Center for Computer Research in Music and Acoustics June 1983

Department of Music
Report No. STAN-M-14

TECHNIQUES FOR DIGITAL FILTER DESIGN AND SYSTEM IDENTIFICATION WITH APPLICATION TO THE VIOLIN

by

Julius Orion Smith III

The computational modeling of natural signals has been a very active area of research since the emergence of digital technology. One class of signal models consists of a linear filter controlled by signals much simpler than the one being modeled. In this thesis, such an approach is taken to the modeling of the violin. In the course of applying the modern repertoire of techniques, various problems had to be surmounted to achieve a musically useful violin model at a computational expense easily affordable with present technology. As a first contribution, it was found that separate models are desirable for the violin body, string, and bow-string interaction. Each of these components presents a challenging problem in signal modeling.

The model for the violin body requires the solution of a very demanding digital filter design problem. Given that one wishes to keep complexity down, how can the *most important* features of the violin-body frequency-response be captured in a minimum-cost digital filter? Several contributions to the design of audio digital filters were developed to this end. For example, conformal mapping techniques are utilized to "stretch" the important main air and wood resonances of the violin body over a larger area of the frequency domain, so that they are easy to fit with a filter design algorithm. Smoothing according to the critical bands of hearing is used to prevent the model from trying to follow less important fine detail in the frequency response. The final model for the violin body contains only eight "poles" and eight "zeros," but its frequency response (fitted to physical measurements) contains resonances at the main air and wood resonances, and at the so-called "singing formant." In addition, ancillary contributions to digital filter design were developed involving the choice of error minimized. Example areas include log-magnitude approximation and joint phase and magnitude optimization for rational filters using the Hankel norm of the frequency-response error.

The vibrating string is in some ways more challenging to model then the violin body. One source of difficulty is that the string "remembers" vibrational energy much longer than does the violin body; also, the many "harmonics" of a vibrating string can each be considered an important resonance of the linear filter which models the string. At the complexity of the violin body, only four harmonics could be sustained. Nevertheless, by constraining the string filter to an efficient recursive form, it is possible to obtain hundreds of resonances with even less complexity than that of the violin body model. Furthermore, methods for fitting such a model to recorded measurements of a vibrating string are presented. A method for simulating the interaction of a bow with the string model was developed based on the work of McIntyre and Woodhouse.

This thesis was submitted to the Department of Electrical Engineering and the Committee on Graduate Studies of Stanford University in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

This research was supported by the Hertz Foundation and the Hertz Foundation and System Development Foundation under Grant SDF #345. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of Stanford University, any agency of the U. S. Government, or of sponsoring foundations.

Acknowledgements

I wish to thank my advisors Prof.'s Gene Franklin, Martin Morf, and James Angell for their help and encouragement during the course of this research. Also, thanks to Prof. Kailath for providing a very enriching environment at ISL, by getting the students together and by bringing others to Stanford from all over the world. Inestimable thanks are due also to Prof. John Chowning who allowed me free run of the facilities at the Stanford Center for Computer Research in Music and Acoustics (CCRMA). Prof. Chowning's emphasis on interdisciplinary research has significantly influenced this work. I also wish to thank Dr.'s Burrus and Pearson of Rice University for their technical influence while I was an undergraduate there.

I was very fortunate to have the opportunity to collaborate on occasion with Ben Friedlander, Max Mathews, Jont Allen, Lennart Ljung, Earl Schubert, Norman Pickering, and Narendra Gupta — my most respectful thanks to them.

For guidance and aid in the areas of signal processing, numerical analysis, system design, machine architecture, and just about any topic I can think of, a hearty thanks to Andy Moorer.

Over the years I have had many fruitful interactions with the students at ISL and CCRMA. Although I cannot begin to mention them all, I would like to say thanks to David Jaffe, Ken Shoemake, Bill Schottstaedt, John Gordon, Tovar, Chris Chafe, John Strawn, Andy Schloss, Bob Shannon, Richard Gooch, Paul Titchener, Arye Nehorai, Hanoch Levari, Boaz Porat, Martin Vitterli, David Shaw, Dick Gabriel, and Ron Goldman. Also, I wish to express thanks to the Stanford staff members who helped me through the "twisty little passages" of academic life; in particular, Rachel Levy at ISL was extremely helpful and always cheerful.

In my opinion, one of the most promising topics in this thesis is Hankel-norm approximation. If not for Lloyd Trefethen and Martin Gutknecht, I might have passed over it entirely.

The extraordinarily elegant type-setting in this work is due to Prof. Don Knuth, author of the TEX program. Also, I appreciate the TEX macros provided by Dick Gabriel, lead guitarist of *The Wizards*.

One thing that can be said about this thesis is that it is *long* and covers a fairly wide scope. This is due primarily to the fact that for about three years I had no other responsibilities but my own research. For this rare and precious opportunity, I thank the Hertz Foundation.

Table of Contents

Techniques for Digital Filter Design and System Identification, with Application to the Violin

By
Julius Orion Smith III

Information Systems Laboratory
Department of Electrical Engineering
Stanford University, Stanford, California, 94305

This thesis is about signal modeling. The signal which receives primary attention is the sound of the violin. The model consists of a simple algorithm which can be easily implemented in a computer or integrated digital circuit. The model produces a digital signal (a series of numbers representing loudspeaker position, for example) from a set of natural controls (such as bowing specifications). The modeling techniques, however, are not limited to the violin. In other applications, the signal might represent speech, a radio broadcast, seismic disturbance, light propagation, thermal gradient, structural stress, aircraft motion, stock market level, weather patterns, or the sensation in a robot's hand.

The model for the violin is based on underlying physics, with certain simplifications. The simplifications achieve two objectives. First, we desire a model which is computationally efficient. This rules out, for example, the numerical integration of a large system of differential equations. Second, the model is to be judged solely on the quality of the sound it produces. From this point of view, many aspects of the physical violin become unimportant, and we wish to exclude them from the model whenever possible.

It was discovered that there are three important components of the violin which deserve individual consideration. (1) The *violin body* contributes to the timbre of the sound by shaping the spectrum in a fixed way. This function is analogous to the effect of the vocal tract in speech production. In a sense, it determines the "vowel" which is spoken by the instrument in each register. However, in the case of the violin, the "mouth" is always of a fixed shape. The body of the violin is accordingly modeled as a linear time-invariant filter. For this step, considerable expertise is required in digital filter design. Chapter 1 is devoted to this aspect of the problem, and several new techniques in filter design are presented. (2) The string of the violin has a "life of its own" which does not absorb gracefully into the chosen model for the body. This is because the vibrating portion of the string changes

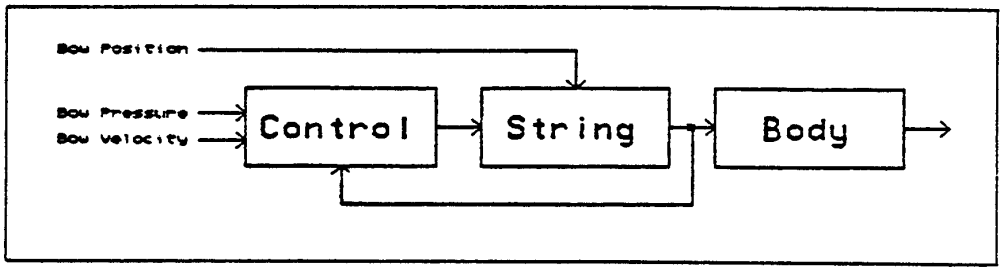

Figure 1. A schematic for the violin.

length during normal play, making it time-varying, and because the string has much greater "memory" than the body of the instrument. The string is still well-modeled as a linear filter, but standard filter types are far too expensive to consider in this context. Consequently, a special linear filter structure for the string is developed which maintains low complexity while providing time-varying filters of very high order. Methods from Chapter 2 are used to calibrate this model to recorded violin data, and the string simulator itself is derived in Chapter 3 from basic physics. (3) The interaction of the string with the violin bow was found to be a very important determinant of the sound of the violin, and a good model of bowed-string behavior is essential. The bow-string interaction is nonlinear, and therefore cannot be incorporated into the body or string models. Fortunately, since the string model corresponds well to underlying physics, it is straightforward to add a bowing mechanism which also mirrors the physics of bowed strings. The three basic elements of the violin model, and their interconnection, are shown in Fig. 1.

It should be emphasized that this thesis concentrates on signal modeling *techniques* rather than the violin per se. The violin problem provides motivation, focus, and a means for measuring the benefit of one technique over another. While the violin model stands alone as a practical contribution, the general techniques developed for its calibration to recorded data are the main topics of research.

Itinerary of Topics

The flow of ideas in the dissertation is as follows. First, the rational filter design problem is posed in a general yet wholly practical setting. The problem is examined first from an "approximation theory" point of view, in which existence and uniqueness of the solution are investigated. The approximation problem proves to be nonlinear and difficult to solve, and a theorem is given indicating how difficult the problem can be in the general case. Having laid to rest the possibility of a practical general solution, the problem is modified in various ways to make its solution tractable. Some modifications

are designed to retain properties of the original formulation as much as possible, while others are based on dropping aspects of the original formulation which are not important for certain applications. Still others stem almost entirely from the desire for a simple robust solution, and the nature of fit with these methods is discussed. Next, the problem is reformulated in the time domain, and generalizations made possible by this are introduced. Finally we come to the violin. After discussing aspects of hearing which lead to a novel adaptation of the filter-design error criterion, a selection of methods is used to obtain a digital filter which simulates the body of the violin. Then a parsimonious model for the vibrating string is proposed, and its parameters are estimated from recorded data. As a postscript, a method for "bowing" the artificial violin is described, and several directions for improvement are suggested.

Summary

The two basic problems addressed are

- Obtaining a digital filter with a prescribed frequency response.
- Parsimonious modeling of nearly periodic signals.

The filters considered can be represented by a rational transfer function of the form

$$H(z) = \frac{B(z)}{A(z)} = \frac{b_0 + b_1 z^{-1} + \cdots + b_{n_b} z^{-n_b}}{1 + a_1 z^{-1} + \cdots + a_{n_a} z^{-n_a}},$$

having n_b *zeros* and n_a *poles*. The quality of fit is considered primarily in the frequency domain, but some methods minimize time-domain errors. Methods are compared on the basis of the error minimized and computational robustness. The design of a model for the body of the violin is considered as a practical example.

The class of time-series to be modeled consists of a quasi-periodic deterministic signal plus noise. In this case, the deterministic part of the model transfer function looks like

$$H(z) = \frac{B(z)}{1 + z^{-P} A(z)},$$

where P is the "period" of the signal. The periodic structure of the signal is "factored out" leaving the remaining information to be captured by the model coefficients. The bowed string is taken as an example to which this model is applied.

In Chapter 1, the filter-design problem is discussed. The desired frequency response is assumed to be continuous and causal. The initial formulation produces filters which match both the magnitude and phase of the desired frequency response. Since this is relatively

difficult for rational filters, a variety of compromise methods is included also. Finally, various new ancillary methods for filter design are presented.

In Chapter 2, a more general formulation is treated wherein a filter is fit to given input/output data. This is the "system identification" paradigm. The identification algorithms are better suited to time-varying filter design than the methods of Chapter 1, and they are appropriate when given input/output data. They also provide an error signal which can be further modeled, and here statistical time-domain models begin to play an important role. The added flexibility of the time-domain formulation is accompanied, however, by greatly reduced flexibility in the choice of optimality criterion. For this reason, the maximum amount of time-invariant structure is first extracted using Chapter 1 methods when attempting to model the violin.

Chapter 3 is a case study in modeling the violin. The problem is divided at the "bridge" into two parts: the resonating body and the bowed string. The *body* is regarded as a linear time-invariant filter whose transfer function relates a force input to radiated sound pressure at a point in space. Techniques from Chapter 1 are tried and compared for this application. The measure of fit is designed to reflect the characteristics of human hearing. The *bowed string* is modeled as a special linear time-varying filter producing a force output in response to a simulated string excitation. The model developed for the string achieves very high order with only a few degrees of freedom, and it is sufficiently general to provide dispersive, frequency-dependent losses due to string stiffness and yielding terminations. Techniques from Chapter 2 are used to determine the parameters of this model from recorded data. Finally, refinements to the basic model are discussed, including a mechanism for simulating the interaction of a bow with the string model.

The appendices contain technical reference material, such as long proofs, and Appendix E gives some fundamental background including definitions of many terms not defined in the main text.

Chapter 1

Methods for Digital Filter Design

1.1. Introduction

The problem of fitting a digital filter to a given spectrum may be formulated as follows:

1.1.1. Problem $\hat{H}^*$

Given a continuous complex function $H(e^{j\omega})$, $-\pi < \omega \leq \pi$, corresponding to a causal* desired frequency-response,† find a stable digital filter,† of the form

$$\hat{H}(z) \triangleq \frac{\hat{B}(z)}{\hat{A}(z)}, \tag{1.1}$$

where

$$\begin{aligned} \hat{B}(z) &\triangleq \hat{b}_0 + \hat{b}_1 z^{-1} + \cdots + \hat{b}_{\hat{n}_b} z^{-\hat{n}_b} \\ \hat{A}(z) &\triangleq 1 + \hat{a}_1 z^{-1} + \cdots + \hat{a}_{\hat{n}_a} z^{-\hat{n}_a}, \end{aligned} \tag{1.2}$$

with $\hat{n}_b, \hat{n}_a$ given, such that some norm† of the error

$$J(\hat{\theta}) \triangleq \left\| H(e^{j\omega}) - \hat{H}(e^{j\omega}) \right\|$$

is minimum with respect to the filter coefficients

$$\hat{\theta}^T \triangleq \left(\hat{b}_0, \hat{b}_1, \ldots, \hat{b}_{\hat{n}_b}, \hat{a}_1, \hat{a}_2, \ldots, \hat{a}_{\hat{n}_a} \right),$$

which are constrained to lie in a subset $\hat{\Theta} \subseteq \Re^{\hat{N}}$, where $\hat{N} \triangleq \hat{n}_a + \hat{n}_b + 1$. When explicitly stated, the filter coefficients may be complex, in which case $\hat{\Theta} \subseteq C^{\hat{N}}$.

---

* $H(e^{j\omega})$ is said to be *causal* if $h(n) \triangleq \int_{-\pi}^{\pi} H(e^{j\omega}) e^{j\omega n} \frac{d\omega}{2\pi} = 0$ for $n < 0$.

† Defined in Appendix E.

The approximate filter $\hat{H}$ will be *constrained* to be stable, and since positive powers of z do not appear in $\hat{B}(z)$, stability implies causality. Consequently, the impulse response of the model $\hat{h}(n)$ is zero for $n < 0$. If H were noncausal, all impulse-response components $h(n)$ for $n < 0$ would be approximated by zero.

The stability restriction on $\hat{H}(z)$ needs to be made more precise:

Definition 1.1. The circle of radius $1 - \delta$ in the complex plane is denoted by

$$\Gamma_\delta \triangleq \{z \in C \,|\, |z| = 1 - \delta\}, \qquad 0 \leq \delta < 1.$$

Definition 1.2. The disk of radius $1 - \delta$ in the complex plane is denoted by

$$D_\delta \triangleq \{z \in C \,|\, |z| \leq 1 - \delta\}, \qquad 0 \leq \delta < 1.$$

The unit circle Γ_0 and unit disk D_0 are denoted more simply as Γ and D, respectively.

For stability of the filter $\hat{H}$, it is necessary to restrict the set of denominator coefficients $\{\hat{a}_1, \ldots, \hat{a}_{\hat{n}_a}\}$ so that the roots of $\hat{A}(z)$ lie in D. The filter is said to be *marginally stable* when at least one root of $\hat{A}(z)$ lies on Γ and all roots are in D. When the filter is to be used with a continuously supplied input, the roots of $\hat{A}(z)$ must lie strictly inside the unit circle Γ for stability, with Γ excluded.

Definition 1.3. A rational filter of the form $\hat{H}(z) = \hat{B}(z)/\hat{A}(z)$ is said to be *strictly stable* of order δ if the roots of $\hat{A}(z)$ are confined to D_δ for some $0 < \delta < 1$.

Strictly stable filters are almost always desired in practice. However, for some applications, such as spectrum analysis applied to sinusoids in white noise, the poles of a good model can be very close to the unit circle, and one is led naturally to choosing the open unit disk $|z| < 1$ as the allowed domain for the poles. Even in these cases, however, the use of finite-time observations (or finite energy signals) justifies a strict stability constraint. The assumption that $H(e^{j\omega})$ is continuous on the unit circle is itself form of strict stability—for it implies $|H(e^{j\omega})|$ is bounded. Without strict stability, much extra work is required in the theory, and numerical problems are more likely in practice. Since there seems to be little motivation to the contrary, *strict stability will be assumed* in the sequel unless stated otherwise. The stability margin δ is an arbitrary *fixed* constant which will not be explicitly mentioned in all cases.

Problem $\hat{H}^*$ is then to find a (strictly) stable $\hat{n}_a$-pole, $\hat{n}_b$-zero digital filter which minimizes some norm of the error in the frequency-response. This is fundamentally *rational approximation of a complex function of a real (frequency) variable, with constraints on the poles.*

1.1.2. Summary of Chapter 1

While the filter-design problem has been formulated quite naturally, it is difficult to solve in practice. The strict stability assumption yields a compact space of filter coefficients $\hat{\Theta}$, leading to the conclusion that a best approximation $\hat{H}^*$ *exists* over this domain. Unfortunately, the norm of the error $J(\hat{\theta})$ typically is not a *concave*† function of the filter coefficients on $\hat{\Theta}$. This means that algorithms based on gradient descent may fail to find an optimum filter due to their premature termination at a suboptimal local minimum of $J(\hat{\theta})$. It is shown that this is a very serious difficulty which makes many (if not most) of the currently available methods unreliable in general.

Fortunately, there is at least one norm whose global minimization may be accomplished in a straightforward fashion without need for initial guesses or ad hoc modifications of problem $\hat{H}^*$—the *Hankel norm.*† The CF method for digital filter design, described in this chapter, is based on the Hankel norm. It does not suffer from non-concavity of the error surface, and the approximation it finds is spontaneously *stable* without imposing coefficient constraints in the algorithm. It appears that methods based on Hankel-norm approximation are the *only* methods which can solve problem $\hat{H}^*$ to within an arbitrary tolerance without requiring exhaustive searching over the filter coefficient space $\hat{\Theta}$.

An alternative to using Hankel-norm approximation is to reformulate problem $\hat{H}^*$ so that it can be solved by *linear* or *concave* techniques. Methods along these lines abandon the use of a *norm* applied to the frequency-response error $H - \hat{H}$ as a quantity to be minimized. Examples include

- *Pseudo-norm*† minimization: (Pseudo-norms can be zero for nonzero functions.) For example, Padé approximation falls in this category. In Padé approximation, the first $\hat{n}_a + \hat{n}_b + 1$ samples of the impulse-response $h(n)$ of H are matched exactly, and the error in the remaining impulse-response samples is ignored.
- *Ratio Error*: Minimize $\| H(e^{j\omega})/\hat{H}(e^{j\omega}) \|$ subject to $\hat{B}(z) = 1$. Minimizing the L^2 norm of the ratio error yields the class of methods known as *linear prediction* techniques. Since $\| e^{j\theta} E(e^{j\omega}) \| = \| E(e^{j\omega}) \|$, by the definition of a norm, it follows that $\| H/\hat{H} \| = \| |H|/|\hat{H}| \|$; thus *ratio error methods ignore the phase of the approximation.* It is also evident that they tend to make $|\hat{H}(e^{j\omega})|$ larger than $|H(e^{j\omega})|$.* For this reason, ratio-error methods are considered most appropriate

† Defined in Appendix E.

* $|\hat{H}|$ cannot go to infinity since the constraint $\hat{B}(z) = 1$ and the stability constraint imply that $\ln|\hat{H}(e^{j\omega})|$ is zero-mean by the argument principle [186,150].

for modeling the *spectral envelope* of $|H(e^{j\omega})|$. It is well known that these methods are fast and exceedingly robust in practice, and this explains in part why they are used almost exclusively for some data-intensive applications such as speech modeling and "modern spectrum analysis." In some applications, such as adaptive control or forecasting, the fact that linear prediction error is minimized can justify their choice.

- *Equation error*: Minimize $\| \hat{A}(e^{j\omega})H(e^{j\omega}) - \hat{B}(e^{j\omega}) \| = \| \hat{A}(e^{j\omega})(H(e^{j\omega}) - \hat{H}(e^{j\omega})) \|$. When the L^2 norm of equation-error is minimized, the problem becomes solving a set of $\hat{N} = \hat{n}_a + \hat{n}_b + 1$ linear equations. Equation error can be viewed as a frequency-response error which has been weighted by $|\hat{A}(e^{j\omega})|$; thus large errors can be tolerated where the poles of the optimum approximation approach the unit circle. While this makes the frequency-domain formulation seem ill-posed, in the time-domain, *linear prediction error* is minimized in the L^2 sense, and in certain applications this is ideal. Equation-error methods can be viewed as generalizing ratio-error methods to include zeros.

- Conversion to *real-valued approximation*: For example, *power spectrum matching*, i.e., minimization of $\| |H(e^{j\omega})|^2 - |\hat{H}(e^{j\omega})|^2 \|$, is possible using the Chebyshev or L^∞ norm.† Similarly, *linear-phase* filter design can be carried out with some guarantees, since again the problem reduces to real-valued approximation on the unit circle. The essence of these methods is that the *phase-response* is eliminated from the error measure, as in the norm of the ratio error, in order to convert a complex approximation problem into a real one. Real rational approximation of a continuous curve seems to be solved in principle only under the L^∞ norm.

- *Decoupling poles and zeros*: An effective example of this approach is Kopec's method which consists of using ratio error to find the poles, computing the error spectrum $E = H/\hat{H}$, inverting it, and fitting poles again (to $1/E(e^{j\omega})$). There is a wide variety of methods which first fit poles and then zeros. None of these methods produce optimum filters, however, in any normal sense.

In addition to these modifications, sometimes it is necessary to reformulate the problem in order to achieve a different goal. For example, in some audio applications, it is desirable to minimize the *log-magnitude* frequency-response error. This is due to the way we hear spectral distortions in many circumstances. A technique which accomplishes this objective to the first order in the L^∞ norm is presented in this chapter.

Sometimes the most important spectral structure is confined to an interval of the frequency domain. A question arises as to how this structure can be accurately modeled

† Defined in Appendix E.

while obtaining a cruder fit elsewhere. A technique based on *conformal mapping* is presented in this chapter. It is especially valuable in connection with methods which are intrinsically unable to minimize a *weighted* norm of the frequency-response error, such as the Hankel-norm method.

There are several methods which produce $\hat{H}(z)\hat{H}(z^{-1})$ instead of $\hat{H}(z)$ directly. A *fast spectral factorization* technique is presented which is useful in conjunction with methods of this category. Roughly speaking, a size $2\hat{n}_a$ polynomial factorization is replaced by an FFT and a size $\hat{n}_a$ system of linear equations.

In the next section, some basic results on the general formulation of problem $\hat{H}^*$ are developed. Next, the possibility of solving the problem under specific norms is considered, and associated methods are discussed. Finally, the auxiliary techniques outlined above are presented.

1.2. Possibility of Solution to Problem $\hat{H}^*$

In this section, the basic issues of existence and uniqueness of a solution to problem $\hat{H}^*$ are examined. It is shown that a solution always exists, but that uniqueness can be guaranteed only in special cases.

Definition 1.4. The space of complex-valued functions continuous on the unit circle Γ in the complex plane, is denoted $C_0(\Gamma)$. The causal subspace of $C_0(\Gamma)$ is denoted $C_0^+(\Gamma)$, and the anti-causal subspace is denoted $C_0^-(\Gamma)$. That is, if $H(e^{j\omega}) \in C_0(\Gamma)$, then

$$H(e^{j\omega}) = H^+(e^{j\omega}) + H^-(e^{j\omega}),$$

where

$$H^+(e^{j\omega}) = \sum_{n=0}^{\infty} h(n)e^{-j\omega n} \quad \in C_0^+(\Gamma)$$

$$H^-(e^{j\omega}) = \sum_{n=-\infty}^{-1} h(n)e^{-j\omega n} \quad \in C_0^-(\Gamma).$$

In problem $\hat{H}^*$, the desired frequency-response $H(e^{j\omega})$ is assumed to lie in $C_0^+(\Gamma)$ which is regarded as a normed linear space,† with the scalars being complex numbers.

Definition 1.5. $\mathcal{H}_{\hat{n}_b,\hat{n}_a}$ denotes the set of all rational functions of the form $\hat{H}(z) = \hat{B}(z)/\hat{A}(z)$ as in (1.2) of problem $\hat{H}^*$, where the roots of $A(z)$ lie inside the closed disk $\mathcal{D}_\delta$, $0 < \delta < 1$.

† Defined in Appendix E.

Proposition 1.6. $\mathcal{H}_{\hat{n}_b,\hat{n}_a}$ is *dense* in $C_0^+(\Gamma)$.

Proof. The Weierstrass approximation theorem states that polynomials in z are uniformly dense in $C_0(\Gamma)$ [57]. This theorem can be generalized to show that for each $\hat{n}_a \geq 0$, the best approximation from $\mathcal{H}_{\hat{n}_b,\hat{n}_a}$ uniformly approaches an arbitrary element of $C_0^+(\Gamma)$ on the unit circle as $\hat{n}_b \to \infty$ (Walsh [75]).

Note that $\mathcal{H}_{\hat{n}_b,\hat{n}_a}$ does not form a closed subspace of continuous functions on the unit circle, since if $\hat{H}_1(z), \hat{H}_2(z) \in \mathcal{H}_{\hat{n}_b,\hat{n}_a}$, then

$$\hat{H}_1(z) + \hat{H}_2(z) = \frac{\hat{A}_2(z)\hat{B}_1(z) + \hat{A}_1(z)\hat{B}_2(z)}{\hat{A}_1(z)\hat{A}_2(z)} \notin \mathcal{H}_{\hat{n}_b,\hat{n}_a},$$

in general. However, it is possible to reparametrize $\mathcal{H}_{\hat{n}_b,\hat{n}_a}$, for $\hat{n}_b \geq \hat{n}_a - 1$, by means of the *partial fraction expansion* such that for each set of *fixed* poles in D_δ, a subspace of $C_0^+(\Gamma)$ is generated by using the pole residues as basis-vector coefficients. For the case $\hat{n}_b = \hat{n}_a - 1$, such a subspace will be denoted $\mathcal{H}_{\hat{n}_a}^\pi$:

Definition 1.7. $\mathcal{H}_{\hat{n}_a}^\pi$ denotes the set of all rational functions from $\mathcal{H}_{\hat{n}_a-1,\hat{n}_a}$ each having the fixed set of poles $\{p_i\}_1^{\hat{n}_a}$.

$$\mathcal{H}_{\hat{n}_a}^\pi \triangleq \left\{ \hat{H} \in \mathcal{H}_{\hat{n}_a-1,\hat{n}_a} \,\middle|\, \hat{H}(z) = \sum_{k=1}^{\hat{n}_a} \frac{R_k}{1 - p_k z^{-1}} \right\}.$$

1.2.1. Existence

Lemma 1.8. For any set of *fixed* complex numbers $\{p_i\}_1^{\hat{n}_a} \in D_\delta$, $0 < \delta < 1$, the function

$$\left\| \hat{B}(e^{j\omega}) \right\|_A \triangleq \left\| \frac{\hat{B}(e^{j\omega})}{\hat{A}(e^{j\omega})} \right\|,$$

where $\hat{A}(p_i) = 0, i = 1, \ldots, \hat{n}_a$, defines a *norm* on $\mathbb{R}^{\hat{n}_b+1}$.

Proof. Since all roots of $\hat{A}(z)$ are in D_δ ($|z| \leq 1 - \delta$), we have

(1) $\left\| \hat{B} \right\|_A \geq 0. \quad \left\| \hat{B} \right\|_A = 0 \iff \hat{B}(e^{j\omega}) \equiv 0 \iff \hat{b}_i = 0, i = 0, \ldots, \hat{n}_b$

(2) $\left\| c\hat{B} \right\|_A = \left\| \frac{c\hat{B}}{\hat{A}} \right\| = |c| \cdot \left\| \frac{\hat{B}}{\hat{A}} \right\| = |c| \cdot \left\| \hat{B} \right\|_A$

(3) $\left\| \hat{B}_1 + \hat{B}_2 \right\|_A = \left\| \frac{\hat{B}_1 + \hat{B}_2}{\hat{A}} \right\| = \left\| \frac{\hat{B}_1}{\hat{A}} + \frac{\hat{B}_2}{\hat{A}} \right\| \leq \left\| \frac{\hat{B}_1}{\hat{A}} \right\| + \left\| \frac{\hat{B}_2}{\hat{A}} \right\| = \left\| \hat{B}_1 \right\|_A + \left\| \hat{B}_2 \right\|_A.$

Thus the defining properties of a norm are satisfied. ∎

Lemma 1.9. The domain of $J(\hat{\theta})$ can be restricted, without loss of generality, to a compact subset $\hat{\Theta}$ of $\Re^{\hat{N}}$.

Proof. We must show that the minimum of $J(\hat{\theta}) = \| H(e^{j\omega}) - \hat{B}(e^{j\omega})/\hat{A}(e^{j\omega}) \|$ occurs over a compact domain $\hat{\Theta} \subseteq \Re^{\hat{N}}$.

Since the roots of $\hat{A}(z)$ lie in D_δ, the poles of $\hat{H}(z)$ lie in a compact subset of $\Re^{\hat{n}_a}$. The coefficients of $\hat{A}(z)$ are a continuous function of the roots of $\hat{A}(z)$, and therefore they also lie in a compact subset of $\Re^{\hat{n}_a}$ (Rudin [154], Thm. 4.14). This set is also compact relative to $\Re^{\hat{N}}$ [154, Thm. 2.33].

It remains to be shown that $\{\hat{b}_0, \ldots, \hat{b}_{\hat{n}_b}\}$ can be made compact. Since $\hat{\theta} = 0$ is an admissible approximation, an optimum value $\hat{\theta}^*$ must satisfy $J(\hat{\theta}^*) \leq J(0)$. Let $\hat{\Theta} = \{\hat{\theta} \,|\, J(\hat{\theta}) \leq J(0)\}$ (sometimes called the *level set* corresponding to $\hat{\theta} = 0$). Then for $\hat{\theta} \in \hat{\Theta}$, we have

$$J(0) = \| H \| \geq \left\| H - \hat{H} \right\| \geq \left\| \hat{H} \right\| - \| H \|$$

$$\Rightarrow \quad \left\| \hat{H} \right\| \leq 2 \| H \| .$$

Let

$$\left\| \hat{B}(e^{j\omega}) \right\|_A \triangleq \left\| \frac{\hat{B}(e^{j\omega})}{\hat{A}(e^{j\omega})} \right\| .$$

By Lemma 1.8, this defines a norm on $\Re^{\hat{n}_b+1}$. By the *norm equivalence theorem*, it is "equivalent" to any other norm on $\Re^{\hat{n}_b+1}$ (Gohberg [142], p. 197). I.e., there exist positive real numbers c and C such that

$$c \| \cdot \|' \leq \| \cdot \|_A \leq C \| \cdot \|' ,$$

for any norm $\| \cdot \|'$, where $\| \cdot \|$ denotes $\| x \|$ for arbitrary x. Let $\| \cdot \|' = \| \cdot \|_1$.† Then there exist positive constants c and C such that

$$c \left\| \hat{B} \right\|_1 \leq \left\| \hat{B} \right\|_A \leq C \left\| \hat{B} \right\|_1 .$$

The constants c and C depend on $\hat{A}$ and the norm $\| \cdot \|$ used for problem $\hat{H}^*$. Since $\| \hat{B} \|_A \leq 2 \| H \|$, it follows that $c \| \hat{B} \|_1 \leq 2 \| H \|$, or

$$\sum_{n=0}^{\hat{n}_b} \left| \hat{b}_n \right| \leq \frac{2}{c} \left\| H(e^{j\omega}) \right\| .$$

Thus the space $\{\hat{b}_n\}_0^{\hat{n}_b}$ of numerator coefficients is *bounded*. If it is also closed, then it is compact [154]. Let $\hat{B}$ be a limit point of the sequence $\hat{B}_n$ where $\| H - \hat{B}_n/\hat{A} \| \leq \| H \|$ for

† Defined in Appendix E.

each n. Then $\lim_{n\to\infty} \| \hat{B} - \hat{B}_n \| = 0$ for any norm. Also,

$$\left\| \hat{B} - \hat{B}_n \right\|_A = \left\| \frac{\hat{B}}{\hat{A}} - \frac{\hat{B}_n}{\hat{A}} \right\| = \left\| \left(H - \frac{\hat{B}}{\hat{A}} \right) - \left(H - \frac{\hat{B}_n}{\hat{A}} \right) \right\|$$
$$\geq \left\| H - \frac{\hat{B}}{\hat{A}} \right\| - \left\| H - \frac{\hat{B}_n}{\hat{A}} \right\| \geq \left\| H - \frac{\hat{B}}{\hat{A}} \right\| - \| H \|.$$

Consequently,

$$0 = \lim_{n\to\infty} \left\| \hat{B} - \hat{B}_n \right\|_A \geq \left\| H - \frac{\hat{B}}{\hat{A}} \right\|_A - \| H \|$$
$$\Rightarrow \quad \left\| H - \frac{\hat{B}}{\hat{A}} \right\| \leq \| H \| \Rightarrow \quad \hat{B} \in \hat{\Theta}.$$

Thus $\hat{\Theta}$ is closed as well as bounded, and therefore compact. ∎

Lemma 1.10. The error measure $J(\hat{\theta})$ is uniformly continuous on $\hat{\Theta}$.

Proof. Since $\hat{\Theta}$ is compact, it suffices to show continuity of J at an arbitrary point of $\hat{\Theta}$ [154, Thm. 4.19]. Let

$$\underline{D} \triangleq \begin{pmatrix} D_{\hat{B}} \\ D_{\hat{A}} \end{pmatrix} \in \hat{\Theta}, \qquad \| \underline{D} \| = 1$$

define an arbitrary direction in $\hat{\Theta}$, and define

$$D_{\hat{B}}(z) = \sum_{n=0}^{\hat{n}_b} d_{\hat{B}}(n) z^{-n}$$
$$D_{\hat{A}}(z) = \sum_{n=0}^{\hat{n}_a} d_{\hat{A}}(n) z^{-n},$$

where $d_{\hat{A}}(n), d_{\hat{B}}(n)$ are the elements of $\underline{D}_{\hat{A}}, \underline{D}_{\hat{B}}$, respectively. Then J is continuous at $\hat{\theta}$ if for all such $\underline{D}$, $\lim_{t\downarrow 0} J(\hat{\theta} + t\underline{D}) = J(\hat{\theta})$. We have

$$J(\hat{\theta} + t\underline{D}) = \left\| H - \frac{\hat{B} + tD_{\hat{B}}}{\hat{A} + tD_A} \right\| = \left\| H - \frac{\hat{B}}{\hat{A}} + \frac{\hat{B}}{\hat{A}} - \frac{\hat{B} + tD_{\hat{B}}}{\hat{A} + tD_A} \right\|$$

which implies

$$\left| \left\| H - \frac{\hat{B} + tD_{\hat{B}}}{\hat{A} + tD_A} \right\| - \left\| H - \frac{\hat{B}}{\hat{A}} \right\| \right| \leq \left\| \frac{\hat{B}}{\hat{A}} - \frac{\hat{B} + tD_{\hat{B}}}{\hat{A} + tD_A} \right\| = t \left\| \frac{\hat{B}D_A - \hat{A}D_{\hat{B}}}{\hat{A}^2 + t\hat{A}D_A} \right\|. \qquad (1.3)$$

Since $\hat{A}$ is bounded away from zero on the unit circle, and since $\hat{A}, \hat{B}, D_{\hat{A}}, D_{\hat{B}}$ are continuous bounded polynomials in $e^{j\omega}$, it follows that the last term above goes to zero with t. Hence, $J(\hat{\theta})$ is continuous. ∎

Dividing both sides of (1.3) by t, we obtain the following side result.

Corollary 1.11. The directional derivative of J at $\hat{\theta}$ in the direction $\underline{D}$, $\|\underline{D}\| = 1$, is bounded in magnitude by

$$\left| \frac{\partial J(\theta)}{\partial \theta}(\hat{\theta}, \underline{D}) \right| \leq \left\| \frac{\hat{B}D_{\hat{A}} - \hat{A}D_{\hat{B}}}{\hat{A}^2} \right\|.$$

This follows from the definition of the directional derivative

$$\frac{\partial J(\theta)}{\partial \theta}(\hat{\theta}, \underline{D}) \triangleq \lim_{t \downarrow 0} \frac{J(\hat{\theta} + t\underline{D}) - J(\hat{\theta})}{t}.$$

Corollary 1.12. The set of values assumed by the error measure $J(\hat{\theta})$ on $\hat{\Theta}$ forms a compact subset of $\Re^1$. (This follows from the fact that the image of a compact set under a continuous mapping is compact [154, Thm. 4.14].)

Theorem 1.13. Let S be a closed and bounded subset of $\Re^n$. If $f : \Re^n \mapsto \Re^1$ is *continuous* on S, then f has at *least* one minimizer in S.

Proof. See Rudin [154], Thm. 2.28.

Theorem 1.14. Problem $\hat{H}^*$ has at least one solution.

Proof. In view of Thm. 1.13, it is only necessary to show that the domain $\hat{\Theta}$ of $J(\hat{\theta})$ is compact, and that $J(\hat{\theta})$ is continuous with respect to the coefficient vector $\hat{\theta}^T = (\hat{b}_0, \ldots, \hat{b}_{\hat{n}_b}, \hat{a}_1, \ldots, \hat{a}_{\hat{n}_a})$ on $\hat{\Theta}$. Lemmas 1.9 and 1.10 provide this information. ∎

1.2.2. Uniqueness

Proposition 1.15. If $J(\hat{\theta}^*) = 0$, then the optimum transfer function $\hat{H}^*(z)$ is unique.

Proof. From the definition of a norm, we have $\|H - \hat{H}\| = 0$ if and only if $H(e^{j\omega}) = \hat{H}(e^{j\omega})$ almost everywhere. With $\hat{H}(z)$ strictly stable, $\hat{H}(e^{j\omega})$ cannot be modified at points of measure zero,* and so it is uniquely determined. Since $\hat{H}(z)$ is rational and analytic for $|z| > 1 - \delta$, it is uniquely determined by $\hat{H}(e^{j\omega})$. ∎

* In the case of nonstrict stability, a pair of poles and zeros can converge toward cancellation at the unit circle in such a way as to leave one arbitrary point in $H(e^{j\omega})$ (Rice [57]). This is the basis for ARMA modeling of sinusoids in white noise.

Proposition 1.16. $\hat{H}^*$ is unique under the L^2 norm† whenever $\hat{E}^*(e^{j\omega}) \triangleq H(e^{j\omega}) - \hat{H}^*(e^{j\omega})$ is uncorrelated with $\tilde{H}(e^{j\omega}) \triangleq \hat{H}^*(e^{j\omega}) - \hat{H}(e^{j\omega})$ for all $\hat{H} \neq \hat{H}^*$.

Proof. Let $\langle X, Y \rangle$ denote the inner product† of the functions $X(e^{j\omega})$ and $Y(e^{j\omega})$ on the unit circle. Then

$$\begin{aligned} J_2^2(\hat{\theta}) &= \| H - \hat{H} \|_2^2 = \| \hat{H}^* + \hat{E}^* - \hat{H} \|_2^2 = \| \tilde{H} + \hat{E}^* \|_2^2 = \left\langle \tilde{H} + \hat{E}^*, \tilde{H} + \hat{E}^* \right\rangle \\ &= \| \hat{E}^* \|_2^2 + \left\langle \tilde{H}, \hat{E}^* \right\rangle + \left\langle \hat{E}^*, \tilde{H} \right\rangle + \| \tilde{H} \|_2^2 \\ &= \| \hat{E}^* \|_2^2 + \| \tilde{H} \|_2^2, \end{aligned}$$

which is minimum if and only if $\| \tilde{H} \| = 0 \Rightarrow \hat{H}(e^{j\omega}) \equiv \hat{H}^*(e^{j\omega})$. ∎

Note that uniqueness of the approximate transfer function does not imply uniqueness of the parameter vector $\hat{\theta}^* = (\hat{b}_0^*, \ldots, \hat{b}_{\hat{n}_b}^*, \hat{a}_1^*, \ldots, \hat{a}_{\hat{n}_a}^*)$. If, however, $\hat{A}^*(z)$ and $\hat{B}^*(z)$ have no common roots, then uniqueness of $\hat{H}^*$ implies uniqueness of $\hat{\theta}^*$.

Proposition 1.17. If the poles of $\hat{H}(z)$ in problem $\hat{H}^*$ are fixed, or if $\hat{n}_a = 0$ (no poles), and if a *strictly concave norm*† is used for the error measure, then the solution $\hat{H}^*(z)$ is *unique.*

Proof. Suppose there are two best approximations $\hat{H}^*(z) = \hat{B}^*(z)/\hat{A}^*(z)$ and $\tilde{H}^*(z) = \tilde{B}^*(z)/\hat{A}^*(z)$. Then by strict concavity,

$$\left\| H - \frac{\hat{H}^* + \tilde{H}^*}{2} \right\| = \left\| \frac{H - \hat{H}^*}{2} + \frac{H - \tilde{H}^*}{2} \right\| < \frac{1}{2} \left\| H - \hat{H}^* \right\| + \frac{1}{2} \left\| H - \tilde{H}^* \right\| = \left\| H - \hat{H}^* \right\|.$$

Thus,

$$\frac{\hat{H}^* + \tilde{H}^*}{2} = \frac{\hat{B}^* + \tilde{B}^*}{2\hat{A}^*}$$

is a better approximation. This contradicts the assumed optimality of $\hat{B}^*$ and $\tilde{B}^*$. ∎

Corollary 1.18. If there are two solutions to problem $\hat{H}^*$ with pre-assigned poles then there is an infinite number of solutions.

Since all L^p norms are strictly concave for $1 < p < \infty$, we have the following.

Corollary 1.19. The solution to problem $\hat{H}^*$ with pre-assigned poles is unique under all L^p norms for $1 < p < \infty$.

† Defined in Appendix E.

† Defined in Appendix E.

Walsh [75] has proved the above corollary for $p = \infty$ also.

For the case where poles and zeros must be optimized, i.e., for problem $\hat{H}^*$, *the solution is not unique* in general. This result was established for the L^∞ norm by Martin Gutknecht and Lloyd Trefethen in the summer of 1982 [39].

1.2.3. Approximation over a Discrete Set of Frequencies

So far, only continuous frequency has been considered. In practice, however, it is often necessary to work with discretized frequency. Below it is shown that problem $\hat{H}^*$ is not fundamentally altered by replacing the unit circle with a dense set of points on the unit circle.

Discrete-frequency approximation is defined as minimizing

$$J(\hat{\theta}) = \left\| H(e^{j\omega_k}) - \hat{H}(e^{j\omega_k}) \right\|,$$

where $\omega_k, k = 1, \ldots, N$ form a discrete set of frequencies on which the error is measured. Note that under these conditions, we are really dealing with a *pseudo-norm* since the error could have zero-crossings at each ω_k without vanishing identically.

Lemma 1.20. A best discrete-frequency approximation always *exists*.

Proof. The proof of Thm. 1.14 goes through with minor modifications. In essence, working over a discrete set of frequencies does not alter the compactness of $\hat{\Theta}$.

Note that in [3] the opposite is claimed, i.e., that solutions need not exist over a discrete frequency grid. Since this is not shown in the paper, I can only suspect that the reason has to do with a lack of strict stability. A counter-example to existence due to Walsh involves an isolated point at $z = 0$, which does not apply to problem $\hat{H}^*$.

The following is similar to a theorem of Cheney [13] for real approximation in the L^∞ norm. However, it is much stronger thanks to the strict stability assumption.

Theorem 1.21. Let $H(e^{j\omega})$ be continuous and let $\hat{H}_k^*(z) \in \mathcal{H}_{\hat{n}_b,\hat{n}_a}$ be a best approximation to H over the set of k points uniformly distributed over the unit circle Γ. Then

$$\lim_{k\to\infty} \hat{H}_k^* = \hat{H}^*,$$

exists and is a best approximation to H on Γ. The distribution of discrete points on the unit circle may be arbitrary provided it becomes dense in Γ as $k \to \infty$.

Proof. Let $\hat{\theta}_k^*$ denote the vector of filter coefficients corresponding to each $\hat{H}_k^*$. By Lemma 1.9, these vectors lie in the compact set $\hat{\Theta}$. Consequently, the sequence $\hat{\theta}_k^*$ has a

subsequence which converges to, say, $\hat{\theta}^* \in \hat{\Theta}$. By Lemma 1.10, $J(\hat{\theta}_k^*) = \|H - \hat{H}_k^*\|$ is uniformly continuous on $\hat{\Theta}$ which implies $\lim_{k\to\infty} J(\hat{\theta}_k^*) = J(\hat{\theta}^*)$.

Suppose $\hat{H}^*$ is not a best approximation. Then there exists $\tilde{\theta}^* \in \hat{\Theta}$ such that $J(\hat{\theta}^*) - J(\tilde{\theta}^*) = \epsilon > 0$. By the asymptotic density of the points ω_k, and by continuity, there exists an integer K sufficiently large so that $|J(\hat{\theta}_k^*) - J(\hat{\theta}^*)| < \epsilon/4$ for all $k > K$. Similarly, there exists a sequence $\tilde{\theta}_k^* \in \hat{\Theta}$ such that $\lim_{k\to\infty} \tilde{\theta}_k^* = \tilde{\theta}^*$ and an integer $\tilde{K}$ for which $|J(\tilde{\theta}_k^*) - J(\tilde{\theta}^*)| < \epsilon/4$ for all $k > \tilde{K}$. Let $M = \max\{K, \tilde{K}\}$. Then for all $k > M$ we have $J(\hat{\theta}_k^*) - J(\tilde{\theta}_k^*) > \epsilon/2 > 0$. This contradicts the assumed optimality of $\hat{\theta}_k^*$. ∎

1.2.4. Feasibility of Gradient/Newton Descent for General Norms

The applicability of algorithms based on gradient descent is governed by the nature of the error surface $J(\hat{\theta})$. All gradient-based methods are "local" in the sense that they only measure "slope" and perhaps "curvature" at the point of the error surface corresponding to the current parameter estimate $\hat{\theta}$ (cf. Appendix E). Thus the gradient methods tend to terminate at the bottom of the first "valley" of the error surface encountered. If the error surface has many valleys, there is usually no guarantee that the first one encountered has the deepest bottom. This is why concavity is important for gradient algorithms—concavity guarantees that the entire surface of $J(\hat{\theta})$ over $\hat{\Theta}$ is one big valley.

In the case of linear approximation, every norm is concave with respect to the parameters. For problem $\hat{H}^*$, $H - \hat{H}$ is linear in the parameters when $\hat{n}_a = 0$, i.e., when the filter has a polynomial transfer function $\hat{H}(z) = \hat{B}(z)$. To see this, note that for $\lambda \in [0,1]$, $\check{\lambda} = 1 - \lambda$,

$$J(\lambda\hat{B}_1 + \check{\lambda}\hat{B}_2) \triangleq \left\| H - (\lambda\hat{B}_1 + \check{\lambda}\hat{B}_2) \right\| \leq \lambda \left\| H - \hat{B}_1 \right\| + \check{\lambda} \left\| H - \hat{B}_2 \right\| \triangleq \lambda J(\hat{B}_1) + \check{\lambda} J(\hat{B}_2).$$

This proves the following:

Proposition 1.22. Problem $\hat{H}^*$ is solvable by gradient methods when there are no poles in the approximation, i.e., when $\hat{n}_a = 0$. In other terms, finite-impulse-response (FIR) filter design can be carried out under any norm using gradient methods.

For the case of rational approximation, concavity is another story. In fact, most results along these lines are negative. For example, we have the following.

Theorem 1.23. Let K be a positive integer. Then for *any* discrete-frequency norm, there exists an order $8K$ FIR filter $H(z)$ and a frequency-grid size N such that the one-pole approximation-error norm

$$J(r) = \left\| H(e^{j\omega}) - \frac{1}{1 - rz^{-1}} \right\|,$$

has K local minima.

Proof. This is the subject of Appendix A.

Corollary 1.24. Problem $\hat{H}^*$ is not concave under any discrete-frequency norm, and there is no upper bound on the number of locally best approximations. Consequently, no gradient-based method can be guaranteed to converge for a filter-design problem with one or more poles in the approximation.

It is worth noting that for all practical purposes, problem $\hat{H}^*$ is not solvable by gradient techniques under any norm whatsoever, since any computer implementation must use discretized frequency. The result can also be extended to most continuous-frequency norms, due to strict stability. These results imply that when modeling spectra with poles and zeros, gradient-descent, Newton-descent and any other algorithms which require concavity cannot be relied upon unless further restrictions on the desired spectrum are possible, or unless good initial estimates are available. Yahagi [79] has found empirically that rational filter orders above 6 or 7 are difficult to design with L^2 norm criteria. Box and Jenkins [135] (in the maximum likelihood ARMA modeling context) recommend "extensive" plotting of the error surface $J(\hat{\theta})$. This of course approaches exhaustive search of the parameter space.

1.2.5. Computational Methods

Some generally useful Fortran programs for filter design are available in the IEEE Programs for Digital Signal Processing collection [169]. The program for "Least-P IIR filter design" by A. G. Deczky comes closest to solving problem $\hat{H}^*$ as defined. It allows the design of recursive filters with arbitrarily specified *magnitude frequency response* and *group delay*. A weighted sum of weighted L^p norms† of errors in these two quantities is minimized. While this is not precisely a norm on $H - \hat{H}$, it has the basic property of matching both phase and magnitude of the desired frequency-response. The Fletcher-Powell algorithm [171], a popular general-purpose gradient-based nonlinear optimization procedure, is used to search for the best approximation to the desired frequency-response. The *poles* and *zeros* of the filter are the parameters which are explicitly optimized by the algorithm—hence $\hat{\Theta}$ is concave. Because the optimization is nonlinear, one is not guaranteed that the best solution is always obtained since the error norm may exhibit local minima which will halt the gradient descent. Therefore, the initial pole and zero locations, which the user must provide, can be crucial to the quality of the design. Consequently, it is advisable to obtain an initial approximate filter design by the more robust methods discussed later in this chapter.

† Defined in Appendix E.

In Appendix A, it is shown that one-pole approximation to a length $2N+1$ FIR filter has at most N local minima. Similar upper bounds can be obtained when the desired frequency response corresponds to a rational filter, or some other convenient analytical expression. In such cases, it may be feasible to carry out a limited-resolution exhaustive search of the error surface. If the parameter space $\hat{\Theta}$ can be partitioned into a set of domains on which the error J is locally concave, then gradient methods can be initialized in each of these regions, and the results for each region may be compared to find the global minimizer. Methods for nonlinear optimization can be found in [161,163,174,179].

1.3. Minimization of the L^2 Norm

The L^2 error norm,

$$J_2(\hat{\theta}) = \left\| H(e^{j\omega}) - \hat{H}(e^{j\omega}) \right\|_2 \triangleq \int_{-\pi}^{\pi} \left| H(e^{j\omega}) - \hat{H}(e^{j\omega}) \right|^2 \frac{d\omega}{2\pi}$$
$$= \sum_{n=0}^{\infty} \left| h(n) - \hat{h}(n) \right|^2 \triangleq \left\| h(n) - \hat{h}(n) \right\|_2 ,$$

is a common choice for general filter-design algorithms. It can be interpreted as the square-root of the error *energy*. Most such methods are based on gradient descent [20,64,79]. In this section, some properties of problem $\hat{H}^*$ under the L^2 norm are developed.

1.3.1. Least-Squares FIR Filter Design

The first case considered is finite impulse-response (FIR) filter design, i.e., $\hat{n}_a = 0$ in (1.2) and the filter is of the form

$$\hat{H}(z) = \hat{B}(z) = b_0 + b_1 z^{-1} + \cdots + b_{\hat{n}_b} z^{-\hat{n}_b} .$$

The error measure becomes

$$J_2^2(\hat{\theta}) = \left\| H - \hat{B} \right\|_2^2 = \sum_{n=0}^{\hat{n}_b} (h(n) - b_n)^2 ,$$

and the unique minimum is given by

$$\hat{b}_n^* = h(n), \quad n = 0, 1, \ldots, \hat{n}_b .$$

Thus FIR filter design under the L^2 norm reduces to *Padé approximation* (see §1.8.7). It can be viewed as a consequence of the orthogonality of the functions $e^{j\omega n}, n = 0, 1, \ldots$ under the inner product.

1.3.2. Least-Squares Recursive Filter Design with Fixed Poles

Suppose $\hat{n}_b = \hat{n}_a - 1 \;\Rightarrow\; \hat{H} \in \mathcal{H}^T_{\hat{n}_a}$, and that the roots $\{p_i\}_1^{\hat{n}_a}$ of $A(z)$ are distinct. (As usual, $|p_i| \leq 1-\delta$ for some $0 < \delta < 1$.) Then $\hat{H}(z)$ can be written

$$\hat{H}(z) = \sum_{i=1}^{\hat{n}_a} \frac{R_i}{1 - p_i z^{-1}}, \tag{1.4}$$

and the values R_i are the residues of the poles of $\hat{H}(z)$.

Approximation with fixed poles amounts to optimizing the residues in (1.4) when $\hat{n}_b = \hat{n}_a - 1$. For $\hat{n}_b \geq \hat{n}_a$, the problem is similar except that the first $\hat{n}_b - \hat{n}_a + 1$ samples of the impulse-response will be matched exactly (by the FIR quotient).

Definition 1.25. The space of all functions analytic in $|z| \geq 1$ is denoted $H^{-\infty}$.

Definition 1.26. Let $H(z) \perp G(z)$ denote orthogonality on the unit circle, i.e.,

$$H(z) \perp G(z) \;\Leftrightarrow\; \langle H, G\rangle = 0 \;\Leftrightarrow\; \int_{-\pi}^{\pi} H(e^{j\omega})\overline{G(e^{j\omega})}\frac{d\omega}{2\pi} = 0.$$

The following lemma is adapted from Walsh [75].

Lemma 1.27. If $H(z) \in H^{-\infty}$, and if $H(1/\bar{a}) = 0$, where $|a| < 1$, then

$$H(z) \perp G(z) \triangleq \frac{1}{1 - az^{-1}}.$$

Proof. We must show $\langle G, H\rangle = 0$. Assume $a \neq 0$. Since $H(1/\bar{a}) = 0$, let $H(z) = (1 - \bar{a}^{-1}z^{-1})H'(z)$. Then

$$\begin{aligned}
\langle G, H\rangle &\triangleq \int_{-\pi}^{\pi} G(e^{j\omega})\overline{H(e^{j\omega})}\frac{d\omega}{2\pi} = \int_{-\pi}^{\pi} \frac{\overline{H(e^{j\omega})}}{1 - ae^{-j\omega}}\frac{d\omega}{2\pi} = \int_{-\pi}^{\pi} H'(e^{-j\omega})\frac{1 - a^{-1}e^{j\omega}}{1 - ae^{-j\omega}}\frac{d\omega}{2\pi} \\
&= \int_{-\pi}^{\pi} H'(e^{-j\omega})a^{-1}e^{j\omega}\frac{ae^{-j\omega} - 1}{1 - ae^{-j\omega}}\frac{d\omega}{2\pi} = -\int_{-\pi}^{\pi} H'(e^{-j\omega})a^{-1}e^{j\omega}\frac{d\omega}{2\pi} \\
&= -\frac{1}{2\pi j a}\oint_{\Gamma} H'(z^{-1})dz = 0
\end{aligned}$$

by the Cauchy residue theorem [138] since $H'(z^{-1})$ is analytic for $|z| \leq 1$. If $a = 0$, then $H(\infty) = 0 \;\Rightarrow\; h(0) = 0$, and we must show $\langle 1, h(1)z^{-1} + h(2)z^{-2} + \cdots\rangle = 0$ which follows immediately from the orthogonality of distinct powers of z on Γ. (Alternatively, note that $\langle H(z), 1\rangle = h(0)$ by the definition of the Fourier transform.) ∎

Lemma 1.28. If $H(z), G(z) \in H^{-\infty}$, then for all $F(z) \in H^{-\infty}$,

$$H(z) \perp G(z) \;\Rightarrow\; H(z) \perp F(z)G(z).$$

Proof. If $H(z) \perp G(z)$, then

$$0 = \int_{-\pi}^{\pi} H(e^{j\omega})G(e^{-j\omega})\frac{d\omega}{2\pi} = -\frac{1}{2\pi j}\oint_{\Gamma} H(z)G(z^{-1})\frac{dz}{z},$$

which implies the integrand has no poles in $\mathcal{D}$. But this remains true if $G(z)$ is replaced by $F(z)G(z)$ since $F(z^{-1})$ has no poles in $\mathcal{D}$. ∎

Theorem 1.29 (Walsh [75]). Let $H(z)$ be analytic in $|z| \geq 1$, then the solution to problem $\hat{H}^*$ with $\hat{n}_a$ *fixed* stable poles and $\hat{n}_a - 1$ zeros is the member of $\mathcal{H}_{\hat{n}_a}^{\pi}$ which interpolates $H(z)$ at the points $1/\overline{p_k}, k = 1, \ldots, n_a$, where $\{p_k\}$ are the pre-assigned poles of $\hat{H}$.

Proof. This is a direct consequence of the orthogonality relations developed in Lemmas 1.27 and 1.28. ∎

The above theorem is readily extended to the case $\hat{n}_b \geq \hat{n}_a$. When the pre-assigned poles are not distinct, the interpolation applies to successive derivatives of $H(z)$ at $1/p_i$ for each non-simple pole p_i. See Chapter 9 of Walsh [75] for a more complete discussion.

1.3.3. Least Squares Recursive Filter Design

This section addresses the solution of problem $\hat{H}^*$ under the L^2 norm. It is shown that there is no upper bound to the number of locally best approximations, and a construction is given exhibiting an arbitrary set of local minima in the one-pole case.

Theorem 1.30. Given any set of K distinct stable one-pole filters,

$$\hat{H}_i(z) = \frac{1}{1 - r_i z^{-1}}, \quad 0 < r_i < 1, \quad i = 1, 2, \ldots K,$$

there exists a bounded causal filter $H(z)$ having each $\hat{H}_i(z)$ as a locally best approximation under the L^2 norm. If $H(z)$ is taken from the set of order $2K$ FIR filters,

$$H(z) = h(0) + h(1)z^{-1} + \cdots + h(2K)z^{-2K},$$

and if the curvature of the squared L^2 error norm

$$c_i \triangleq \frac{\partial^2 J_2^2(r)}{\partial r^2}(r_i) > 0, \quad i = 1, \ldots, K,$$

is specified for each $\hat{H}_i$, then $H(z)$ is unique and is given by

$$\begin{pmatrix} h(1) \\ h(2) \\ \vdots \\ h(2K) \end{pmatrix} = \begin{pmatrix} 1 & 2r_1 & 3r_1^2 & \cdots & 2Kr_1^{2K-1} \\ 0 & 2 & 6r_1 & \cdots & 2K(2K-1)r_1^{2K-2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 2r_K & 3r_K^2 & \cdots & 2Kr_K^{2K-1} \\ 0 & 2 & 6r_K & \cdots & 2K(2K-1)r_K^{2K-2} \end{pmatrix}^{-1} \begin{pmatrix} \alpha_1 \\ \beta_1 \\ \vdots \\ \alpha_K \\ \beta_K \end{pmatrix}, \tag{1.5}$$

with $h(0) = 1$, and

$$\alpha_i \triangleq \sum_{n=0}^{\infty} n r_i^{2n-1} = \frac{r_i}{(1-r_i^2)^2},$$

$$\beta_i \triangleq \sum_{n=0}^{\infty} n(2n-1) r_i^{2n-2} - \frac{c_i}{2} = \frac{1+3r_i^2}{(1-r_i^2)^3} - \frac{c_i}{2}.$$

The proof is given in Appendix A. An immediate extension is the following.

Theorem 1.31. Let K be a positive integer, and $\hat{n}_b, \hat{n}_a$ be given with $\hat{n}_a \geq 1$. Then there exists an $H(e^{j\omega})$ with finite L^2 norm so that $\| H - \hat{H} \|_2$ has at least K local minima.

The existence of multiple local minima in the general case was stated without proof by Chui, Smith, and Su in [15]. The explicit construction (1.5) can be used to extend to arbitrary norms, as discussed in Appendix A.

1.4. Minimization of the L^∞ Norm

1.4.1. Chebyshev FIR Filter Design

The design of FIR digital filters can be treated as a special case of designing recursive filters with fixed poles. (The poles are simply fixed at $z = 0$). Since the theory for preassigned poles is essentially the same, there is little reason to consider FIR design separately.

1.4.2. Chebyshev Recursive Filter Design with Fixed Poles

Suppose again that $\hat{n}_b = \hat{n}_a - 1$, and that the poles $p_i, i = 1, \ldots, \hat{n}_a$ of $\hat{H}(z)$ are distinct and fixed. In such a case $\hat{H} \in \mathcal{H}_{\hat{n}_a}^{\pi}$ can be expanded as

$$\hat{H}(z) = \sum_{k=1}^{\hat{n}_a} \frac{R_k}{1 - p_k z^{-1}}. \tag{1.6}$$

Problem $\hat{H}^*$ then reduces to finding the residues R_k which minimize $\|H - \hat{H}\|_\infty$. It is convenient to define the following *basis functions* for $\mathcal{H}^\pi_{\hat{n}_a}$:

$$G_i(z) = \frac{1}{1 - p_i z^{-1}}, i = 1, \ldots, \hat{n}_a .$$

Definition 1.32. An Nth order *Chebyshev set* on the unit circle Γ is a set of N functions for which every linear combination has at most $N-1$ zeros on the unit circle.

The span of such functions is sometimes called a *Haar subspace.*

An important feature of Chebyshev sets is the *interpolation property.* The $\hat{n}_a \times \hat{n}_a$ matrix $[G_l(e^{j\omega_k})]$ is guaranteed to be nonsingular for every set of distinct points $\omega_k, k = 1, \ldots, \hat{n}_a$ when $\{G_l\}$ is a Chebyshev set [57]. This fact allows $\hat{H}(e^{j\omega})$ to interpolate through arbitrary values over any $\hat{n}_a$ points of the unit circle.

Lemma 1.33. The functions $G_i(e^{j\omega}), i = 1, \ldots, \hat{n}_a$ form a Chebyshev set of order $\hat{n}_a$ on the unit circle.

Proof. This is immediate from the one-to-one correspondence between the form (1.6) and rational filters with $\hat{n}_a$ poles and $\hat{n}_a - 1 = \hat{n}_b$ zeros. ∎

Theorem 1.34. The solution to problem $\hat{H}^*$ with $\hat{n}_a$ fixed poles and $\hat{n}_a - 1$ zeros is *unique.*

Proof. See Chapter 2 of Lorentz [49] where it is proved that if a best approximation with a linear combination of functions from a Chebyshev set exists, then it is unique. ∎

An interesting characterization of the L^∞ solution, which connects it with least-squares, is given by the following.

Theorem 1.35 (Kolmogorov, Rivlin, Shapiro). A filter $\hat{H}(z) \in \mathcal{H}^\pi_{\hat{n}_a}$ with fixed poles is a best Chebyshev approximation to $H(z) \in C_0(\Gamma)$ if and only if there are r points $\{e^{j\omega_k}\}_1^r$ on Γ and r numbers $w_1 > 0, \ldots, w_r > 0$, with $\sum w_k = 1$, such that

$$\sum_{k=1}^{r} w_k [H(e^{j\omega_k}) - \hat{H}(e^{j\omega_k})] \overline{G_i(e^{j\omega_k})} = 0, \qquad i = 1, \ldots, \hat{n}_a, \tag{1.7}$$

where

$$\hat{H}(z) = \sum_{k=1}^{\hat{n}_a} R_k G_k(z),$$

and $\{p_k\}_1^{\hat{n}_a} \subseteq \mathcal{D}_\delta$ are the fixed poles of $\hat{H}(z)$. We have also that $\hat{n}_a + 1 \leq r \leq 2\hat{n}_a - 1$ and that

$$\left| H(e^{j\omega_k}) - \hat{H}(e^{j\omega_k}) \right| = \left\| H - \hat{H} \right\|_\infty, \qquad k = 1, \ldots, r .$$

Proof. See Rivlin and Shapiro [59] or Lorentz [49], Chapter 2.

In the case of real-valued approximation, there are $r = \hat{n}_a + 1$ "extremal points" [29].

The characterization above can be viewed as a *weighted least squares* solution where the weight function consists of an impulse at each extremal point $e^{j\omega_k}$. If the weights w_k and their locations $e^{j\omega_k}$ were known, then the optimal approximation could be easily computed by solving the "normal equations" (1.7). This is the basis of *Lawson's Method* [29,30]. The method is *guaranteed to converge* [58], and its rate of convergence is approximately linear.

1.4.3. Chebyshev Recursive Filter Design

In the case of trying to minimize the Chebyshev (L^∞) norm of the *complex* spectral error $\| H(e^{j\omega}) - \hat{H}(e^{j\omega}) \|_\infty$, multiple local minima and saddle-points may arise [33,34,3,78,30]. (See also Appendix A). Consequently, there seems to be no "local" algorithm (e.g. gradient or Newton method) which is guaranteed to find an optimal $\hat{H}$ from an arbitrary starting point. Moreover, it has been recently established by Gutknecht and Trefethen that L^∞ rational approximation on the unit disk does not have a unique solution [38].

However, when good initial approximations are available, there are algorithms for finding a locally best approximation. A descent method has been developed by Gutknecht [33], written in Algol60 for the CDC6500. A version of Lawson's algorithm for complex rational approximation developed by Ellacott and Williams is presented in [30]. Also, Alliney has implemented Lawson's algorithm in CDC CYBER-76 Fortran [3], and his paper is written from a filter-design point of view. In spite of the lack of theoretical guarantees, the method has been reported to give satisfactory numerical results in some circumstances [3]. Further information on this technique is given in [34] and [78].

Though complex rational approximation on the unit circle has theoretical difficulties, it will be shown in the next section that it is possible to find the *unique* optimum complex Chebyshev approximation to $H(e^{j\omega})$ out of a larger class of rational functions which consists of functions from $\mathcal{H}_{\hat{n}_b,\hat{n}_a}$ augmented with poles from outside the unit circle. If $H \in C_0^+(\Gamma)$ (causal), and if the approximation error is small, then these unstable poles (which can be considered as generators of the noncausal part of $\hat{H}$) can be deleted with little effect on the error. Such an approximation happens to minimize the *Hankel norm.*

1.5. Minimization of the Hankel Norm

In contrast to all norms thus far considered, the *Hankel norm* leads to a satisfactory solution of problem $\hat{H}^*$ under general conditions. The material presented below is adapted from a paper coauthored with Martin Gutknecht and Lloyd Trefethen [36].

1.5.1. The CF Method for Hankel-Norm Minimization

The algorithm we have developed based on Hankel norm minimization is called the *CF algorithm*. The CF method minimizes the *Hankel norm*† of the frequency-response error,

$$J_H(\hat{\theta}) = \left\| H(e^{j\omega}) - \hat{H}(e^{j\omega}) \right\|_H,$$

when $\hat{n}_b \geq \hat{n}_a - 1$. The resulting filter is always *stable*, and it typically closely approaches the optimum rational Chebyshev approximation when $H(e^{j\omega}) \in C_0^+(\Gamma)$. Another advantage of this method is that the error associated with all filter orders up to a desired maximum are available with no extra computation (with $\hat{n}_a - \hat{n}_b$ fixed). Thus, when *order identification* is required, or when very efficient designs are called for in terms of order versus error, the CF method is well suited. Perhaps most important, however, is that the CF method does not suffer from the non-concavity of the error surface $J_H(\hat{\theta})$. Unlike gradient and Newton descent methods, it goes directly to the unique optimum solution, to within computational errors which can be made arbitrarily small by increasing various array sizes in the implementation.

The CF method uses the desired impulse-response $h(n)$ and therefore classifies as a *time-domain* filter-design method. It is based on an extension due primarily to Takagi [67] of a classical theorem in complex analysis of Carathéodory and Fejér [10].

Techniques related to the CF method are used in the *model order reduction* problem. A presentation of some of this work may be found in [43,44,31,32,198]. Note that in these works, the starting-point is a known *rational* digital filter $H(z) = B(z)/A(z)$ which is to be approximated by a lower-order rational filter. In the special case of starting with an FIR filter (no poles), their problem reduces to the one solved by the CF method. From this point of view, the CF method may be considered to offer two advantages relative to previous work. First, it is formulated so as to yield filters with an arbitrary number of poles and zeros; i.e. the restriction that there be $\hat{n}_a - 1$ zeros when there are $\hat{n}_a$ poles is removed. Secondly, the above references propose methods which include a partial fraction expansion. Experience has shown that this can severely limit the length of the original impulse-response which can be used, since this length is the size of the polynomial which must be factored.

The present method circumvents partial fraction expansion by means of the fast spectral factorization technique of §1.9.2, applicable whenever the number of poles and zeros inside the unit circle are known *a priori*. As a result, the CF method can typically be applied to a much longer impulse-response given equal computational environments. Most

† Defined in Appendix E.

of the computation time (about 90%) is spent in tri-diagonalizing the Hankel matrix of the impulse-response. Since the Hankel structure of the matrix is not exploited in this step, it is felt that very significant speed increases are possible.

It appears that the CF method provides an excellent means of initializing L^∞ approximation schemes, since the nature of the approximation tends to be nearly "equal ripple."

1.5.2. Theoretical Basis of the CF Method

Assume an ideal causal impulse-response $h(n)$ $(n = 0, 1, \ldots)$ is given, corresponding to a stable ideal transfer function

$$H(z) \triangleq \sum_{n=0}^{\infty} h(n) z^{-n}, \qquad H(e^{j\omega}) \in C_0^+(\Gamma).$$

Stability implies that this series converges uniformly, so $H(z)$ can be approximated arbitrarily closely by taking a partial sum of sufficiently high order. In practice, it may be preferable to apply a bandlimited window [196] rather than truncate, and the possibly modified impulse-response values are denoted by $\{h_K(n)\}_0^K$ and the corresponding transfer function by $H_K(z)$,

$$H_K(z) \triangleq \sum_{n=0}^{K} h_K(n) z^{-n}.$$

Problem $\hat{H}^*$ under the Hankel norm is then to find the stable rational digital filter

$$\hat{H}(z) = \frac{\hat{B}(z)}{\hat{A}(z)} = \frac{\sum_{k=0}^{\hat{n}_b} \hat{b}_k z^{-k}}{1 + \sum_{k=1}^{\hat{n}_a} \hat{a}_k z^{-k}}$$

which minimizes

$$J_H(\hat{\theta}) = \left\| H(e^{j\omega}) - \hat{H}(e^{j\omega}) \right\|_H,$$

where the Hankel norm is defined in Appendix E.

The key to Hankel-norm minimization is that it is easy to determine the best *Chebyshev* (L^∞) approximation $\tilde{H}^*$ out of the larger class $\tilde{\mathcal{H}}_{\hat{n}_b,\hat{n}_a}$ of functions which are of the form

$$\tilde{H}(z) \triangleq \frac{\tilde{\hat{B}}(z)}{\hat{A}(z)} \triangleq \frac{\sum_{k=-\infty}^{\hat{n}_b} \hat{b}_k z^{-k}}{1 + \sum_{k=1}^{\hat{n}_a} \hat{a}_k z^{-k}} \tag{1.8}$$

where the zeros of $z^{\hat{n}_a}\hat{A}(z)$ still lie inside the unit circle Γ. The class $\tilde{\mathcal{H}}_{\hat{n}_b,\hat{n}_a}$ may be regarded as an extension of the filters in $\mathcal{H}_{\hat{n}_b,\hat{n}_a}$ to include noncausal impulse-response

components. The CF method consists of computing this extended best L^∞ approximation $\tilde{H}^*$, and truncating it to obtain the CF approximant $\hat{H}^{(CF)} \in \mathcal{H}_{\hat{n}_b,\hat{n}_a}$. The truncation of the noncausal part produces a filter which minimizes J_H, as discussed in Appendix B. One way to perform this truncation is to express $\tilde{H}^*$ in the parametric form (1.8) and delete the terms with negative k in the numerator [37]. A better method, which is employed here, is to compute the impulse-response (Laurent series on Γ) for (1.8). If $\hat{n}_b \geq \hat{n}_a - 1$, then one simply truncates all noncausal terms, and what remains is the impulse-response for a function in $\mathcal{H}_{\hat{n}_b,\hat{n}_a}$. For $\hat{n}_b < \hat{n}_a - 1$, a slight modification of this procedure is necessary, as described in the next section.

In our experience, the resulting approximations are quite close to optimal in the Chebyshev sense, especially when $H(e^{j\omega})$ is smooth. Because the complex Chebyshev measure-of-fit is sensitive to both phase and magnitude errors, the noncausal part of the Chebyshev approximation is small when the optimum error is small. Indeed, if $\| H_K(e^{j\omega}) - \tilde{H}^*(e^{j\omega}) \|_\infty = \lambda$, then

$$\begin{aligned} \left| h_K(n) - \tilde{h}^*(n) \right| &\triangleq \left| \int_{-\pi}^{\pi} \left(H_K(e^{j\omega}) - \tilde{H}^*(e^{j\omega}) \right) e^{j\omega n} \frac{d\omega}{2\pi} \right| \\ &\leq \int_{-\pi}^{\pi} \left| H_K(e^{j\omega}) - \tilde{H}^*(e^{j\omega}) \right| \frac{d\omega}{2\pi} \leq \lambda, \end{aligned}$$

where $\tilde{h}^*(n), n \in (-\infty, \infty)$ is the impulse-response of the Chebyshev approximation $\tilde{H}^*$. Thus λ is an upper bound on the error of each individual sample of the impulse-response error, implying, in particular, $|\tilde{h}^*(n)| \leq \lambda, n < 0$. Moreover, $\tilde{h}^*(n)$ tapers *exponentially* to zero as $n \rightarrow -\infty$. However, the truncation error is often much smaller than this. For some estimates on its size, see [71] and [72].

Since the CF approximant and the ideal filter are both stable, the time-domain error approaches zero exponentially. However, by the above inequality, *at no time sample can the error in the impulse-response exceed* λ, *the optimal Chebyshev spectral error.* In general, problem $\hat{H}^*$ must provide fits both in the time-domain and the frequency-domain, since frequency-response phase corresponds to relative time information. However, the above bound on the impulse-response error is better than one normally sees.

The method for computing $\tilde{H}^*$ is based on a theorem developed by Takagi [67], Achieser [1], Clark [17], and Adamjan, Arov, and Krein [2], for which an elementary proof is given in [72]. Appendix B gives a simplified derivation of the main results for the case of real finite-order filter design. For a detailed presentation of the Takagi theory, see also [35]. The polynomial case ($\hat{n}_a = 0$) was settled earlier by Carathéodory and Fejér [10].

The theorem makes use of the *singular value decomposition* of the *Hankel matrix* formed from the windowed impulse-response $\{h_K(n)\}_{n=0}^K$. The values $h_K(n)$ may be complex. By definition, the Hankel matrix corresponding to an impulse-response $\{h_K(n)\}_{n=0}^\infty$ is

the infinite matrix having $h_K(i+j)$ at the intersection of the ith row and jth column ($i,j = 0,1,2,\ldots$). To obtain general type $(\hat{n}_b, \hat{n}_a)$ approximations, we introduce the parameter

$$\nu \triangleq \hat{n}_b - \hat{n}_a + 1$$

and define the Hankel matrix entry (i,j) as $h_K(i+j+\nu)$,

$$\mathbf{H}_{\nu,K} \triangleq \begin{pmatrix} h_K(\nu) & h_K(\nu+1) & \cdots & h_K(K) \\ h_K(\nu+1) & & & 0 \\ \cdot & & & \cdot \\ \cdot & & h_K(K) & \cdot \\ \cdot & & & \cdot \\ h_K(K) & 0 & \cdots & 0 \end{pmatrix}, \tag{1.9}$$

where $h_K(k) \triangleq 0$ for $k < 0$.

The singular value decomposition of $\mathbf{H}_{\nu,K}$ may be expressed as

$$\mathbf{H}_{\nu,K} = U\Sigma V^*, \tag{1.10}$$

where U, V are unitary matrices, and Σ is a diagonal matrix with nonnegative diagonal elements $\sigma_0, \ldots, \sigma_{K-\nu}$ arranged in order of decreasing magnitude [202]. These elements of Σ are called the *singular values* of $\mathbf{H}_{\nu,K}$. (Note that it is customary to number the singular values from 1 rather than 0. Our choice is made to simplify notation. Also, we refer to σ_n as the n^{th} singular value, although it is the $(n+1)$st element of the sequence.) The left and right *singular vectors* corresponding to σ_n are the n^{th} columns of U and V, respectively, and we denote them by

$$\underline{U}_n \triangleq \big(u_n(0), \ldots, u_n(K-\nu)\big)^T,$$
$$\underline{V}_n \triangleq \big(v_n(0), \ldots, v_n(K-\nu)\big)^T.$$

If σ_n is not a simple singular value, then $\underline{U}_n$ and $\underline{V}_n$ are not unique, but this does not matter in the theorem below.

When the impulse-response is real, $\mathbf{H}_{\nu,K}$ is a real symmetric matrix, and in this case $\sigma_n = |\lambda_n|$, where λ_n is the n^{th} eigenvalue of $\mathbf{H}_{\nu,K}$ by magnitude ($|\lambda_0| \geq |\lambda_1| \geq \cdots \geq |\lambda_{K-\nu}|$). Moreover, in this case one may assume

$$\underline{V}_n = \underline{U}_n \operatorname{sign}(\lambda_n).$$

Thus in the case of a real impulse-response (i.e. for real symmetric matrices), each singular vector is also an eigenvector and vice versa.

The basic result on which the Hankel-norm methods are based is the following [72].

Theorem 1.36. H_K has a unique best Chebyshev approximation $\tilde{H}^*$ out of $\tilde{\mathcal{H}}_{\hat{n}_b,\hat{n}_a}$, and the error function $E(z) = H_K(z) - \tilde{H}^*(z)$ is an allpass filter whose gain at each frequency is equal to the $\hat{n}_a{}^{th}$ singular value of the Hankel matrix $\mathbf{H}_{\nu,K}$, i.e.,

$$\left\| H_K - \tilde{H}^* \right\|_\infty = \sigma_{\hat{n}_a},$$

where $\sigma_{\hat{n}_a} \triangleq 0$ for $\hat{n}_a > K - \nu$. $\tilde{H}^*$ is given by

$$\tilde{H}^*(z) = H_K(z) - \sigma_{\hat{n}_a} z^{-\nu} \frac{U_{\hat{n}_a}(z)}{V_{\hat{n}_a}(z^{-1})}, \tag{1.11}$$

where $U_{\hat{n}_a}(z)$ and $V_{\hat{n}_a}(z)$ are formed from the $\hat{n}_a{}^{th}$ singular vectors of $\mathbf{H}_{\nu,K}$ as

$$U_{\hat{n}_a}(z) \triangleq \sum_{n=0}^{K-\nu} u_{\hat{n}_a}(n) z^{-n}, \qquad V_{\hat{n}_a}(z) \triangleq \sum_{n=0}^{K-\nu} v_{\hat{n}_a}(n) z^{-n}.$$

This remarkable theorem implies that every stable linear system (even infinite-order) admits a decomposition into the sum of a noncausal filter from the class $\tilde{\mathcal{H}}_{\hat{n}_b,\hat{n}_a}$ plus an allpass filter, i.e.,

$$H_K(z) = \tilde{H}^*(z) + \sigma_{\hat{n}_a} z^{-\nu} \frac{U_{\hat{n}_a}(z)}{V_{\hat{n}_a}(z^{-1})}.$$

In the proof of the theorem, this equation follows immediately from taking the z-transform of the equation $\mathbf{H}_{\nu,K} \underline{V}_{\hat{n}_a} = \sigma_{\hat{n}_a} \underline{U}_{\hat{n}_a}$, which follows from (1.10). (See Appendix B.) What is nontrivial to show, however, is that the number of poles of $\tilde{H}^*$ inside the unit circle is at most $\hat{n}_a$. This key step was apparently first taken by Takagi [67] using a result of Schur.

For systems having a real impulse-response, the decomposition may be written

$$H_K(z) = \tilde{H}^*(z) + \lambda_{\hat{n}_a} z^{-\nu} \frac{V_{\hat{n}_a}(z)}{V_{\hat{n}_a}(z^{-1})},$$

where $V_{\hat{n}_a}(z)$ is formed from the eigenvector $\underline{V}_{\hat{n}_a}$ as above.

1.5.3. The CF algorithm

Given a finite-length impulse response $h_K(n)$, the CF method consists of the following steps. For simplicity, we assume in this description that $h_K(n)$ is real.

(1) Compute the $\hat{n}_a + 1$ smallest and the $\hat{n}_a + 1$ largest eigenvalues of the symmetric Hankel matrix $\mathbf{H}_{\nu,K}$ of (1.9). This can be accomplished by tridiagonal reduction followed by Sturm sequencing, and routines are provided for this in EISPACK [199, subroutines Tred1 and Tridib].

(2) Order the eigenvalues so that $|\lambda_0| \geq |\lambda_1| \geq \ldots \geq |\lambda_{K-\nu}|$, *find* $\lambda_{\hat{n}_a}$, *and compute a corresponding eigenvector* $\underline{V}_{\hat{n}_a}$. This can be done rapidly by inverse iteration [199, subroutines Tinvit and Trbak1].

