Norbert Wiener and the development of mathematical engineering

Thomas Kailath

1. Introduction

'There is nothing better than concrete cases for the morale of a mathematician. Some of these cases are to be found in mathematical physics and the closely related mathematical engineering...'

So wrote Norbert Wiener in 1949, in an obituary of G. H. Hardy, who reputedly would have shuddered at the thought. [Ironically, Hardy's own major field of number theory has been of great importance for many applications, for example in secure data communications.] This paper will attempt to describe how one particular concrete problem in Wiener's own work – solving the Wiener–Hopf equations encountered in astrophysics – led him, and then a vast host of followers, to chart out several new areas of investigation, and to develop a very significant body of knowledge, which can well go by the name Mathematical Engineering. In the era of the PC, the Internet and the World Wide Web, few of us can be unaware that mathematical engineering has come to play a major role in the world around us. And with this has come, as this paper will describe, an increasing recognition of the seminal role of Norbert Wiener's ideas and influence in the development of this field.

It must be said that the term 'Mathematical Engineering' does not enjoy the currency that the name 'Mathematical Physics' does. Being a younger field, its proponents still focus on more specialized descriptions such as Information Theory, Communications, Control, Signal Processing, Computational Complexity, Image Processing, etc. The names 'System Theory' or even 'Mathematical System Theory' have been advanced but are not universally accepted. However this author believes that the increasing intermingling of the fields mentioned above, with many tools and techniques being successfully applied across them, as well as the tremendous opportunities ahead of them in the Information Era, could well lead to the adoption of Wiener's suggestion. And be that as it may, Wiener's early vision and pioneering contributions will, as mentioned earlier, loom even larger with time. Already in 1962, in a special issue commemorating the 50th anniversary of the effective existence of the IEEE (Institute of Electrical and Electronics Engineers), Lotfi Zadeh, winner of the 1995 IEEE Medal of Honour wrote *'If one were asked to name a single individual who above anyone else is responsible for the conception of system theory, the answer would undoubtedly be 'Norbert Wiener', even though Wiener has not concerned himself with system theory as such, nor has he been using the term "system theory" in the sense employed in this paper'*.

There are many of Wiener's results that have come to be important in mathematical engineering. Although Wiener was apparently never quite secure about his place in the pantheon of scientific innovators, he seemed not to have such doubts about the significance of the particular concrete problem that we shall concentrate on in this paper. This is the Wiener–Hopf equation, which Wiener first encountered in 1931 in astrophysics and then a decade later in the problem of anti-aircraft fire control. I hope to indicate how Wiener's work on this equation has led to the development of a remarkably broad, and deep, range of studies.

In the next section, we shall describe Wiener's beautiful technique of spectral factorization for solving it. In § 3 and 4, we shall note that Wiener encountered the equation again in solving a problem in anti-aircraft fire control and how his 1942 report on this project introduced two fundamental ideas that radically changed the way engineers approached important classes of problems. First is that the communication of information must be formulated as a statistical problem; the second, the introduction of optimization criteria to obtain the limits of performance and replace the earlier 'trial and error' approach to design. From these two ideas has grown the huge flood of activity noted earlier. However to narrow the scope we shall return in the remainder of the paper to a specific problem studied by Wiener – filtering signals out of noisy observations. After describing his results, we shall turn to some of the mathematical developments following from it. First we shall show in

The work was supported by the Advanced Research Projects Agency of the Department of Defense and was monitored by the Air Force Office of Scientific Research under Grant No. F49620-93-1-0085.
First published in the *Proceedings of Symposia in Pure Mathematics*, 1996, 60 (tentative), published by the American Mathematical Society.
Thomas Kailath is Hitachi America Professor of Engineering at Stanford University, Stanford, CA 94305-4055, USA.

§ 6, how, while Wiener was not quite successful in extending his results to the case of multiple time series, this can be done by the introduction of state-space and Markov process descriptions. Somehow Wiener himself never really focused on the Markov property. As we shall see, the state-space description introduces the concept of recursive solution algorithms and enables straight-forward extension to nonstationary/time-variant versions of the filtering problem. In § 7, we shall examine a widely used finite-time prediction problem, which has been generalized by using the concept of displacement structure. Finally in § 8, we shall introduce the nonlinear filtering problem, which is still open, but for which a key tool is martingale theory, the prototype for which was the Wiener (Brownian motion) stochastic process.

2. The Wiener–Hopf equation

Given $\phi_{11}(\cdot)$ and $\phi_{12}(\cdot) \in \mathcal{L}_1\ (\infty, -\infty)$, $\phi_{11}(\cdot)$ even and positive definite, find $k(\cdot)$ such that

$$\phi_{12}(t)=\int_0^\infty \phi_{11}(t-\tau)k(\tau)\mathrm{d}\tau, \qquad t\geq 0 \tag{1}$$

where $k(t) = 0$, $t < 0$. Equations of this type were apparently first introduced by Hvol'son (1894, Leningrad) while studying the scattering of light by milk glass, and later (ca 1920), by E. A. Milne, K. Schwarzschild and others in problems in astrophysics.

While such equations attracted considerable attention, explicit analytical solution was long thought to be impossible, a point of view that, in the words of M. G. Krein (1958) 'was refuted in 1931 by the brilliant achievement of E. Hopf and N. Wiener'. In fact, so striking in its ingenuity and simplicity was their solution that not only the technique, but the equation itself came to be known by the name 'Wiener–Hopf'. Moreover a decade later, Wiener encountered the equation again, with even more significant consequences!

The genesis of these events was apparently (Pincus (1981)) a summer evening with E. Hopf at Wiener's country home in New Hampshire, where characteristically Wiener asked his visitor about challenging open problems in his field. The next morning Wiener came down for breakfast with a 'solution'. It in fact needed some reworking, which was done and the result was published in 1931, with the unusual (for mathematicians) nonalphabetical ordering of names. We shall briefly outline the Wiener–Hopf idea (ignoring various technical assumptions, for which see, e.g., Krein (1958)).

The tools used in the solution had in fact largely been developed by Wiener himself in collaboration with R. E. A. C. Paley. First were the properties in the complex plane of Fourier transforms of one-sided (or 'causal') functions. One motivation for Wiener's interest in such topics may have been his study of electrical networks and his 1931 patent (with Y. W. Lee, Sc. D., Elec. Engg.) on a new cascade realization (based on Laguerre functions), of network impedance functions, known to mathematicians as Caratheodory (or positive-real) functions. In any case, for us the key property is that (again under certain conditions) a function $k(\cdot)$ such that $k(t) = 0$, $t < 0$, has a Fourier transform $\mathcal{K}(\omega)$ whose extension to the complex plane, $\mathcal{K}(s)$, $s = \sigma + i\omega$, is analytic in the RHP, $\sigma > 0$.

The second tool was the closely related fact that such extensions of nonnegative even functions

$$0 \leq \Phi(\omega) = \Phi(-\omega) \qquad -\infty < \omega < \infty$$

have unique canonical factorizations

$$\Phi(s) = \Psi(s)\Psi^*(-s^*), \qquad s = \sigma + i\omega \tag{2}$$

where $\Psi(s)$ is analytic and bounded in the RHP, $\sigma > 0$, while $\Psi^{-1}(s)$ is analytic in the RHP, $\sigma > 0$. It may be noted that Wiener had recognized such functions, $\Phi(\cdot)$, as Fourier transforms of autocorrelation (real, even and positive definite) functions and described their significance as power spectral density functions. An equivalent result was discovered by Khinchin in 1934, leading to the so-called famous Wiener–Khinchin theorem. [It turns out that they had been anticipated by no less than Albert Einstein, in a two-page 1914 sketch answering a question raised by a friend on measuring the 'power' of meteorological time-series!] Now for the Wiener–Hopf technique.

We first extend the equation (1) to the whole line by introducing the function

$$g(t)=\phi_{12}(t)-\int_0^\infty k(\tau)\phi_{11}(t-\tau)\mathrm{d}\tau \qquad -\infty<t<\infty.$$

By the fact that (1) only holds for $t \geq 0$, we know that $g(\cdot)$ is one-sided, but 'anticausal',

$$g(t) = 0 \qquad t \geq 0$$

unlike the unknown 'causal' function $k(\cdot)$,

$$k(t) = 0 \qquad t < 0.$$

Of course now we have two unknown functions, $g(\cdot)$ and $k(\cdot)$, but give us a moment. Take Fourier transforms in the complex plane to get

$$G(s) = \Phi_{12}(s) - \mathcal{K}(s)\Phi_{11}(s).$$

Now use the canonical factorization (2) to write

$$\frac{G(s)}{\Psi^*(-s^*)}=\frac{\Phi_{12}(s)}{\Psi^*(-s^*)}-\mathcal{K}(s)\Psi(s)$$

But by construction $G(s)$ and $1/\Psi^*(-s^*)$ are analytic in the LHP, $\sigma < 0$, while $\mathcal{K}(s)$ and $\Psi(s)$ are analytic in the RHP, $\sigma > 0$. Equivalently the time function obtained by

inverse Fourier transformation (IFT) of $G(s)/\Psi^*(-s^*)$ will be zero for $t \geq 0$ (anticausal), while the time function corresponding to $\mathcal{K}(s)\Psi(s)$ will be zero for $t < 0$ (causal). This means that the latter time function must be equal to the $t \geq 0$ portion of the IFT of $\Phi_{12}(s)/\Psi^*(-s^*)$, leading to the famous formula for the solution of the Wiener-Hopf equation (1)

$$\mathcal{K}(s) = \frac{1}{\Psi(s)} \int_0^\infty dt e^{-st} \oint \frac{\Phi_{12}(p)}{\Psi^*(-p^*)} e^{pt} \frac{dp}{2\pi i}. \quad (3)$$

We shall return to this equation in § 5, after an apparent digression.

