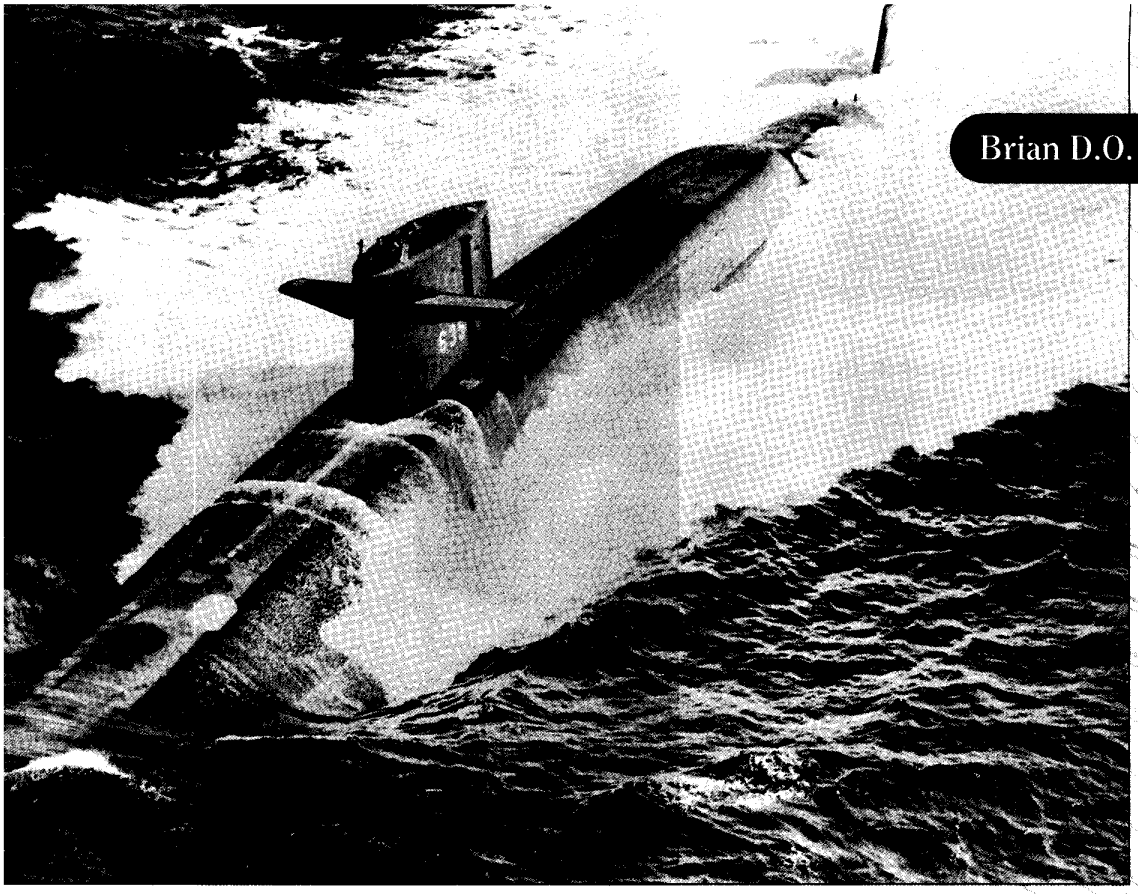

Brian D.O. Anderson

From Wiener to Hidden Markov Models

This article has several aims. First, we seek to trace out some common properties of three types of filters, all obtained from considering various stochastic models. These are Wiener filters, Kalman filters and Hidden Markov Model (HMM) filters. Our thesis is that by forgetting the detailed mathematics (which differs greatly between the filters), unifying features stand out. This is especially so in respect of the forgetting of old data, the forgetting of initial conditions, and the securing of protection from round-off error effects overpowering the calculations.

Second, we aim to clearly differentiate the concept of fixed-lag smoothing from filtering, and expose the comparative advantages and disadvantages. Once again, there are common properties which allow a unified viewpoint. We focus especially on characterizations of a maximally effective smoothing lag, and identification of the SNR circumstances under which smoothing is especially beneficial.

Anderson (brian.anderson@anu.edu.au) is with the Department of Systems Engineering, Research School of Information Sciences and Engineering, Australian National University, Canberra, Australia. *The author wishes to acknowledge the funding of the activities of the Cooperative Research Centre for Robust and Adaptive Systems by the Australian Commonwealth Government under the Cooperative Research Centres Program, the US Army Research Office, Far East, Tokyo, and the Office of Naval Research, Washington. This work was originally presented at WAC'98 (World Automated Congress), Anchorage, Alaska, May 1998.*

 0272-1708/99/$10.00©1999IEEE

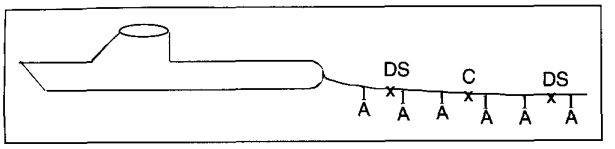

Fig. 1. Submarine with towed array.

Third, we identify several open problems.

In accordance with these aims, there are very few equations in the paper. We also motivate the paper by recording a successful real-world example in which the most accurate form of model is a nonlinear partial differential equation. Such an example emphasizes the great practical utility of filtering and smoothing.

The third section recalls the Wiener and Kalman filters, and the section following that explains smoothing in relation to filtering, especially fixed-lag smoothing. Hidden Markov models are then covered, and we finish with some concluding remarks.

A Filtering Problem

To motivate the article, we shall begin by describing a significant applications problem in the filtering area.

Shape Estimation of a Towed Array

Fig. 1 is a diagrammatic representation of a submarine trailing a towed array. A towed array is a cable on which are located a large number of acoustic sensors (labeled with the letter A), and the purpose of the acoustic sensors is to listen for other vessels. Use of the towed array rather than location of acoustic sensors on board the submarine allows use of an effectively much bigger antenna, and lessens difficulties associated with self-noise generated by the submarine. Satisfactory use of the collection of acoustic sensor signals, however, requires knowledge of the shape of the array. The known motion of the submarine together with equations of motion of a towed cable (a nonlinear partial differential equation) allow the generation of an *estimate* of the shape of the array, but this will be deficient for at least two reasons. First, the equations of motion of the towed cable are only approximations of reality, i.e., there is modeling error, and second, there may well be currents giving rise to further forces on the array, and thus distortions of it.

For this reason, it is desirable to find techniques for improving the determination of array shape, and, to this end, one can contemplate using inclusion along the array of some depth sensors and compasses (labeled DS and C), which would provide some sort of (noisy) measurement information relevant to determining the shape of the array.

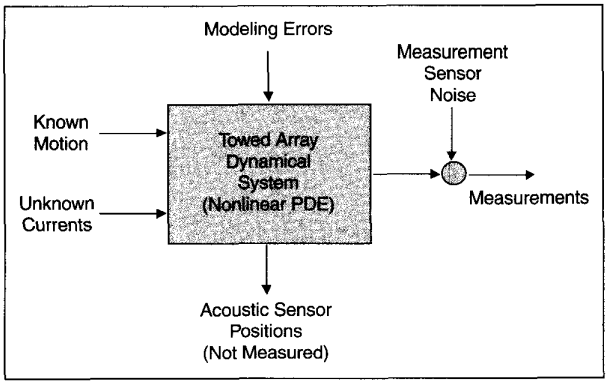

Fig. 2. Abstract representation of array shape estimation problem.

At this stage, one could represent the situation abstractly via Fig. 2. The inputs on the left include the known motion, which is the submarine motion. The rest of the diagram is self-evident.

In order to obtain the acoustic sensor positions, a *filter* is needed. The signals driving the filter are the measurements from the sensors as well as the motion of the submarine, and, of course, the equations defining the filter in some way will depend on the model of the towed array. Filtering theory in the sense of Wiener [1] and Kalman [2], [3] attempts to provide a technique for determination of a filter, and for predicting the performance of the filter (as measured, for example, by the mean square error in the estimates produced by the filter).

A description of the towed array problem can be found in [4].