(3) Evaluate the frequency-response of the optimum (noncausal) Chebyshev approximation (1.11) at $L \gg \hat{n}_b + \hat{n}_a + 1$ *equally spaced points along the unit circle,*

$$\tilde{H}^*(e^{j\omega_k}) = H_K(e^{j\omega_k}) - \lambda_{\hat{n}_a} e^{-j\nu\omega_k} \frac{V_{\hat{n}_a}(e^{j\omega_k})}{V_{\hat{n}_a}(e^{-j\omega_k})}$$

$$\omega_k = \frac{2\pi k}{L}, \quad k = 0, 1, \ldots, L-1 .$$

It is preferable to choose L equal to a power of 2 to allow the use of the Fast Fourier Transform (FFT) for this and the next step. Note that since $h_K(n)$ is real, $\tilde{H}^*(e^{j\omega_k}) = \overline{\tilde{H}^*(e^{-j\omega_k})}$, so that only $L/2+1$ values need to be computed.

(4) Inverse Fourier transform $\tilde{H}^*(e^{j\omega_k})$ *to obtain the impulse response of the extended rational Chebyshev approximation,*

$$\tilde{h}^*(n) = FFT^{-1}\{\tilde{H}^*(e^{j\omega_k})\} = \frac{1}{L}\sum_{k=0}^{L} \tilde{H}^*(e^{j\omega_k}) e^{j\omega_k n} .$$

The first $L/2$ samples, $n = 0, \ldots, L/2-1$, correspond to the causal part.

For $\nu \geq 0$ *(*$\hat{n}_b \geq \hat{n}_a - 1$*):*

(5) Window $\tilde{h}^*$, *selecting the causal part, to obtain the impulse response of the Hankel-norm approximation,*

$$\hat{h}^{(CF)}(n) = \begin{cases} \tilde{h}^*(n), & n = 0, \ldots, L/2-1 \\ 0, & n = L/2, \ldots, L-1 . \end{cases}$$

(6) Convert the nonparametric impulse response $\hat{h}^{(CF)}$ *to parametric form* $\{\hat{a}_i, \hat{b}_j\}$, $i = 1, \ldots, \hat{n}_a$, $j = 0, \ldots, \hat{n}_b$ *by Prony's method (defined in §1.7.2).*

For $\nu < 0$ *(*$\hat{n}_b < \hat{n}_a - 1$*):*

(5′) Window $\tilde{h}^*$ *as*

$$h^{(CF)}(n) = \begin{cases} \tilde{h}^*(n+\nu), & n = 0, \ldots, L/2-1 \\ 0, & n = L/2, \ldots, L-1 . \end{cases}$$

(6′) Convert the nonparametric impulse response $h^{(CF)}$ *to parametric form* $\{\hat{a}_i, c_j\}$, $i = 1, \ldots, \hat{n}_a$, $j = 0, \ldots, \hat{n}_a - 1$ *by Prony's method, and set* $\hat{b}_j = c_{j-\nu}, j = 0, \ldots, \hat{n}_b$.

1.5.4. Practical Considerations

The CF algorithm is defined on the basis of a prescribed order $(\hat{n}_b, \hat{n}_a)$, and in step 2 above, an error measure $|\lambda_{\hat{n}_a}|$ associated with this order is revealed. An alternative is to prescribe only the difference between the number of poles and zeros (ν), and then decide on the final order after the eigenvalues of $\mathbf{H}_{\nu,K}$ have been inspected. This alternative can lead to the most cost-effective filter designs. For many desired filters, the sequence $\{|\lambda_k|\}_0^{K-\nu}$ drops sharply in magnitude over some small interval, and values of $\hat{n}_a$ in this vicinity give efficient designs in terms of order versus error.

A related consideration is that one should ensure $|\lambda_{\hat{n}_a}| < |\lambda_{\hat{n}_a-1}|$, since otherwise a degeneracy will occur in which $\tilde{H}^*$ has fewer than $\hat{n}_a$ stable poles (in the quotient of (1.11), poles and zeros coalesce on the unit circle). This problem often comes up when $H_K(z)$ is an even function ($h_K(n) = 0$ for n odd) or is an odd function ($h_K(n) = 0$ for n even). It is easily circumvented by taking $(\hat{n}_b, \hat{n}_a)$ of the form $(odd, even)$ if H_K is even, and $(even, even)$ if H_K is odd [70]. There are also instances in which $\tilde{H}^*$ has reduced order due to the extreme elements of the eigenvector being zero ($v_{\hat{n}_a}(0) = v_{\hat{n}_a}(K-\nu) = 0$). For a complete treatment of possible degeneracies, see [67,35].

Due to the sampling of the frequency axis inherent in the FFT, the nonparametric impulse response obtained from $\tilde{H}^*(e^{j\omega})$ is really proportional to $\sum_{l=-\infty}^{\infty} \tilde{h}^*(n + lL)$. Therefore, in steps 3 and 4, it is necessary that the FFT size L be sufficiently large that time-aliasing is negligible. Since the poles of $\tilde{H}^*$ do not lie on the unit circle, increasing L sufficiently will reduce the time-aliasing error to any desired level. If d is the smallest distance from a pole of $\tilde{H}^*$ to the unit circle, then we desire $(1-d)^{\frac{L}{2}} \approx 0$.

Since the pole radii are not known in advance, it is useful to estimate the amount of time-aliasing after the fact by means of the formula

$$\mu_{ta} \triangleq \frac{L}{m+1} \frac{\sum_{n=L/2}^{L/2+m} \tilde{h}^{*2}(n)}{\sum_{n=0}^{L-1} \tilde{h}^{*2}(n)}$$

where m is a positive integer less than $L/2$. (The value $m = L/16$ works well.) This is a normalized ratio of the energy where zero is expected and the total energy. We have $0 \leq \mu_{ta} \leq 1$. When $\mu_{ta} \approx 0$, the amount of time aliasing is negligible.

In step 6, if the eigenvector is numerically accurate, and if μ_{ta} is small, then $\hat{h}^{(CF)}$ is by construction the impulse response of an $\hat{n}_a$-pole $\hat{n}_b$-zero rational filter (and simliarly for $h^{(CF)}$ in step (6′)). In this situation, it does not matter very much what norm is minimized in obtaining the parametric form of the filter. For this purpose we have chosen Prony's method [8,61,186], in which the A and B coefficients are obtained separately by solving two

systems of Toeplitz equations of order $\hat{n}_a$ and $\hat{n}_b + 1$ respectively. Code for the solution of Toeplitz linear equations may be found in [206].

Note that steps 3–6 perform the spectral factorization needed to select the causal part of $\tilde{H}^*(z)$. This approach can be applied to any spectral factorization problem where the number of poles and zeros of the causal part is known in advance (see §1.9.2). An alternate method for fast spectral factorization (based on the FFT and properties of the ramp cepstrum) has been proposed by Henrici [178] and was used in [37]; however, Henrici's method suffers from time-aliasing generated by zeros near the unit circle in addition to that due to poles. Our method is only sensitive to poles near the unit circle.

1.5.5. Weighted CF Approximation

It is possible to introduce a limited *complex weighting function* on the frequency-response error of the CF approximation. Let the desired weight function be $W(e^{j\omega})$, where $W(z)$ is a low-order rational transfer function with M zeros and N poles. Then a weighted approximation to $H(e^{j\omega})$, having $\hat{n}_b$ zeros and $\hat{n}_a$ poles, is found by the following steps.

(1) Divide by the desired weighting to produce $H_W(e^{j\omega}) \triangleq H(e^{j\omega})/W(e^{j\omega})$.

(2) Apply the CF method to obtain $\hat{H}_W$, an approximation to $H_W(e^{j\omega})$ consisting of $\hat{n}_b - M$ zeros and $\hat{n}_a - N$ poles.

(3) Multiply $\hat{H}_W(z)$ by $W(z)$ to obtain $\hat{H}(z)$, the final type $(\hat{n}_b, \hat{n}_a)$ weighted approximation.

Since the error $E_W(e^{j\omega}) = H_W(e^{j\omega}) - \hat{H}_W(e^{j\omega})$ is uniformly weighted by the CF algorithm, the final error $E(e^{j\omega}) = H(e^{j\omega}) - \hat{H}(e^{j\omega}) = W(e^{j\omega})E_W(e^{j\omega})$ is weighted by the complex function $W(e^{j\omega})$.

One use of such a weighting strategy is in preserving deep spectral nulls, which tend to fill in (on a dB scale) when the impulse-response corresponding to $H(z)$ is windowed. Similarly, any "known" rational structure may be factored out, leaving the CF algorithm to "fine-tune" the frequency-response fit in a near-minimax sense.

Another reason to consider rational weighting is that the CF algorithm performs particularly well on smooth frequency-response functions. If a low-order rational modification exists which significantly reduces the spectral dynamic range of the desired frequency-response, it will most probably result in a more efficient design.

A completely different approach to installing an error weighting feature is given in §1.9.1.

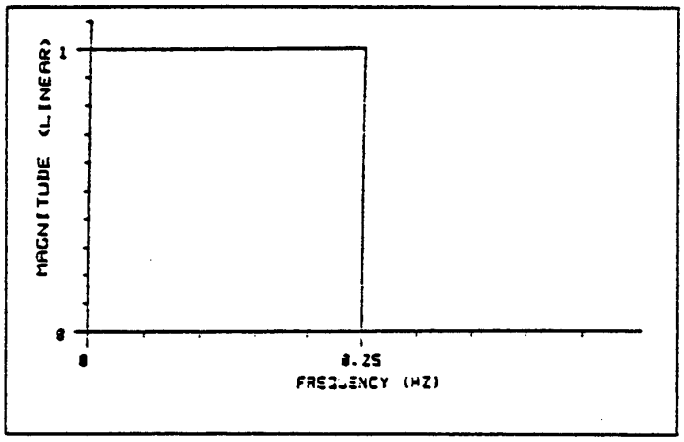

Figure 1.1. Ideal lowpass filter magnitude frequency response.

1.5.6. Computed Examples

In this section, the CF algorithm is used for recursive lowpass filter design. The first example is the design of a *minimum-phase* recursive lowpass, and the second example is for *linear-phase*.

While classical methods are typically best suited for IIR lowpass design [196], the nature of the approximation at a discontinuity in the frequency domain gives an important benchmark in the behavior of any general filter-design algorithm. A spectral discontinuity is somewhat pessimal for the CF method, however, for the method is most effective with a desired frequency response which is smooth. We also use this example to illustrate in detail the various steps of the CF algorithm.

Minimum-Phase Recursive Lowpass Filter Design

In Fig. 1.1 is shown the ideal lowpass filter magnitude frequency response for a cutoff frequency of one-fourth the sampling rate.

In order to obtain a practical "ideal" minimum-phase impulse response corresponding to Fig. 1.1, we begin with the function

$$H(\omega) = \begin{cases} 0\ \text{dB}, & 0 \le \omega < \pi/2 \\ -30\ \text{dB}, & \omega = \pi/2 \\ -60\ \text{dB}, & \pi/2 < \omega \le \pi \end{cases}$$

as the desired magnitude frequency response. Thus, we replace the ideal transfer characteristic by one which steps down 60dB in the frequency domain. This function is then

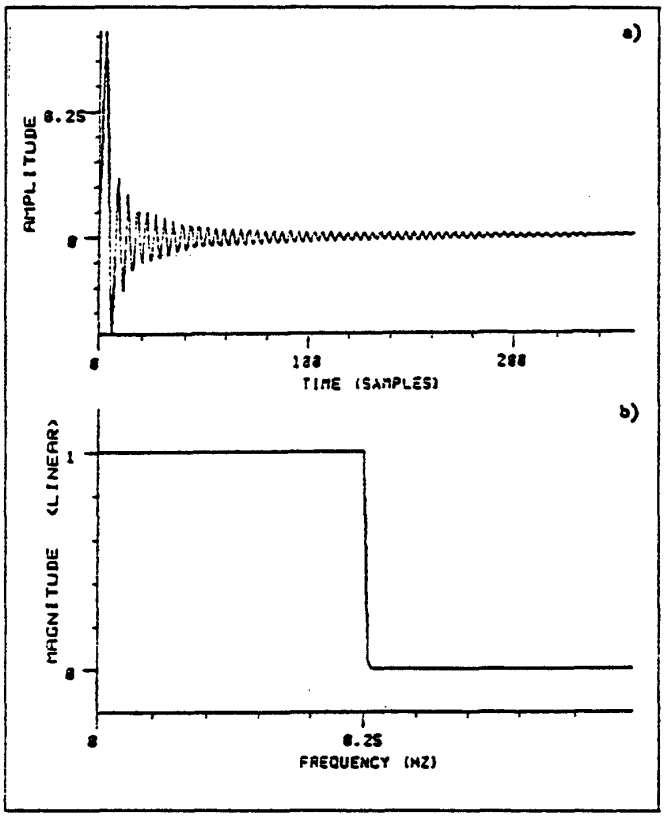

Figure 1.2. Minimum-phase ideal lowpass filter obtained by windowing the real cepstrum of the impulse response.

a) Impulse response.

b) Magnitude frequency response.

sampled at equally spaced frequencies. For this example, 129 points are used, corresponding to an FFT of length 256. Next, the real-cepstrum method [169,191] is used to create the minimum-phase complex spectrum exhibiting this magnitude curve. The use of two samples rather than one in the discontinuity serves to reduce time-aliasing. The inverse FFT of the spectrum so obtained yields the initial desired impulse response, and this is shown in Fig. 1.2a. The magnitude spectrum of Fig. 1.2a is shown in Fig. 1.2b, illustrating the fact that little distortion is incurred at the sample points during the conversion from zero-phase to minimum-phase.

The next step is to window the "ideal" impulse response to the length K desired for use in the CF algorithm. In this case, we choose $K = 79$. The method selected for this

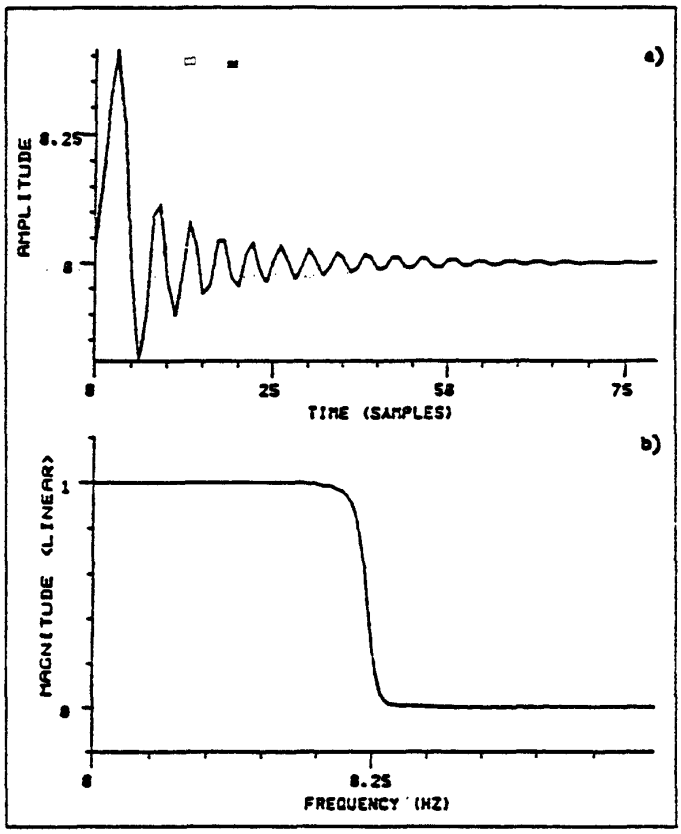

Figure 1.3. Hamming-windowed minimum-phase lowpass filter.

a) Impulse response.
b) Magnitude frequency response.

windowing consists of multiplying the function of Fig. 1.2a by half of a Hamming window. The resulting impulse response and corresponding magnitude spectrum are shown in Fig. 1.3.

We now use the CF method to obtain a 7-pole, 6-zero digital filter which approximates the filter of Fig. 1.3. First, the 80 by 80 Hankel matrix is formed, and its 16 extreme eigenvalues are computed. The magnitudes of all 80 eigenvalues are plotted in Fig. 1.4. The seventh eigenvalue modulus is $|\lambda_7| = 0.019$. This number provides the magnitude of the allpass error in the optimum noncausal Chebyshev filter, and equals the Hankel norm of the final approximation error. Thus we expect about two percent error in the magnitude of the passband. The internal FFT size was chosen to be $L = 512$.

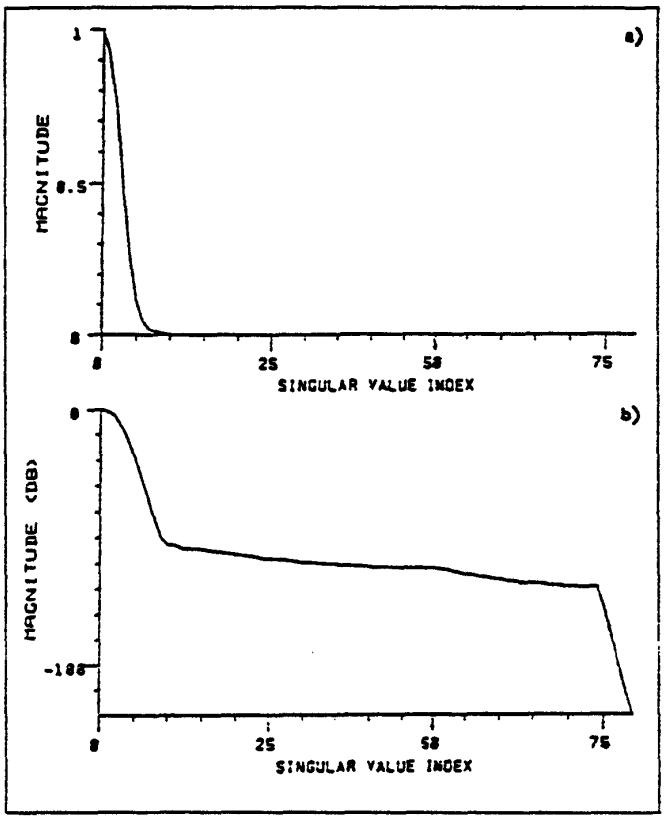

Figure 1.4. Singular values of Hankel matrix $\mathbf{H}_{0,80}$ of windowed minimum-phase filter.

a) Linear scale.

b) DB scale.

Figure 1.5a shows the magnitude error $|H_K(e^{j\omega_k})| - |\tilde{H}^*(e^{j\omega_k})|$ in the optimum extended rational Chebyshev approximation. When the noncausal part of $\tilde{h}^*$ is dropped to obtain $\hat{h}^{(CF)}$, the magnitude error becomes that shown in Fig. 1.5b. Note how slightly the magnitude error for the optimum Hankel approximation extends past the bounds for the optimum Chebyshev error.

The causal impulse response $\hat{h}^{(CF)}$ of the optimum Hankel approximation is finally converted to a set of recursive filter coefficients, via Prony's method applied to the first 80 samples of $\hat{h}^{(CF)}$. The error due to this conversion is $\| \hat{h}^{(CF)} - h^{(CF)} \|_2 = 0.00012$, where $\hat{h}^{(CF)}$ denotes the impulse response obtained nonparametrically, and $h^{(CF)}$ denotes the impulse response of the filter computed by Prony's method. (The norm is measured

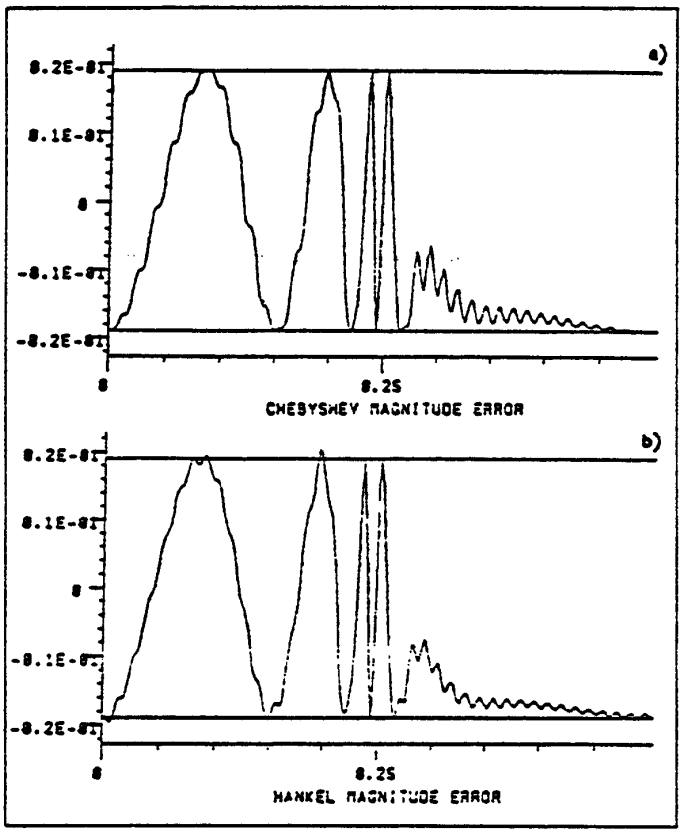

Figure 1.5. Magnitude frequency response error.
a) Optimum Chebyshev Approximation.
b) CF Approximation (Optimum Hankel norm).

over the first 512 samples of each impulse response.) The good match by Prony's method indicates numerical success of the preceding steps, and that L is sufficiently large.

The final frequency response, overlayed with the desired frequency response, is shown in Fig. 1.6a. Notice that the error is nearly equal ripple at about two percent in the passband, as expected.

The filter obtained using *equation-error* minimization on the same target spectrum $H_K(e^{j\omega})$ as for the CF method is shown in Fig. 1.6b. We chose the equation-error method as a standard for comparison because algorithms in this class (such as Prony's method) seem to be the only other way to obtain unique rational approximations which fit both phase and magnitude and which do not suffer from the possibility of convergence to suboptimal solutions. The equation-error algorithm used is a fast version of the one outlined in [77],

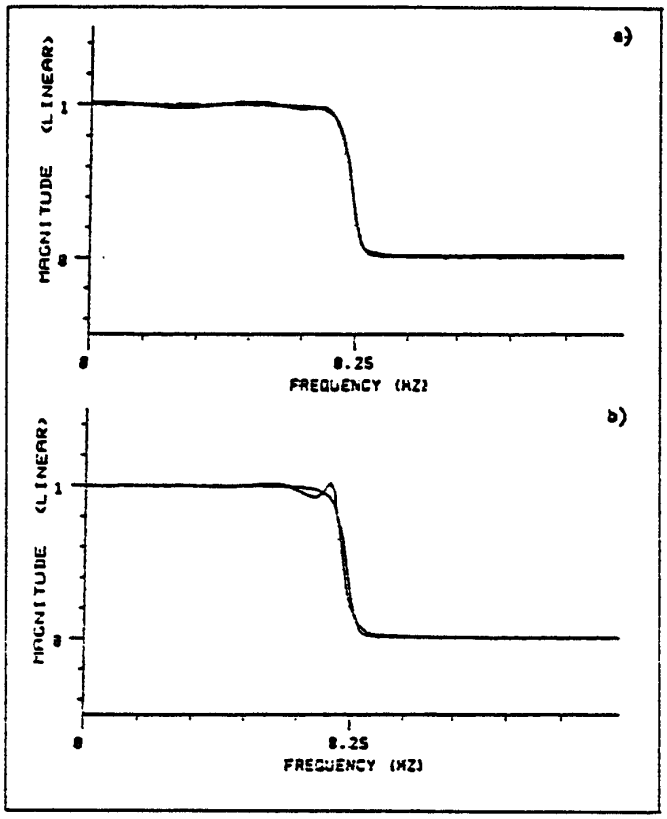

Figure 1.6. Magnitude frequency response fit for the minimum-phase case.

a) CF.

b) Equation error.

Note that there is more error near the passband edge with equation-error minimization, due to the presence of poles nearby. (The equation error is defined as $A(e^{j\omega})(\hat{H}(e^{j\omega}) - B(e^{j\omega})/A(e^{j\omega}))$, which gets weighted toward zero near roots of $A(z)$.) On a Foonly F2 computer, in single precision floating point, the equation-error solution required approximately 2.5 seconds of CPU time, while the CF algorithm took approximately 70 seconds (with 60 seconds spent in the tri-diagonalization of the 80 by 80 Hankel matrix).

Although the CF method does not attempt to minimize any kind of log-spectral error, it is often the case in filter design that such an error is most appropriate. For completeness we show the CF and equation-error magnitude fits on a dB vertical scale in Fig. 1.7. On a log vertical scale, the equation-error method may be preferable to the CF method due to better stop-band rejection.

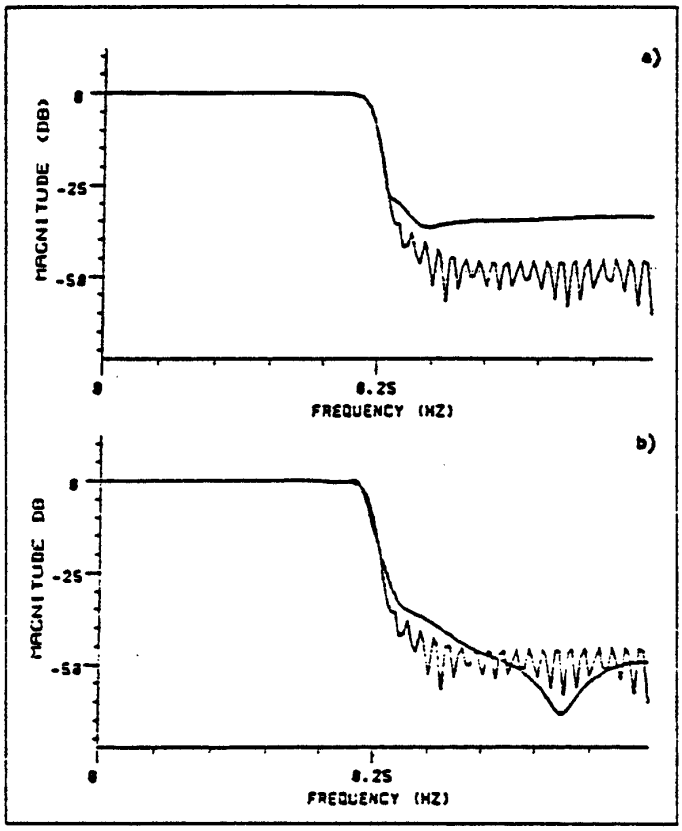

Figure 1.7. Magnitude frequency response fit for the minimum-phase case (dB scale).

a) CF.

b) Equation error.

Figure 1.8 compares the pole-zero plots for the CF and equation-error methods. The large difference between the two plots suggests that use of the equation-error solution as an initial guess for a gradient-descent algorithm, which explicitly minimizes $|| H(e^{j\omega}) - \hat{H}(e^{j\omega}) ||$ with respect to pole positions, may not be effective in general.

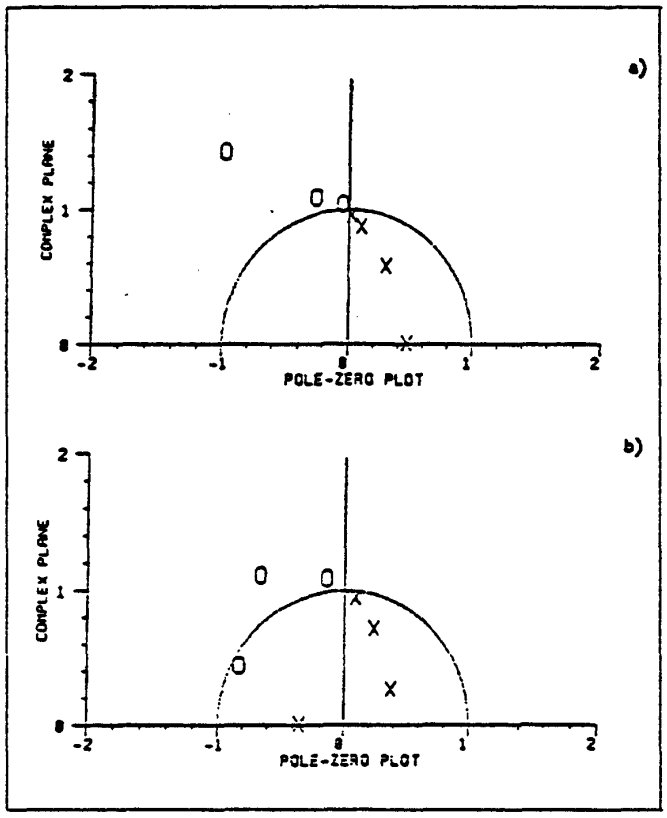

Figure 1.8. Poles and zeros for the minimum-phase case (X = pole, O = zero).

a) CF.

b) Equation error.

Linear-Phase Recursive Lowpass Filter Design

In this example, the goal is to design a *linear phase* recursive lowpass filter. Since the CF method requires a finite impulse response as a starting point, it is good to have an initial target impulse response which is optimal in some sense. The Parks-McClellan algorithm [52,196] provides optimal FIR filters in the sense that the Chebyshev norm of the spectral magnitude error is minimized over filters with exactly linear phase. Since the CF method takes an FIR filter into an IIR filter, preserving the spectrum in a nearly optimal Chebyshev sense, the Parks-McClellan algorithm provides a good initial condition for this problem. Furthermore, our experience indicates that the amount of computational effort in the two methods is comparable, with the CF algorithm being somewhat more expensive. Thus the Parks-McClellan algorithm is a well-matched supplement to the CF algorithm.

We begin with an optimum FIR lowpass filter of length $K = 21$. The passband ranges from $f = 0$ to one-tenth the sampling rate $f = f_s/10$, and the stopband is defined from $f = f_s/5$ to $f = f_s/2$. The singular values of the Hankel matrix for this problem are plotted in Fig. 1.9. In Fig. 1.10, a comparision between the CF method and the equation-error method is given for the case of a 7-pole, 6-zero approximation to the optimum order 20 FIR filter. The FFT size used is $L = 256$. Figure 1.11 gives the same comparison on a dB vertical scale. The impulse response fit for the two methods is shown in Fig. 1.12, and the poles and zeros are displayed in Fig. 1.13. In this example, the CF method clearly out-performs the equation-error method.

The CF method, unlike equation error, does not in principle suffer from desiring a non-minimum-phase impulse response. In practice, however, there can be problems. The most obvious reason to prefer minimum-phase designs is that the order K of the target impulse response h_K must in general be larger in the non-minimum-phase case, for a given spectral magnitude resolution. Also, in the specific instance of linear-phase design, the numerical behavior of the CF algorithm is poor. This is why only a length 21 target impulse response was selected (with the CF algorithm implemented in 36-bit single precision). Nevertheless, high quality recursive linear phase filter design, as illustrated in this example, is possible with the CF algorithm, although high-precision calculations are called for. We do not fully understand the nature of the numerical difficulty, and perhaps the problem can be recast to avoid it.

It should be mentioned that the equation-error method would perform better if a fortuitous amount of negative bulk delay were added to the desired impulse response [278]. In other words, the target impulse-response should be shifted left to make it noncausal. At some point the design would go unstable and phase linearity would be lost in reflecting the unstable pole back inside the unit circle. However, there would be some optimum amount of time shift for the impulse response.

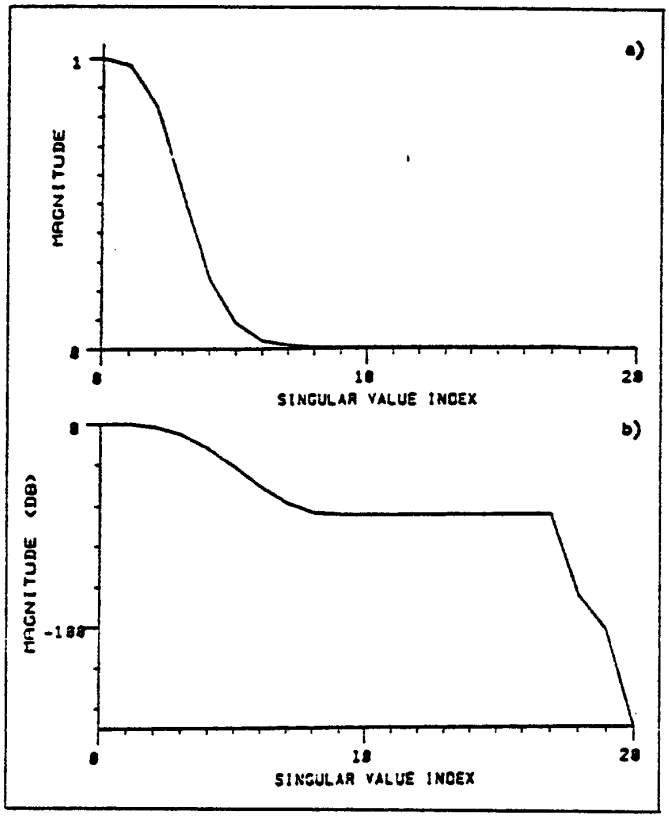

Figure 1.9. Singular values of the Hankel matrix $\mathbf{H}_{0,21}$ corresponding to the optimum FIR linear-phase impulse response.
a) Linear scale.
b) DB scale.

This marks the end of discussion of methods which attempt to solve problem $\hat{H}^*$. The remaining methods solve a modified version of the problem.

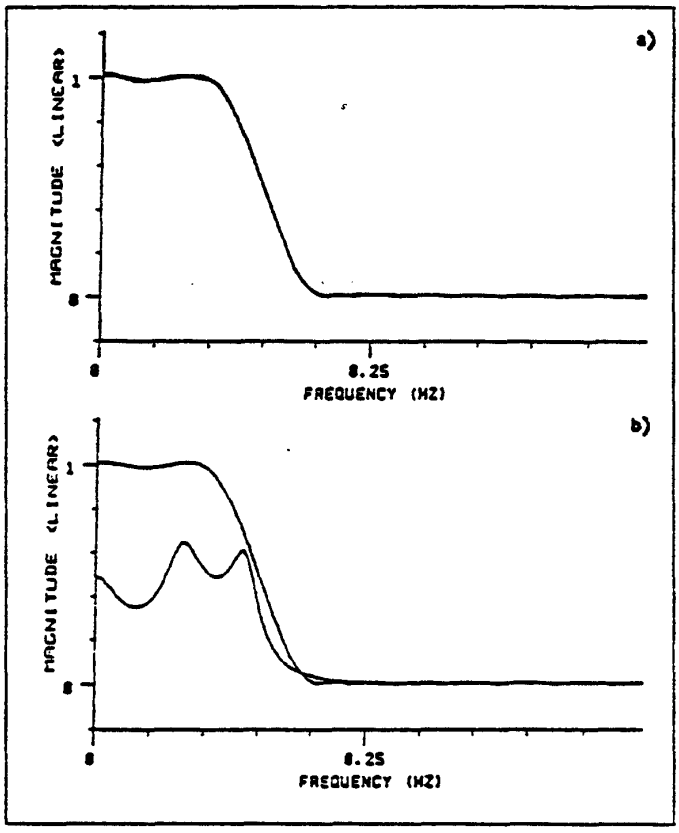

Figure 1.10. Magnitude frequency response fit for the linear-phase case.

a) CF.

b) Equation error.

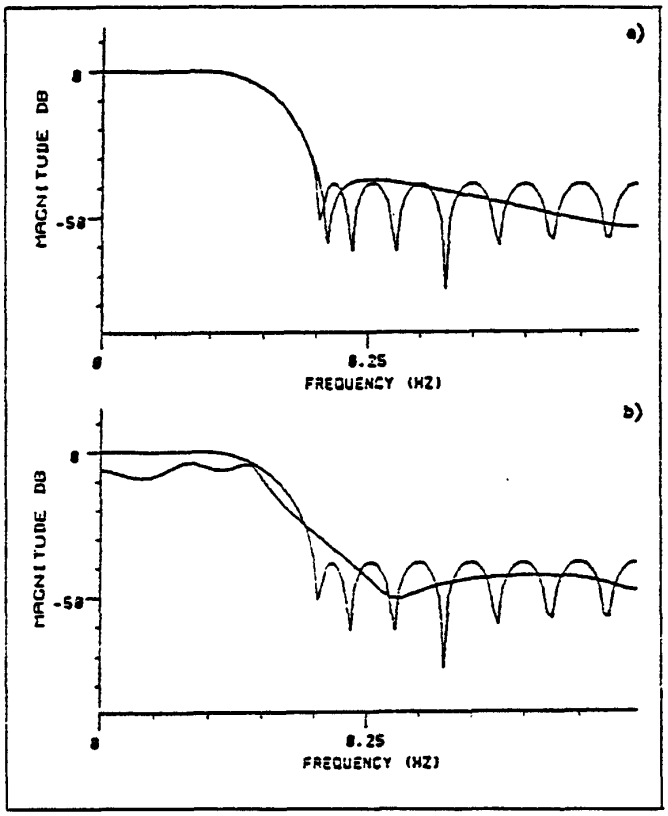

Figure 1.11. Magnitude frequency response fit for the linear-phase case (dB scale).

a) CF.

b) Equation error.

1.6. Minimization of the L^2 Ratio-Error Norm

As we have seen, problem $\hat{H}^*$ is a difficult problem in its general form. It is not always necessary to obtain all of the features provided by a solution to problem $\hat{H}^*$. For example, there are applications in which the phase of the approximation is not important. Also, it is not always essential to have both poles and zeros in the filter. For these situations, it is advantageous to reformulate the problem in order to relax the requirements and facilitate computational methods.

A class of techniques for fitting an *only poles* $\hat{H}(z) = 1/\hat{A}(z)$ to a desired $H(z)$, ignoring the phase of the approximation, can be classified as *ratio-error* methods. The L^2 norm is typically used for this error, yielding *linear prediction methods* [186,185]. The error measure

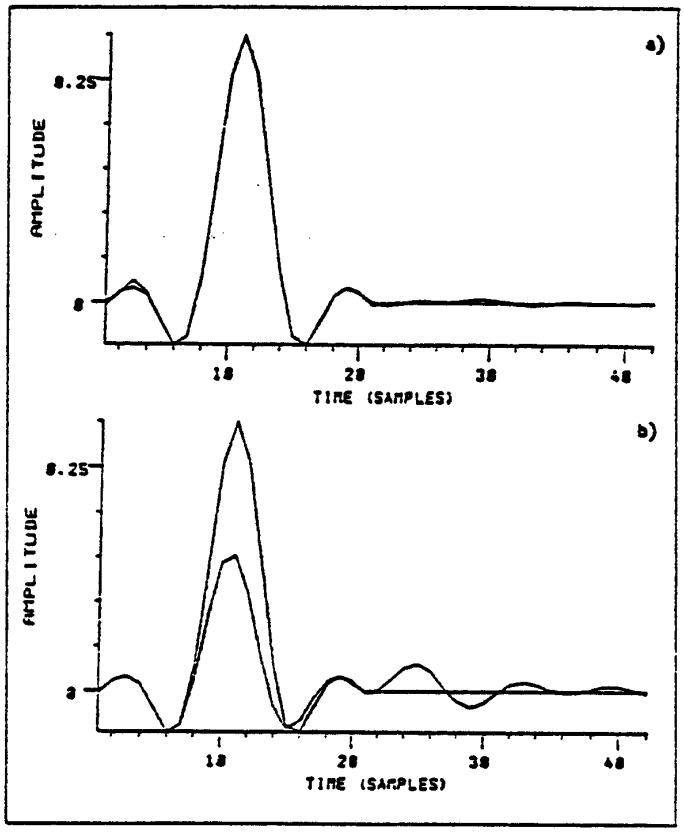

Figure 1.12. Impulse response fit for the linear-phase case.
a) CF.
b) Equation error.

to be minimized is

$$J_R(\hat{\mathbf{a}}) \triangleq \left\| \frac{H(e^{j\omega})}{\hat{H}(e^{j\omega})} \right\|_2^2 \triangleq \left\| \hat{A}(e^{j\omega}) H(e^{j\omega}) \right\|_2^2 \triangleq \left\| \hat{E}(e^{j\omega}) \right\|_2^2 = \| \hat{e}(n) \|_2^2 = \sum_{n=0}^{\infty} \hat{e}^2(n). \quad (1.12)$$

1.6.1. The Autocorrelation Method

Since

$$\hat{e}(n) = \hat{a} * h(n) = h(n) + \hat{a}_1 h(n-1) + \hat{a}_2 h(n-2) + \cdots + \hat{a}_{\hat{n}_a} h(n-\hat{n}_a),$$

the time-domain ratio-error $\hat{e}(n)$ can be viewed as the error in linearly predicting $h(n)$ from the past $\hat{n}_a$ samples of $h(n)$. Since the norm sums all the squared prediction errors $\hat{e}(n)$, this amounts to the *autocorrelation method* of linear prediction [186,185].

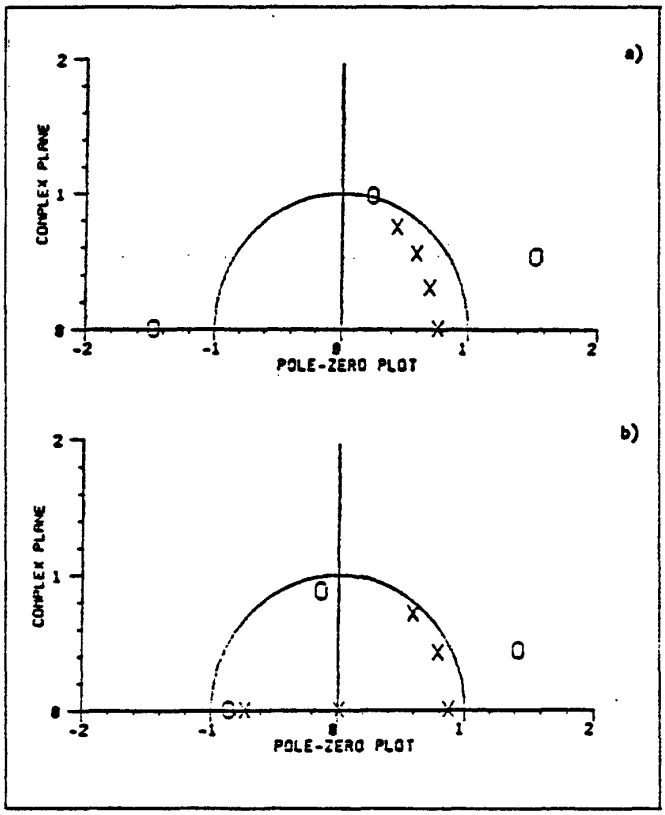

Figure 1.13. Poles and zeros for the linear-phase case (X = pole, O = zero).

a) CF.

b) Equation error.

1.6.2. The Covariance Method

Suppose $H(z)$ is a sum of complex exponentials,

$$H(z) = \sum_{k=1}^{\hat{n}_a} \frac{R_k}{1 - p_k z^{-1}} = \frac{B(z)}{A(z)},$$

where $B(z)$ is of order $\hat{n}_a - 1$ and $A(z)$ is of order $\hat{n}_a$, and that the problem is to find the poles p_k of $A(z)$ given either $H(e^{j\omega}), -\pi < \omega \leq \pi$, or $h(n), n = 0, 1, 2, \ldots$. This is a basic problem in *system identification* which has its roots in a problem posed in 1795 by Prony

[53]. Let $G(z) = 1/A(z)$, and denote the corresponding impulse-response by $g(n)$. Then $H(z) = B(z)G(z)$, and (1.12) becomes

$$J_R(\hat{A}) \triangleq \left\| \hat{A}(e^{j\omega})H(e^{j\omega}) \right\|_2^2 \triangleq \left\| \hat{A}(e^{j\omega})G(e^{j\omega})B(e^{j\omega}) \right\|_2^2. \tag{1.13}$$

At the answer $\hat{A} = A$,

$$J_R(A) = \left\| B(e^{j\omega}) \right\|_2^2 = \sum_{n=0}^{\infty} b_n^2 = \sum_{n=0}^{\hat{n}_a - 1} b_n^2. \tag{1.14}$$

We see that $\hat{A}^*$ will not generally be equal to A since other values of $\hat{A}$ may reduce J_R below $\| B \|_2^2$. However, if the time summation in (1.14) is taken from $n = \hat{n}_a$, then $J_R(A) = 0$, i.e., the error measure goes to zero at the true solution. I.e., we minimize

$$J_{R_e}(\hat{A}) \triangleq \sum_{n=\hat{n}_a}^{\infty} (\hat{a} * h)^2(n).$$

It is easy to show that if $B \not\equiv 0$, and if $A(z)$ is not constant, then J_{R_e} can be zero *only* at the true solution. This is the *covariance method* of linear prediction, having the well-known property of being able to exactly model a sum of decaying complex exponentials [186]. A class of methods which computes the residues R_k in addition to the poles is called *Prony's Method* (discussed in §1.7.2).

1.6.3. Kopec's Method

A nice method based on linear prediction for finding poles *and zeros* is one apparently first proposed by Kopec [186]. Kopec's method consists of the following steps:

- Given $H(z)$, compute an allpole model $1/\hat{A}(z)$ by minimizing $\| \hat{A}(e^{j\omega})H(e^{j\omega}) \|_2$.
- Compute the error spectrum $\hat{E}(e^{j\omega}) \triangleq \hat{A}(e^{j\omega})H(e^{j\omega})$.
- Compute an allpole model $1/\hat{B}(z)$ for $\hat{E}^{-1}(e^{j\omega})$ by minimizing

$$\left\| \hat{B}(e^{j\omega})\hat{E}^{-1}(e^{j\omega}) \right\|_2 = \left\| \frac{\hat{B}(e^{j\omega})}{\hat{A}(e^{j\omega})} H^{-1}(e^{j\omega}) \right\|_2.$$

The basic definition of ratio error implies that the model will tend to fit the spectral envelope of the desired frequency-response. Since the first step of Kopec's method captures the spectral envelope, the "nulls" and "valleys" are in some sense "saved" for the next step which computes zeros. When computing the zeros, these "dips" in the spectrum have become "peaks". Thus, in Kopec's method, the poles model the upper spectral envelope, while the zeros model the lower spectral envelope.

1.7. Minimization of the L^2 Equation-Error Norm

We have seen that minimizing ratio error leads to a simplified filter-design formulation at the price of (1) losing the ability to match phase, (2) having to accept a fit to the spectral envelope, and (3) allowing only poles in the approximation. A different simplification of problem $\hat{H}^*$, which allows pole-zero modeling and simultaneous matching of spectral phase and magnitude, gives the *equation-error* methods [8,7,14,20,54,68,9,77]. Most of these can be classified as variations of *Prony's* method [53,186]. Equation error is used almost exclusively in the *system identification* context [99, 95] (see Chapter 2).

In this case, we minimize the L^2 norm of the *equation-error*

$$J_E(\hat{\theta}) \triangleq \left\| \hat{A}(e^{j\omega})H(e^{j\omega}) - \hat{B}(e^{j\omega}) \right\|_2 . \tag{1.15}$$

Since equation-error is *linear in the parameters*, it presents a very simple optimization problem, resulting in a set of $\hat{n}_b + \hat{n}_a + 1$ linear equations to solve. A complete derivation is given in Chapter 2.

Note, however, that (1.15) can be expressed as

$$J_E(\hat{\theta}) = \left\| \left| \hat{A}(e^{j\omega}) \right| \left| H(e^{j\omega}) - \hat{H}(e^{j\omega}) \right| \right\|_2 . \tag{1.16}$$

Thus the L^2 equation-error norm is the same as the L^2 norm for problem $\hat{H}^*$, viz. $\| H(e^{j\omega}) - \hat{H}(e^{j\omega}) \|_2$, with a weight function given by the poles of the filter to be designed. Since the the poles of a good model tend toward regions of high spectral energy, or toward "corners" in the spectrum, it is evident that the equation-error criterion assigns less importance to the most prominent or structured spectral regions. On the other hand, far away from the roots of $\hat{A}(z)$, good fits to *both phase and magnitude* can be expected.

A problem with equation-error methods is that stability of the filter design is not guaranteed. When an unstable design is encountered, the standard remedy is to reflect unstable poles inside the unit circle, leaving the magnitude response unchanged while modifying the phase of the approximation in an ad hoc manner. This requires polynomial factorization to find the filter poles, which is typically much more work than the filter design.

An alternative to pole-reflection in unstable designs (which destroys the phase-response fit) is to repeat the filter design employing a *bulk delay*. This amounts to replacing $H(e^{j\omega})$ by

$$H_\tau(e^{j\omega}) \triangleq e^{-j\omega\tau} H(e^{j\omega}), \quad \tau > 0,$$

and minimizing $\| \hat{A}(e^{j\omega})H_\tau(e^{j\omega}) - \hat{B}(e^{j\omega}) \|_2$. This effectively *delays* the desired impulse response, i.e., $h_\tau(n) = h(n - \tau)$. As the bulk delay is increased, the likelihood of obtaining an unstable design decreases, for reasons discussed in the next paragraph.

Unstable designs are especially likely when $H(e^{j\omega})$ is *noncausal*. Due to the form of $\hat{H}$, the only way noncausal impulse-response components can be generated is by means of unstable poles. In the case of noncausal H, the equation error method is faced with either suffering the entire desired impulse-response for $n < 0$ as error, or moving poles outside the unit circle to obtain a nonzero approximation there. Since there are no constraints on where the poles of $\hat{H}$ can be, it is perfectly reasonable to expect unstable designs for noncausal desired frequency-response functions.

In the other direction, experience has shown that best results are obtained when $H(z)$ is *minimum phase*, i.e., when the analytic continuation of $H(e^{j\omega})$ has all zeros inside the unit circle. For a given magnitude, $|H(e^{j\omega})|$, minimum phase gives the maximum concentration of impulse-response energy near the origin $n = 0$. Consequently, the impulse-response tends to start large and decay immediately. For non-minimum phase H, the impulse-response $h(n)$ may be small for the first several samples, and the equation error method can yield very poor filters in these cases (see §1.5.6). To see why this is so, consider a desired impulse-response $h(n)$ which is zero for $n \leq \hat{n}_b$, and arbitrary thereafter. Transforming J_E^2 into the time domain yields

$$\begin{aligned} J_E^2(\hat{\theta}) &= \left\| \hat{a} * h(n) - \hat{b}(n) \right\|_2^2 \\ &= \sum_{n=0}^{\infty} \left(\hat{a} * h(n) - \hat{b}(n)\right)^2 \\ &= \sum_{n=0}^{\hat{n}_b} \hat{b}_n^2 + \sum_{n=\hat{n}_b+1}^{\infty} \left(\hat{a} * h(n)\right)^2, \end{aligned}$$

where "$*$" denotes convolution,† and the additive decomposition is due the fact that $\hat{a} * h(n) = 0$ for $n \leq \hat{n}_b$. In this case the minimum occurs for $\hat{B}(z) = 0 \Rightarrow \hat{H}(z) \equiv 0$! Clearly this is not a particularly good fit. Thus the introduction of bulk-delay to guard against unstable designs is limited by this phenomenon.

It should be emphasized that for causal minimum-phase $H(e^{j\omega})$, equation-error methods are very effective. It is simple to convert a desired magnitude response into a causal minimum-phase frequency-response by use of cepstral techniques [169], and this is highly recommended when minimizing equation error.

1.7.1. A Fast Frequency-Domain Equation-Error Method

The algorithm below is a frequency-domain equation error method. The particular implementation is a fast version of the one outlined in [77].

† Defined in Appendix E.

Given a desired spectrum $H(e^{j\omega_k})$ at equally spaced frequencies $\omega_k = 2\pi k/N, k = 0,\ldots,N-1$, with N a power of 2, it is desired to find a rational digital filter with $\hat{n}_b$ zeros and $\hat{n}_a$ poles,

$$\hat{H}(z) \triangleq \frac{\hat{B}(z)}{\hat{A}(z)} \triangleq \frac{\sum_{k=0}^{\hat{n}_b} b_k z^{-k}}{\sum_{k=0}^{\hat{n}_a} a_k z^{-k}}, \tag{1.17}$$

normalized by $a_0 = 1$, such that

$$J_E^2 = \sum_{k=0}^{N-1} \left| \hat{A}(e^{j\omega_k})H(e^{j\omega_k}) - \hat{B}(e^{j\omega_k}) \right|^2$$

is minimized.

Since J_E^2 is a quadratic form, the solution is readily obtained by equating the gradient to zero. An easier derivation follows from minimizing equation error in the time domain and making use of the orthogonality principle. This may be viewed as a system identification problem where the known input signal is an impulse, and the known output is the desired impulse response (cf. Chapter 2). A formulation employing an *arbitrary* known input is valuable for introducing complex weighting across the frequency grid, and this general form is presented. A detailed derivation appears in Chapter 2, and here only the final algorithm is given:

Given spectral output samples $Y(e^{j\omega_k})$ and input samples $U(e^{j\omega_k})$, we minimize

$$J_E^2 = \sum_{k=0}^{N-1} \left| \hat{A}(e^{j\omega_k})Y(e^{j\omega_k}) - \hat{B}(e^{j\omega_k})U(e^{j\omega_k}) \right|^2.$$

If $|U(e^{j\omega_k})|^2$ is to be used as a weighting function in the filter-design problem, then we set $Y(e^{j\omega_k}) = H(e^{j\omega_k})U(e^{j\omega_k})$.

Let $\underline{x}[n_1:n_2]$ denote the column vector determined by $x(n), n = n_1,\ldots,n_2$ filled in from top to bottom, and let $T(x[n_1:n_2])$ denote the size $n_2 - n_1 + 1$ symmetric Toeplitz matrix consisting of $\underline{x}[n_1:n_2]$ in its first column. A nonsymmetric Toeplitz matrix may be specified by its first column and row, and we use the notation $T(\underline{x}[n_1:n_2], \underline{y}^T[m_1:m_2])$ to denote the $n_2 - n_1 + 1$ by $m_2 - m_1 + 1$ Toeplitz matrix with left-most column $\underline{x}[n_1:n_2]$ and top row $\underline{y}^T[m_1:m_2]$. The inverse Fourier transform of $X(e^{j\omega_k})$ is defined as

$$x(n) = \mathrm{FFT}^{-1}\left\{X(e^{j\omega_k})\right\} \triangleq \frac{1}{N}\sum_{k=0}^{N-1} X(e^{j\omega_k})e^{j\omega_k n}.$$

The scaling by $1/N$ is optional since it has no effect on the solution. We require three

correlation functions involving U and Y,

$$R_{uu}(n) \triangleq \mathrm{FFT}^{-1}\left\{\left|U(e^{j\omega_k})\right|^2\right\}$$
$$R_{yy}(n) \triangleq \mathrm{FFT}^{-1}\left\{\left|Y(e^{j\omega_k})\right|^2\right\}$$
$$R_{yu}(n) \triangleq \mathrm{FFT}^{-1}\left\{Y(e^{j\omega_k})\overline{U(e^{j\omega_k})}\right\}$$
$$n = 0, 1, \ldots, N-1,$$

where the overbar denotes complex conjugation, and four corresponding Toeplitz matrices,

$$R_{yy} \triangleq T(R_{yy}[0:\hat{n}_a - 1])$$
$$R_{uu} \triangleq T(R_{uu}[0:\hat{n}_b])$$
$$R_{yu} \triangleq T(R_{yu}[-1:\hat{n}_b - 1], R_{yu}^T[-1:-\hat{n}_a])$$
$$R_{uy} \triangleq R_{uy}^T,$$

where negative indices are to be interpreted mod N, e.g., $r_{yu}(-1) = r_{yu}(N-1)$.

The solution is then

$$\hat{\theta}^* = \begin{pmatrix} \hat{B}^* \\ \hat{A}^* \end{pmatrix} = \begin{pmatrix} R_{uu} & R_{yu} \\ R_{uy} & R_{yy} \end{pmatrix}^{-1} \begin{pmatrix} R_{yu}[0:\hat{n}_b] \\ R_{yy}[1:\hat{n}_a] \end{pmatrix},$$

where

$$\hat{B}^* \triangleq \begin{pmatrix} \hat{b}_0^* \\ \vdots \\ \hat{b}_{\hat{n}_b}^* \end{pmatrix}, \qquad \hat{A}^* \triangleq \begin{pmatrix} \hat{a}_1^* \\ \vdots \\ \hat{a}_{\hat{n}_a}^* \end{pmatrix}.$$

1.7.2. Prony's Method

There are several variations on equation-error minimization, and some confusion in terminology exists. We use the definition of *Prony's method* given by Markel and Gray [186]. It is equivalent to "Shank's method" [61,8]. In this method, one first computes the denominator $\hat{A}^*(z)$ by minimizing

$$J_S^2(\hat{\theta}) = \sum_{n=\hat{n}_b+1}^{\infty} \left(\hat{a} * h(n) - \hat{b}(n)\right)^2$$
$$= \sum_{n=\hat{n}_b+1}^{\infty} \left(\hat{a} * h(n)\right)^2.$$

This step is equivalent to minimization of *ratio error* for the all-pole part $\hat{A}(z)$, with the first $\hat{n}_b + 1$ terms of the time-domain error sum discarded. When $\hat{n}_b = \hat{n}_a - 1$, it coincides with the covariance method of linear prediction [186,185]. This idea for finding the poles by "skipping" the influence of the zeros on the impulse-response shows up in the stochastic case under the name of *modified Yule-Walker equations* [9,92,93].

Now, Prony's method consists of next solving problem $\hat{H}^*$ in L^2 with the pre-assigned poles given by $\hat{A}^*(z)$. In other words, the numerator $\hat{B}(z)$ is found by minimizing

$$\left\| H(e^{j\omega}) - \frac{\hat{B}(e^{j\omega})}{\hat{A}^*(e^{j\omega})} \right\|_2,$$

where $\hat{A}^*(e^{j\omega})$ is now known. This hybrid method is not as sensitive to the time distribution of $h(n)$ as is the pure equation-error method. In particular, the degenerate equation-error example above (in which $\hat{H} \equiv 0$ was obtained) does not fare so badly using Prony's method.

1.7.3. The Padé-Prony Method

Another variation of Prony's method, described by Burrus and Parks in [8] consists of using *Padé* approximation to find the numerator $\hat{B}^*$ after the denominator $\hat{A}^*$ has been found as before. Thus, $\hat{B}^*$ is found by matching the first $\hat{n}_b + 1$ samples of $h(n)$, viz., $\hat{b}_n^* = \hat{a}^* * h(n), n = 0 \ldots, \hat{n}_b$. This method is faster, but does not generally give as good results as the previous version. In particular, the degenerate example $h(n) = 0, n \leq \hat{n}_b$ gives $\hat{H}^*(z) \equiv 0$ here as did pure equation error. This method has been applied in the stochastic case by Cadzow [9].

On the whole, when $H(e^{j\omega})$ is causal and minimum phase (the ideal situation for just about any stable filter-design method), the variants on equation-error minimization described in this section perform very similarly. They are all quite fast, relative to algorithms which attempt to solve problem $\hat{H}^*$, and the equation-error method given in §1.7.1 is generally fastest.

Chapter 2 is almost exclusively concerned with generalized equation-error methods. They are important because of their simplicity and robustness, and are ideally suited for time-varying signal modeling and on-line signal forecasting.

1.8. Other Choices of Error

In this section, other special error criteria are considered. After a synopsis of methods considered up to now, techniques for linear-phase filter design, linear and log power-response matching, phase-only approximation, Padé approximation, and classical analog filter design will be discussed.

1.8.1. Summary of Methods So Far

We have seen that problem $\hat{H}^*$ is difficult to solve, mainly because all norms of the frequency-response error,

$$J(\hat{\theta}) = \left\| H(e^{j\omega}) - \hat{H}(e^{j\omega}) \right\| = \left\| H(e^{j\omega}) - \frac{\hat{B}(e^{j\omega})}{\hat{A}(e^{j\omega})} \right\|$$

lead to a highly nonlinear optimization problem with respect to $\hat{A}$. The one exception to this general rule is optimization under the *Hankel norm*, but here the known methods are limited to desired functions H corresponding to a finite-order rational transfer function.

By limiting the set of filters to those with no zeros, and by modifying the error criterion considerably, the *ratio-error* methods, minimizing

$$J_R(\hat{A}) = \left\| \frac{H(e^{j\omega})}{\hat{H}(e^{j\omega})} \right\| = \left\| \hat{A}(e^{j\omega}) H(e^{j\omega}) \right\|,$$

were obtained. However, in addition to being limited to minimum-phase (and therefore phase-insensitive) all-pole filters, the magnitude of the approximation tends to be biased upward such as to follow the envelope of $|H(e^{j\omega})|$.

By introducing the unseemly weighting function $|\hat{A}(e^{j\omega})|$ into problem $\hat{H}^*$, thus minimizing

$$J_E(\hat{\theta}) = \left\| \left| \hat{A}(e^{j\omega}) \right| \left(H(e^{j\omega}) - \hat{H}(e^{j\omega}) \right) \right\| = \left\| \hat{A}(e^{j\omega}) H(e^{j\omega}) - \hat{B}(e^{j\omega}) \right\|,$$

we obtained the highly practical *equation-error* methods. However, these methods do not generally yield stable approximations, and they can give poor results when $H(e^{j\omega})$ is non-causal or non-minimum phase.

We turn now to another class of modifications to problem $\hat{H}^*$ which are based on phase-insensitive L^∞ approximation.

It is an important fact that real rational L^∞ approximation is essentially a solved problem. There are at least four different types of algorithm which are all guaranteed to converge monotonically to the solution. These are the Remez multiple exchange algorithm [58,21], simplex-type methods [4,40,54], the differential correction algorithm [6,28,11,12,42,50], and Lawson's algorithm [3,58]. One might ask how problem $\hat{H}^*$ can be modified to yield a *real* rational approximation problem. There are at least two ways: linear phase filter design and power frequency-response approximation. Applications of real rational approximation to these two cases have been pursued in [52,27,21,4,40,54]. In addition, it is possible to fit *group-delay*† disregarding magnitude using real rational approximation methods [21,18].

† Defined in Appendix E.

1.8.2. Linear-Phase Filter Design

Linear phase filters are considered very desirable in many situations because they have constant group delay. This means that the time delay of each frequency component of the input signal, due to filtering, is identical for all frequencies. In waveform coding, for example, it is sometimes important to preserve wave shape as much as possible, and in these cases phase linearity is needed.

If $\hat{H}(e^{j\omega})$ is linear phase, then it can be expressed as

$$\hat{H}(e^{j\omega}) = e^{j\omega\tau}G(e^{j\omega}),$$

where τ is the bulk delay of the filter, and $G(e^{j\omega})$ is real (zero-phase). Assuming a real impulse-response $\hat{h}(n)$ is desired, $\hat{H}(e^{j\omega}) = \overline{\hat{H}(e^{-j\omega})}$ which implies $G(e^{j\omega}) = G(e^{-j\omega})$ which, in the time domain, gives

$$\begin{aligned} g(n) &\triangleq \int_{-\pi}^{\pi} G(e^{j\omega})e^{j\omega n}\frac{d\omega}{2\pi} \\ &= 2\int_{0}^{\pi} G(e^{j\omega})\cos(\omega n)\frac{d\omega}{2\pi} \qquad (1.18) \\ \Rightarrow \quad g(n) &= g(-n). \end{aligned}$$

Thus the zero-phase part must have a symmetric impulse-response. (A purely imaginary impulse-response is equivalent in practice to a real impulse-response, in which case $G(e^{j\omega}) = -G(e^{-j\omega}) \Rightarrow g(n) = -g(-n)$. An example of this case is given by Hilbert transform filters [169,52,196].) For stable filters, this symmetry (or anti-symmetry) can only be had with finite-impulse-response (FIR) filters. Thus stable, recursive, linear-phase filter design is a contradiction in terms! However, it is reasonable to ask for the stable recursive filter which approximates $|H(e^{j\omega})|$ closely and is *approximately* linear phase. In §1.5.6 an example of this type is carried out using the CF and equation-error methods. For this purpose, the CF method appears to be the best known method. There does not seem to be a literature on this application, perhaps because it *is* very difficult without Hankel-norm methods, and even then there are numerical difficulties (see §1.5.6).

Linear phase FIR filter design under the L^∞ norm has been available for a long time. Of the four approaches that could be used, the Remez multiple exchange algorithm seems to be the most widely used at present [169,52,196].

1.8.3. Approximation of Power Frequency-Response—Problem $|\hat{H}^*|^2$

The other natural conversion of problem $\hat{H}^*$ into a real rational L^∞ approximation problem is as follows:

Problem $|\hat{H}^*|^2$

Given a continuous real non-negative function $G(e^{j\omega}) \triangleq |H(e^{j\omega})|^2$, $-\pi < \omega \leq \pi$, corresponding to a desired frequency-response magnitude squared, find a strictly stable digital filter, of the form

$$\hat{H}(z) \triangleq \frac{\hat{B}(z)}{\hat{A}(z)},$$

such that some norm of the error

$$\begin{aligned} J(\hat{\theta}) &\triangleq \left\| G(e^{j\omega}) - \hat{G}(e^{j\omega}) \right\| \\ &\triangleq \left\| G(e^{j\omega}) - \frac{\hat{D}(e^{j\omega})}{\hat{C}(e^{j\omega})} \right\| \\ &\triangleq \left\| \left| H(e^{j\omega}) \right|^2 - \left| \hat{H}(e^{j\omega}) \right|^2 \right\| \end{aligned}$$

is minimum, subject to the polynomials $\hat{D}(e^{j\omega})$ and $\hat{C}(e^{j\omega})$ being real and non-negative.

Thus we obtain an optimum *power gain* in the resulting filter. In many applications, this is not a compromise. As in the ratio-error and linear-phase methods, phase is completely eliminated from the error criterion.

Problem $|\hat{H}^*|^2$ is solvable using the L^∞ norm, since it reduces to real rational approximation. Most formulations of real rational approximation apply to rational functions over the real interval $[-1,1]$ rather than the unit circle. This is not a source of difficulty since all type $(\hat{n}_a, \hat{n}_b)$ rational filters have power frequency-response functions which are type $(\hat{n}_a, \hat{n}_b)$ rational functions of $\cos(\omega n)$. (Just use the "folding" technique of (1.18) on the numerator and denominator separately.) The substitutions $T_k(x) = \cos(k\omega)$, $\omega = \cos^{-1}(x), x \in [-1,1]$ convert the rational function of $\cos(k\omega)$ into a (non-negative real) rational function of x in the real interval $[-1,1]$ (by way of the Chebyshev polynomials $T_k(x)$ [132,155]). Alternatively, the optimization can be carried out explicitly in terms of the rational function in $\cos(k\omega)$.

In problem $|\hat{H}^*|^2$, the parameters $\hat{\theta}$ become the coefficients of the "squared" filter transfer function

$$\hat{G}(e^{j\omega}) \triangleq \left| \hat{H}(e^{j\omega}) \right|^2 = \hat{H}(e^{j\omega})\hat{H}(e^{-j\omega}) = \hat{H}(z)\hat{H}(z^{-1})\Big|_{z=e^{j\omega}}.$$

This representation (obtained by analytic continuation) shows that all poles and zeros of $\hat{G}(z)$ occur in reciprocal pairs, e.g., a pole at $z = p$ implies a pole at $z = 1/p$. To obtain a stable filter $\hat{H}(z)$, the poles must be inside the unit circle, thus specifying which half to select in the factorization $\hat{H}(z)\hat{H}(z^{-1})$. A method for performing this factorization efficiently is

given in §1.9.2. The zeros are not so constrained, but they are generally chosen as the half inside the unit circle to provide *minimum phase*. This allow $\hat{H}(z)$ to be *causally invertible*. (Once the poles have been found, however, one is free do compute the zeros by any of a wide variety of methods, since this is an easy problem.) Zeros on the unit circle must be of even order, for otherwise points of negative magnitude frequency-response would ensue, and half of these are selected for $\hat{H}(z)$.

Expanding $\hat{G}(z) = \hat{H}(z)\hat{H}(z^{-1})$ into a Laurent series converging on the unit circle yields a polynomial

$$\hat{G}(z) = \cdots + \hat{g}(k)z^{-k} + \cdots + \hat{g}(1)z^{-1} + \hat{g}(0) + \hat{g}(1)z + \cdots + \hat{g}(k)z^{k} + \cdots$$

which is "mirror symmetric," i.e., the coefficient of z^k equals the coefficient of z^{-k}. Conversely, every mirror-symmetric polynomial has roots in reciprocal pairs (since substituting $z \leftarrow z^{-1}$ in $\hat{G}(z)$ yields $\hat{G}(z^{-1}) = z^{\nu}\hat{G}(z)$ for some ν). It is generally not difficult to enforce this symmetry in specific algorithms, for we can use expression (1.18) to halve the size of the coefficient space and force symmetry (as is done in FIR linear-phase filter design [52,196]).

Unfortunately, mirror symmetry is not sufficient for non-negativity on the unit circle. To see this, note that on the circle,

$$\begin{aligned}\hat{G}(e^{j\omega}) &= \cdots + \hat{g}(k)e^{-j\omega k} + \cdots + \hat{g}(1)e^{-j\omega} + \hat{g}(0) + \hat{g}(1)e^{j\omega} + \cdots + \hat{g}(k)e^{j\omega k} + \cdots \\ &= \hat{g}(0) + 2\hat{g}(1)\cos(\omega) + \cdots + 2\hat{g}(k)\cos(k\omega) + \cdots.\end{aligned} \tag{1.19}$$

Clearly, symmetry forces $\hat{G}(e^{j\omega})$ to be real. However, it must also be non-negative, a condition which is not guaranteed by the form of the expression. Consider the function

$$\hat{G}^{+}(z) \triangleq \hat{g}(0) + 2\hat{g}(1)z^{-1} + \cdots + 2\hat{g}(k)z^{-k} + \cdots, \tag{1.20}$$

which we call the *causal image* of $G(z)$. Non-negativity implies (1) $\hat{G}(e^{j\omega}) = Re\{\hat{G}^{+}(e^{j\omega})\} \geq 0$. Since $\hat{G}^{+}(z)$ must converge on the unit circle, it converges outside the unit circle. Therefore, (2) $\hat{G}^{+}(z)$ is analytic outside the unit circle. These two conditions are equivalent to $\hat{G}^{+}(z)$ being *positive real*. Appendix C proves this and other facts regarding positive real functions.

While enforcing symmetry is not difficult, there remains the problem of enforcing non-negativity of $\hat{G}(e^{j\omega}) = |\hat{H}(e^{j\omega})|^2$, or, alternatively, constraining $\hat{G}^{+}(z)$ of (1.20) to be positive real. Since the fit we can obtain is of the L^{∞} type, $\hat{G}(e^{j\omega})$ will *oscillate uniformly* about $G(e^{j\omega})$. Consider that often in practice the desired squared-magnitude $G(e^{j\omega})$ is close to zero for some frequencies. It is therefore likely that $\hat{G}(e^{j\omega})$ will "ripple through zero" in these frequency regions; unless $G(e^{j\omega}) \gg 0$ for all ω, it is highly likely that we will obtain $\hat{G}(e^{j\omega}) < 0$ for some ω. If we proceed with the factorization (which still works

in principle), we obtain a filter $\hat{H}(z)$ with complex coefficients. But then $\hat{H}(z)\overline{\hat{H}(z^{-1})}$ is no longer the squared modulus of $\hat{H}(z)$ on the unit circle. (Complex filters have power gain given by $\hat{H}(e^{j\omega})\overline{\hat{H}(e^{j\omega})}$ which analytically continues to $\hat{H}(z)\overline{\hat{H}(\overline{z^{-1}})}$.) Consequently, $|\hat{H}(e^{j\omega})|^2$ need bear no resemblance to $|H(e^{j\omega})|^2$ when $\hat{G}(e^{j\omega}) < 0$ for some ω.

One way to restore the positive real condition (non-negativity of $|\hat{H}(e^{j\omega})|^2$) is to replace $\hat{g}(0)$ by $\hat{g}(0) - \min_\omega \hat{G}(e^{j\omega})$. This lifts the approximation uniformly until it becomes non-negative. Of course, a reasonable fit remains only when a good fit was obtained initially. In section §1.8.5, another method for avoiding negative squared-magnitude approximations is given.

Of the four basic techniques mentioned for real rational approximation, the *simplex method* seems to be best suited for problem $|\hat{H}^*|^2$, since imposing the constraints $\hat{D}(e^{j\omega}) \geq 0, \hat{C}(e^{j\omega}) \geq 0$, where $\hat{G}(e^{j\omega}) = \hat{D}(e^{j\omega})/\hat{C}(e^{j\omega})$, is very natural in the linear programming context [54,26]. *

1.8.4. Mapping problem $|\hat{H}^*|^2$ onto problem $\hat{H}^*$

Given $G(e^{j\omega}) = |H(e^{j\omega})|^2$, one can form $G^+(e^{j\omega})$ (cf. (1.20)) using the spectral factorization method of §1.9.2. This is then a causal continuous function which can be directly approximated by the techniques given previously for problem $\hat{H}^*$. The question arises as to what relationship the optimum solution $\hat{G}^{+*}$ of problem $\hat{H}^*$ has with the desired solution $\hat{G}^*$ for problem $|\hat{H}^*|^2$. It turns out that the solutions are essentially equivalent for the *Hankel norm* and the L^2 *norm*.

For the Hankel norm, we have the error measure

$$J_H(\hat{\theta}^+) = \left\| G^+(e^{j\omega}) - \hat{G}^+(e^{j\omega}) \right\|_H .$$

By definition, the Hankel norm is invariant with respect to non-causal modifications (see Appendix E). Thus if we define

$$G_H^+(z) = \sum_{n=0}^{\infty} g(n)e^{-j\omega n}, \qquad \hat{G}_H^+(z) = \sum_{n=0}^{\infty} \hat{g}(n)e^{-j\omega n},$$

where $G(e^{j\omega}) = 2Re\{G_H^+(e^{j\omega})\} - g(0)$ and where $\hat{G}(e^{j\omega}) = 2Re\{\hat{G}_H^+(e^{j\omega})\} - \hat{g}(0)$, then

$$J_H(\hat{\theta}^+) = \left\| G_H^+(e^{j\omega}) - \hat{G}_H^+(e^{j\omega}) \right\|_H = \left\| G(e^{j\omega}) - \hat{G}(e^{j\omega}) \right\|_H = J_H(\hat{\theta}).$$

* A method for magnitude approximation based on Fletcher-Powell descent (which is said to work successfully for filter orders as high as 24 and which is also said to perform better than the one in [26] for simple filter types) is given in [47,24].

Thus with $G_H^+(e^{j\omega})$ defined as above, the *CF method* can be used to obtain a solution to problem $|\hat{H}^*|^2$. However, we have not constrained $\hat{G}_H^+(z)$ to be positive real.

For the L^2 norm, a similar relationship occurs. However, in this case, we define the causal image of the power frequency-response by

$$G_2^+(e^{j\omega}) = 2^{\frac{1}{2}} g(0) + 2\sum_{n=1}^{\infty} g(n)e^{-j\omega n}, \qquad \hat{G}_2^+(e^{j\omega}) = 2^{\frac{1}{2}} \hat{g}(0) + 2\sum_{n=1}^{\infty} \hat{g}(n)e^{-j\omega n}.$$

In this case, $G(e^{j\omega}) = Re\{G_2^+(e^{j\omega})\} + (1-2^{\frac{1}{2}})g(0)$ and $\hat{G}(e^{j\omega}) = Re\{\hat{G}_2^+(e^{j\omega})\} + (1-2^{\frac{1}{2}})\hat{g}(0)$. By the orthogonality of distinct powers of z on the unit circle, we have

$$\begin{aligned}
J_2^2(\hat{\theta}^+) &= \left\| G_2^+(e^{j\omega}) - \hat{G}_2^+(e^{j\omega}) \right\|_2^2 \\
&= \left\| 2^{\frac{1}{2}} g(0) + 2\sum_{n=1}^{\infty} g(n)e^{-j\omega n} - 2^{\frac{1}{2}} \hat{g}(0) - 2\sum_{n=1}^{\infty} \hat{g}(n)e^{-j\omega n} \right\|_2^2 \\
&= 2\big(g(0) - \hat{g}(0)\big)^2 + 4\sum_{n=1}^{\infty} \big(g(n) - \hat{g}(n)\big)^2 \\
&= 2\sum_{n=-\infty}^{\infty} \big(g(n) - \hat{g}(n)\big)^2 \\
&= 2\left\| \sum_{n=-\infty}^{\infty} g(n)e^{-j\omega n} - \sum_{n=-\infty}^{\infty} \hat{g}(n)e^{-j\omega n} \right\|_2^2 \\
&= 2\left\| G(e^{j\omega}) - \hat{G}(e^{j\omega}) \right\|_2^2 \\
&= 2J_2^2(\hat{\theta}),
\end{aligned}$$

thus solving problem $|\hat{H}^*|^2$ under the L^2 norm. Again, the positive real condition is not guaranteed. Furthermore, in this case, there seems to be no known algorithm which is globally convergent (cf. Appendix A).

1.8.5. Approximation of Log-Magnitude Frequency-Response

The human ear can be effectively modeled as a spectrum analyzer, and this spectrum analyzer seems to measure the *weighted log-magnitude spectrum* of sounds, to a first-order approximation (see Chapter 3). Similarly, the eye is sensitive primarily to the log of light intensity. In most perceptual phenomena, subjective scales tend to be logarithmically related to the associated physical quantities ("Weber's law").

In view of these observations, it is natural to desire a filter-design method which minimizes some norm of the *weighted log-magnitude spectral error*. Since the log of a

spectrum admits the decomposition

$$\ln H(e^{j\omega}) = \ln \left| H(e^{j\omega}) \right| + j\angle H(e^{j\omega}),$$

it follows that such a method should be applicable to designing filters with a prescribed *phase-response* or *phase-delay* (with the latter being simply the linear weighting of the former by the function $-1/\omega$, as discussed in Appendix E).

The problem is then to find the filter $\hat{H}^*(z)$ which minimizes

$$\begin{aligned} J_L(\hat{\theta}) &\triangleq \left\| \ln G(e^{j\omega}) - \ln \hat{G}(e^{j\omega}) \right\|_\infty \\ &\triangleq \left\| \ln |H(e^{j\omega})|^2 - \ln |\hat{H}(e^{j\omega})|^2 \right\|_\infty \\ &= 2 \left\| \ln \left| \frac{H(e^{j\omega})}{\hat{H}(e^{j\omega})} \right| \right\|_\infty \qquad \text{(Log Ratio Error)}, \end{aligned} \tag{1.21}$$

Note that this is equivalent to minimizing $\| \ln |H(e^{j\omega})|^p - \ln |\hat{H}(e^{j\omega})|^p \|$ for any $p > 0$. In view of previous discussion, since the log-magnitude error is *real*, the L^∞ norm or Hankel norm [73] should be considered good candidates for obtaining a globally convergent algorithm.

Log-magnitude approximation under the L^∞ norm has been investigated by Deczky [21], where he also applies the same method to approximating *group-delay*. This necessitated some generalizations of the L^∞ approximation theory. The *characterization theorem* for real rational Chebyshev approximation states that a type $(\hat{n}_a, \hat{n}_b)$ rational function is a best approximation if and only if there are $\hat{N} + 1 = \hat{n}_a + \hat{n}_b + 2$ points in the interval of approximation at which the error achieves its maximum, and if the error has alternating signs on these points. Thus the "equiripple" property of Chebyshev approximations is fundamental. (In the complex case, the rippling translates to "winding" of the error function around zero.) Essentially, Deczky showed that the characterization theorem for Chebyshev rational approximation holds without modification for approximation by the log-magnitude of rational functions. (This is not the case for group-delay approximation.) More generally, he showed that the characterization theorem holds for magnitude-approximation by any continuous monotonic function of real rational functions.

For obtaining a Chebyshev approximation to the log-magnitude frequency response, Deczky chose the *Remez multiple exchange* algorithm [58]. The Remez algorithm consists basically of finding $\hat{N} + 1$ points of maximum error, and interpolating to produce an error on these points which alternates in sign with a constant magnitude. The error between these points may jump to wild values, and so this process is repeated. (In the real rational case the alternating error is guaranteed to decrease monotonically.) It is straightforward

to find the extrema of the error curve in order to find the points where error alternation will next be forced. However, the interpolation problem is another nonlinear and difficult step. Deczky recommends the use of Newton's method for the interpolation process. Even if the interpolation step is always successful, there does not appear to exist a proof of global convergence, as there is in the case of ordinary real rational L^∞ approximation.

A method which retains the theoretical guarantees is based on a first-order approximation of the logarithm itself. Equation (1.21) can be written

$$\begin{aligned} J_L(\hat{\theta}) &= \left\| \ln \left| \frac{\hat{H}}{H} \right|^2 \right\|_\infty \\ &= \left\| \left(\left| \frac{\hat{H}}{H} \right|^2 - 1 \right) - \frac{1}{2} \left(\left| \frac{\hat{H}}{H} \right|^2 - 1 \right)^2 + \frac{1}{3} \left(\left| \frac{\hat{H}}{H} \right|^2 - 1 \right)^3 - \cdots \right\|_\infty \\ &\approx \left\| \left| \frac{\hat{H}}{H} \right|^2 - 1 \right\|_\infty = \left\| \frac{|H|^2 - |\hat{H}|^2}{|H|^2} \right\|_\infty , \end{aligned} \tag{1.22}$$

when $|\hat{H}(e^{j\omega})|^2 \approx |H(e^{j\omega})|^2$. Thus when good fits are possible, (1.22) closely approximates the log spectral deviation at the optimum solution. Also, the problem is reduced to the previous case of power approximation with the *linear* weighting by the known function $|H(e^{j\omega})|^2$. Thus any real rational L^∞ approximation method which allows a linear weighting on the error, such as the Remez exchange algorithm [58], can be used to solve this problem.

This method has been tried in the polynomial (FIR) case, using code for the Remez exchange algorithm adapted from [52,196], and has been found to perform satisfactorily in practice (see Chapter 3). The nature of the weighting makes it unlikely that a negative magnitude spectrum will occur in the computed filter. Note, however, that we must have $H(e^{j\omega}) \neq 0$ for all ω. If $H(e^{j\omega}) \approx 0$ for some ω, then we can use the weighting $1/(|H|^2 + \epsilon)$ in place of $1/|H|^2$.