3. The problem of anti-aircraft fire control

Seeking eagerly to contribute to the War effort, Wiener submitted a proposal to the National Defense Research Council (NDRC) for a novel parallel computing machine for solving partial differential equations, as arising for example in aerofoil design problems. However, the proposal was deemed unlikely to be completed in a reasonable time frame. Wiener cast about for a more relevant problem and hit upon the problem of anti-aircraft fire control. In Dec. 1940, he sought and won a ($2350!) project on '*design of a lead or prediction apparatus in which, when one member follows the track of an airplane, another anticipates where the airplane is to be after a fixed lapse of time*'. Under certain assumptions, the problem can be reduced to one of finding an effective method of approximating an exponential function by a rational function of suitable order, which Wiener proposed to do by using ideas from his earlier (ca 1930) work with Y. W. Lee on network synthesis. However working with the engineer hired for the project, Julian Bigelow, Wiener came to see in a few months, that because of noise and of model uncertainties, a satisfactory engineering solution demanded explicit consideration of the *statistical nature* of the problem, and also the use of an *optimization criterion*. Effectively assuming that the airplane trajectories were sample functions of a stationary and ergodic random process, and imposing the requirements of a *linear apparatus* and of the *mean-square-error criterion* led him again to the Wiener–Hopf integral equation. So the solution was at hand! But unlike the apocryphal mathematician, who promptly drops a problem when he reduces it to one that has already been solved, Wiener realized that much more was needed for a real engineering solution.

In fact, Wiener's project reports make a beautiful (and perhaps the first) case study in how to use the theory of *optimal* solutions in the real world. First, he confined himself to the practically most significant case: rational power spectral densities, necessary for '*the actual realization in the field by a finite electrical or mechanical structure*'. Here are a few other quotations: '*the detailed design of a filter involves ... choices ... which must be justified economically*'. '*Let us see ... the stages that are needed. The first stage determines the irreducible error, i.e., the error which cannot be reduced by any delay whatever. ... Next, ... , we can determine a reasonable delay, such that the delay error is not large in comparison with the irreducible error, the sense of 'large' depending on the problem.*' Writing about the assumption of stationarity, he reassures the user that while the available '*statistical information will in fact never be complete, as our information does not run indefinitely for back into the past, it is a legitimate simplification of the facts to assume that the available information runs back much farther into the past than we are called upon to predict the future*'. And so on. Few mathematicians before Wiener had such an interest in the many issues arising in actual implementation.

Despite all this, however, the results actually obtained in field trials were not satisfactory, and the project was terminated. But Wiener wrote up the theoretical results obtained in the course of the work, along with some mathematical preliminaries and his philosophical views on the nature of his new approach to the field of communication engineering, in a remarkable 1942 monograph 'Extrapolation, Interpolation and Smoothing of Stationary Time Series'. It was later declassified and published in 1949 by J. Wiley and the MIT Press. The classified report was widely circulated, while the engineers of the day were hard pressed to follow its contents, it was closely studied by mathematicians such as N. Levinson and R. Philips at MIT, and H. W. Bode, C. E. Shannon, R. Blackman and others at Bell Laboratories. They all wrote up expositions and variations of Wiener's results, some of which we shall mention later.

While preparing his report, Wiener was made aware (by W. Feller) that some-what earlier, and apparently as a purely mathematical investigation, the famous Soviet mathematician A. N. Kolmogorov had studied the prediction problem for general discrete-time stationary processes. The approaches were quite different: Wiener's concrete and focused on applications, Kolmogorov's more abstract and more general. Though later both the methodologies – concrete and abstract – became useful in applications, it was Wiener's work that had a greater impact on electrical engineering, far beyond the original problem and solution.

4. Two new paradigms

In fact, Wiener was quite aware (and perhaps too much so) of the revolutionary significance of his work and ideas, and he stated his views quite emphatically in his 1942 monograph and in various other talks and papers. several passages from these works are quoted here, to show the clarity and prescience of Wiener's thinking.

We begin with his clear and forceful statement of the **statistical nature of the communication problem**. Thus on p. 2 of his monograph:

> *'Communication engineering concerns itself with the transmission of messages. For the existence of a message, it is indeed essential that variable information be transmitted. The transmission of a single fixed item of information is of no communicative value. We must have a repertory of possible messages, and over this repertory a measure determining the probability of these messages.'*
>
> *'A message need not be the result of a conscious human effort for the transmission of ideas. For example the records of current and voltage kept on the instruments of an automatic substation are as truly messages as a telephone conversation.'*

And then on p. 4:

> *'A statistical method, as for example a method of extrapolating a time series into the future is judged by the probability with which it will yield an answer correct within certain bounds, or by the mean (taken with respect to probability) of some point*, i.e., *function or norm of the error contained in its answer.'*
>
> *'No apparatus for conveying information is useful unless it is designed to operate, not on a particular message, but on a set of messages, and its effectiveness is to be judged on the way performs on the average on messages of this set. ... The apparatus to be used for a particular purpose is that which gives the best result 'on the average' in an appropriate sense of the word 'average'.'*

To those now familiar with Information Theory, presented by Shannon in 1948, and Signal Detection Theory as presented by P. M. Woodward and others in the early 1950s, it is worth emphasizing that the above passages were written in 1942. As one major illustration of their influence, we may note Shannon's gracious acknowledgement at the end of his magnificent 1948 paper founding Information Theory:

> *'Credit must also be given to Professor Norbert Wiener, whose elegant solution of the problems of prediction and filtering has considerably influenced the writer's thinking in this field.'*

But there is more. Going beyond the statistical problems he had studied, Wiener stressed the possibility of formulating a variety of engineering problems as *optimization* problems, so that *performance limits* may be calculated and *reasonable* solutions sought; this stands in sharp contrast to many earlier *trial and error* methods:

> *'These specifications give us an optimum filter to fit the situation exactly, whereas the earlier methods designed filters to certain specifications concerning passbands, sharpness of cut-off*, etc., *which stood in no obvious relation to the actual demands of a problem and had to be adjusted to these by the process of trial and error* [N. Y. Acad. Sci. lecture, Oct. 1946].'

and again from his 1942 monograph:

> *'Prediction and filtering do not exhaust the capacity of our methods. They may be applied whenever an ideally desirable linear operation ... is in fact not strictly realizable, although an approximation may be realized.'*

In particular, Wiener suggested that his ideas could be applied to the design of compensators for control systems. This suggestion was picked up in the Ph D research of R. Newton at MIT, who, along with L. Gould and J. Kaiser, wrote a book on the topic entitled *Analytical Design of Linear Feedback Controls*, Wiley, 1957. From it, we may quote passages that clearly echo Wiener's early insights (in the fifties, authors were not as gender conscious as they are now):

> *'... The analytical design procedure has several advantages over the trial and error method, the most important of which is the facility to detect immediately and surely an inconsistent set of specifications. The designer obtains a 'yes' or 'no' answer to the question of whether it is possible to fulfill any given set of specifications; he is not left with the haunting thought that if he had tried this or that form of compensation he might have been able to meet the specifications.'*
>
> *'... Even if the reader never employs the analytical procedure directly, the insight that it gives him into linear system design materially assists him in employing the trial and error design procedure.'*

However more than 50 years later, Wiener's message is apparently still worthy of repetition. Thus, let us quote from a very recent text book by M. Green and D. Limebeer, *Linear Robust Control*, Prentice-Hall, 1995:

> *'One does not want to waste time trying to solve a problem that has no solution, nor does one want to accept specification compromises without knowing that these are necessary. A further benefit of optimization is that it provides an absolute scale of merit against which any design can be measured – if a design is already all but perfect, there is little point in trying to improve it further.'*
>
> *'The aim of this book is to develop a theoretical framework within which one may address complex design problems with demanding specifications in a systematic way.'*

We continue with some further remarks from this book on the Wiener ideas, though the authors here cite

also the work of R. E. Kalman, which we shall discuss in a later section. From p. 2, we quote (note the 'free ride' that Hopf gets; he perhaps was completely unaware of these applications):

Wiener–Hopf–Kalman optimal control

'*The first successes with control system optimization came in the 1950s with the introduction of the Wiener–Hopf–Kalman (WHK) theory of optimal control. At roughly the same time the United States and the Soviet Union were funding a massive research program into the guidance and maneuvering of space vehicles. As it turned out, the then new optimal control theory was well suited to many of the control problems that arose from the space program.*'

'*... [A] revolutionary feature of the WHK theory is that it offers a true synthesis procedure. Once the designer has settled on a quadratic performance index to be minimized, the WHK procedure supplies the (unique) optimal controller without any further intervention from the designer.*'

However, the major motivation for Green and Limebeer's 1995 book was the fact that they were introducing a new (so-called H_∞) formulation:

'*In contrast to experience with aerospace applications, it soon became apparent that there was a serious mismatch between the underlying assumptions of the WHK theory and industrial control problems. Accurate models are not routinely available and most industrial plant engineers have no idea as to the statistical nature of the external disturbances impinging on their plant*

Worst-case control: H_∞ optimization

'*H_∞ optimal control is a frequency-domain optimization and synthesis theory that was developed in response to the need for a synthesis procedure that explicitly addresses questions of modeling errors.*'

Ironically, the new H_∞ theory can be regarded as Wiener–Kalman theory, but in *Krein space* rather than in Hilbert space. More on this later.

