# Some Geometric Questions in the Theory of Linear Systems

ROGER W. BROCKETT, FELLOW, IEEE

*Abstract*—In this paper we discuss certain geometrical aspects of linear systems which, even though they arise in the case of single input–single output systems, do not seem to have been explicitly recognized and studied before. We show, among other things, that the set of minimal, single input–single output, linear systems of degree $n$, when topologized in the obvious way, consists of $n+1$ connected components. The Cauchy index (equivalently, the signature of the Hankel matrix) characterizes the components and the geometry of each component is investigated. We also study the effect of various constraints such as asking that the system be stable or minimum phase.

## I. INTRODUCTION

CLASSICAL control theory deals with single input–single output systems characterized by a transfer function which is rational. In this theory there appear many representations of the rational functions, e.g., by means of partial fractions, pole and zero locations, numerator and denominator coefficients, Nyquist loci, etc. In fact, one may say that a large part of the mathematical technique used in classical control theory is concerned with passing back and forth between these various representations. This is necessary because each representation is convenient for investigating certain phenomena and inconvenient for others.

If one's concern is with the identification or tracking of linear systems it is not clear which, if any, of these representations is to be preferred. In this paper we undertake a study regarding the possible choices of coordinate systems in which to describe rational functions without common factors, including all the above possibilities as well as others. This work has as its goal the determination of adequate sets of coordinate systems for describing rational functions and rules for switching back and forth between possible coordinate systems. Our primary motivation in this paper is to develop a theory which will be useful in the study and implementation of system identification algorithms. We will return to this point later.

The main result to be established here is that the set of rational functions of a fixed degree $n$ (without common factors) is, when topologized in a natural way, actually the disjoint union of $n+1$ open sets. It is impossible to pass from one of these open sets to another without passing

through a region of common factors. As a refinement of this we investigate the set of rational functions without common factors, with denominators of degree $n$, and with numerators of fixed degree $m$ ($< n$), whose poles and zeros are constrained to lie in open, simply connected subsets of the complex plane.

The only prior work on this class of problems is apparently [1]–[3] of the references. Glover's work [2], [3] investigates certain related questions for multivariable systems and discusses the performance of system identification algorithms with this theory in mind. Precise definitions of topological terms used here can be found, for example, in [4].

## II. SECOND-ORDER SYSTEMS

Instead of starting directly with the general situation, we consider second-order systems characterized by a transfer function

$$g_2(s) = \frac{q_1 s + q_0}{s^2 + p_1 s + p_0}.$$

There is a four-parameter family of second-order systems but each choice of a four-tuple $(q_0, q_1, p_0, p_1)$ does not result in a second-order system because of the possibility of common factors. In order to avoid common factors we must parameterize $g_2(s)$ very carefully.

Consider the parameterization

$$g_2(s) = \frac{r(s + \sigma)\cos\theta + r\sin\theta}{(s + \sigma)^2 + \nu}.$$

In this case we see that as $\sigma$ varies over $(-\infty, \infty)$ and as $r$ ranges over $(0, \infty)$ no common factors appear or disappear. Thus, as a first attempt to understand the common factor problem we fix $\sigma = 0$ and $r = 1$. This gives a two-parameter family

$$g_{2r}(s) = \frac{s\cos\theta + \sin\theta}{s^2 + \nu}.$$

The values of $\theta$ and $\nu$ which correspond to common factors satisfy

$$\nu = -\tan^2\theta.$$

Since $\theta$ takes on values between $0$ and $2\pi$ with $0$ and $2\pi$ corresponding to the same point we think of $\theta$ as ranging over the circle $S^1$. Thus, we think of $(\theta, \nu)$ as ranging over

a cylindrical surface (see Fig. 1). The locus of common factors is then as indicated. It is important to note that the locus of common factors separates the space into three connected components, labeled here I, II, and III, and that these distinct components have essentially different geometries, regions I and III being simply connected while region II is not.

Notice that regions I and III are mapped into each other by the change of variables $\theta \to \theta + \pi$. Since $\nu$ is negative, in these regions the poles are real. A short calculation shows that the residues at these poles are both positive in region I and both negative in region III. In region II, however, the poles may be real or complex. If they are real, the residues have opposite signs.