1.8.6. Phase Approximation

As has been shown, a variety of possibilities appear when phase is ignored in the approximation. A natural complement to such techniques is the ability to design an *allpass* filter with an optimized phase, phase-delay, or group-delay. (These terms are defined in Appendix E.) An allpass filter in cascade will not alter the magnitude approximation. The phase of the allpass is fit to the desired phase minus the phase response of the phase-insensitive approximation. The overall resulting filter design cannot be optimal, in general, since phase and magnitude approximation have been decoupled. There are numerous

methods for the design of filters with a prescribed phase-response (usually in the form of group delay), e.g., [5,20,21,56,74,80].

It should be noted that none of these methods is guaranteed to provide an optimum phase-response approximation. The closest approach to this goal is given by Deczky [21], in which the L^∞-type theory is extended as far as possible.

Two general and apparently effective methods are those by Bernhardt [5] and Yegnanarayana [80]. Yegnanarayana obtains a more uniform error in group-delay approximation, especially near 0 Hz and half the sampling rate, than that in [5]. Also, his method is *much* easier to program, and lends itself to implementation on array processors (requiring only the commonly available modules DFT, array exponentiation, and Durbin's recursion.) While Yegnarayana's method is only defined for group-delay approximation, it is simple to generalize to phase and phase-delay approximation.*

1.8.7. Padé Approximation

Another class of methods which do not minimize a true norm of the error $\|H(e^{j\omega}) - \hat{H}(e^{j\omega})\|$ is based on *Padé approximation*. In these methods, a "pseudo-norm†" of the error $H(e^{j\omega}) - \hat{H}(e^{j\omega})$ is minimized, which is given by

$$J_P(\hat{\theta}) \triangleq \sum_{n=0}^{\hat{n}_a+\hat{n}_b} \left(h(n) - \hat{h}(n)\right)^2 .$$

This is then a truncated l^2 norm in which the first $\hat{n}_a + \hat{n}_b + 1$ samples of the impulse-response are matched exactly. From this point of view, Padé approximation provides a *time domain* method.

Padé approximation is also the best approximation of the transfer function $H(z)$ at $z = 0$ in the L^∞ sense. This is because it matches the maximum number of Taylor expansion coefficients about $z = 0$ (the first $\hat{n}_a + \hat{n}_b + 1$ impulse-response values). There is a theorem on *real* rational approximation [14,16,76] which states:

Theorem 1.37. If $H(z)$ is any continuous function in a neighborhood $[0, \delta]$, for $\delta > 0$, and $\hat{H}_\epsilon$ is the optimum L^∞ $(\hat{n}_b, \hat{n}_a)$ approximation of $H(z)$ on $[0, \epsilon]$, then $\{\hat{H}_\epsilon\}$ converges uniformly to the $(\hat{n}_b, \hat{n}_a)$ Padé approximation of $H(z)$ as $\epsilon \to 0$.

---

* In the algorithm described in [80], replace equation (21) by equation (20) for phase approximation, and for phase-delay approximation do the same except mutiply through (20) by $1/\omega$. Division by zero does not occur since (20) is a sum of sines. Both these modifications lead to replacing the initial cosine transform by a sine transform (thus to use the FFT, an anti-symmetric function is prepared for the first step rather than a symmetric one).

† Defined in Appendix E.

1.8.8. Classical Digital Filter-Design Techniques

There is a body of filter-design techniques based on *analog* methods [196]. These methods include filters of the type Butterworth, Chebyshev, inverse Chebyshev, and Cauer (or elliptic function) filters. The bilinear transform $(s \leftarrow (z-1)/(z+1))$ is used to map the analog design from the s-plane onto the z-plane, thus converting the analog filter coefficients to those of a digital filter. Since the bilinear transform maps the frequency axis without modification of the spectral values, equal-ripple filters map to equal-ripple filters, and the optimality of L^∞ designs (elliptic and Chebyshev) is preserved. However, the classical design methods provide only a prototype *lowpass* filter, and are not intended for general spectral curves. Moreover, it is not preferable to carry out general filter design in the s-plane because there the frequency-response is defined over an infinite frequency interval. Apart from this consideration, approximation over the s-plane is essentially equivalent to approximation over the z-plane (except in some specialized cases). Thus, the classical analog filter-design techniques do not seem to offer any new methods for solving problem $\hat{H}^*$.

1.9. Special Tools and Techniques

As a final topic in digital filter design, some techniques which are often needed for the effective application of a design method are presented. The first is a conformal-mapping procedure which can be used to enlarge the domain of applicability of some filter-design methods, and the second is a new method for spectral factorization which has been found to be faster and more reliable than polynomial factorization methods.

1.9.1. Applications of Conformal Mapping

Conformal mapping techniques can be useful for

(1) Scaling a filter to be used at a different sampling rate.

(2) Providing a spectral weight-function for filter-design methods lacking this flexibility.

For adapting a digital filter to a new sampling rate, linear stretching of the frequency axis, or *frequency scaling*, is required. Frequency scaling is simple for analog filters [196]: Substituting $s\omega_2/\omega_1$ into the analog transfer function $H(s)$ maps the frequency ω_1 to the frequency ω_2. Elsewhere, the frequency axis is linearly stretched by the factor ω_2/ω_1 as desired. For digital filters, however, the analogous substitution $z \leftarrow z^{\omega_2/\omega_1}$ produces an *irrational filter*. Thus some other method must be found.

There is a class of transformations from the unit circle to itself which preserves the order of a rational function. These are known as the *bilinear transformations*, having the

form

$$z \leftarrow S_\rho(z) \triangleq \frac{z+\rho}{\rho z+1}, \quad \rho \in \Re, \quad |\rho| < 1, \tag{1.23}$$

The variable z in a rational transfer function $H(z)$ is replaced by the first-order allpass filter $S_\rho(z)$. The restriction to real ρ is necessary to yield real coefficients when mapping a real filter $H(z)$ to $H(S_\rho(z))$. The restriction to $|\rho| < 1$ ensures that the unit disk $\mathcal{D}$ maps to itself. (For $|\rho| > 1$, $\mathcal{D}$ maps to the region $|z| \geq 1$.) Thus $|\rho| < 1$ is necessary to preserve stability of the mapped filter. In such a case, $S_\rho(z)$ is analytic in $\mathcal{D}$, and on the boundary Γ we have

$$\left|S_\rho(e^{j\omega})\right| = \left|\frac{e^{j\omega}+\rho}{\rho e^{j\omega}+1}\right| = \left|\frac{e^{j\omega}+\rho}{\rho+e^{-j\omega}}\right| = \left|\frac{e^{j\omega}+\rho}{\overline{e^{j\omega}+\rho}}\right| = 1.$$

Thus $S_\rho(z)$ is a Schur function.*

A plot of the frequency-mapping behavior of $S_\rho(z)$ for several values of ρ is shown in Fig. 1.14.

Relation to s-plane Frequency Scaling

If the bilinear transform is to be used for frequency-scaling, it is useful to relate it to the s-plane frequency-scaling formulas. The standard mapping from the z-plane to the s-plane is given by

$$z \leftarrow \frac{a+s}{a-s}, \quad a = \omega_s \cot\left(\frac{\omega_z T}{2}\right), \tag{1.24}$$

where $T = 1/f_s$ is the sampling period, ω_z is "digital frequency" and ω_s is "analog frequency." I.e., the point $z = e^{j\omega_z}$ in the z-plane maps to the point $s = j\omega_s$ in the s-plane. This is the most general first-order mapping which takes $\omega_z = 0$ to $\omega_s = 0$ and $\omega_z = \pi f_s$ to $\omega_s = \infty$ [150]. I.e., 0 Hz maps to 0 Hz, and the highest digital frequency is mapped to the highest analog frequency. (This mapping is discussed in more detail in Appendix C.)

Having mapped the digital frequency domain onto the analog frequency domain, giving us the transfer function $H((a+s)/(a-s))$, we can apply the frequency-scaling formulas for analog filters to obtain

$$s \leftarrow cs',$$

thus arriving at the filter $H((a+cs')/(a-cs'))$. The constant c is a convenient frequency scaling parameter. The value $c = \omega_1/\omega_0$ maps ω_0 to ω_1. Finally, we map back to the

* A *Schur function* is defined as a complex function analytic and of modulus not exceeding unity in $\mathcal{D}$.

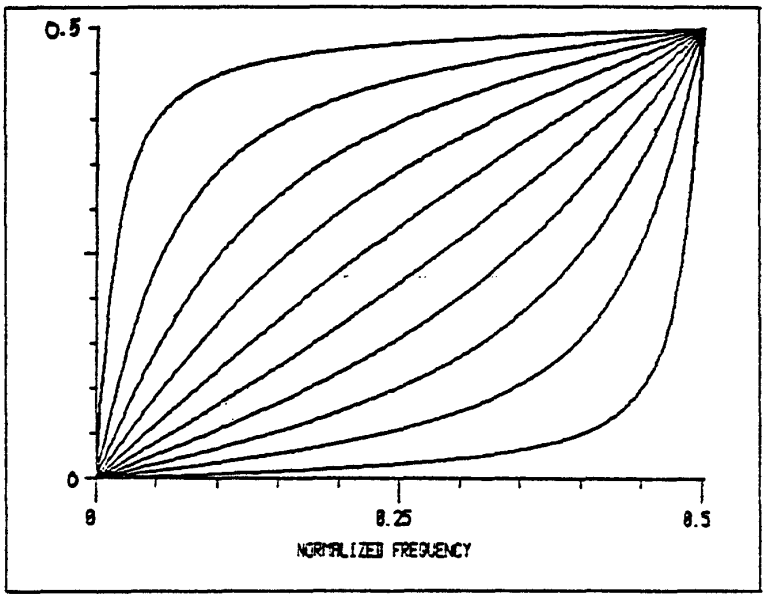

Figure 1.14. Frequency-mapping curves generated by $S_\rho(z)$ with ρ ranging from -0.9 to 0.9 in steps of 0.2. The top curve corresponds to $\rho = 0.9$, and the bottom curve corresponds to $\rho = -0.9$. The functions are generated using the expression $\varphi = \tan^{-1}\{(1-\rho^2)\sin(\omega)/[(1+\rho^2)\cos(\omega) - 2\rho]\}$, where ω is radian input frequency (0 to π), and φ is the image of ω under the mapping $S_\rho(e^{j\omega})$. The plotted values are divided by 2π.

z-plane using the inverse of the mapping (1.24),

$$s' \leftarrow a\frac{z-1}{z+1}.$$

The combined effect of the mappings is of the form $S_\rho(z)$ in (1.23) where

$$\rho = \frac{1-c}{1+c}.$$

The stability constraint translates to $c > 0$, which can be interpreted as requiring each frequency to map to a frequency of the same sign. The resultant flexibility obtainable with the mapping $S_\rho(z)$ is the mapping of any frequency in the open interval $(0, f_s/2)$ to any other frequency in that interval. (The points $f = 0, f = f_s/2$ are fixed-points of the map.) The upper unit semi-circle and the lower unit semi-circle are compressed and stretched together to maintain conjugate symmetry in the frequency response $H(e^{j\omega})$ so that the filter coefficients remain real.

Specific Frequency Mapping

In other situations, one may need to know the mapping $S_\rho(z)$ which takes a particular frequency* φ to the frequency θ. If these frequencies lie in $(0, \pi)$, then the mapping constant is found, after some manipulation, to be given by

$$\rho = \frac{\sin\left(\frac{\theta-\varphi}{2}\right)}{\cos\left(\frac{\theta+\varphi}{2}\right)} \tag{1.25}$$

For low frequencies, we have the approximation

$$\rho \approx \frac{\theta-\varphi}{\theta+\varphi}.$$

The frequency-scaling constant for this mapping is found by solving $\rho = (1-c)/(1+c)$, which, for low frequencies, gives $c \approx \varphi/0$, or $\varphi \approx c\theta$. Thus at low frequencies, the frequency-scaling formulation is an accurate way to map specific frequencies. This is not the case at high frequencies.

To map the other way from θ to φ, we interchange these frequencies in (1.25), and evidently the *inverse map is obtained by negating* ρ. In general,

$$\rho > 0 \;\Rightarrow\; \text{Stretch}(\theta > \varphi)$$
$$\rho < 0 \;\Rightarrow\; \text{Compression}(\theta < \varphi)$$

in the low-frequency region.

In practice, the mapping may be implemented by substituting

$$z^{-1} \leftarrow \frac{\rho + z^{-1}}{1+\rho z^{-1}}$$

into the transfer function $H(z)$ which is conventionally a function of z^{-1}. The inverse transformation is given by

$$z^{-1} \leftarrow \frac{z^{-1}-\rho}{1-\rho z^{-1}}, \tag{1.26}$$

i.e., ρ is simply negated.

* The term "frequency" is used loosely here to mean the angle in the z-plane corresponding to that frequency. Thus if the sampling rate is f_s Hz, the true frequency f in Hz corresponds to the angle $\theta = 2\pi f/f_s$, and here θ is referred to as the frequency.

Application to Filter Design

In some filter-design methods, such as those based on the Hankel norm, it is difficult to introduce a weighting function on the error. In these cases an effective error weighting can by provided by the conformal map $S_\rho(z)$ given in (1.26).

Consider the situation in which the most important structure in the frequency response is at low-frequencies. This arises quite often in high-quality audio work, where the sampling rates are large. For example, it has been observed that the most important information-carrying regions of the spectrum of speech vowels are the first three or four formants [186] which tend to lie largely below 4KHz. If speech is to be modeled at audio-quality sampling rates, such as 40KHz, it is very difficult to get a low-order filter whose frequency response closely follows the first four formants. Markel and Gray [186] give the rule of thumb that for speech one should take the sampling rate in KHz and add five or so. At 40KHz this is an order 45 filter. By using the circle-to-circle transform $S_\rho(z)$, the first four formants can be nicely modeled with a filter of order 8. An example of this nature is given in Chapter 3 where conformal mapping is used to model the body of the violin.

The basic procedure is the following:
(1) Map the desired spectrum via the substitution $e^{j\omega} \leftarrow S_\rho(e^{j\omega})$.
(2) Design a filter.
(3) Map the filter to original frequency coordinates via $z \leftarrow S_{-\rho}(z)$.

More specifically:

(1) Replace the target spectrum $H(e^{j\omega_k}), k = 0, \ldots, N$ by

$$H_\rho(e^{j\omega_k}) \triangleq H\left(\frac{e^{j\omega_k} + \rho}{\rho e^{j\omega_k} + 1}\right), \quad k = 0, \ldots, N.$$

The value of $H_\rho(e^{j\omega_k})$ is assigned the value of H at

$$e^{j\omega\varphi_k} \triangleq \frac{e^{j\omega_k} + \rho}{\rho e^{j\omega_k} + 1}$$

$$\Rightarrow \quad \varphi_k = \tan^{-1}\left(\frac{(1-\rho^2)\sin(\omega_k)}{2\rho + (1+\rho^2)\cos(\omega_k)}\right) \triangleq \frac{2\pi k_\rho}{N}.$$

Thus the kth element of the H_ρ array is assigned the k_ρth element of the H array. Since k_ρ is not an integer in general, one may wish to do some type of interpolation on the values of H for better accuracy.

(2) Fit a digital filter to $H_\rho(e^{j\omega})$ to obtain $\hat{H}_\rho(z)$.

(3) Compute

$$\hat{H}(z) = \hat{H}_\rho\left(\frac{z-\rho}{1-\rho z}\right)$$

as the final approximate filter.

Note that if the number of zeros does not equal the number of poles in step (2), the final filter $\hat{H}(z)$ *will* generally have the same number of poles and zeros. For example, if we fit an all-pole filter,

$$\hat{H}_\rho(z) = \frac{1}{1 + \hat{a}_\rho(1)z^{-1} + \cdots + \hat{a}_\rho(\hat{n}_a)z^{-\hat{n}_a}},$$

with $\hat{n}_a$ poles, then the $\hat{n}_a$ "implicit" zeros at $z = 0$ will move away from the origin under the mapping $S_{-\rho}$. Since the numerical conditioning of the mapping is poor when all roots are at one point, it is best to fit filters with an equal number of poles and zeros unless there is a good reason to do otherwise. Experience has shown that the multiple roots tend to map to a small circle about the point where they would go given perfect numerical precision. The phenomenon appears similar to the one observed when factoring polynomials with repeated roots. It is said in this case that the radius of the circle is of the order $\epsilon^{1/n}$, where ϵ is the machine epsilon and n is the number of repeated roots (see also Ortega [192] pp. 43–44). Indeed, even if the mapping does not cause this problem, the root-finding necessary to look for it may well introduce it. (The frequency response gave direct evidence, however, in the computed examples.) One scheme for dealing with spreading repeated roots is to take their average (center of the circle) and recompute the associated polynomial since this is evidently a good estimate of where they belong. Another possibility is to implement the filter in mapped form; this entails replacing delay cells in the filter structure by first-order allpass filters.

The inverse mapping is numerically poor in the case of multiple roots as discussed above, and also for high order filters (say above 15th order in 36-bit floating point), especially when the poles of the filter are clustered and/or near the unit circle. It was found that best results are obtained by performing a *partial-fraction expansion* on the mapped filter prior to undoing the map. This allows the numerical behavior for an arbitrarily large filter to be the same as for second order. Since the poles and zeros of the filter are typically well separated in the mapped domain, the partial fraction expansion is rarely troublesome.

In Chapter 3, the first-order conformal map is applied to audio filter design. It turns out that this mapping can be used to obtain a very close approximation to *Bark frequency units*. The Bark frequency scale has the property that the critical bands of hearing are made to have the same bandwidth throughout the spectrum. Thus the conformal mapping device provides a means for making the "variability" of a filter frequency-response uniform with respect to the frequency-resolution of the ear.

Finally, note that filters can be designed at a reduced sampling rate (so that low-frequency spectral structure is easier to model), followed by frequency-scaling back to the original sampling rate (using conformal mapping). This is in some sense the "dual" of the filter-design technique discussed above.

1.9.2. Fast Spectral Factorization

Spectral factorization is needed in the identification methods which match spectral power and also in the Hankel-norm method. In this section, a method based on the FFT is described.

Problem Statement

Given $\{G(e^{j\omega_k})\}_0^{N-1}$ or $\{d_n\}_0^{n_b}$ and $\{c_n\}_1^{n_a}$, where

$$G(z) \triangleq \frac{D(z)}{C(z)} \triangleq \frac{\sum_{n=0}^{n_b} d_n(z^n + z^{-n})}{1 + \sum_{n=1}^{n_a} c_n(z^n + z^{-n})}$$
$$\triangleq H(z)H(z^{-1}) \triangleq \frac{B(z)B(z^{-1})}{A(z)A(z^{-1})} \triangleq \left(\frac{\sum_{n=0}^{n_b} b_n z^{-n}}{1 + \sum_{n=1}^{n_a} a_n z^{-n}}\right)\left(\frac{\sum_{n=0}^{n_b} b_n z^{n}}{1 + \sum_{n=1}^{n_a} a_n z^{n}}\right)$$

and upper bounds for n_a, n_b, find $\{b_n\}_0^{n_b}, \{a_n\}_1^{n_a}$ such that the roots of $A(z)$ and $B(z)$ lie in the region $\mathcal{D} = \{z \in C \,|\, |z| \leq 1\}$.

Solution

If $D(z)$ and $C(z)$ are given, an accurate and straightforward solution to this problem is simply to find their roots. The roots of $A(z)$ are the roots of $C(z)$ inside the unit circle, and the roots of $B(z)$ are those of $D(z)$ inside the unit circle. However, root-finding is a computationally expensive operation, and it does not apply when G has been obtained nonparametrically as a sampled spectrum. A more economical solution can be based on the additive decomposition

$$G(z) = G^+(z) + G^+(z^{-1}) = \frac{Q(z)}{A(z)} + \frac{Q(z^{-1})}{A(z^{-1})},$$

where

$$Q(z) \triangleq q_0 + q_1 z^{-1} + \cdots + q_{n_q} z^{-n_q}, \quad n_q \leq \max\{n_a - 1, n_b\}$$
$$G^+(z) \triangleq \frac{g(0)}{2} + \sum_{n=1}^{\infty} g(n) z^{-n}.$$

Given $G^+(z)$, the problem factors naturally into two parts: the determination of $A(z)$ followed by solving for $B(z)$.

The proposed approach to solving the spectral factorization problem is based on an *approximate nonparametric method for obtaining $G^+(z)$ in time domain form.* This is accomplished by means of the inverse FFT applied to the function $D(e^{j\omega_k})/C(e^{j\omega_k})$ to obtain $g(n), n = 0, 1, 2, \ldots$. Once $G^+(z)$ has been obtained in this form, a natural choice for finding $A(z)$ and $Q(z)$ is *Prony's method* (described in §1.7.2). Note that only an upper bound for n_q need be specified, and this is known from the problem specification. In the case where $G(e^{j\omega_k})$ is given, we can compute

$$D(e^{j\omega_k}) = B(e^{j\omega_k})B(e^{-j\omega_k}) = A(e^{j\omega_k})A(e^{-j\omega_k})G(e^{j\omega_k})$$

by point-wise multiplication in the frequency domain.

The problem is now reduced to polynomial spectral factorization. A solution to this sub-problem which actually carries out the nonlinear optimization necessary to factor $D(z)$ into the form $B(z)B(z^{-1})$, and which converges quadratically, is given by Wilson's method [204, 205].

In the stochastic case, the polynomial $D(z)$ arises as the autocorrelation of a *moving average* process. The semi-infinite covariance matrix R of this process is banded and Toeplitz, with the entries $R[i, j] = d_{i-j}$, $i, j = 1, 2, \ldots$. Thus the coefficients $\{d(n)\}_{-n_b}^{n_b}$ define each row (centered about the main diagonal). The Cholesky factorization $R = UU^T$ [168] of this matrix yields $\{b_n\}_0^{n_b}$ on each row of the matrix square root U. U is Toeplitz with the top row being $(b_0, b_1, \ldots, b_{n_b}, 0, \ldots)$. A fast algorithm based on this fact was developed by Friedlander [172] using a square-root normalized lattice-filter estimation algorithm to perform a fast Cholesky decomposition of the finite-data covariance matrix R_N $(N \times N)$ which approaches R as $N \to \infty$. The algorithm requires N iterations to find the nonzero coefficients of the bottom row of U_N where $R_N = U_N U_N^T$. The rate at which the algorithm converges is tied to the rate at which the *partial correlation function* [135] approaches zero. This rate is asymptotically of the order $(1-\rho)^N$, where ρ is the minimum distance of a zero of $D(z)$ to the unit circle [135]. In practice, the approximation error tends to approach zero somewhat linearly with N initially [172], and after a large number of iterations, the error stops decreasing due to accumulated round-off errors.

The proposed approach consists of applying the same nonparametric technique used to obtain $G^+(z)$. By taking the inverse FFT of

$$F(e^{j\omega_k}) \triangleq \frac{1}{C(e^{j\omega_k})} = \frac{1}{B(e^{j\omega_k})B(e^{-j\omega_k})},$$

we obtain the autocorrelation $f(n)$ of a purely autoregressive process. The normalized AR coefficients $\{\tilde{b}_n\}_1^{n_b}$ can be efficiently computed by means of the *Durbin recursion* [186], and

we have $b_n^2 = \tilde{b}_n^2 f(0)/\sigma^2$, where $\sigma^2 = 1 + \tilde{b}_1^2 + \cdots + \tilde{b}_{n_b}^2$. In this method, zeros close to the unit circle give rise to *time aliasing* in the AR autocorrelation (see §1.5.4). The amount of this aliasing is asymptotically of the order $(1-\rho)^N$, where N is the FFT length, and ρ is the minimum distance of a zero of $D(z)$ to the unit circle. Thus the approximation error is similar in principle to that in the Cholesky method.

Algorithm Summary

The FFT-based spectral factorization technique, for the case where $G(z)$ is given, is accomplished by the following steps:

(1) Evaluate the initial power spectrum $G(z)$ at $N \gg n_b + n_a + 1$ equally spaced points along the unit circle to obtain $G(e^{j\omega_k})$, where

$$\omega_k \triangleq \frac{2\pi k}{N}, \quad k = 0, 1, \ldots, N-1 .$$

It is preferable to choose N equal to a power of 2 to allow the use of the FFT. Note that since $g(n)$ is real and symmetric, $G(e^{j\omega_k}) = G(e^{-j\omega_k})$, so that only $N/2+1$ real values are needed.

(2) Inverse Fourier transform $G(e^{j\omega_k})$ to obtain the autocorrelation function $g(n)$,

$$g(n) = \mathrm{FFT}^{-1}\left\{G(e^{j\omega_k})\right\} \triangleq \frac{1}{N}\sum_{k=0}^{N} G(e^{j\omega_k})e^{j\omega_k n} .$$

Since $G(n)$ is real and symmetric, a special version of the FFT can be used [169].

(3) Window $g(n)$, to obtain the causal image $g^+(n)$,

$$g^+(n) = \begin{cases} g(n)/2, & n = 0 \\ g(n), & n = 1, \ldots, N/2-1 \\ 0, & n = N/2, \ldots, N-1 . \end{cases}$$

(4) Convert the nonparametric impulse response g^+ to parametric form $\{a_i, q_j\}$, $i = 1, \ldots, n_a$, $j = 0, \ldots, n_q$ by Prony's method (defined in §1.7.2).

(5) Compute

$$F(e^{j\omega_k}) \triangleq \frac{1}{A(e^{j\omega_k})A(e^{-j\omega_k})G(e^{j\omega_k})}$$

at $N \gg n_b + 1$ equally spaced points along the unit circle as in step 1.

(6) Evaluate the corresponding autocorrelation function

$$f(n) = FFT^{-1}\left\{F(e^{j\omega_k})\right\}$$

as in step 2.

(7) Convert the autoregressive autocorrelation $f(n)$ to parametric form $\{b_i\}$, $i = 1,\ldots,n_b$, by the Durbin recursion [186].

In the case where $C(z), D(z)$ are given, steps 5 and 6 can be applied twice with $F = 1/C$ and $F = 1/D$ respectively.

Computed Examples

Two test cases will be presented to compare the performance of the FFT-based spectral factorization method to the Cholesky-type method proposed in [172] for the polynomial case. All computations were performed in single-precision 36-bit floating point on a Foonly F2 computer emulating a PDP10. The mantissa size is 27 bits.

In the first test, we have

$$G^+(z) = 8004 + 2491z^{-1} + 622z^{-2} + 85z^{-3}$$
$$B(z) = 85 + 27z^{-1} + 7z^{-2} + z^{-3}$$

This example is also computed in three other papers [162,172,205]. Table 1.1 gives a comparison of the two methods for this example.

Cholesky		
Steps	Error	Time
16	4.05×10^{-7}	0.20
32	4.05×10^{-7}	0.27

FFT			
Size	Error	Time	Aliasing
16	3.33×10^{-7}	0.08	5.46×10^{-5}
32	9.74×10^{-8}	0.10	2.46×10^{-8}

Table 1.1. Comparison between the Cholesky and FFT factorization methods for example 1. In the Cholesky method, the Steps column gives the number of iterations and corresponds to the effective size of the covariance matrix which is factored. In the FFT method, the Size column gives the FFT length N. In both cases, the error is computed as the sample standard deviation (root mean squared error) between true and approximate factorization parameters $b_n, n = 0,\ldots,n_b$. The computation time is in seconds, and it is accurate only to a few 60ths of a second. The time aliasing measure, defined in §1.5.4, is based on the middle 10% of the FFT buffer (square-root of the relative energy in $F(e^{j\omega_k})$ for $k \in [N/2 \pm N/20]$).

This example is very easy for both methods since the zeros of $G^+(z)$ are far from the unit circle.

The second test involves a zero pair very close to the unit circle. The zeros are at radius $0.99^{\frac{1}{2}}$ and angle $\pm\pi/4$. We have

$$G^+(z) = 3.9401 - 2.786z^{-1} + 0.99z^{-2}$$
$$B(z) = 1 - 1.4z^{-1} + 0.99z^{-2}$$

Table 1.2 gives the results for this test:

Cholesky		
Steps	Error	Time
32	0.0204145	0.40
64	0.0087674	0.85
128	0.0030524	1.50
256	0.0006634	3.33
512	0.0000470	6.55
1024	0.0000023	12.47
2048	0.0000022	25.27

FFT			
Size	Error	Time	Aliasing
32	0.0350000	0.07	0.998
64	0.0028618	0.17	0.905
128	0.0016984	0.32	0.965
256	0.0002847	0.60	0.860
512	0.0005897	1.38	0.282
1024	0.0000265	2.53	0.293
2048	0.0000263	5.37	0.019
4096	0.0000275	10.78	0.000709
8192	0.0000295	22.63	0.000760

Table 1.2. Comparison between the Cholesky and FFT factorization methods for example 2. The columns of the table have the same meaning as described in Table 1.1.

Table 1.2 shows that the error levels off in each method as the computational effort is increased. The Cholesky method reaches an asymptotic error which is an order of magnitude smaller than that for the FFT method. This may be due to the square-root normalization employed in the Cholesky method. However, the FFT obtains its final error of 3×10^{-5} with a compute time of 2.5 sec., while the Cholesky method requires more than 7 seconds to achieve this error. Since the Cholesky method has complexity of order N while the FFT method has complexity of order $N \log N$, the relative efficiency of the FFT method cannot hold as N becomes arbitrarily large. It should be noted that on the F2, the square roots in the Cholesky method add significantly to the coefficient of N in its complexity.

Although the details are not presented here, an example was tried using an order eight $B(z)$ (example 2 of [172]). In this case, the asymptotic error of the Cholesky method was near 0.002 while the FFT method achieved 0.0009 for size 1024 (in spite of a time-aliasing level of 0.6!). Thus the relative numerical performance is problem dependent.

In conclusion, the FFT method appears to be more cost-effective for rough approximations in a typical general-purpose computing environment (when storage minimization is

not an issue), giving about an order of magnitude smaller error for a given computation time. Also, the FFT method is probably the best choice for minicomputer/array-processor work-stations. Both methods are well-suited for initializing a quadratically convergent algorithm such as Wilson's method. For hardware systems, the Cholesky method with square-root table look-ups is probably the best choice, unless an FFT facility is available for other reasons.

Sharpening the Factorization

In this section, a method for sharpening the precision of a polynomial spectral factor is given. This is needed when spectral factorization methods, such as those of the previous section, reach an asymptotic error level which is too high above machine precision to be acceptable.

There is an iterative algorithm for computing the square root x of a positive real number $\xi = x^2$ which proceeds as follows:

Given $x_0 > 0$, compute x_k as

$$x_{k+1} \leftarrow \frac{1}{2}\left(x_k + \frac{\xi}{x_k}\right).$$

It is known that

$$\lim_{k\to\infty} x_k = x = \xi^{\frac{1}{2}},$$

and that the number of correct significant digits doubles each iteration [168].

The obvious generalization of this recursion was applied to the spectral factorization problem:

Given $B_0(z)$, compute $B_k(z)$ as

$$B_{k+1}(z) \leftarrow \frac{1}{2}\left(B_k(z) + \frac{G^+(z)}{z^{n_b} B_k(z^{-1})}\right). \tag{1.27}$$

For example 2 above, this algorithm gave about an order of magnitude improvement in the error, but it too suffers from slow asymptotic convergence (i.e., the error did not approach the machine epsilon even with 10000 iterations). For the order 8 case of [172], used to sharpen the result of a size 1024 FFT method, it gave only about a factor of 3 improvement in the error after 1024 iterations. Thus the convergence is worse than linear (according to empirical observations). However, it is about twice as fast per iteration as the Cholesky method for low order $B(z)$, and it can be used to obtain better results than can be had with the FFT or Cholesky method alone in these cases. Note that the recursion

(1.27) cannot be used without a good initialization since it has no means of placing all the roots inside the unit circle in one factor. A fixed-point of the recursion is obtained for any factorization of the form $G(z) = B(z)B(z^{-1})$, and there are 2^{n_b} such factorizations, corresponding to different groupings of the roots of $G(z)$ inside and outside the unit circle.

1.10. Summary and Conclusions

The rational filter-design problem, for general continuous desired spectra, has been examined in the frequency domain. The initial goal was to minimize some norm of the frequency-response error $\|H(e^{j\omega}) - \hat{H}(e^{j\omega})\|$. It was demonstrated that a best solution always exists (provided that the filter poles be restricted away from the unit circle), but that the solution is not generally unique. The general problem was shown to be difficult to solve due to the existence of an arbitrary number of locally best approximations. In spite of this discouraging property, a stable, unique solution can always be computed when $H(z)$ is of the form $Y(z)/U(z)$ (i.e. rational) using the Hankel error norm.

Several sub-optimal formulations were described. The minimization of L^2 ratio-error led to exceedingly simple "linear prediction" algorithms whose solutions are unique, robust, and useful in practice. They tend to model the spectral envelope, disregard phase, and are not optimum in any normal sense when both poles and zeros are designed. In certain contexts they minimize linear prediction-error energy.

The minimization of equation-error also gave a simple method which designs poles and zeros, and fits both phase and magnitude. Drawbacks to equation-error design include (1) weighting of the frequency-response error by $|\hat{A}(e^{j\omega})|$, the denominator of the designed filter, (2) extreme sensitivity to excess phase in the desired impulse response (i.e., minimum phase is almost required), and (3) lack of assured stability; when the design is unstable, one must reflect poles inside the unit circle and give up the phase fit, or the filter must be re-designed after adding a linear-phase term to the desired phase response.

The fact that real rational approximation is solved for the Chebyshev error norm was exploited to obtain a method for power frequency-response approximation. This method dovetailed nicely with an approximate log-power approximation scheme which is valuable for audio and video applications.

Various other techniques were mentioned with the intent of classifying them within the general framework adopted here. The use of first-order conformal maps for frequency-scaling and error-weighting in digital filter design was described. Finally, a fast spectral factorization method was presented.

As is usual in a general setting, there is no one method which out-performs the others for all problems. As a general rule, if one requires a uniform fit of both phase and magnitude,

the Hankel-norm method should be considered. While the Hankel-norm method is the strongest in the sense of solving the ideal filter-design problem defined in the introduction, it is the most computationally expensive and numerically demanding. If the desired frequency response is close to minimum phase, and smooth, then the equation-error method may be an adequate choice. If phase is irrelevant, then high-quality magnitude fits can be obtained using the Chebyshev methods (or equation-error method with constructed minimum phase). For audio applications in which phase is unimportant, weighted conformally-mapped log-magnitude approximation may be the most desirable (cf. Chapter 3).

In principle, rational filter design is still an open problem when the desired frequency response is allowed to be an arbitrary continuous complex curve. By this is meant there is no algorithm which is guaranteed to find an optimum solution without exhaustive search over all possible filter coefficients. In practice, however, the Hankel-norm method has essentially solved it, albeit with relatively cumbersome computations. A practical solution of the problem under other (perhaps more desirable) norms, is still outstanding. Also, there probably exists a significantly more efficient Hankel-norm method, for there is much structure as yet unexploited in the currently existing algorithms.

Chapter 2

Methods for System Identification

2.1. Introduction

In this chapter, several prevalent identification algorithms are derived and compared in a unified way. The point of departure will be a time-invariant filter-design problem which is adapted to the time-varying case by casting the design algorithm into time-recursive form. The necessity of time-recursive form severely restricts the choice of error minimized by the algorithm. In the terminology of Chapter 1, only the L^2 *norms* of *equation error* and *ratio error* will be employed.

In the previous chapter, the filter design problem was discussed almost exclusively in the frequency domain. This allowed maximum flexibility in minimizing frequency-response error. In other ways, however, frequency-domain design is restrictive. For example, it may be desired to represent a system which changes over time. In these cases it is more effective to ascertain the model from time-domain input-output data rather than from the frequency-response (which, strictly speaking, no longer exists). Even for the case of a time-invariant filter, a specification in terms of input-output data may arise in practice, such as when the spectrum of the input signal is not invertible. Another advantage of time-domain formulation is that the modeling error appears as a *signal* which itself may be further explored for structure, or statistically summarized and included in the model.

The present chapter discusses filter design in the time domain from the *system identification* point of view. System identification is broadly defined as the determination of a mathematical model for a dynamic system on the basis of input-output data. The majority of activity in this area, however, is concerned with fitting *rational linear* filters—the same as in Chapter 1. The main difference is that they are formulated exclusively in the time domain to allow tracking of time-varying systems. Nonetheless, in system identification *theory*, systems are modeled as slowly changing *time-invariant* filters, and the analysis is carried out for constant parameters. Thus the general principles discussed in Chapter 1 still apply. An elementary introduction to system identification is given in [99,81]. The book by Goodwin and Payne [95] gives a wide-ranging overview of the field and contains 250 references to the literature. The very recent book by Ljung and Soderstrom [111] gives

extensive information on the formulation and analysis of recursive-in-time identification algorithms.

Another aspect of the system identification approach is that it is usually concerned with modeling *stochastic processes*. The measured input and output signals are typically assumed to be correlated Gaussian noise, and the error signal in the time domain is invariably some form of *prediction error*. The ultimate goal of the model is to decorrelate, or "whiten" the prediction error, for the best that can be done is to "explain" all correlation structure in the input-output data with the model. While this setting has extensive application in practice, it will not be pursued in this chapter. The reason is that a somewhat deterministic approach will more clearly bring out connections between identification algorithms and the methods of Chapter 1.

It should be noted that this dissertation is concerned with the *quality of the model* and not its performance in-context. For example, identification algorithms are used in *adaptive control* to adjust the parameters of a controller in real time. In such a situation, the error to be minimized is a *control performance error*, and this can be quite different from minimizing a norm of the frequency-response error. Thus the full capability of identification algorithms will not be utilized in the context of time-domain filter design.

2.2. Summary of Chapter 2

First the basic system-identification problem is defined, followed by its solution in a simple case. This solution is equivalent to the equation-error method discussed in Chapter 1. Next the model is extended to allow further modeling of the equation error as linearly filtered white noise. This extension gives rise to almost all of the different identification algorithms, and their often subtle distinctions, for it poses a new and difficult nonlinear optimization problem. As we saw in Chapter 1, choice of equation error leads to a simple solution. The further modeling of equation error as an ARMA* process makes the problem difficult again. The error criterion is examined in the frequency domain to illustrate connections with the methods of Chapter 1. Some issues in "reduced-order" modeling are also considered.

Two main problem formulations are discussed, based on linear regression and gradient descent techniques. The regression formulation is used to derive the well-known methods of Instrumental Variables, Weighted Least Squares, Generalized Least Squares, and Extended Least Squares. In addition, the regression formulation is given for the multi-input, multi-output case since it is almost no extra work to do so. Next, the Gauss-Newton optimization method is applied to yield a generalized identification algorithm which reduces to the methods of Maximum Likelihood, a new form of Instrumental Variables, and Extended Least Squares, when certain approximations are applied to quantities in the general algorithm. To round out the picture of identification methods for system modeling, a comparison is given between the robust equation-error method and the nonlinear output-error method represented by the Steiglitz-McBride algorithm. The algorithms are converted to time-recursive form by a single substitution in the exact recursive "off-line" algorithm. This is believed to provide a simpler derivation of the recursive identification algorithms than was previously available. Furthermore, the derivation is applied to a generalized identification algorithm, giving a new class of recursive methods. Finally, methods for reducing the computational complexity of the recursive updates are reviewed, and convergence acceleration schemes, based on improving the recursively estimated gradient, are discussed.

2.3. The Identification Problem

* "Autoregressive moving average" meaning "rationally filtered white noise."

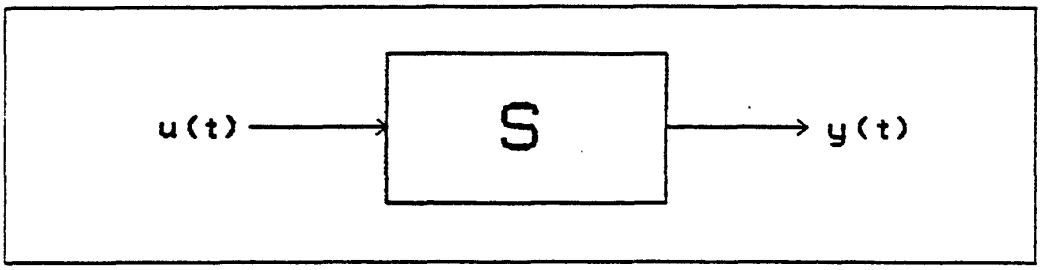

Figure 2.1. General schematic of the system identification problem. The system S is identified given input signal u_t and output signal y_t.

Figure 2.1 gives a conceptual schematic for the system identification problem. We have a system S with input signal u_t, and output signal y_t. Given u_t and y_t, for some range of time t, we wish to deduce S.

2.3.1. Choice of Model Structure

We consider only the class of stable, linear, time-invariant models which can be represented by a difference equation of the form

$$\begin{aligned} y_t = {} & b_1 u_{t-1} + \cdots + b_{n_b} u_{t-n_b} \\ & - a_1 y_{t-1} - \cdots - a_{n_a} y_{t-n_a} \end{aligned} \tag{2.1}$$

This is a set of causal digital filters with n_a poles and $n_b - 1$ zeros. The input and output signals u_t,y_t are considered to be real-valued scalar processes, unless otherwise noted, and the discrete-time index t is always taken to be an integer. The methods to be discussed can be extended to the multi-variable case without essential difficulty. The absence of a direct path from u_t to y_t is not necessary for identification, but is customary in control to avoid feedback degeneracies when u_t is made a function of y_t. This convention will be observed in this chapter since it is ubiquitous in the identification literature. To connect with the notation of chapter 1, a one-sample shift of the time origin of the input signal u_t is all that is required. We assume n_a and n_b are known, since order identification is typically a by-product of parameter estimation for a variety of orders.

Denoting the unit sample delay operator by d, ($d^n y_t = y_{t-n}$), we may rewrite (2.1) as

$$A(d) y_t = B(d) u_t , \tag{2.2}$$

where

$$\begin{aligned} A(d) &= 1 + a_1 d + \cdots + a_{n_a} d^{n_a} \\ B(d) &= b_1 d + \cdots + b_{n_b} d^{n_b} . \end{aligned} \tag{2.3}$$

Note that the delay-operator polynomials $A(d)$ and $B(d)$ are closely related to the z transforms of the coefficient sequences. If the z-transform of the time-series x_t is defined by

$$X(z) = \sum_{t=-\infty}^{\infty} x_t z^{-t}, \tag{2.4}$$

then in the frequency domain,

$$A(z^{-1})Y(z) = B(z^{-1})U(z), \tag{2.5}$$

where $U(z)$ and $Y(z)$ are the z-transforms of the time sequences u_t and y_t, respectively. The model is stable when all roots of $A(z^{-1})$ lie inside the unit circle of the complex plane.

The model equation (2.2) may also be written

$$y_t = \frac{B(d)}{A(d)} u_t \tag{2.6}$$

where the meaning of division by a delay-operator polynomial may be defined by polynomial "long division," or by the equivalence to z-transforms.

Yet another representation of the difference equation (2.1) which allows application of *linear regression* techniques is given by

$$y_t = \varphi_t^T \theta,$$

where

$$\begin{aligned} \varphi_t^T &= (-y_{t-1}, \ldots, -y_{t-n_a}, u_{t-1}, \ldots, u_{t-n_b}) \\ \theta^T &= (a_1, \ldots, a_{n_a}, b_1, \ldots, b_{n_b}). \end{aligned} \tag{2.7}$$

The linear regression formulation will be used extensively in deriving various identification algorithms.

2.3.2. Error Criterion

Now that a class of systems is specified, a *criterion* for judging model quality is needed. Let $\hat{A}(d)$ and $\hat{B}(d)$ denote the estimates of $A(d)$ and $B(d)$. Then given these estimates and the known input u_t, we can compute an estimate of y_t (denoted $\hat{y}_t$) by means of the difference equation (2.1). From this model output, we can form an error criterion in a number of ways. Two prevalent error definitions are the *output error* and *equation error*. Output error is defined by

$$\hat{v}_t = y_t - \frac{\hat{B}(d)}{\hat{A}(d)} u_t.$$

Output error is the same error as minimized in problem $\hat{H}^*$ of Chapter 1. Equation error is defined by

$$\hat{\epsilon}_t = \hat{A}(d)y_t - \hat{B}(d)u_t = y_t - \varphi_t^T\hat{\theta}, \tag{2.8}$$

where $\hat{\theta}^T = \left(\hat{a}_1, \ldots, \hat{a}_{n_a}, \hat{b}_1, \ldots, \hat{b}_{n_b}\right)$.

In addition to being *linear in the parameters* $\{\hat{a}_i, \hat{b}_i\}$, equation error has the interpretation as a *linear prediction error*, and this is appropriate for some applications such as adaptive control. On the other hand, when the model is to be estimated separately and used as a *proxy* for the true system, as opposed to a real-time signal-tracking situation, then output error is typically a better error measure. For example, consider the simple example where $\hat{B}(d) = 1$, $\hat{A}(d) = 1 - \hat{a}d$, and $u_t = \delta_t$, where δ_t is the Kronecker delta function.† Then for $t > 0$, $\hat{v}_t = y_t - \hat{a}^t$ and $\hat{\epsilon}_t = y_t - \hat{a}y_{t-1}$. Minimizing equation error obtains the value of $\hat{a}$ which is good for predicting a sample of y from the previous sample; it does not necessarily find an $\hat{a}$ which when raised to the power t provides a good approximation to y_t. Thus, equation error is desirable for tracking and predicting the output of systems in real time using the true past outputs, while output error is desirable for modeling the transfer function of the system.

In the algorithms to be discussed, the L^2 *norm of the equation error* will be used exclusively as a quantity to be minimized. This is due mainly to the fact (as discussed in Chapter 1) that output error is rarely solvable by gradient techniques. It is also due to the relative ease with which the equation error L^2 norm can be minimized recursively for time-varying applications. Near the end of this chapter, however, the Steiglitz-McBride algorithm for output-error minimization will be described.

Given data from time $t = 1$ to N, the estimate of θ is obtained by minimizing $J(\hat{\theta}_N)$ with respect to $\hat{\theta}_N$, where

$$J(\hat{\theta}_N) = \sum_{t=1}^{N} \hat{\epsilon}_t^2. \tag{2.9}$$

Note that in Chapter 1 notation, we would write J_E^2 for this loss function. In this chapter, however, J will suffice to denote the squared L^2 equation-error norm since a variety of norms will not be considered.

2.3.3. Least Squares Solution for the Noiseless Case

Using equation error, $J(\hat{\theta}_N)$ is a quadratic form in the parameter estimates $\hat{\theta}_N$. Quadratic minimization problems are readily solved by a linear system of equations. Com-

† Defined in Appendix E.

puting the gradient of $J(\hat{\theta}_N)$ with respect to $\hat{\theta}_N$ and equating to zero gives

$$\begin{aligned} 0 = \frac{\partial J(\hat{\theta}_N)}{\partial \hat{\theta}} &= \sum_{t=1}^{N} \frac{\partial \hat{\epsilon}_t^2}{\partial \hat{\theta}} = \sum_{t=1}^{N} 2\frac{\partial \hat{\epsilon}_t}{\partial \hat{\theta}}\hat{\epsilon}_t = \sum_{t=1}^{N} 2(-\varphi_t)(y_t - \varphi_t^T \hat{\theta}_N) \\ \Rightarrow \quad \sum_{t=1}^{N} \varphi_t \varphi_t^T \hat{\theta}_N &= \sum_{t=1}^{N} \varphi_t y_t . \end{aligned} \tag{2.10}$$

The solution is then

$$\hat{\theta}^{(LS)} = \left(\sum_{t=1}^{N} \varphi_t \varphi_t^T \right)^{-1} \sum_{t=1}^{N} \varphi_t y_t . \tag{2.11}$$

The solution exists and is unique whenever the above matrix inversion exists, and the solution is a global minimum of $J(\hat{\theta}_N)$ if the second derivative matrix is positive definite. The second derivative matrix (or *Hessian*) is given by

$$\frac{\partial^2 J(\hat{\theta}_N)}{\partial \hat{\theta}^2} = 2 \sum_{t=1}^{N} \varphi_t \varphi_t^T . \tag{2.12}$$

Since the Hessian is a sum of outer products (which is always positive semi-definite), we have that the solution $\hat{\theta}^{(LS)}$ is a global minimizer of $J(\hat{\theta}_N)$ whenever it exists.

If there are no errors in the measurement of y_t, no disturbances in the system except u_t, and if the true system S is exactly represented by the difference equation (2.1), then $\hat{\epsilon}_t$ can be made identically zero by taking N sufficiently large that the Hessian is invertible in (2.11). When u_t is such that the vectors $\{\varphi_t, t = 1, \ldots, n_a + n_b\}$ are linearly independent, then $N = n_a + n_b$ produces the solution. For example, u_t could be an impulse or white noise, or any other waveform which generates $n_a + n_b$ linearly independent φ vectors. In the identification literature, the condition that u_t be *persistently exciting* of order $n_a + n_b$ is imposed to ensure invertibility of the Hessian matrix [95].

2.3.4. Modeling Stochastic Input Components

In practice there are always some sources of inaccuracy such as instrumentation errors in the measurement of y_t and thermal or load disturbances in the system. In addition, most systems are not exactly representable by (2.1) due to nonlinearities, higher order than provided by the model, and/or the presence of essentially distributed parameters. Thus the set of equations generated by (2.1) for $N > n_a + n_b$ is inevitably inconsistent for a real problem.

Figure 2.2 gives the schematic for system identification in the presence of disturbance noise. In addition to u_t, S, and y_t, there are unknown disturbances which may or may not

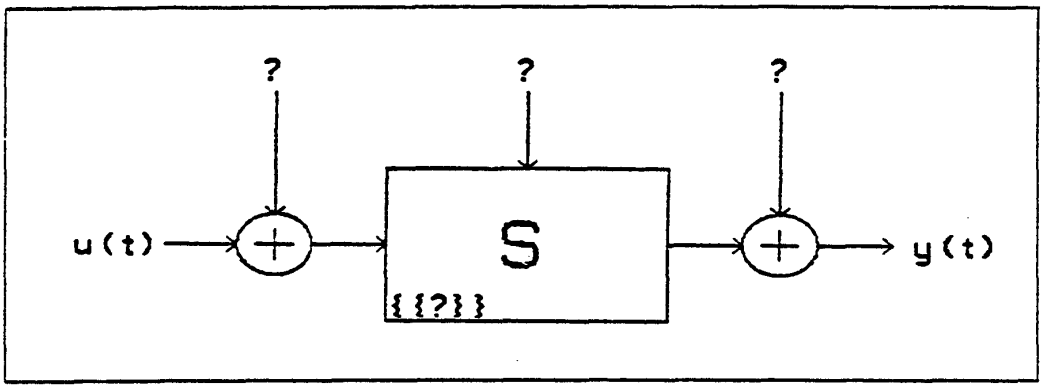

Figure 2.2. System Identification paradigm when noise is included.

be random in nature. Given u_t and y_t, we wish to deduce S in a way that is insensitive to the sources of noise. Since we are considering only linear time-invariant systems, it is meaningful to redraw Fig. 2.1 as shown in Fig. 2.3.

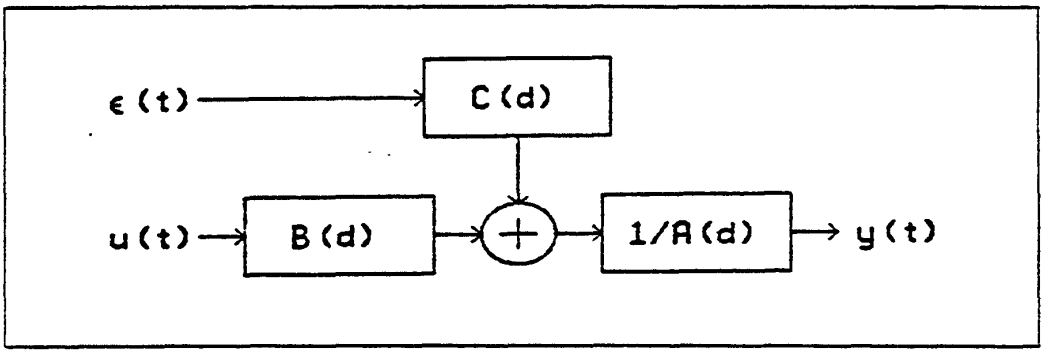

Figure 2.3. Implicit model structure assumed when minimizing equation-error to identify S.

Here the net disturbance is taken to be a single white noise source ϵ_t which arrives at the output via the transfer function $A(d)H_\epsilon(d)$. This figure is drawn according to the definition of equation error, i.e.,

$$A(d)y_t = B(d)u_t + H_\epsilon(d)\epsilon_t\,. \tag{2.13}$$

If $H_\epsilon(d)$ corresponds to a linear system of the form

$$H_\epsilon(d) = \frac{C(d)}{F(d)}, \tag{2.14}$$

where

$$\begin{aligned} C(d) &= 1 + c_1 d + \cdots + c_{n_c} d^{n_c} \\ F(d) &= 1 + f_1 d + \cdots + f_{n_f} d^{n_f}, \end{aligned} \tag{2.15}$$

then we can represent random equation-error disturbances with rational power spectral densities [83].

Interpretation of the Generalized Error

A unifying characteristic of all identification algorithms to be discussed is that

$$\hat{\epsilon}_t = \frac{\hat{A}(d)y_t - \hat{B}(d)u_t}{\hat{H}_\epsilon(d)}$$

is made to approach a white noise sequence. In other terms, the power spectral density of $\hat{\epsilon}_t$ is "flattened." Letting $H(d) = B(d)/A(d)$ and $\hat{H}(d) = \hat{B}(d)/\hat{A}(d)$, we find

$$\hat{\epsilon}_t = \frac{\hat{A}(d)}{\hat{H}_\epsilon(d)}\Big(y_t - \hat{H}(d)u_t\Big) = \frac{\hat{A}(d)}{\hat{H}_\epsilon(d)}\Big(H(d) - \hat{H}(d)\Big)u_t + \frac{\hat{A}(d)}{A(d)}\frac{H_\epsilon(d)}{\hat{H}_\epsilon(d)}\epsilon_t.$$

If ϵ_t is uncorrelated with u_t, as is commonly assumed, then the error functional minimized is given by

$$J = \int_{-\pi}^{\pi} \left|\frac{\hat{A}(e^{j\omega})}{\hat{H}_\epsilon(e^{j\omega})}\right|^2 \left|H(e^{j\omega}) - \hat{H}(e^{j\omega})\right|^2 U(e^{j\omega})\frac{d\omega}{2\pi} + \sigma_\epsilon^2 \int_{-\pi}^{\pi} \left|\frac{\hat{A}(e^{j\omega})}{A(e^{j\omega})}\right|^2 \left|\frac{H_\epsilon(e^{j\omega})}{\hat{H}_\epsilon(e^{j\omega})}\right|^2 \frac{d\omega}{2\pi},$$

where $U(e^{j\omega})$ is the power spectral density of the known input u_t, and σ_ϵ^2 is the variance of ϵ_t.

From this expression, we see that the frequency-response error $H - \hat{H}$ is weighted by $U|\hat{A}|/|\hat{H}_\epsilon|$, similar to the *equation error* formulation of chapter 1, and the noise path is modeled by minimizing a form of *ratio error*. By defining the noise transfer function in the "output error" sense,

$$y_t = H(d)u_t + H_v(d)v_t \;\Rightarrow\; H_v(d) \triangleq \frac{H_\epsilon(d)}{A(d)},$$

the error criterion appears as

$$J = \int_{-\pi}^{\pi} \left|H(e^{j\omega}) - \hat{H}(e^{j\omega})\right|^2 \frac{U(e^{j\omega})}{\left|\hat{H}_v(e^{j\omega})\right|^2}\frac{d\omega}{2\pi} + \sigma_v^2 \int_{-\pi}^{\pi} \left|\frac{H_v(e^{j\omega})}{\hat{H}_v(e^{j\omega})}\right|^2 \frac{d\omega}{2\pi},$$

and in this form, the weighted-output-error and ratio-error components are simplified.

Special Cases

For example, if the only significant uncertainties arise from measurement errors in y_t which are uncorrelated, then the appropriate form of $H_\epsilon(d)$ is

$$H_\epsilon(d) = A(d). \qquad (\textit{White Measurement Error})$$

This case coincides with problem $\hat{H}^*$ of Chapter 1 for the L^2 norm.

Internal disturbances may enter the feedback loops of the system through paths different from the input u_t. If there are no resonant modes other than those of the system, and if the internal noise source is white, equation (2.14) reduces to

$$H_\epsilon(d) = C(d). \qquad (\textit{White Internal Disturbance})$$

Thus the noise transfer function $H_\epsilon(d)$ is a polynomial. When this is not the case, we may estimate $H_\epsilon(d)$ as a polynomial as long as the orders of $A(d)$, and $B(d)$ are sufficiently augmented. That is,

$$\begin{aligned} & A(d)y_t = B(d)u_t + \frac{C(d)}{F(d)}\epsilon_t \\ \Rightarrow \quad & A(d)F(d)y_t = B(d)F(d)u_t + C(d)\epsilon_t \\ \Rightarrow \quad & A'(d)y_t = B'(d)u_t + C(d)\epsilon_t, \end{aligned} \qquad (2.16)$$

and the poles belonging exclusively in the noise transfer function appear as common factors in $A'(d)$, $B'(d)$, or "pole-zero cancellations." In practice, this will occur only approximately, and common factors must be defined in terms of closeness of the poles and zeros. A different phenomenon which complicates this procedure occurs when the model is of larger order than the real system; superfluous poles and zeros in the identification process also tend to cancel each other, although they have been observed in practice to do so near the origin in the z-plane.

In the event that the equation error $Ay - Bu$ is itself a white noise sequence (i.e. $H_\epsilon(d) = 1$, corresponding physically to white noise injection at a single feedback summation node or similar special circumstance), then minimization of $J(\hat{\theta}_N)$ to obtain the solution $\hat{\theta}^{(LS)}$ given by (2.11) results in a *strongly consistent* estimate of the true parameters θ [106]; that is,

$$\hat{\theta}_N^{(LS)} \underset{N\to\infty}{\to} \theta \qquad \text{w.p.1.}$$

In contrast, *correlated* equation error ($H_\epsilon(d) \neq 1$) causes *asymptotic bias* in the parameter estimates; ($\hat{\theta}^{(LS)}$ will not approach θ as the measurement time span is increased, and is said

to be an *inconsistent* estimator). Thus it seems best to include some modeling of the noise in order to guard against bias in the model-complete case.* The remainder of this chapter will discuss a class of techniques for identifying parameters of the model

$$A(d)y_t = B(d)u_t + C(d)\epsilon_t \tag{2.17}$$

which allows for the possibility of correlated noise. The symbol ϵ_t will always stand for an uncorrelated sequence, and correlated equation error will be denoted by ε_t.

2.3.5. Reduced-Order Identification

There is one particularly likely case in which Fig. 2.3 is misleading, and that is when the model cannot represent the true system. In this case, ϵ_t is very much a function of u_t, and even when ϵ_t can be made pseudo-uncorrelated by some choice of $H_\epsilon(d)$, the results are u_t dependent. To see this, consider the system given by

$$A(d)y_t = B(d)u_t \tag{2.18}$$

and the model

$$\hat{A}(d)y_t = \hat{B}(d)u_t + \hat{\epsilon}_t\,. \tag{2.19}$$

Then we can solve for $\hat{\epsilon}_t$ as

$$\hat{\epsilon}_t = \hat{A}(d)\left(\frac{B(d)}{A(d)} - \frac{\hat{B}(d)}{\hat{A}(d)}\right)u_t\,. \tag{2.20}$$

Hence, minimizing the energy of $\hat{\epsilon}$ depends heavily on the input u_t, unless it is possible to obtain $B(d)/A(d) = \hat{B}(d)/\hat{A}(d)$.

In the case of a reduced-order model, it is not likely that minimizing equation error will recover any true system parameters, since the reduced-order poles and zeros must distribute between the actual poles and zeros in a compromising fashion. In some cases, such as in problems of control, one does not require physically meaningful model parameters but rather the best *simulation* of the system in terms of output prediction. In this case, ϵ_t may be viewed as a *prediction error*, and the minimization of its energy is still optimum,

* The model-complete case is defined as the situation in which the true system is representable by the model in the absence of noise, i.e., there exists some θ such that when $\hat{\theta} = \theta$ the model is exact. This is a ubiquitous assumption in the analysis of identification algorithms, especially output-error methods. It is not, however, often realistic. When the model set is not complete, the problem is sometimes called the "reduced-order modeling" situation.

from a linear prediction point of view, even when the model cannot represent the system. In addition, the prediction error paradigm can be applied to a much larger class of systems than we consider here, such as nonlinear and distributed systems [82,85,86,87,108,95].

With extra care, it is still possible in some cases to identify true system parameters with a reduced-order model. The essential idea is to eliminate influence of some system modes from the measured input-output data. For example, suppose the system has an indefinite number of resonances in its frequency response, and identification of only the center frequency and bandwidth of the first few resonances is desired. Examples of such systems include vibrating strings and reverberant rooms. The object is to perform identification in the frequency band of interest without sensitivity to high-frequency characteristics of the system. If the input u_t is at our disposal, then we may select u_t to be *bandlimited* such that only the first few resonances are excited. The level of disturbance noise present and the accuracy of identification desired determine the energy of u_t required [95]. If the disturbance noise is too strong to be dominated by a large input signal, or if the input signal cannot be chosen to be bandlimited, then the nearly same effect may be had by *lowpass filtering* the input and output signals u_t, y_t[97]; in this case the high-frequency disturbance noise is removed from y_t as well as the high-frequency system dynamics.

While lowpass filtering u_t and y_t may serve to enable identification of the low-frequency system poles, the low-frequency zeros are not generally found. For example, suppose the system is given by

$$\begin{aligned} y_t &= \frac{B(d)}{A(d)}u_t + \frac{C(d)}{F(d)}\epsilon_t \\ &= \frac{B(d)W(d)}{A(d)V(d)}u_t + \frac{C(d)S(d)}{F(d)T(d)}\epsilon_t \\ &= \frac{B'(d)}{A(d)}u_t + \frac{W'(d)}{V(d)}u_t + \frac{C'(d)}{F(d)}\epsilon_t + \frac{S'(d)}{T(d)}\epsilon_t, \end{aligned} \tag{2.21}$$

where the low-frequency poles are contained in the polynomials $A(d)$ and $F(d)$, and the low-frequency zeros are grouped in $B(d)$ and $C(d)$. Then linearly lowpass filtering y_t, u_t to obtain $\bar{y}_t, \bar{u}_t$, such that the high-frequency modes are suppressed, gives

$$\bar{y}_t = \frac{B'(d)}{A(d)}\bar{u}_t + \frac{C'(d)}{F(d)}\bar{\epsilon}_t, \tag{2.22}$$

where $\bar{\epsilon}_t$ is defined by the above equation. If this reduced scenario can be exactly identified, then the poles are correct but the zeros have moved such that

$$\begin{aligned} B(d) &= B'(d)V(d) + W'(d)A(d) \\ C(d) &= C'(d)T(d) + S'(d)F(d). \end{aligned}$$

The filtering of the measurements has caused this apparent change in the zeros of the system, for each zero is a function of all resonances present. The benefit of pre-filtering is providing that the system lie in the model space, which allows $\bar{e}_t$ to be meaningfully assumed uncorrelated with $\bar{u}_t$.

In summary, a reduced-order model may be fit to "reduced-order measurements." By identifying system poles in relatively narrow frequency intervals, the modes of a large system can be found using only low-order identification algorithms. This is an important option to keep in mind because numerical experience indicates that large order systems (greater than say 25 poles with 36-bit floating arithmetic) can be difficult to identify recursively (described later) due to round-off error.

2.4. The Regression Formulation

In the following section, we treat the multi-input/output case, since it is formally so similar to the single-input/output case. If there are p outputs and q inputs, it is useful to set up the quantities such that y_t is $p \times 1$, u_t is $q \times 1$, a_i is $p \times p$, and b_i is $p \times q$.

Consider the equation error formulation

$$A(d)y_t = B(d)u_t + e_t \tag{2.23}$$

where e_t may be correlated even when $A(d)$ and $B(d)$ are correct. Given estimated quantities, we will write this equation as

$$\hat{A}(d)y_t = \hat{B}(d)u_t + \hat{e}_t\,. \tag{2.24}$$

Casting (2.23) in the regression form as before in equation (2.7) gives

$$y_t^T = \varphi_t^T\theta + e_t^T\,, \tag{2.25}$$

where

$$\begin{aligned} \varphi_t^T &= \left(-y_{t-1}^T, \ldots, -y_{t-n_a}^T, u_{t-1}^T, \ldots, u_{t-n_b}^T\right) && (1 \times pn_a + qn_b) \\ \theta^T &= (a_1, \ldots, a_{n_a}, b_1, \ldots, b_{n_b}) && (p \times pn_a + qn_b) \end{aligned} \tag{2.26}$$

and e_t may be correlated. Let $N_p = pn_a + qn_b$ denote the number of elements of φ_t.

Writing out (2.25) for N successive samples beginning with $t = 1$ gives a system of N linear matrix equations in θ, viz.,

$$\begin{aligned} y_1^T &= \varphi_1^T\theta + e_1^T \\ y_2^T &= \varphi_2^T\theta + e_2^T \\ &\vdots \\ y_N^T &= \varphi_N^T\theta + e_N^T \end{aligned} \tag{2.27}$$

or

$$Y_N = \Phi_N\theta + E_N, \tag{2.28}$$

where

$$\begin{aligned} Y_N^T &= (y_1, \ldots, y_N) && (p \times N) \\ E_N^T &= (e_1, \ldots, e_N) && (p \times N) \\ \Phi_N^T &= (\varphi_1, \ldots, \varphi_N). && (N_p \times N) \end{aligned} \tag{2.29}$$

Since we wish to solve for θ, we convert to a square system of equations by premultiplying both sides of (2.29) by an $N_p \times N$ matrix which we denote by Z_N^T.

$$Z_N^T Y_N = Z_N^T \Phi_N \theta + Z_N^T E_N, \tag{2.30}$$

with

$$Z_N^T = (z_1, \ldots, z_N). \qquad (N_p \times N) \tag{2.31}$$

Now, if $Z_N^T\Phi_N$ is invertible, which is possible for $N \geq N_p$, then we can write

$$\theta = \left(Z_N^T\Phi_N\right)^{-1} Z_N^T Y_N - \left(Z_N^T\Phi_N\right)^{-1} Z_N^T E_N. \tag{2.32}$$

Since E_N is unknown, it is natural to define the estimate of θ by

$$\hat{\theta}_N = \left(Z_N^T\Phi_N\right)^{-1} Z_N^T Y_N = \left(\sum_{t=1}^{N} z_t\varphi_t^T\right)^{-1} \sum_{t=1}^{N} z_t y_t^T. \tag{2.33}$$

The error associated with this estimator is then

$$\tilde{\theta}_N = \hat{\theta}_N - \theta = \left(Z_N^T\Phi_N\right)^{-1} Z_N^T E_N. \tag{2.34}$$

The parameter estimate is perfect whenever

$$Z_N^T E_N = 0. \tag{2.35}$$

Geometrically, this condition may be interpreted as the requirement that each column of E_N (i.e., the sequence of equation errors for each component of y_t) be orthogonal to all N_p rows of Z_N^T. It can be shown that $Z_N^T \hat{E}_N = 0$ always holds, where $\hat{E}_N = Y_N - \Phi_N^T \hat{\theta}_N$, and that $\hat{E}_N^T \hat{E}_N$ is minimum for $Z_N^T = \Phi_N^T$. In this solution, Z_N "steers" the projection onto the rows of Φ_N such that the line of projection from Y_N to $\hat{Y}_N = \Phi_N^T \hat{\theta}_N$ is orthogonal to Z_N.

2.4.1. The Instrumental Variables Technique

Above, we showed that $\hat{\theta}_N = \theta$ whenever $Z_N^T E_N = 0$. Less stringent conditions which ensure that $\hat{\theta}_N$ *asymptotically* approaches θ are

$$\begin{aligned} &1) \quad \lim_{N\to\infty} det\left(\frac{1}{N}\sum_{t=1}^{N} z_t \varphi_t^T\right) \neq 0 \\ &2) \quad \lim_{N\to\infty} \frac{1}{N}\sum_{t=1}^{N} z_t e_t^T = 0. \end{aligned} \tag{2.36}$$

Since we may interpret the normalized sum of $z_t \varphi_t^T$ as the sample correlation of z_t and φ_t, condition 1) requires that the vectors z_t and φ_t be correlated in all components. Similarly, condition 2) states that the vectors z_t and e_t must be uncorrelated.

Any matrix Z_N^T satisfying 1) and 2) above is called an *instrumental variables matrix*, and the N_p elements of z_t are called the *instrumental variables*. We denote estimators of the instrumental variables class by $\hat{\theta}^{(IV)}$.

2.4.2. Choice of Instrumental Variables

When the identification is performed open-loop, and when the model can represent the true system, then it is reasonable to expect that the disturbance e_t is uncorrelated with the input. (If the model cannot truly represent the system, then e_t must contain modeling error as well as disturbance noise, and the modeling error should not be assumed independent of the input.) In this case, a plausible choice for the instrumental variables is u_t for an appropriate range of t. *E.g.*, set

$$z_t^T = \left(u_{t-n_b-1}^T, \ldots, u_{t-n_b-n_a}^T, u_{t-1}^T, \ldots, u_{t-n_b}^T\right), \tag{2.37}$$

and for "almost all u_t," (2.36) is satisfied [88]. If u_t may be chosen arbitrarily, then a good choice is independently generated white noise. This has the two-fold advantage of persistently exciting all system modes [95], and of being uncorrelated with e_t. This special case of instrumental variables is often called a *correlation method* [99].

The choice of z_t most commonly associated with the instrumental variables method is

$$z_t^T = \left(\hat{y}_{t-1}^T, \cdots, \hat{y}_{t-n_a}^T, u_{t-1}^T, \cdots, u_{t-n_b}^T\right) \tag{2.38}$$

where

$$\begin{aligned}\hat{y}_t = &\hat{b}_1 u_{t-1} + \cdots + \hat{b}_{n_b} u_{t-n_b} \\ &- \hat{a}_1 \hat{y}_{t-1} - \cdots - \hat{a}_{n_a} \hat{y}_{t-n_a}.\end{aligned} \tag{2.39}$$

The idea here is to produce a $\hat{y}_t$ which is uncorrelated with e_t yet correlated with y_t, thus satisfying (2.36). Since u_t is already presumed uncorrelated with e_t, we simply derive $\hat{y}_t$ from u_t in some reasonable fashion, in this case from an *a priori* model estimate such as might be given by $\hat{\theta}^{(LS)}$.

In the case of *recursive* instrumental variables, in which model estimates are updated for each t, one may calculate $\hat{y}_t$ via

$$\begin{aligned}\hat{y}_t = &\hat{b}_1(t-\tau_e) u_{t-1} + \cdots + \hat{b}_{n_b}(t-\tau_e) u_{t-n_b} \\ &- \hat{a}_1(t-\tau_e) \hat{y}_{t-1} - \cdots - \hat{a}_{n_a}(t-\tau_e) \hat{y}_{t-n_a}.\end{aligned} \tag{2.40}$$

The delay parameter τ_e is chosen sufficiently large so that the parameter estimates from time $t - \tau_e$ are not significantly correlated with e_t. For example, when it is assumed that $e_t = C(d)\epsilon_t$, where ϵ_t is white, and $C(d)$ is a filter polynomial of order n_c, then we set $\tau_e > n_c$. Another strategy for suppressing the fact that $\hat{\theta}$ is a function of $\hat{\epsilon}$ is to lowpass filter $\hat{\theta}$ to make it somewhat constant relative to $\hat{\epsilon}$. Further aspects of the instrumental variables method may be found in [130,131,88,99,95,81]. The form (2.38) is also the basis for Landau's "output error" method [102]; note, however, that our definition of output error is very different from that used in the model-reference control literature.

2.4.3. Return to Least Squares

If the equation error at time s for the optimum model is uncorrelated with u_t and y_t for $t < s$ (or more generally, for $s - \max\{n_a, n_b\} \leq t < s$), then it follows that φ_t is uncorrelated with e_t. Also, the vector "most correlated" with φ_t is φ_t itself. Thus in the case of uncorrelated equation errors, we may satisfy (2.36) by choosing $Z_N = \Phi_N$, and the resulting estimator is just

$$\hat{\theta}^{(LS)} = \left(\Phi_N^T \Phi_N\right)^{-1} \Phi_N^T Y_N = \left(\sum_{t=1}^{N} \varphi_t \varphi_t^T\right)^{-1} \sum_{t=1}^{N} \varphi_t y_t \tag{2.41}$$

which is the least squares estimator (2.11) obtained in the noiseless case.

2.4.4. Weighted Least Squares

Suppose the equation error e_t is correlated, but the covariance is known, *i.e*,

$$\mathcal{E}\left\{E_N E_N^T\right\} = R_{EE_N} = \Sigma_N \Sigma_N^T, \tag{2.42}$$

where Σ_N is the ***matrix square root*** of the covariance R_{EE_N} (obtainable by the Cholesky decomposition [168]). Then pre-multiplying both sides of the regression equation $Y_N = \Phi_N \theta + E_N$ (2.28) by Σ_N^{-1}, we obtain

$$\Sigma_N^{-1} Y_N = \Sigma_N^{-1} \Phi_N \theta + \Sigma_N^{-1} E_N$$

or

$$\tilde{Y}_N = \tilde{\Phi}_N \theta + \tilde{E}_N . \tag{2.43}$$

Now, since

$$\mathcal{E}\left\{\tilde{E}_N \tilde{E}_N^T\right\} = \mathcal{E}\left\{\Sigma_N^{-1} E_N E_N^T \Sigma_N^{-T}\right\} = \Sigma_N^{-1} \mathcal{E}\left\{E_N E_N^T\right\} \Sigma_N^{-T} = \Sigma_N^{-1} \Sigma_N \Sigma_N^T \Sigma_N^{-T} = I, \tag{2.44}$$

we may apply plain least squares (2.41) to the pre-multiplied regression equation (2.43) to obtain the consistent estimator

$$\begin{aligned} \hat{\theta} &= \left(\tilde{\Phi}_N^T \tilde{\Phi}_N\right)^{-1} \tilde{\Phi}_N^T \tilde{Y}_N = \left(\Phi_N^T \Sigma_N^{-T} \Sigma_N^{-1} \Phi_N\right)^{-1} \Phi_N^T \Sigma_N^{-T} \Sigma_N^{-1} Y_N \\ &= \left(\Phi_N^T R_{EE_N}^{-1} \Phi_N\right)^{-1} \Phi_N^T R_{EE_N}^{-1} Y_N . \end{aligned} \tag{2.45}$$

This is therefore an instrumental variables method with

$$Z_N^T = \Phi_N^T R_{EE_N}^{-1} . \tag{2.46}$$

It is also the optimal form of ***weighted least squares*** (WLS) [95]. In general, the weighted least squares technique allows an arbitrary matrix to replace R_{EE_N}. In the deterministic formulation of WLS, we have

$$Y_N = \Phi_N \theta + W_N \tilde{E}_N ,$$

where W_N is an arbitrary positive-definite weighting matrix. The general weighted least squares solution is given by

$$\hat{\theta}^{(WLS)} = \left(\Phi_N^T W_N^{-1} \Phi_N\right)^{-1} \Phi_N^T W_N^{-1} Y_N .$$

When W_N is diagonal, it has the effect of modulating the weight of the individual error samples $\hat{\epsilon}_t$, yielding the weighted least squares of Gauss [173].

Note that multiplying through the regression formula by Σ_N^{-1} above corresponds to *pre-filtering* y_t and u_t with linear time-varying filter which depends on the correlation function of e_t and on N. If R_{EE_N} is chosen to be lower-triangular, then this filtering is causal.

If e_t is stationary, then R_{EE_N} is *Toeplitz* in which case

$$R_{EE_N}^{-1} = L_1 U_1 + L_2 U_2 \tag{2.47}$$

where L_1,L_2 are lower-triangular Toeplitz matrices, of order N, and U_1,U_2 are upper-triangular Toeplitz matrices [114].

Thus, for stationary equation error e_t, the instrumental-variables pre-filtering by $R_{EE_N}^{-1}$ may be accomplished by two time-invariant filtering operations in parallel, each the cascade of a causal and anti-causal section.

If R_{EE_N} is *banded* of order n_e (i.e., $R_{EE_N}[i,j] = 0$ for $|i-j| > n_e$), then for $N \gg n_e$, it is the case that $R_{EE_N}^{-1}$ is approximately Toeplitz [144]. In this case, Σ_N^{-1} is asymptotically Toeplitz, and therefore the pre-filtering for the least squares method (2.43) is ultimately time-invariant as N goes to infinity.

An explicit form for the time-invariant pre-filtering which asymptotically decorrelates e_t is available from the model when the model is assumed to represent the system. If we assume the model (2.13) where $e_t = H_\epsilon(d)\epsilon_t$, then we have, in the scalar case,

$$A(d)\frac{1}{H_\epsilon(d)}y_t = B(d)\frac{1}{H_\epsilon(d)}u_t + \epsilon_t \tag{2.48}$$

or

$$A(d)y_t^f = B(d)u_t^f + \epsilon_t \tag{2.49}$$

where y_t^f and u_t^f are the pre-filtered data samples.

2.4.5. Generalized Least Squares

If we assume $H_\epsilon(d) = 1/F(d)$, then we have the relation $F(d)e_t = \epsilon_t$, which allows estimation of $F(d)$ from e_t by means of an *autoregression*. This is the basis of the method called *generalized least squares* [98,99,95,81]. We use the following loop of equations to

estimate θ:

$$\begin{aligned} &1) \quad \hat{u}_t^f = \hat{F}(d)u_t \\ &\qquad \hat{y}_t^f = \hat{F}(d)y_t \\ &2) \quad \hat{A}(d)\hat{y}_t^f = \hat{B}(d)\hat{u}_t^f + \hat{\epsilon}_t \\ &3) \quad \hat{\epsilon}_t = \hat{A}(d)y_t - \hat{B}(d)u_t \\ &4) \quad \hat{F}(d)\hat{\epsilon}_t = \hat{\epsilon}_t \end{aligned} \tag{2.50}$$

The initial value of $\hat{F}(d)$ is chosen arbitrarily (usually 1, i.e., $\hat{f}_i = 0, i = 1, \ldots, n_f$). Equation 1) is the pre-filtering step, 2) estimates $\hat{A}(d)$ and $\hat{B}(d)$ using the formula for $\hat{\theta}^{(LS)}$ (2.11), 3) computes the correlated equation error implied by the current estimates, and 4) estimates the $\hat{F}(d)$ corresponding to the current $\hat{\epsilon}_t$. These steps are then repeated until convergence is obtained.

2.4.6. Extended Least Squares

In the case of $H_\epsilon(d) = C(d)$, the pre-filtering is computed via

$$\begin{aligned} y_t^f &= \frac{1}{C(d)}y_t = y_t - c_1 y_{t-1}^f - \cdots - c_{n_c} y_{t-n_c}^f \\ u_t^f &= \frac{1}{C(d)}u_t = u_t - c_1 u_{t-1}^f - \cdots - c_{n_c} u_{t-n_c}^f, \end{aligned} \tag{2.51}$$

or equivalently,

$$\varphi_t^f = \frac{1}{C(d)}\varphi_t = \varphi_t - c_1\varphi_{t-1}^f - \cdots - c_{n_c}\varphi_{t-n_c}^f. \tag{2.52}$$

Thus if $C(d)$ can be estimated, we can apply standard least squares to the pre-filtered signals u_t^f, y_t^f in order to obtain a consistent estimate of $A(d)$ and $B(d)$.

Estimation of $C(d)$ may be accomplished by augmenting the regression formulation (2.25) such that

$$\begin{aligned} \hat{\varphi}_t^T &= \left(-y_{t-1}^T, \ldots, -y_{t-n_a}^T u_{t-1}^T, \ldots, u_{t-n_b}^T, \hat{\epsilon}_{t-1}^T, \ldots, \hat{\epsilon}_{t-n_c}^T\right) \\ \hat{\theta}^T &= \left(\hat{a}_1, \ldots, \hat{a}_{n_a}, \hat{b}_1, \ldots, \hat{b}_{n_b}, \hat{c}_1, \ldots, \hat{c}_{n_c}\right). \end{aligned} \tag{2.53}$$

in which case we again have the model

$$y_t = \theta^T\varphi_t + \epsilon_t \tag{2.54}$$

where the error ϵ_t is white, thus making it is permissible to "whiten" $\hat{\epsilon}_t$. However, we find that in this formulation the matrix $\Phi_N = (\hat{\varphi}_1, \ldots, \hat{\varphi}_N)$ contains $\hat{\epsilon}_0$ through $\hat{\epsilon}_{N-1}$, and

these values are determined by $\hat{\theta}_N$ which is what we are trying to compute. Thus we are forced into some sort of iterative or *relaxation* method [95] where we alternately compute $\hat{\theta}_N$ given $\hat{E}_N$ and $\hat{E}_N$ given $\hat{\theta}_N$. A number of options exist as to how this augmented estimation might be carried out, and the *gradient descent* formulation, to which we now turn, is most helpful in providing insight into the alternatives.

2.5. The Gradient Approach for Offline Identification

The derivations given above for the Instrumental Variables, Correlation, Weighted Least Squares, and Generalized Least Squares methods were all based on the regression formulation. An alternative approach to least squares identification is direct minimization of the cost function

$$J(\hat{\theta}_N) = \sum_{t=1}^{N} \hat{\epsilon}_t^T \hat{\epsilon}_t \tag{2.55}$$

using gradient descent techniques. This was done previously to get (2.11) for the scalar deterministic case. With this approach, we will obtain a simplified derivation for recursive Maximum Likelihood, Extended Least Squares, and Instrumental Variables.