Wiener and Kalman Filtering

Wiener Filtering

Wiener filtering theory [1] is probably the first attempt to provide optimal filters in situations where signals are characterised by random processes. Indeed, the preface to [1] states: "Largely because of the impetus gained during World War II, communication and control engineering have reached a very high level of development today .... The point of departure may well be the recasting and verifying of the theories of control and communication ... *on a statistical basis* (italics added)." Later in the preface, it is noted that the book was first published during the war as a classified report, and used mathematical developments which in the main were new, but had origins in works of Kolmogorov and Kosulajeff published between 1939 and 1941 on interpolation and extrapolation in random stationary sequences. The prototypical situation considered by a Wiener filter is illustrated in Fig. 3.

Fig. 3a denotes the signal model. This comprises a noise process, the input noise process, which is normally of zero mean, gaussian, and known spectrum; a known linear time-invariant stable system, whose output is the signal process; a measurement noise process, also normally zero mean, gaussian and of known spectrum; and a measurement process, which is the signal plus measurement noise. Most commonly, the input noise process and the measurement noise process are independent processes, and so the spectrum of the measurement process will be the sum of the signal spectrum and the measurement noise spectrum. Also, without any real loss of generality, the input noise process can be taken as white (if it is not white, a shaping filter driven by white noise can be regarded as generating the input process, and then the shaping filter can be combined with the linear, time-invariant, stable system between the input noise process and the signal). The measurement noise process is often taken as white. The Wiener filter, exhibited in Fig. 3b, is the device which reconstructs in real time a best estimate of the signal from the measurement process. (A definition of the best estimate is given below for an example.) The Wiener filter itself is a causal system; i.e., its output at any time t depends on inputs up to that time, and it is a stable system.

Kalman Filter

The Kalman filter [2], [3] is a variant on the Wiener filter, see Fig. 4. The variation is in several respects; a) in addition to an input noise process, there can be a known input; b) the system link-

ing the input noise to the signal is assumed to be finite dimensional; c) the system linking the input noise to the signal is not assumed to be time-invariant or stable (however, if it is unstable, it must be switched on at some finite time, rather than the infinitely remote past); d) input and measurement noise processes are almost universally assumed to be white.

Obviously in this case both the noisy measurement of the signal as well as the known signal component of the system input can be acted upon by the filter to produce an estimate. Also, the estimate is not just of the varable labeled "signal," but of the whole state of the finite dimensional linear system. Without going into detailed commentary it should be intuitively clear that with a best estimate of the state, one should be in a good position to come up with a good if not optimal estimate of the signal also.

The great advances in the Kalman filter are firstly the introduction of time-variation in the signal model and non-stationarity in the underlying random processes, and, secondly, the recognition that the whole of a state vector may be of interest as an object of estimation, rather than just the entity in Fig. 4 named "signal." The "cost" in using the Kalman filter, as opposed to the Wiener filter, may be the restriction to finite-dimensional systems. Aside from these shifts of viewpoint, one can only admire the introduction of Riccati equations as a key tool, and the sophisticated stability results (depending on time-varying generalizations of the equally sophisticated concepts of complete controllability and complete observability).

Nor was the Kalman filter in 1960 simply an academic exercise. Navigation problems, especially for space applications, were certainly problems requiring filtering with the underlying system a time-varying one [5].

By the 1960's, sampled-data systems were part of the thinking of many electrical engineers, and it was as natural to cast Kalman filtering theory in a discrete-time framework as in a continuous-time framework. The discrete-time framework allowed mental distinguishing of time-updating and measurement updating, which makes the connection with centuries old work of Gauss clearer—for an historical account, see [6].

Like the Wiener filter, the Kalman filter is supposed to apply just to linear systems. It is, however, natural to try to push it further. Most nonlinear extensions of the Kalman filter at best are based on an approximation involving some form of linearization. An example of a nonlinear extension of a Kalman filter is a receiver of FM radio signals constructed using a phased-locked loop [7]. When the receiver is operating above the threshold signal to noise ratio, a linearized description of its operation is valid, and the threshold signal to noise ratio effectively defines the boundary of the region in which the linearization assumption is valid.

General theories of nonlinear filtering have not been regarded as particularly successful, and most theory tries to build on use of one or several simultaneous Kalman filters, together with linearization assumptions.

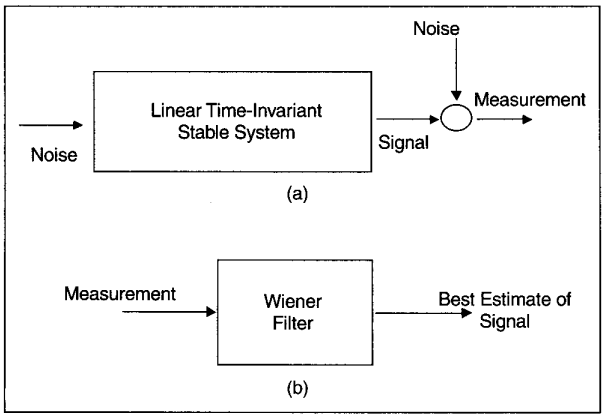

Fig. 3. (a) Signal model, (b) Basic set-up for Wiener filtering.

Common Aspects of the Wiener and Kalman Filters

There is clearly some overlap between the Wiener filter and the Kalman filter. Consider for example the simple model of Fig. 5a. This qualifies as a signal model amenable to treatment by the Wiener filtering and Kalman filtering theory. While Kalman filtering theory does not guarantee that the Kalman filter will be a time-invariant linear system, there are circumstances in which time-invariance will be obtained (linearity is always obtained). The example fits one of those circumstances. Fig. 5b depicts the optimal filter (optimal according to both Wiener and Kalman filtering theories), with optimality in the sense that the mean square error $E[s(t) - \hat{s}(t)]^2$ is minimized for the particular choice of causal filter linking $z(\cdots)$ to $\hat{s}(\cdots)$. The minimum value of this error in this instance is $\sqrt{(a^2 + q)} - a$.

Key Properties of the Kalman Filter

For the purposes of this article, we want to record several key properties of the Kalman filter. These properties are not universally obtained; however, they are obtained in most circumstances, and strict theorems are available defining circumstances under which they are obtained; see, for example, [2], [3], [8].

In relation to the performance of the filter,

(a) Old measurements are forgotten exponentially fast; put another way, the best estimate of the state at the particular time t depends in an exponentially decaying fashion on prior measurements.

(b) Initial state information of the Kalman filter is forgotten exponentially fast. What does this mean? At some time, the Kalman filter has to be turned on. It is a linear finite dimensional system, and it has to be given an initial state. The best initial state is probably the one equal to the unknown state of the system be-

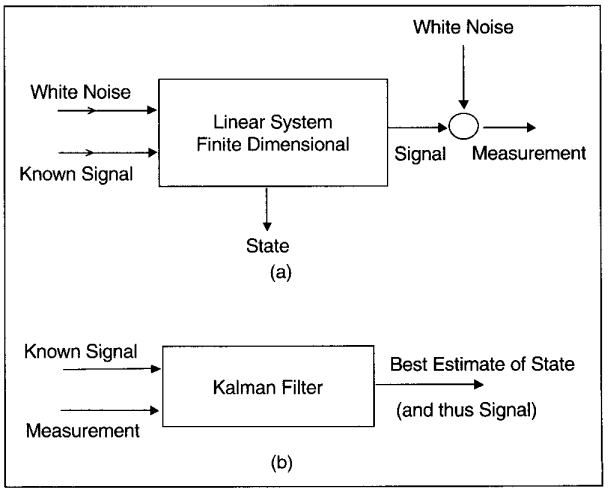

Fig. 4. (a) Signal model, (b) Basic set-up for Kalman filtering.

ing filtered. However, one must reckon with the possibility that an inappropriate initial state for the Kalman filter is selected. The point of the remark is that any damage caused by inappropriate selection is forgotten exponentially fast.

(c) Round-off and similar errors can only accumulate to a limited extent. Suppose the Kalman filter is implemented on a computer. Then at virtually every step in the calculations, one must assume that some little error is introduced. If the calculations run on in time, one should have an a priori concern as to whether these errors accumulate in an ultimately damaging way. The point of the remark is that normally they do not.