Here, however, as a final quotation from Wiener's own monograph, let us present one that illustrates Wiener's keen insight into what mathematical refinements are relevant to engineers. Engineering textbooks are often still concerned about the level of rigor at which to present Fourier Theory – should one worry about pointwise convergence, L_2 convergence, etc? Wiener cuts right to the heart of the matter:

'*Up to the present we have been treating the Fourier series as a purely formal expression without any regard to whether it converges or not. ... Now it is obvious that no physical quantity can be observed for a single precise value of t. ... Thus all functions of t are for the physicist averages over small ranges of t rather values of a precise point of t.*' [He goes on to very simply show that] '*all local averages of the formal [Fourier] series [of f(t)] converge to the local average of f(t). As we have pointed out, this is all that we need to make a practical employment of the Fourier series for f(t).*'

This is the approach that later in 1948 L. Schwartz elaborated in his theory of distributions which, among many other things, made rigorous the widespread use of impulsive functions by engineers. I cannot resist remarking that Wiener also uses the 'Sampling Theorem for Bandlimited Functions', first derived by J. M. Whittaker in 1915, but to this day widely cited as Shannon's Sampling Theorem. There are several other little gems in Wiener's report, now fortunately also available in an MIT Press paperback entitled 'Time Series'. But let us now return to some more explicit mathematical problems.

5. The Wiener filter

The appearance in 1949 of Wiener's monograph, and of Shannon's work on Information Theory, generated a huge wave of activity in, to use Wiener's phrase, '*mathematical engineering*'. A striking range of mathematics has and is being used in these studies, as can be seen by glancing through, for example, the IEEE Transactions on Information Theory, on Automatic Control, on Signal Processing, on Circuits and Systems, on Image Processing, etc. Many of the algorithms were too complex for implementation till about five to ten years ago, but the pace of technology insertion is accelerating.

In the rest of this paper, we shall discuss one of the directions most closely related to Wiener's own work, focusing on some problems left open by him: spectral factorization of matrix-valued rational spectral densities; finite-time problems/nonstationary processes; nonlinear estimation/random signal detection. The idea is to give a glimpse of part of the wide range of consequences arising from just one of Wiener's mathematical contributions.

We begin with a review of the so-called (causal and noncausal) Wiener filters. A filter is a selective device, e.g., one that allows the transmission of certain frequency regions (the passband) and rejects certain other regions (the stop band). In filtering signals out of a combination of the signals and additive noise, the conventional solution was a device that rejects as much of the noise as possible while passing the signal through undistorted. If as is usually the case, the signal is bandlimited (i.e., occupies a limited frequency range) while the noise is wideband, the solution is the so-called Ideal Filter, with unit gain in the frequency regions of the signal and zero gain elsewhere. Wiener pointed out that if

the signal is a random process, one might better be more concerned about regions where the power spectrum of the signal is much larger than that of the noise, with proper weighting to be assigned to different frequency regions depending upon the choice of an optimization criterion. The most fruitful has been Wiener's own choice of a least-squares criterion.

To be specific, consider the problem of estimation the value at a given time t of a stochastic signal process $s(\cdot)$ given the noisy observations

$$y(\tau) = s(\tau) + v(\tau) \qquad -\infty < \tau < \infty \qquad (4)$$

by using a linear time-invariant filter $h(\cdot)$ such that

$$\hat{s}(t) = \int_{-\infty}^{\infty} h(t-\tau) y(\tau) d\tau$$

satisfies (E denotes expectation)

$$E|s(t) - \hat{s}(t)|^2 = \text{minimum.}$$

Under the assumption that the processes $\{s(\cdot), v(\cdot)\}$ are zero mean and jointly stationary, it is not hard to check that the optimum filter must satisfy the equation

$$\phi_{sy}(t) = \int_{-\infty}^{\infty} h(t-\tau) \phi_{yy}(\tau) d\tau$$

where the $\phi(\cdot)$ are the autocorrelation functions

$$\phi_{sy}(t) = Es(\sigma + t) y^*(\sigma), \qquad \phi_{yy}(t) = Ey(\sigma + t) y^*(\sigma).$$

This equation is easily solved by Fourier transformation to yield (using capitals for the transforms)

$$H(\omega) = \frac{\Phi_{sy}(\omega)}{\Phi_{yy}(\omega)}.$$

When the signal and noise processes are uncorrelated with each other, this reduces to

$$H(\omega) = \frac{\Phi_{ss}(\omega)}{\Phi_{ss}(\omega) + \Phi_{vv}(\omega)}. \qquad (5)$$

To reflect the fact that the noise has a much higher frequency range than the signal, engineers often assume that the noise is *white*,

$$\Phi_{vv}(\omega) = R > 0, \quad -\infty < \omega < \infty. \qquad (6)$$

Such noise processes of course fall out of the scope of the usual theory of stationary processes, since for one thing, the IFT of $\Phi_{vv}(\omega)$ is a delta function. And, in fact, since to have finite power all physical processes must have spectra that decay to zero as $\omega \to \infty$, white noise is nonphysical as well. Yet it is widely used, as a model, for several important (mathematical *and* physical) reasons that we do not have time to explain here. However, it is interesting to see Wiener's free use of the white noise model, which is important for many reasons, both theoretical and practical:

> '*As for (the noise), we shall take a case which, although not formally contained in the theory we have given, constitutes the limiting case of it, and one of the greatest importance in practice. This is the case in which the noise input is due to a shot effect and has an equipartition of power in frequency. Theoretically, of course, this is not strictly realizable, as it would demand an infinite power; practically as in the case of Plank's law in optics, it may hold within the limits of observation up to frequencies of a magnitude so great that they are no longer of interest for our particular problem.*'

To return to our problem, with the white noise assumption, we can see that

$$H(\omega) = \frac{\Phi_{ss}(\omega)}{\Phi_{ss}(\omega) + R} \to \frac{\Phi_{ss}(\omega)}{R} \text{ as } R \to \infty.$$

On the other hand, as $R \to 0$, $H(\omega)$ tends to the ideal filter (with unit gain in the passband of the signal), but this is not the optimum filter when $R \neq 0$. As $R \to \infty$, we note that $H(\omega) \to \Phi_{ss}(\omega)/R)$, so that the filter tends to reinforce frequencies where $\Phi_{ss}(\omega)/R > 1$, and to suppress the signal in regions where $\Phi_{ss}(\omega)/R < 1$. This fits, in hindsight, with our intuition, but the theory is necessary to tell us what to do for arbitrary values of R, or for nonwhite (often called coloured) noise.

The above solution, though given by Wiener, does not use the Wiener–Hopf equation, because we are assuming that the process $y(\cdot)$ is observerd over all time instants, past as well as future. When the observations of $y(\cdot)$ are restricted to the past, we have to solve a Wiener–Hopf equation, which when (4) and (6) hold, takes the form

$$Rk(t) + \int_0^{\infty} k(\tau) \phi_{ss}(t-\tau) d\tau = \phi_{ss}(t), \quad t \geq 0. \qquad (7)$$

With the white noise assumption, the formula for the solution takes the striking form (apparently first noted by Yovits and Jackson (1960))

$$\mathcal{K}(\omega) = 1 - R^{1/2} \Psi^{-1}(\omega). \qquad (8)$$

In other words, the *canonical spectral factorization* completely defines the estimator in the additive white noise case!

One might justly wonder about the physical significance of the canonical factorization, and the stochastic problem allows a nice (and far reaching) interpretation, first given by Bode and Shannon (1950), and independently (and in somewhat more general form) in a lesser-known paper of Zadeh and Ragazzini, also appearing in 1950. These authors noted that passing the observations process $y(\cdot)$ through a linear filter with transfer function $\Psi^{-1}(\omega)$ gives us a process $e(\cdot)$ with power spectrum (using well-known formulas),

$$\Phi_{ee} = \Psi^{-1}(\omega) \Phi_{yy}(\omega) \Psi^{-*}(-\omega) = I.$$

Therefore, we can interpret the first term $\Psi^{-1}(\omega)$ in the general Wiener–Hopf formula (3)

$$\mathcal{K}(\omega)=\frac{1}{\Psi(\omega)}\left[\int_0^\infty dt e^{-j\omega t}\oint\frac{\Phi_{12}(p)}{\Psi^*(-p^*)}e^{pt}\frac{dp}{2\pi i}\right],$$

as allowing us to replace the observations process $y(\cdot)$ by a much simpler stochastic process $e(\cdot)$, for which the problem of estimating a related stochastic process $s(\cdot)$ turns out to be much simpler: when $\phi_{11}(\cdot)$ in the Wiener–Hopf equation

$$\phi_{12}(t)=\int_0^\infty\phi_{11}(t-\tau)k(\tau)d\tau,\quad t\geq 0$$

is a delta function, the solution is immediate: $k(t)=\phi_{12}(t)$, $t\geq 0$.