In the case of white equation error ($e_t = \epsilon_t \Rightarrow n_c = 0$), the gradient calculation proceeds exactly as in the case (2.10) for deterministic least squares, and the resulting estimator is $\hat{\theta}^{(LS)}$ given by (2.11). As shown previously, if e_t is correlated with φ_t, then $\hat{\theta}^{(LS)}$ gives an inconsistent estimator for θ.

The more interesting case which we examine now is where $e_t = C(d)\epsilon_t$. Consider again the augmented regression (2.53) in which $\hat{C}(d)$ is appended to $\hat{\theta}$ and $\hat{\epsilon}$ is included in φ to give $\hat{\varphi}$. In this case, the minimization goes as in (2.10) except that the gradient of $\hat{\epsilon}_t$ with respect to $\hat{\theta}_N$, which we denote by $\hat{\psi}_t$, is more involved.

2.5.1. Computing the Gradient

For simplicity, we give the derivation of $\hat{\psi}_t$ for the case of scalar $\hat{\epsilon}_t$. We use the convention that the gradient of a scalar with respect to a vector is a column vector. Thus $\hat{\psi}_t = \partial\hat{\epsilon}_t/\partial\hat{\theta}$ is $N_p \times 1$ where now $N_p = n_a + n_b + n_c$. We have

$$\hat{\psi}_t = \frac{\partial \hat{\epsilon}_t}{\partial \hat{\theta}} = \frac{\partial}{\partial \hat{\theta}}\left(y_t - \hat{\varphi}_t^T\hat{\theta}\right) = -\frac{\partial \hat{\varphi}_t^T\hat{\theta}}{\partial \hat{\theta}} = -\hat{\varphi}_t - \frac{\partial \hat{\varphi}_t^T}{\partial \hat{\theta}}\hat{\theta} \tag{2.56}$$

where

$$\begin{aligned}\frac{\partial \hat{\varphi}_t^T}{\partial \hat{\theta}} &= \frac{\partial}{\partial \hat{\theta}}(-y_{t-1},\ldots,-y_{t-n_a},u_{t-1},\ldots,u_{t-n_b},\hat{\epsilon}_{t-1},\ldots,\hat{\epsilon}_{t-n_c}) \\ &= \left(0,\ldots,0,0,\ldots,0,\frac{\partial \hat{\epsilon}_{t-1}}{\partial \hat{\theta}},\ldots,\frac{\partial \hat{\epsilon}_{t-n_c}}{\partial \hat{\theta}}\right) \\ &= \left(0,\ldots,0,0,\ldots,0,\hat{\psi}_{t-1},\ldots,\hat{\psi}_{t-n_c}\right).\end{aligned} \tag{2.57}$$

Therefore,

$$\frac{\partial \hat{\varphi}_t^T}{\partial \hat{\theta}}\hat{\theta} = \hat{c}_1\hat{\psi}_{t-1} + \cdots + \hat{c}_{n_c}\hat{\psi}_{t-n_c} \tag{2.58}$$

which implies

$$\hat{\psi}_t = -\hat{\varphi}_t - \hat{c}_1\hat{\psi}_{t-1} - \cdots - \hat{c}_{n_c}\hat{\psi}_{t-n_c} = \frac{-\hat{\varphi}_t}{\hat{C}(d)} = -\hat{\varphi}_t^f. \tag{2.59}$$

or

$$\hat{C}(d)\hat{\psi}_t = -\hat{\varphi}_t. \tag{2.60}$$

Equation (2.59) gives a recursion for $\hat{\psi}_t$ using $\hat{C}(d)$, and we are now in a position to compute the gradient of $J(\hat{\theta}_N)$.

$$\begin{aligned}\frac{\partial J(\hat{\theta}_N)}{\partial \hat{\theta}} &= \sum_{t=1}^{N}\frac{\partial \epsilon_t^2}{\partial \hat{\theta}} = \sum_{t=1}^{N} 2\frac{\partial \epsilon_t}{\partial \hat{\theta}}\epsilon_t = \sum_{t=1}^{N} 2\hat{\psi}_t\left(y_t - \hat{\varphi}_t^T\hat{\theta}_N\right) = 2\sum_{t=1}^{N}\hat{\psi}_t y_t - 2\sum_{t=1}^{N}\hat{\psi}_t\hat{\varphi}_t^T\hat{\theta}_N \\ &= -2\sum_{t=1}^{N}\varphi_t^f y_t + 2\sum_{t=1}^{N}\varphi_t^f\hat{\varphi}_t^T\hat{\theta}_N.\end{aligned} \tag{2.61}$$

2.5.2. The Second-Derivative Matrix

A derivation similar to that of (2.61) gives the second derivative to be

$$\begin{aligned}\frac{\partial^2 J(\hat{\theta}_N)}{\partial \hat{\theta}^2} &= 2\sum_{t=1}^{N}\left(\hat{\psi}_t\hat{\psi}_t^T + \hat{\epsilon}_t\frac{\partial^2 \hat{\epsilon}_t}{\partial \hat{\theta}^2}\right) \\ &= 2\sum_{t=1}^{N}\hat{\psi}_t\hat{\psi}_t^T - 2\sum_{t=1}^{N}\hat{\epsilon}_t\left(\frac{1}{\hat{C}(d)}\left(\frac{\partial \hat{\varphi}_t^T}{\partial \hat{\theta}} + \left(\frac{\partial \hat{\varphi}_t^T}{\partial \hat{\theta}}\right)^T\right)\right)\end{aligned} \tag{2.62}$$

where

$$\begin{aligned}\frac{\partial \hat{\varphi}_t^T}{\partial \hat{\theta}} &= \left(0,\ldots,0,0,\ldots,0,\hat{\psi}_{t-1},\ldots,\hat{\psi}_{t-n_c}\right) \\ &= -\frac{1}{C(d)}(0,\ldots,0,0,\ldots,0,\hat{\varphi}_{t-1},\ldots,\hat{\varphi}_{t-n_c}).\end{aligned} \tag{2.63}$$

2.5.3. Solving for Extreme Points

Equating (2.61) to zero and solving for $\hat{\theta}_N$ gives

$$\hat{\theta}_N = \left(\sum_{t=1}^{N} \hat{\psi}_t \hat{\varphi}_t^T \right)^{-1} \sum_{t=1}^{N} \hat{\psi}_t y_t = \left(\sum_{t=1}^{N} \hat{\varphi}_t' \hat{\varphi}_t^T \right)^{-1} \sum_{t=1}^{N} \hat{\varphi}_t' y_t = \left(Z_N^T \Phi_N \right)^{-1} Z_N^T Y_N , \tag{2.64}$$

with $Z_N^T = \left(\hat{\psi}_1, \ldots, \hat{\psi}_N \right)$. Thus the solution obtained by taking the derivative and setting it to zero is simply an *instrumental variables* form in which the instrumental variables are given by the derivative of $\hat{\epsilon}_t$ with respect to $\hat{\theta}$. Furthermore, we see from (2.59) that the explicit formula for the negative gradient is exactly the pre-filtered data vector we used previously as a means for decorrelating the equation error in the case of optimally weighted least squares. However, in this instance the parameter vector θ has been augmented to include the unknown pre-filtering coefficients $\{c_1, \ldots, c_{n_c}\}$, and it may be somewhat surprising that the pre-filtering is also applied to the values of $\hat{\epsilon}_t$ stored in $\hat{\varphi}_t$.

From inspection of the Hessian (2.62), assuming $J(\hat{\theta}_N)$ to be sufficiently smooth, we see that if $\hat{C}(d)$ is stable and $\| \hat{\epsilon} \| \ll \| \hat{\varphi} \|$, then a local minimum to the loss function is obtained by using (2.64). This suggests using $\hat{\theta}^{(LS)}$ to obtain an initial estimate.

The solution (2.64) is only formal, however, since $\hat{C}(d)$ appears explicitly in the left-hand side, and is required to compute $\hat{\psi}_t$ in the right-hand side. Therefore, it is still necessary to employ an iterative strategy for obtaining $\hat{\theta}_N$. The most straightforward procedure is to iterate (2.64) such that each solution is used in calculating the next, i.e.,

$$\hat{\theta}_N^{(IV)}(i+1) = \left(\sum_{t=1}^{N} \hat{\psi}_t(i) \hat{\varphi}_t^T(i) \right)^{-1} \sum_{t=1}^{N} \hat{\psi}_t(i) y_t . \tag{2.65}$$

This is a relaxation method resulting from equating the gradient to zero. Each iteration coincides with a weighted least squares method in which the optimum weighting matrix has been estimated.

2.5.4. An Approximate Newton's Method

It is also possible to directly apply *Newton's Method* [161,163,174,143,179] (derived in Appendix E):

$$\begin{aligned} \hat{\theta}_N(i+1) &= \hat{\theta}_N(i) - \left(\frac{\partial^2 J}{\partial \theta^2}(\hat{\theta}_N(i)) \right)^{-1} \frac{\partial J}{\partial \theta}(\hat{\theta}_N(i)) \\ &= \hat{\theta}_N(i) - \left(\sum_{t=1}^{N} \hat{\epsilon}_t'(i) \hat{\epsilon}_t'(i)^T + \hat{\epsilon}_t'(i) \hat{\epsilon}_t''(i) \right)^{-1} \sum_{t=1}^{N} \hat{\epsilon}_t(i) \hat{\epsilon}_t'(i) , \end{aligned} \tag{2.66}$$

where $\hat{\epsilon}_t(i)$ is the model error at time t given parameters $\hat{\theta}_N(i)$, and the prime denotes differentiation with respect to the parameters $\hat{\theta}_N(i)$. This is a second order gradient descent method for nonlinear optimization, and it converges to the answer in one step of i for the case of quadratic $J(\hat{\theta}_N)$ (as happens when $C(d) = 1$). The parameter estimate $\hat{\theta}$ is incremented in the so-called *Newton direction* for each i. In the case of uncorrelated equation error, for which $C(d) = 1$, we have exact equality above since the Taylor expansion of $J(\theta)$ is truly second order in θ. In the present circumstance, the change in succesive estimates $\hat{\theta}_N(i)$ gives some measure of the higher order terms and hence a measure of the non-quadratic nature of the loss function $J(\hat{\theta}_N)$.

For this method, we need explicit calculation of the second derivative matrix which was given by (2.62). When $\hat{\theta}$ is close to θ, the last term in (2.62) may be neglected to give

$$\frac{\partial^2 J(\hat{\theta}_N)}{\partial \hat{\theta}^2} \approx 2 \sum_{t=1}^{N} \hat{\psi}_t \hat{\psi}_t^T \tag{2.67}$$

Using this approximation, the *Gauss-Newton Method* is obtained [174],

$$\hat{\theta}_N^{(ML)}(i+1) = \hat{\theta}_N^{(ML)}(i) - \left(\sum_{t=1}^{N} \hat{\psi}_t(i) \hat{\psi}_t^T(i) \right)^{-1} \sum_{t=1}^{N} \hat{\psi}_t(i) \hat{\epsilon}_t(i) \tag{2.68}$$

where $\hat{\psi}_t(i)$ and $\hat{\epsilon}_t(i)$ are computed using $\hat{\theta}(i)$. That is, on each batch iteration, the $\hat{C}(d)$ polynomial from the previous iteration is used to perform time-invariant pre-filtering throughout the current iteration. It turns out that this same algorithm can be obtained from the maximum likelihood point of view for both the case of known and estimated error covariance R_{EE_N} [95, p.93]. For this reason the superscript "ML" is used to denote the estimate, and we call it the *Maximum Likelihood Method* [82, 95]. The Gauss-Newton method also arises in the context of the *prediction error formulation* [82,85,86,87,108,95].

The approximate algorithm (2.68) is widely recommended (see e.g. [95]) over the more precise version (2.66) for two reasons. First, since the cost surface is not actually quadratic, the true Newton direction is only accurate *near the optimum* θ; the quality of approximation (2.67) is greatest near the optimum also. Second, it is important (especially initially in the search) that the matrix alteration of the gradient be positive definite in order that the adaption step always be "down hill," and the proposed approximation guarantees this since it forces symmetry in the matrix (of course there must be at least as many $\hat{\psi}_t$ vectors as there are parameters). Thus using the approximate form of the second derivative (2.67) guarantees a positive definite gradient transformation and provides the Newton direction in the final iterations to give "quadratic convergence" [161].

2.5.5. Further Approximations

When the noise e_t is not very correlated, then $C(d) \approx 1$ in which case a sensible approximation to the gradient $\hat{\psi}_t$ is the vector $-\hat{\varphi}_t$ (see (2.59)). This gives the so-called *Extended Matrix Method* or *Extended Least Squares* [99,95,81]:

$$\hat{\theta}_N^{(ELS)}(i+1) = \hat{\theta}_N^{(ELS)}(i) + \left(\sum_{t=1}^{N} \hat{\varphi}_t(i) \hat{\varphi}_t^T(i) \right)^{-1} \sum_{t=1}^{N} \hat{\varphi}_t(i) \hat{\epsilon}_t(i) \tag{2.69}$$

This is the same form as ordinary least squares but with $\hat{\theta}$ augmented to include $\hat{C}(d)$ and φ augmented to include $\hat{\epsilon}$. It is the same as $\hat{\theta}^{(ML)}$ except for the absense of pre-filtering by $1/C(d)$. By expanding $\hat{\epsilon}_t(i)$ as $y_t - \hat{\varphi}_t^T(i)\hat{\theta}_N(i)$, we have

$$\hat{\theta}_N^{(ELS)}(i+1) = \left(\sum_{t=1}^{N} \hat{\varphi}_t(i) \hat{\varphi}_t^T(i) \right)^{-1} \sum_{t=1}^{N} \hat{\varphi}_t(i) y_t \tag{2.70}$$

This form shows that $\hat{\theta}^{(IV)}$ (as well as $\hat{\theta}^{(ML)}$) reduces to $\hat{\theta}^{(ELS)}$ when the gradient approximation $\psi \approx -\varphi$ is employed.

2.5.6. Convergence of the Offline Identification Techniques

A principal advantage of the gradient formulation is in knowing that the estimate will always be improved by moving some amount in the direction of the negative gradient. However, since the higher order terms of the Taylor expansion of $J(\hat{\theta}_N)$ are nonzero, the optimal step-size to take in this direction is not clear. The step-size used in many nonlinear optimization schemes is that which reaches a local minimum of the loss function in the search direction. The solution $\hat{\theta}^{(IV)}$ above conforms to this policy by finding a local extremum of $J(\hat{\theta}_N)$, conditioned on $\hat{C}(d)$. The solution $\hat{\theta}^{(ML)}$ also takes this approach to step-size determination, but by finding a point where the "local quadratic approximation" to $J(\hat{\theta}_N)$ reaches a minimum, also conditioned on $\hat{C}(d)$. If the loss surface $J(\hat{\theta}_N)$ is reasonably smooth, then each iteration will result in an improvement of the parameter estimate. Finally, since $\hat{\theta}^{(ELS)}$ may be interpreted as a version of $\hat{\theta}^{(ML)}$ in which the gradient $\hat{\psi}_t$ is approximated by $-\hat{\varphi}_t$, its adaption policy may be seen as equivalent, using an inferior quadratic approximation relative to $\hat{\theta}^{(ML)}$.

In addition to the possibility of having an improper step-size, there is the danger that any of these schemes may halt prematurely due to a local minimum in $J(\hat{\theta}_N)$ which is not a global minimum. When $C(d) = 1$, this is not possible since the error is then linear in the parameters, and the solutions $\hat{\theta}^{(IV)}$, $\hat{\theta}^{(ML)}$, and $\hat{\theta}^{(ELS)}$ give the unique optimum parameter estimate in one step. For each fixed value of $\hat{C}(d)$ used in each iteration, the error is linear

in the parameters so that the minimum of $J(\hat{\theta}_N)$ is again unique and therefore global. Thus the primary question left undecided is whether or not $\hat{C}(d)$ itself will converge to the globally optimum polynomial $C(d)$, and this depends on the concavity of the cost surface $J(\hat{\theta}_N)$ with respect to c_i. There seem to be no general results concerning the existence of sub-optimal local minima in $J(\hat{\theta}_N)$ for the general ARMAX model (2.17). However, if there is no u_t input ($n_b = 0$), in which case we are fitting an ARMA model, then it is known that all local minima of $J(\hat{\theta}_N)$ are global minima when the model can represent the true system [119,84]; when n_a and n_c in the ARMA model are correct, then there is only one minimum of $J(\hat{\theta}_N)$, and for excessively high model orders, the extraneous local minima correspond to pole-zero cancellations in $C(d)/A(d)$. These remarks for the model-complete ARMA case apply only to *unconstrained* local minima. Sub-optimal local minima can still appear at the boundary of the stability domain.

Convergence may be aided in practice by the following measures. First, it is necessary that $\hat{C}(d)$ have a stable inverse at all times since otherwise $\hat{\epsilon}_t$ grows exponentially. We factor the $\hat{C}(d)$ polynomial after each iteration and contract the zeros to lie inside the unit circle if necessary (this may be done in a way that preserves angles in the complex plane by exponentially weighting the $\{c_i\}$ coefficients via $c_i \leftarrow \lambda^i c_i$, for some fixed $\lambda \in [0,1]$). Second, a good initial value for $\hat{C}(d)$ is $\hat{C}(d) = 1$ (i.e., $\hat{c}_i = 0, i = 1,\ldots,n_c$), and a good initial value for $\hat{\theta}_N(0)$ is that obtained by the more robust least squares method $\hat{\theta}^{(LS)}$ (which always gives the smallest value of $\hat{E}^T\hat{E}$ possible using $\hat{A}(d), \hat{B}(d)$ alone). Finally, Box and Jenkins [135] recommend that the loss function $J(\hat{\theta}_N)$ be plotted "extensively" as a function of the various components of $\hat{\theta}$. In this way, sub-optimal local minima and convergence against the unit circle may be detected.

2.5.7. Output Error Minimization

The output-error algorithms estimate the parameters of a rational linear time-invariant system given input-output data in the presence of uncorrelated measurement noise. Thus the measured output y_t is assumed to be of the form

$$y_t = \frac{B(d)}{A(d)}u_t + v_t, \qquad t = 1,2,\ldots \tag{2.71}$$

where u_t is the input signal, v_t is stationary white noise uncorrelated with u_s for all s, and the unknown linear system is representable by finite order polynomials in the unit-sample delay operator d:

$$\begin{aligned} A(d) &= 1 + a_1 d + \cdots + a_{n_a} d^{n_a} \\ B(d) &= b_1 d + \cdots + b_{n_b} d^{n_b}. \end{aligned} \tag{2.72}$$

We assume that the system is stable, i.e., that all the roots of $A(z^{-1})$ lie inside the unit circle in the complex plane. It is also assumed that system observations y_t and u_t are

available for an arbitrary length of time. Accordingly, the analysis will be given for the case of infinitely long data records.

The model is written as

$$y_t = \frac{\hat{B}(d)}{\hat{A}(d)} u_t + \hat{v}_t, \tag{2.73}$$

where

$$\hat{A}(d) = 1 + \hat{a}_1 d + \cdots + \hat{a}_{\hat{n}_a} d^{\hat{n}_a}$$
$$\hat{B}(d) = \hat{b}_1 d + \cdots + \hat{b}_{\hat{n}_b} d^{\hat{n}_b},$$

and we define

$$n^* = \min\{\hat{n}_a - n_a, \hat{n}_b - n_b\}.$$

In the case of white measurement noise v_t, we wish to minimize

$$J_{oe}(\hat{\theta}) = \bar{\mathcal{E}}_t\{\hat{v}_t^2\} \triangleq \lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} \hat{v}_t^2 \tag{2.74}$$

with respect to the vector of parameters $\hat{\theta}^T = (\hat{a}_1, \ldots, \hat{a}_{\hat{n}_a}, \hat{b}_1, \ldots, \hat{b}_{\hat{n}_b})$. Such a procedure is known as an *output error* identification technique. We must restrict $\hat{A}(z^{-1})$ to have roots inside the unit circle in order for the limit (2.74) to exist.

Using the assumption that v_t is uncorrelated with u_s for all s, we have

$$J_{oe}(\hat{\theta}) = \bar{\mathcal{E}}_t\left\{\left[\left(\frac{B(d)}{A(d)} - \frac{\hat{B}(d)}{\hat{A}(d)}\right)u_t\right]^2\right\} + \sigma_v^2.$$

where

$$\sigma_v^2 \triangleq \bar{\mathcal{E}}_t\{v_t^2\}$$

is the output noise variance which is independent of $\hat{\theta}$.

In the frequency domain, this is equivalent to minimizing

$$J_{oe}(\hat{\theta}) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \left|\frac{B(e^{j\omega})}{A(e^{j\omega})} - \frac{\hat{B}(e^{j\omega})}{\hat{A}(e^{j\omega})}\right|^2 U(e^{j\omega})d\omega + \sigma_v^2.$$

where

$$U(e^{j\omega}) \triangleq \sum_{k=-\infty}^{\infty} \bar{\mathcal{E}}_t\{u_t u_{t+k}\}e^{-jwk}$$

is the power spectral density of u_t.

If $n^* = 0$, and if u_t is persistently exciting of order $n_a + n_b$, then the unique minimum of $J_{oe}(\hat{\theta})$ is given by the true system parameters [126].

If $n^* > 0$ and u_t is persistently exciting of order $\max\{\hat{n}_a + n_b, n_a + \hat{n}_b\}$, then the unique minimum is attained by a continuum of parameter vectors $\hat{\theta}$ such that

$$\hat{A}(d) = A(d)L(d)$$
$$\hat{B}(d) = B(d)L(d),$$

and the polynomial

$$L(d) = 1 + l_1 d + \cdots + l_{n^*} d^{n^*}$$

is of order n^* with arbitrary coefficients l_i [126].

In the case $n^* < 0$, minimization of $J_{oe}(\hat{\theta})$ yields an optimal weighted least squares fit to the system frequency response, where the power spectral density of u_t appears as the weighting function in the frequency domain. If u_t is spectrally flat, such as an impulse or white noise, then the mean squared error between the system and model impulse responses is minimized.

A difficulty with output error methods is that they present a nonlinear optimization problem. This is due to the fact that the loss function $J_{oe}(\hat{\theta})$ is not a quadratic form in the parameters $\hat{a}_i$, or equivalently, because v_t is not linear in the parameters a_i. Currently, the conditions for convergence are unknown. Moreover, it is shown in Appendix A that when $\hat{n}_a < n_a$, there can exist multiple local minima in J_{oe} which can make gradient-based descent algorithms of limited utility.

The Steiglitz-McBride Algorithm

The Steiglitz-McBride algorithm is an iterative application of an *equation-error* minimization where u_t and y_t are filtered between iterations by an estimate of $1/\hat{A}(d)$. This converts equation error $\hat{A}_k(d)y_t - \hat{B}_k(d)u_t$ into the error $\hat{A}_k(d)y_t/\hat{A}_{k-1}(d) - \hat{B}_k(d)u_t/\hat{A}_{k-1}(d)$ which reduces to $y_t - \hat{B}_k(d)u_t/\hat{A}_k(d)$ upon convergence of $\hat{A}$.

The Steiglitz-McBride algorithm is given by the following iteration for $k = 0, 1, 2, \ldots$.

$$\hat{\theta}_{k+1} = \overline{\varepsilon}_t\left\{\varphi_t^f(\hat{\theta}_k)\varphi_t^f(\hat{\theta}_k)^T\right\}^{-1}\overline{\varepsilon}_t\left\{\varphi_t^f(\hat{\theta}_k)y_t^f(\hat{\theta}_k)\right\}, \tag{2.75}$$

where

$$\hat{\theta}_k^T = \left(\hat{a}_1(k), \ldots, \hat{a}_{\hat{n}_a}(k), \hat{b}_1(k), \ldots, \hat{b}_{\hat{n}_b}(k) \right)$$

$$\varphi_t^f(\hat{\theta}_k)^T = \left(-y_{t-1}^f(\hat{\theta}_k), \ldots, -y_{t-n_a}^f(\hat{\theta}_k), u_{t-1}^f(\hat{\theta}_k), \ldots, u_{t-n_b}^f(\hat{\theta}_k) \right)$$

$$y_t^f(\hat{\theta}_k) = \frac{1}{\hat{A}_k(d)} y_t$$

$$u_t^f(\hat{\theta}_k) = \frac{1}{\hat{A}_k(d)} u_t$$

$$\hat{A}_k(d) = 1 + \hat{a}_1(k)d + \cdots + \hat{a}_{\hat{n}_a}(k)d^{\hat{n}_a}$$

$$\hat{B}_k(d) = \hat{b}_1(k)d + \cdots + \hat{b}_{\hat{n}_b}(k)d^{\hat{n}_b}.$$

Note that $\hat{\theta}_{k+1}$ minimizes

$$J_{ef}(\hat{\theta}_{k+1}) = \overline{\varepsilon}_t \left\{ \hat{\varepsilon}_t^f(\hat{\theta}_k)^2 \right\}$$

where

$$\begin{aligned} \hat{\varepsilon}_t^f(\hat{\theta}_k) &= y_t^f(\hat{\theta}_k) - \varphi_t^f(\hat{\theta}_k)^T \hat{\theta}_{k+1} \\ &= \hat{A}_{k+1}(d) y_t^f(\hat{\theta}_k) - \hat{B}_k(d) u_t^f(\hat{\theta}_k) \\ &= \frac{\hat{A}_{k+1}(d)}{\hat{A}_k(d)} \hat{v}_t. \end{aligned}$$

Thus, if $\hat{\theta}_k$ converges, then the output error criterion is minimized. Unfortunately, convergence even to the nearest local minimum has not been shown [128]. It is straightforward to extend the Steiglitz-McBride algorithm to time-recursive form. It was observed empirically that the $\hat{A}$ estimate used in the pre-filtering should be delayed by at least $\hat{n}_a$ samples relative to the pre-filter input signals.

In practice, the algorithm does apparently get stuck at sub-optimal local minima more often than not. However, when it does work, it tends to give much superior modeling of the frequency response. Figure 2.4a shows the result of applying Weighted Least Squares to white noise filtered by $H(z) = 1 + 0.7z^{-14}$. The weight function (an order 6 Butterworth lowpass with the -3dB point at $f = f_s/8$) causes the algorithm to focus on the first two "resonances" in the frequency response. Fig. 2.4b shows the result after 3 iterations of the Steiglitz-McBride algorithm. It is clear from this figure that output error can yield a better match in the band of interest. This example was fortunate, however, for in other similar cases, the fit changed only slightly or even became worse. This emphasizes the practical importance of guaranteed convergence to an optimum solution. Indeed, much of the work of Chapter 1 was motivated by this example and others like it.

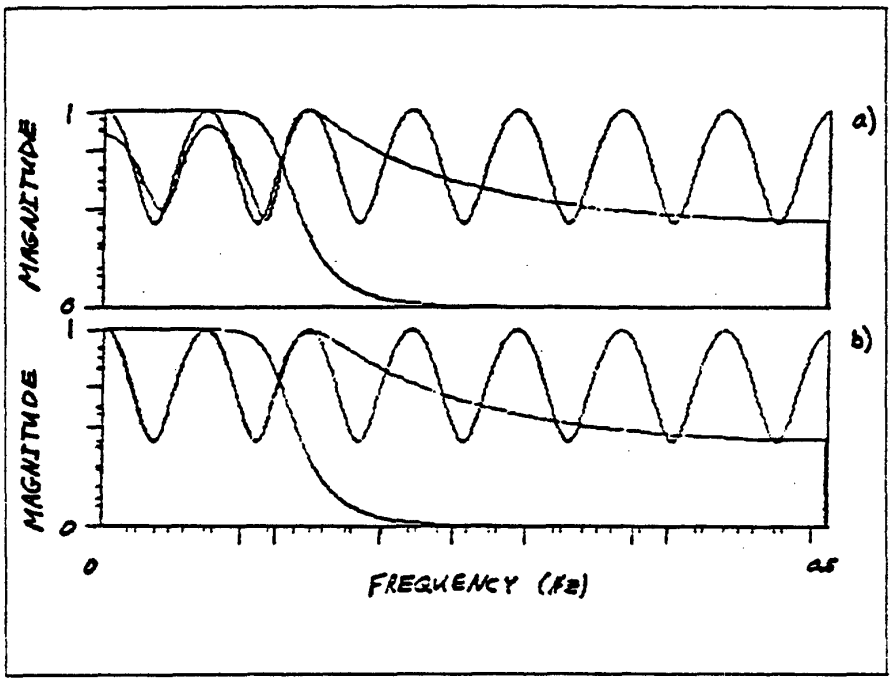

Figure 2.4. Comparison of equation-error minimization and output-error minimization using the Steiglitz-McBride algorithm. Each plot gives an overlay of the true frequency response magnitude (the regular series of arches), the weight function (a gentle lowpass characteristic), and the 5-pole, 4-zero model frequency response magnitude.

a) Equation error.

b) Output error.

2.5.8. Summary of Offline Identification Algorithms

We have derived the off-line versions of Least Squares ($\hat{\theta}^{(LS)}$), Optimally Weighted Least Squares, Generalized Least Squares, Extended Least Squares ($\hat{\theta}^{(ELS)}$), Instrumental Variables ($\hat{\theta}^{(IV)}$), Maximum Likelihood ($\hat{\theta}^{(ML)}$), and the Steiglitz-McBride algorithm. All methods (other than Steiglitz-McBride) differ only in the manner in which the correlation of the noise e_t is handled. The Steiglitz-McBride algorithm provides a means for transforming an equation-error method into an output-error method. The two general frameworks discussed were the regression formulation and gradient descent.

In the following sections, we shall be primarily concerned with the methods $\hat{\theta}^{(LS)}$, $\hat{\theta}^{(IV)}$, $\hat{\theta}^{(ELS)}$, and $\hat{\theta}^{(ML)}$, which can all be expressed as

$$\begin{aligned}\hat{\theta}_N(i+1) &= \hat{\theta}_N(i) - \left(\sum_{t=1}^{N} \hat{\epsilon}_t'(i)\hat{\epsilon}_t'(i)^T\right)^{-1} \sum_{t=1}^{N} \hat{\epsilon}_t(i)\hat{\epsilon}_t'(i), \\ &= \hat{\theta}_N(i) + \left(\sum_{t=1}^{N} z_t(i)\xi_t^T(i)\right)^{-1} \sum_{t=1}^{N} z_t(i)\hat{\epsilon}_t(i),\end{aligned} \tag{2.76}$$

where

$$\begin{aligned}\hat{\epsilon}_t(i) &= y_t - \hat{\varphi}_t^T(i)\hat{\theta}_N(i) \\ \hat{\varphi}_t^T(i) &= \left(-y_{t-1}^T, \ldots, -y_{t-n_a}^T, u_{t-1}^T, \ldots, u_{t-n_b}^T, \hat{\epsilon}_{t-1}^T(i), \ldots, \hat{\epsilon}_{t-n_c}^T(i)\right) \\ \hat{\theta}_N^T(i) &= \left(\hat{a}_1(i), \ldots, \hat{a}_{n_a}(i), \hat{b}_1(i), \ldots, \hat{b}_{n_b}(i), \hat{c}_1(i), \ldots, \hat{c}_{n_c}(i)\right),\end{aligned} \tag{2.77}$$

and the individual methods are obtained from the following table:

Method	$z_t(i)$	$\xi_t(i)$
$\hat{\theta}^{(LS)}$	φ_t	φ_t
$\hat{\theta}^{(ELS)}$	$\hat{\varphi}_t(i)$	$\hat{\varphi}_t(i)$
$\hat{\theta}^{(IV)}$	$\hat{\varphi}_t^f(i)$	$\hat{\varphi}_t(i)$
$\hat{\theta}^{(ML)}$	$\hat{\varphi}_t^f(i)$	$\hat{\varphi}_t^f(i)$
General	$-\hat{\epsilon}_t'(i)$	$-\hat{\epsilon}_t'(i)$

Table 2.1

where

$$\hat{\varphi}_t^f(i) = \hat{\varphi}_t(i) - \hat{c}_1(i)\hat{\varphi}_{t-1}^f(i) - \cdots - \hat{c}_{n_c}(i)\hat{\varphi}_{t-n_c}^f(i) \tag{2.78}$$

is (precisely) the negative gradient of $\hat{\epsilon}_t(i)$ with respect to $\hat{\theta}_N(i)$ when the model is of the form (2.76). For more general models, the gradient of the error at time t with respect to the parameters $\hat{\theta}_N(i)$ is denoted $\hat{\epsilon}_t'(i)$. These methods may all be viewed as forms of the Gauss-Newton method for nonlinear optimization with different gradient approximations.

2.6. Recursive Computation of Offline Methods

The algorithms derived above can made time-recursive, where the parameter estimate $\hat{\theta}$ is updated for each t. In this section we will consider only one pass through the N data points. Therefore, we will drop the pass-number i from the equations for notational simplicity. The initial value $\hat{\theta}_N(i)$, from the previous pass, will be denoted $\hat{\theta}_0$, and the final estimate $\hat{\theta}_N(i+1)$, obtained at the end of pass i, is written as $\hat{\theta}_N$. Then the four off-line

algorithms described in the preceding section may be written as

$$\hat{\theta}_N = \hat{\theta}_0 + \left(\sum_{t=1}^{N} z_t \xi_t^T\right)^{-1} \sum_{t=1}^{N} z_t \hat{\epsilon}_t, \tag{2.79}$$

which is simply (2.76) with the pass number omitted. Defining

$$R_N \triangleq \sum_{t=1}^{N} z_t \xi_t^T, \qquad G_N \triangleq \sum_{t=1}^{N} z_t \hat{\epsilon}_t, \tag{2.80}$$

we have the time recursions

$$\begin{aligned} R_t &= R_{t-1} + z_t \xi_t^T \\ G_t &= G_{t-1} + z_t \hat{\epsilon}_t, \end{aligned} \tag{2.81}$$

and the relation

$$\hat{\theta}_t = \hat{\theta}_0 + R_t^{-1} G_t \qquad t = 1, \ldots, N. \tag{2.82}$$

We derive a recursive in time update for $\hat{\theta}_t$ as follows:

$$\begin{aligned} \hat{\theta}_t &= \hat{\theta}_0 + R_t^{-1} G_t \\ &= \hat{\theta}_0 + R_t^{-1}(G_{t-1} + z_t \hat{\epsilon}_t) \\ &= \hat{\theta}_0 + R_t^{-1}\Big(R_{t-1}\Big(\hat{\theta}_{t-1} - \hat{\theta}_0\Big) + z_t \hat{\epsilon}_t\Big) \\ &= \hat{\theta}_0 + R_t^{-1} R_{t-1}\Big(\hat{\theta}_{t-1} - \hat{\theta}_0\Big) + R_t^{-1} z_t \hat{\epsilon}_t \\ &= \hat{\theta}_0 + R_t^{-1}\Big(R_t - z_t \xi_t^T\Big)\Big(\hat{\theta}_{t-1} - \hat{\theta}_0\Big) + R_t^{-1} z_t \hat{\epsilon}_t \\ &= \hat{\theta}_{t-1} - R_t^{-1} z_t \xi_t^T\Big(\hat{\theta}_{t-1} - \hat{\theta}_0\Big) + R_t^{-1} z_t \hat{\epsilon}_t \\ &= \hat{\theta}_{t-1} + R_t^{-1} z_t\Big(\hat{\epsilon}_t - \xi_t^T\Big(\hat{\theta}_{t-1} - \hat{\theta}_0\Big)\Big). \end{aligned} \tag{2.83}$$

2.6.1. Recursive LS, ELS, and IV

In $\hat{\theta}^{(IV)}$ and $\hat{\theta}^{(ELS)}$, we have $\xi_t = \hat{\varphi}_t$ which implies

$$\begin{aligned} \hat{\epsilon}_t - \xi_t^T(\hat{\theta}_{t-1} - \hat{\theta}_0) &= (y_t - \hat{\varphi}_t^T \hat{\theta}_0) - \hat{\varphi}_t^T \hat{\theta}_{t-1} + \hat{\varphi}_t^T \hat{\theta}_0 \\ &= y_t - \hat{\varphi}_t^T \hat{\theta}_{t-1}. \end{aligned} \tag{2.84}$$

Thus the exact recursive forms of the offline methods $\hat{\theta}^{(IV)}$ and $\hat{\theta}^{(ELS)}$ (and also $\hat{\theta}^{(LS)}$ when $\hat{\varphi}$ is replaced by φ) are given by

$$\begin{aligned} \hat{\theta}_t &= \hat{\theta}_{t-1} + R_t^{-1} z_t\Big(y_t - \hat{\varphi}_t^T \hat{\theta}_{t-1}\Big) \\ &= \hat{\theta}_{t-1} + R_t^{-1} z_t \hat{\epsilon}_t(t-1) \end{aligned} \tag{2.85}$$

and Table 2.1.

It is interesting to note that the "driving residual" of the off-line version $\hat{\epsilon}_t \triangleq y_t - \hat{\varphi}_t^T \hat{\theta}_0$ has become

$$\hat{\epsilon}_t(t-1) = y_t - \hat{\varphi}_t^T \hat{\theta}_{t-1} \tag{2.86}$$

in the recursively updated version Otherwise the structure is very similar to the batch version (2.76).

2.6.2. Interpretation of ELS as a Limited-Step Newton's Method

The recursive algorithm (2.85) may be interpreted as a *limited-step Newton's method* [161] for each t, since

$$\begin{aligned} \frac{\partial \hat{\epsilon}_t^2(t-1)}{\partial \hat{\theta}_{t-1}} &= 2\hat{\epsilon}_t(t-1)\frac{\partial \hat{\epsilon}_t(t-1)}{\partial \hat{\theta}_{t-1}} \\ &= -2\hat{\epsilon}_t(t-1)\hat{\varphi}_t \end{aligned} \tag{2.87}$$

using the fact that $\hat{\varphi}_t$ is not a function of $\hat{\theta}_{t-1}$. Thus, in the recursion (2.85), an exact "instantaneous" gradient of the prediction error $\hat{\epsilon}_t(t-1)$ with respect to the latest parameter estimate $\hat{\theta}_{t-1}$ is computed, given of course the fixed pre-filter $\hat{C}_0(d)$ which applies to the whole batch for $t = 1, \ldots, N$.

Since the Hessian,

$$\frac{\partial^2 \hat{\epsilon}_t^2(t-1)}{\partial \hat{\theta}_{t-1}^2} = 2\hat{\varphi}_t \hat{\varphi}_t^T \tag{2.88}$$

is not invertible, a true instantaneous quadratic descent is not possible. However, in $\hat{\theta}^{(ELS)}$, the average Hessian is used, for the update can be written as

$$\hat{\theta}_t^{(ELS)} = \hat{\theta}_{t-1}^{(ELS)} - \frac{1}{t}\overline{\epsilon}_t\left\{\frac{\partial^2 \hat{\epsilon}_t^2(t-1)}{\partial \hat{\theta}_{t-1}^2}\right\}^{-1}\frac{\partial \hat{\epsilon}_t^2(t-1)}{\partial \hat{\theta}_{t-1}} \tag{2.89}$$

where

$$\overline{\epsilon}_t\{x_k\} \triangleq \frac{1}{t}\sum_{k=1}^{t} x_k \tag{2.90}$$

denotes time averaging.

Thus $\hat{\theta}^{(ELS)}$ is a form of limited-step Newton's method [161]. The search direction at each time instant is given by the instantaneous direction of steepest descent multiplied by the estimate of the inverse Hessian based on all data up to time t. The step-size factor $1/t$ weights each individual step by the inverse of the total number of steps so far; this essentially averages the individual Newton directions, and from (2.69) we see that it is exactly equivalent to first averaging the instantaneous gradients and then multiplying by the inverse of the final average Hessian to obtain a single (batch) Newton step.

2.6.3. Recursive Maximum Likelihood

The one case of (2.83) where we did not get a cancellation of the term involving $\hat{\theta}_0$ was with $\hat{\theta}^{(ML)}$. In this case, substituting according to Table 2.1 into (2.83), we have,

$$\begin{aligned}\hat{\theta}_t^{(ML)} &= \hat{\theta}_{t-1}^{(ML)} + R_t^{-1}\hat{\varphi}_t^f\Big(\hat{\epsilon}_t - \hat{\varphi}_t^{fT}\Big(\hat{\theta}_{t-1}^{(ML)} - \hat{\theta}_0^{(ML)}\Big)\Big) \\ &= \hat{\theta}_{t-1}^{(ML)} + R_t^{-1}\hat{\varphi}_t^f\Big(y_t - \hat{\varphi}_t^{fT}\hat{\theta}_{t-1}^{(ML)} + \hat{\varphi}_t^{fT}\Big(\hat{\theta}_0^{(ML)} - \hat{\theta}_{t-1}^{(ML)}\Big)\Big).\end{aligned} \tag{2.91}$$

Thus $\hat{\theta}_{t-1}^{(ML)}$ is asymptotically equivalent in form to the simpler cases since eventually $\hat{\theta}_0 - \hat{\theta}_{t-1} \to 0$ as the number of batch iterations tends to infinity. Also, when the equation error e_t is not strongly correlated, so that $\hat{\varphi}_t^f \approx \hat{\varphi}_t$, then $\hat{\theta}^{(ML)}$ again reduces to $\hat{\theta}^{(ELS)}$ (as does $\hat{\theta}^{(IV)}$).

We may note one further interpretation of the $\hat{\theta}^{(ML)}$ recursion as follows. By linearity, pre-filtering may be factored outside the expression $y_t^f - \hat{\varphi}_t^{fT}\hat{\theta}_0$ to obtain $\hat{C}_0^{-1}(d)(y_t - \hat{\varphi}_t^T\hat{\theta}_0) = \hat{C}_0^{-1}(d)\hat{\epsilon}_t = \hat{\epsilon}_t^f$. Similarly, we find that $y_t^f - \hat{\varphi}_t^{fT}\hat{\theta}_{t-1} = \hat{C}_0^{-1}(d)\hat{\epsilon}_t(t-1) = \hat{\epsilon}_t^f(t-1)$. Therefore, introducing $y_t^f - y_t^f$ into (2.91) yields

$$\hat{\theta}_t^{(ML)} = \hat{\theta}_{t-1}^{(ML)} + R_t^{-1}\hat{\varphi}_t^f\Big(\hat{\epsilon}_t - \hat{\epsilon}_t^f + \hat{\epsilon}_t^f(t-1)\Big). \tag{2.92}$$

Thus $\hat{\theta}^{(ML)}$ is effectively driven by three residuals which are

$$\begin{aligned}\hat{\epsilon}_t &= y_t - \hat{\varphi}_t^T\hat{\theta}_0^{(ML)} \\ -\hat{\epsilon}_t^f &= \frac{-1}{\hat{C}_0(d)}\Big(y_t - \hat{\varphi}_t^T\hat{\theta}_0^{(ML)}\Big) \\ \hat{\epsilon}_t^f(t-1) &= \frac{1}{\hat{C}_0(d)}\Big(y_t - \hat{\varphi}_t^T\hat{\theta}_{t-1}^{(ML)}\Big)\end{aligned} \tag{2.93}$$

This concludes the derivation of the recursive forms of the off-line identification algorithms. No approximations have been made, and these recursive forms are exactly equivalent to their off-line counterparts. Thus the off-line algorithms (2.76) may be implemented recursively by means of (2.83) making repeated passes through the data until convergence is achieved.

2.6.4. Eliminating the Initial Estimate from the Recursions

The offline methods discussed so far step $\hat{\theta}_N$ in a direction which reduces residual energy $J(\hat{\theta}_N) = \hat{E}_N^T\hat{E}_N$ as computed over the entire data record. As a result, a fixed

value of $\hat{\theta}$ is used over all the data on each pass. For large N, this can be inefficient since fewer samples may give almost as good a step in $\hat{\theta}$. In other words, there is a trade-off between accuracy in the estimation of the descent step versus taking steps often enough to give descent along the loss surface at a reasonable pace. In this section, algorithms are derived which *simultaneously* update the search direction and the parameter estimates on which the search direction is conditioned. These are the algorithms which are known as *recursive identification algorithms.*

Consider that in the recursive versions of the off-line methods, the pre-filtering by $\hat{C}_0^{-1}(d)$ is fixed throughout each pass. However, within a pass, new and presumably better estimates of $\hat{C}(d)$ are being explicitly computed. It seems reasonable to expect that incorporating these latest estimates into the recursive algorithm will improve its rate of convergence. Also, if the system is changing over time, it is better to use a "local" estimate of $C(d)$ in order to get a good gradient computation.

There are a variety of ways to infuse the latest information about θ into the recursive algorithms. One approach is to extend the interpretation of $\hat{\theta}^{(ELS)}$ as a instantaneous quadratic gradient descent method by removing the fixation of $\hat{C}(d)$ to $\hat{C}_0(d)$ and re-computing the gradient. Another is to apply the Robbins-Monro algorithm from the theory of stochastic approximation [124]. However, we give instead a derivation based on the recursively upated offline maximum likelihood method $\hat{\theta}^{(ML)}$.

Equation (2.92) gives the off-line version of $\hat{\theta}^{(ML)}$ which was obtained by applying Newton's method to the minimization of $J(\hat{\theta}_N)$. We wish to *use the latest parameter estimates wherever possible in the algorithm.* Accordingly, we replace $\hat{\theta}_0$ by $\hat{\theta}_{t-1}$ and $\hat{C}_0(d)$ by $\hat{C}_{t-1}(d)$ in the offline recursion (2.92). With these substitutions, $\hat{\epsilon}_t$ becomes $\hat{\epsilon}_t(t-1)$, and $\hat{\epsilon}_t^f$ becomes $\hat{\epsilon}_t^f(t-1)$. Therefore, we get a cancellation of the two filtered residual terms, and $\hat{\theta}^{(ML)}$ becomes

$$\begin{aligned} \hat{\theta}_t^{(RML)} &= \hat{\theta}_{t-1}^{(RML)} - R_t^{-1}\hat{\psi}_t(t-1)\hat{\epsilon}_t(t-1) \\ R_t &= R_{t-1} + \hat{\psi}_t(t-1)\hat{\psi}_t^T(t-1) \end{aligned} \tag{2.94}$$

where

$$\begin{aligned} \hat{\epsilon}_{t-i}(t-1) &= y_{t-i} - \hat{\varphi}_{t-i}^T(t-1)\hat{\theta}_{t-1} \\ \hat{\varphi}_t^T(t-1) &= (-y_{t-1},\ldots,-y_{t-n_a},u_{t-1},\ldots,u_{t-n_b},\hat{\epsilon}_{t-1}(t-1),\ldots,\hat{\epsilon}_{t-n_c}(t-1)). \end{aligned} \tag{2.95}$$

Thus we proceed as if we were recursively computing the offline maximum likelihood estimates, refining the initial conditions to be the latest estimates as we go; i.e., $\hat{\theta}_0$ is replaced by $\hat{\theta}_{t-1}$ wherever it appears. We call this the *Recursive Maximum Likelihood Algorithm*, and denote the parameter estimate by $\hat{\theta}^{(RML)}$.

The gradient $\hat{\psi}_t(t-1)$ is by definition

$$\begin{aligned}\hat{\psi}_t(t-1) &= \frac{\partial \hat{\epsilon}_t(t-1)}{\partial \hat{\theta}_{t-1}} \\ &= -\hat{\varphi}_t(t-1) - \Big(0,\ldots,0,0,\ldots,0,\hat{\psi}_{t-1}(t-1),\ldots,\hat{\psi}_{t-n_c}(t-1)\Big)\hat{\theta}_{t-1} \\ &= -\hat{\varphi}_t(t-1) - \hat{c}_1(t-1)\hat{\psi}_{t-1}(t-1) - \cdots - \hat{c}_{n_c}(t-1)\hat{\psi}_{t-n_c}(t-1) \\ &= -\frac{1}{\hat{C}_{t-1}(d)}\hat{\varphi}_t(t-1).\end{aligned} \tag{2.96}$$

Due to the shift structure of $\hat{\varphi}_t(t-1)$, the above filtering may be implemented with smaller complexity by the following computation.

$$\begin{aligned}y_t^f(\tau) &= \frac{1}{\hat{C}_{t-1}(d)}y_t = y_t - \hat{c}_1(\tau)y_{t-1}^f(\tau) - \cdots - \hat{c}_{n_c}(\tau)y_{t-n_c}^f(\tau) \\ u_t^f(\tau) &= \frac{1}{\hat{C}_{t-1}(d)}u_t = u_t - \hat{c}_1(\tau)u_{t-1}^f(\tau) - \cdots - \hat{c}_{n_c}(\tau)u_{t-n_c}^f(\tau) \\ \hat{\epsilon}_t^f(\tau) &= \frac{1}{\hat{C}_{t-1}(d)}\hat{\epsilon}_t(\tau) = \hat{\epsilon}_t(\tau) - \hat{c}_1(\tau)\hat{\epsilon}_{t-1}^f(\tau) - \cdots - \hat{c}_{n_c}(\tau)\hat{\epsilon}_{t-n_c}^f(\tau) \\ \hat{\psi}_t^T(\tau) &\triangleq \Big(y_{t-1}^f(\tau),\ldots,y_{t-n_a}^f(\tau),-u_{t-1}^f(\tau),\ldots,-u_{t-n_b}^f(\tau),-\hat{\epsilon}_{t-1}^f(\tau),\ldots,-\hat{\epsilon}_{t-n_c}^f(\tau)\Big),\end{aligned} \tag{2.97}$$

where in the algorithm $\tau = t-1$.

Note that the above algorithm is not truly recursive in that the gradient requires computation of $\hat{\varphi}_k(t-1)$, for $k = 1,\ldots,t$, and the computation of $\hat{\varphi}_t(t-1)$ alone requires recomputing all the residuals from time $t = 1$ using the latest model estimate $\hat{\theta}_{t-1}$. However, if the zeros of $\hat{C}(d)$ are well outside the unit circle, then the gradient recursion rapidly forgets past $\hat{\varphi}$ vectors. Also, since the change in $\hat{\theta}_t$ over n_c successive samples should be small for $t \gg n_c$, it is reasonable to expect that $\hat{\epsilon}_t(t-i) \approx \hat{\epsilon}_t(t-j)$ for small i and j as long as the computation of $\hat{\epsilon}_t$ is strictly stable (*i.e.* the recursive filtering used in computing $\hat{\epsilon}_t$ from u_t and y_t must have approximately finite memory, and preferably fairly short memory). Thus we consider using

$$\begin{aligned}y_t^f &= y_t - \hat{c}_1(t-1)y_{t-1}^f - \cdots - \hat{c}_{n_c}(t-1)y_{t-n_c}^f \\ u_t^f &= u_t - \hat{c}_1(t-1)u_{t-1}^f - \cdots - \hat{c}_{n_c}(t-1)u_{t-n_c}^f \\ \hat{\epsilon}_t^f &= \hat{\epsilon}_t(t-1) - \hat{c}_1(t-1)\hat{\epsilon}_{t-1}^f - \cdots - \hat{c}_{n_c}(t-1)\hat{\epsilon}_{t-n_c}^f \\ \hat{\psi}_t^T &= \Big(y_{t-1}^f,\ldots,y_{t-n_a}^f,-u_{t-1}^f,\ldots,-u_{t-n_b}^f,-\hat{\epsilon}_{t-1}^f,\ldots,-\hat{\epsilon}_{t-n_c}^f\Big).\end{aligned} \tag{2.98}$$

With these approximations, we obtain an algorithm that is truly recursive in time. The residuals in $\hat{\varphi}$ are now just the prediction errors computed on the past n_c updates, and the recursive filtering for the gradient estimate also just uses previously computed quantities. In the case of the $\hat{\varphi}$ vector, this time-shift structure is necessary for using the complexity $O(N_p)$ update algorithms to be discussed shortly.

To summarize, we have derived a Gauss-Newton algorithm which is truly recursive in time and which allows simultaneous updating of the parameter estimates, the gradient search direction, and (unlike the offline versions) the parameter estimates used in computing both the gradient and equation error. The search direction is akin to the Newton direction using a Hessian estimate based on all the data up to time t, and an "instantaneous" gradient of $\hat{\epsilon}_t^2(t-1)$ with respect to $\hat{\theta}_{t-1}$.

Miscellaneous Relations and Terminology

In the case $n_c = 0$, $\hat{\theta}^{(RML)}$ reduces to exact recursive least squares $\hat{\theta}^{(LS)}$. Also, when the $\hat{C}(d)$ polynomial is fixed to $\hat{C}_0(d)$, and when $\hat{\varphi}_t$ is made to contain $\{\hat{\epsilon}_{t-i}(0), i = 1, \ldots, n_c\}$, we obtain the exact recursive update for the off-line extended least squares method $\hat{\theta}^{(ELS)}$. If $\hat{\theta}_t$ is constant, the approximations used in obtaining $\hat{\theta}^{(RML)}$ from $\hat{\theta}^{(ML)}$ become exact. Thus if the estimate $\hat{\theta}_t^{(RML)}$ converges, it becomes equivalent to the recursive offline $\hat{\theta}^{(ML)}$ except for state information accumulated prior to convergence. This information consists of an additive difference in the matrix R_t and memory of old residuals and gradient vectors via the recursive filtering (2.98). If the average signal powers of u_t and y_t remain nonzero, then the R_t discrepancy becomes arbitrarily small. If the final $\hat{C}^{-1}(d)$ is strictly stable, then the influence of residual and gradient estimates prior to convergence will also die away. In other terms, when convergence at a finite time is assumed, the input is persistently exciting, and the final $C(d)$ is strictly minimum-phase, then the truly recursive maximum likelihood method is equivalent to the offline version.

The present form of the algorithm for $\hat{\theta}^{(RML)}$ is called *Recursive Maximum Likelihood* (RML) [95] or RML2 [125]. When the gradient $\hat{\psi}_t$ is approximated as $-\hat{\varphi}_t$, then we get what is called *RML1* [125], *Approximate Maximum Likelihood* (AML) [119,127], or *Recursive Extended Least Squares* (RELS). When the true gradient is used only as one of the two vectors in the R_t update, then a form of *Recursive Instrumental Variables* (RIV) is obtained. When $n_c = 0$ the result is *Recursive Least Squares* (RLS) which is equivalent to the off-line version, since all approximations above pertain only to the computations needing $\hat{C}(d)$. By considering the more general case in which $H_\epsilon(d) = 1/F(d)$, it is possible also to incorporate *Recursive Generalized Least Squares* (RGLS). In this case, the gradient of $\hat{\epsilon}_t$ is more complicated, but still obtained by linearly filtering components of $\hat{\varphi}_t$. For discussion

of the more general case

$$A(d)y_t = \frac{B(d)}{G(d)}u_t + \frac{C(d)}{F(d)}\epsilon_t, \tag{2.99}$$

as well as general prediction-error algorithms, see [108,111].

2.6.5. A Generalized Recursive Gauss-Newton Method

Using the same steps as in (2.83), we can obtain an exact recursion which does not depend on the form of ϵ_t, and only the sum-of-squares structure is used. The recursively updated Gauss-Newton method becomes

$$\begin{aligned}\hat{\theta}_t &= \hat{\theta}_{t-1} - R_t^{-1}\hat{\epsilon}_t'(0)\left[\hat{\epsilon}_t(0) + \hat{\epsilon}_t'(0)^T\left(\hat{\theta}_{t-1} - \hat{\theta}_0\right)\right] \\ R_t &= R_{t-1} + \hat{\epsilon}_t'(0)\hat{\epsilon}_t'(0)^T,\end{aligned} \tag{2.100}$$

where $\hat{\epsilon}_t(0)$ is the model error at time t given parameters $\hat{\theta}_0$, and the prime denotes differentiation with respect to the parameters $\hat{\theta}_0$.

As in the previous section, we eliminate the initial estimate $\hat{\theta}_0$ by using the latest parameter estimate $\hat{\theta}_{t-1}$ in its place. This gives

$$\begin{aligned}\hat{\theta}_t &= \hat{\theta}_{t-1} - R_t^{-1}\hat{\epsilon}_t'(t-1)\hat{\epsilon}_t(t-1) \\ R_t &= R_{t-1} + \hat{\epsilon}_t'(t-1)\hat{\epsilon}_t'(t-1)^T,\end{aligned} \tag{2.101}$$

where now, $\hat{\epsilon}_t(t-1)$ denotes the model error at time t obtained using the previous parameter estimate $\hat{\theta}_{t-1}$, and $\hat{\epsilon}_t'(t-1)$ denotes the gradient of $\hat{\epsilon}_t(t-1)$ with respect to $\hat{\theta}_{t-1}$.

2.6.6. Forgetting the Past

When performing a nonlinear optimization by means of the recursive Gauss-Newton method, it is usually helpful to discard the influence of early gradient estimates, since they are typically inaccurate. In this case, the update for R_t becomes

$$R_t = \lambda_t R_{t-1} + \hat{\epsilon}_t'(t-1)\hat{\epsilon}_t'(t-1)^T,$$

where $0 < \lambda_t \leq 1$ is the "forgetting factor." The exponential time constant corresponding to $\lambda_t = \lambda$ is $1/(1-\lambda)$ samples, and thus $1/(1-\lambda_t)$ can be interpreted as the "memory" or "averaging capacity" of the algorithm at time t. If the corresponding off-line Newton's method is taken to be

$$\hat{\theta}_N(i+1) = \hat{\theta}_N(i) - \left(\sum_{t=1}^{N} w_t\left[\hat{\epsilon}_t'(i)\hat{\epsilon}_t'(i)^T + \hat{\epsilon}_t'(i)\hat{\epsilon}_t''(i)\right]\right)^{-1}\sum_{t=1}^{N} w_t\hat{\epsilon}_t(i)\hat{\epsilon}_t'(i),$$

where $w_t = \lambda_t \lambda_{t+1} \cdots \lambda_{N-1}$, then it is easy to verify that the exact recursive update (2.100) for $\hat{\theta}$ is the same. The $\hat{\theta}_t$ recursion (2.101) is unchanged also. Only the R_t update is affected.

When the input and output signal are stationary, and when the parameter estimation is nonlinear, it is generally recommended that λ_t start out small and ultimately approach unity. This is so that early gradient estimates are given less weight in the solution. The use of a forgetting-factor which starts small and grows toward unity can be considered a convergence acceleration technique; other acceleration methods will be discussed in a later section. A simple way to implement a growing forgetting-factor is to define

$$\lambda_t = c\lambda_{t-1} + (1 - c),$$

where $0 \leq c \leq 1$ controls the rise of λ_t to 1, and λ_0 determines the initial "memory." For example, with $c = 0.99$ and $\lambda_0 = 0.9$, one may say heuristically that the averaging-time of the identification algorithm (affecting R_t) rises from about 10 samples to infinity with a time-constant of 100 samples.

When tracking time-varying parameters, λ_t can be set to correspond to the rate of change in $\hat{\theta}_t$. A trade-off appears between averaging-time and parameter bandwidth. For example, setting $\lambda_t = 0.999$ allows on the order of 1000 time samples to be accumulated in R_t, and therefore the true system parameters should be relatively constant over a span of 1000 samples. It would seem possible to set λ_t as a function of "time-variation indicators" such as the short-term prediction error statistics.

2.6.7. Summary of Recursive Identification Algorithms

The generalized recursive Gauss-Newton (RGN) algorithm may be prescribed as follows (using simplified notation from above):

$$\begin{aligned} \hat{\epsilon}_t &= y_t - \hat{\varphi}_t^T \hat{\theta}_{t-1} \\ R_t &= \lambda_t R_{t-1} + z_t \xi_t^T \\ \hat{\theta}_t &= \hat{\theta}_{t-1} + R_t^{-1} z_t \hat{\epsilon}_t, \end{aligned} \tag{2.102}$$

where

$$\begin{aligned} \hat{\varphi}_t^T &= (-y_{t-1}, \ldots, -y_{t-n_a}, u_{t-1}, \ldots, u_{t-n_b}, \hat{\epsilon}_{t-1}, \ldots, \hat{\epsilon}_{t-n_c}) \\ \hat{\theta}_t^T &= \Big(\hat{a}_1(t), \ldots, \hat{a}_{n_a}(t), \hat{b}_1(t), \ldots, \hat{b}_{n_b}(t), \hat{c}_1(t), \ldots, \hat{c}_{n_c}(t) \Big), \end{aligned} \tag{2.103}$$

and the special cases are given by the following table.

Method	z_t	ξ_t
$\hat{\theta}^{(RELS)}$	$\hat{\varphi}_t$	$\hat{\varphi}_t$
$\hat{\theta}^{(RIV)}$	$\hat{\varphi}_t^f$	$\hat{\varphi}_t$
$\hat{\theta}^{(RML)}$	$\hat{\varphi}_t^f$	$\hat{\varphi}_t^f$
General	$-\hat{\epsilon}_t'(t-1)$	$-\hat{\epsilon}_t'(t-1)$

Table 2.2

where

$$\begin{aligned}
\hat{\varphi}_t^f &= \left(-y_{t-1}^f, \ldots, -y_{t-n_a}^f, u_{t-1}^f, \ldots, u_{t-n_b}^f, \hat{\epsilon}_{t-1}^f, \ldots, \hat{\epsilon}_{t-n_c}^f\right) \\
y_t^f &= y_t - \hat{c}_1(t-1)y_{t-1}^f - \cdots - \hat{c}_{n_c}(t-1)y_{t-n_c}^f \\
u_t^f &= u_t - \hat{c}_1(t-1)u_{t-1}^f - \cdots - \hat{c}_{n_c}(t-1)u_{t-n_c}^f \\
\hat{\epsilon}_t^f &= \hat{\epsilon}_t - \hat{c}_1(t-1)\hat{\epsilon}_{t-1}^f - \cdots - \hat{c}_{n_c}(t-1)\hat{\epsilon}_{t-n_c}^f \\
\hat{\epsilon}_t'(t-1) &= \frac{\partial \hat{\epsilon}_t}{\partial \hat{\theta}_{t-1}}.
\end{aligned} \tag{2.104}$$

The recursive least squares method $\hat{\theta}^{(RLS)}$ is given by any of the first three methods with $n_c = 0$.

2.7. Accelerating Convergence

As mentioned previously, the use of a forgetting factor λ_t can improve the rate of convergence in the methods for which $n_c > 0$. In this section, convergence acceleration based on improving the approximate gradient used in the RGN algorithm will be discussed.

Recall from (2.95) that within $\hat{\varphi}_t$, $\{\hat{\epsilon}_{t-1}(t-1), \ldots, \hat{\epsilon}_{t-n_c}(t-1)\}$ is, in a sense, approximated by $\{\hat{\epsilon}_{t-1}(t-2), \ldots, \hat{\epsilon}_{t-n_c}(t-n_c-1)\}$. This approximation allows the simple insertion of $\hat{\epsilon}_t(t-1)$ into a "delay line" within $\hat{\varphi}_t$, avoiding the computation of all residuals $\hat{\epsilon}_k(t-1)$, $, k = 1, \ldots, t$ from the beginning of time for each update. We now examine techniques which improve this approximation at greater computational cost.

2.7.1. Use of the A Posteriori Residuals

For a small increase in computation (having complexity $\mathcal{O}(N_p)$), we can use instead the a posteriori residual estimates $\{\hat{\epsilon}_{t-1}(t-1), \ldots, \hat{\epsilon}_{t-n_c}(t-n_c)\}$ which give a slightly better approximation still exhibiting the important shift structure. Thus at time t, after $\hat{\theta}_t$ is

updated using $\hat{\varphi}_t$ and $\hat{\epsilon}_t(t-1)$, we compute $\hat{\epsilon}_t(t) = y_t - \hat{\varphi}_t^T\hat{\theta}_t$ which is then placed into the $\hat{\varphi}$ vector where $\hat{\epsilon}_t(t-1)$ was used before.

This modification transforms (2.102) into

$$\begin{aligned} \hat{\epsilon}_t(t-1) &= y_t - \hat{\varphi}_t^T\hat{\theta}_{t-1} \\ R_t &= R_{t-1} + z_t\xi_t^T \\ \hat{\theta}_t &= \hat{\theta}_{t-1} + R_t^{-1}z_t\hat{\epsilon}_t(t-1) \\ \hat{\epsilon}_t(t) &= y_t - \hat{\varphi}_t^T\hat{\theta}_t, \end{aligned} \tag{2.105}$$

where

$$\hat{\varphi}_t^T = (-y_{t-1},\ldots,-y_{t-n_a},u_{t-1},\ldots,u_{t-n_b},\hat{\epsilon}_{t-1}(t-1),\ldots,\hat{\epsilon}_{t-n_c}(t-n_c)). \tag{2.106}$$

Simulations show this extra computation to significantly improve the early convergence of the RGN algorithm. As an example, consider the system

$$y_t - 0.8\,y_{t-1} = u_t + 0.034(\epsilon_t + 0.7\,\epsilon_{t-1}) \tag{2.107}$$

where u_t and ϵ_t are independently generated Gaussian noise with unit variance. The scaling of the disturbance by 0.034 results in a signal to noise ratio in y_t of about 10dB.

Table 2.3 lists the sample standard deviation of the prediction errors $\hat{\epsilon}_t$ for the first two sets of 256 points. The sample standard deviations are computed using the formula

$$\hat{\sigma}_\epsilon^2(t_i:t_f) = \frac{1}{t_f - t_i}\sum_{t=t_i}^{t_f}\hat{\epsilon}_t^2. \tag{2.108}$$

In the cases using convergence acceleration, the sample standard deviation is denoted by $\hat{\sigma}_\epsilon^{(A)}(t_i:t_f)$.

Method	$\hat{\sigma}_\epsilon(1{:}256)$	$\hat{\sigma}_\epsilon^{(A)}(1{:}256)$	$\hat{\sigma}_\epsilon(257{:}512)$	$\hat{\sigma}_\epsilon^{(A)}(257{:}512)$
$\hat{\theta}^{(RELS)}$	0.0379	0.0240	0.0223	0.0206
$\hat{\theta}^{(RIV)}$	0.0703	0.0243	0.0250	0.0205
$\hat{\theta}^{(RML)}$	0.0373	0.0237	0.0218	0.0206

Table 2.3. Comparison of prediction-error variances with and without convergence acceleration (superscript A).

2.7.2. Backtracking

Equations (2.96) and (2.97) give an exact recursion for the gradient of the prediction error with respect to the model parameters. Equation (2.98) gives the corresponding approximation used in the RGN algorithms. It is possible to use approximate gradients which fall between these two extremes. The idea is to reach back N_B samples before the current time t and recompute the gradient using the latest parameter estimate. If one reaches back to time 0, then the exact gradient is obtained. If one reaches back one sample, the convergence acceleration using a posteriori residuals is obtained. This experiment was carried out for a variety of values of N_B. The results indicated that $N_B = 1$ (yielding a posteriori residuals) gave nearly all the improvement that is to be had with this technique. Greater values of N_B gave negligible further improvement in the rate of convergence. Thus there is something fundamental about $N_B = 1$ which deserves further explanation. This experiment was suggested by Lennart Ljung.

2.8. Efficient Recursive Updates

In this section we review a method for reducing the computations per time sample in the RGN alorithm to the complexity $O(N_p^2)$. The next section gives a method which brings the complexity down to $O(N_p)$.

The RGN algorithm, as given in (2.102), is dominated computationally by the inversion of $R_t = R_{t-1} + z_t \xi_t^T$. The inversion of a general $N_p \times N_p$ matrix is of complexity $O(N_p^3)$. Since the update of R_t is a "rank 1 correction," we may update its inverse explicitly with $O(N_p^2)$ operations using the so-called "matrix inversion lemma,"

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1}, \tag{2.109}$$

with $A = \lambda_t R_{t-1}$, $B = z_t$, $C = I$, and $D = \xi_t^T$.

Defining

$$P_t \triangleq R_t^{-1}, \tag{2.110}$$

we have

$$\begin{aligned} P_t &= (\lambda_t R_{t-1} + z_t \xi_t^T)^{-1} \\ &= \left[P_{t-1} - \frac{P_{t-1} z_t \xi_t^T P_{t-1}}{\lambda_t + \xi_t^T P_{t-1} z_t} \right] \frac{1}{\lambda_t}. \end{aligned} \tag{2.111}$$

Note that R_0 cannot be chosen as a singular matrix. When it is felt that R_0 should be zero, an arbitrarily good approximation is δI where δ is a small positive number. This implies using the initial condition $P_0 = \delta^{-1} I$.

Computation may be saved by using the fact that

$$P_t z_t = \frac{P_{t-1} z_t}{\lambda_t + \xi_t^T P_{t-1} z_t}. \tag{2.112}$$

2.9. "Fast" Recursive Updates

There is yet more structure in the update of R_t which allows us to further reduce computational complexity. Note that in the algorithms $\hat{\theta}^{(IV)}$, $\hat{\theta}^{(ELS)}$, and $\hat{\theta}^{(ML)}$ the $\hat{\varphi}_t$ vector may be partitioned into three sections, each of which acts as a "delay line" or "shift register." That is, at time t, the way $\hat{\varphi}_t$ is updated for the next cycle is to shift the elements of each partition (corresponding to y, u, and $\hat{\epsilon}$) down one place, and insert the three samples $-y_t$, u_t, and $\hat{\epsilon}_t$ in the vacated positions. As a result, it is possible to perform updates of the quantity $P_t z_t$ with $O(3N_p)$ operations. This result is applicable also to any case of the generalized RGN algorithm (2.101) for which the gradient of the error with respect to the parameters exhibits this time-shift structure. To specify the procedure, we quote a lemma from Ljung and Falconer [110] which is based on the properties of low-displacement-rank matrices introduced by Morf [115,116,117]. The lemma provides an $O(N_p)$ update for the quantity $R_t^{-1} z_t$ which we denote as k_t.

Let ς_t and η_t be two sequences of $p \times 1$ vectors such that $\varsigma_j = \eta_j = 0$ for $j \leq 0$, and let

$$z_t = \begin{pmatrix} \varsigma_{t-1} \\ \vdots \\ \varsigma_{t-n} \end{pmatrix}, \xi_t = \begin{pmatrix} \eta_{t-1} \\ \vdots \\ \eta_{t-n} \end{pmatrix}.$$

Then the quantity

$$k_t = \left(\sum_{j=1}^{t} z_j \xi_j^T + \delta I \right)^{-1} z_t$$

can be determined recursively as

$$e_t = \eta_t + A_{t-1}^T \xi_t \tag{2.113.1}$$

$$A_t = A_{t-1} - k_t \epsilon_t^T (t-1) \tag{2.113.2}$$

$$C_t = C_{t-1} + \xi_t \varsigma_t^T \tag{2.113.3}$$

$$\epsilon_t = \varsigma_t - C_t^T k_t \tag{2.113.4}$$

$$\Sigma_t = \Sigma_{t-1} + \epsilon_t \epsilon_t^T (t-1) \tag{2.113.5}$$

$$\bar{k}_t = \begin{pmatrix} \Sigma_t^{-1} \epsilon_t \\ k_t + A_t \Sigma_t^{-1} \epsilon_t \end{pmatrix}. \tag{2.113.6}$$

Partition $\bar{k}_t$ as

$$\bar{k}_t \triangleq \begin{pmatrix} m_t \\ \mu_t \end{pmatrix} \begin{matrix} (np \times 1) \\ (p \times 1) \end{matrix}. \tag{2.113.7}$$

Let

$$\rho_t(t-1) = \varsigma_{t-n} + B_{t-1}^T \xi_{t+1} \tag{2.113.8}$$

$$B_t = \Big(B_{t-1} - m_t \rho_t^T (t-1) \Big) \Big(I - \mu_t \rho_t^T (t-1) \Big)^{-1} \tag{2.113.9}$$

$$k_{t+1} = m_t - B_t \mu_t . \tag{2.113.10}$$

The initial conditions can be taken as

$$k_1 = A_0 = C_0 = B_0 = 0, \tag{2.113.11}$$

$$\Sigma_0 = \delta I . \tag{2.113.12}$$

This result can be applied to our formulation by re-ordering the elements of $\hat{\varphi}_t$ and $\hat{\theta}_t$

such that

$$\hat{\varphi}_t = \begin{pmatrix} -y_{t-1} \\ u_{t-1} \\ \hat{\epsilon}_{t-1} \\ \vdots \\ -y_{t-n} \\ u_{t-n} \\ \hat{\epsilon}_{t-n} \end{pmatrix} = \begin{pmatrix} \eta_t \\ \vdots \\ \eta_{t-n} \end{pmatrix},$$

and

$$\hat{\theta}_t^T = \left(\hat{a}_1, \hat{b}_1, \hat{c}_1, \ldots, \hat{a}_n, \hat{b}_n, \hat{c}_n\right),$$

where $n = n_a = n_b = n_c$. In the present formulation, $p = 2$ for $\hat{\theta}^{(RLS)}$, and $p = 3$ for $\hat{\theta}^{(IV)}$, $\hat{\theta}^{(ELS)}$, and $\hat{\theta}^{(ML)}$. Using (2.113) to obtain k_t, we have $\hat{\theta}_t = \hat{\theta}_{t-1} + k_t\hat{\epsilon}_t$, where $\hat{\epsilon}_t = y_t - \hat{\varphi}_t^T\hat{\theta}_{t-1}$ as before. These computations are required only for $\hat{\theta}^{(RML)}$ since for $\hat{\theta}^{(RLS)}$, $\hat{\theta}^{(RIV)}$, and $\hat{\theta}^{(RELS)}$, $\hat{\theta}_t$ lies along the top row of A_t^T.

Note closely the coupling of the equations due to the third component of the e_t vector in (2.113.1) being a feed-around and negation of the first component. In the *a posteriori* residuals versions, the top row of Equation (2.113.2) must be computed before Equation (2.113.1) can be finished, and then (2.113.2) can be finished.

In the case of arbitrary n_a, n_b, and n_c, the fast updates may be implemented using a permutation matrix on $\bar{k}_t$ above [110]. The gradient vector is obtained by the same recursion (2.104) as before, giving reordered elements in $\hat{\varphi}_t^f$.

In the symmetric cases ($\hat{\theta}^{(LS)}$, $\hat{\theta}^{(ELS)}$, $\hat{\theta}^{(ML)}$), the recursion for the cross-covariance C_t can be eliminated. Simply set $\xi = z$ in (2.113), and replace (2.113.3) and (2.113.4) by $\epsilon_t = \varsigma_t - A_t^T z_t$.

2.10. Convergence of Recursive Identification Algorithms

It is important to consider the convergence properties of the recursive identification algorithms. Since they were derived as approximations to the offline algorithms, one may suspect that their convergence behavior is similar. This is in fact the case, although it is not easy to show. For a detailed discussion of the convergence of recursive identification algorithms, see Ljung and Soderstrom [111]. The bottom line is that the Recursive Maximum Likelihood (RML) algorithm has the best convergence properties. It can be shown that under ordinary conditions the RML algorithm will converge to a local minimum of the loss function, or to the stability boundary. Intuitively, this is due to the fact that the approximate negative gradient $\hat{\varphi}_t/\hat{C}(d)$ used in RML becomes the true negative gradient when the parameter estimate $\hat{\theta}_t$ is fixed. The Extended Least Squares (ELS) algorithm, on

the other hand, does not always converge to a local minimum of the loss function. This is because the absence of the filtering of $\hat{\varphi}_t$ by $1/\hat{C}(d)$ can cause the gradient to "point the wrong way" and actually prevent convergence.

Convergence of RML

In order to guarantee convergence of the RML algorithm, as given by equation (2.94), the following conditions must be met [111]:

- $1/\hat{C}_t(d)$ must be strictly stable for each time-step t. If the algorithm produces a parameter vector $\hat{\theta}_t$ such that $1/\hat{C}_t(d)$ is unstable, then the roots of $\hat{C}(z^{-1})$ must be projected inside the unit circle.
- The matrix $\frac{1}{t}R_t$ must be strictly positive definite for each $t > 0$. This can be guaranteed by choosing $R_0 = \delta I$ for some small $\delta > 0$. It is also necessary that the input data be persistently exciting of sufficiently high order, i.e., that $\mathcal{E}\{\varphi_t\varphi_t^T\}$ be nonsingular. This ensures that the initial R_0 is "forgotten" and that the eigenvalues of $\frac{1}{t}R_t$ are strictly positive asymptotically.
- The forgetting-factor λ_t must asymptotically approach unity. When $\lambda_t < 1$, convergence cannot occur since the algorithm is "throwing away" past information. For fixed $\lambda_t = c$, the parameter estimates $\hat{\theta}_t$ have a nonzero asymptotic variance which approaches zero as c approaches 1.

In addition to the above restrictions on the algorithm, the following limits involving the data u_t and y_t must exist for each fixed $\hat{\theta}$ such that $1/\hat{C}(d)$ is stable. As before, let the time-averaging operator be denoted by

$$\bar{\mathcal{E}}\{x_t\} \triangleq \lim_{t\to\infty} \frac{1}{t} \sum_{k=0}^{t-1} x_k .$$

Then the required limits are

$$\bar{\mathcal{E}}\left\{\hat{\psi}_t\hat{\epsilon}_t\right\} = f(\hat{\theta})$$

$$\bar{\mathcal{E}}\left\{\hat{\psi}_t\hat{\psi}_t^T\right\} = G(\hat{\theta})$$

$$\bar{\mathcal{E}}\{(1 + |y_t|^3)\} < \infty .$$

The first two conditions state that when the parameter estimate $\hat{\theta}_t$ is "frozen," then the update directions for $\hat{\theta}_t$ and R_t are asymptotically mean-stationary. The third condition is satisfied whenever y_t is bounded, as is always the case in practice.

2.11. Conclusion

In this chapter, many tools of system identification have been derived in a unified and simplified manner. In addition, we examined the relationship of the identification algorithms to the methods of Chapter 1. Various interconnections among the algorithms were discussed. A generalized recursive Gauss-Newton method was defined which can be used in nonlinear problems, and which reduces to the RML, ELS or RLS algorithms when the problem structure permits and when the gradient is approximated in a particular way. It was noted that the gradient approximation associated with use of the a posteriori residuals performed as well in practice as the exact gradient, and much better than the gradient approximation corresponding to the prediction error.

Chapter 3

Modeling the Violin

"Making an instrument is one of music's greatest joys. Indeed, to make an instrument is in some strong sense to summon the future. It is, as Robert Duncan has said of composing, 'a volition. To seize from the air its forms.' Almost no pleasure is to be compared with first tones, tests, and perfections of an instrument one has just made. Nor are all instruments invented and over with, so to speak. The world is rich with models—but innumerable forms, tones, and powers await their summons from the mind and hand. Make an instrument—you will learn more in this way than you can imagine."

— from *Lou Harrison's Music Primer* [246,251]

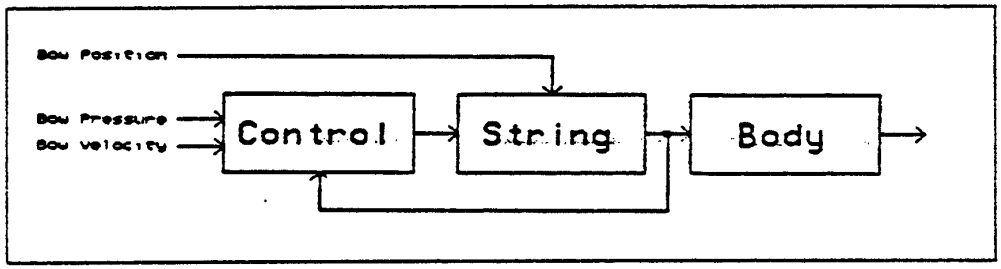

Figure 3.1. Complete violin model to be identified from input-output measurements.

3.1. Introduction

Techniques from the previous two chapters will now be used in the construction of a model for the violin. The violin model will be as shown in Fig. 3.1.

Thus the violin is decomposed into three principal parts—the input excitation, the string, and the body. The body will be modeled by a rational digital filter using the techniques of chapters 1 and/or 2. The string will be modeled as a special type of linear filter which will be introduced in this chapter; the methods of the first two chapters will still apply. The excitation is an external input which will not be treated in detail, but a good first-order approximation is supplied, and a high-quality method based on the physics of bowed strings is described.

The first part of the chapter pertains to modeling the body, and the second part to modeling the string. In the next section, some issues associated with the perception of modeling error are discussed. Based on this information, a pre-processing procedure is developed which will be applied to measured violin frequency-response data. Next, an experiment is described in which the input and output of the violin body were recorded simultaneously, and the results of this recording are discussed. From the measured input-output data, the empirical frequency-response of the violin body is formed. A variety of rational filters are then calibrated to the measured frequency response using methods of Chapters 1 and 2. It turns out that Chapter 1 methods are most successful in modeling the body because they can take advantage of spectral pre-processing. Next a special model for the vibrating string is derived. It's main virtue is that it provides an extremely high-order

model with only a few degrees of freedom. In this case, methods from Chapter 2 prove to be best suited for estimating the parameters of the string from recorded data. Finally, a bowing mechanism is described, and some possible extensions to the overall model are discussed.

3.2. Minimizing Audible Error

> *"Psychologists interested in perception have, ever since the experiments of Adelbert Ames, known that we tend to perceive what we want to perceive. This, in turn, is conditioned by what we have already perceived and by the framework we have been taught to perceive in. We don't just see something 'out there.' Our perceptive apparatus conditions, filters out, focuses the raw data that initiates the process of seeing."*
>
> — Richard A. Lanham [279]

When attempting to fit a parametric model to a naturally occurring audio filter, aspects of human auditory perception should be respected by the measure of fit employed. For example, the ear is much more sensitive to spectral amplitude than phase. Also, the "critical bands" of the ear have a strong bearing on the importance of fine-structure in the spectrum, with higher frequencies being less individually resolved. Knowledge of the discernability of various spectral modifications allows more efficient use of the degrees of freedom in a model. Accordingly, in this section, some relevant findings from psychoacoustics research will be reviewed. These facts will be used to define an *error criterion* which is well-suited to audio modeling.