This shows then that there are three basic types of second-order systems which we may parameterize in the following way:

$$g_{\mathrm{I}}(s) = \frac{e^{\alpha_1}}{s + \lambda_1} + \frac{e^{\alpha_2}}{s + \lambda_1 + e^{\lambda_2}}, \qquad (\alpha_1, \alpha_2, \lambda_1, \lambda_2) \in \Re^4$$

$$g_{\mathrm{II}}(s) = \frac{e^{\alpha}\left[(s + \sigma)\cos\theta + \sin\theta\right]}{(s + \sigma)^2 + \ln(e^{\beta} + e^{-\tan^2\theta})}, \qquad (\alpha, \beta, \sigma) \in \Re^3, \theta \in S^1$$

$$g_{\mathrm{III}}(s) = \frac{-e^{\alpha_1}}{s + \lambda_1} + \frac{-e^{\alpha_2}}{s + \lambda_1 + e^{\lambda_2}}, \qquad (\alpha_1, \alpha_2, \lambda_1, \lambda_2) \in \Re^4.$$

A given second-order system can be represented by exactly one of these formulas whereas each of these expressions defines a second-order system. There is no compelling reason for the choice of functions used here. We have simply made one choice which clearly reveals the geometry of the three regions.

## III. THE CAUCHY INDEX

By rat$(n)$ we mean the set of rational functions of the form

$$g_\alpha(s) = \frac{q_{n-1}^\alpha s^{n-1} + \cdots q_0^\alpha}{s^n + p_{n-1}^\alpha s^{n-1} + \cdots + p_0^\alpha}$$

with no common factors. We regard rat$(n)$ as a metric space with the metric

$$\rho(g_1, g_i) = \left(q_{n-1}^1 - q_{n-1}^2\right)^2 + \cdots + \left(q_0^1 - q_2^2\right)^2$$
$$+ \left(p_{n-1}^1 - p_{n-1}^2\right)^2 + \cdots + \left(p_0^1 - p_0^2\right)^2.$$

One way to visualize rat$(n)$ is as an open subset of $\Re^{2n}$ with its usual topology. That is,

$$\mathrm{rat}(n) = \{(q_{n-1}, \cdots, q_0, p_{n-1}, \cdots, p_0) : q(s) \text{ and}$$
$$p(s) \text{ have no common factors}\}$$

and the condition of not having a common factor simply subtracts a closed set from $\Re^{2n}$.

We now describe a way of associating an integer between $-n$ and $n$ with each element of rat$(n)$. The basic



Fig. 1. Illustrating the three types of second-order systems.

idea is due to Cauchy but it is convenient to give it a new twist. Regard $g(\cdot)$ as defining a mapping of $\Re^1$ into the circle $S^1$ according to the rule

$$\theta = -2\tan^{-1}g(x).$$

This associates with each rational function $g(\cdot)$ a continuous curve on a cylinder (see Figs. 2 and 3). The ambiguity of the definition of the inverse tangent is resolved by restricting it to lie between $-\pi/2$ and $\pi/2$. Since $g(-\infty) = g(\infty) = 0$ we may identify the ends of the cylinder so as to obtain from each $g(\cdot)$ a closed curve on a torus.

Associated with each rational function is an integer, the number of times this curve winds around the torus in the positive direction as $x$ increases from $-\infty$ to $\infty$. This winding number will be called the *Cauchy index*. A glance at the partial fraction expansion of $g(s)$ will convince the reader that in the case of unrepeated poles, the Cauchy index is simply the number of real poles with positive residues, minus the number of real poles with negative residues.

*Lemma 1:* The Cauchy index is a continuous map of rat$(n)$ into the real numbers.

*Proof:* Let $x_n = (1, x, x^2, \cdots, x^n)'$. Then in an obvious notation

$$g(x) = \frac{\langle \tilde{q}, x_n \rangle}{\langle \tilde{p}, x_n \rangle}.$$

For fixed $x$, the map of $\Re^{2n}$ into $S^1$ defined by

$$(q_{n-1}, \cdots, q_0, p_{n-1}, \cdots, p_0) : \mapsto -2\tan^{-1}g(x)$$

is continuous at all points except where $q(\cdot)$ and $p(\cdot)$ vanish simultaneously. Thus, on rat$(n)$ this function is continuous.