One might wonder if there is a 'loss of information' in going from the original observed process $y(\cdot)$ to the white process $e(\cdot)$? The answer is no, because by the fact that the canonical factor $\Psi(s)$ and its inverse $\Psi^{-1}(s)$ have the property that they are analytic in the RHP, one can pass (recall the Paley–Wiener results quoted in § 3) from $y(\cdot)$ to $e(\cdot)$ and from $e(\cdot)$ back to $y(\cdot)$ by causal and stable linear operations. Since

$$\mathcal{F}^{-1}\{\Phi_{ee}(\omega)\}=Ee(\tau+t)e^*(\tau)=\delta(\tau)=0,\ \tau\neq 0$$

the value of $e(\cdot)$ at any instant is uncorrelated with its values at any other instant, and therefore every observation, $e(t)$, brings *new* information, which cannot be said about a (corrected) nonwhite process. The process $e(\cdot)$ is called the *innovations process* of $y(\cdot)$; we shall return to it in a more general context in § 8. The innovations concept has been useful in extensions of Wiener filtering theory to nonstationary processes and to nonlinear problems, see e.g., Kailath (1970), Davis (1977) and Fujisaki *et al.* (1971). We may note that, Kolmogorov's more general approach to the discrete-time prediction problem (1939), (1941) was based on the use of the innovations process, which avoids (or rather trivializes, as we noted earlier) the use of the Wiener–Hopf equation. Thus somewhat ironically, Kolmogorov's more abstract approach ultimately became more powerful than Wiener's more concrete approach, a phenomenon, mathematicians may be pleased to know, that is not uncommon in applications.

6. Extensions: Matrix spectral factorization

Wiener's monograph inspired various attempts at extensions – to finite-time nonstationary problems, and to vector-valued processes in particular. When observations are only available over a finite time, say $(0, t)$ rather than $(-\infty, t)$, the W–H equation is replaced by one of 'W–H type',

$$h(t,s)+\int_0^t h(t,\tau)\phi(\tau-s)d\tau=\phi(t-\tau),\quad 0\leq s\leq t. \tag{9}$$

No general methods were or are known for its solution, and a vast literature developed on various special cases, tricks, etc; so much so that a 1958 editorial by P. Elias urged no more work on '*Two Famous Papers*'. One generic title was 'The Optimum Linear Mean Square Filter for Separating Sinusoidally Modulated Triangular Signals from Randomly Sampled Stationary Gaussian Noise, with Applications to a Problem in Radar'. (The other: 'Information theory, Photosynthesis and Religion'.)

The apparent mess was cleaned up by the use, by R. E. Kalman in 1960, of the *state-space description* of process with rational spectral densities. Such descriptions are actually of much older vintage, see, e.g. a paper by and were used by Wang and Uhlenbeck (1930) on Markov models for noise processes. Doob wrote two long papers on them in 1944 and 1949 but, alas, did not mention them in his very influential 1953 book! Had he done so many developments might have occurred much earlier.

The so-called Kalman (or sometimes Kalman-Bucy) filter has been widely discussed in a host of a papers and textbooks, e.g., Anderson and Moore (1979), Kailath (1981). It gains its power, as just noted, from the introduction of state-space models, which turns out to be equivalent to modelling stochastic processes as linear combinations of the components of a vector-valued Markov process. Briefly, we model a scalar process $s(\cdot)$ with an n-th order rational spectral density as

$$\begin{cases}s(t)=Hx(t)\\ \dot{x}(t)=Fx(t)+u(t)\end{cases}\qquad t\geq t_0 \tag{10}$$

where $H\in\mathcal{C}^{1\times n}$, $F\in\mathcal{C}^{n\times n}$ are known matrices, $u(\cdot)$ is an $n\times 1$ vector-valued zero-mean white noise process, with

$$\langle u(t), u(s)\rangle\triangleq Eu(t)u(s)^*=Q\delta(t-s),$$

and the initial *state*, $x(t_0)$, is such that

$$\langle x(t_0), 1\rangle=Ex(t_0)=0,\ \langle x(t_0), x(t_0)\rangle=\Pi_0,$$
$$\langle x(t_0), u(t)\rangle=0.$$

The matrices $Q\in\mathcal{C}^{n\times n}$ and $\Pi_0\in\mathcal{C}^{n\times n}$ are also assumed to be known.

We use the inner product notation to follow Kolmogorov in assuming that (zero-mean) random variables defining a (second-order) stochastic process live in a Hilbert space (or Hilbert module, when the random variables are vector-valued). Of course we are stretching this formulation when we deal with white noise processes, but rigor can be restarted by regarding a white noise process as the formal derivative of a process with orthogonal increments.

Though the linear system relating the stochastic input process $u(\cdot)$ to the output stocastic proces is time-invariant, the process $s(\cdot)$ will in general be nonstationary, because the 'transients' arising from the fact that the input is switched on at time t_0 and does not begin in the remote past. In fact, it is not hard to see that

$$\Pi(t) = \langle x(t), x(t) \rangle$$

will obey

$$\dot{\Pi}(t) = F\Pi(t) + \Pi(t)F^* + Q, \quad \Pi(t_0) = \Pi_0. \tag{13}$$

However, when F is 'stable', i.e., all its eigenvalues have strictly negative real parts, then it turns out that the process $s(\cdot)$ will be stationary if the initial state variance is chosen as

$$\Pi(t_0) = \Pi,$$

where Π is the unique nonnegative definite solution of the (Lyapunov) equation

$$0 = F\Pi + \Pi F^* + Q. \tag{14}$$

Now since we have introduced matrix notation, we can as well take the step of regarding $s(\cdot)$ as a $p \times 1$ vector-valued process, so that $H \in \mathcal{C}^{p \times n}$. The filtering problem is now a 'multichannel' problem of attempting to find

$\hat{s}(t)$ = the linear least-squares estimate of $s(t)$ given $\{y(\tau), t_0 \leq \tau < t\}$,

where

$$y(t) = s(t) + v(t), \quad t \geq 0 \tag{15}$$

and $v(\cdot)$ is a white noise process such that

$$\left\langle \begin{bmatrix} u(t) \\ v(t) \end{bmatrix}, \begin{bmatrix} u(s) \\ v(s) \end{bmatrix} \right\rangle = \begin{bmatrix} Q & S \\ S^* & R \end{bmatrix} \delta(t-s), \tag{16}$$

where $S \in \mathcal{C}^{n \times p}$ and $R \in \mathcal{C}^{p \times p}$ are also assumed to be known *a priori*. The presence of the additive white noise is essential to get useful results, and so it is assumed that R is strictly positive definite, $R > 0$.

It is widely believed that the reason for the greater scope of the Kalman theory (applying to vector-valued but finite-dimensional) stationary processes also to (finite-dimensional) nonstationary process is in fact that it starts, as above, with a model for the process $s(\cdot)$ rather than with power-spectral or covariance data. However this is not true – the Wiener and Kalman approaches become equivalent if one carries over the state-space characterization of the process to its power-spectra and/or covariance functions.

We shall illustrate this now by using the state-space model to solve a problem not satisfactorily resolved by Wiener and several later researchers – finding an effective way of computing the canonical factorization of a rational power spectral density function matrix. We start by noting that since the transfer function from the input white noise processes $\{u(\cdot), v(\cdot)\}$ to the output process $y(\cdot)$ is

$$[H(i\omega I - F)^{-1} \quad I], \tag{17}$$

the power-spectral density function of $y(\cdot)$ can be computed as

$$\Phi_{yy}(\omega) = [H(i\omega I - F)^{-1} \quad I] \begin{bmatrix} Q & S \\ S^* & R \end{bmatrix} \begin{bmatrix} (-i\omega I - F^*)^{-1} H^* \\ I \end{bmatrix}. \tag{18}$$

An alternative expression can be found by taking the Fourier transform of the covariance function:

$$\phi_y(\tau) = \langle y(t), y(t-\tau) \rangle = R\delta(t-\tau) + He^{F\tau}N1(\tau) + N^*e^{F^*\tau}H^*1(\tau), \tag{19}$$

where

$$N = \Pi H^* + S \tag{20}$$

and $1(\cdot)$ is the (Heaviside) step function

$$1(t) = \begin{cases} 1 & t > 0 \\ 1/2 & t = 0 \\ 0 & t < 0 \end{cases}$$

The Fourier transform of $\phi_y(\cdot)$ is

$$\Phi_{yy}(\omega) = [H(i\omega I - F)^{-1} \quad I] \begin{bmatrix} 0 & N \\ N^* & R \end{bmatrix} \begin{bmatrix} (-i\omega I - F^*)^{-1} H^* \\ I \end{bmatrix}. \tag{21}$$

Comparing (18) and (21) shows that different 'central' matrices could be used to specify $\Phi_y(\omega)$, some nonnegative definite as in (18), while, the one in (21) is indefinite. It is natural to ask how we can characterize the nonuniqueness of the central matrix? The answer is that we can use *any* central matrix of the form

$$M = \begin{bmatrix} Q + FZ + ZF^* & S + ZH^* \\ S^* + HZ & R \end{bmatrix}, \quad Z = Z^* \tag{22}$$

i.e., where Z is any Hermitian matrix. The choices $Z = 0$ and $Z = \Pi$ give the previous expressions (18) and (21). The fact that any such M yields $\Phi_{yy}(\omega)$ can be verified by a direct (but tedious) calculation. However, a nicer and more useful derivation can be obtained by allowing the random variables to live in an indefinite (Krein space), rather than in Hilbert space.