At the outset, the domain of audio signals considered will be restricted to a subset which is fairly well understood. Since the main purpose at hand is the modeling of linear filters in nature, we may focus on perceivable differences in *frequency response* functions. For an audio filter with a short impulse response, such as the body of a violin, the frequency response will be dealt with primarily in terms of its effect on *steady-state tones*. That is, we may compare the outputs of the true filter and the model for a relatively small set of stationary or periodic excitations. A natural generalization of this method of comparison, necessary when no *a priori* restriction is possible on the class of input signals, is to treat the true and approximate frequency response functions as steady-state spectra themselves. This is most reasonable for models driven by a periodic impulse train or white noise, and valid only for systems having a short impulse response (i.e. when transient effects in the filter have negligible effect on the perceived sound). For systems with long impulse responses, such as concert halls, this approach is highly incomplete since the temporal structure of a long impulse response is an important determinant of quality. Although the time-structure of the impulse response is contained in the phase of the frequency response, it is no longer analogous to the phase of a steady-state tone spectrum. Thus it is important to distinguish between *spectral-magnitude* and *reverberant* properties of an acoustical filter.

For the purposes of studying the perceivable differences between steady-state tones, the ear may be modeled as a *spectrum analyzer* with limited frequency resolution. This idea dates back to Helmholtz who, in 1863, provided the foundations for this point of view in his classic treatise *On the Sensations of Tone* [252]. Since then, much research has been done to refine the basic spectrum analyzer model of the ear. Two excellent books on this subject are *Aspects of Tone Perception* by R. Plomp [263], and volume IV of the *Handbook of Perception* [248].

3.2.1. The Importance of Phase

It is generally known that the phase relations among the components of a sum of sinusoids do not contribute significantly to the perceived sound. However, a closer look at the nature of phase perception will be useful for understanding the limitations of an approximation which ignores phase. The following summary for the case of sums of sinusoids will serve to highlight some of the main points [263]:

- Phase differences are more noticeable at higher frequencies.
- Discrimination between two complex harmonic tones on the basis of phase is easier at low pitches.
- Phase relations which significantly modify the "peakiness" of the time waveform (the ratio of maximum and minimum of the amplitude envelope) can result in different timbres. Conversely, if the phase of a spectrum is changed in a way which does not appreciably alter the amplitude envelope, then typically no difference is perceived.
- In a tone consisting of 10 harmonics with amplitudes inversely proportional to frequency, the greatest phase discrimination is observable between the case of a sum of sines (or cosines)

$$\sin(\omega t) + \frac{1}{2}\sin(2\omega t) + \frac{1}{3}\sin(3\omega t) + \frac{1}{4}\sin(4\omega t) + \cdots + \frac{1}{10}\sin(10\omega t),$$

and a harmonic sum in which sine and cosine are alternated

$$\sin(\omega t) + \frac{1}{2}\cos(2\omega t) + \frac{1}{3}\sin(3\omega t) + \frac{1}{4}\cos(4\omega t) + \cdots + \frac{1}{10}\cos(10\omega t).$$

The perceived difference between a sum of sines and a sum of cosines in this situation is negligible.

- In the previous case, changing the slope of the spectrum magnitude by a small amount has a more pronounced effect on perception than does the maximally different phase modification. This amount varies from person to person, and in a

study involving eight subjects, the change in slope comparable with the maximal phase distortion, in the context of timbre discrimination, ranged from 0.2 dB per octave to 2.7 dB per octave.

- Discrimination based on phase seems to be concentrated within "critical bands" (see below).

These effects may be explained to a degree in terms of critical bands, nerve-cell firing-rate modulation, and "combination tones" due to mild nonlinearities in the response of the ear [263].

In view of the above facts, an important simplification available when modeling systems with *short impulse responses* is that *phase information can be largely ignored.* Further considerations which support the unimportance of phase errors are that

(1) When the position of the listener changes relative to the source, the phase of the spectrum is modified (when the sound radiates from other than a point source, or reflections are present).

(2) In a reverberant sound field, the phase of the spectrum received by the ear is randomized.

In cases where phase is unimportant, it is typically best to convert the desired frequency response into the corresponding *minimum phase* frequency response. This allows most methods for system identification and filter design to give their best results in terms of magnitude fit. Again it is emphasized that phase-response can be ignored only for systems with impulse responses which are short compared with transients in the input signal.

3.2.2. Perception of Phase-Delay and Group-Delay Distortion

The previous discussion of phase effects applies only to the case of periodic or stationary excitation. For *transient* sounds, a more relevant measure of phase-response distortion is given in terms of the *group delay* and *phase delay.* (See Appendix E for definitions of these quantities.) There is only a small amount of data available on the perception of such distortion, and a summary is given by Preis [267]. The conclusions (based on narrow-band signal tests) are that

- At low frequencies (below 500 Hz), deviations in group-delay on the order of a few milliseconds are imperceptible.
- At high frequencies, the ear becomes more sensitive to delay distortion. For example, between 1 and 5 KHz, group-delay errors greater than ± 0.5 msec can be perceived under sensitive test conditions.

In complex contexts, such as natural speech or music, the threshold would be more than twice the nominal values above [267].

For the case of lowpass filtering, Bloom and Preis have determined the phase-distortion audibility threshold for two filter cut-off frequencies [247]. The excitation signal used was (effectively) an impulse. Their results were that

- At 4KHz cut-off, two seventh-order elliptic function filters in cascade give audible phase distortion. Four cascade eighth-order Butterworth filters give audible phase distortion. (The trials always used one or more *pairs* of cascade filter sections.)
- At 15KHz cut-off, up to eight seventh-order elliptic function filters in cascade give *no* audible phase distortion.

Thus, while phase-sensitivity increases from low to middle frequencies, at very high frequencies (near the limits of hearing), the ear becomes relatively insensitive to this type of distortion.

3.2.3. Frequency Resolution

The resolving power of the ear as a spectrum analyzer varies with frequency. Since hearing is not a linear process, it is impossible to arrive at a definition of frequency resolution which holds in all circumstances. However, in the context of discriminating steady-state timbre, a reasonable definition can be made based on *critical bands*. The concept of critical bandwidth in the ear is itself rather vaguely defined, being somewhat dependent on the particular experimental procedure employed.

From measurements of critical bandwidth based on masking [248,249,274], a reasonable approximation to the frequency-resolution of the ear is given by

$$B_c(f) \approx \begin{cases} 100\text{Hz}, & f < 500\text{Hz} \\ f/5, & f \geq 500\text{Hz}. \end{cases} \tag{3.1}$$

From Plomp [263] we have the following general guidelines regarding the frequency-resolution of the ear:

- "For complex sounds with equal amplitude components, the ear is able to identify these partials as long as their frequencies are separated by more than 15–20% (first 5 to 7 harmonics of a complex tone), with a minimal frequency distance of about 60 Hz."
- This resolving power is consistent with critical bandwidth measurements made using direct masking.

- When tones are not simultaneous, the critical bandwidths can be as small as half the bandwidth measured for simultaneous tones (due to lateral suppression).

Note that the ear can resolve two sinusoids much closer together than a critical bandwidth. However, the evidence is that when a complex spectrum is perceived as a whole, the resolution of the ear corresponds to critical bands. This distinction is analogous to the distinction between "acute" and "peripheral" vision of the eye.

3.2.4. Perception of Amplitude Spectrum

Since phase changes in a steady-state tone are so weakly perceived, the perception of distortion in the amplitude spectrum of a tone is central to determining an appropriate frequency-response error measure to be used in the modeling of short-memory systems.

One successful approach to defining such a distortion measure is based on a model for *loudness summation* proposed by Zwicker and Scharf [274]. The central premise is that the *loudness* of a complex sound is derivable from its *excitation pattern*. This is based on the observation [272] that as the frequency spread of an ensemble of sinusoids is increased, the loudness remains constant until the overall bandwidth exceeds that of a critical band. The computation of loudness from the power spectrum proceeds as follows [263]:

- The power spectrum is converted to an excitation pattern (dB SPL vs. log frequency). The excitation pattern is defined as the masking pattern plus 3 dB for low and middle frequencies, and 6 dB is added for high frequencies. (These corrections correspond to just noticeable differences in level for narrow-band noise.)
- The excitation pattern is weighted (above 2 KHz) according to the frequency response of the middle ear.
- The frequency axis is warped such that critical bandwidth is independent of center-frequency (the so-called *Bark* frequency scale [274]).
- The power in each critical band is summed and converted to a *specific loudness* for the band. Based on experimental results, a factor of 2 in specific loudness is made to correspond to a 12 dB shift in the excitation pattern.
- Zero loudness is accounted for by assuming a physiological background noise at the hearing threshold (which is inaudible). The hearing threshold is not uniform with respect to frequency.
- The loudness estimate, in *sones*, is given by the integral of the specific loudness curve along the Bark frequency axis.

While this model is carefully constructed on the basis of psychophysical measurements and provides insight into the mechanics of aural processing, a simpler procedure, also due

to Zwicker [273], has found greater use in practice. In the simpler technique, loudness is computed directly from the output of *1/3-octave bandpass filters*. (1/3-octave filters give spectral resolution comparable to critical bands.) A program for performing this computation is given in [260] and the method has been accepted as an international standard (Method B of ISO Recommendation R-532).

For the purpose of predicting dissimilarity between steady-state tones, Plomp states [263, p. 94] that

> "for stimuli with modest differences in amplitude spectrum, this [elaborate loudness summation] procedure appears to give predictions that are hardly better than a much simpler procedure directly based on the sound pressure levels within 1/3 octaves."

In [258], the findings of [265,266] are summarized to state that "differences in sound spectrum, measured in one-third octave bands, was a good first-order approximation of the physical correlate of timbre dissimilarity."

Plomp's measure of timbral dissimilarity between two tones i and j is computed as

$$J_p \triangleq \left(\sum_{k=1}^{15} (L_{ik} - L_{jk})^p \right)^{1/p}, \qquad p = 2, \tag{3.2}$$

where L_{ik} is the sound pressure level of signal i at the output of the kth 1/3-octave bandpass filter. Plomp notes that $p = 1$ gave slightly better results than $p = 2$, but that the value of p "appears not to be very critical."

A particularly relevant experiment conducted by Plomp [262,263] consisted of measuring the *perceived difference* in timbre between two steady-state harmonic tones, and comparing this with the difference measured using (3.2). Nine tones were generated by replicating a single period from a note at 349 Hz played on nine orchestral instruments, and these tones were scaled to have equal loudness. For the perceptual difference measurement, ten subjects listened to three tones and were asked to decide which pair was most *similar* and which pair was most *dissimilar*. From these comparisons, the dissimilarity associated with each tone pair was derived. *Multidimensional scaling* techniques were then used to form a Euclidean "timbre space" in which linear distance corresponds to dissimilarity. Each tone appears as a point in this space. The correlation coefficient for the two distance measures was 0.8 for $p = 2$ and 0.86 for $p = 1$. Thus the quantitative distance measure (3.2) correlates well with the qualitative perception of dissimilarity in harmonic spectra.

The timbre space could be projected into three dimensions with only 2.3% "stress" (a measure of discarded information in multidimensional scaling [255]). This value of stress

is said to correspond to an "excellent" goodness of fit. Similarly, *principal components analysis* [259] was applied to the 15-dimensional manifold corresponding to (3.2); the first three principal dimensions accounted for 90.4% of the spectral variance. The correlation coefficients between the three principal axes of the timbre space and the space generated by (3.2) were 0.993, 0.987, and 0.912, in order of importance. Thus, when the degrees of freedom in the model is reduced to three, excellent agreement is obtained between these qualitative and quantitative distance measures.

3.2.5. Perception of Formant Resonances

Some research has been done directly on the perception of *formants* in the spectrum of speech sounds [170, 250]. The following is a summary from Flanagan [170].

- The JND ("Just Noticeable Difference") for the *overall intensity* of a vowel is about 1.5 dB.
- The JND for the *overall intensity* of wide-band noise is about 0.4 dB for sensation levels above 30 dB.
- The JND for the *intensity of the second formant* of a near-neutral vowel is about 3 dB.
- The JND for the *intensity of a harmonic* in the "valley" between formants can be as much as +13 dB to $-\infty$ dB (i.e. zero amplitude).
- The JND for the *bandwidth of a formant* is on the order of 20 to 40%.*
- The JND for the *fundamental frequency* of a vowel (male speech) is about 0.3 to 0.5% (5 to 9 cents).
- For a filtered white noise, the minimum perceptible Q is about 5 dB for a two-pole resonance and 8 dB for a two-zero anti-resonance.

We have briefly reviewed some aspects of perception of steady-state spectra. Awareness of these properties of hearing can be valuable in obtaining efficient models for audio applications.

* It has been observed [280] that for models of the singing voice, the first formant bandwidth should be more accurate than this, while high-frequency formants can be less accurately modeled. Such a rule-of-thumb seems also to be true of violin body models.

3.3. Pre-Processing for Time-Invariant Audio Spectra

In this section, a method is described for smoothing steady-state spectra so as to eliminate information of small perceptual significance. The smoothing will be applied to the violin frequency-response data described in the following section. The main advantage of such pre-processing is that a model can be more readily fit to the most important features. On the other hand, it is difficult to define a pre-processing strategy for frequency-response functions which does not give up some important attributes of the original. This is because perceptually important characteristics of a frequency response depend on the particular signal used as input.

The first data-reduction step is the *elimination of phase* information. This is done initially by taking the squared magnitude of the frequency response to obtain the *power frequency-response*. For modeling methods which are sensitive to phase, a *minimum-phase* frequency-response will be constructed [169].

The second step is *smoothing according to critical bands* of the ear. For this step, a "moving average" filter is applied across the power frequency-response which grows in length as it progresses to higher frequencies. The filter length grows so that spectral power is averaged over roughly a critical bandwidth. The equation for the filter length (in Hz) at each frequency is given by (3.1). Thus for frequencies below 500 Hz, the smoothing is over a fixed 100 Hz interval, and for higher frequencies the smoothing extends over an interval which is 20% of the frequency at the mid-point of the window.

The third pre-processing step consists of *warping the frequency axis* to *normalize* critical bandwidths as much as possible subject to the constrained form of the mapping. Thus the frequency axis is made to approximate the Bark frequency scale [274]. The frequency-warping is restricted to the class of first-order conformal maps as discussed in Chapter 1 (§1.9.1). This restriction is necessary so that a digital filter fit to the warped spectrum can be unwarped without increasing its order. Another function of the conformal mapping can be to provide an effective *weight function* on the frequency-response error. For example, by stretching the low-frequency axis more than necessary to achieve constant critical bandwidths, increased emphasis is placed on the fit at low frequencies. This happens because the fine-structure in the spectrum at low frequencies is spread out over a much larger interval, and it becomes "easier to follow" with the frequency response of a digital filter.

Note that steps (1) to (3) produce data similar to the power output of a one-third octave filter bank. As discussed in the previous section, spectral distance-measures based on power in one-third octave bands correspond closely to perceived timbre for steady-state tones.

As a fourth step, one may apply a rational *pre-emphasis* function as discussed in §1.5.5. However, pre-emphasis is not included in the general pre-processing specification because it must be carried out with characteristics of the driving-signal and the error-criterion in mind. The purpose of such an "equalization" is to effect a weight-function on the spectral error in the modeling procedure. For example, in modeling the violin body, a weighting of this sort is desirable to force greater accuracy on low-frequency detail in the filter response.

Algorithm Summary

Let all frequency-responses be defined on N equi-spaced points on the unit circle, including the point $z = 1$. Also assume that all impulse response functions are real so that only $N_s = \lfloor N/2 \rfloor + 1$ points are needed in computations (the upper half of the unit circle is taken). Define $\omega_k = 2\pi k/N$. Let T denote the sampling period in seconds, and $f_s = 1/T$ the sampling rate in Hertz. Then the pre-processing consists of the following steps:

(1) *Convert to power response.*
Replace the desired frequency response $H(e^{j\omega_k})$ by

$$H_p(e^{j\omega_k}) \triangleq \left| H(e^{j\omega_k}) \right|^2, \quad k = 0, \ldots, N_s - 1 .$$

(2) *Smooth according to critical bands.*
Filter the desired power response H_p to obtain

$$H_s(e^{j\omega_k}) = \frac{1}{N'_a(k)} \sum_{m=-N_l(k)}^{N_u(k)} H_p(e^{j\omega(k+m)}), \quad k = 0, \ldots, N_s - 1 ,$$

where

$$N_l(k) \triangleq \min\left\{ \left\lfloor \frac{N_a(k)}{2} \right\rfloor, k \right\}$$

$$N_u(k) \triangleq \min\left\{ \left\lfloor \frac{N_a(k) - 1}{2} \right\rfloor, N_s - 1 - k \right\}$$

$$N'_a(k) \triangleq N_l(k) + N_u(k) + 1 ,$$

and

$$N_a(k) \triangleq \begin{cases} \left\lfloor N\frac{100}{f_s} + \frac{1}{2} \right\rfloor, & k < N\frac{500}{f_s} \\ \left\lfloor \frac{k}{5} + \frac{1}{2} \right\rfloor, & k \geq N\frac{500}{f_s} . \end{cases}$$

(3) *Warp the frequency axis to improve low-frequency resolution.*

Replace the smoothed power response H_s by

$$H_\rho(e^{j\omega_k}) \triangleq H_s\left(\frac{e^{j\omega_k}+\rho}{\rho e^{j\omega_k}+1}\right), \quad k=0,\ldots,N_s-1\,.$$

where

$$\rho \triangleq \frac{\sin(\pi(f_i-f_r)T)}{\cos(\pi(f_i+f_r)T)},$$

and f_r is an arbitrary reference frequency in Hz, f_i is the desired image of f_r under the mapping, and T is the sampling period. Note that the value of $H_\rho(e^{j\omega_k})$ is assigned the value of H_s at

$$e^{j\omega\varphi_k} \triangleq \frac{e^{j\omega_k}+\rho}{\rho e^{j\omega_k}+1}$$

$$\Rightarrow \quad \varphi_k = \tan^{-1}\left(\frac{(1-\rho^2)\sin(\omega_k)}{2\rho+(1+\rho^2)\cos(\omega_k)}\right) \triangleq \frac{2\pi k_\rho}{N}\,.$$

Thus the kth element of the H_ρ array is assigned the k_ρth element of the H_s array. Since k_ρ is not an integer in general, one may wish to use interpolation of the H_s values.

This completes the pre-processing. The remaining modeling steps are as follows:

(4) *Fit a digital filter to $H_\rho(e^{j\omega})$ to obtain $\hat{H}_\rho(z)$.*

This filter will be mapped back into the original frequency domain by means of the inverse of the mapping used in step (3).

(5) Compute

$$\hat{H}(z) = \hat{H}_\rho\left(\frac{z-\rho}{1-\rho z}\right)$$

as the final approximate filter.

Discussion

Note that in step (2), low-frequency resonances with bandwidths less than 100 Hz will be flattened somewhat. This is undesirable since it can significantly alter the balance of the low-order partials relative to higher partials when a periodic excitation is present. When sharp low-frequency resonances are present, it may be advisable to reduce or eliminate the low-frequency smoothing.

Heuristically, ρ is set in step (3) to horizontally stretch the low-frequency spectrum. A good first choice is that which makes critical bands have constant bandwidths over the entire

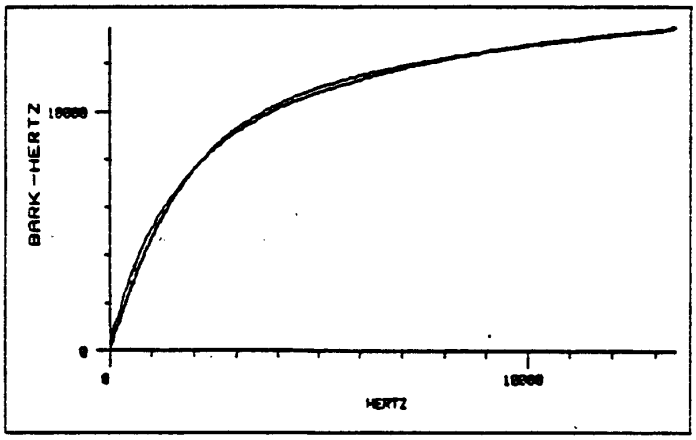

Figure 3.2. Overlay of the frequency mapping provided by the Bark scale with that provided by an optimum first-order conformal map. This figure holds only for a sampling rate of $f_s = 27$ KHz. The Bark-frequencies have been multiplied by $(f_s/2)/24$ to facilitate comparison with the conformal map frequency-scale.

spectrum. In this case, fine-structure in the power-spectrum is made more "perceptually uniform." The frequency-scaling which accomplishes this is the *Bark frequency scale* [274]. There are approximately 25 critical bands covering a range of zero to 13.5 KHz in the Bark scale data published by Zwicker.*

Of course, the mapping is highly constrained, and one can only approximate the Bark scale. What is surprising, however, is how close the mapping can come to the Bark scale (which is based on measured psychophysical data). An example match for the (maximum available) sampling rate 27 KHz is shown in Fig. 3.2. The fit is even better at lower sampling rates due to Bark frequency-mapping being closer to linear. Table 3.1 gives a list of Bark-scale mapping-constants for various sampling rates.

* The Bark scale values at $0, 1, 2, \ldots, 24$ correspond respectively to frequencies (in Hz) 0, 50, 150, 250, 350, 450, 570, 700, 840, 1000, 1170, 1370, 1600, 1850, 2150, 2500, 2900, 3400, 4000, 4800, 5800, 7000, 8500, 10500, and 13500. These values were interpolated using cubic splines. The author is grateful to John Grey and John Gordon for making this function and associated software available.

Bark Frequency Conversion			
f_s	f_i	ρ	error
6	0.983	0.3572604	0.0078
8	1.159	0.4218740	0.0081
10	1.337	0.4765069	0.0089
12	1.500	0.5177548	0.0088
14	1.646	0.5490850	0.0087
16	1.781	0.5746753	0.0092
18	1.910	0.5965544	0.0102
20	2.036	0.6159932	0.0111
22	2.164	0.6339815	0.0120
24	2.295	0.6507166	0.0126
26	2.423	0.6657497	0.0132
27	2.484	0.6724681	0.0135

Table 3.1. Table of mapping constants which provide approximations to the Bark frequency scale for various sampling rates. The sampling rate f_s and the image-frequency f_i are in KHz. The image-frequency f_i is the image of $f_r = 500$ Hz. The value of f_i was optimized by least-squares to within 1 Hz using a bisection method [174]. The conformal mapping constant ρ can be used directly in step (3) of the pre-processing procedure. The error measure is the root-mean-square deviation between the Bark scale and the frequency-scale generated by the mapping, divided by $f_s/2$.

An example of the practical performance is given in Fig. 3.3. A sum of 24 sinusoids was generated at frequencies $1, 2, 3, \ldots, 24$ Bark. In other words, the sinusoids are spaced in frequency by critical bandwidths. The spectrum of this signal is shown in Fig. 3.3a. After mapping the frequency at 500 Hz to frequency 2484 Hz (corresponding to $f_s = 27$ KHz), the spectrum appears as shown in Fig. 3.3b. As one can see, the mapping gives an excellent approximation to the Bark frequency scale.

The frequency-warping according to critical bands does not, in principle, eliminate any information contained in the data. In practice, however, the size N_s of the spectrum array should be sufficiently large so that adjacent spectral samples can be accurately linearly interpolated. Due to the excellent agreement between the Bark scale and the conformally mapped frequency scale, it is reasonable to perform the mapping *before* the smoothing. In the warped frequency coordinates, the smoothing is uniform and therefore easier to implement.* However, at sampling rates higher than 27 KHz, it may be better to smooth first, since the mapping to Barks becomes less accurate at higher sampling rates.

One might also consider using the Bark scale to set the size of the smoothing filter in step (2). This is indeed possible. The slope of the Bark-to-Hz curve may be taken as

* Uniform smoothing can be accomplished by a much wider variety of techniques. See, for example, Oppenheim and Schafer [191] on cepstral ("homomorphic") smoothing.

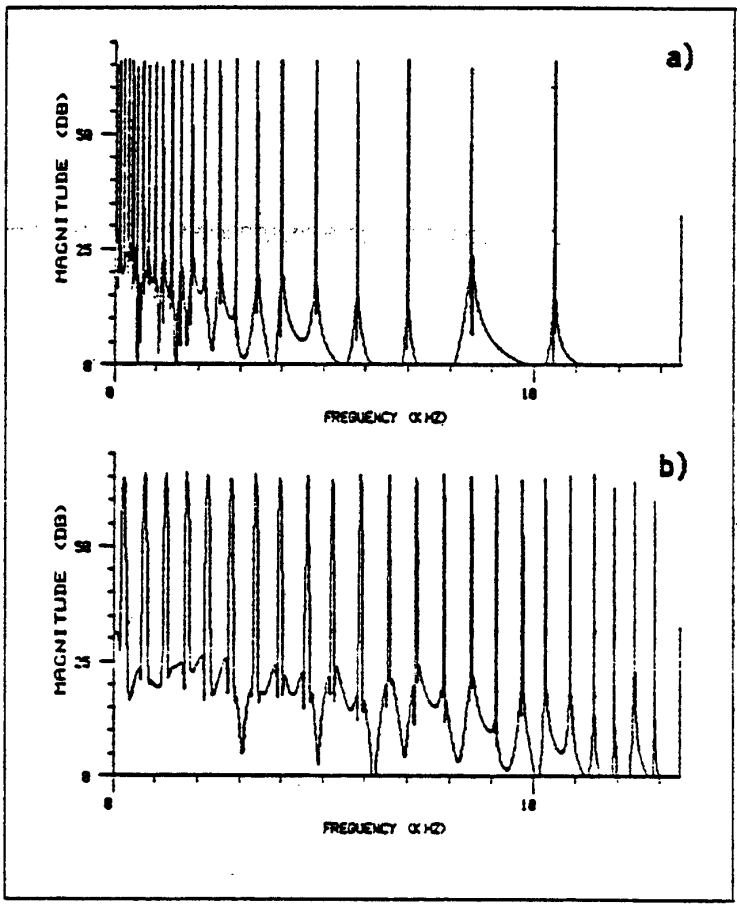

Figure 3.3. Performance of the Bark-conformal-map at 27 KHz on a sum of 24 sinusoids spaced according to critical bandwidths.

a) Spectrum prior to mapping.

b) Mapped spectrum. The apparent difference in the heights of the spectral lines is due to insufficient plot resolution; such error due to the frequency mapping procedure is not visible to the eye. The FFT size is 4096 and a Hamming window was used. The mapping constant is $\rho = 0.6724681$.

a measure of "instantaneous critical bandwidth" which can be used to set the FIR filter length. This, however, was found to be unnecessary effort. Figure 3.4 shows an overlay of the Bark critical bandwidth data with the simple estimate proposed in step (2). It was felt that the curves are sufficiently close that the simple form would suffice. Another reason is that the smoothing according to critical bands is not rigorously justifiable. The function in

step (2) is merely a reasonable point-of-departure in the search for a successful smoothing strategy.

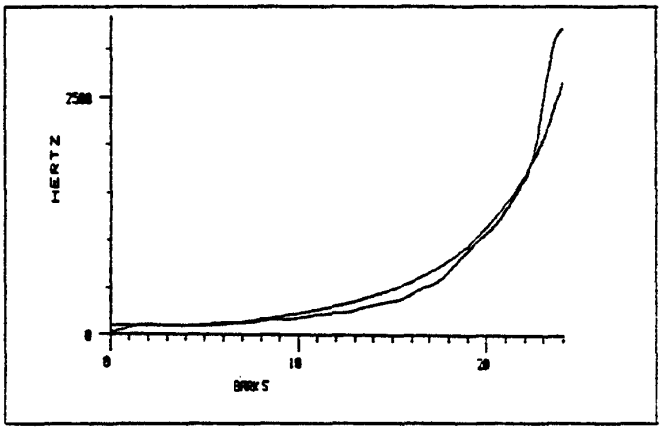

Figure 3.4. Comparison of the critical bandwidths estimated from the Bark-scale with those generated using the piecewise linear approximation in (3.1). There are 256 points displayed. The discrete Bark function was interpolated using cubic splines and differentiated to obtain critical bandwidth.

3.4. Violin Frequency-Response Measurement

The next goal is to obtain a measured frequency-response for the body of a violin. This will become $H(e^{j\omega})$ in problem $\hat{H}^*$, and a model for the body will be designed using the methods of Chapter 1. Also, input-output measurements will be used in applying the methods of Chapter 2. The data-collection task can be very delicate. Norman Pickering [237] summarizes the problem as follows:

> "The acoustic spectrum of a stringed musical instrument is a complex affair, covering a significant frequency range of nearly 10 KHz and a dynamic range of about 60 dB. It consists of from 50 to 100 successive resonant peaks, with Q factors which vary from 5 to over 100. Furthermore, the spectrum of a given instrument varies with time, humidity, and the state of adjustment to an extent which may make the difference between musical acceptability and the lack of it. It is startling to observe the audible effect of small changes in amplitude in limited portions of the frequency range. Despite the fact that the spectrum is so very far from 'flat,' a change of less than 2 dB in any one of the principal resonances is clearly recognizable by a skilled player."

These observations indicate the requirements for measurement as well as modeling precision. The claimed JND of about 2 dB for a major resonance must be carefully interpreted. One of the effects of a change in body resonance is a change in the "feel" of the instrument. A strong body-resonance coupled to the string provides greater energy dissipation in that frequency region (possibly even a "wolf note"). Some strongly coupled resonances do not correspond to efficiently radiating body modes. Thus it is possible that noticeable differences in the playability arise more from changes in responsiveness of the strings. Also, it is typically not possible to change one mode without affecting the higher modes corresponding to the same geometry. For example, brightness may be affected by the attenuation of an entire series of modes corresponding to a specific physical modification. Recall also (§3.2.5) that a 2 dB change in the middle-formant amplitude of a neutral spoken vowel is imperceptible. These points are brought up because of the fact that the experimental results to be discussed later indicate that a much wider tolerance in the resonances is allowable when the sound quality is the only concern.

3.5. Measuring Violin-Body Input-Output Data

In this section a method for measuring the input-output characteristics of the violin body is described. This leads to an empirical frequency response which will be approximated by a rational digital filter using techniques of Chapter 1. It also provides signals suitable for system identification techniques described in Chapter 2. The first issue is what to measure. The output signal is more straightforward, so it is considered first. It should be noted that the ultimate use of the measurements will be to provide a musically useful model. This relaxes some of the requirements for rigor in experimental technique. The main criterion for judging a set of measurements is how well they capture musically important information.

3.5.1. The Output Signal

Loosely speaking, the output signal is "what you hear." The pre-processing procedure developed in §3.3 is designed to convert (in an approximate manner) from *sound pressure* in the air to the excitation envelope along the basilar membrane. Thus we wish to measure sound pressure radiated by the violin body. This should ideally be done for a representative set of listening positions in an anechoic room. . . .

Based on the simple observation that the sound of a violin does not change very drastically when one changes position, only one measurement is made at one point in space in our experiment. It is desirable to place the microphone sufficiently far from the violin body that all radiating elements are represented, yet close as possible to maximize the signal-to-noise ratio of the recording. As a compromise, we chose a point approximately one foot from the top plate, roughly over its mid-point. However, this positioning was varied for ease of play.

The chief differences noted in the measured sound-pressure spectrum due to changing the observation point was the movement of spectral nulls. The violin body is a distributed source, and changing the point of observation changes the phase relationships among the rays in the sum. Since these changes obviously have a small second-order effect on the steady-state sound, they should be largely ignored or modeled only statistically. This argument supports the use of critical-band smoothing. Nulls due to summing spatially distributed sources become more dense at high frequencies, and thus more smoothing is called for at higher frequencies, as occurs in the critical-band smoothing algorithm.

3.5.2. The Input Signal

The input to the violin body is at the bridge. The input from a given string has a *force* and *velocity* component in *three dimensions.*

Due to the geometry of the bridge, there is poor coupling of motion parallel to the string, and this component is ignored. Since the bow moves along a line tangential to the top of the bridge, one would expect that the main motion of the string is in this direction (to be known as the *horizontal* bridge excitation). It is common practice to measure string motion in this direction only. However, the component of force at the bridge normal to the top plate (the "vertical component") is known to be significant [281], and if it is to be ignored, pains should be taken to avoid exciting it. It would be better to identify *two* transfer functions—one from the vertical and one from the horizontal directions of excitation. Since the string is the least linear element of the entire system, the tension modulation in the string (which is a rectified version of the waves propagating on the string) can provide a

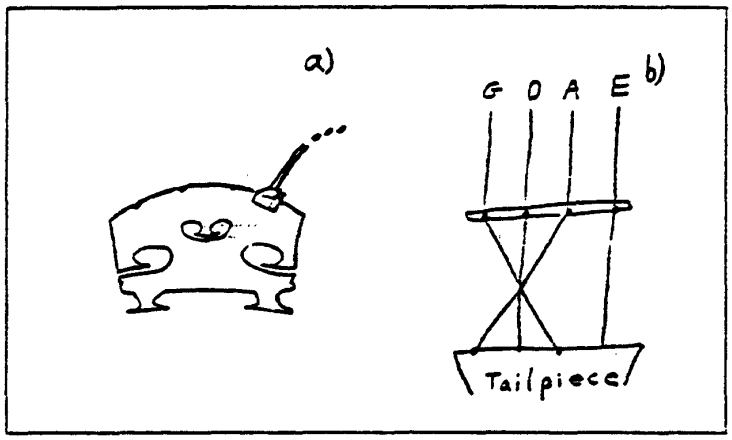

Figure 3.5. Violin modifications for measuring force at the bridge.
a) Violin bridge with piezoelectric transducer attached.
b) Diagram of how the G and A strings are swapped at the tailpiece.

significant contribution to the vertical excitation at the bridge. We attempt to measure only the transfer function associated with horizontal excitation.

Since the bridge is a relatively rigid termination, *force* is the primary input variable. We will consider force at the bridge as the controllable input, and it will be measured in the lateral direction, tangential to the top of the bridge.

3.5.3. Description of Recording Apparatus

The sound-pressure was measured using a high-quality PZM audio condenser microphone. The Barcus-Berry Hot Dot is a "sub-miniature" piezoelectric transducer which was employed for the measurement of lateral force at the string-bridge termination. A special bridge constructed for this purpose is shown in Fig. 3.5a. The piezoelectric crystal was epoxied to the bridge to provide a transverse termination for the G string. The bridge had a small rectangle of material removed (with a pen-knife) so that the center of the sensitive face of the crystal would be aligned with the top of the bridge, and so the crystal would have solid mechanical support. On the face of the transducer, a small metal rod was epoxied to provide a better string termination. This rod had a small groove etched into it, and it was mechanically similar to the metal bridge-rods found on many acoustic guitars. The resulting termination appeared to be quite steady, allowing no observable slipping of the string. The G and A strings are swapped at the tailpiece in order to introduce a lateral "bias" force on the pickup, as shown in Fig. 3.5b. This configuration was inspired by the electronic violin built by Max Mathews [229].

Three pre-amps were tried with the transducer: a voltage-to-voltage amplifier, an integrating current-to-voltage amplifier (known as a "charge amp" [118,177]), and a (non-integrating) current-to-voltage amplifier. The first two gave comparable results, and the observed force waveforms at the bridge were very similar to those published in the literature. The third, which provided a +6 dB per octave pre-emphasis, provided the best data. The bridge and microphone signals were low-passed to 20 KHz and simultaneously digitized at 44.642 KHz with a 14-bit A/D converter.

3.5.4. Results

Several types of excitation were tried. The main purpose of course was to obtain a good estimate of the frequency response. This was not a trivial task. The main problem seemed to be obtaining an excitation at the bridge which excited all frequencies equally. Bowed excitations were tried, but these are insufficient because the lowest pitch is around 200 Hz, and sampling the frequency response at 200 Hz intervals is far too sparse. It is possible, however, to combine bowed excitations at a dense set of pitches to construct a spectrally rich source. (One simply adds the bridge input measurements from several recordings, and adds the corresponding microphone outputs together to form a spectrally rich input-output pair, by superposition. One could even go so far as to optimize the coefficients of a linear combination of diverse signal types so as to optimize the effective excitation.) Also, glissandos were recorded, giving a "chirp response." With the strings wrapped in cloth to suppress their vibration, we recorded bow-noise response (the bow being slid close to the bridge at very light pressure to produce a smooth hissing sound), raucous squawks (producing rather isolated bow-slips which are like impulses), the response to striking the transducer in the lateral direction with a metal rod (a hefty screw-driver actually), and finally, the response to plucking near the bridge with a guitar pick. Of all these, the guitar-pick excitation turned out to yield the best data. The worst data came from the metal-rod excitation, presumably due to the direct sound of the metal-to-metal elastic collision between the rod and the bridge element installed on the transducer. The squawks produced results in agreement with the plucking.*

* Omission of discussion of the other recordings does not necessarily imply they are bad approaches. In the case of bow-noise excitation, for example, the recording was rejected due to poor signal levels, and the glissando recording had clipping in many places. The recordings were hard to get because the new digitizer interface would crash the time-sharing system every third trial or so, causing the loss of in-core text to people editing files. (Experiences such as this are enough to make a person pursue pure theory forever.) The idea of constructing a superposition of harmonic excitations was not tried due to time-constraints. Basically, I declared the data collection phase to be finished once I got similar results from two different methods.

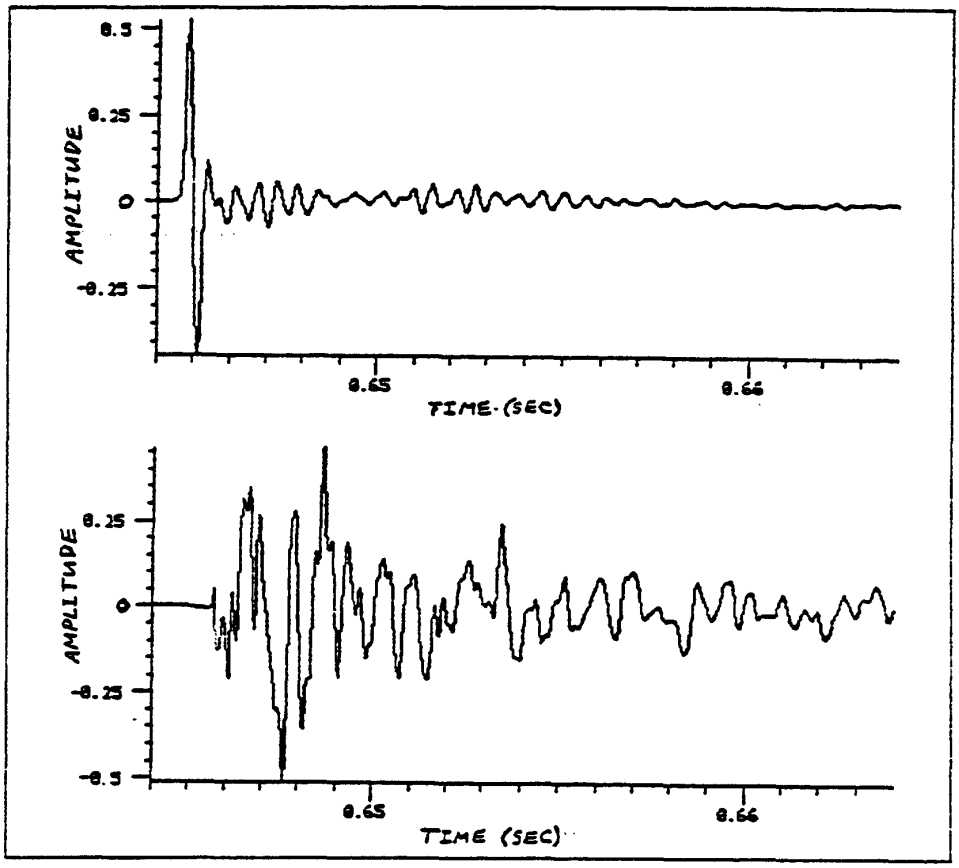

Figure 3.6. Time-domain input-output pair for the case of a plucked damped violin G-string.

a) Force derivative at the bridge.

b) Sound pressure from the body.

Figure 3.6 shows an input-output pair for the case of a plucked, damped string. Thick cloth was wrapped around all the strings in order to eliminate their contribution to the response, with about two inches exposed near the bridge. The G-string was plucked with a heavy guitar pick, near the bridge, in a regular up-and-down fashion at the rate of about one per second. Thus we recorded several "up-picks" and several "down-picks" in alternation. The recording of several samples was useful for determining repeatability and linearity of the entire process, which was found to be fairly good. The spectra of the input and output signals are shown in Fig. 3.7. Also shown is the noise floor, obtained by Fourier transforming the window of silence immediately preceding the pluck for both input and output.

Note the significant time ($\approx$ 3 msec) it takes for the body output to develop full amplitude. Apparently, the sound must propagate a few feet within the body before quasi-steady-state is established. Consequently, the violin body is somewhat far from being a minimum-phase filter. Minimum-phase models, such as will be constructed shortly, may give a slightly "harsh" sound even when the amplitude response is accurately modeled. However, minimum-phase models are (1) much easier to compute, (2) much less expensive to implement, and (3) usually adequate in reverberant and/or ensemble contexts. One place, however, where the finite rise-time of the body response may be quite important, is in the perception of attack in vigorous bowing styles (e.g. martelé [244]). We should keep in mind that a minimum-phase body-model impulse-response needs a more gradual attack envelope.

The input spectrum (in dB) is subtracted from the output spectrum to yield the frequency-response estimate shown in Fig. 3.8a. One can see the very dense structure in the frequency-response, and the presence of many deep nulls. It should be noted that detailed behavior of this frequency-response varies considerably from recording to recording, especially above 5 KHz, although the general envelope does not. By subtracting the shown frequency response from that obtained from a different recording, it was found that the difference is generally flat up to 5 KHz, after which it begins to look noise-like with standard deviations steadily rising up to a level of around 12 dB at 20 KHz. This is thought to be due in part to the relatively weak excitation energy at high frequencies, and to the movement of spectral nulls associated with the exact positioning of the instrument with respect to the microphone. (Half a cycle at 5 KHz corresponds to about one inch of sound propagation in air.)

Next, the sampling rate is reduced by the factor 2/5 to yield $f_s = 17.857$ KHz, and the resulting frequency response is shown in Fig. 3.8b. Sampling-rate conversion was done for several reasons. First, it is thought that the frequency-response detail above 8 KHz is not very critical, with the main requirement being to provide the correct roll-off so that brightness is unchanged. Second, the measurements at high-frequency are not very repeatable experimentally, and so their reliability is suspect. Third, 22 KHz is a lot of bandwidth to fit with a digital filter, especially when it is desired to obtain a close match at low frequencies. (The main air and wood resonances of the violin body are both below 500 Hz.) Fourth, the conformal mapping technique of Chapter 1 (§1.9.1) can be used to rescale the final filter to high sampling rates.

3.8. Pre-Processed Violin Data

The effect of smoothing the measured frequency response according to critical bands is shown in Fig. 3.9a. Next, the optimum first-order conformal map is applied to warp the

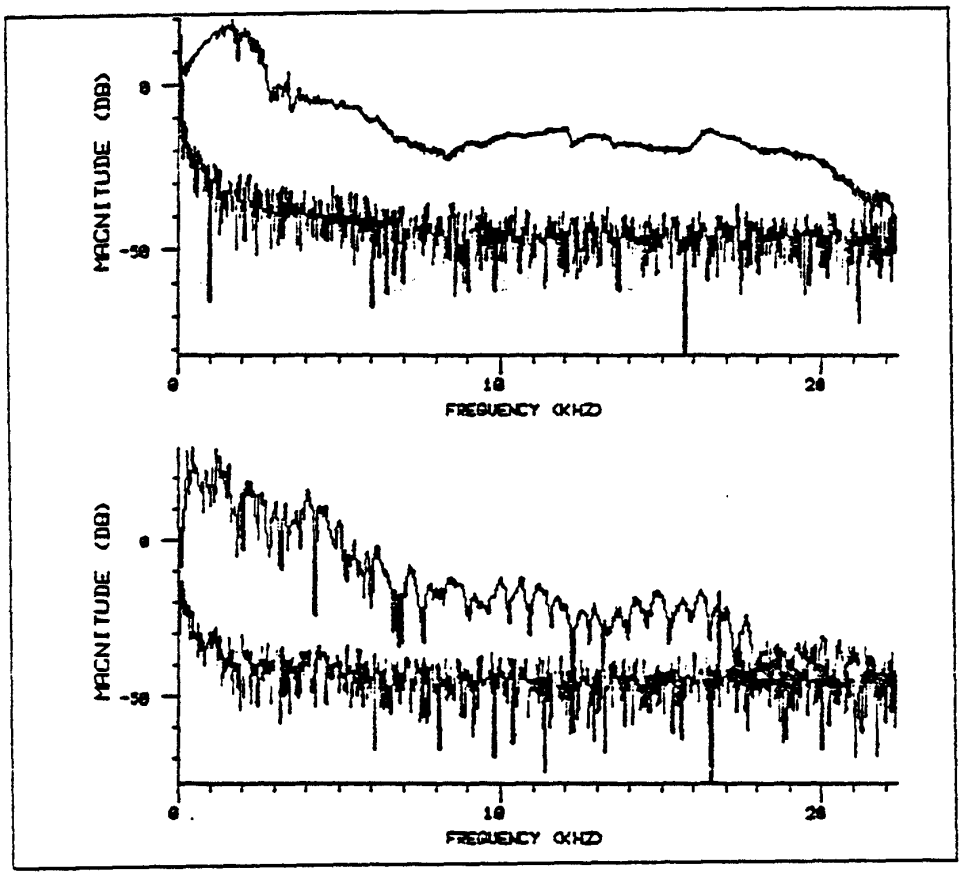

Figure 3.7. Frequency-domain input-output pair for the case of a plucked damped violin G-string.

a) Force derivative at the bridge overlayed with noise floor.

b) Sound pressure from the body overlayed with noise floor.

frequency axis as shown in Fig. 3.9b. This is the desired frequency-response to which a digital filter will be fit using a variety of spectral modeling methods.

Note that the main wood and air resonances (below 500 Hz), though well resolved, have been "rounded" by the smoothing according to critical bands. A different smoothing was tried in which frequencies below 500 Hz were not smoothed at all, and the smoothing above 500 Hz was the same as in the figure. The frequency responses were compared aurally using the experiment described in §3.7.7. The difference was almost imperceptible. However, when the fundamental passed through one of the two major resonant frequencies, a slightly greater "reverberant swell" was noticeable in the case with no low-frequency smoothing. The difference was considered musically unimportant.

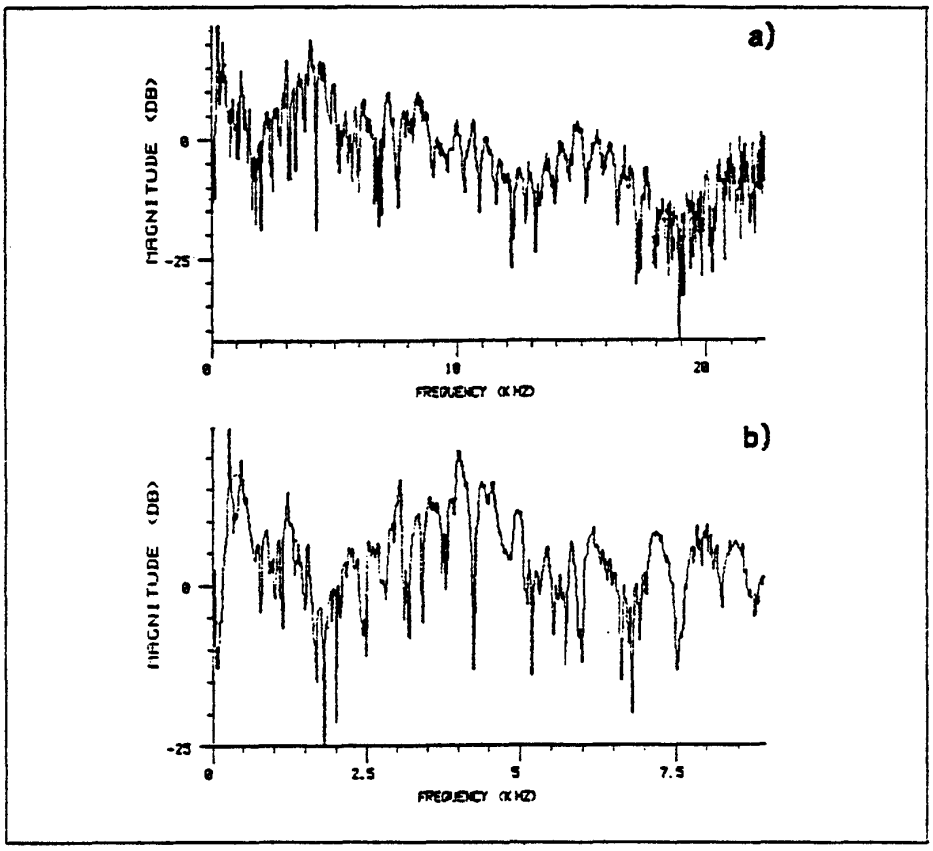

Figure 3.8. Frequency-response measurement for the case of a plucked damped violin G-string.

a) Result for 44.642 KHz sampling rate.

b) Result after resampling to 17.857 KHz.

Generally, the smoothing is most reasonable at high frequencies where there are many densely-spaced peaks and valleys in the frequency response. The assumption is that the individual amplitudes are not important in an ensemble of partials falling within a critical bandwidth. On the other hand, the dense modulation of the high-frequency power-response provides independent temporal amplitude modulation of partial amplitudes when vibrato is present. While this leads outside of the steady-state point of view, it is a source of richness of sound which one may wish to recover in some other way. Since reverberation is often described in terms of the statistics of the peak/valley distribution, it may be that a simple reverberator with a similar spectral variance can recapture the perceptually significant features of an irregular high-frequency power response. In any case, it is not feasible at

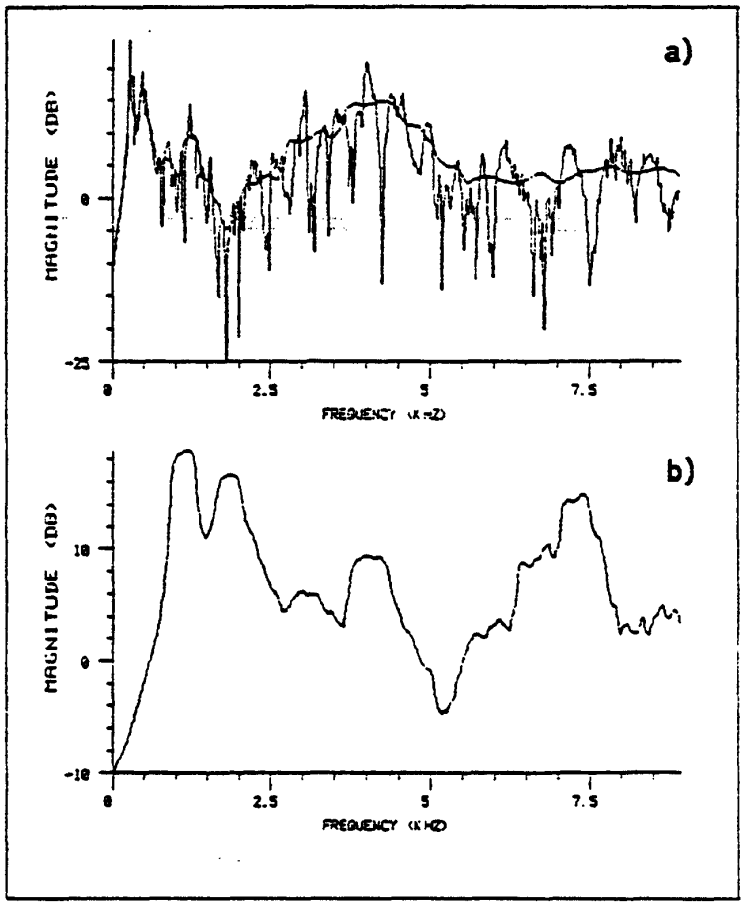

Figure 3.9. Illustration of frequency-response pre-processing.
a) Overlay of orignal data with the critical-band smoothed data.
b) Smoothed data after frequency-warping.

present to place two poles at each peak of the frequency response, for filter orders well into the hundreds would ensue for the violin body at quality sampling rates.

3.7. Performance of Various Modeling Methods

The task now is to find a digital filter of low complexity which has nearly the same frequency-response as the pre-processed curve of Fig. 3.9b. While a priori considerations could narrow the choice of method considerably, a wide range of methods will be applied, primarily to illustrate their characteristics on a common problem. (This section serves as

a "computed examples" section for Chapter 1.) Various frequency-response error norms, shown for each method, are defined in Appendix D. To fix the comparisons with respect to order, only 8 poles and 8 zeros will be allowed for each filter-design method. Ordinarily, one would specify instead an upper bound of 17 degrees of freedom (the sum of the numbers of poles and zeros plus a gain-matching scale-factor). This is not appropriate here because the conformal mapping will equalize the number poles and zeros; for example, a 16-pole filter with no zeros becomes a 16-pole, 16-zero filter when mapped back to the normal frequency axis.

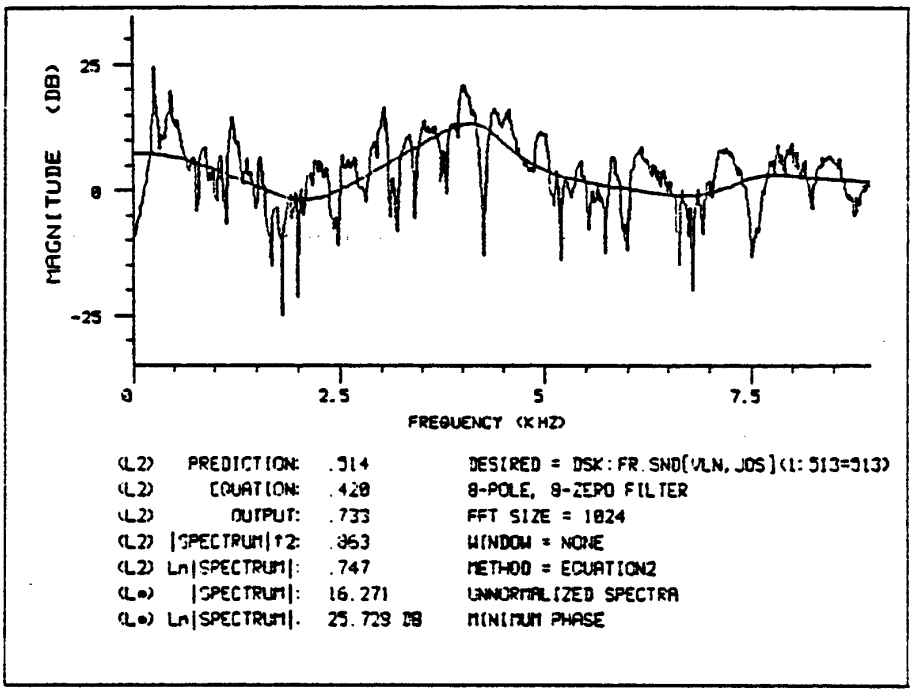

Figure 3.10. Frequency-response fit using RLS

3.7.1. A System-Identification Model

The use of time-domain system-identification methods may seem ideal because we have input-output data available. It turns out, however, that the pre-processing defined for the frequency-domain methods is very much necessary for a reasonable fit using 8 poles and 8 zeros. Figure 3.10 shows the results of Recursive Least Squares (RLS) on the input-output data used to obtain the desired frequency response. The figure lists the error norms defined in Appendix D. The input/output signals were modified so that the desired frequency response is minimum phase (with the desired amplitude response left unaltered). This is typically unavoidable when minimizing equation error, as discussed in Chapter 1, §1.7.1. As is evident, there is no resolution of the important main air and wood resonances (below 500 Hz), and the fit elsewhere is crude. Note that RLS is equivalent to the fast frequency-domain equation-error method described in Chapter 1; results for the frequency-domain version, including the pre-processing, will be given shortly.

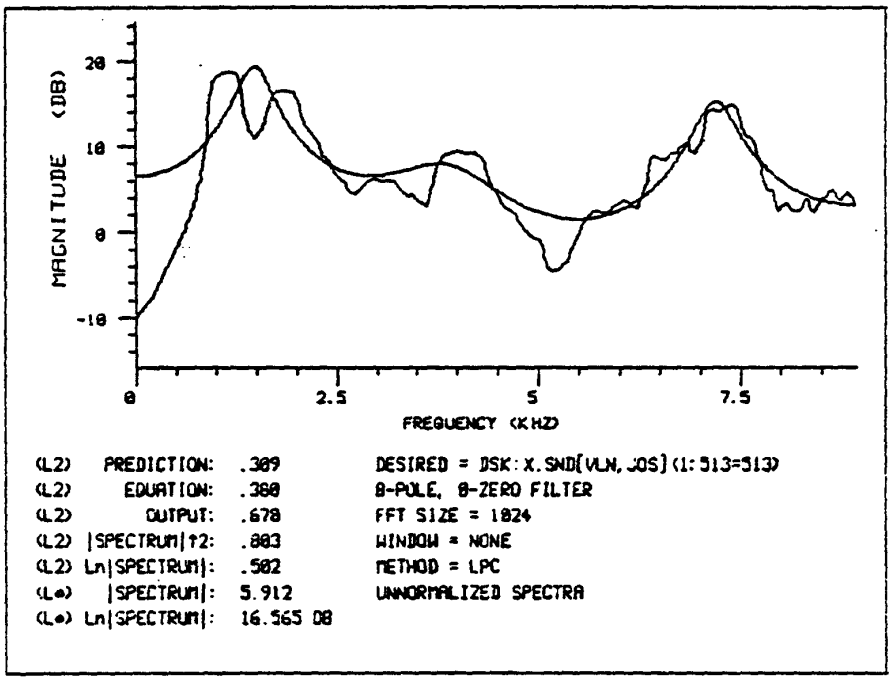

Figure 3.11. Frequency-response fit using linear prediction.

3.7.2. Linear Prediction

One of the most popular spectral modeling techniques is that of linear prediction. The two primary methods of linear prediction are the covariance method and the autocorrelation method. In the present circumstance, in which a desired power response is posed, the two methods are equivalent. Thus, the squared amplitude response is inverse-Fourier-transformed to provide the autocorrelation of the impulse response, and this is converted to an 8-pole filter by the Durbin recursion [188]. The frequency-response fit is shown in Fig. 3.11. Note that linear prediction techniques provide a minimum-phase model, and the phase of the desired spectrum is discarded in forming the autocorrelation function. One can see that the main air and wood resonances are still not resolved.

As an aside, the 16-pole linear-prediction fit to the *unsmoothed* (but warped) desired frequency response is shown in Fig. 3.12. Linear prediction methods characteristically yield a model of the *spectrum envelope*, as this figure clearly shows. Since frequency-warping was done, this is really twice the complexity of the other models presented. In practice, one may wish to implement the inverse conformal map $(\rho + z^{-1})/(1 + \rho z^{-1})$ in place of each

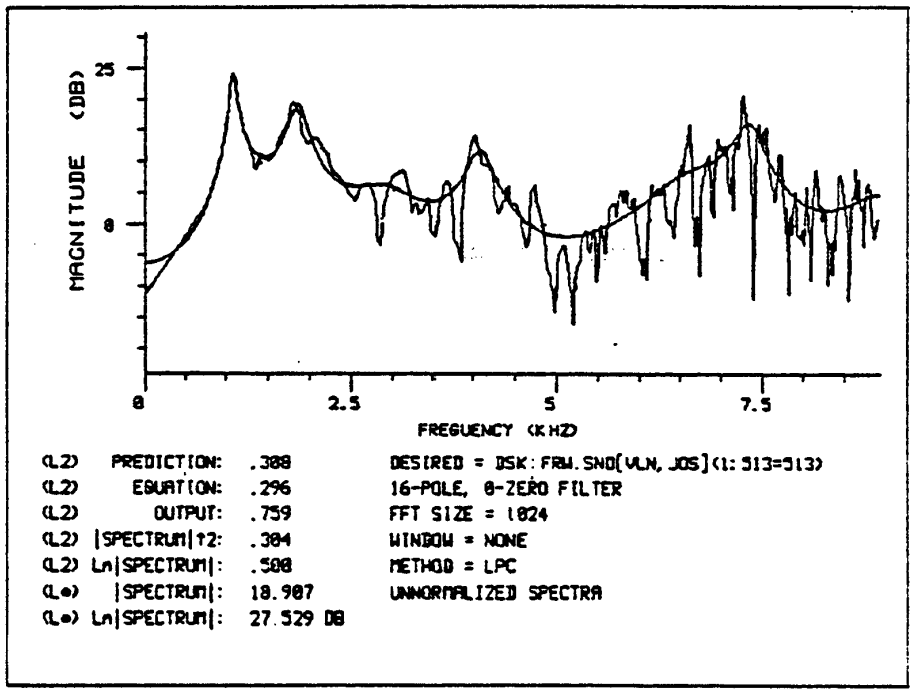

Figure 3.12. Frequency-response fit using 16-pole linear prediction on the unsmoothed desired frequency-response.

unit-sample delay in the mapped filter structure. This would allow real-time frequency scaling, for example. In such a situation, the 16-pole linear-prediction filter would be a good choice.

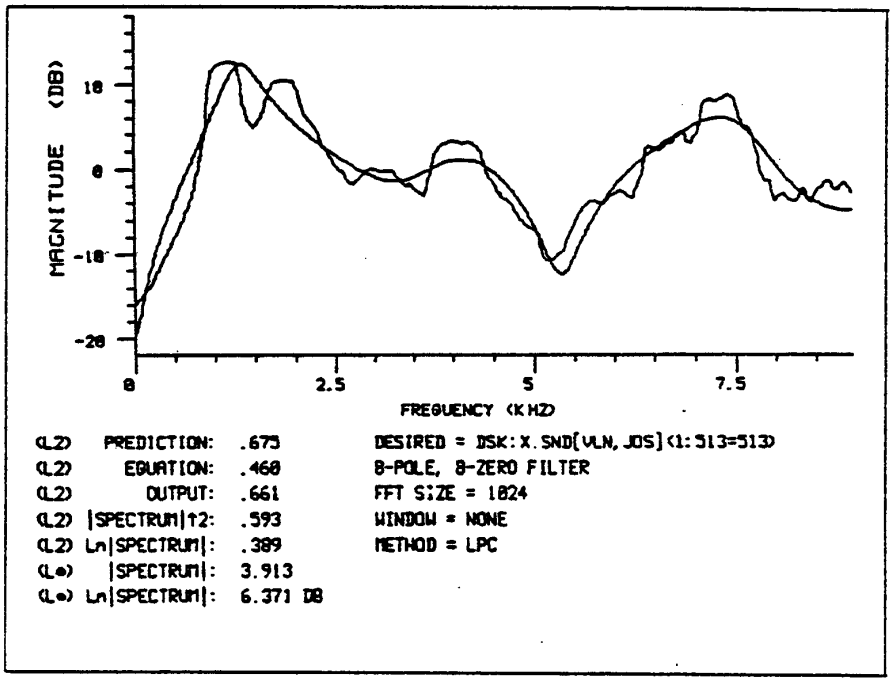

Figure 3.13. Frequency-response fit using Kopec's Method.

3.7.3. Kopec's Method

Kopec's method is a variation on linear prediction which allows zeros as well as poles in the model [186]. First, an 8-pole fit is obtained by the linear prediction method. Then the error spectrum is inverted and an 8-pole fit is obtained to this, giving the numerator of the model transfer function. The results are shown in Fig. 3.13. The spectra have been normalized to have 0dB mean since the natural scaling places the approximation along the lower spectral envelope. If zeros are estimated before the poles, then the natural scaling gives an approximate filter which follows the upper spectral envelope. The model is minimum-phase because it is a ratio of two stable allpole filters obtained by linear prediction.

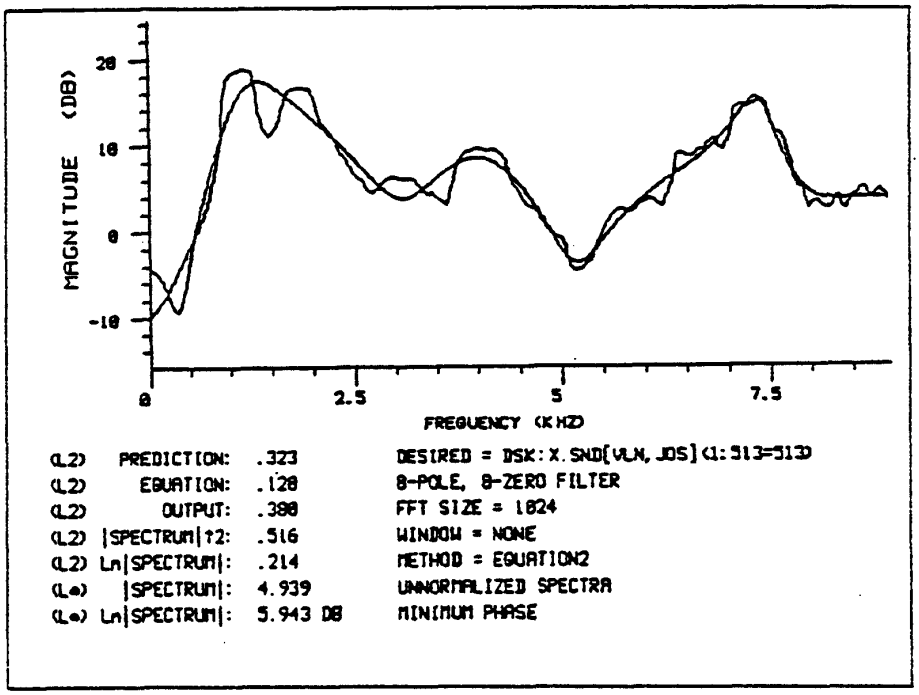

Figure 3.14. Frequency-response fit by minimizing equation error.

3.7.4. L^2 Equation-Error Minimization

Figure 3.14 gives the results of applying the fast frequency-domain equation-error method given in chapter 1 (§1.7.1). The results are almost identical to those obtained by Shank's method and the Padé-Prony method (also discussed in Chapter 1), and to save space they are not shown. For these three methods, it was necessary to prepare a *minimum phase* impulse response corresponding to the pre-processed desired frequency response, and the cepstral method was used for this [169].

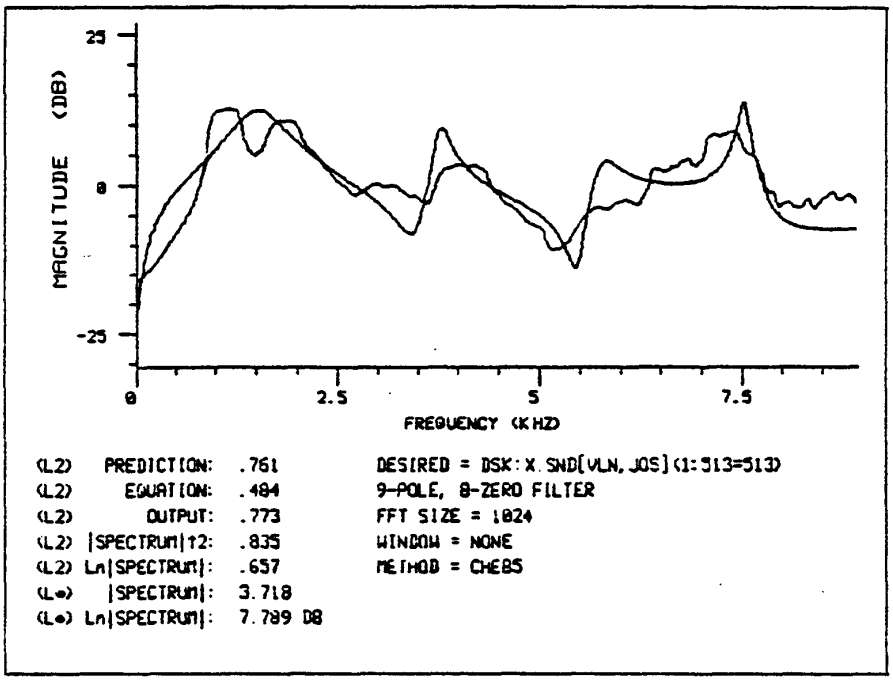

Figure 3.15. Frequency-response fit by the log-magnitude method.

3.7.5. Log-Magnitude Spectrum Matching

Figure 3.15 gives the results of applying the approximate log-magnitude matching method presented in chapter 1 (§1.8.5). Since Chebyshev approximation was implemented in the polynomial case only, the idea behind Kopec's method was used to obtain a rational filter. First, 8 zeros were fit, then the error spectrum was inverted and fit with 9 zeros (giving the poles of the model). Nine zeros were used in the second step instead of eight because because the error after fitting 8 poles was nearly equiripple. (On a log scale this error is merely reversed in sign and used as a desired function for the second step.) As a result, the second eighth-order fit is nearly zero. (In general, the optimum nth order Chebyshev approximation to the error of an nth order Chebyshev approximation is identically zero.) The error for a 16-pole fit is shown in Fig. 3.16. One can see the nearly equal-ripple error on a dB scale. Since Chebyshev approximation is feasible in the rational case, better results are to be expected when a true optimum rational approximation is found.

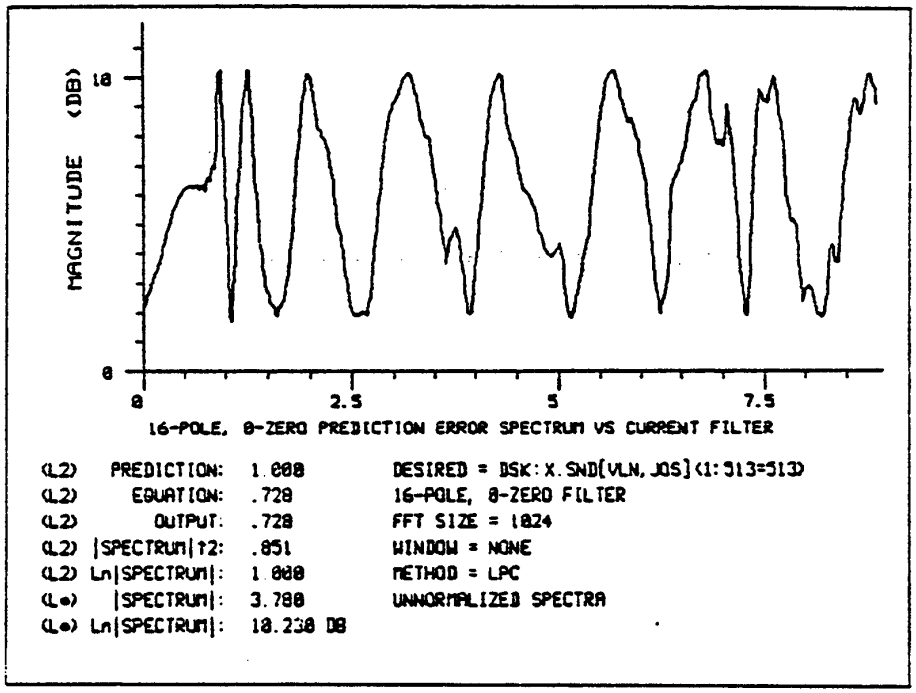

Figure 3.18. Frequency-response error for the log-magnitude method.

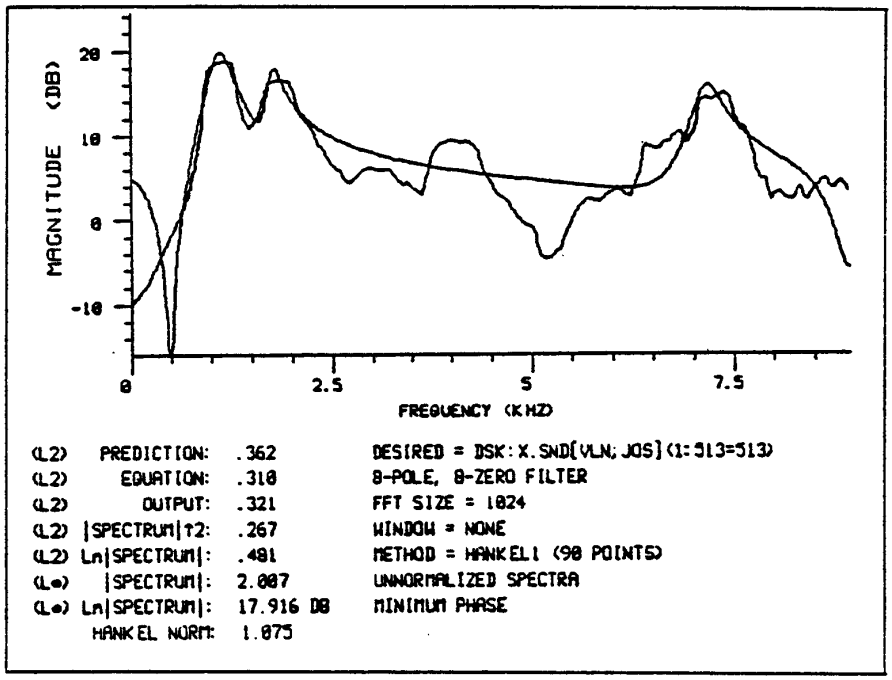

Figure 3.17. Frequency-response fit by the Hankel-norm method.

3.7.6. Hankel-Norm Minimization

Figure 3.17 gives the results of applying the Hankel norm method presented in chapter 1 (§1.8.5). This is the only eighth-order case in which good resolution of the main air and wood resonances is obtained. The small resonance near the center, on the other hand, is pretty much ignored. This is understandable in light of the appearance of the fit on a *linear* magnitude scale, shown in Fig. 3.18. Optimum Hankel-norm approximations tend to be approximate Chebyshev approximations, and on a linear scale, one can see the nearly equal-ripple error in the amplitude response. Also, the flattening of the resonances due to smoothing has been compensated by the nature of the fit.

It is unfortunate that no method is known for introducing an arbitrary weight function for Hankel-norm methods, for then it would be possible to extend the approximate log-spectral matching idea to the case of Hankel-norm approximation.

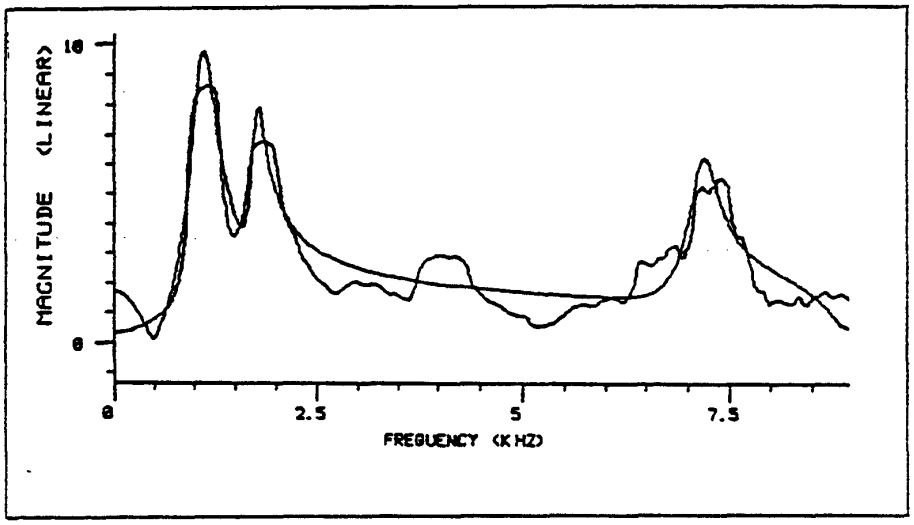

Figure 3.18. Frequency-response fit by the Hankel-norm method on a linear vertical scale.

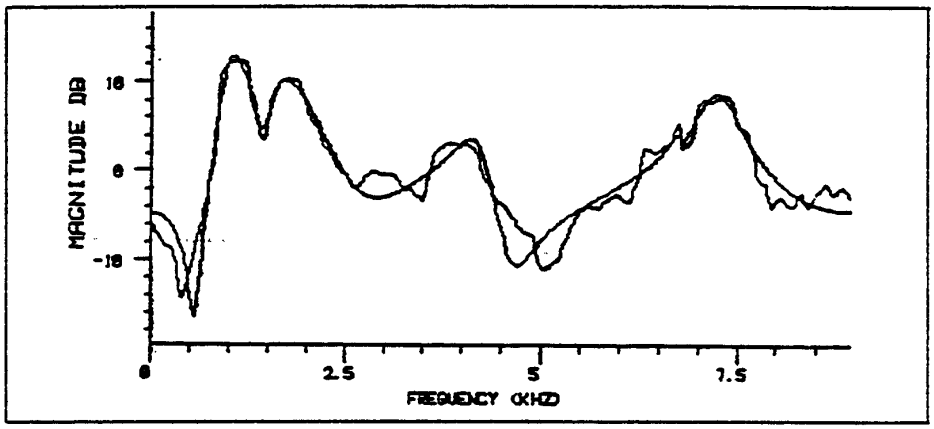

Figure 3.19. Frequency response of final violin body model in mapped form.

3.7.7. Conclusions Regarding the Body Model

Several of the filter-design methods from Chapter 1 have been applied to the pre-processed frequency-response data for the violin body. By limiting the comparisons to a predetermined set of conditions, neither the filter-design methods nor the violin body received full consideration. The purpose of this was to provide a comparison of the filter-design methods. The set of conditions was chosen, however, to be close to those used in obtaining the filter which was selected for use in the final violin model. The final choice was an eighth-order Hankel-norm design using somewhat more frequency warping than provided by the Bark-scale case. This filter is shown in Fig. 3.19. Although the comparisons are not presented here, various orders and smoothing strategies were tried and compared in arriving at the final filter.

A particularly illuminating test consisted of driving the body model with an impulse train which rose in pitch from 200 Hz to 400 Hz over 2 seconds time. This signal has all harmonics present in equal amounts, and the rising pitch ensures that the entire frequency response is represented in the output sound. These signals were compared in informal listening tests. The overall conclusion is that small changes in the shape of the body-filter frequency response do not affect the quality of the sound in a musically significant way. Also, the effect of critical-band smoothing on the original data did not change the first-order sound quality; it had a noticeable effect, however, on the perception of "roughness" in the high-frequency spectrum. It was decided that either the fine details of the frequency-response of the violin body are of second-order importance, or that conclusions cannot be drawn when the bowed string is replaced by a very artificial excitation, such as an impulse-train. Nonetheless, this test was the basis for the choice of such a low-order model for the

violin body. The only features retained in the model are the main air and wood resonances, the "singing formant," and the overall spectral-envelope. The comparisons should be repeated using the measured bridge excitation for naturally played notes. At present, however, the choice of 8th order filters, and the smoothing used, indicate preliminary conclusions regarding the necessary quality of fit to a measured violin frequency response when sound quality is the only consideration.

3.8. A Parametric Model for the Vibrating String

We now turn from the violin body to the string. The string is in many ways more important than the body for determining the sound of a violin. This was graphically demonstrated by Max Mathews [229] who built a pleasant sounding "violin" consisting of a bowed string on a metal mounting (non-resonating), and a *single* electronic resonance in the vicinity of 3 KHz (the "singing formant"). While it did not sound exactly like a violin in *A:B* comparisons, it did evoke the illusion of a member of the violin family.

A problem in modeling vibrating strings is that they are very high-order systems. As a rough estimate, the number of poles needed in the model is twice the number of harmonics produced by the string. At high sampling rates (say 40 KHz) and low pitches (say 100 Hz), the complete number of poles becomes very large (400!). The situation is further aggravated by the fact that these poles are very close to the unit circle. Numerical experience indicates that no method of filter design can be expected to model closely such a large-order, lightly-damped system on typical computer systems. Besides, who would want to? A major goal of modeling is to capture the useful essence of the physical phenomenon in a computationally efficient form.

This section derives a parametric model for the vibrating string which provides most musically important features of real strings at much-reduced complexity. This model will be then applied to the case of bowed strings. The bowed-string simulator consists of a linear digital filter which models the vibrating string, together with a means of exciting the string like a bow. Two approaches to driving the string model are considered. The first is similar to that used in linear predictive speech coding, and consists only of a series of impulses which convey pitch and amplitude information. A more elaborate method yields a much improved model of bow-string interaction; specifically, the input variables are the pressure, velocity, and position of the bow on the string.

3.8.1. The Wave Equation for an Ideal String

The wave equation and its solution in terms of "traveling waves" were derived by d'Alembert in 1747 [139]. Eight years later, Daniel Bernoulli demonstrated the solution

in terms of "standing waves" [139]. Bernoulli claimed that any displacement of the string could be expressed as an infinite sum of sinusoidal harmonics—or standing waves. Euler objected to this claim on the grounds that a sum of sinusoids could not express an arbitrary curve. In this way, a controversy was born which led to Fourier theory and the foundations of modern analysis.

For an ideal string, we have the following *wave equation*.

$$w_{tt}(x,t) = c^2 w_{xx}(x,t), \tag{3.3}$$

where $w(x,t)$ denotes the transverse displacement of the string at the point x along the string at time t in seconds. If the length of the string is L, then x is taken to lie between 0 and L. The partial derivative notation used above is defined by

$$w_{uv} \triangleq \frac{\partial}{\partial u}\left(\frac{\partial w}{\partial v}\right).$$

The constant c is given by $c^2 = T/\rho$ where T is the string tension, and ρ is the mass per unit length of the string. An elegant derivation of the wave equation is given by Morse [257]. The wave equation implies that *the transverse acceleration of a point on the string (w_{tt}) is proportional to the curvature* of the string at that point (w_{xx}).*

The general *traveling-wave* solution to (3.3) is given by

$$w(x,t) = \varphi(x - ct) + \psi(x + ct). \tag{3.4}$$

This solution form is interpreted as the sum of two fixed wave-shapes traveling in opposite directions along the string. The specific waveshapes are determined by the initial shape $w(x,0)$ and the initial velocity $w_t(x,0)$ of the whole string.