We now make the observation that included in the possible Cauchy indices are $-n$, $-n+2, \cdots, n-2$, $n$. These possibilities show up in the rational functions

$$g_k(x) = \sum_{r=-k}^{n-k} \frac{r}{s + r}, \qquad k = n, n-1, \cdots, 0$$

which have winding number $n - 2k$.

## IV. THE NUMBER OF CONNECTED COMPONENTS

In the case of first-order systems we have a representation

$$g(s) = \alpha / (s + \lambda), \qquad \alpha \neq 0.$$

Regarding the parameters $\alpha$ and $\lambda$ as points in the plane,
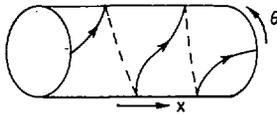
Fig. 2.   Typical curve on the cylinder.
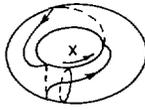


Fig. 3.   The curve on the torus.

we see that the set of all first-order systems consists of two connected components; one for $\alpha > 0$ and one for $\alpha < 0$. The respective winding numbers are one and minus one. We indicated in Section II evidence to the effect that the set of all second-order systems consist of three connected components. In this section we show that rat($n$), the set of all $n$th degree proper rational functions without common factors and with monic denominators, has $n+1$ connected components.

We carry out this proof in a number of short steps beginning with the following lemma.

*Lemma 2:* The set rat($n$) has at least $n+1$ distinct connected components.

*Proof:* From Lemma 1 we see that the Cauchy index is a continuous function on rat($n$). Since it is integer valued and takes on at least $n+1$ different values as $g$ varies over rat($n$) (from the example) we see that there must be at least $n+1$ distinct connected components of rat($n$).

We now introduce a more precise notation. By rat($p,q$) we understand the subset of rat($p+q$) corresponding to rational functions whose Cauchy index is $p-q$. We regard rat($p,q$) as a topological space with the topology described above.

*Theorem 1:* rat($p,q$) is arcwise connected. Thus, the set of all real proper rational functions of degree $n$ (without common factors) is the disjoint union of exactly $n+1$ arcwise connected open sets.

The previous lemma establishes that there are at least $n+1$ connected components; thus if we prove that rat($p,q$) is connected, we will have established the result. We introduce the following special terminology. An element $g$ of rat($n$) is called *typical* if all poles and all zeros are unrepeated and have distinct real and imaginary parts except for the real parts of complex conjugate poles and zeros and the imaginary parts of real poles and zeros. Elements of rat($n$) which are not typical are called *atypical*.

*Lemma 3:* It is possible to find in any neighborhood $N$ of $g \in$ rat($n$) an element $g_1 \in N(g)$ such that $g_1$ is typical.

*Proof:* The property of being atypical means that there is one or more algebraic relation satisfied by the coefficients of $g_1$. Thus, it is impossible to fill a neighborhood of $g$ with such elements.

*Lemma 4:* Any element $g \in$ rat($p,q$) can be deformed in rat($p,q$) to an element with all poles and zeros either purely real or purely imaginary.

*Proof:* We have observed that rat($n$), and hence rat($p,q$), is an open subset of $\mathcal{R}^{2n}$. Of course $\mathcal{R}^{2n}$ is locally arcwise connected (see [4, p. 55]) and thus rat($p,q$) is as well. Given $g$ we can by Lemma 3 and the local arcwise connectedness of rat($p,q$) deform it into a rational function which is typical. Then deform the typical element as indicated in Fig. 4; i.e., by moving the complex poles and zeros horizontally to the imaginary axis. This process involves no pole-zero cancellation because the imaginary parts of the poles and zeros are distinct. Thus, this procedure defines a path in rat($p,q$).

*Lemma 5:* Any element of rat($p,q$) can be deformed within rat($p,q$) to one with exactly $|p-q|$ real poles, exactly $|p-q|-1$ real zeros, the poles and zeros interlacing, and the complex poles and zeros purely imaginary.

*Proof:* If there are two poles on the real axis which are not separated by a zero we may move them together without introducing common factors. When they are together we may split them off into the complex plane. Following the device of the proof of Lemma 4 we may make them purely imaginary. The same is true of the zeros. Thus, we eliminate any adjacent poles and any adjacent zeros occuring on the real axis.