In a Krein space we can have elements such that $\langle u, u \rangle$ is indefinite or even zero. For example, we could have

$$\left\langle \begin{bmatrix} u^0(t) \\ v^0(t) \end{bmatrix}, \begin{bmatrix} u^0(s) \\ v^0(s) \end{bmatrix} \right\rangle \triangleq \begin{bmatrix} Q^0 & S^0 \\ S^{0*} & R^0 \end{bmatrix} \delta(t-s) = \begin{bmatrix} 0 & N \\ N^* & R \end{bmatrix} \delta(t-s). \tag{23}$$

In view of this, let us add to the original $\{u(\cdot), v(\cdot)\}$, elements $\{u^0(\cdot), v^0(\cdot)\}$ such that (in an obvious notation)

$$\begin{cases} \dot{x}(t) + \dot{x}^0(t) = F(x(t) + x^0(t)) + G(u(t) + u^0(t)) \\ y(t) + y^0(t) = H(x(t) + x^0(t)) + v(t) + v^0(t) \end{cases} \tag{24}$$

where

$$\left\langle \begin{bmatrix} u(t) \\ v(t) \end{bmatrix}, \begin{bmatrix} u^0(t) \\ v^0(t) \end{bmatrix} \right\rangle = 0 \text{ and } \langle y^0(t), y^0(\tau) \rangle = 0. \tag{25}$$

This can be done using earlier formulas (see (5)–(6)) to note that

$$\Phi_{y^0y^0}(\omega) = [H(i\omega I - F)^{-1} \quad I]$$

$$\begin{bmatrix} 0 & \Pi^0 H^* + S^0 \\ H\Pi^0 + S^{0*} & R^0 \end{bmatrix} \begin{bmatrix} (-i\omega I - F^*)^{-1} H^* \\ I \end{bmatrix}, \quad (26)$$

where $\Pi^0 = \Pi^{0*}$ is such that

$$F\Pi^0 + \Pi^0 F^* + Q^0 = 0. \quad (27)$$

From these we conclude that $\Phi_{y^0y^0}(\omega)$ will be identically zero if we choose

$$S^0 = -\Pi^0 H^*,\ Q^0 = -F\Pi^0 - \Pi^0 F^* \text{ and } R^0 = 0. \quad (28)$$

Finally setting $Z = -\Pi_0$ gives

$$\left\langle \begin{bmatrix} u(t)+u^0(t) \\ v(t)+v^0(t) \end{bmatrix}, \begin{bmatrix} u(s)+u^0(s) \\ v(\tau)+v^0(\tau) \end{bmatrix} \right\rangle$$

$$= \begin{bmatrix} Q+Q^0 & S+S^0 \\ S^*+S^0 & R+R^0 \end{bmatrix} \delta(t-s)$$

$$\begin{bmatrix} Q+FZ+ZF^* & S+ZH^* \\ S^*+HZ & R \end{bmatrix} \delta(t-s) \triangleq M\delta(t-s) \quad (29)$$

exactly as claimed above (see (22)). We see now that the arbitrary matrix Z can be interpreted as the (negative of) the state-variance matrix of a process with zero s-spectrum. However, so far we only have a formal calculation. The significant theorem is the so-called KYP Lemma (see Willems and Trentelman for a recent discussion):

Theorem (KYP Lemma) *When $\Phi_y(s) > 0$, $s = j\omega$, then there exists a $Z = Z^*$ such that the central matrix is non-negative definite (i.e., it is the covariance matrix of a collection of genuine random variables).*

- We do not need F to be stable; a weaker condition from linear system theory, a subject developed in the engineering literature of the last 30 years, will suffice: the pair $\{F, H\}$ should be detectable, i.e., it should be such that $[H^*sI - F^*]$, $s = \sigma + i\omega$, should be full rank for all $\sigma \geq 0$. Here however we shall, for simplicity, stay with the assumption that F is stable so that we are dealing with a stationary process $y(\cdot)$. There are important corollaries characterizing matrix *positive-real* (Caratheodory) and *bounded-real* (Schur) functions, which are widely encountered in applications. We note also that the Krein space interpretation introduced above can be used to give a simple geometric proof of the lemma; it also leads to several other results, e.g., unifying H_2 and H_∞ control (see, e.g., Hassibi *et al.* (1996)).

However while the KYP Lemma is an important result, which is why we mentioned it here, it is not necessary to use it to obtain a spectral factorization of $\Phi_{yy}(\omega)$. To this end, note that although we cannot make any assertions on the positivity of the central matrix M, defined in (28), the fact that

$$\Phi_{yy}(\omega) = [H(i\omega I - F)^{-1} \quad I]\, M \begin{bmatrix} (-i\omega I - F^*)^{-1} H^* \\ I \end{bmatrix} > 0 \quad (30)$$

shows that M has at least p positive eigenvalues. [Note that, for each ω, the above expression is the product of a $p \times (n+p)$, an $(n+p) \times (n+p)$ and an $(n+p) \times p$ matrix, which is positive definite (i.e., has p positive eigenvalues). Therefore the central $(n+p) \times (n+p)$ matrix, M, must have at least p positive eigenvalues.]

Now that we have shown that the matrix M has at least p positive eigenvalues for all choices of Z, it is interesting to ask whether Z can be chosen so that M has only p positive eigenvalues and no negative eigenvalues, i.e., if Z can be chosen so that M has minimal rank p. To see that this is indeed possible recall the easily verified factorization (recall that we have assumed the invertibility of R).

$$M = \begin{bmatrix} Q+FZ+ZF^* & S+ZH^* \\ S^*+HZ & R \end{bmatrix}$$

$$= \begin{bmatrix} I & K \\ 0 & I \end{bmatrix} \begin{bmatrix} \Delta & 0 \\ 0 & R \end{bmatrix} \begin{bmatrix} I & 0 \\ K^* & I \end{bmatrix}, \quad (31)$$

where

$$\Delta(Z) \triangleq Q + FZ + ZF^* - (S + ZH^*)(R(S + ZH^*)^* \quad (32)$$

$$K \triangleq (S + ZH^*)R^{-1}. \quad (33)$$

Therefore $\Phi_{yy}(\omega)$ in (30) can be written as

$$\Phi_{yy}(\omega) = H(i\omega - F)^{-1}\Delta(Z)(-i\omega - F^*)^{-1}H^* +$$

$$[I + H(i\omega - F)^{-1}K]R[I + H(i\omega - F)^{-1}\mathrm{K}]^* \quad (34)$$

The second term on the RHS is $p \times p$ and non-negative definite, so we can immediately obtain a factorization by choosing Z so that it satisfies

$$0 = \Delta(Z) = FZ + ZF^* - (S + ZH^*)R^{-1}(S + ZH^*)^* \quad (35)$$

The only issue is whether the resulting spectral factor has a well-defined inverse, viz., one that when extended into the complex plane is analytic in the right half plan (cf. (2)). There is an interesting result here. The nonlinear algebraic equation (35), which for reasons explained below is called an Algebraic Riccati Equation (or ARE), has many solutions. However it can be shown that when F is stable (or even just when $\{F, H\}$ is detectable), there is one and only one non-negative solution, say P; moreover, this solution is such that the spectral factor

$$\Psi(s) \triangleq [H(sI - F)^{-1}K + I]R^{1/2},\ K = S + PH^*R^{-1} \quad (36)$$

and its inverse

$$\Psi^{-1}(s) \triangleq [H(sI - F)^{-1}KR^{1/2} + R^{1/2}]^{-1}$$

$$= R^{-1/2}[I - H(sI - F + KH)^{-1}K], \quad (37)$$

are both analytic in the right half plane. There are several computationally effective methods of finding the desired nonnegative definite solution of the ARE – a good source is the reprint volume edited by Bittanti *et al.* (1995). So the introduction of the ARE, first done in the Kalman theory, overcame what had been regarded as one of the stumbling blocks to the Wiener theory. A minor quibble is that the factorization is expressed in terms of the parameters $\{F, G, H, Q, R, S\}$ of a particular model for the process rather than in terms of the spectral data. Now for the state-space model, the covariance and the spectral density are fixed by (*cf.* (19)–(21)) by $\{H, F, N\}$. To use this data, all we need to do is to choose the central matrix M not as in (31), but as (*cf.* (19)–(22)).

$$\begin{bmatrix} 0+FZ+ZF^* & N+ZH^* \\ N^*+HZ & R \end{bmatrix} \tag{38}$$

Then proceeding as before, the rank of this matrix can be dropped by choosing Z so that it satisfies

$$0 = FZ + ZF^* - (N + ZH^*)R^{-1}(N + ZH^*)^* \tag{39}$$

which will lead to a factorization of the form

$$\Phi_{yy}(s) = [I + H(sI - F)^{-1}(N + ZH^*)]R^{-1}[\ \dots\ ]^* \tag{40}$$

The particular choice that will give a factor with an inverse that is analytic in the right-hand plane can be shown to be the unique negative semidefinite, say $-\Sigma$, solution of the ARE (39). The corresponding factor is therefore

$$\Psi(s) = [I + H(sI - F)^{-1}K]R^{1/2}, \tag{41}$$

where we define

$$K \triangleq N - \Sigma H^*, \tag{42}$$

and

$$\Sigma \geq 0,\ 0 = F\Sigma + \Sigma F^* + (N - \Sigma H^*)R^{-1}(N - \Sigma H^*)^* \tag{43}$$

The reader may have wondered that we used the *same* symbols $\Psi(s)$ and K as in the earlier formula (36) – the reason is that the canonical factorization is unique! [This implies the interesting identity $\Pi = P + \Sigma$, which we shall not explore here.]