Given any solution of the form (3.4), we may successively differentiate the terms of the equation to obtain a solution in terms of traveling velocity waves, acceleration waves, etc. It turns out that for bowed and plucked strings, *acceleration waves* are a convenient choice. This is because a pluck gives rise to a pair of acceleration (or curvature) impulses propagating outward from the pluck-point. A bowed string may be described, to the first order, as a periodically plucked string in which one of these two curvature impulses is eliminated.

If the string is rigidly fixed to $w = 0$ at $x = 0$, say, then in equation (3.3) the constraint $w(0,t) = 0$ holds for all t. The traveling displacement waves of (3.4) must then satisfy

$$\varphi(-ct) = -\psi(ct), \tag{3.5}$$

* Actually, curvature is defined as $\pm w_{xx}/(1 + w_x^2)^{3/2}$ [157]. However, when the maximum slope on the string is much less than unity ($|w_x| \ll 1$), the curvature and the second derivative in x are approximately equal. We will refer to w_{xx} as the curvature mainly for descriptive convenience.

If $\varphi(\cdot)$ is chosen arbitrarily, then $\psi(\cdot)$ must be the same waveform flipped about the horizontal and vertical axes. An equivalent point of view is that ψ is the reflection of φ from the termination. At a rigid termination, an arbitrary incident wave is reflected such that it emerges negated and time-reversed. A curvature impulse merely changes sign upon reflection from a rigid termination since it has negligible width.

3.8.2. A Description of One-Dimensional Propagating Waves

An analytical approach will be described which is well-suited to modeling vibrating strings as *linear filters*. We desire an *input-output* representation of the string, because we wish to drive the string with a simple function, such as an impulse, to produce natural waveforms.

Assume that the string is driven at the point x_i on the string by the acceleration function $u(t)$. Let the output acceleration, observed at the point x_o, be denoted $y(t)$. Then we have $u(t) \equiv w_{tt}(x_i, t)$ and $y(t) \equiv w_{tt}(x_o, t)$.

Let $u(t)$ be an acceleration impulse (defined in Appendix E) with amplitude A at time $t = 0$. This is expressed by writing

$$u(t) = A\delta(t).$$

We assume for the moment that the string is *ideal* which means the curvature impulse does not attenuate or "spread" as it propagates along the string. If the observation point x_o is located a distance $d = |x_i - x_o|$ from the point of input, in the direction of travel for the impulse, then at time $t = d/c$ there will be an output impulse. This output is written as

$$y(t) = A\delta\left(t - \tilde{d}\right),$$

where $\delta(t - \tilde{d})$ denotes an impulse occurring at time $\tilde{d}$, and $\tilde{d} \triangleq d/c$. In this chapter, a tilda placed above a distance value denotes the time it takes to travel that distance at speed c. In the frequency domain, the *Laplace transform* of the first output impulse is

$$Y(s) \triangleq \int_{-\infty}^{\infty} y(t)e^{-st}dt = Ae^{-s\tilde{d}}.$$

For discrete time, simply set $t = nT$ where T is the sampling period, and define the x positions along the string to be at integer multiples of cT, the distance traversed during one sample period. If $d = mcT$, then

$$y(nT) = A\delta(nT - mT).$$

The frequency domain representation is given in the discrete-time case by the *z-transform* of the time-domain expression,

$$Y(z) \triangleq \sum_{n=-\infty}^{\infty} y(nT) z^{-nT} = Az^{-d} = Az^{-mT}.$$

The discrete-time and continuous-time representations are related in the frequency domain by the substitution $z = e^{sT}$. When working with discrete time, it is convenient to use the definitions

$$y(n) = A\delta(n-m), \qquad Y(z) = Az^{-m},$$

where the sampling interval has been normalized to 1. In this situation, the substitution $z = e^{sT}$ converts properly to the analog frequency domain. In particular, the *spectrum* is obtained by setting $z = e^{j\omega T}$ in the z-transform, where $\omega = 2\pi f$, and f is frequency in Hz.

Let $x(n)$ be an arbitrary waveform having the z-transform $X(z)$. The notation $x(n) \leftrightarrow X(z)$ means that $x(n)$ and $X(z)$ are transform pairs. By the *shift theorem* for z-transforms [164],

$$x(n) \leftrightarrow X(z) \quad \Rightarrow \quad x(n-k) \leftrightarrow z^{-k}X(z).$$

This implies that the term z^{-k} may be interpreted as a *delay operator* which delays a signal by k time samples.

In deriving the string model, the following *delay-operator* notation will be used.

$$z^{-\tau}x(t) \triangleq x(t-\tau).$$

Thus time is in seconds and the unnormalized notation for the z-transform is used. This notation facilitates writing down the string transfer function by inspection in simple cases. In later sections, where discrete-time effects are of interest, the sampling rate will be normalized to unity so that the time indices will be integers.

3.8.3. Non-Rigid Terminations and Distributed Losses

Non-rigid terminations and lossy, stiff strings are relatively difficult to analyze when the wave equation is used as a starting point [257]. Therefore, we take a simpler approach involving some approximations which are reasonable for audio applications. First, *losses and phase-dispersion due to string imperfections are absorbed into termination losses.* This causes little or no change in the perceived output signal. For example, if the string is exponentially damped, this can be provided by simple scaling by $\rho < 1$ at a termination point; this replaces a continuously decaying exponential envelope by an exponential "staircase"

which steps down once per period. Similarly, if the string is dispersive, a variable delay (as a function of frequency) may be inserted at the string termination to alter the round-trip travel time as though the speed of propagation on the string were being changed as a function of frequency. The second approximation is that *a yielding termination is represented as a linear time-invariant filter of low order.* This termination filter includes all absorbed string imperfections. In general, as long as the transformation from one period to the next can be well approximated by a low-order filtering operation, this approach can be effective. Since the ear is not very sensitive to a rearrangement of phase in a quasi-periodic signal, lumping the losses and dispersiveness of a string into the terminations does not have serious audible consequences. This is the key to economy in the model for vibrating strings to be discussed.

If the termination loss is purely "resistive", i.e. nondispersive, then an incident waveform $\varphi(ct)$ will produce a reflection expressible at time t (and $x = 0$) as $-\rho\varphi(-ct)$. A curvature impulse goes into the termination as $\delta(t)$ and emerges as $-\rho\delta(t)$. The general case of a linear lossy termination is obtained by replacing ρ with an arbitrary integro-differential operator in continuous time, or delay-operator expression in discrete time. The "reflection" of a curvature impulse becomes the impulse response of the termination-filter. Further reflections produce multiple convolutions of the termination impulse response. In general, the amplitude response of the termination controls string damping as a function of frequency, and the phase delay of the termination adds to the effective string length as a function of frequency. Thus the ideal string can be perturbed in a parsimonious fashion to provide harmonics which decay at different rates and/or overtones which are not harmonically related to the fundamental.

We will now find a general expression representing the effects of filtering due to yielding terminations and string imperfections. For a doubly-terminated, ideal string of length L, the period of oscillation is $P = 2L/c$. When a filtering operation is introduced on the string, the signal is no longer periodic, but we assume that this filtering is so "mild" that the signal is close to periodic. Let the total round-trip filtering be denoted by $H_l(z)$. Then each "period" is the previous period filtered by $H_l(z)$. If $P(z)$ equals the Fourier transform of the first period, then the next period has the Fourier transform $z^{-P}H_l(z)P(z)$. Thus the output transform for all time is given by

$$\begin{aligned} Y(z) &= P(z) + z^{-P}H_l(z)P(z) + z^{-2P}H_l^2(z)P(z) + \cdots \\ &= \frac{P(z)}{1 - z^{-P}H_l(z)}. \end{aligned} \tag{3.6}$$

3.8.4. Transfer Function for the Non-Ideal String

Using the above concepts, a transfer function for the doubly-terminated linear string will be derived which includes frequency-dependent damping and inharmonic overtones.

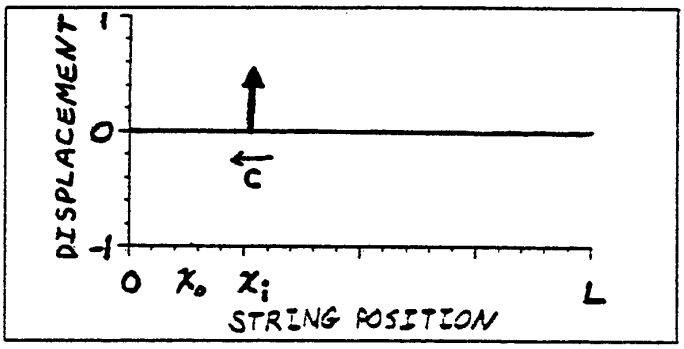

Figure 3.20. Initial acceleration (or curvature) impulse traveling to the left on the string. The point of input is x_i, and the output is observed at position x_o.

For convenience, the "left" termination (at $x = 0$) will be called the *bridge*, and the "right" termination (at $x = L$) will be called the *nut*. The nut and bridge reflection transfer functions will be denoted $H_n(z)$ and $H_b(z)$, respectively.

Suppose that the string is initialized with a solitary acceleration impulse of amplitude A traveling, say, toward the bridge ($x = 0$) from the point $x = x_i$ at time $t = 0$ as shown in Fig. 3.20. Then $u(t) = A\delta(t)$. Let the observation point on the string be at $x = x_o$, where for definiteness x_o is assumed to lie to the left of x_i (*i.e.* $x_o < x_i$). As before, we define $\tilde{x} = x/c$ for an arbitrary distance x, in order that distance be normalized to correspond to time in seconds.

The first output impulse occurs when the impulse has traveled from x_i to x_o. This will be at time $t = (x_i - x_o)/c = \tilde{x}_i - \tilde{x}_o$. Thus, the first contribution to the output is the term $Az^{-(\tilde{x}_i - \tilde{x}_o)}\delta(t)$. The impulse continues left until it reaches the bridge, where it is filtered and reflected (a change of sign is typically included in $H_b(z)$). Next it propagates to the right with "amplitude" $AH_b(z)$, and when it reaches the point x_o, the term $AH_b(z)z^{-(\tilde{x}_i - \tilde{x}_o)}z^{-2\tilde{x}_o}\delta(t) = AH_b(z)z^{-(\tilde{x}_i + \tilde{x}_o)}\delta(t)$ appears at the output. The pulse continues to the right, reflects at the nut, and returns to x_i at time $P = 2L/c$ with amplitude $AH_b(z)H_n(z)$. Thus the first period of the output waveform is

$$y(t) = Az^{-(\tilde{x}_i - \tilde{x}_o)}\delta(t) + AH_b(z)z^{-(\tilde{x}_i + \tilde{x}_o)}\delta(t), \qquad 0 \le t \le P.$$

For the remaining periods, we invoke (3.6) to obtain

$$y(t) = \frac{z^{-(\tilde{x}_i - \tilde{x}_o)} + H_b(z)z^{-(\tilde{x}_i + \tilde{x}_o)}}{1 - z^{-P}H_b(z)H_n(z)}A\delta(t), \qquad t \ge 0.$$

The transfer function of the string is therefore

$$H_s^-(x_i, x_o, z) \triangleq \frac{Y(z)}{U(z)} = \frac{z^{-(\tilde{x}_i - \tilde{x}_o)} + H_b(z) z^{-(\tilde{x}_i + \tilde{x}_o)}}{1 - z^{-P} H_b(z) H_n(z)}. \tag{3.7}$$

An analogous derivation yields the string transfer function for the case of an acceleration impulse traveling initially toward the nut, again for the case $x_o < x_i$.

$$H_s^+(x_i, x_o, z) \triangleq H_n(z) z^{-P} \frac{z^{\tilde{x}_i + \tilde{x}_o} + H_b(z) z^{\tilde{x}_i - \tilde{x}_o}}{1 - z^{-P} H_b(z) H_n(z)}. \tag{3.8}$$

If an impulse is input to the string with an initial velocity of propagation other than $\pm c$, then the transfer function becomes a linear combination of $H_s^+(x_i, x_o, z)$ and $H_s^-(x_i, x_o, z)$. Thus the general transfer function is given by

$$H_s(x_i, x_o, \sigma, z) \triangleq \gamma H_s^+(x_i, x_o, z) + (1 - \gamma) H_s^-(x_i, x_o, z), \tag{3.9}$$

where γ is determined by the initial position and velocity of the input impulse. It will be shown that for the plucked string, observed between the bridge and the pick, we set

$$\gamma = \frac{1}{2}, \qquad \text{(Plucked String)},$$

and for the steady-state Helmholtz description of the bowed string, observed between the bridge and the bow, we set

$$\gamma = \begin{cases} 0, & \text{(Down-Bow)} \\ 1, & \text{(Up-Bow)} \end{cases}.$$

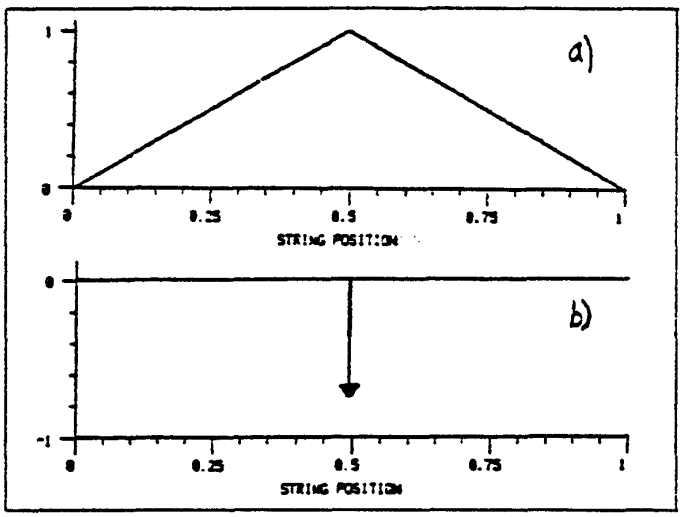

Figure 3.21.
a) Initial string displacement for the ideal plucked string.
b) Corresponding initial acceleration distribution.

3.9. Application to Bowed Strings

In this section, the description of bowed-string motion due to Helmholtz will be used to derive a model for bowed strings which uses the string model of the previous section along with an impulse-train excitation. The ideal plucked string is treated also, because it turns out to be simply the superposition of a "down-bow" and an "up-bow" of Helmholtz bowed-string waveforms.

3.9.1. The Plucked String

A diagram for the ideal plucked string is shown in Fig. 3.21a. Prior to time 0, the string is pulled away from equilibrium by a point force at $x = x_i$, to give two straight string segments of equal tension meeting at a corner. Let D denote the maximum initial displacement of the string, $D = w(x_i, 0)$. The initial velocity of the string is assumed zero.

Twice differentiating the initial string shape with respect to x gives the initial curvature

distribution,*

$$w_{xx}(x,0) = -\frac{DL}{x_i(L-x_i)}\delta(x-x_i).$$

From the wave equation (3.3), the equivalent acceleration input is

$$u(t) \triangleq w_{tt}(x_i,0)\delta(t) = -\frac{DLc^2}{x_i(L-x_i)}\delta(t) \triangleq 2A\delta(t).$$

This initial acceleration is shown schematically in Fig. 3.21b.

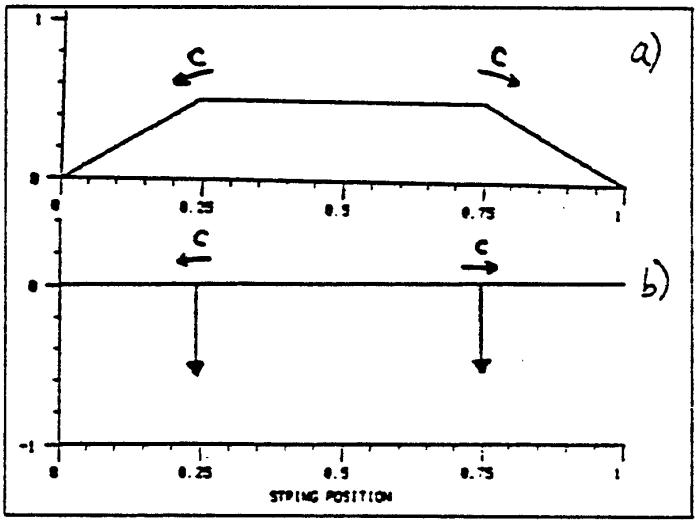

Figure 3.22. Plucked-string displacement and corresponding acceleration distribution at time $t = P/8 = (1/8)2L/c$.

a) Displacement.

b) Acceleration.

Since the initial velocity of the string is zero, the curvature impulse is also initially at rest. Therefore, for $t > 0$, the impulse splits into a left-going and a right-going impulse, each with amplitude A. The string displacement and corresponding acceleration distribution for a time one-eighth period after $t = 0$ are shown in Fig. 3.22.

* $w_{xx}(x,0) \triangleq \lim_{h\to 0}[w_x(x+h,0) - w_x(x-h,0)]/(2h)$.

The transfer function of the string (3.9) can be used to write the output acceleration, observed at $x = x_o$, as $y(t) = H_s(x_i, x_o, 1/2, z)2A\delta(t)$, and therefore the transfer function of the plucked string is given by

$$H_s(x_i, x_o, 1/2, z) = \frac{z^{-(\tilde{x}_i - \tilde{x}_o)} + H_b(z)z^{-(\tilde{x}_i + \tilde{x}_o)} + H_n(z)z^{\tilde{x}_i + \tilde{x}_o - P} + H_l(z)z^{\tilde{x}_i - \tilde{x}_o - P}}{1 - z^{-P}H_l(z)}, \tag{3.10}$$

where $H_l(z) \triangleq H_b(z)H_n(z)$.

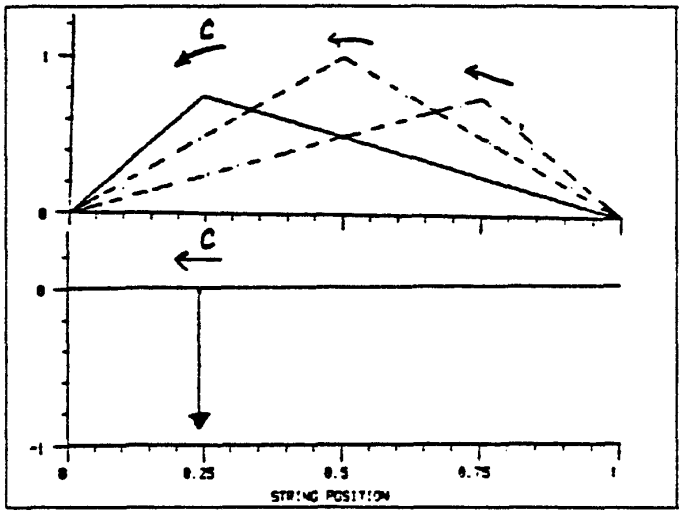

Figure 3.23.

a) String displacement for the ideal bowed string.

b) Corresponding acceleration distribution on the string.

3.9.2. The Helmholtz Bowed-String Model

The first-order model of bowed string motion derived by Helmholtz [252] is as shown in Fig. 3.23. The string consists of two straight linear segments *at all times*. The corner at the intersection of the two segments propagates along a parabolic arc with speed c. Since the linear string segments have zero curvature, the wave equation implies that the acceleration must be zero everywhere except at the corner where it is impulsive. This impulse simply shuttles back and forth on the string. In the plucked string there were two such impulses.

Thus in terms of acceleration waves, the model for the bowed string is even simpler than that of the plucked string.

Let D be the peak displacement of the string when the corner is at the midpoint $x = L/2$, and let $t = 0$ correspond to this configuration. Then the curvature distribution is given by

$$w_{xx}(x,0) = \frac{4D}{L}\delta(x - L/2),$$

which is the same as the plucked string for $x_i = L/2$. By requiring the curvature to propagate with constant magnitude, one obtains the peak displacement

$$D(x) = 4D(L/2)x(L-x)/L^2$$

at each point $x \in [0, L]$ on the string—a parabola, in agreement with Helmholtz. The acceleration input is then

$$u(t) \triangleq w_{tt}(x_i,0)\delta(t) = \frac{4Dc^2}{L}\delta(t) \triangleq A\delta(t).$$

For $t > 0$, there is *either* a left-going *or* a right-going impulse, depending on the direction of bowing. This is necessary to produce the steady-state string motion described by Helmholtz. We arbitrarily assume that the bowing is "down" and that the impulse is initially left-going.

The transfer function of the string for "down-bowing" is then

$$y(t) = H_s(x_i, x_o, 0, z) = H_s^-(x_i, x_o, z) = \frac{z^{-(\tilde{x}_i - \tilde{x}_o)} + H_b(z)z^{-(\tilde{x}_i + \tilde{x}_o)}}{1 - z^{-P}H_l(z)}. \tag{3.11}$$

The transfer function for "up-bowing" (the impulse initially traveling to the right) is found by subtracting (3.11) from the transfer function of the plucked string (3.10).

3.9.3. Bowed Strings as Periodically Plucked Strings

Helmholtz bowed-string behavior amounts to a single curvature impulse traversing the string, while a plucked string entails two such impulses. A simple physical description of the bowed string based on this idealized picture is as follows.

Imagine the string drawn from rest by the bow until the tension in the string overcomes the static friction force, and the string releases from the bow. Assume the dynamic coefficient of friction is zero. Then the string has been "plucked" and two curvature impulses propagate outward from the bowing point, just as in Fig. 3.22. Assume the bow is closer to the bridge than to the nut. Then the impulse traveling initially toward the bridge returns

first to the bow. Consider the time just after this impulse has returned and passed under the bow. Then the string is moving in the direction of the bow at a constant velocity (still behaving as an ideal plucked string). If the bow velocity happens to be the same as that of the string, then they move together as if the string were stuck to the bow. Thus, with no work whatever done by the bow (and therefore no new "pluck"), the string has returned to a "stuck" condition, where now the static coefficient of friction applies. But now when the impulse from the nut returns to the bow, it finds the string *terminated* by the bow. If this termination is absorptive, then the secondary impulse disappears. Now there is only one impulse on the string which continues on to the nut, reflects, and returns to the bow. Since the string is back to the initial displacement, the static friction cannot hold, and the string slips free once again, repeating the process.

If the string attenuates (nondispersively) the propagating curvature impulse, then the bow can compensate by providing a small pluck *once per period.* This means that the excitation function is simply an impulse train. Our first-order model for bowed strings is based on this idealization. The transfer function (3.11) is simply driven by an impulse train at the desired pitch and amplitude. Note that if the impulse train has constant amplitude (a "rectangular envelope"), then the string output has an "exponential attack" followed by a sustain, and then an "exponential release." When the impulse train ceases suddenly, the resulting effect is like lifting the bow from the string.

3.9.4. Helmholtz Crumples

Helmholtz described the main deviations from the simple mode of bowed-string operation as "crumples". By twice integrating the acceleration waveform to get transverse string displacement, one obtains the "leaning sawtooth" waveform as observed by Helmholtz. Crumples appear as little ripples in the sawtooth which correspond to those harmonics having nodes at the bowing point. Helmholtz recognized crumples as missing Fourier components, and reasoned that losses in the string can explain them. When the string waveform is considered as a steady-state natural-mode oscillation of the string, it seems clear that the bow cannot replenish energy in those modes having a node under the bow.

It is interesting to observe that these missing harmonics are the same as those in the ideal plucked string. (To see this, replace all termination filters in the numerator of (3.10) with -1 and compute the frequency response of the numerator, which reduces to $4\sin(\omega\tilde{x}_i)\sin(\omega\tilde{x}_o)$.)

Thus one can simulate the effect of crumples by interpolating between the extremes of plucked-string and bowed-string in the general transfer function (3.9). The parameter γ in $H_s(x_i, x_o, \gamma, z)$ is selected between 0 and 1/2 for a down-bow, and between 1/2 and 1 for an up-bow.

3.10. General Capabilities of the String Model

3.10.1. Frequency-Dependent Damping and Inharmonicity

Equation (3.9) gives the general transfer function for the doubly terminated string. When the string is plucked, struck, or otherwise initialized and then allowed to vibrate freely, the numerator of the transfer function can be considered as part of the initial conditions for vibration. For free vibration, the damping and spacing of the partials are governed solely by the denominator of (3.9)

$$H(z) \triangleq \frac{1}{1 - z^{-P} H_l(z)}, \tag{3.12}$$

where $H_l(z) \triangleq H_n(z)H_b(z)$ is the total round-trip "loop filter" associated with the string and its terminations (the bulk delay P being factored out).

The frequency response of the string simulator is then

$$H(e^{j\omega}) = \frac{1}{1 - e^{-j\omega P} H_l(e^{j\omega})}.$$

The loop gain is equal to

$$G_l(f) \triangleq \left| H_l(e^{j2\pi f}) \right|,$$

and the effective loop length is equal to

$$T_l(f) \triangleq P + D_l(f) \quad (seconds),$$

for each sinusoidal frequency f, where

$$D_l(f) \triangleq -\frac{\angle H_l(e^{j\omega})}{\omega}, \quad \omega = 2\pi f,$$

is the phase delay of H_l in seconds.*

Since the freely vibrating string is only quasi-periodic, it does not consist of discrete sinusoids. Essentially there are many narrow "bands" of energy decaying to zero at different rates. When these energy bands are centered at frequencies which are an integer multiple of a lowest frequency, they will be referred to as *harmonics*. When the frequency components are not necessarily uniformly spaced, the term *partial* will be used to emphasize

* For example, consider $H(z) = z^{-\tilde{x}}$. The phase delay is $-\angle H(e^{j\omega})/\omega = -\angle e^{-j\omega\tilde{x}}/\omega = \omega\tilde{x}/\omega = \tilde{x} = x/c$.

the possibilty of inharmonicity. Consider a partial at frequency f Hz circulating in the loop. On each pass through the loop, it suffers an attenuation equal to the loop amplitude response, $G_l(f)$. Since the round-trip time in the loop equals $P+D_l(f)$ seconds, the number of trips through the loop after t seconds is equal to $t/(P+D_l(f))$. Thus the *attenuation factor* at frequency f Hz and time t seconds, is given by

$$\alpha_f(t) \triangleq G_l(f)^{\frac{t}{P+D_l(f)}}, \tag{3.13}$$

For example, an initial partial amplitude A at time 0 becomes amplitude $A\alpha_f(t)$ at time t, where f is the frequency of the partial in Hz.

The *time constant* of an exponential decay is traditionally defined as the time when the amplitude has decayed to $1/e \approx 0.37$ times its initial value. The time constant at frequency f is found by equating (3.13) to e^{-t/τ_f} and solving for τ_f, which gives

$$\tau_f = \frac{-t}{\ln \alpha_f(t)} = -\frac{(P+D_l(f))}{\ln G_l(f)} \quad (seconds). \tag{3.14}$$

For audio, it is normally more useful to define the time constant of decay as the time it takes to decay -60 dB, or to 0.001 times the initial value. In this case, we equate (3.13) to 0.001 and solve for t. This value of t is often called t_{60}. Conversion from τ_f to $t_{60}(f)$ is accomplished by

$$t_{60}(f) = \ln(1000)\tau_f \approx 6.91\tau_f. \tag{3.15}$$

For example, if a sinusoid at frequency f Hz has amplitude A at time 0, then at time $t_{60}(f)$ it has amplitude $A\alpha_f(t_{60}(f)) = A/1000$, or it is 60 dB below its starting level.

The above analysis describes the attenuation due to "propagation" on the string. It does not, however, incorporate the fact that sinusoids which do not "fit" on the string are quickly destroyed by self-interference. Any signal may be fed into the string, but after the input ceases, the remaining energy quickly assumes a quasi-periodic nature. Thus, even if the string were initialized with a random shape, after a very short time the primary frequencies present are those which have an integral number of periods in $P+D_l(f)$ seconds. The lowest such frequency provides the fundamental or pitch frequency of the note, and it is defined as the minimum positive solution of

$$f_1 = \frac{1}{P+D_l(f_1)}. \tag{3.16}$$

Experience has shown that f_1 corresponds well with the perceived pitch of the freely vibrating string.

The higher partials are solutions of

$$f_k \triangleq \frac{k}{P + D_l(f_k)}, \qquad k = 1, 2, \ldots, \frac{P}{2T}, \tag{3.17}$$

and the decay factor at time t for the k^{th} partial is

$$a_k(t) = G_l(f_k)^{\frac{t}{P+D_l(f_k)}}. \tag{3.18}$$

3.10.2. Stiff Strings

An important capability of the string-loop filter $H_l(z)$ is the effective shortening of the string length at high frequencies to simulate stiffness effects. Measurements of the cello A-string suggest that stiffness is the dominant source of nonlinearity in the phase delay of the filter $H_l(z)$ [231, Fig. 2]. The theory of stiff strings [257, p. 170] indicates that stiffness creates a stretching of the partials according to the approximate formula

$$f_k \approx k f_0 \left[1 + \delta + \left(\frac{1 + k^2\pi^2}{8}\right)\delta^2\right] \triangleq k f_0 s(k), \qquad k = 1, 2, \ldots, \qquad k^2 < \frac{4}{\pi^2\delta^2},$$

where f_0 is near the fundamental frequency, and k is the partial number. The parameter δ has been called the *coefficient of inharmonicity*; if $\delta = 0$, then perfect harmonicity results. The phase delay desired for $H_l(z)$ when the string is tuned to f_1 is given by solving

$$f_k = \frac{k}{P + D_l(f_k)} = \frac{k f_1 s(k)}{s(1)} \;\Rightarrow\; D_l(f_k) = \frac{s(1)}{f_1 s(k)} - P.$$

A robust technique for designing filters with a prescribed phase delay may be obtained by a simple modification of the procedure given in [80], as described in Chapter 1, §1.8.6.

3.10.3. Passive Terminations

Another source of energy loss in the string is yielding terminations. When a string resonance is close in frequency to a body resonance, the coupled resonator effect [257] causes the effective damping and length of the string to change near the common resonant frequency [221]. This effect has been observed to alter the resonant frequency of a violin string by 10 to 30 cents* [236], an amount extending well outside the tolerances of good

* A cent is defined as one-hundredth of a semitone. The frequency one cent higher than f_0 is given by $2^{1/1200} f_0 \approx 1.0006 f_0$

intonation. In extreme cases, a phenomenon known as the *wolf note* occurs [220,215]. A wolf note manifests as beating between the normal modes of coupled oscillation between the string and body.

While the wolf note is undesirable, the pulled tuning of the various partials of the string can be considered an important determinant of the sound of the string. Therefore, we desire a realistic model of the termination filtering, especially at frequencies near strong body resonances which are coupled to the string at the bridge. Coupling can be assumed sufficiently weak that string resonances are only altered by a weak shift of resonant frequency and damping. In particular, the resonance need not split into two distinct resonances so as to produce a wolf note. This relaxes the requirements on the phase delay of the loop filter.

The *driving-point impedance* of the string termination is defined as the ratio of the force to the velocity at the point where the string drives the bridge. When the termination is passive, the driving-point impedance is *positive real* (see Appendix C). It is convenient to work with the *specific* impedance which is defined as the ratio of the driving-point impedance to the wave impedance of the string (ρc). Since the termination is assumed to be close to rigid, the specific driving-point impedance is large compared to unity.

In general, there are two important driving-point impedance functions, corresponding to the horizontal and vertical modes of bridge excitation. In the violin family, it is often assumed that the transverse mode of vibration is dominant because the bow displaces the string along this direction. However, near body resonances, where the impedance behavior is most important, incident vibration in the transverse direction is reflected with a different polarization [281]. This is caused by bridge-coupled resonances having a directionality other than tangential to the top of the bridge. For further discussion of this phenomenon, see [221].

Let $R_b(z)$ denote the z-transform of the specific driving-point impedance of the bridge termination. It is assumed that $R_b(z)$ is positive real, a condition which is preserved when the bilinear or matched-z transform is used to map the continuous-time impedance to discrete time (cf. Appendix C). To maintain alignment between the string resonant frequencies and the structural resonant frequencies of the body and bridge, the matched-z transformation should be used along with anti-aliasing filtering. If only a statistical distribution of resonances represented by $R_b(z)$ is needed, then the bilinear transform may be adequate. In either case, $R_b(z)$ satisfies

1) z real $\Rightarrow R_b(z)$ real,

2) $|z| \geq 1 \Rightarrow Re\{R_b(z)\} \geq 0$.

Given $R_b(z)$ we wish to find the corresponding reflection transfer function $H_b(z)$.

Adapting an argument from [257] yields

$$H_b(z) = \frac{1 - R_b(z)}{1 + R_b(z)}, \qquad R_b(z) = \frac{1 - H_b(z)}{1 + H_b(z)}, \tag{3.19}$$

The squared modulus of the reflection frequency response is then

$$\left|H_b(e^{j\omega})\right|^2 = \frac{1 - 2Re\{R_b(e^{j\omega})\} + \left|R_b(e^{j\omega})\right|^2}{1 + 2Re\{R_b(e^{j\omega})\} + \left|R_b(e^{j\omega})\right|^2} \leq 1, \tag{3.20}$$

since R_b is positive real. Note that conversely, if $R_b(z)$ is not positive real, then (3.20) shows that for some frequency $\left|H_b(e^{j\omega})\right| > 1$. Consequently, if there are no other losses in the string, *the string transfer function $H_l(z) = -H_b(z)$ is stable if and only if the reflection transfer function $R_b(z)$ is positive real.* The string is *strictly* stable if and only if $R_b(z)$ is *strictly* positive real.

It is shown in Appendix C that all positive-real functions have an equal number of poles and zeros, all of which are inside the unit circle (stable and minimum phase). This property holds only for the filter representing the reflection at a termination—the filtering which takes place along the string is stable but *not* minimum-phase in general.

Since typically the string termination is a stiff support, we have $|R_b(e^{j\omega})| \gg 1$, which implies

$$\left|H_b(e^{j\omega})\right| = 1 - \epsilon(\omega)$$

where $\epsilon(\omega) > 0$ is small. Since $|H_b(e^{j\omega})|$ is close to the stability boundary, a small deviation in amplitude response can translate into a relatively large change in the time-constant of decay at a given frequency for the vibrating string.

3.11. Practical Extensions of the String Model

In this section, various enhancements of the string model are discussed which improve its practical value for music applications. These improvements were suggested by David Jaffe at the time he was realizing his computer music work "Silicon Valley Breakdown." The refinements which follow were initially applied to the Karplus-Strong algorithm [254], which, when configured for plucked string sounds, may be considered a special case of the string model developed here. For a complete account of our extensions to the Karplus-Strong algorithm, see [253].

Previously, the analysis was done using notation which was closely connected to a simple physical picture. Now, however, practical implementation issues are the main topic, and so we will start speaking of time in *samples* rather than seconds. Accompanying this will be what was termed the *normalized* notation for z-transforms (cf. §3.8). Time indices will be denoted by n rather than t, and all times are in samples unless explicitly multiplied by T, the sampling period. A frequency denoted f is still in Hz, and $fT = f/f_s$ gives normalized frequency corresponding to the unit sampling rate. For example, $D_l(e^{j\omega})$ stands for the phase delay of the loop filter in *samples* at frequency $\omega = 2\pi fT$, where f is in Hz.

3.11.1. Fine-Tuning the String

If $H_l(z)$ is not measured separately for each note, then the string has tuning problems. Since the fundamental frequency is $f_1 = f_s/(P+D_l(f_1))$, the allowed pitches are quantized, especially at high frequency. For large values of P (low pitches), the difference between the pitch at P and $P+1$ is very slight. However for high pitches, P and $P+1$ yield very different pitches and tuning becomes crude.

This problem can be solved by adjusting the phase delay of the loop filter H_l. Recall that the fundamental frequency satisfies

$$f_1 = \frac{f_s}{P + D_l(f_1)}.$$

Suppose f_1 turns out to be mistuned due to the integer quantization of P. To make up the difference between f_1 and the desired frequency, we need to introduce into the feedback loop a filter which can contribute a small delay without altering the loop gain. This implies an *allpass* filter is needed. It turns out that a first-order allpass provides the necessary tuning freedom. The first-order allpass filter has the difference equation

$$y(n) = Cu(n) + u_{n-1} - Cy_{n-1}, \tag{3.21}$$

and transfer function

$$H_a(z) \triangleq \frac{C + z^{-1}}{1 + Cz^{-1}},$$

where C is the only coefficient to be set. For stability, we must have $|C| < 1$. The amplitude response is given by

$$G_a(f) \triangleq \left| H_a(e^{j\omega}) \right| = \frac{\left| C + e^{-j\omega} \right|}{\left| 1 + Ce^{-j\omega} \right|} = 1, \qquad \omega = 2\pi f.$$

The use of an allpass ensures that no modification of the decay rate will take place. The loop gain is $G_a(f)G_l(f) = G_l(f)$ as before.

The transfer function of the whole string (ignoring the numerator) is now

$$H(z) \triangleq \frac{1}{1 - H_a(z)H_l(z)}.$$

The phase delay of H_a is selected to tune f_1 to the precise desired frequency. This requires only the ability to select phase delays between 0 and T seconds, i.e., one sample's worth.

The phase delay, in samples, of the first-order allpass H_a is given by

$$\begin{aligned} D_a(f) &\triangleq -\frac{\angle H_a(e^{j\omega})}{\omega} \\ &= -\frac{1}{\omega}\angle\frac{C + e^{-j\omega}}{1 + Ce^{-j\omega}} \\ &= \frac{\angle\left(1 + Ce^{-j\omega}\right)}{\omega} - \frac{\angle\left(C + e^{-j\omega}\right)}{\omega} \\ &= \frac{1}{\omega}\tan^{-1}\left(\frac{-C\sin(\omega)}{1 + C\cos(\omega)}\right) - \frac{1}{\omega}\tan^{-1}\left(\frac{-\sin(\omega)}{C + \cos(\omega)}\right). \end{aligned} \tag{3.22}$$

When the arguments to arctangent above have magnitude less than unity, the power series expansion [132]

$$\tan^{-1}(x) = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \cdots, \qquad |x| < 1$$

holds. Thus the *low-frequency* phase delay is approximable by

$$D_a(f) \approx \frac{\sin(\omega)}{\omega(C + \cos(\omega))} - \frac{C\sin(\omega)}{\omega(1 + C\cos(\omega))} \approx \frac{1}{C+1} - \frac{C}{1+C} = \frac{1-C}{1+C}. \tag{3.23}$$

A plot of the exact phase delay is given in Fig. 3.24 for 17 values of C equally spaced between -0.999 and 0.999, inclusive. Note that delays between 0 and 1 sample can be provided somewhat uniformly across the frequency axis. A delay of 0 samples corresponds to $C = 1$, where the pole and zero of $H_a(z)$ cancel to give $H_a(z) \equiv 1$. However, pole-zero cancellation is inadvisable in practice, since roundoff errors may yield an unstable filter, or one which tends to overflow in fixed-point arithmetic. Therefore, it is preferable to shift the range of one-sample delay control to the region $\epsilon \leq D_a \leq (1 + \epsilon)$, for some small nonnegative ϵ ($0 < \epsilon \ll 1$). It is preferable not to shift very far since the phase-delay curves are less flat in the region beyond one sample's delay.

Note that the delay curves below the one-sample level in Fig. 3.24 correspond to slightly flattened upper partials, while the delay curves above the one-sample level correspond to slightly sharpened upper partials. The timbre change due to slight systematic shifting of

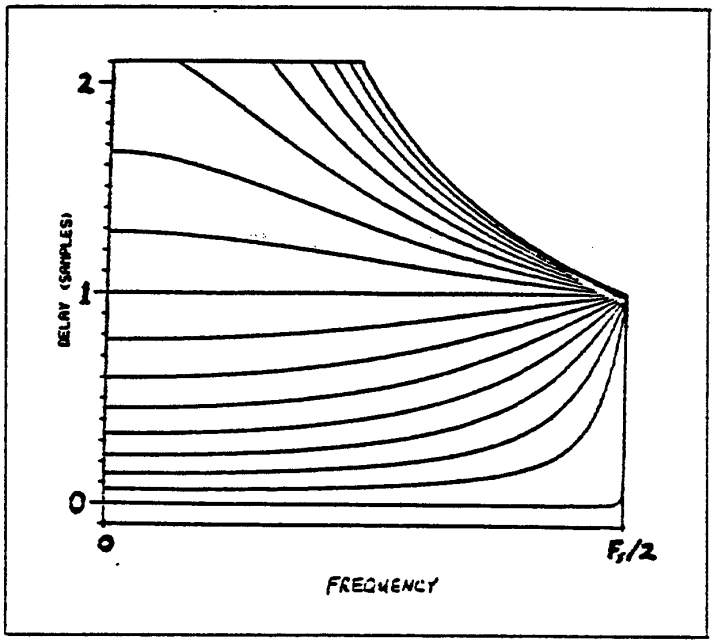

Figure 3.24. Phase delay for the fine-tuning allpass filter $H_a(z) = (C + z^{-1})/(1 + Cz^{-1})$.

the upper partials, of an amount less than one sample period, was found to be hardly noticeable in our implementation.

To precisely tune the instrument to a desired fundamental frequency f_1, let P_1 equal f_s/f_1, the real value for the period of the first partial, in samples, which would give perfect tuning. Then we desire $P + D_l(f_1) + D_a(f_1) = P_1$. The integer buffer length P and the delay $D_a(f_1)$ required from the allpass filter become

$$\begin{aligned} P &\triangleq \lfloor P_1 - D_l(f_1) - \epsilon \rfloor \\ D_a(f_1) &\triangleq P_1 - P - D_l(f_1), \end{aligned} \tag{3.24}$$

where $\epsilon > 0$ is the offset which shifts $D_a(f_1)$ into the range $[\epsilon, 1 + \epsilon]$.

We next solve for the filter coefficient C in (3.22) as a function of $D_a(f_1)$. Taking the tangent of both sides, and using an identity for the tangent of a difference leads to the quadratic equation in C,

$$C^2 \sin(\omega_1 D_a(f_1) + \omega_1) + 2C \sin(\omega_1 D_a(f_1)) + \sin(\omega_1 D_a(f_1) - \omega_1) = 0,$$

where $\omega_1 \triangleq 2\pi f_1$. The solution is found, after some manipulation [276], to be

$$C = \frac{-\sin(\omega_1 D_a(f_1)) \pm \sin(\omega_1)}{\sin(\omega_1 D_a(f_1) + \omega_1)}.$$

We have introduced an extra root by producing a quadratic equation. The previous approximation (3.23) indicates that the + sign should be taken. Therefore, the final solution is

$$C = \frac{\sin(\omega_1) - \sin(\omega_1 D_a(f_1))}{\sin(\omega_1 D_a(f_1) + \omega_1)} = \frac{\sin\left(\frac{\omega_1 - \omega_1 D_a(f_1)}{2}\right)}{\sin\left(\frac{\omega_1 + \omega_1 D_a(f_1)}{2}\right)} \tag{3.25}$$

which can be approximated, at low frequencies, by

$$C \approx \frac{1 - D_a(f_1)}{1 + D_a(f_1)}. \tag{3.26}$$

3.11.2. Approximating Nonlinearities

The amplitude of the synthetic string signal is proportional to the amplitude of its input driving function. This is an unsatisfactory control in simulating the timbral effect of dynamic level as it occurs in the case of a real stringed instrument [277]. This is due in part to the fact that real strings are *nonlinear*, and the harmonic and intermodulation distortion due to nonlinearity can contribute significantly to the timbre. The first-order effect of nonlinearities in a real string is an increase in the "brightness" of the tone. Harmonic distortion can be simulated in the string model by means of a filter which boosts the high-frequencies present in the waveform. Since the complete harmonic series is usually present in applications of the string model for which the phase delay of the loop filter $H_l(z)$ is constant, any harmonic distortion can, in principle, be simulated. When the string model is used with inharmonic partial overtones, however, the simulation of nonlinear operation fails to produce intermodulation distortion as required for full realism.

In this section, a means for providing dynamic level control using a simple low-pass filter is described. A higher dynamic level is implemented by increasing the effective spectral *bandwidth* of a tone in order to boost its apparent intensity. This technique was developed jointly with David Jaffe.

The bandwidth is controlled by means of a one-pole low-pass filter applied to the initial period (before it is fed into the string). This filter will be referred to as the "dynamics filter". The difference equation of the dynamics filter is

$$y(n) = (1 - R)u(n) + Ry(n - 1)$$

and its transfer function is

$$H_d(z) \triangleq \frac{1-R}{1-Rz^{-1}}, \tag{3.27}$$

where R is a real number between 0 and 1, computed as a function of fundamental frequency f_1 and the desired dynamic level L. When a series of notes at pitch f_1 is played while R is moved gradually toward 1, a diminuendo is approximated in terms of both decreasing loudness and spectral bandwidth reduction.

We define the *dynamic level* L as a *bandwidth* between 0 and $f_s/2$. If L is small, the spectrum is more low-passed, corresponding to a softer dynamic level. Similarly, large L gives a bright spectrum corresponding to louder notes. It is not sufficient to use a fixed low-pass filter for all pitches since low-pitched notes would then be louder than high-pitched notes. Rather, for a given dynamic level, R must be changed with pitch to yield a uniform perceived loudness. While this is a difficult problem in general, a good approximation is obtained by varying R so that the amplitude of the fundamental frequency component is always the same.

It remains to be shown how R is computed for a given pitch f_1 and dynamic level L. The main steps are as follows: First, a one-pole low-pass filter is designed having bandwidth L. Next the gain of this filter at a "middle" frequency is computed. Finally, the dynamics filter is computed as a one-pole low-pass having this gain at the desired fundamental f_1. The remainder of this section gives the equations needed for these steps.

The reference frequency f_m is chosen as the logarithmic middle (geometric mean) of the range to be used (a function of the particular musical context and the sampling frequency),

$$f_m = e^{\frac{1}{2}(\ln(f_u)+\ln(f_l))} = (f_u f_l)^{\frac{1}{2}}$$

where f_u is the upper pitch limit ($< f_s/2$), and f_l is the lower pitch limit.

The one-pole low-pass having bandwidth L is given by

$$H_L(z) \triangleq \frac{1-R_L}{1-R_L z^{-1}},$$

where

$$R_L \triangleq e^{-\pi L T}.$$

The substitution $R_L = e^{-\pi L T}$ is a somewhat standard approximate formula for mapping bandwidth to pole radius; it derives from applying the *matched-z* transformation [196] to the conventional bandwidth approximation (in terms of damping factor) for an analog two-pole resonator.

The gain of the low-pass filter H_L at the reference frequency is defined as

$$G_L(f_m) = \left| H_L(e^{j2\pi f_m T}) \right| = \frac{1-R_L}{\left|1 - R_L e^{-j2\pi f_m T}\right|},$$

where T is the sampling period.

Now, for any desired fundamental frequency f_1, R is computed so as to provide gain $G_d(f_1) = G_L(f_m)$. In other words, *all fundamental-frequency components are made to have the same amplitude.* The value of R is found by solving

$$G_L(f_m) = \frac{1-R}{\left|1 - Re^{-j2\pi f_1 T}\right|}.$$

Squaring both sides of this equation, and solving the resulting quadratic polynomial in R yields

$$R = \frac{1 - G_L^2(f_m)\cos(2\pi f_1 T)}{1 - G_L^2(f_m)} \pm 2G_L(f_m)\sin(\pi f_1 T)\frac{\left(1 - G_L^2(f_m)\cos^2(\pi f_1 T)\right)^{\frac{1}{2}}}{1 - G_L^2(f_m)}.$$

We use whichever value is less than 1 in magnitude for stability.

A family of frequency-response curves for H_d is shown in Fig. 3.25 for six fundamental frequencies in octave steps from $f_1 = 100$ Hz to $f_1 = 3200$ Hz. The dynamic level in each case is $L = 100$ Hz. A vertical line is drawn to each curve at the fundamental frequency to which it applies. The reference frequency f_m is set to 282.84 Hz (the geometric mean of $f_l = 20$ Hz and $f_u = f_s/2$), and the sampling rate is $f_s = 8000$ Hz.

To add to the effect of simulated dynamics, it is sometimes helpful to increase the damping on the low soft notes, using a "resistive" loss factor $g < 1$ in cascade with $H_l(z)$ [277].

3.11.3. Coupled Strings

The illusion of a sympathetically vibrating string can be created by exciting one copy of the string simulator by a small percentage of the output from the main string.

The effect of several sympathetic strings can be created simply by a bank of parallel "sympathetic strings" as defined above, each tuned to an arbitrary frequency. Each string output can be added (highly attenuated) to the input of every other string to provide full coupling. The output of the whole assembly is just the sum of the string outputs.

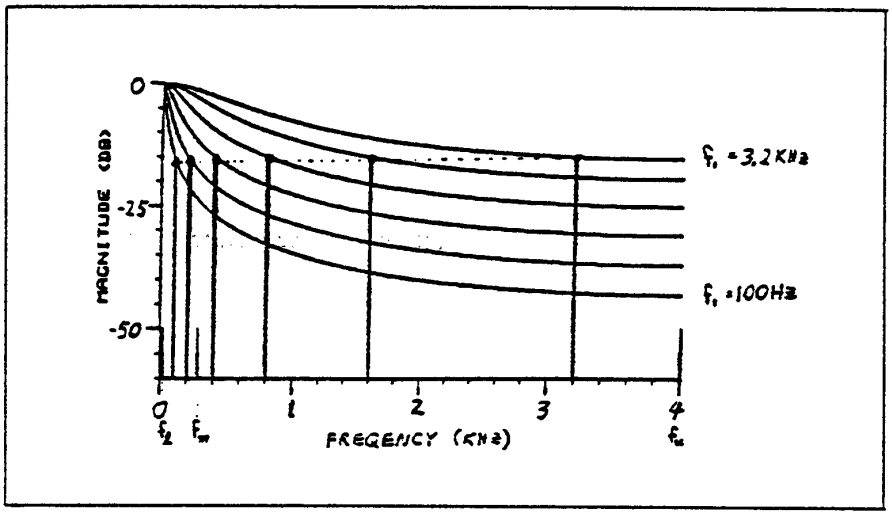

Figure 3.25. Frequency response of $H_d(z) = (1-R)/(1-Rz^{-1})$ computed at dynamic level $L = 100$ for six values of f_1.

3.11.4. Moving the Point of Excitation

An examination of the general string transfer function (3.7–3.9) (cf. also §3.9) shows that the principle effect of the numerator is to introduce severe losses at nearly harmonically spaced intervals in the spectrum. If the pluck-point x_i divides the length of the string L, then harmonics numbered $L/x_i, 2L/x_i, \ldots$ are suppressed completely in the ideal string. Thus the "pick" or "bow" position can be easily simulated by zeros uniformly distributed over the spectrum of driving excitation. These zeros may be placed on a circle in the z-plane. As the radius of this circle approaches one, the effect becomes more and more pronounced. This has been found to simulate convincingly the effect of plucking or bowing a string at varying distances from the bridge. The string input is filtered with a "comb filter," H_c having the difference equation

$$y(n) = u(n) - \gamma u(n - \mu P_1),$$

where μ is the fraction of the string between the bridge and point of excitation, P_1 is the period of the played string, and $\gamma \in [0, 1]$ is used to vary the amount of the effect (which is strongest for $\gamma = 1$ and weakest for $\gamma = 0$). For example, in the case of a harmonically vibrating string ($D_l(f)$ constant), $\mu = 1/2$ causes the even harmonics to be removed, and the effect is that of driving a string at its midpoint. Similarly, when $\mu = 1/10$, every tenth harmonic is suppressed, and the effect is like playing a tenth of the way up the string. With $\mu = 1/P_1$, the filter approximates a differentiator, creating a sharp *sul ponticello* sound.

3.12. String-Loop Identification

This section treats the problem of identifying the string loop filter $H_l(z)$. which represents the total filtering accumulated in the propagation once up and down the string (one period).

As was discussed in the derivation of equation (3.6), the string transfer function can be expanded in a formal power series to yield

$$H(z) = \frac{U(z)}{1 - z^{-P}H_l(z)} = U(z) + z^{-P}H_l(z)U(z) + z^{-2P}H_l^2(z)U(z) + \cdots,$$

where $U(z)$ may be thought of as the numerator of $H(z)$, or as the z-transform of the initial period of vibration (in which case $H(z)$ is the output transform). This expression illustrates the interpretation of $H_l(z)$ as a filter which is applied successively at periodic intervals to generate the output waveform from the initial conditions.

3.12.1. Error Criterion

As in any modeling problem, we must first think about the error criterion. The basic approach is to state what would be ideal to minimize, and then whittle it down by various sensible approximations until it is tractable with known approximation techniques.

Magnitude Error

An important role of the amplitude response $|H_l(e^{j\omega})|$ in the string loop is to provide frequency-dependent damping in the string. Such damping causes different decay rates of the various partials when the string is left to vibrate freely. A natural choice of error would then be the error in the time-constant of decay. In equation (3.14), (§3.10.1), it was shown that the decay time-constant is given by

$$\tau_\omega = -\frac{P - \angle H_l(e^{j\omega})/\omega}{\ln\left|H_l(e^{j\omega})\right|} \quad \text{(samples)},$$

where P is the loop bulk-delay in samples. The *time-constant error* is then defined by

$$\hat{E}_\tau(\omega) \triangleq \frac{P - \angle \hat{H}_l(e^{j\omega})/\omega}{\ln\left|\hat{H}_l(e^{j\omega})\right|} - \frac{P - \angle H_l(e^{j\omega})/\omega}{\ln\left|H_l(e^{j\omega})\right|} = \frac{P\ln\left|H_l(e^{j\omega})\right| - P\ln\left|\hat{H}_l(e^{j\omega})\right|}{\ln\left|H_l(e^{j\omega})\right|\ln\left|\hat{H}_l(e^{j\omega})\right|}, \quad (3.28)$$

where $\hat{H}_l(e^{j\omega})$ is the approximate frequency response, and $H_l(e^{j\omega})$ is the true frequency response of the string-loop filter. We assume that P is known. Since P is typically much

larger than the phase-delay of the string-loop filter, the terms $\angle\hat{H}_l(e^{j\omega})/\omega, \angle H_l(e^{j\omega})/\omega$ are both much less than P and have been dropped.

The denominator of (3.28) may be interpreted as a weighting which approaches infinity as the loop gain approaches one. This is appropriate since such gains correspond to very long time constants, and a small change in magnitude produces a large change in the time constant. This weight function is still provided to a good extent if we approximate $|\hat{H}_l(e^{j\omega})|$ by $|H_l(e^{j\omega})|$ in the denominator of (3.28). The advantage of this modification is that it will reduce the error criterion to one discussed in Chapter 1, viz.,

$$\hat{E}_\tau(\omega) \approx P\frac{\ln\left|H_l(e^{j\omega})\right| - \ln\left|\hat{H}_l(e^{j\omega})\right|}{\ln^2\left|H_l(e^{j\omega})\right|} \approx \frac{P}{2}\frac{\ln\left|H_l(e^{j\omega})\right|^2 - \ln\left|\hat{H}_l(e^{j\omega})\right|^2}{\ln^2\left|H_l(e^{j\omega})\right|}. \quad (3.29)$$

In the terminology of Chapter 1, this error classifies as a ***weighted log-power frequency-response error***, and a method was given in §1.8.5 which minimizes the L^∞ norm of its first-order Taylor-series approximation with respect to $|\hat{H}_l|$. Thus, the final magnitude error criterion is given by

$$J_\tau(\hat{\theta}) \triangleq \left\| \frac{\left|H_l(e^{j\omega})\right|^2 - \left|\hat{H}_l(e^{j\omega})\right|^2}{\left|H_l(e^{j\omega})\right|^2 \ln^2\left|H_l(e^{j\omega})\right|} \right\|_\infty \quad (3.30)$$

where $\hat{\theta}$ is the vector of filter coefficients for $\hat{H}_l(z)$. The final error inside the norm also happens to be the first-order term in the Taylor expansion of (3.28) with respect to $|\hat{H}_l|$ about $|H_l|$, making the explicit approximation from (3.28) to (3.29) unnecessary.

It is an important benefit that the final form (3.30) can be minimized by an algorithm which is guaranteed to converge monotonically to an optimum solution. Furthermore, this is true even for rational filters $\hat{H}_l$, which, as Chapter 1 discusses at length, is typically rare in filter design.

Phase Error

We have only arrived at a means for obtaining a good *magnitude* approximation. We now consider the *phase*-response error, and its importance.

An important role of the loop-filter phase response is to provide inharmonic partials in the freely vibrating string. The ear is very sensitive to slight relative perturbations in the frequency of sinusoids (cf. §3.2). Let ω_k denote the true radian frequency of the kth partial overtone of the string, and let $\hat{\omega}_k$ denote the kth partial frequency of the string model. Also, let $D_l(\omega) = -\angle H_l(\omega)/\omega$ denote the phase-delay of $H_l(z)$. Then minimizing

the worst-case relative deviation in partial-tuning means to minimize (cf. §3.10.1)

$$\left\| \frac{\omega_k - \hat{\omega}_k}{\omega_k} \right\|_\infty = \left\| 1 - \frac{\hat{\omega}_k}{\omega_k} \right\|_\infty = \left\| 1 - \frac{2\pi k f_s/(P + \hat{D}_l(\hat{\omega}_k))}{2\pi k f_s/(P + D_l(\omega_k))} \right\|_\infty$$

$$= \left\| 1 - \frac{P + D_l(\omega_k)}{P + \hat{D}_l(\hat{\omega}_k)} \right\|_\infty = \left\| \frac{\hat{D}_l(\hat{\omega}_k) - D_l(\omega_k)}{P + \hat{D}_l(\hat{\omega}_k)} \right\|_\infty \approx \frac{1}{P} \left\| \hat{D}_l(\omega_k) - D_l(\omega_k) \right\|_\infty .$$

Thus a good phase error criterion is given by

$$J_f(\hat{\theta}) \triangleq \left\| \hat{D}_l(\omega) - D_l(\omega) \right\|_\infty = \left\| \frac{\angle H_l(e^{j\omega}) - \angle \hat{H}_l(\omega)}{\omega} \right\|_\infty . \tag{3.31}$$

Minimizing the error in the tuning of the string overtones thus corresponds to minimizing the error in the *phase delay* of the string-loop filter. Assuming the magnitude response is obtained by log-power matching, as discussed above, one can then design an allpass filter having the optimum phase-delay approximation (cf. §1.8.6 of Chapter 1). This does not produce an overall optimal solution since the phase and magnitude have been matched separately in two independent sections of the filter $\hat{H}_l(z)$. However, the increase in filter order required may be offset by the greater relevance of the two error criteria. A further advantage is that the relative accuracy in decay-rate and partial-tuning can be easily controlled.

The ability to work with the error criteria (3.30) and (3.31) depends on having the true string-loop frequency response $H_l(e^{j\omega})$ to use as a desired function. A method for measuring $H_l(e^{j\omega})$ will be discussed later. Alternatively, it is possible to form string models directly from the recorded behavior of the freely vibrating string, as discussed in the next two subsections.

3.12.2. A Linear Prediction Approach

The structure

$$H(z) = \frac{1}{1 - z^{-P} H_l(z)} \tag{3.32}$$

where $H_l(z)$ is a low order string-loop transfer function of the *finite impulse response* (FIR) class

$$H_l(z) \triangleq h_l(0) + h_l(1)z^{-1} + \cdots + h_l(N)z^{-N}$$

can be estimated from the behavior of a freely vibrating by means of a modified form of *linear prediction*. To see this, note that in the time domain, (3.32) implies

$$y(n) = H(d^{-1})u(n) = \frac{u(n)}{1 - d^P H_l(d^{-1})} = u(n) + H_l(d^{-1})y(n - P), \tag{3.33}$$

where d is the unit-sample delay operator.* If the string is vibrating freely at time n, then the input $u(n)$ is finished, which gives

$$y(n) = H_l(d^{-1})y(n-P).$$

The modeling problem is then to find the coefficients of $H_l(d^{-1})$. This can be done by minimizing the energy of the signal

$$\hat{e}(n) \triangleq y(n) - \hat{H}_l(d^{-1})y(n-P).$$

with respect to the coefficients $\hat{h}_l(i), i = 1, \ldots, N$. This is nothing more than a linear prediction problem in which the prediction takes place over a span of P samples instead of the usual one sample. The various styles of solution to the one-step linear prediction problem may be carried over to the P-step case with no particular surprises, and the reader is referred to Markel and Gray [186] for a comprehensive (though advanced) treatment of the one-step case. We will solve the P-step linear prediction formulation as a special case of the system-identification approach, presented in the next section.

The P-step linear-prediction formulation has not received a lot of attention in the literature, although forms closely related to it exist. In [160], a similar form was proposed for the purpose of eliminating the fine-structure from the spectrum of the residual output of a conventional linear predictor for voiced speech. Also, related models have been applied to seasonally varying time-series [135]. We have also applied it successfully to *pitch tracking*; in this application, the coefficients $\{h_l(i)\}_0^N$ exhibit a peak corresponding to the estimated period. Pitch tracking can also be based on the average phase-delay of $H_l(z)$. The basic structure (3.32) can provide a parsimonious model for a wide class of quasi-periodic signals.

3.12.3. A System Identification Approach

The results of §3.10 and Appendix C imply that the reflection transfer function corresponding to passive terminations is a *Schur function*.† As such, it must have the same number of zeros as poles. Thus it is desirable to allow pole-zero modeling of the string-loop transfer function, so that more physically accurate models are possible. Rational models can be estimated directly from string behavior by means of *system identification* methods (described in Chapter 2).

* This delay-operator notation was used extensively in Chapter 2. Recall that $d^k x(n) \triangleq x(n-k)$, and that d can be replaced by z^{-1} to convert the delay-operator function into a transfer function.

† Defined in Appendix E.

The free vibration of the string is expressed by equation (3.33) with the input $u(n)$ set to zero, giving

$$y(n) = H_l(d^{-1})y(n-P).$$

$H_l(z)$ is assumed to be a linear transfer function which is to be modeled with a *rational* transfer function of the form

$$H_l(z) \triangleq \frac{\hat{B}(z)}{\hat{A}(z)}, \tag{3.34}$$

where

$$\begin{aligned} \hat{B}(z) &\triangleq \hat{b}_0 + \hat{b}_1 z^{-1} + \cdots + \hat{b}_N z^{-N} \\ \hat{A}(z) &\triangleq 1 + \hat{a}_1 z^{-1} + \cdots + \hat{a}_N z^{-N}, \end{aligned} \tag{3.35}$$

and the order N is given. As developed in Chapter 2, applying the basic system-identification formulation means to minimize *equation error*

$$\hat{\epsilon}(n) = \hat{A}(d^{-1})y(n) - \hat{B}(d^{-1})y(n-P) \tag{3.36}$$

in the time domain under the L^2 norm. If we set $\hat{A}(d^{-1}) = 1$, then (3.36) reduces to the linear prediction formulation discussed in the previous section. In the more general case, the solution is given by a standard least-squares procedure which is derived in detail in Chapter 2 (§2.4, "The Regression Formulation").

When considering the application of system-identification methods, the question arises as to the utility of the more general formulation,

$$\hat{C}(d^{-1})\hat{\epsilon}(n) = \hat{A}(d^{-1})y(n) - \hat{B}(d^{-1})y(n-P).$$

The addition of the polynomial $\hat{C}(d^{-1})$ allows greater flexibility in the modeling of the noise term $\hat{\epsilon}(n)$. This will be of value only if the residual $\hat{\epsilon}(n)$ is to be used in the model in some way. For example, one might wish to characterize $\hat{\epsilon}(n)$ by its *statistical behavior* and feed it into the model according to the difference equation

$$y(n) = \frac{\hat{B}(d^{-1})}{\hat{A}(d^{-1})}y(n-P) + \frac{\hat{C}(d^{-1})}{\hat{A}(d^{-1})}e(n),$$

where $e(n)$ is statistically similar to $\hat{\epsilon}(n)$. Note that in practice, $\hat{\epsilon}(n)$ will depend on the nature of the initial excitation of the string used to generate $y(n)$.

3.12.4. Practical Issues

In both string modeling formulations, the solution is at first sight independent of the initial excitation of the string. This occurs because one is really only comparing one

period* to the next to find the filtering operation which takes one to the other. In practice, however, the excitation is important. For example, if the string is initially excited only at its fundamental frequency, then the string-loop frequency response can only be measured at that frequency. Similarly, when all partials are present, the string-loop frequency response is "sampled" only at these frequencies. In the case of modeling the violin body, harmonic excitations proved to be insufficient for identification because the resonance structure of the body cannot be outlined by the overtones of any single note in the range of the instrument, and the losses at the string termination have potentially the same general frequency-distribution as the body resonances.

The ideal experiment consists of sending an impulse into the string, and measuring its shape after one trip around the string loop. This only works if the impulse response of the string loop is less than one period (a reasonable assumption), since otherwise the measured impulse response will be "time-aliased."

Another possibility is the use of *pizzicato* recordings, i.e., the response to a pluck. In this case, two curvature impulses, traveling in opposite directions, are initialized in the string. This is less desirable since the measured spectrum will have nulls at all multiples of the first frequency having a node at the pluck-point. However, if the pluck-point is chosen very close to the bridge, then the first null will be at a very high frequency, e.g., above the 20th harmonic.

A means for obtaining a *nonparametric* estimate of the string-loop frequency response, to which the error criteria of §3.12.1 can be applied, is as follows. In a recording of N samples of the freely vibrating string, $y(n), n = 1, \ldots, N$, take $N - P$ samples in the time range $[1, N - P]$ and call the DFT of this segment $Y_i(e^{j\omega})$. Now take the same amount of data in the time range $[P + 1, N]$ and call its DFT $Y_o(e^{j\omega})$. It is helpful to use spectral smoothing in these DFT's in order to reduce side-lobe oscillation and spurious errors in the spectra. Now, the string-loop frequency response estimate is given by

$$\hat{H}_l'(e^{j\omega}) \triangleq \frac{Y_o(e^{j\omega})}{Y_i(e^{j\omega})}.$$

The prime on $\hat{H}_l'(e^{j\omega})$ is used to distinguish the nonparametric estimate from the parametric representation $\hat{H}_l(z)$ in terms of a rational filter. $\hat{H}_l'(e^{j\omega})$ can then be used as a desired frequency response in a filter-design technique.

Another practical issue is that of *pre-emphasis*. In both the system identification and linear prediction methods, there is an implicit weighting function on the frequency-response

* Again, the term "period" is used loosely. The signal is only approximately periodic. If it were exactly periodic, the string-loop transfer function would be 1.

error which is proportional to the excitation power spectrum. This can be seen by looking at the definition of equation error in the frequency domain:

$$\begin{aligned}\left|E(e^{j\omega})\right|^2 &= \left|\hat{A}(e^{j\omega})Y(e^{j\omega}) - \hat{B}(e^{j\omega})e^{-j\omega P}Y(e^{j\omega})\right|^2 \\ &= \left|\hat{A}(e^{j\omega})\right|^2\left|Y(e^{j\omega})\left(1 - e^{-j\omega P}\frac{\hat{B}(e^{j\omega})}{\hat{A}(e^{j\omega})}\right)\right|^2 \\ &= \left|\hat{A}(e^{j\omega})\right|^2\left|U(e^{j\omega})\right|^2\left|\frac{1 - e^{-j\omega P}\hat{H}_l(e^{j\omega})}{1 - e^{-j\omega P}H_l(e^{j\omega})}\right|^2,\end{aligned}$$

where $U(e^{j\omega})$ is the spectrum of the initial string excitation. If there is significantly less energy at high frequencies (as is the case after a short time of free vibration) then the high-frequency error may be larger. This was found to be a real problem in practice. A means of counteracting this behavior is to apply *pre-emphasis* to the data, e.g., work with

$$Y_p(e^{j\omega}) \triangleq C(e^{j\omega})Y(e^{j\omega}),$$

where $C(e^{j\omega})$ is a low-order polynomial obtained as the solution to one-step linear prediction modeling of Y. Thus Y_p is a prediction error spectrum, and tends to be flat, especially in the middle of the analysis time frame.

3.12.5. Performance on Pizzicato Data

Figure 3.26 shows a recording of a plucked violin G-string. The upper curve is the waveform at the bridge. Note how the main pulse decays and spreads over time. The loop filter $\hat{H}_l(z)$ must provide this behavior.

To equalize the spectral content, the time-waveform of Fig. 3.26a was filtered by a third-order linear-prediction inverse filter. The resulting time-waveform and spectrum are shown in Fig. 3.27. This signal is the prediction error from the linear prediction. Note that at high frequencies the signal does not have sharply defined partials or harmonics. This proved to lead to a string-loop filter with excessive loss at high frequencies.

The nonparametric method in which the FFT of one segment is divided by the FFT of a segment one period earlier gave unusable results due to excessive variance in the frequency-response estimate (even when heavy smoothing was used). For example, the loop-filter gain exceeded unity at many places in the spectrum. This was unfortunate since a good nonparametric estimate would allow the error criteria (3.30) and (3.31) of §3.12.1 to be used. The linear-prediction and system-identification methods, however, gave plausible results, as will be shown.

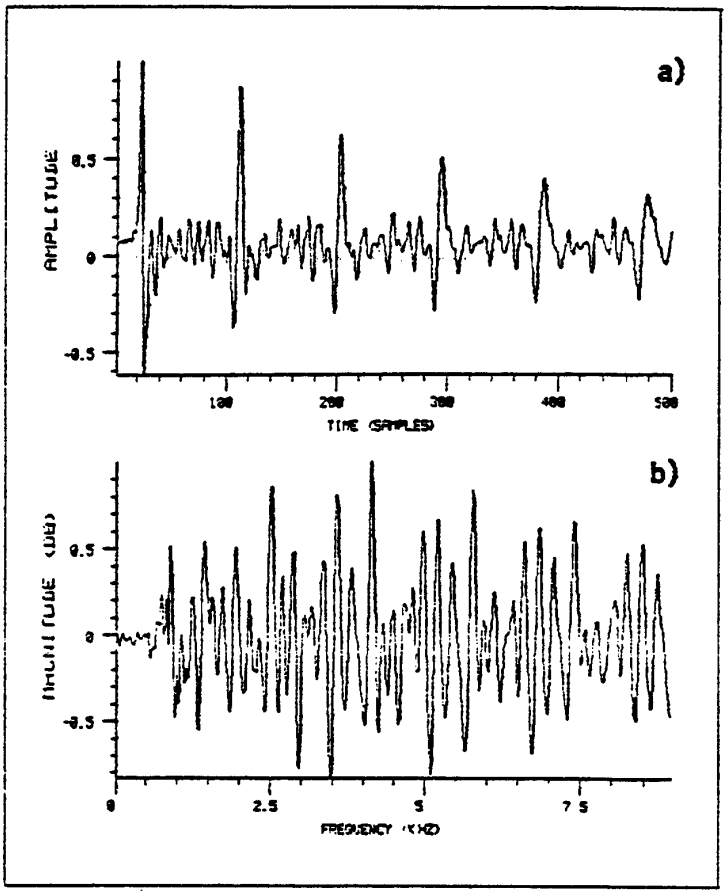

Figure 3.28. Time-domain input-output pair for the case of a plucked open violin G-string.

a) Force derivative at the bridge.

b) Sound pressure from the body.

The results of an order 8 and an order 20 string-loop filter estimate, obtained using the periodic linear prediction method, are overlayed in Figure 3.28. The plot shows the coefficients of the predictor, which can be interpreted as the appearance of an impulse after one round-trip on the string. The curves have been aligned to show the agreement obtained over their common delay range.

The performance of the string model using the smaller filter is shown in Figure 3.29. It is evident that the overall exponential decay is reasonable, but that the high frequencies die out too quickly. This is thought to be due to the lack of high-frequency energy in the

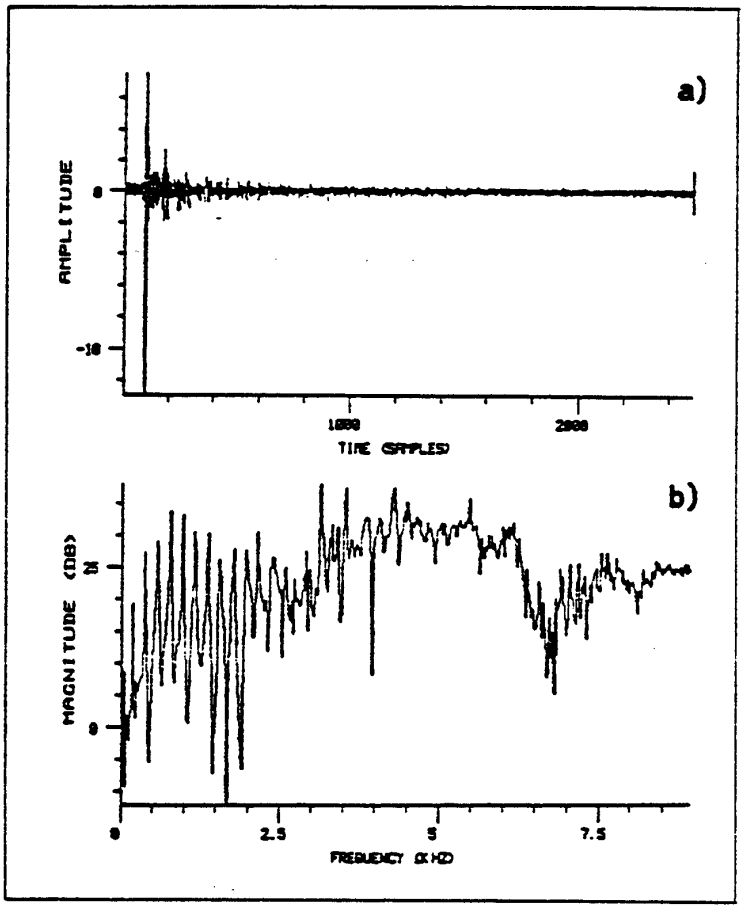

Figure 3.27. The same time-domain input as in Fig. 3.26a after applying third-order linear-predictive pre-emphasis.

a) Force derivative at the bridge.

b) Corresponding spectrum.

original recording. This type of trouble could be also be caused by stiffness; if the predictor does not span more than the range of effective string-length change over all frequencies, then those frequencies which are "out of reach" of the predictor appear uncorrelated and will be attenuated severely.

Figure 3.30 shows an overlay of three string-loop frequency responses obtained using the system identification approach. Each filter has the same number of poles as zeros, and the three cases are orders 6, 7, and 8. In each case, the bulk delay was adjusted by trial and error to maximize the gain of the string-loop filter at 0 Hz (theoretically 0 dB).

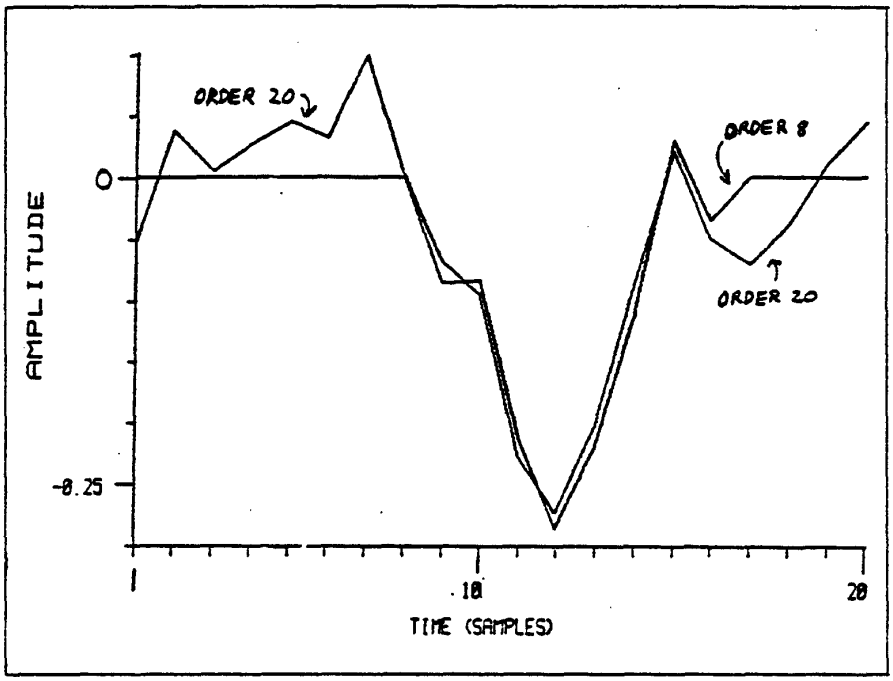

Figure 3.28. Two string-loop filter coefficient sets obtained by the linear prediction method. The FIR loop filters are of order 8 and 20, respectively.

This was a very important step due to an attenuation phenomenon discussed in Chapter 1 (§1.7.1). In addition to the three pole-zero cases, the order 8 FIR string-loop frequency response is included for comparison. Again, high frequencies appear decorrelated, and the high-frequency fit is somewhat random. The low-frequency identification, however, is quite consistent. Figure 3.31 illustrates the response of the string model when the order 8 recursive filter is used as the string-loop filter.

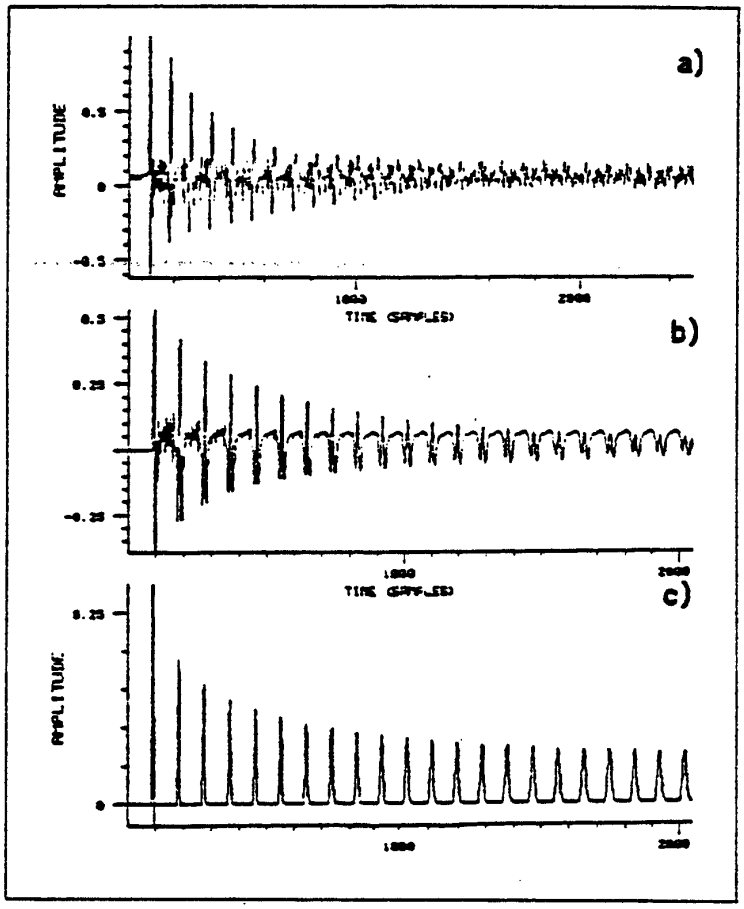

Figure 3.29. Performance of the order 8 FIR string-loop filter.
a) Original pizzicato recording.
b) String-model response initialized with the first period of a).
c) String-model impulse response.

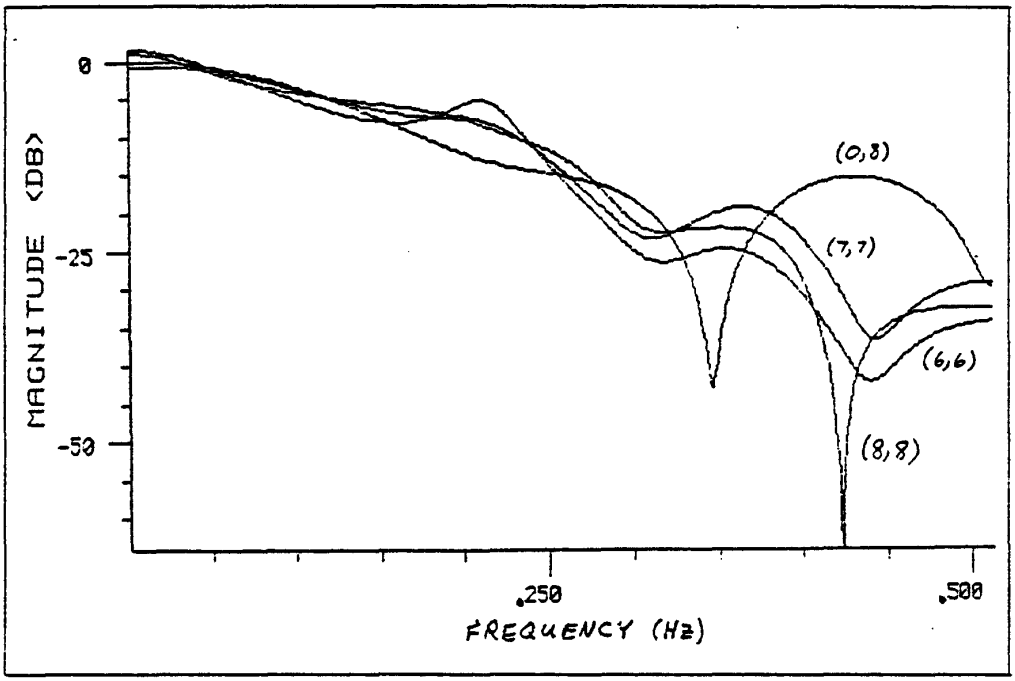

Figure 3.30. Three string-loop filter frequency-response magnitudes obtained by the system-identification method, and one filter obtained by the linear-prediction method. The string-loop filters have (pole,zero) orders (6,6), (7,7), (8,8), and (0,8), as marked in the figure. The horizontal axis ranges from 0 Hz to half the sampling rate.