The number of poles on the real axis is, of course, $|p-q|$ since otherwise we would not have the correct Cauchy index. The number of zeros can be $|p-q|+1$ or $|p-q|-1$ since the total number for a typical element is $p+q-1$ and an even number of them are complex. If there are $|p-q|+1$ we proceed as indicated in Fig. 5. If the rightmost zero is $(\eta s + 1)$ then we let $\eta$ change sign to bring the zero through infinity reappearing on the left. Now there are two adjacent zeros and we may split them off into the complex plane and hence locate them on the imaginary axis.

The following lemma completes the proof of Theorem 1 since it shows that any two elements of rat($p,q$) can be deformed into a particular element and hence into each other.

*Lemma 6:* Any element of rat($p,q$) can be deformed within rat($p,q$) to

$$g(s) = \left[ \operatorname{sgn}(p-q) \right] \sum_{r=0}^{|p-q|} \frac{r}{s+r}$$

$$+ \frac{(s^2+2)(s^2+4)\cdots(s^2+2\nu)}{(s^2+1)(s^2+3)\cdots(s^2+2\nu+1)}$$

where $\nu$ is $\frac{1}{2}(n-|p-q|)$.

*Proof:* By using the path indicated in Fig. 6 we may split adjacent poles on the imaginary axis, thus arranging the poles and zeros in an alternating pattern on the imaginary axis with adjacent poles appearing only on either side of zero. Thus, an application of the deformations of Lemmas 5 and 6 gives a rational function whose real poles and zeros interlace on the positive half-axis and negative half-axis. To get the form given in the lemma
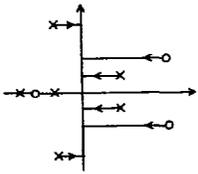
Fig. 4.   Illustrating the proof of Lemmas 4 and 5.
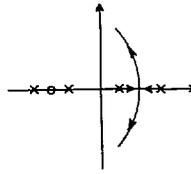


Fig. 5.   Passing a zero through infinity on the real axis.
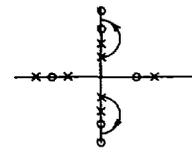


Fig. 6.   A step in producing the interlacing patterns.

statement we only need to shift the poles and zeros along the real and imaginary axes. Clearly, this can be done without leaving $\mathrm{rat}(p,q)$.

## V.   $\mathrm{rat}(p,q)$ AS A DIFFERENTIABLE MANIFOLD

In system identification problems one deals with approximations to actual systems by models which lie in a certain class. Very often the class is the class of linear time-invariant systems of a certain degree. It is desirable —in fact essential in some cases—to eliminate common factors when carrying out system identification. One reason for this is that the measure of the fit is insensitive to the location of a common factor of the form $(s+\alpha)/(s+\alpha)$ and thus the minimization problem becomes excessively ridge-like in one direction if common factors are allowed.

Thus, to deal with the identification problem we want a "good" coordinate system or, failing that, a set of "good" coordinate systems in terms of which we can describe $\mathrm{rat}(p,q)$. In short, we want to carry out the analysis of Section II for systems of arbitrary order.

To begin with let us observe that $\mathrm{rat}(p,q)$ has been constructed as an open subset of $\mathfrak{R}^{2n}$. Thus, at each point of $\mathrm{rat}(p,q)$ there is a local coordinate system, corresponding to the standard coordinate system of $\mathfrak{R}^{2n}$. If we want to construct a real analytic manifold which has $\mathrm{rat}(p,q)$ as its underlying topological space, we merely let the charts consist of all those charts compatible with the standard chart for $\mathfrak{R}^{n}$ (see [4, p. 98]). Thus, we see rat $(p,q)$ admits the structue of a $2(p+q)$-dimensional real analytic manifold. Our task is to identify useful coordinate charts with which to describe the manifold. Also, and this is as yet not completely understood, we want to describe how the local charts fit together to give us $\mathrm{rat}(p,q)$ as a whole.

In "most" of $\mathrm{rat}(p,q)$ partial fraction expansions are very convenient; however in a neighborhood of a repeated pole the partial fraction expansion must be modified in some way thus causing severe difficulties.