Actually, as most readers will recognize, these three key properties are closely related.

Filtering and Smoothing as Alternatives

The purpose of this section is to distinguish, in a technical way, two ways of processing measurements. Any processing of the measurements is often described by the generic term "filtering." However, one can particularize the meaning of the word filtering, to distinguish it from a related concept called "smoothing." This is the point of this section. Also, we will record some distinctions in the properties of filters and smoothers.

The Difference Between Filtering and Smoothing

Consider Fig. 6. This figure depicts one entry of the unknown true state of the system on which the filtering is being performed; it depicts the measurements taken at the output of that system; and it depicts one entry of the filtered estimate of the state, obtained at the output of the Kalman filter. The notation $\hat{x}(t|t)$ is used to denote a filtered estimate. The first occurrence of t signifies the fact that we are estimating $x(t)$. The second occurrence of t signifies that we are achieving this estimate using measurements occurring up to a time t.

It is intuitively obvious that measurements received after time t must contain some sort of information about $x(t)$. If those estimates are in some way usable, we ought to be able to obtain an improved estimate of $x(t)$, i.e., one with lesser mean square error. Fig. 7 illustrates the distinction between filtering and smoothing, where in smoothing, we are using measurements not just up to time t but up to some later time T in order to produce our estimate of $x(t)$. The new estimate is termed $\hat{x}(t|T)$.

Leaving aside for the moment the question of how exactly such an estimate might be constructed, it is important to realize that there is one key disadvantage of working with a smoothed estimate, namely, the estimate is not available in real time, but only with some delay. For a control application, this may be a fatal disadvantage. However, if one is analyzing what happened in an experiment subsequent to that experiment there may be no disadvantage at all.

For the sake of completeness, we should mention also the concept of prediction: using measurements $z(\cdot)$ up till time t, one seeks to estimate not $x(t)$ but $x(t+\delta)$ for some $\delta > 0$. (The quantity δ may be fixed and t a running variable.) Such an estimation may be relevant in, for example, a rendezvous problem, with a moving target with which a rendezvous is sought at a future time. If one can do filtering, it is generally very easy to do prediction, and we will devote almost no attention to it.

Types of Smoothing

There are in fact several different types of smoothing which need to be distinguished. These are termed fixed interval smoothing, fixed point smoothing, and fixed-lag smoothing.

In fixed interval smoothing, t is variable and T is fixed. This corresponds to the situation where one collects some experimental data, and then derives estimates of the state subsequent to the data collection. The data field is fixed. In fixed point smoothing, t is fixed and T is increasing. This means one is after the best possible estimate of the state at a particular point in time, and one uses just as much experimental data as one can collect in order to produce this estimate. Finally, in fixed-lag smoothing, t is variable and T is set to equal $t + \Delta$, with Δ a fixed quantity, termed the lag. Thus fixed-lag smoothing is like filtering with delay. Fig. 8 illustrates how various measurements give rise to a fixed-lag estimate at different time instants.

Of the three types of smoothing, that which is most like filtering is clearly fixed-lag smoothing. Fixed-lag smoothing is treated in the Wiener filtering context in [1] and the Kalman filter context in [8]. Fig. 9 depicts traces of the state of a discrete-time system, a filtered estimate of that state and a fixed-lag estimate of

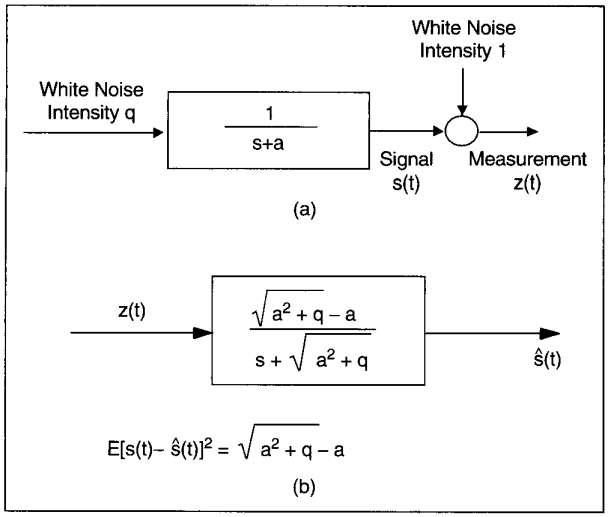

Fig. 5. (a) Simple signal model, (b) Associated Wiener/Kalman filter.

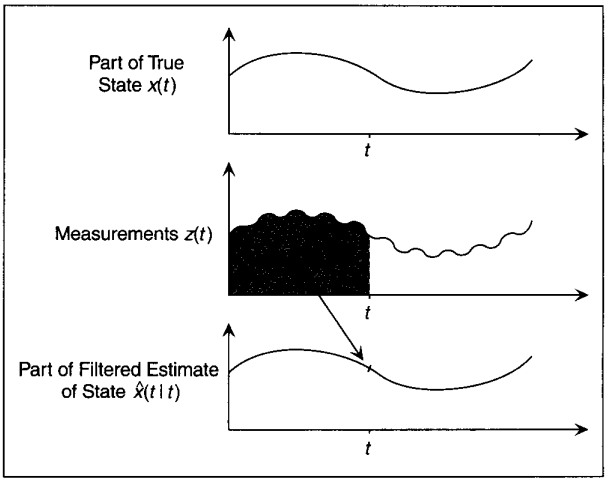

Fig. 6. Representation of filtered estimate dependence on measurements.

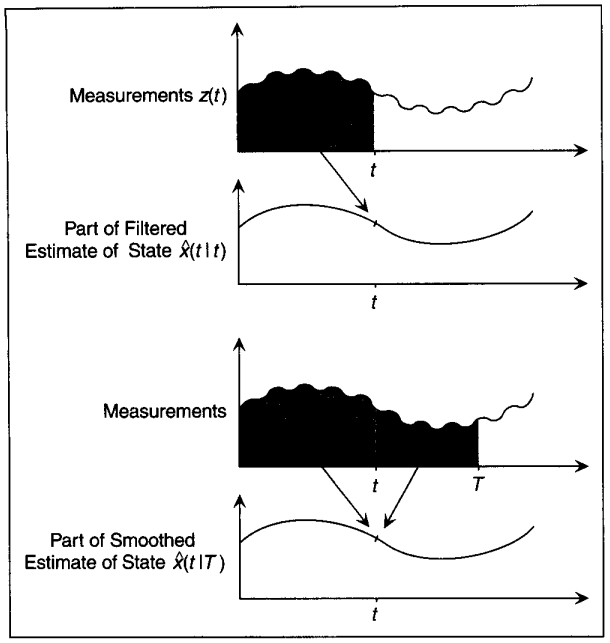

Fig. 7. Contrast of smoothed and filtered estimate dependence on measurements.

that state, together with the error performance of the filter and the smoother. An inspection by eye suggests the greater accuracy of the fixed-lag estimate, and this is confirmed by a calculation of the error variance.

In relation to the towed array problem, it is evident that in filtering, measurement information up to a time t would allow estimates of the acoustic sensor positions at time t and allow listening for other vessels using those sensor estimates. Smoothing would allow a better estimate of acoustic sensor positions, and allow better listening—but there would be a delay. In the demodulation of a frequency modulated signal, as commented earlier, the phase-locked loop is a form of linearized Kalman filter. It has been argued that the incorporation of a pre-emphasis/de-emphasis circuit is equivalent to introducing a fixed-lag smoother [7, section 5.3]. The delay associated with this smoother is of the order of a millisecond, and is imperceptible to the listener.

Our focus in this article will be on fixed-lag smoothing, perhaps originally studied in [1]. Textbook references dealing with all types of smoothing include [8]-[11]. Early important papers include [12]-[15] and the survey [16].

Practical Signal Processing Issues in Fixed-Lag Smoothing

To understand some of the issues arising, we can consider again the earlier example, now depicted in Fig. 10.