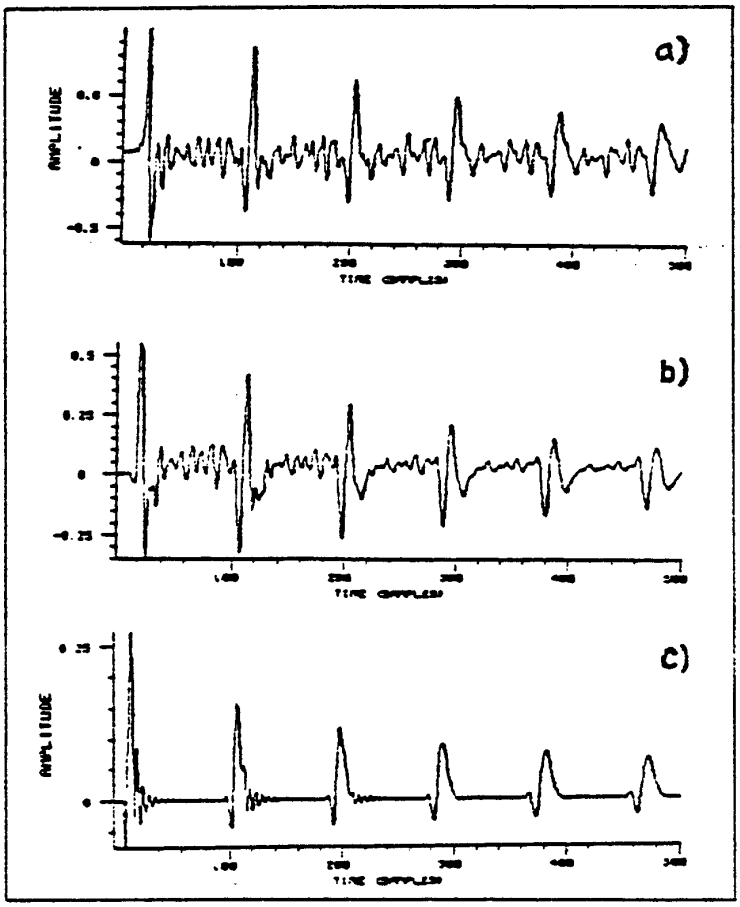

Figure 3.31. Performance of the 8-pole, 8-zero string-loop filter obtained by the system-identification method.

a) Original pizzicato recording.

b) String-model response initialized with the first period of a).

c) String-model impulse response.

3.13. Additional Refinements

"There is no problem, no matter how complex, which cannot, when viewed in the right way, become still more complex."

— Anderson's Law

Thus far, the model for the violin includes only the body and string. For plucked-string instruments, the model is somewhat complete. For bowed strings, however, this is only the beginning. The interaction between the bow and string plays a crucial role in the quality of transient sounds. Even for steady-state tones, the model as it now stands accounts for only a few of the characteristics that are observed with bowed strings. For more realistic behavior, covering a wide range of bowing styles, a more accurate modeling of the bowing process is required.

3.13.1. Bowing the String

As discussed previously, driving the string model with an impulse train provides Helmholtz-type behavior. The main improvement needed for realistic transient behavior is an input to the string which behaves more like a bow. This input is an additive component which models the effect of a bow with a given pressure, differential velocity, and position. The basic physical mechanism is the friction of the bow against the string. The most convenient choice of string simulation quantities is *transverse velocity*, obtainable by one time-integration of the acceleration waves considered previously.

Figure 3.32 shows a diagram of the bowed string which more accurately reflects the physics of bowing. This bowing model has been most extensively developed by McIntyre and Woodhouse [231]. To enhance the physical analogy, the propagation delay from the bow to nut and back has been split into two halves, placing the termination at the nut between them. Similarly, the bridge is located between two delay-lines corresponding to the portion of the string between the bow and bridge.

The operation of the "bow" is as follows: The variables sensed are the incoming left-going and right-going velocity waves which are denoted $v_{il}(n)$ and $v_{ir}(n)$, respectively. Let $v_i(n) = v_{il}(n) + v_{ir}(n)$ denote their sum (which is interpreted as the incoming transverse string velocity at the bowing point). We must solve for the change in velocity $\Delta v(n)$ as a function of the bow pressure, the relative velocity between the bow and string, and the string wave impedance. The velocity correction will be distributed equally in left-going and right-going directions to produce outward velocities $v_{ol} = v_{il} + \Delta v(n)/2$ and $v_{or} = v_{ir} + \Delta v(n)/2$. The net instantaneous string velocity under the bow is $v(n) = v_i(n) + \Delta v(n)$. The force of the bow applied at time-sample n will be $f(n) = F(v(n) - v_b(n))$, where F is the friction curve which relates force and velocity under the bow, at a given bow pressure and bow velocity v_b. This force must also correspond to a velocity shift of the amount $\Delta v(n) = Y f(n)$, where Y is the characteristic wave-admittance of the string. These last two equations are solved simultaneously in practice by finding the intersection of the functions

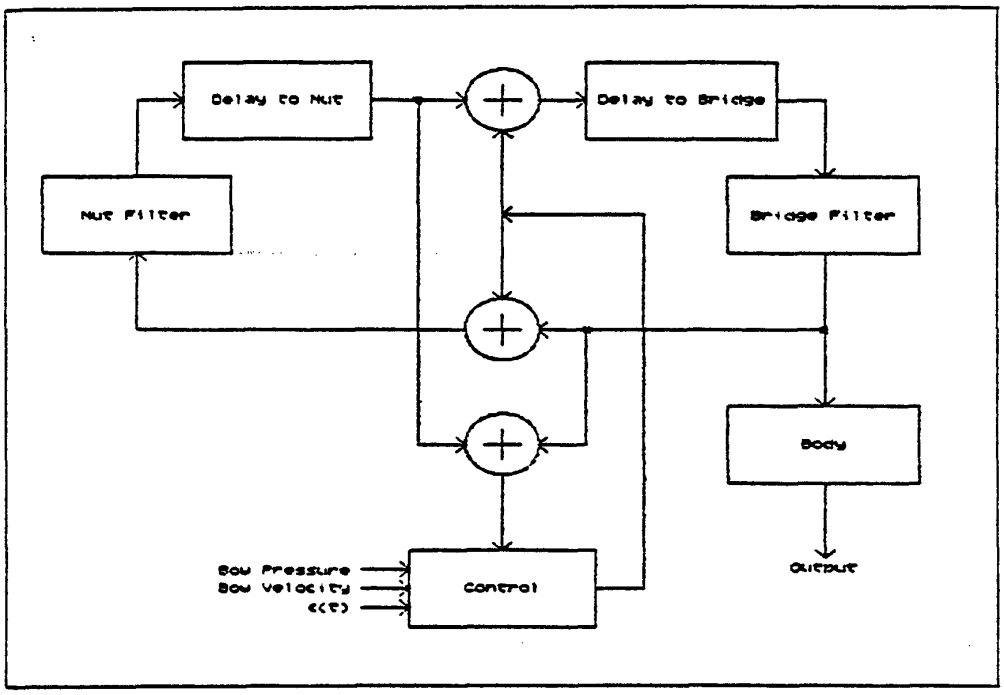

Figure 3.32. System diagram of the string model with the McIntyre-Woodhouse bowing system.

$$f = \Delta v / Y$$
$$f = F(\Delta v + v_i - v_b).$$

If bow force is held constant, then the intersection of the curves can be pre-computed and stored in a table look-up. Typical values of Y and $F(v)$ can be found in [235]. The bow force can be approximately represented by a vertical scaling of the friction curve. Bow position, of course, is represented by the sizes of the delay-lines to the left and right of the bowing point. It is straightforward to extend the above model to include torsional string waves, multiple bow-hairs, and realistic bow-hair dynamics [235].

3.13.2. Spikes

In [232], McIntyre and Woodhouse discuss a phenomenon known as "spikes" which is thought to be responsible for much of the noise which builds up when bowing with heavy force near the bridge. The main effect is that several secondary curvature impulses appear between the main Helmholtz impulses. As bowing pressure is increased further,

these impulses become larger and more widely separated. The secondary impulses appear somewhat random and might be well-modeled by a Poisson or renewal process [148].

The spikes are evidently due to the string slipping on the bow during the time it is normally stuck. A simple explanation would be that the bow velocity exceeds that of the string, and therefore the string slips a little to let the bow get ahead. However, this hypothesis is contradicted by the fact that spikes cannot be produced with a bow contacting the string at only one point [232]. Thus the finite width of the bow seems to play a crucial role in this phenomenon [232].

The mechanism proposed in [232] to explain the source of spikes is as follows. When the string is first captured by the bow, it is trapped at an angle under the bow. As the bow and string move together, the string is bent such that it is no longer a straight line from the bridge to the main Helmholtz corner. At some time, before the main corner returns to the bow to initiate slipping, the string slips enough to straighten out under the bow—a process called "differential slipping". This may happen several times within a period. The greater the bow force, the longer the string waits (and the more it deforms) between slips. The innermost bow-hairs slip most often, since the force on them is typically greater.

To provide spikes, a second bowing unit can be added to the structure of Fig. 3.32 which is separated from the other bow-point by a few samples delay on both the upper and lower rails.

In the simple model, where an acceleration impulse train is used as input, spikes may be simulated with extra impulses which arrive at random times and amplitudes, as in a Poisson "random shock" process.

3.13.3. Adding Vibrato

Another aspect of musical control is the inclusion of vibrato. This is obtained easily in the simplified model by placing frequency (and perhaps amplitude) vibrato on the driving impulse train. (Good results are obtained without varying the delay-line lengths.) In the more general case, using the McIntyre-Woodhouse bowing system, vibrato must be implemented by varying the size of the delay line which represents the length of the string. Since the delay lines are of integer length, unacceptable results are obtained unless an interpolation of delay length is performed. It was found that the technique presented in §3.11.1 can be easily adapted to provide vibrato with no audible distortion due to string quantization.

3.14. Conclusions Regarding the String Model

A computational model for bowed strings has been presented, which captures several of the most musically important aspects of real stringed instruments. This capability is obtained at very low cost compared to other approaches of similar generality. Several schemes for calibrating the model to recorded data were proposed, and some results for pizzicato data were presented. Although approximations were made to arrive at the final structure, it has been found that the performance is very realistic in practice. However, there is still considerable work to be done in making a useful family of bowed strings using the techniques presented.

One area for future work is in experiment design for the methods of §3.12.5. The pilot experiments presented there are by no means final. While demonstrating the basic feasibility of identification methods, they point out areas of practical difficulty such as in obtaining reliable high-frequency estimates of the string-loop frequency response. It would also be much better to measure or infer force at the bridge without imbedding a sensor in the instrument; alterations of the playing properties seem unavoidable with this technique, and it should not be considered for instruments of very high quality. Norman Pickering has employed a phonograph pickup resting on the bridge to measure transverse velocity. Since the vertical force of the string on the bridge is significant, it is suggested that the excitation at the bridge be measured in both the horizontal and vertical directions. Consequently, there would be two body transfer functions, one from each input to the output.

For musical purposes, the most pressing problem is that of control. The complete model requires functions of time corresponding to bow pressure and velocity. It is somewhat awkward to specify bowing style by these functions. What is needed is a versatile library of pressure and velocity curves (dependent on bow position) which cover the useful bowing regimes [241,244]. Once such a library exists, it should be straightforward to find functional approximations. It is conceivable that empirical measurements could provide these data, using, for example, laser interferometry during normal playing. Such experiments could also be used to estimate the statistics of the innovations signal for natural bowed strings.

The innovations signal may be important for "breathing life" into the sound. That is, a purely deterministic simulation of bowed strings may prove to be fatiguing to the ear even when realistic (though smooth) excitation functions are used. Probably the most expedient path to satisfactory stochastic components is by way of trial and error input and/or model perturbations. A real musical instrument may be so inherently complex that a prohibitively large and accurate deterministic model is necessary to obtain an innovations signal which is well-modeled by a pseudo-random sequence.

Appendix A. Non-Concavity of Problem $\hat{H}^*$

In this appendix, it is shown that the recursive filter design problem, called "problem $\hat{H}^*$" in Chapter 1, is extremely difficult regardless of the choice of error norm. In particular, it is shown that there is *no upper bound on the number of locally best approximations.* This means that algorithms based on local searching of the error surface can in general never know when they have reached an optimum solution.

The outline of the proof is as follows. First, a construction is given which provides a "desired" frequency response that gives any number of local minima in one-pole approximation under the L^2 norm. These local minima have controllable curvature. It is next shown that the *norm equivalence theorem* can be used to extend this construction to provide an arbitrary number of locally best approximations under virtually any norms.

Lemma A.1. Given any set of K distinct stable one-pole filters,

$$\hat{H}_i(z) = \frac{1}{1 - r_i z^{-1}}, \quad 0 < r_i < 1, \quad i = 1, 2, \ldots K,$$

there exists a bounded causal filter $H(z)$ having each $\hat{H}_i(z)$ as a locally best approximation under the L^2 norm. If $H(z)$ is taken from the set of order $2K$ FIR filters,

$$H(z) = h(0) + h(1)z^{-1} + \cdots + h(2K)z^{-2K},$$

and if the curvature of the squared L^2 error norm

$$c_i \triangleq \frac{\partial^2 J(r)}{\partial r^2}(r_i) > 0, \quad i = 1, \ldots, K,$$

is specified for each $\hat{H}_i$, then $H(z)$ is unique and is given by

$$\begin{pmatrix} h(1) \\ h(2) \\ \vdots \\ h(2K) \end{pmatrix} = \begin{pmatrix} 1 & 2r_1 & 3r_1^2 & \cdots & 2Kr_1^{2K-1} \\ 0 & 2 & 6r_1 & \cdots & 2K(2K-1)r_1^{2K-2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 2r_K & 3r_K^2 & \cdots & 2Kr_K^{2K-1} \\ 0 & 2 & 6r_K & \cdots & 2K(2K-1)r_K^{2K-2} \end{pmatrix}^{-1} \begin{pmatrix} \alpha_1 \\ \beta_1 \\ \vdots \\ \alpha_K \\ \beta_K \end{pmatrix}, \tag{A.1}$$

with $h(0) = 1$, and

$$\alpha_i \triangleq \sum_{n=0}^{\infty} n r_i^{2n-1} = \frac{r_i}{(1 - r_i^2)^2},$$

$$\beta_i \triangleq \sum_{n=0}^{\infty} n(2n-1) r_i^{2n-2} - \frac{c_i}{2} = \frac{1 + 3r_i^2}{(1 - r_i^2)^3} - \frac{c_i}{2}.$$

Proof. Minimization of $J_2(r)$ gives the same answer as minimizing

$$V(r) \triangleq J_2^2(r) = \left\| H(e^{j\omega}) - \frac{1}{1-re^{-j\omega}} \right\|_2^2 = \int_{-\pi}^{\pi} \left| H(e^{j\omega}) - \frac{1}{1-re^{-j\omega}} \right|^2 \frac{d\omega}{2\pi} = \sum_{n=0}^{\infty} (h(n) - r^n)^2 . \tag{A.2}$$

Necessary conditions for a locally best approximation at $r = r_i$ are that

$$\frac{\partial V(r)}{\partial r}(r_i) = 0 \qquad \frac{\partial^2 V(r)}{\partial r^2}(r_i) = c_i > 0 \tag{A.3}$$

for $i = 1, \ldots, K$. The derivative of $V(r)$ at $r = r_i$ is

$$\frac{\partial V(r)}{\partial r}(r_i) = \frac{\partial}{\partial r_i} \sum_{n=0}^{\infty} (h(n) - r_i^n)^2 = \sum_{n=0}^{\infty} 2(h(n) - r_i^n)(-nr_i^{n-1})$$
$$= 2 \sum_{n=0}^{\infty} n(r_i^{2n-1} - r_i^{n-1} h(n)) = 2\alpha_i - 2 \sum_{n=0}^{2K} n r_i^{n-1} h(n)$$

Differentiating again at $r = r_i$ yields

$$\frac{\partial^2 V(r)}{\partial r^2}(r_i) = 2 \sum_{n=0}^{\infty} n((2n-1) r_i^{2n-2} - (n-1) r_i^{n-2} h(n))$$
$$= 2\beta_i + c_i - 2 \sum_{n=0}^{2K} n(n-1) r_i^{n-2} h(n) .$$

Conditions (A.3) become

$$H'(r_i) = \sum_{n=1}^{2K} r_i^{n-1} (nh(n)) = \alpha_i \qquad H''(r_i) = \sum_{n=1}^{2K} (n-1) r_i^{n-2} (nh(n)) = \beta_i + \frac{c_i}{2} \tag{A.4}$$

where the summation indices have been adjusted for vanishing terms. Since $h(0)$ has no effect on the conditions, it is arbitrary; we take $h(0) = 1$ since $\dot{h}(0) = 1$ for all r_i (nonzero by hypothesis).

Define a new polynomial $G(r)$ by

$$g(n-1) = nh(n), \quad n = 1, \ldots, 2K,$$
$$G(r) = \sum_{n=0}^{2K-1} g(n) r^n = H'(r).$$

Then the problem is to (1) interpolate the order $2K-1$ polynomial $G(r)$ to the K values $G(r_i) = \alpha_i, 1 \leq i \leq K$, and (2) interpolate the K derivatives of $G(r)$ to $G'(r_i) = \beta_i + c_i/2$. Since $G(r) = H'(r)$, we see that the solution is a K-point interpolation of the first and second derivatives of $H(r)$.

The conditions (A.4) become the $2K \times 2K$ matrix equation

$$\begin{pmatrix} 1 & r_1 & r_1^2 & \cdots & r_1^{2K-1} \\ 0 & 1 & 2r_1 & \cdots & (2K-1)r_1^{2K-2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & r_K & r_K^2 & \cdots & r_K^{2K-1} \\ 0 & 1 & 2r_K & \cdots & (2K-1)r_K^{2K-2} \end{pmatrix} \begin{pmatrix} g(0) \\ g(1) \\ \vdots \\ g(2K-1) \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ \beta_1 \\ \vdots \\ \alpha_K \\ \beta_K \end{pmatrix}, \tag{A.5}$$

from which (A.1) follows.

The only remaining issue is whether the matrix is nonsingular in general. While this follows from interpolation theory [19,75], an elementary proof will be given. A square matrix is nonsingular if and only if its null space equals the zero vector [151]. Thus if setting the right-hand side of (A.5) to zero gives an equation in which $g(n) \equiv 0$ is the only solution, then the matrix is always invertible. Setting the right-hand side to zero gives a system of equations corresponding to

$$G(r_i) = 0, \qquad G'(r_i) = 0, \qquad i = 1, \ldots, K, \tag{A.6}$$

where $G'(r_i) \triangleq \frac{dG(r)}{dr}(r_i)$. A solution to (A.6) can be written as

$$G(r) = P(r) \prod_{i=1}^{K} (r - r_i),$$

where $P(r)$ is a polynomial of order not exceeding $K-1$. We can also write $G(r)$ in the form $G(r) = (r - r_i)Q_i(r)$ which can be differentiated to obtain

$$G'(r) = Q_i(r) + (r - r_i)Q_i'(r)$$
$$\Rightarrow \quad G'(r_i) = Q_i(r_i) = P(r_i) \prod_{\substack{k=1 \\ k \neq i}}^{K} (r_i - r_k)$$

Since the r_i are distinct, the requirement $G'(r_i) = 0$ implies

$$P(r_i) = 0, \quad i = 1, \ldots, K .$$

But $P(r)$ is of order not greater than $K - 1$. Hence, P and therefore G must vanish identically. It then follows that a solution to (A.1) exists for all distinct $\{r_i\} \subseteq (0, 1)$. ∎

Note that Chui, Smith, and Su state in [15] that arbitrarily many local minima can exist for higher order filters.

We have constructed an order $2K$ FIR filter which gives rise to an arbitrary set of K local minima in the squared L^2 error norm $V(\hat{\theta})$ for the case of one-pole approximation. Furthermore, the curvature at these points can be set to arbitrary values. With this it is easy to prove the following.

Lemma A.2. For any set of $2K$ distinct real numbers $\{r_i\}_{i=1}^{2K}$, $0 < r_i < 1, K \geq 1$, there exists an order $4K$ polynomial $H(r)$ such that the one-pole L^2 approximation error $J_2(r)$ oscillates with local minima at $r_{2l-1}, l = 1, \ldots, K$ and local maxima at $r_{2l}, l = 1, \ldots, K$. Moreover, the curvature at these extrema can be set arbitrarily and independently.

Proof. Proceed as in the Lemma A.1 using alternating signs for the c_i. ∎

Lemma A.3. Under the conditions of Lemma A.2, at least $\lfloor K/2 \rfloor$ local maxima, $V(r_{2i}) = J_2^2(r_{2i})$ approach infinity as the alternating curvature values c_j approach infinity in magnitude, i.e.,

$$\lim_{\{|c_j|_{j=1}^{2K}\} \to \infty} V(r_i) = \infty ,$$

for at least $\lfloor K/2 \rfloor$ even values of i.

Proof. First note that by construction, $V(r) \geq 0$. Suppose that i is even and that $V(r_i)$ approaches a finite limit as the curvature at each extremum becomes infinite. Then there exists a number M sufficiently large such that with $|c_i| > M$, the second derivative $V''(r)$ changes sign at least three times as r goes from r_i to r_{i+1}, $i = 1, \ldots, 2K - 1$. (See Fig. A.1.) Let N_f equal the number of finite maxima of V. Then N_f is between 0 and K. For each finite maximum (at r_i say) between two minima, there are at least six sign changes in $V''(r)$ as r ranges from r_{i-1} to the r_{i+1}. If $V(r_i)$ goes to infinity, then the sign of $V''(r)$ need only change sign twice on the interval $[r_{i-1}, r_{i+1}]$. The total number of real zeros of $V''(r)$ on $(0, 1)$ is therefore at least $6N_f + 2(K - N_f) - \mu$, where $\mu = 1$ if $V(r_{2K})$ has an infinite limit, and $\mu = 3$ otherwise.

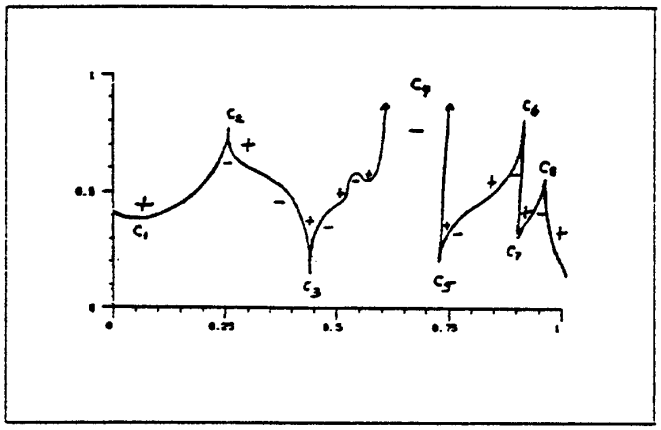

Figure A.1. Illustration of the behavior of a polynomial having finite values at extrema with curvature of infinite magnitude and alternating sign.

By definition,

$$\begin{aligned} V(r) &= \sum_{n=0}^{\infty} \big(h(n) - r^n\big)^2 \\ &= \sum_{n=0}^{\infty} h^2(n) - 2h_n r^n + r^{2n} \\ &= H(1) + P(r) + \frac{1}{1-r^2}, \end{aligned} \tag{A.7}$$

where

$$P(r) = -2 \sum_{n=0}^{4K} h_n r^n$$

is a polynomial of degree at most $4K$ in r. Twice differentiating with respect to r gives

$$V''(r) = \frac{(1-r^2)^3 P''(r) + 6r^2 + 2}{(1-r^2)^3}$$

The term $1 - r^2$ does not change sign on $(0, 1)$. Consequently, $V''(r)$ can have at most $4(K + 1)$ sign changes on $(0, 1)$. Therefore,

$$6N_f + 2(K - N_f) - \mu \leq 4(K + 1) \;\Rightarrow\; N_f \leq \left\lfloor \frac{K+1}{2} \right\rfloor.$$

Hence there must be at least $\lfloor K/2 \rfloor$ local maxima of $V(r)$ which approach infinity as the $2K$ extrema curvatures approach infinity in magnitude. ∎

Lemma A.4. Under the conditions of Lemma A.2, for $K \geq 2$, the curvature magnitudes at the extrema of $V(r)$ can be made to grow in such a way that the local minima remain fixed.

Proof. A Taylor expansion of (A.2) about $r = r_i$ gives

$$\begin{aligned} V(r) &= V(r_i) + V'(r_i)(r - r_i) + \frac{1}{2}V''(r_i)(r - r_i)^2 + \frac{1}{3!}V'''(r_i)(r - r_i)^3 + \cdots \\ &= V(r_i) + \frac{c_i}{2}(r - r_i)^2 + \frac{1}{3!}V'''(r_i)(r - r_i)^3 + \cdots . \end{aligned}$$

Equation (A.7) shows that the singularities of $V(r)$ are at ± 1. Therefore the Taylor expansion is valid throughout $(0, 1)$. If c_i decreases, $V(r)$ decreases everywhere but at $r = r_i$. At each local maximum (r_i for i even), c_i is negative. If the local-minima curvatures are held constant while the local-maxima curvatures increase in magnitude (decrease), then $V(r)$ decreases for all r. Similarly, when only the minima-curvatures increase, $V(r)$ "floats up." Our problem can be solved by defining the curvatures at the local minima in terms of those at the maxima so that the minima are held fixed. The total differential of V at r with respect to the curvature differential $(dc_1, dc_2, \ldots, dc_{2K})$ is given by

$$dV(r) = \frac{\partial V(r)}{\partial c_1}dc_1 + \cdots + \frac{\partial V(r)}{\partial c_{2K}}dc_{2K} = \frac{1}{2}\sum_{i=1}^{2K}(r - r_i)^2 dc_i .$$

To simplify notation, let

$$\begin{aligned} r_h(i) &\triangleq r_{2i}, \qquad & r_v(i) &\triangleq r_{2i-1}, \\ c_h(i) &\triangleq c_{2i}, \qquad & c_v(i) &\triangleq c_{2i-1}, \end{aligned}$$

for $i = 1, 2, \ldots, K$. The subscript "h" denotes "hill", and "v" denotes "valley". Then $c_h(i) < 0$ and $c_v(i) > 0$. The total differential is now

$$dV(r) = \frac{1}{2}\sum_{i=1}^{K}(r - r_h(i))^2 dc_h(i) + \frac{1}{2}\sum_{i=1}^{K}(r - r_v(i))^2 dc_v(i) .$$

In order that the local minima remain fixed as curvature magnitudes increase, we require

$$dV(r_v(j)) = 0, \quad j = 1, \ldots, K ,$$

which implies

$$\sum_{i=1}^{K}(r_v(j) - r_v(i))^2 dc_v(i) = -\sum_{i=1}^{K}(r_v(j) - r_h(i))^2 dc_h(i) \triangleq d\gamma(j) ,$$

or in matrix form,

$$A(K)d\underline{C_v}(K) = d\underline{\Gamma}(K),$$

where

$$\underline{C_v}(K) \triangleq \begin{pmatrix} c_v(1) \\ c_v(2) \\ \vdots \\ c_v(K) \end{pmatrix}, \qquad \underline{\Gamma}(K) \triangleq \begin{pmatrix} \gamma(1) \\ \gamma(2) \\ \vdots \\ \gamma(K) \end{pmatrix},$$

and

$$A(K) \triangleq \begin{pmatrix} 0 & (r_v(1)-r_v(2))^2 & (r_v(1)-r_v(3))^2 & \cdots & (r_v(1)-r_v(K))^2 \\ (r_v(2)-r_v(1))^2 & 0 & (r_v(2)-r_v(3))^2 & \cdots & (r_v(2)-r_v(K))^2 \\ (r_v(3)-r_v(1))^2 & (r_v(3)-r_v(2))^2 & 0 & \cdots & (r_v(3)-r_v(K))^2 \\ \vdots & \vdots & \vdots & & \vdots \\ (r_v(K)-r_v(1))^2 & (r_v(K)-r_v(2))^2 & (r_v(K)-r_v(3))^2 & \cdots & 0 \end{pmatrix}.$$

We must show

(1) $|A(K)| \neq 0$,

(2) $dc_v(i) > 0, \qquad i = 1, \ldots, K$,

For $K = 2$, we have

$$A^{-1}(2) = \frac{1}{(r_v(1)-r_v(2))^4}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

for which conditions (1) and (2) are satisfied. Suppose these conditions are satisfied for some $K \geq 2$. Then note that

$$A(K+1) = \begin{pmatrix} A(K) & a(K) \\ a(K)^T & 0 \end{pmatrix},$$

where

$$a(K)^T \triangleq \left((r_v(K+1)-r_v(1))^2, \ldots (r_v(K+1)-r_v(K))^2 \right).$$

Let

$$d\underline{C_v}(K+1) \triangleq \begin{pmatrix} d\underline{C_v}(K) \\ d\mu \end{pmatrix},$$

where $d\mu > 0$ is an arbitrary positive differential. Then

$$A(K+1)d\underline{C_v}(K+1) = \begin{pmatrix} d\underline{\Gamma}(K) + a(K)d\mu \\ a(K)^T d\underline{C_v}(K) \end{pmatrix} \triangleq d\underline{\Gamma}(K+1).$$

Since every element in $d\underline{\Gamma}(k)$, $a(K)$, and $d\underline{C_v}(K)$ is positive, it follows that $d\underline{\Gamma}(K+1)$ is properly defined. Also, every element of $d\underline{C_v}(K+1)$ is positive by construction. Therefore,

it is possible to increase curvature at all extrema of $V(r)$ without perturbing the values of V at the minima. ∎

Lemma A.5. Under the conditions of Lemma A.2, for $K \geq 2$, the *relative oscillation amplitude*, defined by

$$A_i(V) \triangleq \left| \frac{V(r_{i+1}) - V(r_i)}{V(r_i)} \right|, \qquad i = 1, \ldots, 2K \tag{A.8}$$

can be made arbitrarily large for $\lfloor K/2 \rfloor$ even values of i. In particular, for any number M, there exist curvature values $\{c_j\}_1^{2K}$ such that $A_i(V) > M$ for $\lfloor K/2 \rfloor$ even values of i.

Proof. This is a straightforward consequence of Lemmas A.3 and A.4. ∎

Lemma A.6. Let $V_N(r)$ denote the one-pole squared L^2 approximation error norm defined on a discrete-frequency grid of size N, and let $V(r) = J_2^2(r)$ denote the usual continuous-frequency error norm. Then for any $\epsilon, \delta > 0$, there exists $M(\epsilon, \delta)$ such that

$$|V_N(r) - V(r)| < \epsilon$$

for all $N > M(\epsilon, \delta)$ and for all $|r| \leq R \triangleq 1 - \delta$.

Proof. By direct computation,

$$\begin{aligned} V_N(r) - V(r) &= \sum_{n=0}^{\infty} \left(h(n) - \frac{r^n}{1 - r^N} \right)^2 - (h(n) - r^n)^2 \\ &= \frac{2r^N - r^{2N}}{(1 - r^2)(1 - r^N)^2} + \frac{2r^N}{r^N - 1} \sum_{n=0}^{\infty} h(n) r^n . \end{aligned}$$

Since $|H(e^{j\omega})| < B_H$ for some finite B_H, we have

$$|h(n)| = \left| \int_{-\pi}^{\pi} H(e^{j\omega}) e^{j\omega n} \frac{d\omega}{2\pi} \right| \leq B_H .$$

Consequently,

$$\begin{aligned} |V_N(r) - V(r)| &\leq \left| \frac{2r^N - r^{2N}}{(1 - r^2)(1 - r^N)^2} \right| + 2 \left| \frac{r^N}{r^N - 1} \right| \frac{B_H}{1 - |r|} \\ &\leq \frac{3R^N}{\delta^3} + 2 \frac{R^N B_H}{\delta^2} . \end{aligned}$$

Thus we may set

$$M(\epsilon, \delta) = \ln \left(\frac{\epsilon \delta^3}{2\delta B_H + 3} \right) \Big/ \ln(1 - \delta),$$

and for $N > M(\epsilon, \delta)$, the two norms differ by less than ϵ over the whole range of r. ∎

We can now state the main theorem for which the above lemmas were developed.

Theorem A.7. Let K be a positive integer. Then for any discrete-frequency norm, there exists an order $8K$ FIR filter $H(z)$ and a frequency-grid size N such that the one-pole approximation-error norm

$$J(r) = \left\| H(e^{j\omega_k}) - \frac{1}{1 - re^{-j\omega_k}} \right\|,$$

has K local minima.

Proof. Lemma A.2 constructs an oscillating L^2 error measure, and Lemma A.5 shows that the oscillation can be given arbitrarily wide excursions over half the extrema. Lemma A.6 then allows that Lemma A.5 apply over a sufficiently dense discrete-frequency grid. On a discrete frequency grid, the *norm equivalence theorem* [192] applies which states for each norm there exist positive constants g and G such that

$$gJ_2(r) \leq J(r) \leq GJ_2(r),$$

for all r. Thus J is confined between gJ_2 and GJ_2 as shown in Fig. A.2. If $J_2(r)$ is made to oscillate with sufficiently large amplitude (A.8), then $J(r)$ must oscillate, i.e., if $gJ_2(r_*) > GJ_2(r^*)$ holds, where r_* is an arbitrary constructed maximizer of J_2 and r^* is the greater of its two adjacent constructed minimizers, then $J(r)$ must have at least one local minimum between each pair of maxima constructed for J_2. By Lemma A.5, there exist curvature values c_i such that

$$J_2(r_*) - J_2(r^*) > \frac{G-g}{g} J_2(r^*) \;\Rightarrow\; J_2(r_*) > \frac{G}{g} J_2(r^*) \;\Rightarrow\; gJ_2(r_*) > GJ_2(r^*),$$

as needed to force oscillation of $J(r)$. ∎

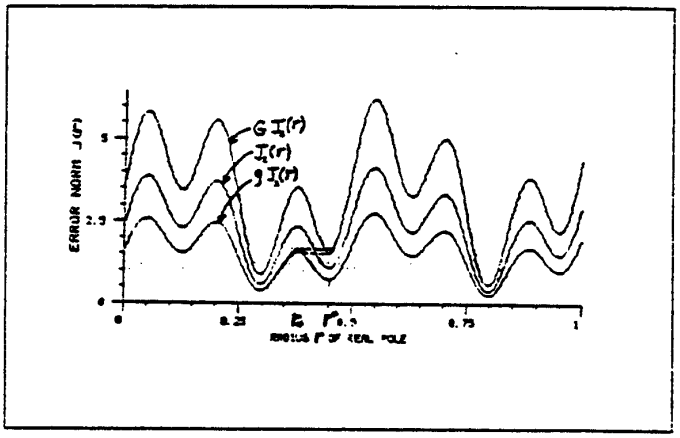

Figure A.2. Region containing an arbitrary norm bounded by scaled L^2 norms (norm equivalence theorem).

Corollary A.8. Problem $\dot{H}^*$ is not concave under any discrete-frequency norm, and there is no upper bound on the number of locally best approximations. Consequently, no gradient-based method can be guaranteed to converge for a filter design problem with one or more poles in the approximation.

While the result was shown only for discrete-frequency norms, it is true also for any continuous-frequency norm which is a uniformly continuous limit of discrete norms. For example, all L^p norms fall in this category.

Appendix B. Optimality of the CF Algorithm

The purpose of this appendix is to prove that the CF method provides an optimal approximation under the Hankel norm of the impulse-response error under certain conditions. The theory is an elementary version of the more sophisticated arguments in Hankel-norm theory [2]. An attempt has been made to simplify as much as possible the development needed to cover the case of real digital filter design.

Given a real, causal, finite-impulse-response sequence $\{h_K(n)\}_0^K$, with $h_K(0), h_K(K) \neq 0$, corresponding to the desired transfer function

$$H_K(z) \triangleq \sum_{n=0}^{K} h_K(n) z^{-n},$$

the CF method computes an approximation to $H_K(z)$ on the unit circle $|z| = 1$ by a rational transfer function

$$\hat{H}(z) \triangleq \frac{B(z)}{A(z)} \triangleq \frac{\sum_{k=0}^{M} b_k z^{-k}}{\sum_{k=0}^{N} a_k z^{-k}}, \tag{B.1}$$

with all poles inside unit circle, and normalized by $a_0 = 1$. We denote by $\mathcal{H}_{M,N}$ the set of all such functions.

As an intermediate step, the CF algorithm determines the best Chebyshev approximation out of the larger class $\tilde{\mathcal{H}}_{M,N}$ of functions which are of the form

$$\hat{H}_\infty(z) \triangleq \frac{\tilde{B}(z)}{A(z)} \triangleq \frac{\sum_{k=-\infty}^{M} b_k z^{-k}}{\sum_{k=0}^{N} a_k z^{-k}} \triangleq \sum_{n=-\infty}^{\infty} \hat{h}_\infty(n) z^{-n}$$

(with $a_0 = 1$), where the zeros of $z^N A(z)$ still lie inside the unit circle.

When $M \geq N - 1$, the CF approximation is simply the causal projection of $\hat{H}_\infty(z)$,

$$\hat{H}(z) \triangleq \sum_{n=0}^{\infty} \hat{h}_\infty(n) z^{-n}.$$

We will show that in this case $\hat{H}$ is optimal with respect to the Hankel norm,

Define

$$\nu \triangleq M - N + 1 \geq 0.$$

The $\nu \geq 0$ restriction is only necessary in the final step below. The *Hankel matrix* corresponding to H_K and ν is defined as

$$\mathbf{H}_{\nu,K} \triangleq \begin{pmatrix} h_K(\nu) & h_K(\nu+1) & \cdots & h_K(K) \\ h_K(\nu+1) & & & 0 \\ \vdots & & & \vdots \\ \vdots & & h_K(K) & \vdots \\ \vdots & & & \vdots \\ h_K(K) & 0 & \cdots & 0 \end{pmatrix}.$$

Let $\{\lambda_n\}_0^{K-\nu}$ denote the eigenvalues of $\mathbf{H}_{\nu,K}$ ordered according to decreasing magnitude $|\lambda_0| \geq |\lambda_1| \geq \ldots \geq |\lambda_{K-\nu}|$, and for each λ_n let $\underline{v}_n = (v_n(0), \ldots, v_n(K-\nu))^T$ denote any corresponding eigenvector,

$$\mathbf{H}_{\nu,K}\underline{v}_n = \lambda_n \underline{v}_n$$

$$\Rightarrow \quad \sum_{i=0}^{K-\nu} h_K(i+k+\nu)v_n(i) = \lambda_n v_n(k), \qquad k = 0,1,\ldots,K-\nu. \tag{B.2}$$

The eigenvalue/eigenvector pair of interest in the CF method is the N^{th}, where N is the number of poles desired in the approximation. It will turn out that λ_N is the Hankel-norm of approximation error associated with the optimum N-pole approximation. We assume $\lambda_{N-1} > \lambda_N$ since otherwise poles can be deleted (N decreased) without increasing the approximation error.

Now take the z-transform of both sides of (B.2), with $n = N$, to obtain

$$\sum_{k=0}^{K-\nu}\sum_{n=0}^{K-\nu} h_K(n+k+\nu)v_N(n)z^{-k} = \lambda_N \sum_{k=0}^{K-\nu} v_N(k)z^{-k} \triangleq \lambda_N V_N(z).$$

The change of summation index $l = n+k+\nu$ yields

$$\begin{aligned}
\lambda_N V_N(z) &= \sum_{n=0}^{K-\nu} v_N(n) \sum_{k=0}^{K-\nu} h_K(n+k+\nu)z^{-k} \\
&= \sum_{n=0}^{K-\nu} v_N(n)z^{n+\nu} \sum_{l=n+\nu}^{n+K} h_K(l)z^{-l} \\
&= \sum_{n=0}^{K-\nu} v_N(n)z^{n+\nu}\left(\sum_{l=0}^{K} h_K(l)z^{-l} - \sum_{l=0}^{n+\nu-1} h_K(l)z^{-l}\right) \\
&\triangleq z^{\nu}V_N(z^{-1})H_K(z) - Q_N(z^{-1}),
\end{aligned}$$

where

$$\begin{aligned} Q_N(z^{-1}) &\triangleq z^{\nu} \sum_{n=0}^{K-\nu} \sum_{l=0}^{n+\nu-1} v_N(n) h_K(l) z^{n-l} \\ &= \sum_{i=1}^{K} \left(\sum_{j=1}^{K} v_N(j-\nu) h_K(j-i) \right) z^i \\ &\triangleq \sum_{i=1}^{K} q_N(i) z^i , \end{aligned}$$

with vacuous sums (e.g. $\sum_0^{-1}$) and out-of-range subscripts (e.g. $v_N(-1)$) defined to be zero.

We have decomposed $H_K(z)$ as

$$H_K(z) = z^{-\nu} \frac{Q_N(z^{-1})}{V_N(z^{-1})} + \lambda_N z^{-\nu} \frac{V_N(z)}{V_N(z^{-1})} . \tag{B.3}$$

To eliminate positive powers of z from the denominator of (B.3), we define

$$\hat{V}_N(z) \triangleq z^{-(K-\nu)} V_N(z^{-1}) = v_N(K-\nu) + v_N(K-\nu-1) z^{-1} + \cdots + v_N(0) z^{-(K-\nu)}$$
$$\hat{Q}_N(z) \triangleq z^{-K} Q_N(z^{-1}) = q_N(K) + q_N(K-1) z^{-1} + \cdots + q_N(1) z^{-(K-1)} .$$

With these polynomials, (B.3) becomes

$$H_K(z) = \frac{\hat{Q}_N(z)}{\hat{V}_N(z)} + \lambda_N z^{-K} \frac{V_N(z)}{\hat{V}_N(z)} \tag{B.4}$$

The first term in the decomposition (B.4) of $H_K(z)$ is an unstable approximation which we denote by

$$\hat{H}_\infty(z) \triangleq \frac{\hat{Q}_N(z)}{\hat{V}_N(z)} ,$$

and the approximation error, $\lambda_N z^{-K} V_N(z)/\hat{V}_N(z)$, is an allpass filter with magnitude λ_N on the unit circle.

Lemma B.1. $\hat{H}_\infty(z) \in \tilde{\mathcal{H}}_{M,N}$.

Proof. By a theorem of Takagi [67], we know that $\hat{V}_N(z)$ has precisely N roots inside the unit circle when $\lambda_{N-1} > \lambda_N$. Therefore, it can be factored as

$$\hat{V}_N(z) = \hat{V}_i(z) \hat{V}_o(z) \tag{B.5}$$

where the N roots of $\hat{V}_i(z)$ are all inside the unit circle, and the $K-\nu-N$ roots of $\hat{V}_o(z)$ are all outside the unit circle. The term $1/\hat{V}_o(z)$ may be written as

$$\begin{aligned}\frac{1}{\hat{V}_o(z)} &= \frac{1}{\hat{v}_o(0)+\cdots+\hat{v}_o(K-\nu-N)z^{-(K-\nu-N)}} \\ &= \frac{z^{K-\nu-N}}{\hat{v}_o(0)z^{K-\nu-N}+\cdots+\hat{v}_o(K-\nu-N)} \\ &\triangleq z^{K-\nu-N}\tilde{V}_o(z) \\ &\triangleq z^{K-\nu-N}\big(\tilde{v}_o(0)+\tilde{v}_o(1)z+\tilde{v}_o(2)z^2+\cdots\big).\end{aligned}$$

Therefore, the numerator of $\hat{H}_\infty(z) = z^{K-\nu-N}\hat{Q}_N(z)\tilde{V}_o(z)/\hat{V}_i(z)$ has highest negative power of z equal to $(K-1)-(K-\nu-N) = M$. Thus we conclude $\hat{H}_\infty(z)$ belongs to $\tilde{\mathcal{H}}_{M,N}$. ∎

The CF approximation is defined as the causal part of $\hat{H}_\infty(z)$ as expanded in a Laurent series convergent on the unit circle $|z| = 1$. Let the operator which projects a transfer function $H(z)$ onto its causal part be denoted by $C\{H(z)\}$. Then the CF approximation can be expressed as

$$\hat{H}(z) \triangleq C\{\hat{H}_\infty(z)\} = C\{\sum_{n=-\infty}^{\infty}\hat{h}_\infty(n)z^{-n}\} \triangleq \sum_{n=0}^{\infty}\hat{h}_\infty(n)z^{-n} \tag{B.6}$$

The CF approximation impulse response is then

$$\hat{h}(n) \triangleq \begin{cases}\hat{h}_\infty(n), & n = 0,1,\ldots \\ 0, & n<0\end{cases}$$

Definition B.2. The *Hankel norm* of a stable filter $H(z) \leftrightarrow h(n)$ is defined as the spectral norm of the associated Hankel matrix. That is, if H is the matrix whose $(i,j)^{th}$ element is $h(i+j)$ $(i,j = 0,1,2,\ldots)$, and $\underline{X} = (x(0),x(1),\ldots)$ is a vector, then the Hankel norm of H is equal to

$$\|H\|_H \triangleq \max_{\underline{X}}\frac{\underline{X}^T\mathrm{H}\underline{X}}{\underline{X}^T\underline{X}},$$

where the maximum is attained by virtue of the stability restriction. Note that a noncausal component of $h(n)$ does not affect the norm.

To show that $\hat{H}(z)$ is an optimum approximation under the Hankel norm, we need the following.

Lemma B.3. The function $\hat{H}_\infty(z)$ is the unique optimum Chebyshev approximation from the class of rational functions $\tilde{\mathcal{H}}_{M,N}$.

Proof. To show that $\hat{H}_\infty(z)$ is an optimum Chebyshev approximation, suppose there exists $\hat{H}'_\infty(z) \in \tilde{\mathcal{H}}_{M,N}$ which is a better approximation in the Chebyshev sense, i.e.,

$$\left\| H_K(e^{j\omega}) - \hat{H}'_\infty(e^{j\omega}) \right\|_\infty < \left\| H_K(e^{j\omega}) - \hat{H}_\infty(e^{j\omega}) \right\|_\infty = \lambda_N .$$

Then we have, by use of inequalities proved in [43,32]

$$\begin{aligned} \lambda_N &> \left\| H_K(e^{j\omega}) - \hat{H}'_\infty(e^{j\omega}) \right\|_\infty \geq \left\| H_K(e^{j\omega}) - \hat{H}'_\infty(e^{j\omega}) \right\|_H \\ &= \left\| H_K(e^{j\omega}) - C\{\hat{H}'_\infty(e^{j\omega})\} \right\|_H \geq \lambda_N , \end{aligned}$$

which provides a contradiction.

If λ_N is an isolated eigenvalue, then $\underline{v}_N$ is unique up to a scalar which does not affect $\hat{H}_\infty(z)$. Consequently, $\hat{H}_\infty(z)$ is unique in this instance. In [2] it is shown that uniqueness holds even in the case of multiple eigenvalues $\lambda_n, n = N, N+1, \ldots, N+\kappa$. Essentially what happens in this case is that all eigenvectors which satisfy (B.2) have a unique ratio of the form $\underline{v}(z)/\underline{v}(z^{-1})$ due to pole-zero cancellations. In [2], the fact that there are no other forms for an optimal approximation from $\tilde{\mathcal{H}}_{M,N}$ is also established. ∎

It is equally immediate to show that the causal part of an optimum Chebyshev approximation is an optimum Hankel-norm approximation.

Lemma B.4. *$\hat{H}(z)$ is the unique optimum Hankel-norm approximation to $H_K(z)$.*

Proof. Using the same argument as in the previous lemma, we have that

$$\begin{aligned} \lambda_N &= \left\| H_K(e^{j\omega}) - \hat{H}_\infty(e^{j\omega}) \right\|_\infty \geq \left\| H_K(e^{j\omega}) - \hat{H}_\infty(e^{j\omega}) \right\|_H \\ &= \left\| H_K(e^{j\omega}) - \hat{H}(e^{j\omega}) \right\|_H \geq \lambda_N . \end{aligned}$$

Therefore equality must hold, and in particular,

$$\left\| H_K(e^{j\omega}) - \hat{H}(e^{j\omega}) \right\|_H = \lambda_N .$$

This establishes the fact that $\hat{H}(z)$ is an optimum Hankel approximant.

Uniqueness follows from the uniqueness of the rank N Hankel matrix which optimally approximates $\mathbf{H}_{\nu,K}$ under the spectral matrix norm [2]. ∎

It only remains to be shown that $\hat{H}(z)$ is a member of $\mathcal{H}_{M,N}$. For this it suffices to show that $\hat{H}(z)$ is of the form (B.1) with all its poles inside the unit circle.

Lemma B.5. *$\hat{H}(z)$ as defined by (B.6) is a member of $\mathcal{H}_{M,N}$.*

Proof. $\hat{H}(z)$ is defined as the causal part of

$$\hat{H}_\infty(z) \triangleq \frac{\hat{Q}_N(z)}{\hat{V}_N(z)}.$$

The factorization (B.5) leads to the additive decomposition

$$\hat{H}_\infty(z) = F(z) + \frac{R_i(z)}{\hat{V}_i(z)} + \frac{R_o(z)}{\hat{V}_o(z)},$$

where

$$\begin{aligned} R_i(z) &= r_i(0) + r_i(1)z^{-1} + \cdots + r_i(N-1)z^{-(N-1)} \\ R_o(z) &= r_o(0) + r_o(1)z^{-1} + \cdots + r_o((K-N)-1)z^{-((K-N)-1)} \\ F(z) &= f(0) + f(1)z^{-1} + \cdots + f(\nu-1)z^{-(\nu-1)}. \end{aligned}$$

If $\nu \leq 0$, then $F(z) \equiv 0$.

since all the roots of $\hat{V}_o(z)$ lie outside the unit circle, The term $R_o(z)/\hat{V}_o(z)$ may be expanded in a Taylor series about $z = 0$ which converges in the unit disk.

$$\begin{aligned} H_-(z) \triangleq \frac{R_o(z)}{\hat{V}_o(z)} &= z\frac{r_o(0)z^{(K-\nu-N)-1)} + \cdots + r_o((K-\nu-N)-1)}{\hat{v}_o(0)z^{K-\nu-N} + \cdots + \hat{v}_o(K-\nu-N)} \\ &= h_-(1)z + h_-(2)z^2 + \cdots. \end{aligned}$$

Therefore, the causal projection of H_- is zero. Analogous considerations show that $F(z)$ and $R_i(z)/\hat{V}_i(z)$ are invariant under causal projection. Consequently,

$$\hat{H}(z) = F(z) + \frac{R_i(z)}{\hat{V}_i(z)}.$$

For $\nu = 0$ ($M = N-1$), $\hat{H}(z)$ is obviously in $\mathcal{H}_{N-1,N}$, since $F(z) \equiv 0$ and the degree of $R_i(z)$ is $N-1$. If $\nu < 0$, corresponding to $M < N-1$, then $\hat{H}_\infty(z) \in \mathcal{H}_{N-1,N} \neq \mathcal{H}_{M,N}$ (in general). On the other hand, for $\nu > 0$ we have

$$\hat{H}(z) = \frac{F(z)\hat{V}_i(z) + R_i(z)}{\hat{V}_i(z)}.$$

Since the degree of $F(z)$ is $\nu - 1$, and the degree of $\hat{V}_i(z)$ is N, we have a maximum numerator degree of $\max\{\nu - 1 + N, N-1\} = \max\{M, N-1\} = M$.

Thus, we conclude

$$\hat{H}(z) \in \mathcal{H}_{M,N}, \qquad M \geq N-1.$$

The previous discussion allows us to state the following.

Theorem B.8. The CF method finds the unique optimum M-zero, N-pole approximation $\hat{H}(z)$ to an arbitrary $H_K(z)$ $(M \geq N-1)$, minimizing the Hankel norm of the error $\| H_K(z) - \hat{H}(z) \|_H$.

Appendix C. Functions Positive Real in the Outer Disk

Any passive driving-point impedance, such as the impedance of a violin bridge, is positive real. It was found in §1.8.3 of Chapter 1 that approximation of magnitude-squared frequency response requires an estimated polynomial to be positive real. Moreover, some convergence proofs for system identification algorithms, particulary ELS (cf. Chapter 2), rely on the condition that a particular transfer function be positive real [102,107,127]. Positive real functions have been extensively studied in the continuous-time case in the context of network synthesis [165,203]. Very little seems to be available on the image of the many properties of positive real functions under translation to discrete time. The purpose of this Appendix is to record facts derived about positive real transfer functions for discrete-time linear systems.

Definition C.1. A complex valued function of a complex variable $f(z)$ is said to be *positive real* (PR) if

1) z *real* $\Rightarrow$ $f(z)$ *real*

2) $|z| \geq 1 \Rightarrow Re\{f(z)\} \geq 0$

We now specialize to the subset of functions $f(z)$ representable as a ratio of finite-order polynomials in z. This class of "rational" functions is the set of all transfer functions of finite-order time-invariant linear systems, and we write $H(z)$ to denote a member of this class. We use the convention that stable, minimum phase systems are analytic and nonzero in the strict outer disk.* Condition 1) implies that for $H(z)$ to be PR, the polynomial coefficients must be real, and therefore complex poles and zeros must exist in conjugate pairs. We assume from this point on that $H(z) \not\equiv 0$ satisfies 1). From 2) we derive the facts below.

Theorem C.2. A real rational function $H(z)$ is PR iff $|z| \geq 1 \Rightarrow |\angle H(z)| \leq \frac{\pi}{2}$.

Proof. Expressing $H(z)$ in polar form gives

$$\begin{aligned} Re\{H(z)\} &= Re\{|H(z)|e^{j\angle H(z)}\} = |H(z)|\cos(\angle H(z)) \\ &\geq 0 \quad \Leftrightarrow \quad |\angle H(z)| \leq \frac{\pi}{2}. \end{aligned} \tag{C.1}$$

since the zeros of $H(z)$ are isolated. ∎

Theorem C.3. $H(z)$ is PR iff $1/H(z)$ is PR.

* The strict *outer disk* is defined as the region $|z| > 1$ in the extended complex plane.

Proof. Assuming $H(z)$ is PR, we have by Thm. C.2,

$$|\angle H^{-1}(z)| = |-\angle H(z)| = |\angle H(z)| \leq \frac{\pi}{2}, \quad |z| \geq 1 .$$

∎

Theorem C.4. A PR function $H(z)$ is analytic and nonzero in the strict outer disk.

Proof. (By contradiction).

Without loss of generality, we treat only n^{th} order polynomials

$$\alpha_0 z^n + \alpha_1 z^{n-1} + \cdots + \alpha_{n-1} z + \alpha_n$$

which are nondegenerate in the sense that $\alpha_0, \alpha_n \neq 0$. Since facts about $\alpha_0 H(z)$ are readily deduced from facts about $H(z)$, we set $\alpha_0 = 1$ at no great loss.

The general (normalized) causal, finite-order, linear, time-invariant transfer function may be written

$$\begin{aligned} H(z) &= z^{-\nu} \frac{b(z)}{a(z)} \\ &= z^{-\nu} \frac{1 + b_1 z^{-1} + \cdots + b_M z^{-M}}{1 + a_1 z^{-1} + \cdots + a_N z^{-N}} \\ &= z^{-\nu} \frac{\prod_{i=1}^{M} (1 - q_i z^{-1})}{\prod_{i=1}^{N} (1 - p_i z^{-1})} \\ &= z^{-\nu} \sum_{i=1}^{N_d} \sum_{j=1}^{\mu_i} \frac{z K_{i,j}}{(z - p_i)^j}, \qquad \nu \geq 0, \end{aligned} \tag{C.2}$$

where N_d is the number of distinct poles, each of multiplicity μ_i, and

$$\sum_{i=1}^{N_d} \mu_i = \max\{N, M\} . \tag{C.3}$$

Suppose there is a pole of multiplicity m outside the unit circle. Without loss of generality, we may set $\mu_1 = m$, and $p_1 = R e^{j\varphi}$ with $R > 1$. Then for z near p_1, we have

$$\begin{aligned} z^{\nu} H(z) &= \frac{z K_{1,m}}{(z - R e^{j\varphi})^m} + \frac{z K_{1,m-1}}{(z - R e^{j\varphi})^{m-1}} + \cdots \\ &\approx \frac{z K_{1,m}}{(z - R e^{j\varphi})^m} . \end{aligned}$$

Consider the circular neighborhood of radius ρ described by $z = R e^{j\varphi} + \rho e^{j\psi}$, $-\pi \leq \psi < \pi$. Since $R > 1$ we may choose $\rho < R - 1$ so that all points z in this neighborhood lie

outside the unit circle. If we write the residue of the factor $(z - R\,e^{j\varphi})^m$ in polar form as $K_{1,m} = C\,e^{j\xi}$, then we have, for sufficiently small ρ,

$$z^\nu H(z) \approx \frac{K_{1,m} R\,e^{j\varphi}}{(z - R\,e^{j\varphi})^m} = \frac{K_{1,m} R\,e^{j\varphi}}{\rho^m\,e^{jm\psi}} = \frac{C\,R}{\rho^m}\,e^{j(\varphi+\xi-m\psi)}. \tag{C.4}$$

Therefore, approaching the pole $R\,e^{j\varphi}$ at an angle ψ gives

$$\lim_{\rho\to 0}\left|\angle H(R\,e^{j\varphi} + \rho\,e^{j\psi})\right| = |\varphi(1-\nu) + \xi - m\psi|, \quad -\pi \le \psi < \pi \tag{C.5}$$

which cannot be confined to satisfy Thm. C.2 regardless of the value of the residue angle ξ, or the pole angle φ (m cannot be zero by hypothesis). We thus conclude that a PR function $H(z)$ can have no poles in the outer disk. By Thm. C.3, we conclude that positive real functions must be minimum phase. ∎

Corollary C.5. In equation (C.2), $\nu = 0$.

Proof. As $|z| \to \infty$, $H(z) \to z^{-\nu} \;\Rightarrow\; |\angle H(z)| \to |\nu \angle z| \le \nu\pi$. Since z can be chosen such that equality holds, Thm. C.2 implies $\nu = 0$. ∎

Corollary C.6. The log-magnitude of a PR function has zero mean on the unit circle.

Proof. This is a general property of stable, minimum-phase transfer functions which follows immediately from the *argument principle* [183,150]. ∎

Corollary C.7. A rational PR function has an equal number of poles and zeros all of which are in the unit disk.

Proof. This really a convention for numbering poles and zeros. In (C.2), we have $\nu = 0$, and all poles and zeros inside the unit disk. Now, if $M > N$ then we have $M - N$ extra poles at $z = 0$ induced by the numerator. If $M < N$, then $N - M$ zeros at the origin appear from the denominator. ∎

Corollary C.8. Every pole on the unit circle of a positive real function must be simple with a real and positive residue.

Proof. We repeat the argument of Thm. C.4 using a semicircular neighborhood of radius ρ about the point $p_1 = e^{j\varphi}$ to obtain

$$\lim_{\rho\to 0}\left|\angle H(e^{j\varphi} + \rho e^{j\psi})\right| = |\varphi + \xi - m\psi|, \qquad \varphi - \frac{\pi}{2} \le \psi \le \varphi + \frac{\pi}{2}. \tag{C.6}$$

In order to have $|\angle H(z)| \le \pi/2$ near this pole, it is necessary that $m = 1$ and $\xi = 0$. ∎

Corollary C.9. If $H(z)$ is PR with a zero at $z = q_1 = e^{j\varphi}$, then

$$H'(z) \triangleq \frac{H(z)}{(1 - q_1 z^{-1})}$$

must satisfy

$$\begin{aligned} H'(q_1) &\neq 0 \\ \angle H'(q_1) &= 0 \end{aligned} \tag{C.7}$$

Proof. We may repeat the above for $1/H(z)$. ∎

Theorem C.10. Every PR function $H(z)$ has a causal inverse z-transform $h(n)$.

Proof. This follows immediately from analyticity in the outer disk [195, pp. 30-36]. However, we may give a more concrete proof as follows. Suppose $h(n)$ is non-causal. Then $\exists k > 0$ such that $h(-k) \neq 0$. We have,

$$H(z) \triangleq \sum_{n=-\infty}^{\infty} h(n)\, z^{-n} = h(-k)z^k + \sum_{n \neq -k} h(n)\, z^{-n}.$$

Hence, $H(z)$ has a pole of order k at infinity and cannot be PR by Thm. C.4. ∎

Theorem C.11. $H(z)$ is PR iff it is analytic for $|z| > 1$, poles on the unit circle are simple with real and positive residues, and $Re\{H(e^{j\theta})\} \geq 0$ for $0 \leq \theta \leq \pi$.

Proof. If $H(z)$ is positive real, the conditions stated hold by virtue of Thm. C.4 and the definition of positive real. We prove the converse:

Since $h(n)$ real $\Rightarrow$ $Re\{H(e^{j\theta})\}$ even in θ, nonnegativity on the upper semicircle implies nonnegativity over the entire circle.

Next, since the function e^z is analytic everywhere except at $z = \infty$, it follows that $f(z) = e^{-H(z)}$ is analytic wherever $H(z)$ is finite. There are no poles of $H(z)$ outside the unit circle due to the analyticity assumption, and poles on the unit circle have real and positive residues. Referring again to the limiting form C.4 of $H(z)$ near a pole on the unit circle at $e^{j\varphi}$, we see that

$$\begin{aligned} H(e^{j\varphi} + \rho\, e^{j\psi}) &\underset{\rho \to 0}{\to} \frac{C}{\rho}\, e^{j(\varphi - \psi)}, \qquad \varphi - \frac{\pi}{2} \leq \psi \leq \varphi + \frac{\pi}{2} \\ &\triangleq \frac{C}{\rho}\, e^{j\theta}, \qquad \theta \triangleq \varphi - \psi, \quad -\frac{\pi}{2} \leq \theta \leq \frac{\pi}{2} \\ \Rightarrow \quad f(z) &\underset{\rho \to 0}{\to} e^{-\frac{C}{\rho} e^{j\theta}} = e^{-\frac{C}{\rho}\cos\theta} e^{-j\frac{C}{\rho}\sin\theta} \\ &\underset{\rho \to 0}{\to} 0 \end{aligned} \tag{C.8}$$

since the residue C is positive, and the net angle θ does not exceed $\pm\pi/2$. From (C.8) we can state that for points z, z' with modulus ≥ 1, we have

$$\forall\, \epsilon > 0,\ \exists\, \delta > 0 \,\big|\, |z - z'| < \delta \ \Rightarrow\ |f(z) - f(z')| < \epsilon. \tag{C.9}$$

Thus $f(z)$ is analytic in the strict outer disk, and continuous up to the unit circle which forms its boundary. By the maximum modulus theorem [138],

$$\sup_{|z|\geq 1} |f(z)| \triangleq \sup_{|z|\geq 1} \left|e^{-H(z)}\right| = \sup_{|z|\geq 1} e^{-Re\{H(z)\}} = \inf_{|z|\geq 1} Re\{H(z)\}$$

occurs on the unit circle. Consequently,

$$\inf_{-\pi<\theta\leq\pi} Re\{H(e^{j\theta})\} \geq 0 \quad \Rightarrow \quad \inf_{|z|\geq 1} Re\{H(z)\} \geq 0 \quad \Rightarrow \quad H(z)\ PR.$$

For example, if a transfer function is known to be strictly stable, then a frequency response with nonnegative real part implies that the transfer function is positive real.

Consideration of $1/H(z)$ leads to analogous necessary and sufficient conditions for $H(z)$ to be positive real in terms of its zeros instead of poles. ∎

C.1. Relation to Stochastic Processes

Theorem C.12. If a stationary random process $\{x_n\}$ has a rational power spectral density $R(e^{j\omega})$ corresponding to an autocorrelation function $r(k) = \mathcal{E}\{x_n x_{n+k}\}$, then

$$R_+(z) \triangleq \frac{r(0)}{2} + \sum_{n=1}^{\infty} r(n) z^{-n} \tag{C.10}$$

is positive real.

Proof. By the representation theorem [83, pp. 98-103] there exists an asymptotically stable filter $H(z) = b(z)/a(z)$ which will produce a realization of $\{x_n\}$ when driven by white noise, and we have $R(e^{j\omega}) = H(e^{j\omega})H(e^{-j\omega})$. We define the analytic continuation of $R(e^{j\omega})$ by $R(z) = H(z)H(z^{-1})$. Decomposing $R(z)$ into a sum of causal and anti-causal components gives

$$\begin{aligned} R(z) = \frac{b(z)b(z^{-1})}{a(z)a(z^{-1})} &= R_+(z) + R_-(z) \\ &= \frac{q(z)}{a(z)} + \frac{q(z^{-1})}{a(z^{-1})} \end{aligned} \tag{C.11}$$

where $q(z)$ is found by equating coefficients of like powers of z in

$$b(z)b(z^{-1}) = q(z)a(z^{-1}) + a(z)q(z^{-1}). \tag{C.12}$$

Since the poles of $H(z)$ and $R_+(z)$ are the same, it only remains to be shown that $Re\{R_+(e^{j\omega})\} \geq 0$, $0 \leq \omega \leq \pi$.

Since spectral power is nonnegative, $R(e^{j\omega}) \geq 0$ for all ω, and so

$$R(e^{j\omega}) \triangleq \sum_{n=-\infty}^{\infty} r(n)e^{j\omega n} = r(0) + 2\sum_{n=1}^{\infty} r(n)\cos(\omega n) = 2Re\{R_+(e^{j\omega})\} \geq 0. \quad \text{(C.13)}$$

∎

C.2. Relation to Schur Functions

Definition C.13. A *Schur function* $S(z)$ is defined as a complex function analytic and of modulus not exceeding unity in $|z| \geq 1$.*

Theorem C.14. Let α be a real number greater than zero. The function

$$S(z) \triangleq \frac{\alpha - R(z)}{\alpha + R(z)} \quad \text{(C.14)}$$

is a Schur function if and only if $R(z)$ is positive real.

Proof.

Suppose $R(z)$ is positive real. Then for $|z| \geq 1$, $Re\{R(z)\} \geq 0 \Rightarrow \alpha + Re\{R(z)\} \geq 0 \Rightarrow \alpha + R(z)$ is PR. Consequently, $\alpha + R(z)$ is minimum phase which implies all roots of $S(z)$ lie in the unit circle. Thus $S(z)$ is analytic in $|z| \geq 1$. Also,

$$\left|S(e^{j\omega})\right|^2 = \frac{\alpha^2 - 2\alpha Re\{R(e^{j\omega})\} + \left|R(e^{j\omega})\right|^2}{\alpha^2 + 2\alpha Re\{R(e^{j\omega})\} + \left|R(e^{j\omega})\right|^2} \leq 1.$$

By the maximum modulus theorem, $S(z)$ takes on its maximum value in $|z| \geq 1$ on the boundary. Thus $S(z)$ is Schur.

Conversely, suppose $S(z)$ is Schur. Solving (C.14) for $R(z)$ and taking the real part on the unit circle yields

$$R(z) = \alpha\frac{1 - S(z)}{1 + S(z)}$$

$$Re\{R(e^{j\omega})\} = \alpha Re\left\{\frac{1 - S(e^{j\omega})}{1 + S(e^{j\omega})} \cdot \frac{1 + S(e^{-j\omega})}{1 + S(e^{-j\omega})}\right\}$$

$$= \frac{\alpha Re\{1 - S(e^{j\omega}) + S(e^{-j\omega}) - \left|S(e^{j\omega})\right|^2}{\left|1 + S(e^{j\omega})\right|^2\}}$$

$$= \alpha\frac{1 - \left|S(e^{j\omega})\right|^2}{\left|1 + S(e^{j\omega})\right|^2} \geq 0.$$

* Classically, this definition is given with $|z| \leq 1$.

If $S(z) = c$ is constant, then $R(z) = (1 - |c|^2)/|1 + c|^2$ is PR. If $S(z)$ is not constant, then by the maximum principle, $S(z) < 1$ for $|z| > 1$. By Rouche's theorem applied on a circle of radius $1 + \epsilon$, $\epsilon > 0$, on which $|S(z)| < 1$, the function $1 + S(z)$ has the same number of zeros as the function 1 in $|z| \geq 1 + \epsilon$. Hence, $1 + S(z)$ is minimum phase which implies $R(z)$ is analytic for $z \geq 1$. Thus $R(z)$ is PR. ∎

C.3. Relation to Functions PR in the Right-Half Plane

Theorem C.15. $Re\{H(z)\} \geq 0$ for $|z| \geq 1 \quad \Leftrightarrow \quad Re\{H(\frac{\alpha+s}{\alpha-s})\} \geq 0$ for $Re\{s\} \geq 0$, where α is any positive real number.

Proof. We shall show that the change of variable $z \leftarrow (\alpha + s)/(\alpha - s)$, $\alpha > 0$, provides a conformal map from the z-plane to the s-plane that takes the region $|z| \geq 1$ to the region $Re\{s\} \geq 0$. The general formula for a bilinear conformal mapping of functions of a complex variable is given by

$$\frac{(z - z_1)(z_2 - z_3)}{(z - z_3)(z_2 - z_1)} = \frac{(s - s_1)(s_2 - s_3)}{(s - s_3)(s_2 - s_1)}. \tag{C.15}$$

In general, a bilinear transformation maps circles and lines into circles and lines [138]. We see that the choice of three specific points and their images determines the mapping for all s and z. We must have that the imaginary axis in the s-plane maps to the unit circle in the z-plane. That is, we may determine the mapping by three points of the form $z_i = e^{j\theta_i}$ and $s_i = j\omega_i$, $i = 1, 2, 3$. If we predispose one such mapping by choosing the pairs $(s_1 = \pm\infty) \leftrightarrow (z_1 = -1)$ and $(s_3 = 0) \leftrightarrow (z_3 = 1)$, then we are left with transformations of the form

$$s = \left(s_2 \frac{z_2 + 1}{z_2 - 1}\right)\left(\frac{z - 1}{z + 1}\right) = \alpha\left(\frac{z - 1}{z + 1}\right) \tag{C.16}$$

or

$$z \leftarrow \frac{\alpha + s}{\alpha - s}, \tag{C.17}$$

Letting s_2 be some point $j\omega$ on the imaginary axis, and z_2 be some point $e^{j\theta}$ on the unit circle, we find that

$$\alpha = j\omega \frac{e^{j\theta} + 1}{e^{j\theta} - 1} = \omega \frac{\sin\theta}{1 - \cos\theta} = \omega \cot(\theta/2) \tag{C.18}$$

which gives us that α is real. To avoid degeneracy, we require $s_2 \neq 0, \infty$, $z_2 \neq \pm 1$, and this translates to α finite and nonzero. Finally, to make the unit disk map to the *left*-half s-plane, ω and θ must have the same sign in which case $\alpha > 0$. ∎

There is a bonus associated with the restriction that α be real which is that

$$z = \frac{\alpha + s}{\alpha - s} \in \Re \quad \Leftrightarrow \quad s = \alpha \frac{z - 1}{z + 1} \in \Re. \tag{C.19}$$

We have therefore proven

Theorem C.16. $H(z)$ PR $\Leftrightarrow H(\frac{\alpha+s}{\alpha-s})$ PR, where α is any positive real number.

The class of mappings of the form C.15 which take the exterior of the unit circle to the right-half plane is larger than the class C.17. For example, we may precede the transformation C.17 by any conformal map which takes the unit disk to the unit disk, and these mappings have the algebraic form of a first order complex allpass whose zero lies inside the unit circle.

$$z \leftarrow e^{j\theta}\frac{w-w_0}{\overline{w_0}\,w-1}, \qquad |w_0|<1 \tag{C.20}$$

where w_0 is the zero of the allpass and the image (also pre-image) of the origin, and θ is an angle of pure rotation. Note that (C.20) is equivalent to a pure rotation, followed by a *real* allpass substitution (w_0 real), followed by a pure rotation. The general preservation of condition 2) in Def. C.1 forces the real axis to map to the real axis. Thus rotations by other than π are useless, except perhaps in some special cases. However, we may precede C.17 by the first order *real* allpass substitution

$$z \leftarrow \frac{w-r}{r\,w-1}, \qquad |r|<1,\ r\ real, \tag{C.21}$$

which maps the real axis to the real axis. This leads only to the composite transformation,

$$z \leftarrow \frac{s+(\alpha\frac{1-r}{1+r})}{s-(\alpha\frac{1-r}{1+r})} \tag{C.22}$$

which is of the form C.17 up to a minus sign (rotation by π). By inspection of C.15, it is clear that sign negation corresponds to the swapping of points 1 and 2, or 2 and 3. Thus the only extension we have found by means of the general disk to disk pre-transform, is the ability to interchange two of the three points already tried. Consequently, we conclude that the largest class of bilinear transforms which convert functions positive real in the outer disk to functions positive real in the right-half plane is characterized by

$$z \leftarrow \pm\frac{\alpha+s}{\alpha-s}. \tag{C.23}$$

Riemann's theorem may be used to show that (C.23) is also the largest such class of conformal mappings. It is not essential, however, to restrict attention solely to conformal maps. The pre-transform $z \leftarrow \overline{z}$, for example, is not conformal and yet PR is preserved.

The bilinear transform is one which is used to map analog filters into digital filters. Another such mapping is called the *matched z transform* [196]. It also preserves the positive real property.

Theorem C.17. $H(z)$ is PR if $H(e^{sT})$ is positive real in the analog sense, where $T > 0$ is interpreted as the sampling period.

Proof. The mapping $z \leftarrow e^{sT}$ takes the right-half s-plane to the outer disk in the z-plane. Also z is real if s is real. Hence $H(e^{sT})$ PR implies $H(z)$ PR. (Note, however, that rational functions do not in general map to rational functions.) ∎

These transformations allow application of the large battery of tests which exist for functions positive real in the right-half plane [203].

C.4. Special Cases and Examples

- The sum of positive real functions is positive real.
- The difference of positive real functions is conditionally positive real.
- The product or division of positive real functions is conditionally PR.
- $H(z)$ PR $\Rightarrow z^{\pm k}H(z)$ not PR for $k \geq 2$.
- $\frac{1}{H(z)} - \frac{1}{2}$ is PR iff $|H(z) - 1| \leq 1$ for $|z| \geq 1$.

Minimum Phase (MP) Polynomials in z

All properties of MP polynomials apply without modification to marginally stable allpole transfer functions (cf. Thm. C.3).

- Every first-order MP polynomial is positive real.
- Every first-order MP polynomial $b(z) = 1 + b_1 z^{-1}$ is such that $\frac{1}{b(z)} - \frac{1}{2}$ is positive real [107].
- A PR second-order MP polynomial with complex-conjugate zeros,

$$\begin{aligned} H(z) &= 1 + b_1 z^{-1} + b_2 z^{-2} \\ &= 1 - (2R\cos\varphi)z^{-1} + R^2 z^{-2}, \quad R \leq 1 \end{aligned}$$

satisfies

$$R^2 + \frac{\cos^2\varphi}{2} \leq 1 .$$

If $2R^2 + \cos^2\varphi = 2$, then $Re\{H(e^{j\omega})\}$ has a double zero at

$$\omega = \cos^{-1}\left(\pm\left(\frac{1-R^2}{2R^2}\right)^{\frac{1}{2}}\right) = \cos^{-1}\left(\pm\frac{\cos\varphi}{2R}\right) = \cos^{-1}\left(\pm\frac{\cos\varphi}{(2+2\sin^2\varphi)^{\frac{1}{2}}}\right).$$

- All polynomials of the form

$$H(z) = 1 + R^n z^{-n}, \quad R \leq 1$$

  are positive real. (These have zeros uniformly distributed on a circle of radius R.)

C.5. Conjectured Properties

The following conjectures are true for analog positive-real functions, but no attempt was made to establish them in the discrete-time case.

- If all poles and zeros of a PR function are on the unit circle, then they alternate along the circle.
- If $B(z)/A(Z)$ is PR, then so is $B'(z)/A'(z)$, where the prime denotes differentiation in z.

Appendix D. Frequency-Domain Error Criteria

Below are the objective error criteria shown in the figures comparing frequency-response functions in Chapter 3. Each L^2 error measure $J(\hat{\theta})$ is normalized to lie between 0 and 1 under normal circumstances. To avoid square root symbols, most measures are given in squared form below.

$$(L^2)\ \text{PREDICTION:}\ J^2(\hat{\theta}) = \frac{\sum_{k=1}^{N_s} \left| \frac{H(e^{j\omega_k})}{\hat{H}(e^{j\omega_k})} \right|^2}{\sum_{k=1}^{N_s} \left| H(e^{j\omega_k}) \right|^2}$$

$$(L^2)\ \ldots.\text{OUTPUT:}\ J^2(\hat{\theta}) = \frac{\sum_{k=1}^{N_s} \left| H(e^{j\omega_k}) - \hat{H}(e^{j\omega_k}) \right|^2}{\sum_{k=1}^{N_s} \left| H(e^{j\omega_k}) \right|^2}$$

$$(L^2)\ ..\text{EQUATION:}\ J^2(\hat{\theta}) = \frac{N_s \sum_{k=1}^{N_s} \left| \hat{A}(e^{j\omega_k}) H(e^{j\omega_k}) - \hat{B}(e^{j\omega_k}) \right|^2}{\sum_{k=1}^{N_s} \left| H(e^{j\omega_k}) \right|^2 \sum_{k=1}^{N_s} \left| \hat{A}(e^{j\omega_k}) \right|^2}$$

$$(L^2)\ |\text{SPECTRUM}|^2\text{:}\ J^2(\hat{\theta}) = \frac{\sum_{k=1}^{N_s} \left| \left| H(e^{j\omega_k}) \right|^2 - \left| \hat{H}(e^{j\omega_k}) \right|^2 \right|^2}{\sum_{k=1}^{N_s} \left| H(e^{j\omega_k}) \right|^4}$$

$$(L^2)\ \text{Ln}|\text{SPECTRUM}|\text{:}\ J^2(\hat{\theta}) = \frac{\sum_{k=1}^{N_s} \left| \ln \left| H(e^{j\omega_k}) \right|^2 - \ln \left| \hat{H}(e^{j\omega_k}) \right|^2 \right|^2}{\sum_{k=1}^{N_s} \ln^2 \left| H(e^{j\omega_k}) \right|^2}$$

$$(L^\infty)\ |\text{SPECTRUM}|\text{:}\ J^2(\hat{\theta}) = \max_k \left\{ \left| \left| H(e^{j\omega_k}) \right|^2 - \left| \hat{H}(e^{j\omega_k}) \right|^2 \right| \right\}$$

$$(L^\infty)\ \text{Ln}|\text{SPECTRUM}|\text{:} J(\hat{\theta}) = 10 \max_k \left\{ \left| \log_{10} \left| H(e^{j\omega_k}) \right|^2 - \log_{10} \left| \hat{H}(e^{j\omega_k}) \right|^2 \right| \right\}$$

Appendix E. Fundamentals

E.1. Digital Filter Theory

In this section, *linearity, time-invariance* and four basic representations of digital filters are defined: the *difference equation coefficients, impulse response, transfer function,* and *frequency response.* Next the concepts of *phase delay, group delay, poles and zeros,* and filter *stability* are defined. This elementary material was taken from course notes for a class given at Stanford by the author in 1979.

Definition E.1. A real *signal* is defined as any real-valued function of the integers. Similarly a complex signal is any complex-valued function of the integers.

Definition E.2. A real *filter* L_n is defined as any real-valued functional of a signal for each integer n. We express the input-output relation of the filter by

$$y(n) = L_n\{x(\cdot)\} \tag{E.1}$$

where $x(\cdot)$ is the entire input signal, and $y(n)$ is the output at time n.

E.1.1. Linearity and Time-Invariance

In everyday terms, the fact that a filter is *linear* means simply that

1) the amplitude of the output is proportional to the amplitude of the input,

and

2) when two signals are added together and fed to the filter, the filter output is the same as if one had put each signal through the filter separately and then added the outputs.

Definition E.3. A filter is said to be *linear* if, for any pair of signals $x_1(\cdot), x_2(\cdot)$ and for all constant gains g, we have

$$\begin{aligned} &1) \quad L_n\{g\,x_1(\cdot)\} = g\,L_n\{x_1(\cdot)\} \\ &2) \quad L_n\{x_1(\cdot) + x_2(\cdot)\} = L_n\{x_1(\cdot)\} + L_n\{x_2(\cdot)\}, \end{aligned} \tag{E.2}$$

for all n. These two conditions are a mathematical statement of the previous definition. For g rational, property 2) implies 1).

Definition E.4. A filter is said to be *time-invariant* if

$$\mathcal{L}_n\{x(\cdot - N)\} = \mathcal{L}_{n-N}\{x(\cdot)\} = y(n-N), \tag{E.3}$$

where $x(\cdot - N)$ is understood to denote the waveform $x(\cdot)$ shifted right (or delayed) by N samples.

From now on, all filters discussed will be linear and time-invariant. For brevity, these will be referred to as *LTI filters*.

E.1.2. Difference Equation

Definition E.5. The *difference equation* for a general linear time-invariant (LTI) digital filter is given by

$$\begin{aligned} y(n) = {} & b_0\,x(n) + b_1\,x(n-1) + \cdots + b_{n_b}\,x(n-n_b) \\ & - a_1\,y(n-1) - \cdots - a_{n_a}\,y(n-n_a) \end{aligned} \tag{E.4}$$

where x is the input signal, y is the output signal, and the constants $\{b_i,\ i = 0, 1, 2, \ldots, n_b\}$, $\{a_i,\ i = 1, 2, \ldots, n_a\}$ are called *difference equation coefficients*, or more simply, *filter coefficients*. When the a and b coefficients are real numbers, then the filter is said to be *real*.

When $n_b = 0, n_a > 0$, the filter is sometimes called an *all-pole, infinite-impulse-response (IIR)*, or *autoregressive (AR)* filter. When $n_a = 0, n_b > 0$, the filter may be called an *all-zero, finite-impulse-response (FIR)*, or *moving average (MA)* filter. When $n_a > 0, n_b > 0$, the filter may be called a *pole-zero, infinite-impulse-response (IIR)*, or *autoregressive moving average (ARMA)* filter. (The terms AR, MA, and ARMA are usually found in connection with filtered stochastic processes.)