The only case in which we have a single natural parameterization for a whole component of $\mathrm{rat}(p,q)$ is when $p$ or $q$ is zero. Say $q$ is zero. Then

$$g(s)=\frac{e^{\alpha_1}}{s+\lambda_1}+\frac{e^{\alpha_2}}{s+\lambda_1+e^{\lambda_2}}+\cdots+\frac{e^{\alpha_n}}{s+\lambda_1+e^{\lambda_2}+\cdots+e^{\lambda_n}}$$

is, for each real choice of $\alpha_i$ and $\lambda_i$, an element of $\mathrm{rat}(n,0)$ and any element of $\mathrm{rat}(n,0)$ has a unique expression of this type. To parameterize $\mathrm{rat}(0,n)$ we insert a minus sign in front of the whole expression.

There is a second case which is more complicated but which can be analyzed in a quite satisfactory way. This is $\mathrm{rat}(n-1,1)\cong\mathrm{rat}(1,n-1)$. In this case we display $2n+1$ coordinate charts corresponding to a division of $\mathrm{rat}(n-1,1)$ into $2n+1$ overlapping pieces. Consider first of all the part of $\mathrm{rat}(n-1,1)$ consisting of those elements with a single pair of complex poles. This set is parameterized by

$$g(s)=\frac{r[(s+\sigma)\cos\theta+\sin\theta]}{(s+\sigma)^2+\ln(e^\nu+e^{-\tan^2\theta})}+\frac{e^{\alpha_1}}{s+\lambda_1}$$

$$+\cdots+\frac{e^{\alpha_n}}{s+\lambda_1+\cdots+e^{\lambda_n}}.$$

There are, in addition, $n$ distinct real pole regions

$$g(s)=\frac{-e^{\alpha_1}}{s+\lambda_1}+\frac{e^{\alpha_2}}{s+\lambda_1+e^{\lambda_2}}+\cdots+\frac{e^{\alpha_n}}{s+\lambda_1+e^{\lambda_2}+\cdots+e^{\lambda_n}}$$

$$g(s)=\frac{e^{\alpha_1}}{s+\lambda_1}-\frac{e^{\alpha_2}}{s+\lambda_1+e^{\lambda_2}}+\cdots+\frac{e^{\alpha_n}}{s+\lambda_1+e^{\lambda_2}+\cdots+e^{\lambda_n}}$$

$$\cdots$$

$$g(s)=\frac{e^{\alpha_1}}{s+\lambda_1}+\frac{e^{\alpha_2}}{s+\lambda_1+e^{\lambda_2}}+\cdots-\frac{e^{\alpha_n}}{s+\lambda_1+e^{\lambda_2}+\cdots+e^{\lambda_n}}.$$

We add to this set of coordinate charts $n$ more charts to cover what happens in the case where one is near a triply repeated pole. Notice that in $\mathrm{rat}(n-1,1)$ a triple pole would be of the form $[\alpha(s+\sigma)^2+\beta(s+\sigma)+\gamma]/(s+\sigma)^3$ with $\gamma>0$. The condition $\gamma<0$ is not compatible with the definition of $\mathrm{rat}(n-1,1)$. Thus, we add as charts

$$g(s)=\frac{(s+\lambda_1)^2+\beta(s+\lambda_1)+\gamma}{(s+\lambda_1)^3+\delta s^2+\epsilon s}+\frac{e^{\alpha_4}}{s+\lambda_1+e^{\lambda_4}}+\cdots$$

$$+\frac{e^{\alpha_n}}{s+\lambda_1+e^{\lambda_4}+\cdots+e^{\lambda_n}}$$

$$g(s) = \frac{e^{\alpha_1}}{s+\lambda_1} + \frac{\alpha(s+\lambda_1+e^{\alpha_2})^2 + \beta(s+\lambda_1+e^{\alpha_2})+\gamma}{(s+\lambda_1+e^{\lambda_2})^3 + \delta s^2 + \epsilon s} + \cdots$$

$$+ \frac{e^{\alpha_n}}{s+\lambda_1+e^{\lambda_2}+\cdots+e^{\lambda_n}}$$

$$\cdots$$

$$g(s) = \frac{e^{\alpha_1}}{s+\lambda_1} + \frac{e^{\alpha_2}}{s+\lambda_1+e^{\lambda_2}} + \cdots$$

$$+ \frac{e^{\alpha_{n-4}}}{s+\lambda_1+e^{\lambda_2}+\cdots+e^{\lambda_{n-4}}} + \frac{\alpha(s+\lambda_n)^2+\beta(s+\lambda_n)+\lambda}{(s+\lambda_n)^3+\delta s^2+\epsilon s}.$$

In these expressions we must restrict $\delta$ and $\epsilon$ to be sufficiently small. These $2n+1$ charts cover all of $\mathrm{rat}(n-1,1)$. There is overlap as indicated schematically in Fig. 7. These coordinate charts enable one to prove that the differentiable manifold $\mathrm{rat}(n-1,1)$ is topologically equivalent to $S^1 \times \mathfrak{R}^{2n-1}$ and thus implicitly give the "shape" of $\mathrm{rat}(n-1,1)$ or $\mathrm{rat}(1,n-1)$.

