the conditional probability distribution function

$$Pr[x_1(t_1) \leq \xi_1 | y(t_0) = \eta(t_0), \ldots, y(t) = \eta(t)] = F(\xi_1) \tag{1}$$

Evidently, $F(\xi_1)$ represents all the information which the measurement of the random variables $y(t_0)$, ..., $y(t)$ has conveyed about the random variable $x_1(t_1)$. Any statistical estimate of the random variable $x_1(t_1)$ will be some function of this distribution and therefore a (nonrandom) function of the random variables $y(t_0)$, ..., $y(t)$. This statistical estimate is denoted by $X_1(t_1|t)$, or by just $X_1(t_1)$ or X_1 when the set of observed random variables or the time at which the estimate is required are clear from context.

Suppose now that X1 is given as a fixed function of the random variables $y(t_0)$, ..., $y(t)$. Then X_1 is itself a random variable and its actual value is known whenever the actual values of $y(t_0)$, ..., $y(t)$ are known. In general, the actual value of $X_1(t_1)$ will be different from the (unknown) actual value of $x_1(t_1)$. To arrive at a rational way of determining X_1, it is natural to assign a penalty or loss for incorrect estimates. Clearly, the loss should be a (i) positive, (ii) nondecreasing function of the estimation error $\epsilon = x_1(t_1) - X_1(t_1)$. Thus we define a *loss function* by

$$L(0) = 0$$

$$L(\epsilon_2) \leq L(\epsilon_1) \leq 0 \text{ when } \epsilon_2 \leq \epsilon_1 \leq 0 \tag{2}$$

$$L(\epsilon) = L(-\epsilon)$$

Some common examples of loss functions are: $L(\epsilon) = a\epsilon^2$, $a\epsilon^4$, $a|\epsilon|$, $a[1 - \exp(-\epsilon^2)]$, etc., where a is a positive constant.

One (but by no means the only) natural way of choosing the random variable X_1 is to require that this choice should minimize the average loss or risk

$$E\{L[x_1(t_1) - X_1(t_1)]\} = E[E\{L[x(t_1) - X_1(t_1)]|y(t_0), \ldots, y(t)\}] \quad (3)$$

Since the first expectation on the right-hand side of (3) does not depend on the choice of X_1 but only on $y(t_0), \ldots, y(t)$, it is clear that minimizing (3) is equivalent to minimizing

$$E\{L[x_1(t_1) - X_1(t_1)]|y(t_0), \ldots, y(t)\} \quad (4)$$

Under just slight additional assumptions, optimal estimates can be characterized in a simple way.

Theorem 1. *Assume that L is of type (2) and that the conditional distribution function $F(\xi)$ defined by (1) is:*

A symmetric about the mean $\bar{\xi}$:

$$F(\xi - \bar{\xi}) = 1 - F(\bar{\xi} - xi)$$

B convex for $\xi \leq \bar{\xi}$:

$$F(\lambda\xi_1 + (1-\lambda)\xi_2) \leq \lambda F(\xi_1) + (1-\lambda)F(\xi_2)$$
$$\text{for all } \xi_1, \xi_2 \leq \bar{\xi} \text{ and } 0 \leq \lambda \leq 1$$

for all ξ_1, $\xi_2 \leq \xi$ and $0 \leq \lambda \leq 1$

Then the random variable $x_1^*(t_1|t)$ which minimizes the average loss (3) is the conditional expectation

$$x_1^*(t_1|t) = E[x_1(t_1)|y(t_0), \ldots, y(t)] \quad (5)$$

Proof: As pointed out recently by Sherman [25], this theorem follows immediately from a well-known lemma in probability theory.

Corollary. *If the random processes $\{x_1(t)\}$, $\{x_2(t)\}$, and $\{y(t)\}$ are gaussian, Theorem 1 holds.*

Proof: By Theorem 5, (A) (see Appendix), conditional distributions on a gaussian random process are gaussian. Hence the requirements of Theorem 1 are always satisfied.

In the control system literature, this theorem appears sometimes in a form which is more restrictive in one way and more general in another way:

Theorem 1-A. *If $L(\epsilon) = \epsilon^2$, then Theorem 1 is true without assumptions (A) and (B).*

Proof: Expand the conditional expectation (4):

$$E[x_1^2(t_1)|y(t_0),\ldots,y(t)] - 2X_1(t_1)E[x_1(t_1)|y(t_0),\ldots,y(t)] + X_1^2(t_1)$$

and differentiate with respect to $X_1(t_1)$. This is not a completely rigorous argument; for a simple rigorous proof see Doob [15], pp. 77-78.

Remarks.(a) As far as the author is aware, it is not known what is the most general class of random processes $x_1(t)$, $x_2(t)$ for which the conditional distribution function satisfies the requirements of Theorem 1.

(b) Aside from the note of Sherman, Theorem 1 apparently has never been stated explicitly in the control systems literature. In fact, one finds many statements to the effect that loss functions of the general type (2) cannot be conveniently handled mathematically.

(c) In the sequel, we shall be dealing mainly with vectorvalued random variables. In that case, the estimation problem is stated as: Given a vector-valued random process $\mathbf{x}(t)$ and observed random variables $\mathbf{y}(t_0)$, $\ldots$, $\mathbf{y}(t)$, where $\mathbf{y}(t) = \mathbf{M}\mathbf{x}(t)$ ($\mathbf{M}$ being a singular matrix; in other words, not all co-ordinates of $\mathbf{x}(t)$ can be observed), find an estimate $\mathbf{X}(t_1)$ which minimizes the expected loss $E[L(\|x(t_1) - X(t_1)\|)]$, $\|\ \|$ being the norm of a vector. Theorem 1 remains true in the vector case also, provided we require that the conditional distribution function of the n coordinates of the vector $\mathbf{x}(t_1)$,

$$Pr[x_1(t_1) \leq \xi_1, \ldots, x_n(t_1) \leq \xi_n | \mathbf{y}(t_0), \ldots, \mathbf{y}(t)] = F(\xi_1, \ldots, \xi_n)$$

be symmetric with respect to the n variables $\xi_1 - \bar{x}i_1$, $\ldots$, $\xi_n - \bar{\xi}_n$ and convex in the region where all of these variables are negative.

4 Orthogonal Projections

5 Models for Random Processes

6 Solution of the Wiener problem

7 The Dual Problem

8 Applications

9 Conclusions