Fig. 11 illustrates the fixed-lag smoother (the notation should be self-explanatory). The transfer function $H(s)$ is given by

$$H(s) = \frac{-\exp(-s\Delta) + \exp(-b\Delta)}{s-b} \quad \left(b = \sqrt{a^2+q}\right). \qquad (1)$$

The impulse response associated with $H(s)$ is a finite impulse response, comprising a growing exponential which falls back to zero at time Δ (see Fig. 12).

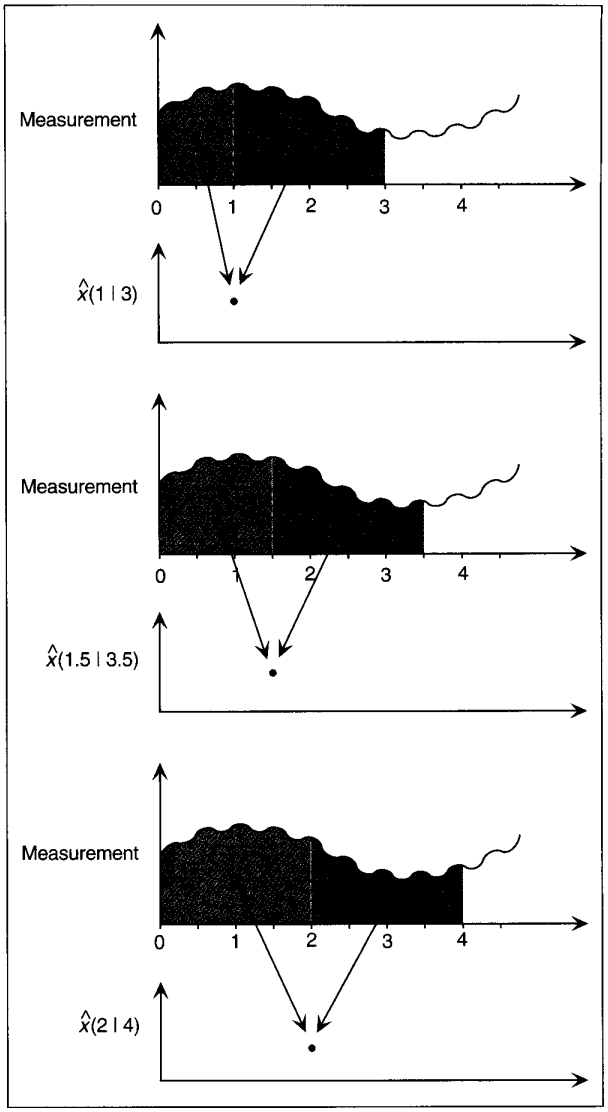

Fig. 8. Representation of generation of a fixed-lag estimate.

It turns out that the implementation of practical fixed-lag smoothers lagged the theory. The core reason was that suggested implementations often concealed an unstable pole-zero cancellation. This unstable pole-zero cancellation can be fairly clearly seen for the above very simple example. The "obvious" way which one might use to implement $H(s)$ is shown in Fig. 13 and because b is positive, there is an unstable pole-zero cancellation. This instability was discovered around 1970 [17]. Successful implementation of fixed-lag smoothers in continuous time was addressed by approximation or by clever switching arrangements incorporating the use over finite intervals of unstable filters, always with zero initial conditions at the start of each finite interval [18]-[20].

In discrete time, the difficulty is nowhere near as great, since the finite impulse response transfer function corresponding to $H(s)$ in discrete time is no longer irrational. It is given by

$$H(z) = \frac{-z^{-\Delta} + b^{\Delta}}{z-b}. \qquad (2)$$

Although there is a pole-zero cancellation appearance, that cancellation can be very easily and legitimately made in this case, and one can actually implement $H(z)$ as $H(z) = -[z^{-(\Delta-1)} + bz^{-(\Delta-2)} + \cdots + b^{\Delta-1}]$. This is evidently a finite impulse response implementation.

Of course, what we have said in relation to the particular example above generalizes to arbitary situations [8], [21], [22].

Comparative Advantages of Smoothing Over Filtering

We have already referred to the key disadvantage of using smoothing as opposed to filtering, including fixed-lag smoothing. This is the delay in obtaining an estimate. The key advantage is the greater accurary in the estimate. This naturally raises the question: "What improvement we can expect?" A subsidiary question is: "How much lag should one use in fixed-lag smoothing to capture all the significant improvement?"

These questions have been addressed in a number of papers; see, e.g., [23]-[26]. The first key conclusion is that *at high signal to noise ratios, smoothing gives greater improvement over filtering than at low signal to noise ratios.* Denote by P_S and P_F the mean square error in estimating the signal with a smoother and with a filter, respectively. Then [23] provides, for a significant family of systems, a bound for the minimum possible value of P_S / P_F in terms of the maximum signal to noise ratio. The bound is depicted in Fig. 14. At low signal to noise ratios then, it is impossible to get much improvement. Note also that the curve does not guarantee that at high signal to noise ratios, there has to be a lot of improvement. It simply indicates that there may be a lot of improvement. Nevertheless, examples supported in various references testify to the conclusion that at high signal to noise ratios, a significant improvement can be expected.

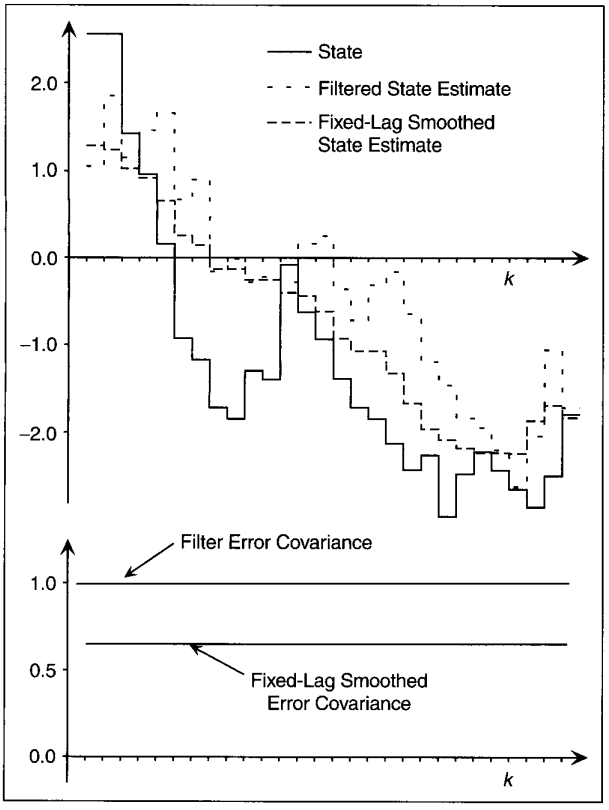

Fig. 9. Simulation data for filtering and fixed-lag smoothing comparison.

(Notice that the SNR can go to infinity either as signal power goes to infinity or noise power goes to zero. In the latter case, P_S and P_F both go to zero, and, in the limit, the issue of improvement is irrelevant. However, for high signal to noise ratios, improvement may nevertheless be very desirable; modern digital communications systems after all do seek extremely low error rates.)

A second key conclusion relates to the choice of Δ for fixed-lag smoothing. If Δ is taken to be several times the dominant time constant of the Wiener or Kalman filter, then one will obtain all the practical improvement that it is possible to obtain using fixed-lag smoothing. A typical curve illustrating the situation is shown in Fig. 15. The case of zero lag corresponds to filtering. As the lag is increased, the mean square error goes down, monotonically in fact, but the benefit from further increases in lag gradually tails off (in fact it tails off exponentially) until a lag is reached at which further increase of Δ is pointless.

Fig. 16 illustrates the situation for the towed array problem where it is assumed that there are three compasses located along the array, at distances 0.3, 0.6 and 0.85 along the normalized length of the array. The curve is drawn from [4]. The particular nature of the dynamical equation of the towed array gives rise to the somewhat strange shape, but it is clear that smoothing gives considerable improvement over filtering, and that a longer lag is better than a short lag.

For comparison, performance of a predictor is also shown. (Here, $x(t)$ must be predicted using measurements ceasing prior to time t.) The mean square error in estimation is normalized by the mean square value of the compass noise.