The cases $\mathrm{rat}(n-2,2)$ and $\mathrm{rat}(2,n-2)$ are even more complicated and are not yet completely understood.

## VI. Restricting the Pole-Zero Locations

Suppose one knows in advance that the poles of the transfer function lie in the left-half plane or within the unit circle. Or, suppose that the systems under consideration are all minimum phase so that the zeros lie in the left-half plane or within the unit circle. We may want to know how *a priori* assumptions of this type affect Theorem 1. Our goal here is to establish a suitable generalization of Theorem 1 so as to be able to handle questions of this type.

*Theorem 2:* Let $R_1$ and $R_2$ be two open, connected subsets of the complex plane which are each bounded and symmetric with respect to the real axis. Then the subset of $\mathrm{rat}(n)$ consisting of those rational functions having all their poles in $R_1$, exactly $r$ zeros in $R_2$, and no other zeros in the finite part of the plane has 1) two connected components if $R_1 \cap R_2 \cap$ {real axis} is empty, and 2) $2(r+1)$ connected components if $R_1 \cap R_2 \cap$ {real axis} is nonempty.

*Proof (sketch):* The idea of the proof is quite similar to the proof of Theorem 1. Typical examples of cases 1) and 2) are indicated in Fig. 8.

In case 1) we see that the Cauchy index must be 1 if $n$ is odd and 0 if $n$ is even, since there can be no interlacing poles and zeros on the real axis. If the Cauchy index is zero we note that

$$g(s) = \frac{q_r s^r + \cdots + q_0}{s^n + p_{n-1}s^{n-1} + \cdots p_0}$$

cannot be deformed into its negative without one or more of the zeros leaving the bounded set $R_2$. In particular we cannot change the sign of the coefficient of $s^r$ in the
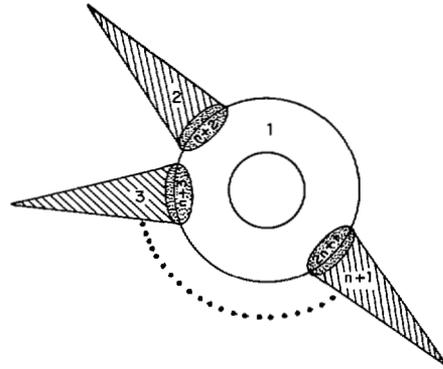


Fig. 7. Fitting together the coordinate charts of $\mathrm{rat}(n-1,1)$.
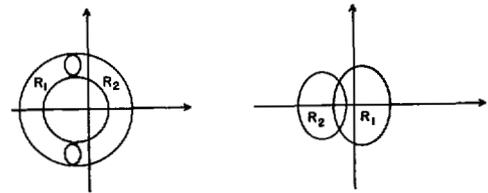


Fig. 8. Illustrating the hypothesis of Theorem 2.

numerator. Thus, if $n$ is even or odd there are at least two connected components—characterized for $n$ even by the sign of the coefficient of $s^r$ in the numerator and by the Cauchy index for $n$ odd. To show that there are only two connected components we note that we can deform any pole-zero pattern into a particular one in the given space as we did in the proof of Theorem 1. Since a pole-zero pattern, together with a scaling determines a rational function the description of the pole-zero deformation is all that is required to complete the proof. Since the proof of Theorem 1 provides one example of this type we omit further details here.

In case 1) we may have any Cauchy index of the correct parity between $-r-1$ and $r+1$, the limiting factor being the possible number of interlacing poles and zeros on the real axis. The question is that of determining how the restriction on the poles and zeros further divides the connected components of $\mathrm{rat}(n)$. To study this we may think of the special case where $R_1$ and $R_2$ are overlapping disks as indicated in Fig. 8. Following the proofs of the lemmas given above we see that everything proceeds as before (slight modification of typical being required) up to the point where we pass a real zero through infinity in the proof of Lemma 5. This is not possible here since the set $R_2$ is bounded. Thus, after we have removed from the real axis all possible poles and zeros as in Lemma 5 four types of configurations are possible.