To close the story of Wiener filtering, let us note that with the canonical factor in hand, we can really write down the optimal filter by using (8) and (37)

$$\mathcal{K}(s) = I - R^{1/2}\Psi^{-1}(s) = H(sI - F + KH)^{-1}K \tag{44}$$

where K can be found either from the model parameters, as in (36), or from the covariance/spectral parameters, as in (42). This is a reasonably explicit formula for the optimal filter, but another advantage of the state-space formulation is that we can readily write down a state-space model for the filter:

$$\dot{\hat{x}}(t) = (F - KH)\,\hat{x}\,(t) + Ky(t),\quad \hat{x}\,(t_0) = 0 \tag{45}$$

$$\hat{s}(t) = H\hat{x}(t) \tag{46}$$

as can be verified by checking that the transfer function from $y(\cdot)$ to $\hat{s}\,(\cdot)$ is exactly as in (44). We have used the notation $\hat{x}\,(\cdot)$ for the state-variable in (45), because in fact we have a bonus: $\hat{x}\,(\cdot)$ is the linear least-mean-squares estimate of the state $x(\cdot)$ itself (under the assumption that $\{F, H\}$ is observable, i.e., $[sI - F^*H^*]$ has full rank for all $s \in C$).

We close with some remarks that, *inter alia*, will fulfill our promise to explain the name ARE. The first remark is obtained by going back to our state-space model, (10) *seq*. Observe that stationarity arose from a particular choice of initial condition, $\Pi(t_0) = \Pi$ defined as the unique matrix such that

$$\Pi \geq 0,\ 0 = F\Pi + \Pi F^* + Q$$

For any other choice of $\Pi(t_0)$, or if F is unstable (so that (14) will not have a solution $\Pi \geq 0$), the process $s(\cdot)$ will be nonstationary, with covariance function

$$Es(t + \tau)s^*(t) = He^{F\tau}\,N(\tau),\ \tau \geq 0$$

where

$$N(t) = \Pi(t)H^*,\ \dot{\Pi}(t) = F\Pi(t) + \Pi(t)F^* + Q,$$

In turns out that the previous discussions can all be extended by now working in terms of covariance functions rather than power-spectral-density functions. The key change is that instead of the algebraic (Riccati) equation

$$P \geq 0,\ 0 = Q + FP + PF^*KRK^*,\ K \equiv (S + PH^*)R^{-1}$$

we shall have the matrix Riccati differential equation,

$$\dot{P}(t) = Q + FP(t) + P(t)F^* - K(t)RH^*(t),\ P(t_0) = \Pi(t_0)$$

$$K(t) \triangleq (S + P(t)H^*)R^{-1}$$

When the state is one dimensional, the resulting quadratically nonlinear equation is the one apparently first studied by Jacopo Francesco, Count Riccati, and later introduced by Legendre and others into the calculus of variations. It was introduced into control theory by R. E. Bellman (1957) and the matrix version by especially R. E. Kalman (1960).

Explicit analytic solution of the Riccati equation is impossible in the matrix case. But fortunately, this is a (nonlinear) initial value problem, so it can be solved via a discretization scheme, e.g., in the naive way,

$$P(t + \delta) = P(t) + \delta[Q + FP(t) + P(t)F^* - K(t)RK^*(t) + 0(\delta),\ t = 0, 1, 2, \dots$$

Now, an important observation is that once the need for some computer-based iterative algorithm is realized, one might further guess that there is no particular need to restrict oneself to time-invariant systems: one can just as easily consider time-variant models,

$$\dot{x}\,(t) = F(t)x(t) + v(t),\ t \geq t_0$$

$$y(t) = H(t)x(t) + v(t)$$

with

$$\left\langle \begin{bmatrix} u(t) \\ v(t) \end{bmatrix}, \begin{bmatrix} u(s) \\ v(s) \end{bmatrix} \right\rangle = \begin{bmatrix} Q(t) & S(t) \\ S^*(t) & R(t) \end{bmatrix} \delta(t-s).$$

Now the (Riccati) iteration is still as before,

$$P(t+\delta) = P(t) + \delta[Q(t) + F(t)P(t) + \ldots] + 0(\delta),\ t = 0, 1, 2, \ldots$$

except that we need to store/know the values of the functions $\{F(\cdot), Q(\cdot), \ldots\}$.

We have thus arrived at the Kalman-(Bucy) filtering algorithm. There is a vast literature on it, with several significant results and issues. Here, we go on to a different kind of extension of Wiener's results, involving finite-time discrete time series. That discussion will lead us to a concept called *displacement structure*, which actually had its roots in studies of the Riccati equation.

7. Beyond state-space models/displacement structure

In one of several different variations of Wiener's problem, his colleague N. Levinson in 1947 studied a finite-time discrete prediction problem, where the Wiener–Hopf equation was replaced by a set of linear equations with a Toeplitz coefficient matrix. He proposed a fast recursive solution, now known as the *Levinson algorithm*, very widely used in geophysical data processing (beginning in the mid-fifties) and in speech processing (beginning in the mid-sixties). Kolmogorov's (1939) formulation of the prediction problem gives an interesting insight into this algorithm, and led to connections with the work of Szegö (1939) and Geronimus (1939) on orthogonal polynomials, of Schur (1917) on H^∞ functions, and then to new results on Toeplitz-like matrices and more generally matrices with displacement structure (see below).

The Kolmogorov Isomorphism: The identity

$$\langle y_k, y_l \rangle = E_{y_k y_l^*} = r_{k-l} = \oint z^k z^{-l} \frac{dF(z)}{2\pi z} = \langle z^k, z^l \rangle_2$$

allows one to form an isometric mapping between the Hilbert space of random variables spanned by $\{y_k\}$, and the Hilbert space of functions on the unit circle spanned by the $\{z^k\}$.

Then the finite-interval prediction problem: find $\{a_{k,j}\}$ to minimize

$$E \| y_k + a_{k,1} y_{k-1} + \ldots + a_{k,m} y_{k-m} \|^2$$

is equivalent to the polynomial approximation problem: find $\{a_{k,j}\}$ to minimize

$$\oint |z^k + a_{k,1} z^{k-1} + \ldots + a_{k,m} z^{k-m}|^2 \frac{dF(z)}{2\pi z}.$$

It turns out that around 1920 Szegö had shown that the minimizing polynomials

$$a_m(z) = z^m + a_{m,1} z^{m-1} + \ldots + a_{m,m}$$

had the nice property that they were orthogonal to each other w.r.t. the measure $F(z)$. Szegö and others went on to make many studies of these orthogonal polynomials. Among other results, in 1939, Szegö and Geronimus independently discovered that these polynomials obeyed a two-term (rather than the usual 3-term) recursion:

$$a_{m+1}(z) = a_{m(z)} - k_{m+1} z\, a_m^{\#}(z),\ a_m^{\#}(z) = \text{the reciprocal polynomial}$$

where $k_{m+1} = -a_{m+1,\,m+1}$, the constant term in $a_{m+1}(z)$. This is in fact almost the same as the recursion discovered by Levinson in 1947, except that to obtain a true recursion one needs to be able to compute k_{m+1} in terms of $\{F(z), a_m(z)\}$. This could have been done by Szegö or Geronimus, had they been interested in actual computation; however they were more interested in the asymptotic properties of the polynomials (in fact, a famous Szegö formula is just the formula discovered by Kolmogorov and Wiener for the irreducible error in prediction). A survey of the connections between orthogonal polynomial theory and linear estimation, and their fascinating continuous-time analogs, can be found in Kailath *et al.* (1978).

Later it was discovered that a more farreaching connection could be made with some of the work of I. Schur, who was well ahead of his time with his interest in computation. In 1917, he wrote a remarkable paper giving a computationally efficient solution to the Caratheodory moment problem that, in effect, also gave a fast algorithm for factorizing Toeplitz matrices; Levinson's algorithm factorizes the *inverse* of a Toeplitz matrix. It turns out that Schur's algorithm offers an alternative to the Levinson algorithm: it is somewhat slower for serial computation, but can be much faster for (software or hardware) parallel implementation!

There are many aspects to these algorithms arising from pursuing the prediction problem. One of the most fascinating is the concept of *displacement structure*. One way of motivating it is by asking questions such as the following:

If there are fast algorithms for factorizing Toeplitz matrices, what about factoring non-Toeplitz matrices that are known to have Toeplitz inverses? Similarly, should it be much harder to factor the non-Toeplitz matrix $T_1 T_2$ or $T_1 T_2^{-1} T_3$ than T_1 (or T_2 or T_3) alone?