Some Open Issues

Above, we have indirectly raised one open issue, namely the extent to which improvement by use of a fixed-lag smoother over a filter may be guaranteed or not at higher signal to noise ratios. At the moment, there is no guarantee. It would be nice to have one.

Another question relates to two dimensional picture processing. Here, it would be desirable to have some general conclusions regarding the benefits of smoothing over filtering.

Hidden Markov Models

Wiener and Kalman filtering theories are concerned with filtering of signals and linear systems. The theory can be pushed to consider some levels of nonlinearity, typically when linearization is applicable, but in no sense do the theories provide a general theory for the filtering of nonlinear systems. There is however a theory which can capture many nonlinear filtering problems, and that is based on hidden Markov models [27], [28].

As noted in the abstract of [28], hidden Markov models were initially introduced in the late 1960's and early 1970's, i.e., between 25 and 30 years ago, and their popularity has slowly grown. To quote from the abstract of [28]: "There are two strong reasons why this has occurred. First, the models are very rich in mathematical structure and hence can form the theoretical basis for use in a wider range of applications. Second the models, when applied properly, work very well in practice with some important applications."

The reasons might just have well been advanced for the Wiener filter and Kalman filter. But for HMMs, the mathematical struc-

ture is very different, and the successful applications are also very different. The mathematical structure does not include spectral factorization or Riccati equations. But it does include the theory of positive matrices (to be distinguished from positive definite matrices), including (as work of this decade has revealed) a form of time-varying extension of Perron-Frobenius theory.

Examples of Hidden Markov Models

Before defining what a hidden Markov model is, let us give several examples. The first is a very old one—the random telegraph wave (see Fig. 17). One assumes that a signal is transmitted which takes the value zero or one. One has available noisy measurements of that signal, and from the noisy measurements, one is required to reconstruct the original signal. (The common noise model is additive white gaussian noise.) The transitions within the original signal are assumed to occur in a Poisson manner, and the number of levels in the original signal (here two) can be generalized to be finite, but not to be infinite.

For the second example, consider the problem of listening from one submarine for the engine of another submarine. One can postulate that the engine of the other submarine has a fundamental frequency lying in one of a finite set of frequency ranges, and the transition probability for movement from one range to another is known. Noisy estimates (in effect noisy measurements) are available of the particular range in which the fundamental frequency of the other submarine lies, and the problem is to properly reconstruct the activity of the other submarine's engine from the estimates.

As a third example, see Fig. 18. At any time instant, the contents of the communication channel must be one of a finite set of possibilities (the number depending on the size of the signal alphabet and the length of the finite impulse response). We can regard the contents of the communications channel as a state, and the state assumes one of a finite set of values. The received signal contains a limited measurement of this state, and the problem is to work out from the received signal exactly what is transmitted.

Sophisticated use of HMMs occurs in speech recognizers; see [28].

Formal Description of a Finite State Hidden Markov Model (Discrete Time)

In this subsection we try to capture in a more abstract framework the contents of the previous examples. There is an underlying state process $X_0, X_1, \cdots, X_k$, and X_k can assume one of the finite set of values, for convenience $1,2,\cdots,N$. The quantity $Pr[X_{k+1} = i \,|\, X_k = j] = a_{ij}$ is a transition probability, and X_k is a Markov process. We denote by A the matrix $[a_{ij}]$. There is also an output process $Y_0, Y_1, \cdots$. We shall assume that Y_k takes one of a finite set of values, for convenience labeled $1,2,\cdots,M$. (There are many examples where Y_k assumes continuous values, for example when it is equal to gaussian noise plus the state. However, for the sake of this paper we shall assume the simpler case of a finite set of values. This finite set incidentally could arise through quantization of a continuum.) The link between the state process and the output process is defined by $Pr[Y_k = m \,|\, X_k = n] = c_{mn}$, and we denote by C the matrix (c_{mn}).

Evidently, then, two matrices whose entries are all probabilities, A and C, describe the hidden Markov model process.

Hidden Markov Model Filter

A hidden Markov model filter is, in quite precise terms, a device for calculating the N-vector whose ith entry is $Pr[X_k = i \,|\, Y_0, Y_1, \cdots, Y_k]$. This means that the filter uses the measurements up to time k to provide the best possible statement concerning the knowledge of the state at time k. For the simple HMM setup that we have described, it is fairly easy to obtain filtering equations by straightforward application of the Bayes' Theorem. The update process involves two steps, incorporating a time update of the state variable with no extra measurements, and then adding in the extra measurement associated with an update. More precisely, let $\Pi_{k/k}$ = vector with ith entry $Pr[X_k = i \,|\, Y_0, \cdots, Y_k]$, and $\Pi_{k+1/k}$ = vector with ith entry $Pr[X_{k+1} = i \,|\, Y_0, \cdots, Y_k]$. Then

$$\Pi_{k+1/k} = A\Pi_{k/k} \tag{3}$$

$$\Pi_{k+1/k+1} = \frac{1}{[1\cdots1]C_{y_{k+1}}\Pi_{k+1/k}} C_{y_{k+1}}\Pi_{k+1/k} \tag{4}$$

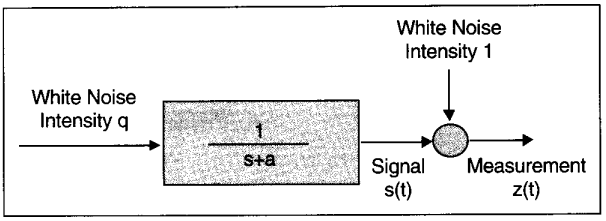

Fig. 10. Simple signal model.

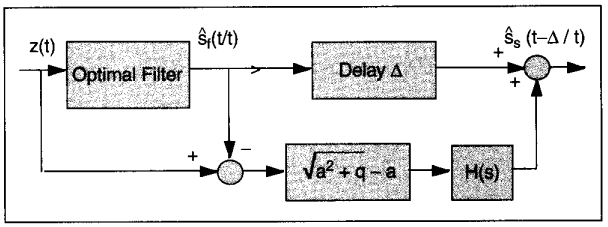

Fig. 11. Fixed-lag smoother for signal model of Fig. 10.

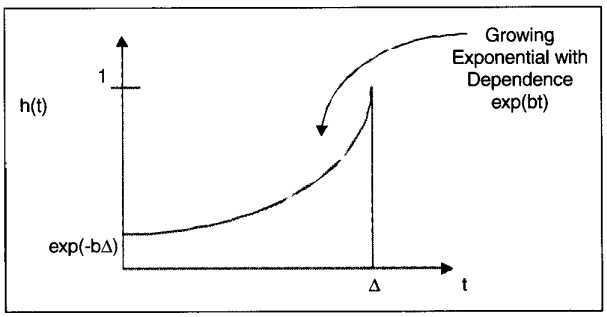

Fig. 12. Impulse response of $H(s)$ block of smoother.

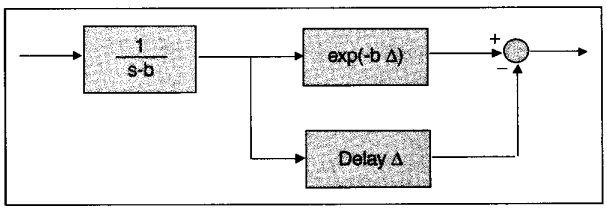

Fig. 13. Obvious, but unsatisfactory, implementation of $H(s)$ of Fig. 12.

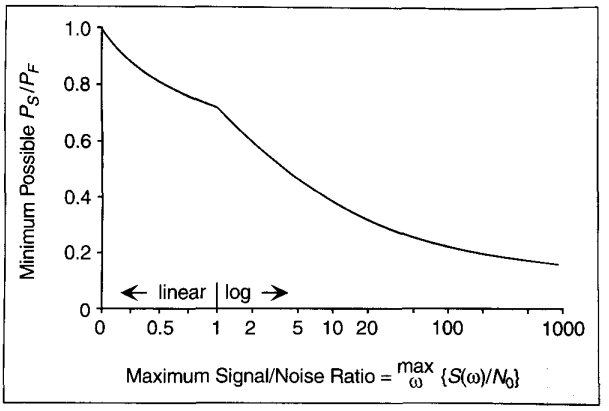

Fig. 14. Smoothing improvement against maximum signal/noise ratio.