1) zero, pole, $\cdots$, pole, zero;    $r+n$ is odd
2) zero, pole, $\cdots$, zero, pole;    $r+n$ is even
3) pole, zero, $\cdots$, zero, pole;    $r+n$ is odd
4) pole, zero, $\cdots$, pole, zero;    $r+n$ is even

where we include the possibility of zero length sequences under 2) and 3).

If $(n+r)$ is odd and $r$ is odd, we have $(r+1)/2$ possible configurations of type 1) and $(r+1)/2$ possible configurations of type 3) for a total of $r+1$. If $(n+r)$ is even and $r$ is odd we have $(r+1)/2$ possible configurations of type 2) and $(r+1)/2$ of type 4) for a total of $r+1$. Finally, if $(n+r)$ is even and $r$ is even we have $r/2+1$ possible configurations of type 2) and $r/2$ possible configurations of type 4). In all cases the total number is $r+1$.

In rat$(p,q)$ $g(s)$ can be deformed into $-g(s)$ if the Cauchy index is zero. Otherwise, each interlacing pole-zero sequence corresponds to two possible Cauchy indices, 1 and $-1$. In the present case each of the sequences 1)–4) again corresponds to two possible Cauchy indices if the Cauchy index is nonzero. However, even if the Cauchy index is zero, in the present case we cannot deform $g(s)$ into $-g(s)$ because we cannot pass a zero through infinity as we did before. Hence, even in this case each interlacing pole-zero configuration corresponds to two connected components. Putting together these remarks we see that there are at least $2(r+1)$ connected components of the subset of rat$(n)$ under consideration here.

The proof that each of the components identified is actually connected proceeds along the same lines as before; further details are omitted.

## VIII. HANKEL MATRICES

These results make contact with the state variable theory of linear dynamical systems in the following way.

Associated with each proper rational function of degree $n$ we have an expansion

$$g(s) = h_1 s^{-1} + h_2 s^{-2} + \cdots$$

and we have a family of Hankel matrices

$$H_m = \begin{bmatrix} h_1 & h_2 & h_3 & \cdots & h_m \\ h_2 & h_3 & h_4 & \cdots & h_{m+1} \\ h_3 & h_4 & h_5 & \cdots & h_{m+2} \\ \vdots & & & & \vdots \\ h_m & h_{m+1} & h_{m+2} & \cdots & h_{2m-1} \end{bmatrix}, \quad m = 1, 2, \cdots.$$

It has been known for a long time [5] and easily seen from the theory of controllability and observability, that the Hankel matrix $H_n$ corresponding to a proper rational function is of full rank if and only if the rational function is an $(n-1)$th degree polynomial divided by an $n$th degree polynomial with no common factors between the numerator and denominator. Since $H_n$ is a symmetric matrix its eigenvalues are real. Moreover, since the coefficients in the expansion of $g(\cdot)$ depend continuously on $g$, we will argue that two rational functions, say $g_1(s)$ and $g_2(s)$, which have associated Hankel matrices with differing numbers of positive eigenvalues cannot be in the same connected component of rat$(n)$. This comes about because to deform $g_1(s)$ into $g_2(s)$ one would, by necessity, deform $H_n^1$ into $H_n^2$. Since $H_n^1$ and $H_n^2$ do not have the same number of positive eigenvalues and since their eigen-

values are real, a zero eigenvalue (hence a singular $H_n$) would be encountered in deforming $H_n^1$ into $H_n^2$. Since $H_n$ being singular implies a common factor this means that there is at least one distinct component of rat$(n)$ for each possible signature of $H_n$.