The answer is that these problems in fact have the same order of complexity as purely Toeplitz problems do. The reason is that what allows fast algorithms for Toeplitz matrices is not their Toeplitzness, which is lost under inversion and under multiplication, but something called *displacement structure*: R has displacement

structure if $R - FRA$, or more generally $\Omega R\Delta - FRA$ has low rank for appropriate (low complexity) matrices $\{\Omega, \Delta, F, A\}$. The interested reader can verify that when $F = A^* = Z$, the lower shift matrix with ones on the first subdiagonal and zeros elsewhere, a Hermitian Toeplitz matrix and its inverse have displace rank less than or equal to 2. It is not hard to show that products, inverses and Schur complements essentially inherit the displacement structure. This fact can be exploited to obtain a generalized Schur algorithm for the fast recursive factorization of such matrices. Moreover there is a very useful physical structure – a cascade network or generalized transmission line – that can be associated with the generalized Schur algorithm, a fact that has lots of implications and applications. We may mention, among others, problems in linear algebra, inverse scattering, coding theory, complex interpolation, matrix completion, etc. Surveys of these results can be found in Kailath (1987) and Kailath and Sayed (1995).

To end this account, though we should note that the initial stimulus for the development of the concept of displacement structure came not from linear algebra, but from the Wiener–Hopf equation itself, as it was further studied by the astronomers V. A. Ambartsumian (1943) in the former Soviet Union, and S. Chandrasekhar in the USA. It will take too long to make those connections here, and we refer the interested reader to the reviews Kailath (1991), Sayed and Kailath (1995) for some the history and for some of the later developments, including links back to the work of I. Schur.

8. Nonlinear estimation

In the late 1950s, Wiener gave a series of lectures on the problem of nonlinear least-mean squares estimation, which were transcribed into a monograph (Wiener (1958)). Wiener proposed to use a so-called 'Volterra series' characterization of nonlinear systems as a sum of linear + quadratic + ... systems. However this approach had several limitations, especially of computational complexity and the difficulty of approximation (how many or which terms should we keep for a particular nonlinear system?).

The success of the state-space models for the linear problems led to a significant effort to try to obtain similar results for the nonlinear case. Thus suppose we have a nonlinear system, in state-space form,

$$\begin{cases} \dot{x}(t) = f(x(t), t, u(t)), t \geq 0 \\ y(t) = h(x(t), t) + v(t) = s(t) + v(t), \text{ say.} \end{cases}$$

The minimum mean-square estimator of $x(t)$ given $\{y(\tau), \tau < t\}$ is no longer linear, and its computation requires full statistical knowledge of the non-Gaussian processes $x(.)$, $z(.)$ and $y(.)$:

$$\hat{s}(t) = E[s(t)\} \mathscr{F}\{y(\tau), \tau < t\}].$$

When $\{x(.), s(.), y(.)\}$ are jointly Gaussian, one has the Kalman filter recursions. But in general, all has an ascending chain of coupled nonlinear equations for which no really satisfactory practical algorithms, or approximations, have been found. Therefore the nonlinear problem is effectively still open.

However there have been several interesting theoretical results. One set arises from the introduction of ideas from martingale theory (with some of the results now being pursued in finance theory and on Wall Street). Martingale theory first enters through the fact that the white Gaussian measurement noise, $v(\cdot)$, of the engineers is the formal derivative of the special process introduced by Wiener in his study of Brownian motion:

$\int_0^t v(\tau)d\tau = W(t)$, the Wiener (-Lévy) process.

The martingale properties of $W(\cdot)$ lead to a striking generalization of the innovations process first introduced in the linear theory. Let us recall from § 5 that with (scalar) observations containing white noise,

$$y(t) = s(t) + v(t), \langle v(t), v(\tau)\rangle = \delta(t - \tau)$$

the optimum linear filter for finding $\hat{s}(\cdot)$ has transfer function (note that now $R \equiv 1$ in (8))

$$K(\omega) = 1 - \Psi^{-1}(\omega).$$

This implies in the time domain that

$$\hat{s}(t) = y(t) - e(t)$$

or that the innovations can be expressed as

$$e(t) = y(t) - \hat{s}(t).$$

Now when we deal with nonlinear operations on white noise, the formal manipulations become harder to justify: linear operations on white noise give smoother processes, but what is the square of white noise? Therefore one now works with integrated processes,

$$Y(t) \triangleq \int_0^t y(\tau)d\tau = \int_0^t s(\tau)d\tau + \int_0^t v(\tau)d\tau = \int_0^t s(\tau)d\tau + W(t)$$

and uses the Ito theory of stochastic integrals, especially as developed by the Japanese and French schools (see e.g., Meyer (1975)).

In this language, one can show (see Kailath (1971), Meyer (1973)) the following: Let

$$Y(t) = \int_0^t s(\tau)d\tau + W(t),$$

with

$$\int_0^T E|s(t)dt < \infty, E[W(t) - W(\tau)z(\tau) = 0, \quad t > \tau.$$

Then, the process $E(\cdot)$ defined as

$$E(t) = Y(t) - \int_0^t \hat{s}(\tau)d\tau$$

where

$$\hat{s}(t) \triangleq E[s(t)F\{y(\tau),\tau \le t\}]$$

is also a Wiener process w.r.t. the (nested) family of sigma fields $(F\{Y(\tau), \tau \le t\}$. The main idea of the proof is to show first that $E(\cdot)$ is a martingale with respect to these sigma fields, and then to show that $E(\cdot)$ and $W(\cdot)$ have the same 'quadratic variation' (again a concept introduced by Wiener). Then a theorem of Levy's gives the result. This is a nice result, since the process $y(\cdot)$ might be much more complicated than $E(\cdot)$; it shows the power of the assumption of additive white noise. Now in the linear case, results from the theory of integral equations enable us to show that (Kailath (1968, 1972))

$$F\{E(\tau), \tau \le t\} = F\{Y(\tau), \tau \le t\}, 0 \le t \le T$$

so that the process $\{Y(\cdot)\}$ and $\{E(\cdot)\}$ are replaceable each by the other, without any loss of information. As mentioned earlier, this was the idea behind the innovations approach to the Wiener filter (Bode-Shannon (1950), Zadeh-Ragazzini (1950)); in the nonstationary finite-time case, the above result allows for a similar approach to the Kalman filter and several related problems (Kailath (1970), Davis (1977)).

Therefore an important question is under what conditions this equality of sigma fields holds in the general case. The problem turned out to be quite difficult (Benes (1976)) and only after attempts by several researchers, did Allinger and Mitter (1981) succeed in proving the equality for the case where $s(\cdot)$ and $W(\cdot)$ are independent of each other and $\int_0^T E|s(t)|^2 dt < \infty$.

However even without the equivalence, the process $E(\cdot)$ leads to several nice results. One is that even though the sigma fields generated by $E(\cdot)$ and $Y(\cdot)$ may not be equivalent, Fujisaki *et al.* (1972) showed that any function measurable w.r.t. the Y sigma fields can be written as a stochastic integral w.r.t. the Wiener process $R(\cdot)$. This then allows for a simpler description of the nonlinear filtering equations: however as mentioned before, they are not useful for actual computation.

Another application that exploits only the fact that $E(\cdot)$ is a Wiener process is a generalized Cameron-Martin formula for the Radon-Nikodym derivative of the measures P_y and P_w induced by the processes $Y(\cdot)$ and $W(\cdot)$:

$$\frac{dP_y}{dP_w} = \exp \int_0^T \hat{s}(t) dY(t) - \frac{1}{2}\int_0^T |\hat{s}(t)|^2 dt.$$

This expression has useful implications for the problem of detecting the presence or absence of a random signal $s(\cdot)$ in the presence of noise. When the signal $s(\cdot)$ is deterministic (and therefore known *a priori*) $\hat{s}(\cdot) \equiv s(\cdot)$, this is a result of Cameron and Martin (1944). It is an interesting and useful fact that for random $s(\cdot)$ the deterministic formula still applies with the unavailable random signal $s(\cdot)$ being replaced by the observable least-squares estimate $\hat{s}(\cdot)$. This allows a lot of the insights and results of estimation theory to be carried over to signal detection theory (Kailath (1969), Davis and Andreadakis (1977)). We may remark that the Cameron-Martin formula arose as a theory of 'linear changes of variables' in Wiener space (the space of sample functions of a Wiener process). The generalized Cameron-Martin formula follows from a nonlinear version of this theory introduced in a seminal paper of Girsanov (1960), which has since been much studied and extended.

9. Concluding remarks

This has been an account of some of the ways in which Nobert Wiener's work and ideas have influenced several engineering developments. The key ideas were his emphasis of the statistical nature of the communication process and his introductions of the use of optimization criteria into system design. I should hasten to add that many other notable researchers (Shannon, Rice, Tukey, Bellman, Pontryagin, Kalman, to name just a few) had major roles in the post-1942 story. Finally, Wiener's own specific mathematical contributions to mathematical engineering are too numerous to cover in a simple article. Here I have described, in a very sketchy way and with some focus on things I know best, some of the wide range of ideas and techniques stimulated by Wiener's work on prediction and filtering.