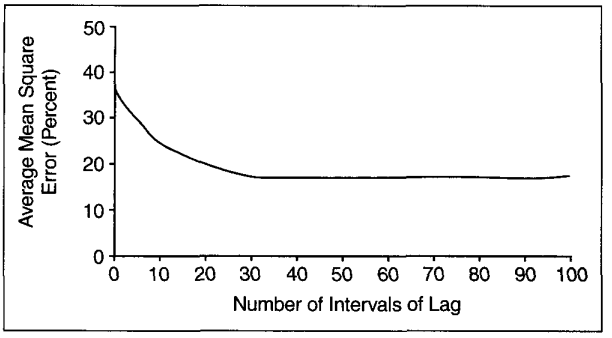

Fig. 15. Variation of smoothing performance with lag.

with $C_{Y_{k+1}} = \text{diag}(c_{p1},\cdots,c_{pN})$ when $Y_{k+1} = p$.

The Forgetting Property

At this stage one can ask similar questions to those which can be asked regarding Wiener and Kalman filters; are old measurements forgotten, is an inappropriate filter initialization forgotten, and are round-off and similar errors guaranteed not to overpower the calculation?

As for Kalman and Wiener filtering problems, the answer is, in general, yes. The qualification is one which can be expressed in technical terms [29]-[32], and in broad terms demands that filtering problems be well posed. The general conclusion is in fact that there is an exponential forgetting property just like that for Kalman and Wiener filtering.

Incidentally, obtaining these conclusions for continuous time hidden Markov models is much more technically difficult.

There is a new angle here, which does not arise in Kalman and Wiener filters, and it should be noted. What we have just said is that the calculations leading to the conditional probability associated with the filtering problem are ones in which an exponential forgetting property is found. Suppose that one focuses on the actual production of a state estimate. Thus, one can define a filtered state estimate by saying that $\hat{X}_{k/k} = J$ if J maximizes $Pr[X_k = i | Y_0,\cdots,Y_k]$ over i. Then it turns out that $\hat{X}_{k/k}$ is determined with a *finite memory*, i.e., $\hat{X}_{k/k}$ depends on $Y_k, Y_{k-1},\cdots,Y_{k-l}$ for some fixed l and all k. At this stage, theory is not available to estimate l easily [33].

A Short Digression

The exponential forgetting property can be established by an elegant extension of the Perron-Frobenius theory on the eigenvalue of matrices of nonnegative or positive entries [34]. The extension, to be found in [35], is roughly as follows. Let $\{A_1,\cdots,A_p\}$ denote a finite set of matrices with positive entries, and let $E_N = D_N D_{N-1}\cdots D_1$, where $D_i \in \{A_1,\cdots,A_p\}$. Then as $N \to \infty$, $E_N \to u_N v^T$ for variable vector u_N and fixed vector v, exponentially fast. (The Perron-Frobenius theorem deals with the case where all the D_i are the same.) There are incidentally extensions of the inhomogeneous product to cope with nonnegative matrices, and such extensions are needed for the application to hidden Markov models; see [32] and [33].

Smoothing

Just as for Wiener and Kalman filters, fixed interval, fixed point and fixed-lag smoothing can all be contemplated for hidden Markov models. In fixed-lag smoothing, one calculates as a recursion in k the vector $\Pi_{k/k+\Delta}$ whose ith entry is given by

$$Pr[X_k = i | Y_0,\cdots,Y_k,Y_{k+1},\cdots,Y_{k+\Delta}]. \tag{5}$$

The calculations required to update this quantity are not greatly different from those applying to filtering.

Some twenty years ago, [36] appeared, which contained a description of a smoother for a random telegraph wave (recall this is a particular sort of hidden Markov model). In particular, this reference investigated on a simulation basis, i.e., by numerical experiment, how the error performance depended on lag, and how it depended on signal to noise ratio. The experimental data showed results that were very similar to those obtained in Wiener and Kalman filtering theory.

In particular, as the smoothing lag increased, performance improved, but the improvement with each increase in lag got less and less, and in effect fell off exponentially. There was a maximum lag beyond which no significant improvement in error performance would be obtained. This lag in fact corresponded to several times the dominant time constant (to the extent to which that could be defined) for the associated filter. Very recent theoretical work has verified that the experimental observations of [36] were instances of a general theory, and that exponential rates could be obtained with the same exponent applying in filtering as applied in the way fixed-lag performance varied with change of lag [37]. (For filtering the exponent is that associated with exponential forgetting of old data.)

The work of [36] also, via simulation, studied the influence of signal to noise ratio on the extent to which smoothing could give benefit over filtering, and concluded that at higher signal to noise ratios, more benefit was to be expected than at low signal to noise ratios.

In more detail [36] examines the effects of change of signal to noise ratio in filtering on a random telegraph wave. The random telegraph wave moved between the two states with transition times determined by a stationary Poisson process, and the observation comprised the state process with additive white gaussian noise. Simulation results drawn directly from [36] record the improvement due to smoothing at different levels of signal to noise ratio, and are represented in Fig. 19. Although the data is limited, one can clearly see that at high signal to noise ratios, the improvement due to smoothing is the greater. (The quantity β is a measure

of the noise standard deviation, and ν is a measure of the frequency with which signal transitions occur, so that small ν corresponds to significant signal energy at DC.)

Recent theoretical work has supported these conclusions. One can construct a family of hidden Markov models indexed by a parameter ϵ. The family is defined by $A(\epsilon)$, the state probability transition matrix, and C, the matrix of transition probabilities linking state to output, with the latter matrix independent of ϵ.

The general form of the matrix $A(\epsilon)$ is

$$\begin{aligned} a_{ij} &= \epsilon\lambda_{ij} \qquad i \neq j \\ &= 1-\epsilon\lambda_{jj} \quad i = j \end{aligned} \tag{6}$$

where $\lambda_{jj} = \sum_{i \neq j} \lambda_{ij}$, $\lambda_{ij} > 0$ for all i, j, and the matrix C has entries which satisfy

$$c_{mn} > 0 \text{ for all } m, n \tag{7}$$

and no two columns of C are identical.

As indicated earlier, for $i = 1,2,\cdots$, one can calculate the quantities $Pr[X_k = i | Y_0,\cdots,Y_k]$ for $k = 1,2$, etc., in a recursive manner. These are the filtered probabilities for the unobserved state. We will adopt for the filter the definition $\hat{X}_{k/k} = J$ if J maximizes $Pr[X_k = i | Y_0,\cdots,Y_k]$ over i. We will say that an error occurs if $\hat{X}_{k/k} \neq X_k$.

It is evident that when ϵ approaches zero, state transitions occur less and less frequently. The covariance of the output process can be calculated from the quantities $A(\epsilon), C$, and, being stationary, it has associated with it a spectrum. As ϵ goes to zero, this spectrum inherits the property from the states that it is more and more concentrated at low frequencies. In a crude sense, one could say that the signal to noise ratio at zero frequency approaches infinity as ϵ approaches zero.

Be that as it may, one can argue theoretically [38] that the probability of filtering error is $O(\epsilon \log \epsilon^{-1})$ as ϵ approaches zero, while more recent arguments [39] show that the probability of smoothing error is simply $O(\epsilon)$ as ϵ aproaches zero. Thus the ratio of the probability of filtering error to the probability of smoothing error, measuring as it does the extent to which smoothing improves matters over filtering, is $O(\log \epsilon^{-1})$. Evidently, this ratio approaches infinity as ϵ goes to zero. In this sense then, the property is verified that at high signal to noise ratios, smoothing offers greater advantage over filtering.