Definition E.6. Equation (E.4) represents only *causal* LTI filters. A filter is said to be *causal* when the output does not depend on any "future" inputs. (In more colorful terms, a filter is causal if it does not "laugh" before it is "tickled.")

Definition E.7. The *maximum time span*, in samples, used in creating each output sample is called the *order* of the filter. In (E.4), the order is the larger of n_b and n_a. Since n_b and n_a in (E.4) are assumed finite, (E.4) represents the class of *finite order* causal LTI filters.

In addition to difference equation coefficients, any LTI filter may be represented in the time domain by its response to a specific signal called the *impulse*.

Definition E.8. The *impulse* is denoted as $\delta(n)$ and is defined by

$$\delta(n) \triangleq \begin{cases} 1, & n = 0 \\ 0, & n \neq 0 \end{cases}.$$

Definition E.9. The *impulse response* of a filter is the response of the filter to $\delta(n)$ and is most often denoted $h(n)$.

Definition E.10. A filter is said to be *stable* if the impulse response $h(n)$ approaches zero as n goes to infinity.

Convolution Representation

If $y(n)$ is the output of an LTI filter with input $x(n)$ and impulse response $h(n)$, then y is the *convolution* of x with h,

$$y(n) = \sum_{i=0}^{n} x(i)h(n-i) \triangleq x * h(n).$$

Since convolution is commutative ($x * h(n) = h * x(n)$), we have also

$$y(n) = \sum_{i=0}^{n} h(i)x(n-i). \tag{E.5}$$

Definition E.11. The *z-transform* of the discrete-time signal $x(n)$ is defined to be

$$X(z) \triangleq \sum_{n=-\infty}^{\infty} x(n)z^{-n},$$

That $x(n)$ and $X(z)$ are transform pairs is expressed by writing $X(z) = Z\{x(n)\}$ or $X(z) \leftrightarrow x(n)$.

Theorem E.12. The *convolution theorem* (Papoulis [195]) states that

$$x * y(n) \leftrightarrow X(z)Y(z). \tag{E.6}$$

In words, *convolution in the time domain is multiplication in the frequency domain.*

Taking the z-transform of both sides of (E.5) and applying the convolution theorem gives

$$Y(z) = H(z)X(z) \tag{E.7}$$

where $H(z)$ is the z-transform of the filter impulse response. Thus the z-transform of the filter output is the z-transform of the input times the z-transform of the impulse response.

Definition E.13. The *transfer function* $H(z)$ of a linear time-invariant discrete-time filter is defined to be the z-transform of the impulse response $h(n)$.

Theorem E.14. The *shift theorem* [195] for z-transforms states that

$$x(n-k) \leftrightarrow z^{-k}X(z).$$

The general difference equation for an LTI filter appears as

$$\begin{aligned} y(n) = b_0\,x(n) + b_1\,x(n-1) + \cdots + b_{n_b}\,x(n-n_b) \\ - a_1\,y(n-1) - \cdots - a_{n_a}\,y(n-n_a), \end{aligned}$$

Taking the z-transform of both sides, denoting the transform by $Z\{\,\}$ gives

$$\begin{aligned} Z\{y(n)\} = b_0\,Z\{x(n)\} + b_1\,z^{-1}Z\{x(n)\} + \cdots + b_{n_b}\,z^{-n_b}Z\{x(n)\} \\ - a_1\,z^{-1}Z\{y(n)\} - \cdots - a_{n_a}\,z^{-n_a}Z\{y(n)\}, \end{aligned}$$

using linearity and the shift theorem. Replacing $Z\{y(n)\}$ by $Y(z)$, $Z\{x(n)\}$ by $X(z)$, and solving for $Y(z)/X(z)$, which equals the transfer function $H(z)$, yields

$$H(z) = \frac{Y(z)}{X(z)} = \frac{b_0 + b_1\,z^{-1} + \cdots + b_{n_b}\,z^{-n_b}}{1 + a_1\,z^{-1} + \cdots + a_{n_a}\,z^{-n_a}}. \tag{E.8}$$

E.1.3. Frequency Response

From (E.8), we have

$$H(z) = \frac{Y(z)}{X(z)},$$

where $X(z)$ is the z-transform of the filter input, $Y(z)$ is the z-transform of the output signal, and $H(z)$ is the filter transfer function.

Definition E.15. The *frequency response* of a linear time-invariant digital filter is defined to be the transfer function, $H(z)$, evaluated on the unit circle, that is, $H(e^{j\omega})$.

The frequency response is a complex-valued function of a real variable. The response at frequency f Hz, for example, is $H(e^{j2\pi fT})$, where T is the sampling period in seconds.

Since every complex number can be represented as a magnitude and angle, the frequency response may be decomposed into two real-valued functions, the *amplitude response* and the *phase response*. Formally, we may define them as follows:

$$\begin{aligned} G(\omega) &\triangleq |H(e^{j\omega})| \\ \Theta(\omega) &\triangleq \angle H(e^{j\omega}) \end{aligned}$$

so that

$$H(e^{j\omega}) = G(\omega)\, e^{j\Theta(\omega)}. \tag{E.9}$$

Thus $G(\omega)$ is the magnitude (or complex modulus) of $H(e^{j\omega})$, and $\Theta(\omega)$ is the phase (or complex angle) of $H(e^{j\omega})$.

Definition E.16. The real valued function $G(\omega)$ is called the filter *amplitude response* or *magnitude frequency response* and it specifies the amplitude *gain* that the filter provides at each frequency.

Definition E.17. The function $G^2(\omega)$ is called the *power response* and it specifies the *power gain* at each frequency.

Definition E.18. The real function $\Theta(\omega)$ is the *phase response* and it gives the phase shift in radians that each input component sinusoid will undergo.

If the filter input and output signals are $x(n)$ and $y(n)$ respectively, then

$$\left|Y(e^{j\omega})\right| = G(\omega)\left|X(e^{j\omega})\right|$$
$$\angle Y(e^{j\omega}) = \Theta(\omega) + \angle X(e^{j\omega}).$$

E.1.4. Phase Delay and Group Delay

The phase response of a filter $\Theta(\omega)$ gives the *radian* phase shift experienced by each sinusoidal component of the input signal. Sometimes it is more meaningful to consider *phase delay* [195].

Definition E.19. The *phase delay* of an LTI filter $H(z)$ with phase response $\Theta(\omega)$ is defined by

$$P(\omega) \triangleq -\frac{\Theta(\omega)}{\omega}.$$

The phase delay gives the *time delay* in seconds experienced by each sinusoidal component of the input signal. For example, in the filter $y(n) \leftarrow x(n) + x(n-1)$, the phase response is $\Theta(\omega) = -\omega T/2$ which corresponds to a phase delay $P(\omega) = T/2$ which is one-half sample.

More generally, if the input to a filter with frequency response $H(e^{j\omega}) = G(\omega)e^{j\Theta(\omega)}$ is

$$x(n) = \cos(\omega n T),$$

then the output is

$$y(n) = G(\omega)\cos(\omega n T + \Theta(\omega)) = G(\omega)\cos\Big(\omega(nT - P(\omega))\Big),$$

and it can be seen that the phase delay expresses phase response as time delay.

In working with phase delay, care must be taken to ensure all appropriate multiples of 2π have been included in $\Theta(\omega)$. We defined $\Theta(\omega)$ simply as the complex angle of the frequency response $H(e^{j\omega})$, and this is not sufficient for obtaining a phase response which can be converted to true time delay. By discarding multiples of 2π, as is done in the definition of complex angle, the phase delay is modified by multiples of the sinusoidal period. Since LTI filter analysis is based on sinusoids without beginning or end, one cannot in principle distinguish between "true" phase delay and a phase delay with discarded sinusoidal periods. Nevertheless, it is convenient to define the filter phase response as a *continuous* function of frequency with the property that $\Theta(0) = 0$ (for real filters). This specifies a means of "unwrapping" the phase response to get a unique phase-delay curve.

Definition E.20. A more commonly encountered representation of filter phase response is called the *group delay*, and it is defined by

$$D(\omega) \triangleq -\frac{d}{d\omega}\Theta(\omega).$$

For linear phase responses, the group delay and the phase delay are identical, and each may be interpreted as time delay.

For any phase function, the group delay $D(\omega)$ may be interpreted as the time delay of the *amplitude envelope* of a sinusoid at frequency ω [195]. The bandwidth of the amplitude envelope in this interpretation must be restricted to a frequency interval over which the phase response is approximately linear. While the proof will not be given here, it should seem reasonable when the process of amplitude envelope detection is considered. The narrow "bundle" of frequencies centered at the carrier frequency ω is translated to 0 Hz. At this point, it is evident that the group delay at the carrier frequency gives the slope of the linear phase of the translated spectrum. But this is a constant phase delay, and therefore it has the interpretation of true time delay for the amplitude envelope.

E.2. Vector Space Concepts

Definition E.21. A set X of objects is called a *metric space* if with any two points p and q of X there is associated a real number $d(p,q)$, called the distance from p to q, such that (a) $d(p,q) > 0$ if $p \neq q$; $d(p,p) = 0$, (b) $d(p,q) = d(q,p)$, (c) $d(p,q) \leq d(p,r) + d(r,q)$, for any $r \in X$ [154].

Definition E.22. A *linear space* is a set of "vectors" X together with a field of "scalars" S with an addition operation $+ : X \times X \mapsto X$, and a multiplication opration $\cdot$ taking $S \times X \mapsto X$, with the following properties: If x, y, and z are in X, and α, β are in

S, then

(1) $x + y = y + x$.

(2) $x + (y + z) = (x + y) + z$.

(3) There exists $\emptyset$ in X such that $0 \cdot x = \emptyset$ for all x in X.

(4) $\alpha(\beta x) = (\alpha\beta)x$.

(5) $(\alpha + \beta)x = \alpha x + \beta x$.

(6) $1 \cdot x = x$.

(7) $\alpha(x + y) = \alpha x + \alpha y$.

The element $\emptyset$ is written as 0 thus coinciding with the notation for the real number zero. A linear space is sometimes be called a linear vector space, or a vector space.

Definition E.23. A *normed linear space* is a linear space X on which there is defined a real-valued function of $x \in X$ called a *norm*, denoted $\| x \|$, satisfying the following three properties:

(1) $\| x \| \geq 0$, and $\| x \| = 0 \quad \Leftrightarrow \quad x = 0$.

(2) $\| cx \| = | c | \cdot \| x \|$, c a scalar.

(3) $\| x_1 + x_2 \| \leq \| x_1 \| + \| x_2 \|$.

The functional $\| x - y \|$ serves as a distance function on X, so a normed linear space is also a metric space.

Note that when X is the space of continuous complex functions on the unit circle in the complex plane, the norm of a function is not changed when multiplied by a function of modulus 1 on the unit circle. In signal processing terms, the norm is insensitive to multiplication by a unity-gain allpass filter (also known as a Blaschke product).

Definition E.24. A *pseudo-norm* is a real-valued function of $x \in X$ satisfying the following three properties:

(1) $\| x \| \geq 0$, and $x = 0 \Rightarrow \| x \| = 0$.

(2) $\| cx \| = | c | \cdot \| x \|$, c a scalar.

(3) $\| x_1 + x_2 \| \leq \| x_1 \| + \| x_2 \|$.

A pseudo-norm differs from a norm in that the pseudo-norm can be zero for nonzero vectors (functions).

Definition E.25. A *Banach Space* is a *complete* normed linear space, that is, a normed linear space in which every Cauchy sequence* converges to an element of the space.

Definition E.26. A function $H(e^{j\omega})$ is said to belong to the space L^p if

$$\int_{-\pi}^{\pi} \left| H(e^{j\omega}) \right|^p \frac{d\omega}{2\pi} < \infty .$$

* A sequence $H_n(e^{j\omega})$ is said to be a *Cauchy sequence* if for each $\epsilon > 0$ there is an N such that $\| H_n(e^{j\omega}) - H_m(e^{j\omega}) \| < \epsilon$ for all n and m larger than N.

Definition E.27. A function $H(e^{j\omega})$ is said to belong to the space H^p if it is in L^p and if its analytic continuation $H(z)$ is analytic for $|z| < 1$. $H(z)$ is said to be in H^{-p} if $H(z^{-1}) \in H^p$.

Theorem E.28 (Riesz-Fischer). The L^p spaces are complete.
Proof. See Royden [153], p. 117.

Definition E.29. A Hilbert space is a Banach space with a symmetric bilinear inner product $\langle x, y \rangle$ defined such that the inner product of a vector with itself is the square of its norm $\langle x, x \rangle = \| x \|^2$.

E.3. Specific Norms

The L^p *norms* are defined on the space L^p by

$$\| F \|_p \triangleq \left(\frac{1}{2\pi} \int_{-\pi}^{\pi} \left| F(e^{j\omega}) \right|^p \frac{d\omega}{2\pi} \right)^{1/p}, \quad p \geq 1. \tag{E.10}$$

L^p norms are technically pseudo-norms; if functions in L^p are replaced by equivalence classes containing all functions equal almost everywhere, then a norm is obtained. Since all functions in problem $\hat{H}^*$ of Chapter 1 are continuous and therefore bounded on the unit circle ($H(e^{j\omega}) \in C_0$), it follows that each equivalence class contains only one function and that $\{H(e^{j\omega})\}$ forms a Banach space under any L^p norm.

The *weighted* L^p *norms* are defined by

$$\| F \|_p \triangleq \left(\frac{1}{2\pi} \int_{-\pi}^{\pi} \left| F(e^{j\omega}) \right|^p W(e^{j\omega}) \frac{d\omega}{2\pi} \right)^{\frac{1}{p}}, \quad p \geq 1, \tag{E.11}$$

where $W(e^{j\omega})$ is real, positive, and integrable. Typically, $\int W = 1$. If $W(e^{j\omega}) = 0$ for a set of nonzero measure, then a pseudo-norm results.

The case $p = 2$ gives the popular *root mean square* norm, and $\| \cdot \|_2^2$ can be interpreted as the total energy of F in many physical contexts.

An advantage of working in L^2 is that the norm is provided by an *inner product*,

$$\langle H, G \rangle \triangleq \int_{-\pi}^{\pi} H(e^{j\omega}) \overline{G(e^{j\omega})} \frac{d\omega}{2\pi}.$$

The norm of a vector $H \in L^2$ is then given by

$$\| H \| \triangleq \langle H, H \rangle^{\frac{1}{2}}.$$

As p approaches infinity in (E.10), the error measure is dominated by the largest values of $|F(e^{j\omega})|$. Accordingly, it is customary to define

$$\| F \|_\infty \triangleq \max_{-\pi < \omega \leq \pi} \left| F(e^{j\omega}) \right|, \tag{E.12}$$

and this is often called the *Chebyshev* or *uniform* norm.

Suppose the L^1 norm of $F(e^{j\omega})$ is finite, and let

$$f(n) \triangleq \frac{1}{2\pi} \int_{-\pi}^{\pi} F(e^{j\omega}) e^{j\omega n} \frac{d\omega}{2\pi}$$

denote the Fourier coefficients of $F(e^{j\omega})$. When $F(e^{j\omega})$ is a filter frequency response, $f(n)$ is the corresponding *impulse response*. The filter F is said to be *causal* if $f(n) = 0$ for $n < 0$.

The norms for impulse response sequences $\| f \|_p$ are defined in a manner exactly analogous with the frequency response norms $\| F \|_p$, viz.,

$$\| f \|_p \triangleq \left(\sum_{n=-\infty}^{\infty} | f(n) |^p \right)^{\frac{1}{p}}.$$

These time-domain norms are called *l^p norms*.

The L^p and l^p norms are *strictly concave* functionals for $1 < p < \infty$ (see below).

By Parseval's theorem, we have $\| F \|_2 = \| f \|_2$, i.e., the L^p and l^p norms are the same for $p = 2$.

The *Frobenious norm* of an $m \times n$ matrix A is defined as

$$\| A \|_F \triangleq \left(\sum_{i=1}^{m} \sum_{j=1}^{n} | a_{ij} |^2 \right)^{\frac{1}{2}}.$$

That is, the Frobenious norm is the L^2 norm applied to the elements of the matrix. For this norm there exists the following.

Theorem E.30. The unique $m \times n$ rank k matrix B which minimizes $\| A - B \|_F$ is given by $U \Sigma_k V^*$, where $A = U \Sigma V^*$ is a singular value decomposition of A, and Σ_k is formed from Σ by setting to zero all but the k largest singular values.

Proof. See Golub and Kahan [176].

The *induced norm* of a matrix A is defined in terms of the norm defined for the vectors $\underline{X}$ on which it operates,

$$\| A \| \triangleq \sup_{\underline{X}} \frac{\| A\underline{X} \|}{\| \underline{X} \|}$$

For the L^2 norm, we have

$$\| A \|_2^2 = \sup_{\underline{X}} \frac{\underline{X}^T A^T A \underline{X}}{\underline{X}^T \underline{X}},$$

and this is called the *spectral norm* of the matrix A.

The *Hankel matrix* corresponding to a time series f is defined by $\Gamma(f)[i,j] \triangleq f(i+j)$, i.e.,

$$\Gamma(f) \triangleq \begin{pmatrix} f(0) & f(1) & f(2) & \cdots \\ f(1) & f(2) & & \\ f(2) & & & \\ \vdots & & & \end{pmatrix}. \tag{E.13}$$

Note that the Hankel matrix involves only causal components of the time series.

The *Hankel norm* of a filter frequency response is defined as the spectral norm of the Hankel matrix of its impulse response,

$$\left\| F(e^{j\omega}) \right\|_H \triangleq \| \Gamma(f) \|_2 .$$

The Hankel norm is truly a norm only if $H(z) \in H^{-?}$, i.e., if it is causal. For noncausal filters, it is a pseudo-norm.

If F is strictly stable, then $|F(e^{j\omega})|$ is finite for all ω, and all norms defined thus far are finite. Also, the Hankel matrix $\Gamma(f)$ is a bounded linear operator in this case.

The Hankel norm is bounded below by the L^2 norm, and bounded above by the L^∞ norm [43,32],

$$\| F \|_2 \leq \| F \|_H \leq \| F \|_\infty ,$$

with equality iff F is an allpass filter (i.e., $|F(e^{j\omega})|$ constant).

E.4. Concavity

Definition E.31. A set S is said to be *concave** if for every vector x and y in S, $\lambda x + (1-\lambda)y$ is in S for all $0 \leq \lambda \leq 1$. In other words, all points on the line between two points of S lie in S.

Definition E.32. The *concave hull* of a set S in a metric space is the smallest concave set containing S.

Definition E.33. A *functional* is a mapping from a vector space to the real numbers $\Re$.

* or *convex*

Every norm is a functional. The norm of the approximation error $J(\hat{\theta})$ in problem $\hat{H}^*$ of Chapter 1 is a functional defined on the subset of $\mathbb{R}^{\hat{n}_a+\hat{n}_b+1}$ in which the filter coefficients are contained.

Definition E.34. A *linear functional* is a functional f such that for each x and y in the linear space X, and for all scalars α and β, we have $f(\alpha x + \beta y) = \alpha f(x) + \beta f(y)$.

Definition E.35. The *norm* of a linear functional f is defined on the normed linear space X by

$$\| f \| \triangleq \sup_{x \in X} \frac{|f(x)|}{\| x \|}.$$

Definition E.36. A functional f defined on a concave subset S of a vector space X is said to be *concave* on S if for every vector x and y in S,

$$\lambda f(x) + (1-\lambda) f(y) \geq f(\lambda x + (1-\lambda) y), \qquad 0 \leq \lambda \leq 1.$$

A concave functional has the property that its values along a line segment lie below or on the line between its values at the end points. The functional is *strictly concave* on S if strict inequality holds above for $\lambda \in (0,1)$. Finally, f is *uniformly concave* on S if there exists $c > 0$ such that for all $x, y \in S$,

$$\lambda f(x) + (1-\lambda) f(y) - f(\lambda x + (1-\lambda) y) \geq c\lambda(1-\lambda) \| x - y \|^2, \qquad 0 \leq \lambda \leq 1.$$

We have

$$\text{Uniformly Concave} \quad \Rightarrow \quad \text{Strictly Concave} \quad \Rightarrow \quad \text{Concave}$$

Definition E.37. A *local minimizer* of a real-valued function $f(x)$ is any x^* such that $f(x^*) \leq f(x)$ in some neighborhood of x.

Definition E.38. A *global minimizer* of a real-valued function $f(x)$ on a set S is any $x^* \in S$ such that $f(x^*) \leq f(x)$ for all $x \in S$.

Definition E.39. A *cluster point* x of a sequence x_n is any point such that every neighborhood of x contains at least one x_n.

E.5. Concave Norms

A desirable property of the error norm minimized by a filter-design technique is *concavity* of the error norm with respect to the filter coefficients. When this holds, the error surface "looks like a bowl," and the *global minimum* can be found by iteratively moving the parameters in the "downhill" (negative gradient) direction. The advantages of concavity are evident from the following classical results [156].

Theorem E.40. If X is a vector space, S a concave subset of X, and f a concave functional on S, then any local minimizer of f is a global minimizer of f in S.

Theorem E.41. If X is a normed linear space, S a concave subset of X, and f a *strictly* concave functional on S, then f has *at most* one minimizer in S.

Theorem E.42. Let S be a closed and bounded subset of $\Re^n$. If $f : \Re^n \mapsto \Re^1$ is *continuous* on S, then f has *at least* one minimizer in S.

Replacing "closed and bounded" with "compact," Thm. E.42 becomes true for a functional on an arbitrary metric space (Rudin [154], Thm. 4.14). (In $\Re^n$, "compact" is equivalent to "closed and bounded" [153].)

Thm. E.42 bears directly on the existence question for a solution to problem $\hat{H}^*$. It implies that only compactness of $\hat{\Theta} = \{\hat{b}_0, \ldots, \hat{b}_{\hat{n}_b}, \hat{a}_1, \ldots, \hat{a}_{\hat{n}_a}\}$ and continuity of the error norm $J(\hat{\theta})$ on $\hat{\Theta}$ need to be shown to prove existence of a solution.

E.6. Gradient Descent

Concavity is valuable in connection with the *Gradient Method* of minimizing $J(\hat{\theta})$ with respect to $\hat{\theta}$.

Definition E.43. The gradient of the error measure $J(\hat{\theta})$ is defined as the $\hat{N} \times 1$ column vector

$$J'(\hat{\theta}) \triangleq \frac{\partial J(\theta)}{\partial \theta}(\hat{\theta}) \triangleq \left(\frac{\partial J(\theta)}{\partial b_0}(\hat{b}_0), \frac{\partial J(\theta)}{\partial b_1}(\hat{b}_1), \ldots, \frac{\partial J(\theta)}{\partial b_{\hat{n}_b}}(\hat{b}_{\hat{n}_b}), \frac{\partial J(\theta)}{\partial a_1}(\hat{a}_1), \ldots, \frac{\partial J(\theta)}{\partial a_{\hat{n}_a}}(\hat{a}_{\hat{n}_a}) \right)^T.$$

Definition E.44. The *Gradient Method* (Cauchy) is defined as follows.

Given $\hat{\theta}_0 \in \hat{\Theta}$, compute

$$\hat{\theta}_{n+1} = \hat{\theta}_n - t_n J'(\hat{\theta}_n), \qquad n = 1, 2, \ldots,$$

where $J'(\hat{\theta}_n)$ is the *gradient* of J at $\hat{\theta}_n$, and $t_n \in \Re$ is chosen as the smallest nonnegative local minimizer of

$$\Phi_n(t) \triangleq J(\hat{\theta}_n - tJ'(\hat{\theta}_n)).$$

Cauchy originally proposed to find the value of $t_n \geq 0$ which gave a global minimum of $\Phi_n(t)$. This, however, is not always feasible in practice.

Some general results regarding the Gradient Method are given below [156].

Theorem E.45. If $\hat{\theta}_0$ is a local minimizer of $J(\hat{\theta})$, and $J'(\hat{\theta}_0)$ exists, then $J'(\hat{\theta}_0) = 0$.

Theorem E.46. The gradient method is a *descent* method, i.e., $J(\hat{\theta}_{n+1}) \leq J(\hat{\theta}_n)$.

Definition E.47. $J : \hat{\Theta} \to \Re^1$, $\hat{\Theta} \subseteq \Re^{\hat{N}}$, is said to be in the class $C_k(\hat{\Theta})$ if all kth order partial derivatives of $J(\hat{\theta})$ with respect to the components of $\hat{\theta}$ are continuous on $\hat{\Theta}$.

Definition E.48. The *Hessian* $J''(\hat{\theta})$ of J at $\hat{\theta}$ is defined as the matrix of second-order partial derivatives,

$$J''(\hat{\theta})[i,j] \triangleq \frac{\partial^2 J(\theta)}{\partial\theta[i]\partial\theta[j]}(\hat{\theta}),$$

where $\theta[i]$ denotes the ith component of θ, $i = 1, \ldots, \hat{N} = \hat{n}_a + \hat{n}_b + 1$, and $[i,j]$ denotes the matrix entry at the ith row and jth column.

The Hessian of every element of $C_2(\hat{\Theta})$ is *symmetric* [158]. This is because continuous second-order partials satisfy

$$\frac{\partial^2}{\partial x_1 \partial x_2} = \frac{\partial^2}{\partial x_2 \partial x_1}.$$

Theorem E.49. If $J \in C_1(\hat{\Theta})$, then any cluster point $\hat{\theta}_\infty$ of the gradient sequence $\hat{\theta}_n$ is necessarily a *stationary point*, i.e., $J'(\hat{\theta}_\infty) = 0$.

Theorem E.50. Let $\tilde{\Theta}$ denote the concave hull of $\hat{\Theta} \subseteq \Re^{\hat{N}}$. If $J \in C_2(\hat{\Theta})$, and there exist positive constants c and C such that

$$c\|\eta\|^2 \leq \eta^T J''(\hat{\theta})\eta \leq C\|\eta\|^2, \tag{E.14}$$

for all $\hat{\theta} \in \hat{\Theta}$ and for all $\eta \in \Re^{\hat{N}}$, then the gradient method beginning with any point in $\hat{\Theta}$ converges to a point $\hat{\theta}^*$. Moreover, $\hat{\theta}^*$ is the unique global minimizer of J in $\Re^{\hat{N}}$.

By the norm equivalence theorem [192], (E.14) is satisfied whenever $J''(\hat{\theta})$ is a *norm* on $\hat{\Theta}$ for each $\hat{\theta} \in \hat{\Theta}$. Since J'' belongs to $C_2(\hat{\Theta})$, it is a symmetric matrix. It is also bounded since it is continuous over a compact set. Thus a sufficient requirement is that J'' be *positive definite* on $\hat{\Theta}$. Positive definiteness of J'' can be viewed as "positive curvature" of J at each point of $\hat{\Theta}$ which corresponds to *strict concavity* of J on $\hat{\Theta}$.

E.7. Taylor's Theorem

Theorem E.51 (Taylor). Every functional $J : \Re^{\hat{N}} \mapsto \Re^1$ in $C_2(\Re^{\hat{N}})$ has the representation

$$J(\hat{\theta} + \eta) = J(\hat{\theta}) + J'(\hat{\theta})\eta + \frac{1}{2}\eta^T J''(\hat{\theta} + \lambda\eta)\eta$$

for some λ between 0 and 1, where $J'(\hat{\theta})$ is the $\hat{N} \times 1$ gradient vector evaluated at $\hat{\theta} \in \Re^n$, and $J''(\hat{\theta})$ is the $\hat{N} \times \hat{N}$ Hessian matrix of J at $\hat{\theta}$, i.e.,

$$J'(\hat{\theta}) \triangleq \frac{\partial J(\theta)}{\partial\theta}(\hat{\theta})$$

$$J''(\hat{\theta}) \triangleq \frac{\partial^2 J(\theta)}{\partial\theta^2}(\hat{\theta})$$

Proof. See Goldstein [143] p. 119. The Taylor infinite series is treated in Williamson and Crowell [158]. The present form is typically more useful for computing bounds on the error incurred by neglecting higher order terms in the Taylor expansion.

E.8. Newton's Method

The gradient method is based on the first-order term in the Taylor expansion for $J(\hat{\theta})$. By taking a second-order term as well and solving the quadratic minimization problem iteratively, *Newton's method* for functional minimization is obtained. Essentially, Newton's method requires the error surface to be close to *quadratic*, and its effectiveness is directly tied to the accuracy of this assumption. For most problems, the error surface can be well approximated by a quadratic form near the solution. For this reason, Newton's method tends to give very rapid ("quadratic") convergence in the last stages of iteration.

Newton's method is derived as follows: The Taylor expansion of $J(\theta)$ about $\hat{\theta}$ gives

$$J(\hat{\theta}^*) = J(\hat{\theta}) + J'(\hat{\theta})(\hat{\theta}^* - \hat{\theta}) + \frac{1}{2}(\hat{\theta}^* - \hat{\theta})^T J''(\lambda\hat{\theta}^* + \check{\lambda}\hat{\theta})(\hat{\theta}^* - \hat{\theta}),$$

for some $0 \leq \lambda \leq 1$, where $\check{\lambda} \triangleq 1 - \lambda$. It is now necessary to assume that $J''(\lambda\hat{\theta}^* + \check{\lambda}\hat{\theta}) \approx J''(\hat{\theta})$. Differentiating with respect to $\hat{\theta}^*$, where $J(\hat{\theta}^*)$ is presumed to be minimum, this becomes

$$0 = 0 + J'(\hat{\theta}) + J''(\hat{\theta})\left(\hat{\theta}^* - \hat{\theta}\right).$$

Solving for $\hat{\theta}^*$ yields

$$\hat{\theta}^* = \hat{\theta} - [J''(\hat{\theta})]^{-1} J'(\hat{\theta}). \tag{E.15}$$

Applying (E.15) iteratively, we obtain the following.

Definition E.52. *Newton's method* is defined by

$$\hat{\theta}_{n+1} = \hat{\theta}_n - [J''(\hat{\theta}_n)]^{-1} J'(\hat{\theta}_n), \quad n = 1, 2, \ldots, \tag{E.16}$$

where $\hat{\theta}_0$ is given as an initial condition.

When $J''(\lambda\hat{\theta}^* + \check{\lambda}\hat{\theta}) = J''(\hat{\theta})$, the answer is obtained after the first iteration. In particular, when the error surface $J(\hat{\theta})$ is a *quadratic form* in $\hat{\theta}$, Newton's method produces $\hat{\theta}^*$ in one iteration, i.e., $\hat{\theta}_1 = \hat{\theta}^*$ for every $\hat{\theta}_0$.

For Newton's method, there is the following general result on the existence of a critical point (i.e. a point at which the gradient vanishes) within a sphere of a Banach space.

Theorem E.53 (Kantorovich). Let $\hat{\theta}_0$ be a point in $\hat{\Theta}$ for which $[J''(\hat{\theta}_0)]^{-1}$ exists, and set

$$R_0 \triangleq \left\| [J''(\hat{\theta}_0)]^{-1} J'(\hat{\theta}_0) \right\|.$$

Let S denote the sphere $\{\hat{\theta} \in \hat{\Theta} \mid \| \hat{\theta} - \hat{\theta}_0 \| \leq 2R_0\}$. Set $C_0 = \| J''(\hat{\theta}_0) \|$. If there exists a number

M such that

$$\left\| J''(\hat{\theta}_1) - J''(\hat{\theta}_2) \right\| \leq \frac{M \left\| \hat{\theta}_1 - \hat{\theta}_2 \right\|}{2},$$

for $\hat{\theta}_1, \hat{\theta}_2$ in S, and such that $C_0 R_0 M \triangleq h_0 \leq 1/2$, then $J'(\hat{\theta}) = 0$ for some $\hat{\theta}$ in S, and the Newton sequence (E.16) converges to it. Furthermore, the rate of convergence is quadratic, satisfying

$$\left\| \hat{\theta}^* - \hat{\theta}_n \right\| \leq 2^{-n+1} (2h_0)^{2^n - 1} R_0 .$$

Proof. See Goldstein [143], p. 143.

E.9. Maxims of Signal Processing

1) Every technique is equivalent to the same operation, once you really understand it.

2) If one technique is superior to another, it is due to more averaging.

3) With sufficiently sophisticated processing, it is no longer necessary to have any input data.

4) Almost all unusual or interesting results are ultimately found to be artifacts of the processing.

References

R.1. Rational Approximation on the Unit Circle

[1] N. I. Achieser, *Theory of Approximation,* Frederick Ungar Publishing Co., New York, 1956.

[2] V. M. Adamjan, D. Z. Arov, and M. G. Krein, "Analytic Properties of Schmidt Pairs for a Hankel Operator and the Generalized Schur-Takagi Problem," *Math. USSR Sbornik,* vol. 15, pp. 31–73, 1971.

[3] S. Alliney and F. Sgallari, "Chebyshev Approximation of Recursive Digital Filters having Specified Amplitude and Phase Characteristics," *Signal Processing,* vol. 2, pp. 317–321, 1980.

[4] J. A. Athanassopoulos and A. D. Waren, "Design of Discrete-time Systems by Mathematical Programming," in *Proc. 1968 Hawaii Int. Conf. Syst. Sci.,* Honolulu, Univ. of Hawaii Press, pp. 224–227, 1968.

[5] P. A. Bernhardt, "Simplified Design of High-Order Recursive Group-Delay Filters," *IEEE Trans. on Acoust., Speech, and Signal Proc.,* vol. ASSP–28, pp. 498–503, 1980.

[6] L. Barrodale, M. J. D. Powell, and F. D. K. Roberts, "The Differential Correction Algorithm for Rational L_∞-Approximation," *SIAM J. Numer. Anal.,* vol. 9, pp. 493–504, Sep. 1972.

[7] F. Brophy and A. C. Salazar, "Recursive Digital Filter Synthesis in the Time Domain," *IEEE Trans. on Acoust., Speech, and Signal Proc.,* vol. ASSP–22, pp. 45–55, 1974.

[8] C. S. Burrus and T. W. Parks, "Time Domain Design of Recursive Digital Filters," *IEEE Trans. on Audio Electroacoust.,* vol. AU–18, pp. 137–141, June 1970.

[9] J. A. Cadzow, "High Performance Spectral Estimation—A New ARMA Method," *IEEE Trans. on Acoust., Speech, and Signal Proc.,* vol. ASSP–28, pp. 524–529, Oct. 1980.

[10] C. Carathéodory and L. Fejér, "Über den Zusammenhang der Extremen von harmonischen Funktionen mit ihrer Koeffizienten und über den Picard-Landauschen Satz.", Rend. Circ. Mat. Palermo, vol. 32, pp. 218-239, 1911.

[11] E. W. Cheney and H. L. Loeb, "Two New Algorithms for Rational Approximation," *Numer. Math.,,* vol. 3, pp. 72–75, 1961.

[12] E. W. Cheney and H. L. Loeb, "On Rational Approximation," *Numer. Math.,,* vol. 4, pp. 124–127, 1962.

[13] E. W. Cheney, *Introduction to Approximation Theory,* McGraw-Hill, New York, 1966.

[14] C. K. Chui and A. K. Chan, "A Two-sided Rational Approximation Method for Recursive Digital Filtering," *IEEE Trans. on Acoust., Speech, and Signal Proc.*, vol. ASSP-27, pp. 141-145, Apr. 1979.

[15] C. K. Chui, p. W. Smith, and L. Y. Su, "A Minimization Problem related to Padé Synthesis of Recursive Digital Filters," in [60], pp. 247-256, 1977.

[16] C. K. Chui, O. Shisha, and P. W. Smith, "Padé Approximants as Limits of Chebyshev Rational Approximants," *J. Approx. Theory,* vol. 12, pp. 201-204, 1974.

[17] D. Clark, "Hankel Forms, Toeplitz Forms and Meromorphic Functions," *Trans. Amer. Math. Soc.,* vol. 134, pp. 109-116, 1968.

[18] R. L. Crane, "All-Pass Network Synthesis," *IEEE Trans. on Circ. Theory,* vol. CT-15, pp. 474-477, Dec. 1968.

[19] P. J. Davis, *Interpolation and Approximation,* Dover, New York, 1975 (orig. 1963).

[20] A. G. Deczky, "Synthesis of Recursive Digital Filters using the Minimum p-Error Criterion," *IEEE Trans. on Audio Electroacoust.*, vol. AU-20, pp. 257-263, 1972.

[21] A. G. Deczky, "Equiripple and Minimax (Chebyshev) Approximations for Recursive Digital Filters," *IEEE Trans. on Acoust., Speech, and Signal Proc.*, vol. ASSP-22, pp. 98-111, 1974.

[22] Y. V. Genin, "On the Role of the Nevanlinna-Pick Problem in Circuit and System Theory," P. H. Delsarte, Y. V. Genin, and Y. Kamp, *Int. J. Circuit Theory Appl.,* vol. 9, no. 2, pp. 177-187, April 1981.

[23] P. DeWilde and H. Dym, "Schur Recursions, Error Formulas, and Convergence of Rational Estimators for Stationary Stochastic Estimators," *IEEE Trans. on Info. Theory,* vol. IT-27, pp. 446-461, July 1981.

[24] M. T. Dolan, "Comments on 'On the Approximation Problem for Recursive Digital Filters with Arbitrary Attenuation Curve in the Pass-Band and Stop-Band' [26] ," *IEEE Trans. on Acoust., Speech, and Signal Proc.*, vol. ASSP-24, pp. 575-577, Dec. 1976.

[25] R. G. Douglas and W. Rudin, "Approximation by Inner Functions," *Pacific J. Math.*, vol. 31, pp. 313-320, 1969.

[26] H. Dubois and H. Leech, "On the Approximation Problem for Recursive Digital Filters with Arbitrary Attenuation Curve in the Pass-Band and Stop-Band," *IEEE Trans. on Acoust., Speech, and Signal Proc.*, vol. ASSP-23, pp. 202-207, Apr. 1975.

[27] D. E. Dudgeon, "Recursive Filter Design using Differential Correction," *IEEE Trans. on Acoust., Speech, and Signal Proc.*, vol. ASSP-22, pp. 443-448, Dec. 1974.

[28] C. B. Dunham, "Stability of Differential Correction for Rational Approximation," *SIAM J. Numer. Anal.,*, vol. 17, no. 5, pp. 639-640, Oct. 1980.

[29] S. Ellacott and J. Williams, "Linear Chebyshev Approximation in the Complex Plane using Lawson's Algorithm," *Math. of Computation,* vol. 30, pp. 35–44, Jan. 1976.

[30] S. Ellacott and J. Williams, "Rational Chebyshev Approximation in the Complex Plane," *SIAM J. Numer. Anal.,* vol. 13, pp. 310–323, June 1976.

[31] Y. V. Genin and S. Kung, "A Two-Variable Approach to the Model Reduction Problem with Hankel Norm Criterion," *IEEE Trans. Circ. and Sys.,* vol. CAS–28, No. 9, pp. 912–924, Sep. 1981.

[32] Y. Genin, "An Introduction to the Model Reduction Problem with Hankel Norm Criterion," in *Proc. European Conf. Circuit Theory and Design,* The Hague, Netherlands, Aug. 1981.

[33] M. H. Gutknecht, "Ein Abstiegsverfahren für gleichmässige Approximation, mit Anwendungen," *Diss. ETH Zürich,* 1973.

[34] M. H. Gutknecht, "Non-Strong Uniqueness in Real and Complex Chebyshev Approximation," *J. Approx. Theory,* vol. 23, pp. 204–213, 1978.

[35] M. Gutknecht, "Rational Carathéodory-Fejér Approximation on a Disk, a Circle, and an Interval," to appear in *J. Approx. Theory* (accepted for publication).

[36] M. H. Gutknecht, J. O. Smith, and L. N. Trefethen, "The Carathéodory-Fejér (CF) Method for Recursive Digital Filter Design," Submitted to the IEEE Trans. on Acoust., Speech, and Signal Proc (accepted for publication).

[37] M. H. Gutknecht and L. N. Trefethen, "Recursive digital filter design by the Carathéodory-Fejér method," Numerical Analysis Manuscript NA-80-01, Computer Science Dept., Stanford Univ., 1980.

[38] M. H. Gutknecht, private communiction.

[39] M. H. Gutknecht and L. N. Trefethen, "Three Papers on Rational Chebyshev Approximation," Res. Rep. no. 82-07, Seminar für Angewandte Mathematik, ETH, CH-8092, Sept. 1982.

[40] H. D. Helms, "Digital Filters with Equiripple or Minimax Responses," *IEEE Trans. on Audio Electroacoust.,* vol. AU–19, pp. 87–94, 1971.

[41] W. Hintzman, "On the Existence of Best Analytic Approximations," in *Approximation Theory II,* G. G. Lorentz, C. K. Chui, and L. L. Schumaker, eds., Academic Press, New York, 1976.

[42] E. H. Kaufman and G. D. Taylor, "Uniform Approximation by Rational Functions Having Restricted Denominators," *J. Approx. Th.,,* vol. 32, pp. 9–26, 1981.

[43] S. Y. Kung, "Optimal Hankel-Norm Model Reductions: Scalar Systems," *Proc. Joint Aut. Control Conf.,* San Francisco, 1980.

[44] S. Y. Kung and D. W. Lin, "A State-Space Formulation for Optimal Hankel-Norm Approximation," *IEEE Trans. Automat. Contr.,* vol. AC–26, No. 4, pp. 942–946, Aug. 1981.

[45] S. Y. Kung and D. W. Lin, "Reduced-Order System Modeling via Singular Value Analysis," to appear.

[46] S. Y. Kung and D. W. Lin, "Optimal Hankel-Norm Model Reductions: Multivariable Systems," *Proc. IEEE Conf. Decis. Control Incl. Symp. Adapt. Processes 19th,*, vol. 1, Albuquerque, NM, Dec. 10–12 1980. Also, *IEEE Trans. Automat. Contr.*, vol. AC–26, No. 4, pp. 832–852, Aug. 1981.

[47] L. S. Lasdon and A. D. Warren, "Optimal Design of Filters with Bounded, Lossy Elements," *IEEE Trans. on Circ. Theory*, vol. CT–13, pp. 175–187, June 1966.

[48] C. M. Lee and F. D. K. Roberts, "A Comparison of Algorithms for Rational l_∞ Approximation," *Math. Comp.*, vol. 27, pp. 111-121, 1973.

[49] G. G. Lorentz, *Approximation of Functions*, Holt, Rinehart, and Winston, New York, 1966.

[50] M. T. McCallig, R. Kurth, and B. Steel, "Recursive Digital Filters with Low Coefficient Sensitivity," *Proc. IEEE Int. Conf. Acoust. Speech and Sig. Proc.*, Washington, D.C., April 1979.

[51] D. J. Newman, "Approximation with Rational Functions," Regional Conf. Series in Math., Conf. Board Math. Sci., no. 41, June 1978.

[52] T. W. Parks and J. H. McClellan, "A Program for the Design of Linear Phase Finite Impulse Response (FIR) Digital Filters," *IEEE Trans. on Audio Electroacoust.*, vol. AU–20, pp. 195–199, Aug. 1972.

[53] R. Prony, "Essai experimental et Analytique sur les lois de la dilatabilité des fluides élastiques et sur celles de la force expansive de la vapeur de l'eau et de la vapeur de l'alcool, à différentes temperatures," J. Ecole Polytech., Paris, vol. 1, pp. 24–76, 1795.

[54] L. R. Rabiner, N. Y. Graham and H. D. Helms, "Linear Programming Design of IIR Digital Filters with Arbitrary Magnitude Function," *IEEE Trans. on Acoust., Speech, and Signal Proc.*, vol. ASSP–22, pp. 117–123, 1974.

[55] L. R. Rabiner and K. Steiglitz, "The Design of Wide-band Recursive and Nonrecursive Digital Differentiators," *IEEE Trans. on Audio Electroacoust.*, vol. AU–18, pp. 204–209, 1970.

[56] B. D. Rakovitch and B. M. Djurich, "Chebyshev Approximation of a Constant Group Delay with Constraints at the Origin," *IEEE Trans. on Circ. Theory*, vol. CT–19, pp. 466–475, Sep. 1972.

[57] J. R. Rice, *The Approximation of Functions, Volume I*, Addison-Wesley, Reading MA, 1964.

[58] J. R. Rice, *The Approximation of Functions, Volume II*, Addison-Wesley, Reading MA, 1969.

[59] T. J. Rivlin and H. S. Shapiro, "A Unified Approach to Certain Problems of Approximation and Minimization," *J. Soc. Indust. Appl. Math.*, vol. 9, pp. 670–699, 1961.

[60] E. B. Saff and R. S. Varga, eds., *Padé and Rational Approximation,* Academic Press, New York, 1977.

[61] J. L. Shanks, "Recursion Filters for Digital Processing," *Geophysics,* vol. 32, pp. 33–51, Feb. 1967.

[62] L. L. Sharf and J. C. Luby, "Statistical Design of ARMA Digital Filters," *IEEE Trans. on Acoust., Speech, and Signal Proc.*, vol. ASSP-27, pp. 240–247, June 1979.

[63] L. M. Silverman and M. Bettayeb, "Optimal Approximation of Linear Systems," to appear.

[64] S. D. Stearns, "Error Surfaces of Recursive Adaptive Filters," *IEEE Trans. on Acoust., Speech, and Signal Proc.*, vol. ASSP-29, pp. 763–766, June 1981.

[65] K. Steiglitz, "Computer-aided Design of Recursive Digital Filters," *IEEE Trans. on Audio Electroacoust.*, vol. AU-18, pp. 123–129, 1970.

[66] "Optimal Convergence of Minimum Norm Approximation in H^p," Numer. Math., vol. 29, pp. 345–362, 1978.

[67] T. Takagi, "On an Algebraic Problem related to an Analytic Theorem of Carathéodory and Fejér and on an Allied Theorem of Landau" and "Remarks on an Algebraic Problem," *Japan J. Math.*, vol. 1, pp. 83–91, 1924, and vol. 2, pp. 13–17, 1925.

[68] P. Thajchayapong and P. J. W. Rayner, "Recursive Digital Filter Design by Linear Programming," *IEEE Trans. on Audio Electroacoust.*, vol. AU-21, pp. 107–112, 1973.

[69] J. Thiran, "Equal-Ripple Delay Recursive Digital Filters," *IEEE Trans. on Circ. Theory,* vol. CT-18, pp. 664–669, Nov. 1971.

[70] L. N. Trefethen, private communication.

[71] L. N. Trefethen, "Near-circularity of the Error Curve in Complex Chebyshev Approximation," *J. Approx. Theory,* vol. 31, pp. 344–367, 1981.

[72] L. N. Trefethen, "Rational Chebyshev Approximation on the Unit Disk," *Numer. Math.*, vol. 37, pp. 297–320, 1981.

[73] L. N. Trefethen and M. H. Gutknecht, "The Carathéodory-Fejér Method for Real Rational Approximation," submitted to *SIAM J. Numer. Anal.*,

[74] R. Unbehauen, "Recursive Digital Low-Pass Filters with Predetermined Phase or Group Delay and Chebyshev Stop-Band Attenuation," *IEEE Trans. on Circ. Theory,* vol. CT-28, pp. 905–912, Sep. 1981.

[75] J. L. Walsh, *Interpolation and Approximation by Rational Functions in the Complex Domain,* American Mathematical Society, Providence RI, 1960.

[76] J. L. Walsh, "Padé Approximants as Limits of Rational Functions of Best Approximation," *J. Math. Mech.*, vol. 13, pp. 305–312, 1964.

[77] B. Widrow, P. F. Titchener, and R. P. Gooch, "Adaptive Design of Digital Filters," *Proc. IEEE Conf. Acoust. Speech Sig. Proc.*, pp. 243–246, 1981.

[78] J. Williams, "Characterization and Computation of Rational Chebyshev Approximation in the Complex Plane," *SIAM J. Numer. Anal.,*, vol. 16, no. 5, pp. 819–827, Oct. 1979.

[79] T. Yahagi, "New Methods for the Design of Recursive Digital Filters in the Time Domain," *IEEE Trans. on Acoust., Speech, and Signal Proc.*, vol. ASSP–29, pp. 245–254, Apr. 1981.

[80] B. Yegnanarayana, "Design of Recursive Group-Delay Filters by Autoregressive Modeling," *IEEE Trans. on Acoust., Speech, and Signal Proc.*, vol. ASSP–30, pp. 632–637, Aug. 1982.

R.2. System Identification

[81] K. J. Astrom and P. Eykhoff, "System Identification, a Survey," *Part 1 (of 2) Preprints 2nd IFAC Symposium on Identification,* Prague, Czechoslovakia, June 15–20, 1970, and *Automatica*, vol. 7, pp. 123–162, 1971.

[82] K. J. Astrom, "Maximum Likelihood and Prediction Error Methods," Contained in [99].

[83] K. J. Astrom, *Introduction to Stochastic Control Theory,* Academic Press, New York, 1970.

[84] K. J. Astrom and T. Soderstrom, "Uniqueness of the Maximum Likelihood Estimates of the Parameters of an ARMA Model," *IEEE Trans. Automat. Contr.*, vol. AC–19, No. 6, pp. 769–773, Dec. 1974.

[85] P. E. Caines, "Prediction Error Identification Methods for Stationary Stochastic Processes," Res. Rep. No. 7516, Univ. of Toronto, Dept. of Elect. Eng., 1975.

[86] P. E. Caines, "Stationary Linear and Nonlinear System Identification and Predictor Set Completeness," Res. Rep. No. 7516, Univ. of Toronto, Dept. of Elect. Eng., 1975.

[87] P. E. Caines and L. Ljung, "Asymptotic Normality and Accuracy of Prediction Error Estimators," Res. Rep. No. 7602, Univ. of Toronto, Dept. of Elect. Eng., 1976.

[88] B. M. Finigan and I. H. Rowe, "Strongly Consistent Parameter Estimation by the Introduction of Strong Instrumental Variables," *IEEE Trans. Automat. Contr.*, vol. AC–19, No. 6, Dec. 1974.

[89] B. Friedlander, "A Recursive Maximum Likelihood Algorithm for ARMA Spectral Estimation," *IEEE Trans. on Info. Theory*, vol. IT–28, pp. 639–646, July 1982.

[90] B. Friedlander, "A Recursive Maximum Likelihood Algorithm for ARMA Line Enhancement," *IEEE Trans. on Acoust., Speech, and Signal Proc.*, vol. ASSP–30, pp. 651–657, Aug. 1982.

[91] B. Friedlander, "System Identification Techniques for Adaptive Signal Processing," *Circuit Systems Signal Process*, vol. 1, no. 1, 1982.

[92] B. Friedlander and B. Porat, "A Non-Iterative Method for ARMA Spectral Estimation," to appear.

[93] B. Friedlander and B. Porat, "A Parametric Technique for Time Delay Estimation," to appear.

[94] B. Friedlander and M. Morf, "Least Squares Algorithms for Adaptive Linear-Phase Filtering," *IEEE Trans. on Acoust., Speech, and Signal Proc.*, vol. ASSP-30, pp. 381-390, June 1982.

[95] G. C. Goodwin and R. L. Payne, *Dynamic System Identification,* Academic Press, New York, 1977.

[96] N. Gupta and R. K. Mehra, "Computational Aspects of Maximum Likelihood Estimation and Reduction in Sensitivity Function Calculations," *IEEE Trans. Automat. Contr.*, vol. AC-19, No. 6, 1974.

[97] N. Gupta, private communication, July 1980.

[98] R. Hastings-James and M. W. Sage, "Recursive Generalized Least Squares Procedure for On-line Identification of Process Parameters," *Proc. IEE*, vol. 116, No. 12, pp. 2057, 1969.

[99] R. Isermann, ed., "System Identification Tutorials," *Preprints 5th IFAC Symposium on Identification,* Darmstadt, Germany, Sept. 24-28, 1979, and *Automatica*, vol. 16, pp. 500-574, 1980.

[100] R. Isermann, ed., *Identification and System Parameter Estimation,* Volumes I and II, *Proceedings 5th IFAC Symposium on Identification,* Darmstadt, Germany, Sept. 24-28, 1979, Pergamon Press, New York, N.Y., 1979.

[101] C. R. Johnson, "The Common Parameter Estimation Basis of Adaptive Filtering, Identification, and Control," *IEEE Trans. on Acoust., Speech, and Signal Proc.*, vol. ASSP-30, pp. 587-595, Aug. 1982.

[102] I. D. Landau, "Unbiased recursive identification using model reference adaptive techniques," *IEEE Trans. Automat. Contr.*, vol. AC-21, pp. 194-202, Apr. 1976.

[103] L. Ljung, "Analysis of Recursive Stochastic Algorithms," *IEEE Trans. Automat. Contr.*, vol. AC-22, pp. 551-575, Aug. 1977.

[104] L. Ljung, "Characterization of the Concept "Persistently Exciting" in the Frequency Domain," Report 7119, Lund Inst. of Tech., Lund, Sweden, Div. of Auto. Control, 1971.

[105] L. Ljung, "Convergence Analysis of Parameteric Identification Methods," *IEEE Trans. Automat. Contr.*, vol. AC-23, pp. 770-783, 1978.

[106] L. Ljung, "Consistency of the Least Squares Identification Method," *IEEE Trans. Automat. Contr.*, vol. AC-21, pp. 779-781, Oct. 1976.

[107] L. Ljung, "On Positive Real Transfer Functions and the Convergence of some Recursive Schemes," *IEEE Trans. Automat. Contr.*, vol. AC-22, pp. 551-575, Aug. 1977.

[108] L. Ljung, "On Recursive Prediction Error Identification Algorithms," Report LiTH-ISY-I-0226, Dept. of Elect. Eng., Linkoping University, S-581 83 Linkoping, Sweden, 1978.

[109] L. Ljung, "On the Consistency of Prediction Error Identification Methods," Contained in [113].

[110] L. Ljung, M. Morf, and D. Falconer, "Fast Calculation of Gain Matrices for Recursive Estimation Schemes," *Int. Journal of Control,*, vol. 27, no. 1, pp. 1–19, 1978.

[111] L. Ljung and T. L. Soderstrom, *Theory and Practice of Recursive Identification,* MIT Press, Cambridge MA, 1983.

[112] K. D. Marshall and J. D. Rankert, "Modal Analysis: Concepts, Measurement, and Modeling Techniques," BF Goodrich Tech. Doc., Dept. 8521, Program no. 75-21-007,Oct. 20, 1978.

[113] R. K. Mehra and D. G. Lainiotis, ed.'s, *System Identification: Advances and Case Studies,* Academic Press, New York, 1976.

[114] M. Morf, class notes for *EE 479,* Dept. of Elect. Eng., Stanford University, Stanford CA, 1979.

[115] M. Morf, *Fast Algorithms for Multivariable Sysytms,* Ph.D. dissertation, Dept. of Elect. Eng., Stanford University, Stanford CA, Aug., 1974.

[116] M. Morf and D. T. L. Lee, "Fast Algorithms for Speech Modeling," Tech. Rep. No. M308–1, Information Systems Lab, Dept. of Elect. Eng., Stanford University, Stanford CA, Dec., 1978.

[117] M. Morf, L. Ljung and T. Kailath, "Fast Algorithms for Recursive Identification," Proc. IEEE Conf. on Decision and Control, Clearwater Beach, Florida, 1976.

[118] H. K. P. Neubert, *Instrument Transducers,* Clarendon Press, Oxford, 1975.

[119] V. Panuska, "Uniqueness of the Parameter Estimates Obtained from an Approximate Maximum Likelihood Identification Algorithm," Proc. Joint Automatic Control Conference, session WP19-1:50, 1977.

[120] B. Porat and B. Friedlander, "An Efficient Algorithm for Output Error Model Reduction," to appear.

[121] B. Porat and B. Friedlander, "Estimation of Spatial and Spectral Parameters of Multiple Sources," to appear.

[122] A. P. Sage and J. L. Melsa, *System Identification,* Academic Press, New York, 1971.

[123] G. N. Saridis, "Comparison of Six On-line Identification Algorithms," *Automatica,* vol. 10, pp. 69–79, 1974.

[124] G. N. Saridis, "Stochastic Approximation Methods for Identification and Control — A Survey," *IEEE Trans. Automat. Contr.*, vol. AC–19, No. 6, Dec. 1974.

[125] T. Soderstrom, I. Gustavson, and L. Jung, "A Comparative Study of Recursive Identification Methods," Report 7427, Lund Inst. of Tech., Lund, Sweden, Dept. of Auto. Control, 1974.

[126] T. Soderstrom, "On the Uniqueness of Maximum Likelihood Identification for Different Structures," Report 7307, Lund Inst. of Tech., Lund, Sweden, Div. of Auto. Control, March 1973.

[127] V. Solo, "The convergence of AML," *IEEE Trans. Automat. Contr.*, vol. AC-24, pp. 958–962, Dec. 1979.

[128] P. Stoica and T. Soderstrom, "The Steiglitz-McBride Algorithm Revisited — Convergence Analysis and Accuracy Aspects," *IEEE Trans. Automat. Contr.*, vol. AC-26, No. 3, pp. 712–717, June 1981.

[129] P. Stoica and T. Soderstrom, "Uniqueness of the Maximum Likelihood Estimates of ARMA Model Parameters—An Elementary Proof," *IEEE Trans. Automat. Contr.*, vol. AC-27, No. 3, pp. 736–738, June 1982.

[130] P. C. Young, "An Instrumental Variable Method for Real-time Identification of a Noisy Process," *Automatica*, vol. 6, pp. 271–287, 1970.

[131] K. Y. Wong and E. Polack, "Identification of Linear Discrete Time Systems using the Instrumental Variable Method," *IEEE Trans. Automat. Contr.*, vol. AC-12, pp. 707–718, Dec. 1967.

R.3. Mathematics and Statistics

[132] M. Abramowitz and I. A. Stegun, Ed., *Handbook of Mathematical Functions,* National Bureau of Standards, Washington D.C., 1964.

[133] L. V. Ahlfors, *Complex Analysis,* McGraw-Hill, New York, 1979.

[134] L. *Calculus,* Holt, Rinehart, and Winston, New York, 1969.

[135] G. E. P. Box, and G. M. Jenkins, *Time Series Analysis,* Holden-Day, San Francisco, CA, 1976.

[136] C. Caratheodory, *Theory of Functions, Volume I,* Chelsea, New York, 1964 (orig. 1950).

[137] C. Caratheodory, *Theory of Functions, Volume II,* Chelsea, New York, 1960 (orig. 1950).

[138] R. V. Churchill, *Complex Variables and Applications,* McGraw-Hill, New York, 1960.

[139] P. J. Davis and R. Hersch, *The Mathematical Experience,* Houghton Mifflin Co., Boston MA, 1981.

[140] N. Draper and H. Smith, *Applied Regression Analysis,* John Wiley and Sons, Inc., New York, 1966.

[141] P. Duren, *Theory of H^p Spaces,* Academic Press, New York, 1970.

[142] I. Gohberg and S. Goldberg, *Basic Operator Theory,* Birkhäuser, Boston, 1981.

[143] A. A. Goldstein, *Constructive Real Analysis,* Harper and Row, New York, 1967.

[144] R. M. Gray, "Toeplitz and Circulant Matrices: II," Tech. Rep. No. SEL–77–011, Information Systems Lab, Dept. of Elect. Eng., Stanford University, Stanford CA, April, 1977.

[145] U. Grenander and G. Szegö, *Toeplitz Forms and their Applications,* University of California Press, Berkeley and Los Angeles CA, 1958.

[146] P. Henrici, "Fast Fourier Methods in Computational Complex Analysis," *SIAM Review,* vol. 21, pp. 481–527, 1979.

[147] *Banach Spaces of Analytic Functions,* Prentice-Hall Inc., Englewood Cliffs, NJ, 1962.

[148] S. Karlin and H. M. Taylor, *A First Course in Stochastic Processes,* Academic Press, New York, 1975.

[149] A. M. Mood, F. A. Grayhill, and D. C. Boes, *Introduction to the Theory of Statistics,* McGraw-Hill, New York,1950,1963, 1974.

[150] Z. Nehari, *Conformal Mapping,* Dover, New York, 1952.

[151] B. Noble, *Applied Linear Algebra,* Prentice-Hall Inc., Englewood Cliffs, NJ, 1969, 1977.

[152] C. R. Rao, *Linear Statistical Inference and its Applications,* John Wiley and Sons, Inc., New York, 1965.

[153] H. L. Royden, *Real Analysis,* Macmillan, London, 1968.

[154] W. Rudin, *Principles of Mathematical Analysis,* McGraw-Hill, New York, 1964.

[155] G. Szegö, *Orthogonal Polynomials,* Amer. Math. Soc., Colloq. Publ. no. 23, New York, 1939.

[156] R. A. Tapia, "Course notes: Foundations of Optimization (Math Sci. 460)," Rice University, TX, 1975.

[157] G. B. Thomas, *Calculus and Analytic Geometry,* Addison-Wesley, Reading MA, 1972.

[158] R. E. Williamson, R. H. Crowell, and H. F. Trotter *Calculus of Vector Functions,* Prentice-Hall Inc., Englewood Cliffs, NJ, 1972.

R.4. Signal Processing and Computational Methods

[159] B. D. O. Anderson and J. B. Moore, *Optimal Filtering,* Prentice-Hall Inc., Englewood Cliffs, NJ, 1979.

[160] B. S. Atal and M. R. Schroeder, "Predictive Coding of Speech Signals and Subjective Error Criteria," *IEEE Trans. on Acoust., Speech, and Signal Proc.*, vol. ASSP-27, pp. 247–254, June 1979.

[161] M. Avriel, *Nonlinear Programming: Analysis and Methods,* Prentice-Hall Inc., Englewood Cliffs, NJ, 1976.

[162] F. L. Bauer, "Ein Direktes Iterationsverfahren Zur Hurwitz-Zerlegung Eines Polynoms," Mitteilung aus dem Mathematischen Institut der Technischen Hochschule Muenchen, *Archiv der Elektrischen,* Ubertragung, vol. 9, no. 1, pp. 285–290, 1955.

[163] M. J. Box, D. Davies, and W. H. Swann, *Nonlinear Optimization Techniques,* ICI Monograph no. 5, Oliver and Boyd, 1969.

[164] R. Bracewell, *The Fourier Transform and its Applications,* McGraw-Hill, New York, 1965.

[165] O. Brune, "Synthesis of a finite two terminal network whose driving point impedance is a prescribed function of frequency," *J. Math. and Phys.*, vol. 10, pp. 191-236, 1931.

[166] J. P. Burg, "Maximum Entropy Spectrum Analysis." Contained in [167].

[167] D. G. Childers, ed., *Modern Spectrum Analysis,* IEEE Press, New York, 1978.

[168] G. Dahlquist and A. Bjorck, *Numerical Methods,* Prentice-Hall Inc., Englewood Cliffs, NJ, 1974.

[169] Digital Signal Processing Committee, ed., *Programs for Digital Signal Processing,* IEEE Press, New York, 1979.

[170] J. L. Flanagan, *Speech Analysis, Synthesis, and Perception,* Springer-Verlag, New York, 1972.

[171] R. Fletcher and M. D. Powell, "A Rapidly Converging Descent Method for Minimization," *Comput. J.*, vol. 6, no. 2, pp. 163–168, July 1963.

[172] B. Friedlander, "A Lattice Algorithm for Factoring the Spectrum of a Moving Average Process," to appear.

[173] K. G. Gauss, *Theory of Motion of Heavenly Bodies,* Dover, New York, 1963.

[174] P. E. Gill, W. Murray, and M. H. Wright, *Practical Optimization,* Academic Press, New York, 1981.

[175] B. Gold and C. M. Rader, *Digital Processing of Signals,* McGraw-Hill, New York, 1969.

[176] G. Golub and W. Kahan, *Calculating the Singular Values and Pseudo-Inverse of a Matrix,* J. SIAM Numer. Anal. (Ser. B), vol. 2, no. 2, pp. 205–224, 1965.

[177] J. G. Graeme, G. E. Tobey, and L. P. Huelsman, ed., *Operational Amplifiers,* McGraw-Hill, New York, 1971.

[178] P. Henrici, "Fast Fourier Methods in Computational Complex Analysis," *SIAM Review,* vol. 21, pp. 481–527, 1979.

[179] S. L. S. Jacoby, J. S. Kowalik, and J. T. Pizzo, *Iterative Methods for Nonlinear Optimization Problems,* Prentice-Hall Inc., Englewood Cliffs, NJ, 1972.

[180] T. Kailath, "A View of Three Decades of Linear Filtering Theory," *IEEE Trans. on Info. Theory,* vol. IT-20, pp. , March 1974.

[181] T. Kailath, *Lectures on Linear Least-Squares Estimation,* Springer-Verlag, New York, 1976.

[182] T. Kailath, *Linear Systems,* Prentice-Hall Inc., Englewood Cliffs, NJ, 1980.

[183] G. E. Kopec, "Speech Analysis by Homomorphic Prediction," S. M. Thesis, Dept. Elec. Eng., MIT, Cambridge MA, 1975.

[184] K. Madsen and J. K. Reid, "Fortran Subroutines for Finding Polynomial Zeros," Report 7986, Computer Science and Systems Division, A.E.R.E. Harwell, Didcot, Oxford England, Feb. 1975.

[185] J. Makhoul, "Linear Prediction: A Tutorial Review," *Proc. IEEE,* vol. 63, pp. 561–580, Apr. 1975.

[186] J. D. Markel and A. H. Gray, *Linear Prediction of Speech,* Springer-Verlag, New York, 1976.

[187] J. D. Markel and A. H. Gray, "A Normalized Digital Filter Structure," *IEEE Trans. on Acoust., Speech, and Signal Proc.,* vol. ASSP-23, pp. 268–277, June 1975.

[188] J. D. Markel and A. H. Gray, "Roundoff Noise Characteristics of a Class of Orthogonal Polynomial Filter Structures," *IEEE Trans. on Acoust., Speech, and Signal Proc.,* vol. ASSP-23, pp. 473–486, Oct. 1975.

[189] J. D. Markel and A. H. Gray, "Fixed-Point Implementation Algorithms for a Class of Orthogonal Polynomial Filter Structures," *IEEE Trans. on Acoust., Speech, and Signal Proc.,* vol. ASSP-23, pp. 486–494, Oct. 1975.

[190] J. D. Markel and A. H. Gray, "Fixed-Point Truncation Arithematic Implementation of a Linear Prediction Autocorrelation Vocoder," *IEEE Trans. on Acoust., Speech, and Signal Proc.,* vol. ASSP-22, pp. 273–282, Aug. 1974.

[191] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing,* Prentice-Hall Inc., Englewood Cliffs, NJ, 1975.

[192] J. M. Ortega, *Numerical Analysis,* Academic Press, New York, 1972.

[193] S. M. Kay and S. L. Marple, "Spectrum Analysis—A Modern Perspective," *Proc. IEEE,* vol. 69, pp. 1380–1419, (278 references to the literature), Nov. 1981.

[194] G. A. Merchant and T. W. Parks, "Efficient Solution of a Toeplitz-Plus-Hankel Coefficient Matrix System of Equations," *IEEE Trans. on Acoust., Speech, and Signal Proc.*, vol. ASSP–30, pp. 40–44, Feb. 1982.

[195] A. Papoulis, *Signal Analysis,* McGraw-Hill, New York, 1977.

[196] L. R. Rabiner and B. Gold, *Theory and Application of Digital Signal Processing,* Prentice-Hall Inc., Englewood Cliffs, NJ, 1975.

[197] L. R. Rabiner, *Digital Processing of Speech Signals,* Prentice-Hall Inc., Englewood Cliffs, NJ, 1978.

[198] L. M. Silverman and M. Bettayeb, "Optimal Approximation of Linear Systems," to appear.

[199] B. T. Smith, et al., *Matrix Eigensystem Routines — EISPACK Guide,* Lecture Notes in Computer Science 6, 2nd ed., Springer-Verlag, New York, 1976.

[200] J. O. Smith, "Introduction to Digital Filters," to appear.

[201] K. Steiglitz, *An Introduction to Discrete Systems,* John Wiley and Sons, Inc., New York, 1974.

[202] G. W. Stewart, "Introduction to Matrix Computations," Academic Press, New York, 1973.

[203] M. E. Van Valkenburg, *Introduction to Modern Network Synthesis,* John Wiley and Sons, Inc., New York, 1960.

[204] G. T. Wilson, "Factorization of the Covariance Generating Function of a Pure Moving-Average Process," *SIAM J. Numer. Anal.*, vol. 6, no. 1, pp. 1–7, March 1969.

[205] G. T. Wilson, "The Factorization of Matricial Spectral Densities," *SIAM J. Appl. Math.*, vol. 23, no. 4, pp. 420–426, Dec. 1972.

[206] S. Zohar, "Fortran Subroutines for Solution of Toeplitz Sets of Linear Equations," *IEEE Trans. on Acoust., Speech, and Signal Proc.*, vol. ASSP–27, pp. 656–658, Dec. 1979. See also vol. ASSP-28, p. 601, and vol. ASSP-29, p. 1212.

R.5. Bowed Strings and the Violin

[207] I. Andersson, "Stick-Slip Motion in One-Dimensional Continuous Systems and in Systems with Several Degrees of Freedom," *Wear,* vol. 69, no. 2, pp. 255–256, June 1981.

[208] A. Benade, *Fundamentals of Musical Acoustics,* Oxford University Press, New York, 1976.

[209] A. Benade, "Fiddle Acoustics: The Summer of 1977," *Catgut Acoust. Soc. News Let.*, no. 28, Nov. 1977.

[210] J. S. Bradley and T. W. W. Stewart, "Comparison of Violin Response Curves Produced by Hand Bowing, Machine Bowing, and an Electromagnetic Driver," *J. Acoust. Soc. Amer.*, vol. 48, no. 2 pt 2, pp. 575–578, Aug 1970.

[211] G. Budzynski and A. Kulowski, "Bowed String as the Two-Terminal Oscillator," *Arch. Acoust.,*, vol. 2, no. 2, pp. 115–120, 1977.

[212] L. Cremer, "Bow Pressure Influence on the Self-Excited Vibrations of a String During Contact," *Acustica*, vol. 30, no. 3, pp. 119–136, March 1974.

[213] L. Cremer, "Fate of the Secondary Waves During Self-Excitation of String Instruments," *Acustica*, vol. 42, no. 3, pp. 133–148, May 1979.

[214] H. Dunnwald, "Research into the Origin of the Wolf-Tone of Violin Instruments," *Acustica*, vol. 41, no. 4, pp. 238-45, Jan. 1979.

[215] I. M. Firth, "The Action of the Cello at the Wolf Tone," *Acustica*, vol. 39, no. 4, pp. 252–263, Mar. 1978.

[216] I. M. Firth and J. M. Buchanan, "Wolf in the Cello," *J. Acoust. Soc. Amer.*, vol. 53, no. 2, pp. 457–463, Feb. 1973.

[217] I. M. Firth, "Mechanical Admittance Measurements on the Sound Post of the Violin, and its Action," *Acustica*, vol. 38, no. 5, pp. 332-9, Jan. 1977.

[218] F. G. Friedlander, "On the Oscillations of the Bowed String," Proc. Cambridge Philos. Soc., vol. 49, pp. 516–530, 1953.

[219] A. Gabrielsson and E. V. Jansson, "Long-Time-Average-Spectra and Rated Qualities of Twenty-Two Violins.," *Acustica*, vol. 42, no. 1, pp. 47–55, March 1979.

[220] C. E. Gough, "The Resonant Response of a Violin G-String and the Excitation of the Wolf Note," *Acustica*, vol. 44, no. 2, pp. 112–123, Feb. 1980.

[221] C. E. Gough, "The Acoustics of Stringed Instruments Studied by String Resonances," *Catgut Acoust. Soc. News Let.*, no. 35, pp. 22 May 1981.

[222] C. E. Gough, "The Theory of String Resonances on Musical Instruments," *Acustica*, vol. 49, no. 2, pp. 124–141, Oct. 1981.

[223] M. Hacklinger, "Violin Timbre and Bridge Frequency Response.," *Acustica*, vol. 39, no. 5, pp. 323-330, April 1978.

[224] C. M. Hutchins, "the Acoustics of Violin Plates," *Scientific American*, vol. 245, no. 4, pp. 126–135, Oct. 1981.

[225] J. B. Keller, "Bowing of Violin String," Comm. Pure Applied Math., vol. 6, pp. 483–495, 1953.

[226] J. Kohut and M. V. Mathews, "Study of Motion of a Bowed Violin String," *J. Acoust. Soc. Amer.*, vol. 49, no. 2 pt 2, pp. 532–537, Feb 1971.

[227] B. Lawergren, "On the Motion of Bowed Violin Strings," *Acustica*, vol. 44, no. 3, pp. 194–220, March 1980.

[228] J. C. Luke, "Measurement and Analysis of Body Vibrations of a Violin," *J. Acoust. Soc. Amer.*, vol. 49, no. 4 pt 2, pp. 1264–1274, Apr 1971.

[229] M. V. Mathews, "An Electronic Violin with a Singing Formant," invited paper presented at the Acoustical Society Conference no. 103, Chicago, Apr. 1982.

[230] M. E. McIntyre, R. T. Schumacher, and J. Woodhouse, "New Results on the Bowed String," *Catgut Acoust. Soc. News Let.*, no. 28, pp. 27–31, Nov. 1977.

[231] M. E. McIntyre and J. Woodhouse, "On the Fundamentals of Bowed String Dynamics," *Acustica*, vol. 43, no. 2, pp. 93–108, Sep. 1979.

[232] M. E. McIntyre, R. T. Schumacher, and J. Woodhouse, "Aperiodicity in Bowed-String Motion," *Acustica*, vol. 49, no. 1, pp. 13–32, Sep. 1981.

[233] M. E. McIntyre and J. Woodhouse, "The Influence of Geometry on Linear Damping," *Acustica*, vol. 39, no. 4, pp. 209–24, March 1978.

[234] M. E. McIntyre and J. Woodhouse, "The Acoustics of Stringed Musical Instruments," *Interdisciplinary Sci. Rev.*,, vol. 3, no. 2, pp. 157–73, June 1978.

[235] M. E. McIntyre, R. T. Schumacher, and J. Woodhouse, "On the Oscillations of Musical Instruments," to appear in JASA.

[236] N. C. Pickering, "Anomalies in the Frequency-Length Functions in Violin Strings," 70^{th} *Conv. of the Audio Eng. Soc.*, New York, Oct. 1981.

[237] N. C. Pickering, "A Computer-Controlled Violin Spectrum Analyzer," Norman C. Pickering, The Norman Pickering Company, Southampton, N.Y., 11968, Apr. 1982.

[238] C. V. Raman, "On the Mechanical Theory of Vibrations of Bowed Strings, etc.," *Indian Assoc. Cult. Sci. Bull.*, vol. 15, pp. 1–158, 1918.

[239] P. M. Ruiz, "A Technique for Simulating the Vibrations of Strings with a Digital Computer," *Master's Thesis*, Univ. Ill., 1969.

[240] J. C. Schelleng, "The Violin as a Circuit," *J. Acoust. Soc. Amer.*, vol. 35, pp. 326–338, 1963.

[241] J. C. Schelleng, "The Bowed String and the Player," *J. Acoust. Soc. Amer.*, vol. 53, pp. 26–41, Jan. 1973.

[242] R. T. Schumacher, "Self-Sustained Oscillations of the Bowed String.," *Acustica*, vol. 43, no. 2, pp. 109-20, Sep. 1979.

[243] R. T. Schumacher, "Ab Initio Calculations of the Oscillations of a Clarinet," *Acustica*, vol. 48, no. 2, pp. 71–85, May 1981.

[244] B. G. Seagrave and J. Berman, *Dictionary of Bowing Terms for Stringed Instruments*, American String Teachers Assoc., 1968, 2nd ed. 1976.

[245] J. O. Smith, "Synthesis of Bowed Strings," invited paper presented at the Acoustical Society Conference no. 103, Chicago, Apr. 1982.

R.8. Acoustics, Psychoacoustics, and Music

[246] R. Banek and J. Scoville, *Sound Designs*, Ten Speed Press, Berkeley CA, 1980.

[247] P. J. Bloom and D. Preis, "Perceptual Identification and Discrimination of Phase Distortions," *Proc. IEEE Int. Conf. Acoust. Speech and Sig. Proc.*, Boston MA, vol. 3, pp. 1396–1399, April 1983.

[248] C. Carterette and M. P. Friedman, ed., *Handbook of Perception, Volume IV: Hearing*, Academic Press, New York, 1978.

[249] J. P. Egan and H. W. Hake, "On the Masking Pattern of a Simple Auditory Stimulus," *J. Acoust. Soc. Amer.*, vol. 22, pp. 622–630, 1950.

[250] O. Ghitza and J. L. Goldstein, "Discrimination of Formant Frequency, Intensity, and Bandwidth in Natural Speech," presented at the Acoustical Society Conference no. 103, Chicago, Apr. 1982.

[251] L. Harrison, *Lou Harrison's Music Primer*, C. F. Peters Corp., 373 Park Ave. South, New York, 10016, 1971.

[252] H. L. F. von Helmholtz, *Die Lehre von den Tonempfindungen als Physiologische Grundlage für die Theorie der Musik*, F. Vieweg und Sohn, Braunschweig, 1863. English translation by A. J. Ellis, *On the Sensations of Tone as a Physiological Basis for the Theory of Music*, Reprinted by Dover Publications, New York, 1954.

[253] D. Jaffe and J. O. Smith, "Extensions of the Karplus-Strong Plucked String Algorithm," submitted to the *Computer Music Journal*.

[254] K. Karplus and A. Strong, "Digital Synthesis of Plucked String and Drum Timbres," submitted to the *Computer Music Journal*.

[255] J. B. Kruskal, "Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis," *Psychometrika*, vol. 29, no. 1, March, 1964.

[256] M. V. Mathews, *The Technology of Computer Music*, MIT Press, Cambridge MA, 1969.

[257] P. M. Morse, *Vibration and Sound*, published by the American Institute of Physics for the Acoustical Society of America, 1976 (1st ed. 1936, 2nd ed. 1948).

[258] J. O. Nordmark, "Frequency and Periodicity Analysis," in C. Carterette and M. P. Friedman, ed., *Handbook of Perception Volume IV: Hearing*, Academic Press, New York, pp. 243–282, 1978.

[259] J. C. Nunnally, *Psychometric Theory*, McGraw-Hill, New York, 1967.

[260] E. Paulus and E. Zwicker, "Programme zur Automatischen Bestimmung der Lautheit aus Terzpegeln oder Frequenzgruppenpegeln," *Acustica*, vol. 27, pp. 253–266, 1972.

[261] W. Piston, *Orchestration*, W. W. Norton and Co., New York, 1955.

[262] R. Plomp, "Timbre as a Multidimensional Attribute of Complex Tones," in R. Plomp and G. F. Smorrenburg, ed., *Frequency Analysis and Periodicity Detection in Hearing,* A. W. Sijthoff, Leiden, pp. 397–411, 1970.

[263] R. Plomp, *Aspects of Tone Sensation,* Academic Press, New York, 1976.

[264] L. C. W. Pols, "Perceptual Space of Vowel-Like Sounds and its Correlation with Frequency Spectrum," in R. Plomp and G. F. Smorrenburg, ed., *Frequency Analysis and Periodicity Detection in Hearing,* A. W. Sijthoff, Leiden, pp. 463–473, 1970.

[265] L. C. W. Pols, L. J. T. van der Kamp, and R. Plomp, "Perceptual and Physical Space of Vowel Sounds," *J. Acoust. Soc. Amer.*, vol. 46, pp. 458–467, 1969.

[266] L. C. W. Pols, H. R. C. Tromp, and R. Plomp, "Frequency Analysis of Dutch Vowels from 50 Male Speakers," *J. Acoust. Soc. Amer.*, vol. 53, pp. 1093–1101, 1973.

[267] D. Preis, "Phase Distortion and Phase Equalization in Audio Signal Processing—A Tutorial Review," *J. Audio Eng. Soc.*, vol. 30, no. 11, pp. 774–794, Nov. 1982.

[268] B. Scharf, "Loudness," in C. Carterette and M. P. Friedman, ed., *Handbook of Perception Volume IV: Hearing,* Academic Press, New York, pp. 187–242, 1978.

[269] J. O. Smith, J. W. Gordon, D. A. Jaffe, B. Mont-Reynaud, A. Schloss, B. Schottstaedt, and P. Wieneke, "Recent Research in Computer Music at CCRMA," CompCon Proc. IEEE Computer Soc., pp. 35–39, San Francisco, Feb. 1982.

[270] J. O. Smith and J. B. Angell, "A Constant Peak-Gain Digital Resonator Tuned by a Single Coefficient," *Computer Music J.*, vol. 6, no. 4, pp. 36–40, 1982.

[271] G. Weinreich, "Coupled Piano Strings," *J. Acoust. Soc. Amer.*, vol. 62, pp. 1474, 1977. Also *Scientific American,* vol. 240, p. 94, 1979.

[272] E. Zwicker, G. Flottorp, and S. S. Stevens, "Critical Bandwidth in Loudness Summation," *J. Acoust. Soc. Amer.*, vol. 29, pp. 548–557, 1957.

[273] E. Zwicker, "Ein graphisches Verfahren zur Bestimmung der Lautstärke und der Lautheit aus dem Terzpegeldiagramm," *Frequenz,* vol. 13, pp. 234–238, 1959.

[274] E. Zwicker and B. Scharf, "A Model of Loudness Summation," *Psych. Rev.*, vol. 72, pp. 3–26, 1965.

R.7. Miscellaneous

[275] J. B. Allen, private communication.

[276] Bernard Mont-Reynaud, private communication.

[277] David A. Jaffe, private communication.

[278] R. P. Gooch, private communication.

[279] R. A. Lanham, *Style: An Anti-Textbook*, Yale University Press, New Haven CT, 1974.

[280] X. Rodet, private communication.

[281] G. Weinreich, private communication.