1. Allinger, D. F. and Mitter, S. K., New Results on the Innovations Problem for Nonlinear Filtering, *Stochastics*, 1981, **4**, 339–348.
2. Ambartsumian, V. A., Diffuse Reflection of Light by a Foggy Medium, *Dokl. Akad. Sci. SSSR*, 1943, **38**, 229–322.
3. Benes, V. E., On Kailath's innovations conjecture, *Bell Syst. Tech J.*, 1976, **55**, 981–1001.
4. Bode, H. W. and Shannon, C. E., A Simplified Derivation of Linear Least Squares Smoothing and Prediction Theory, *Proc. IRE*, 1950, **38**, 417–425.
5. Cameron, R. H. and Martin, W. T., Transformation of Wiener Intergrals under Translations, *Ann. Math.*, 1944, **45**, 368–396.
6. Castl, J. L., Kalaba, R. E. and Murthy, V. K., A New Initial-Value Method for On-Line Filtering and Estimation, *IEEE Trans. Inform. Theory*, 1972, **IT-18**, 515–518.
7. Chandrasekhar, S., On the Radiative Equilibrium of a Stellar Atmosphere, Pt. XXL, *Astrophys. J.*, 1947, **106**, 152–216; Pt. XXII, *Astrophys. J.*, 1948, **107**, 48–72.
8. Chandrasekhar, S., *Radiative Transfer*, Oxford University Press, New York, 1950; Russian transl., IL, Moscow, 1953.
9. Davis, M. H. A., *Linear Estimation and Stochastic Control*, Halsted Press, New York, 1977.
10. Davis, M. H. A. and Andreadakis, E., Exact and Approximate Filtering in Signal Detection: An Example, *IEEE Trans. Inform. Theory*, 1977, **IT-23**, 768–772.
11. Einstein, A., Méthode pour la détermination de valeurs statistiques d'observations concernant dee grandeurs soumises a des fluctuations irrégulières, *Archives des sciences Physiques et Naturelles*, 1914, **37**, 254–256 (A translation and commentary in IEEE Magazine on Acoustics, Speech and Signal Processing, October 1987).

12. Fujisaki, M. and Kallianpur, G. and Kunita, H., Stochastic Differential Equations for the Nonlinear Filtering Problem, *Osaka J. Math.*, 1972, **9**, 19–40.
13. Girsanov, I. V., On Transforming a Certain Class of Stochastic Processes by Absolutely Continuous Changes of Measure, *Th. Prob. App.*, 1960, **5**, 285–301.
14. Hassibi, B., Sayed, A. and Kailath, T., Linear Estimation in Krein Spaces: Parts I and II, *IEEE Trans. Autom. Contr.*, 1996, **AC-41**, 18–49.
15. Kailath, T., An Innovations Approach to Least-Squares Estimation, Part I: Linear Filtering in Additive White Noise, *IEEE Trans Automatic Control*, 1968, **AC-13**, 646–655.
16. Kailath, T., A General Likelihood Ratio Formula for Random Signals in Noise, *IEEE Trans. Inform Theory*, 1969, **IT-15**, no. 3, 350–361.
17. Kailath T., The Innovations Approach to the Detection and Estimation Theory, *Proc. IEEE*, 1970, **58**, 680–695.
18. Kailath, T., Likelihood Ratios for Gaussian Processes, *IEEE Trans. Inform. Theory*, 1970, **IT-16**, no. 3, 276–288.
19. Kailath, T., Some Extensions of the Innovations Theorem, *Bell Syst. Tech. J.*, 1971, **50**, 1487–1494.
20. Kailath, T., A Note on Least-Squares Estimation by the Innovations Method, *J. SIAM Control.*, 1972, **10**, 477–486.
21. Kailath, T., Some New Algorithms for Recursive Estimation in Constant Linear Systems, *IEEE Trans. Inform. Theory*, 1973, **IT-19**, 750–760.
22. Kailath, T., A View of Three Decades of Linear Filtering Theory, *IEEE Trans. Inform. Theory*, 1974, **IT-20**, no. 20, 145–181.
23. Kailath, T. (ed.), *Linear Least-Squares Estimation*, Benchmark Papers in Electrical Engineering, Academic Press, New York, 1977, vol. 17.
24. Kailath, T., *Linear Systems*, Prentice-Hall, New Jersey, 1980.
25. Kailath, T., Ljung, L. and Morf, M., Generalized Krein-Levinson Equations for Efficient Calculation of Fredholm Resolvents of Nondisplacement Kernels, in *Topics in Functional Analysis* (eds Gohberg, I. C. and Kac, M.), Academic Press, New York, 1978, pp. 169–184.
26. Kailath, T., Vieira, A. and Morf, M., Inverses of Toeplitz Operators, Innovations and Orthogonal Polynomials, *SIAM Rev.*, 1978, **20**, 106–119.
27. Kailath, T., Remarks on the Origins of the Displacement Rank Concept, *Appl. Math. Comp.*, 1991, **45**, 193–206.
28. Kailath, T. and Sayed, A., Displacement Structure: Theory and Applications, *SIAM Rev.*, 1995, **37**, 297–386.
29. Kalman, R. E., A New Approach to Linear Filtering and Prediction Problems, *J. Basic Eng.*, 1960, **82**, 34–45.
30. Kalman, R. E., New Methods of Wiener Filtering Theory, in Proc. 1st Symp. Engineering Applications of Random Function Theory and Probability (eds Bogdanoff, J. L. and Kozin, F.), Wiley, New York, 1963, 270–388.
31. Kalman, R. E. and Bucy, R. S., New Results in Linear Filtering and Prediction Theory, *Trans. ASME, Ser. D., J. Basic Eng.*, 1961, **83**, 95–107.
32. Kolmogorov, A. N., Sur l'interpolation et extrapolation des suites stationnaires, *C. R. Acad. Sci.*, 1939, **208**, 2043–2045.
33. Kolmogorov, A. N., *Stationary Sequences in Hilbert Spaces* (in Russian), Bulletin Moscow State Univ., 1941, vol. 2, pp. 1–40. (English translation in Linear Least-Squares Estimation (1977), edited by T. Kailath, Academic Press, NY.)
34. Krein, M. G., Integral Equations on a Half-Axis with Kernel Depending on the Difference of the Arguments, *USP. Math. Nauk.*, 1958, **13**, 3–120; *Am. Math. Soc.* Transl.
35. Levinson, N., The Wiener RMS (Root-Mean-Square) Error Criterion in Filter Design and Prediction, *J. Math. Phys.*, 1947, **25**, 261–278.
36. Ljung, L. and Kailath, T., A Scattering Theory Framework for Fast Least-Squares Algorithms, in *Multivariate Analysis IV* (ed. Krishnaiah, P. R.), North-Holland, Amsterdam, 1977. [Presented at Symposium, June 1975].
37. Ljung, L., Kailath, T. and Friedlander, B., Scattering Theory and Linear Least-Squares Estimation, Part I – Continuous-Item Problems, *Proc. IEEE*, 1976, **64**, 131–139
38. Meyer, P. A., Sur un Probléme de Filtration, Séminaire de Probabilités, Pt. VII, *Lecture Notes in Mathematics*, 1973, **321**, Springer, New York, 223–247.
39. Meyer, P. A., Un cours sur les intégrales stochastiques, Seminaire de Probabilités, X, *Lecture Notes in Math.*, Springer-Verlag, N.Y., 1975, vol. 511, pp. 245–400.
40. Morf, M. and Kailath, T., Square-Root Algorithms for Linear Least-Squares Estimation and Control, *IEEE Trans. Auto. Cont.*, 1975, **AC-20**, 487–497.
41. Pincus, J., Commentary on Über eine klasse integralgleichungen, *Norbert Wiener, Collected Works*, (ed. Masani, P.), MIT Press, 1981, vol. III, pp. 44–53.
42. Redheffer, R. M., On the Relation of Transmission-Line Theory to Scattering and Transfer, *J. Math. Phys.*, 1962, **41**, 1–41.
43. Robinson, E. A., *Multichannel Time-Series Analysis with Digital Computer Programs*, Holden-Day, San Francisco, CA, 1967.
44. Shannon, C. E, *Bell Syst. Tech. J.*, 1948, **27**, 379–423, 623–656.
45. Trentelman, H. L. and Willems, J. C., The Dissipation Inequality and the Algebraic Riccati Equation, in *The Riccati Equation* (eds Bittanti, S., Laub, A. J. and Willems, J. C.), Springer-Verlag, 1991, pp. 197–242.
46. Yovits, M. C. and Jackson, J. L., Linear Filter Optimization with Game Theory Considerations, in *IRE Nat. Conv. Rec.*, 1955, pt. 4, 193–199.
47. Wiener, N., *The Fourier Integral and Certain of Its Applications*, University Press, Cambridge, 1933.
48. Wiener, N., *Extrapolation, Interpolation and Smoothing of Stationary Time Series, with Engineering Applications*, Technology Press, New York and Wiley, 1949 (Originally issued in February 1942, as a classified Nat. Defense Res. Council Rep.).
49. Wiener, N. and Hopf, E., Über eine Klasse Singulärer integralgleichungen, S.-B., Preuss, *Akad. Wiss. Berlin, Phys.-Math.*, Kl. 30/32, 1931, 696–706.
50. Zadeh, L. A. and Ragazzini, J. R., An Extension of Wiener's Theory of Prediction, *J. Appl. Phys.*, 1950, **21**, 645–655.
51. Zadeh, L. A., From Circuit Theory to System Theory, *Proc. IEEE*, 1962, **50**, 856–865.