For the particular case where the matrices $A(\epsilon)$ and C are given by

$$A = \begin{bmatrix} 1-0.2\epsilon & 0.4\epsilon \\ 0.2\epsilon & 1-0.4\epsilon \end{bmatrix} \tag{8}$$

and

$$C = \begin{bmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{bmatrix} \tag{9}$$

Fig. 20 illustrates simulation results for a filter and a fixed-lag smoother with a sizeable lag as a function of ϵ. The coincidence with the theoretical predictions is remarkable over quite a range of values of ϵ.

Maximum A Posteriori Trajectory Determination

In nonlinear filtering problems, it becomes important to distinguish between the most likely state at a particular instant of time and the most likely trajectories occurring over a period of time. Suppose $\hat{X}_{k/k}$ is the most likely state at time k, given the measurements up to time k. Consider the sequence $\hat{X}_{1/1}, \hat{X}_{2/2},\cdots,\hat{X}_{k/k}$. It may well be that this is not the most likely trajectory traced out by $X_1,\cdots,X_k$ given the measurements up to time k, while the determination of this trajectory may clearly be of interest. It may not be the most likely trajectory for two reasons. First, as far as early points on the trajectory are concerned, it is evident that their determination in some sense should involve a smoothing problem rather than a filtering problem. Second, it is easy to conceive of hidden Markov models in which certain state transitions are completely impossible. At the same time, one could well envisage a filter giving two state estimates at successive instants of time which correspond to a state pair within the original model for which the state transition would be barred. A most likely trajectory obviously could not include within it two successive state values between which no transition was ever possible.

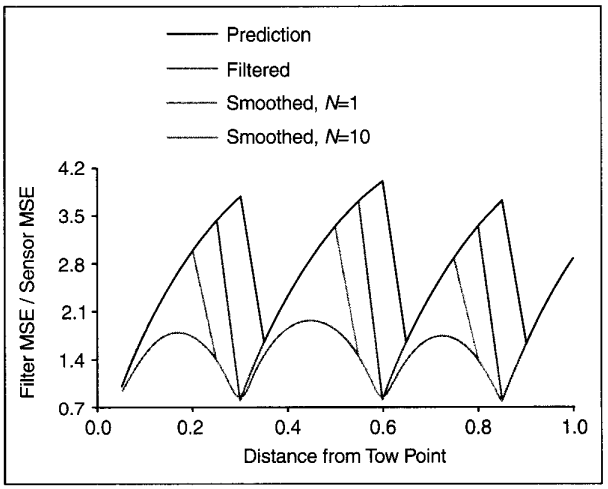

Fig. 16. Performance of towed array estimator.

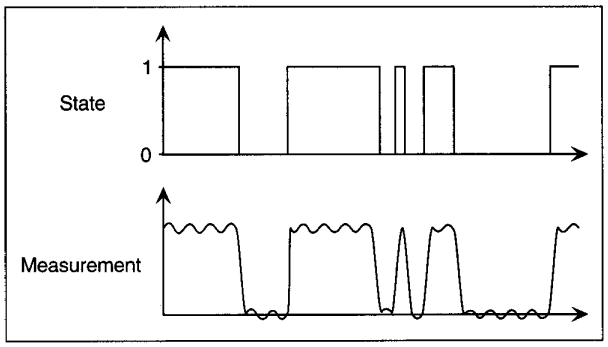

Fig. 17. Noisy measurement of a random telegraph wave.

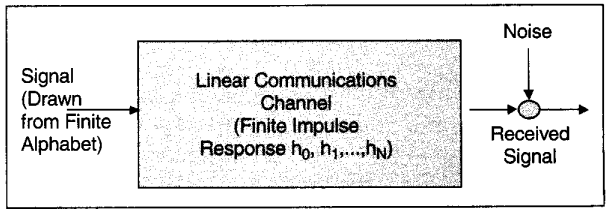

Fig. 18. Channel deconvolution problem.

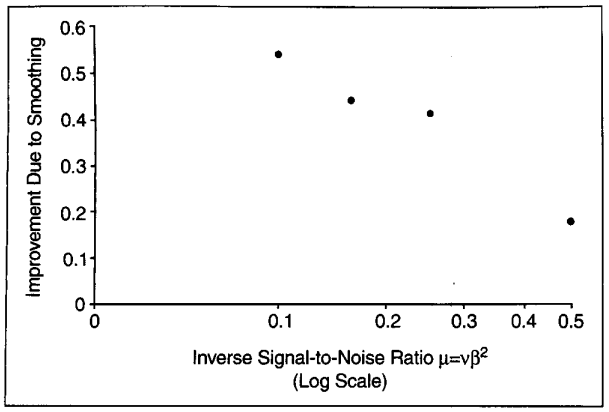

Fig. 19. Simulation data studying smoothing improvement against SNR.

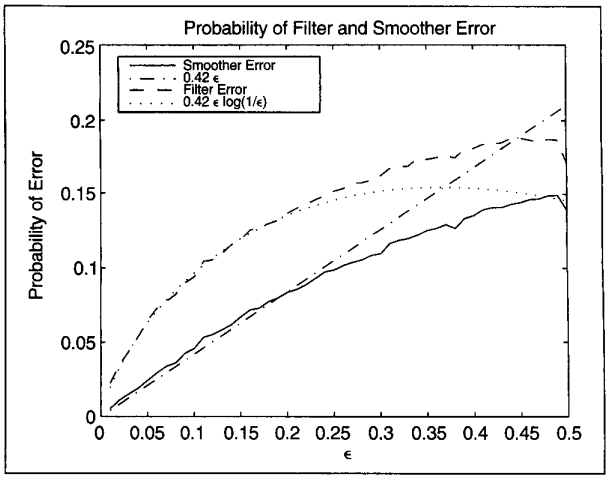

Fig. 20. Comparison of filter and smoother error performance as function of ϵ, theoretical and simulation.

As already indicated, it may well be that we want the most likely path occurring over a time interval, i.e., for some P we may want the sequence

$$X^*_{k-P+1}, X^*_{k-P+2} \cdots, X^*_k$$

which maximizes

$$Pr[X_{k-P+1}, X_{k-P+2} \cdots, X_k |\ Y_0, Y_1, \cdots, Y_k].$$

Finding the sequence is a maximum a posteriori trajectory determination problem. A Viterbi decoder finds such a trajectory, and normal implementation in some way uses dynamic programming.

Practical Viterbi decoders, because of their finite memory, assume that for some Q,

$$\arg\max Pr[X_{k-P+1}, \cdots, X_k |\ Y_0, Y_1, \cdots, Y_k]$$

$$= \arg\max Pr[X_{k-P+1} \cdots, X_k |\ Y_{k-Q} \cdots, Y_k]. \quad (10)$$

This means that they only remember a finite amount of data, and in fact it is assumed that it is only necessary to remember a finite amount of data. Rules of thumb are used to define Q, the amount of data that is remembered.

Evidently, key questions are: "Can the forgetting properties giving rise to the above equality (10) be justified, and more precisely, can rules of thumb help to establish Q?" At this time, the equality (10) cannot be completely theoretically justified. There are however partial results [40] which rely on the convergence of matrix products of the type

$$F_N = D_N \otimes D_{N-1} \otimes \cdots \otimes D_1, \quad (11)$$

where $\otimes$ denotes a max-plus algebra product [41].

Conclusions

The first major message of this paper is that there are striking parallels between Wiener, Kalman, and hidden Markov model filtering in respect of the exponential forgetting of initial conditions, old measurements, and round-off errors. The second major message is that this same exponential rate, for all three filters, governs the choice of a lag for fixed-lag smoothing, where that lag is chosen to secure all the practical advantage that it is possible to secure from fixed-lag smoothing.

The third major message is that at high signal to noise ratios, smoothing gives greater improvement over filtering than at low signal to noise ratios.

In the course of the paper we have indicated some open problems. These include extending the results to two dimensional signals, such as arise in picture processing, and in the hidden Markov model area, justifying and dimensioning finite length Viterbi decoding.

What of the future? The 50 year old words of Wiener quoted earlier remain true: "Communication and control engineering have reached a high level of development today." The level of development of 50 years ago looks puny in comparison with today's level, and the same is likely to be true in another 50 years. As control problems become higher and higher level, dealing more with hybrid systems, and systems of systems, various filtering and estimation problems are still going to remain. But the systems will be less physical in nature, and certainly often not susceptible to methods of linear/quadratic theory (perhaps after linearization). It is here that new ways of modeling, perhaps like HMMs, will become important. (In particular, HMMs seem a very relevant tool for hybrid-systems.) And the modeling schemes which appeal will be those for which there is an associated intuitive framework, possibly with rules of thumb, clarity regarding the nature and consequences of approximations, and some capability to predict performance. These are some of the properties now available for the Kalman filter (though not available when the mathematics was first done three decades ago) that, together with the application drivers, explain its ubiquity. These properties are also becoming available for the HMM. But we must expect further frameworks again, spurred by application problems so far undreamt of.

References

[1] N. Wiener, *Extrapolation, Interpolation and Smoothing of Stationary Time Teries*, MIT Press, 1949.

[2] R.E. Kalman and R.S. Bucy, "New results in linear filtering and prediction theory," *J Basic Eng. Trans. ASME*, Series D, vol. 83, 1961, pp. 95-108.

[3] R.E. Kalman, "A new approach to linear filtering and prediction problems," *J Basic Eng, Trans. ASME*, Series D, vol. 82, 1960, pp. 35-45.

[4] D.A. Gray, B.D.O. Anderson, and R.R. Bitmead, "Towed array shape estimation using Kalman filters Part I—Theoretical Model," *J. Oceanic Engineering*, vol. 18, 1993, pp. 543-556.

[5] P. Swerling, "A proposed stagewise differential correction procedure for satellite tracking and predictions," *J. Astronaut. Sci.*, vol. 6, 1959, pp. 46-59.

[6] H.W. Sorenson, "Least-squares estimation: from Gauss to Kalman," *IEEE Spectrum*, vol. 7, no. 7, 1970, pp. 66-68.

[7] H.L. Van Trees, *Detection, Estimation and Modulation Theory*, vol. 2, John Wiley, 1971.

[8] B.D.O. Anderson and J.B. Moore, *Optimal Filtering*, Prentice Hall Inc., 1979.

[9] J.S. Meditch, *Stochastic Optimal Linear Estimation and Control*, McGraw Hill, 1969.

[10] A.E. Bryson, Jr., and Y.C. Ho, *Applied Optimal Control*, Blaisdell, Mass, 1969.

[11] A.P. Sage and J.L. Melsa, *Estimation Theory with Applications to Communications and Control*, McGraw Hill, 1971.

[12] H.E. Rauch, "Solutions to the linear smoothing problem," *IEEE Trans. Auto Contr.*, vol. AC-8, October 1963, pp. 371-372.

[13] H.E. Rauch, F. Tung, and C.T. Striebel, "Maximum likelihood estimates of linear dynamic systems," *AIAA J.*, vol. 3, 1965, pp. 1445-1450.

[14] D.Q. Mayne, "A solution to the smoothing problem for linear dynamic systems," *Automatica*, 1966, pp. 73-92.

[15] D.C. Fraser and J.E. Potter, "The optimum linear smoother as a combination of two optimum linear filters," *IEEE Trans. Auto. Contr.*, vol. AC-14, 1969, pp. 387-390.

[16] J.S. Meditch, "A survey of data smoothing for linear and nonlinear dynamic systems," *Automatica*, vol. 9, 1973, pp. 151-162.

[17] C.N. Kelly and B.D.O. Anderson, "On the stability of fixed-lag smoothing algorithms," *J. Franklin Instit.*, vol. 291, 1971, pp. 271-281.

[18] S. Chirarattanon and B.D.O. Anderson, "Outline design for stable, continuous-time fixed-lag smoothers," *Electronic Letters*, vol. 8, 1972, pp. 263-264.

[19] S. Chirarattanon and B.D.O. Anderson, "Stable fixed-lag smoothing of continuous-time processes," *IEEE Trans. Inform. Theory*, vol. IT-19, 1973, pp. 25-36.

[20] P.K.S. Tam and J.B. Moore, "Stable realization of fixed-lag smoothing equations for continuous-time signals," *IEEE Trans. Auto. Control*, vol. AC-19, 1974, pp. 84-87.

[21] J.B. Moore, "Discrete-time fixed-lag smoothing algorithms," *Automatica* vol. 9, 1973, pp. 163-174.

[22] S. Chirarattanon and B.D.O. Anderson, "The fixed-lag smoother as a stable finite-dimensional linear system," *Automatica*, vol. 7, 1974, pp. 657-669.

[23] B.D.O. Anderson and S. Chirarattananon, "Smoothing as an improvement on filtering: a universal bound," *Electronics Letters*, vol. 7, 1971, p. 524.

[24] B.D.O. Anderson, "Properties of optimal linear smoothing," *IEEE Trans. Auto. Contr.*, vol. AC-14, 1969, pp. 114-115.

[25] B.D.O. Anderson and S. Chirarattananon, "New linear smoothing formulas," *IEEE Trans. Aut. Contr.*, vol. AC-17, 1972, pp. 160-161.

[26] J.B. Moore and K.L. Teo, "Smoothing as an improvement on filtering in high noise," *Systems and Control Letters*, vol. 8, 1986, pp. 51-54.

[27] R.J. Elliott, L. Aggoun, and J.B Moore, *Hidden Markov Models: Estimation and Control*, Springer Verlag, 1994.

[28] L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, 1989, pp. 257-285.

[29] F. LeGland and L. Mevel, "Geometric ergodicty in Hidden Markov Models," IRISA report, July 1996.

[30] A. Arapostathis and S.I. Marcus, "Analysis of an identification algorithm arising in adaptive estimation of Markov chains," *Math of Control, Signals and Systems*, vol. 3, 1990, pp. 1-29.

[31] L.E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state Markov chains," *Annals of Math Stats*, vol. 37, 1966, pp. 1554-1567.

[32] B.D.O. Anderson, "New developments in the theory of positive systems," in *Systems and Control in the Twenty-First Century*, C.I. Byrnes, B.N. Datta, C.F. Martin, and D.S. Gilliam, eds., Birkhauser, Boston, 1997.

[33] B.D.O. Anderson, "Forgetting properties for hidden Markov models," *Proc. AFOSR/DSTO Workshop*, 1997.

[34] R.A. Horn and C.R. Johnson, *Matrix Analysis*, Cambridge U. Press, 1985.

[35] E. Seneta, *Non-negative matrices and Markov Chains*, Springer Verlag, 1981.

[36] D. Clements and B.D.O. Anderson, " A nonlinear fixed-lag smoother for finite-state Markov processes," *IEEE Trans. Inform. Theory*, vol. IT-21, 1975, pp. 446-452.

[37] L. Shue, B.D.O. Anderson, and S. Dey, "Exponential stability of filters and smoothers for hidden Markov models," *IEEE Trans. Signal Processing*, to appear.

[38] R. Khasminski and O. Zeitouni, "Asymptotic filtering for finite state Markov chains," *Stochastic Processes and their Application*, vol. 63, 1996, pp. 1-10.

[39] L. Shue, F. De Bruyne and B.D.O. Anderson, "Asymptotic smoothing errors for hidden Markov models," submitted.

[40] L. Shue, B.D.O. Anderson, and S. Dey, "On steady-state properties of certain max-plus products," *Proc Amer. Contr. Conf.*, 1998, pp. 1909-1913.

[41] F. Baccelli, G. Cohen, G.J. Olsder, and J-P Quadrat, *Synchronization and Linearity*, John Wiley, 1992.

Brian D.O. Anderson was born in Sydney, Australia, and received his undergraduate education at the University of Sydney, and a Ph.D. degree from Stanford University. He worked in industry in Silicon Valley and served as a faculty member in electrical engineering at Stanford. He is now Professor of Systems Engineering at the Australian National University and Director of the Research School of Information Sciences and Engineering. His interests are in control and signal processing. He is a Fellow of the IEEE, the Royal Society and several similar bodies, and holds four honorary doctorates. He was President of the International Federation of Automatic Control from 1990 to 1993 and is currently President of the Australian Academy of Science. His awards include the 1997 IEEE Control Systems Award.