From Theorem 1 we see that there must be a bijection between the set of Cauchy indices and the set of possible signatures of the associated $n \times n$ Hankel matrices. To determine what this bijection is we only need to check to which signature a particular element of rat$(p,q)$ corresponds. Notice that the realization of

$$g(s) = \sum_{i=1}^{n} \frac{\alpha_i}{s + \lambda_i}, \qquad \lambda_i, \alpha_i \text{ real}$$

which takes the form

$$\dot{x}_i = -\lambda_i \dot{x}_i + \sqrt{|\alpha_i|}\, u;$$

$$y(t) = \sum_{i=1}^{n} (\operatorname{sgn} \alpha_i) \sqrt{|\alpha_i|}\, x_i, \qquad i = 1, 2, \cdots, n$$

results in Hankel matrix of the form

$$H_n = (b, Ab, \cdots, A^{n-1}b)' \Sigma (b, Ab, \cdots, A^{n-1}b)$$

where

$$\Sigma = \operatorname{diag}(\operatorname{sgn} \alpha_1, \cdots, \operatorname{sgn} \alpha_n).$$

Thus, we see that the signature of $H_n$ is equal to the number of real poles with positive residues minus the number of real poles with negative residues. This gives a new "control theoretic" proof of a 19th century theorem of Hermite and Hurwitz (see Gantmacher [5, p. 210]).

*Theorem 3 (Hermite–Hurwitz):* The Cauchy index of $g(\cdot)$ is equal to the signature of the associated Hankel matrix.

## REFERENCES

[1] R. W. Brockett, "The geometry of the space of reduced rational functions," *Notices AMS*, vol. 22, no. 3, p. A417, 1975.
[2] K. Glover, "Structural aspects of system identification," Ph.D. dissertation, Dept. Elec. Eng., Massachusetts Inst. Technol., Cambridge, Sept. 1973.
[3] ——, "Some geometrical properties of linear systems with implications in identification," *1975 IFAC Congr., Preprints*, Aug. 1975.
[4] I. M. Singer and J. A. Thorpe, *Lecture Notes on Elementary Topology and Geometry.* New York: Scott.
[5] F. R. Gantmacher, *Theory of Matrices*, vol. II. New York: Chelsea, 1959.

**Roger W. Brockett** (S'62–M'63–SM'73–F'74) received the Ph.D. degree from Case Western Reserve University, Cleveland, OH.

From 1963 to 1969 he taught in the Department of Electrical Engineering at the Massachusetts Institute of Technology, Cambridge. He is presently a Gordon McKay Professor of Applied Mathematics in the Division of Engineering and Applied Physics, Harvard University, Cambridge, MA, where he participates in the Interdisciplinary Decision and

Control Program through teaching and research. He is coeditor (with H. H. Rosenbrock) of the series *Studies in Dynamical Systems* (Camden, NJ: Nelson) and is author of the book *Finite Dimensional Linear Systems* (New York: Wiley) and coeditor (with D. Q. Mayne) of *Geometric Methods in System Theory* (Dordrect, The Netherlands: Reidel), 1973.

He has held visiting positions at Warwick University, Imperial College, the University of Rome, and Washington University.

Dr. Brockett is recipient of the American Automatic Control Council's Donald P. Eckman Award. In 1975 he held a Guggenheim Fellowship for study in mathematical system theory.

# Max-Min Control Problems: A System Theoretic Approach

MICHAEL HEYMANN, MEIR PACHTER, AND RONALD J. STERN

*Abstract*—In this paper a "max-min controllability" concept for a situation in which two linear control systems are in conflict is introduced and characterized. This concept is employed in solving a max-min linear-quadratic control problem with terminal state constraints and the relationship with differential game theory is discussed.

## I. INTRODUCTION

CONSIDER the following linear system with dual controls:

$$\dot{x} = A(t)x + B_p(t)u + B_e(t)v, \qquad x(t_0) = x_0. \quad (1.1)$$

Here $x = x(t)$ is the state vector in Euclidean space $R^n$, with $x_0$ a specified initial state at time $t_0$. The vectors $u = u(t) \in R^{m_p}$ and $v = v(t) \in R^{m_e}$, regarded, respectively, as the pursuer and evader controls, are required to satisfy $\int_I \|u(t)\|^2 dt < \infty$ and $\int_I \|v(t)\|^2 dt < \infty$ on each compact interval $I \subset [t_0, \infty)$, where $\|\cdot\|$ denotes the Euclidean norm. The matrices $A$, $B_p$, and $B_e$ are assumed to have entries which are real and measurable on $[t_0, \infty)$. For any pair of controls $u$ and $v$ we shall denote by $x(t) = \phi(t, t_0, x_0, u, v)$ the corresponding unique solution of (1.1) emanating from $x_0$ at time $t_0$ ($t \geqslant t_0$).

In situations in which the pursuer and evader are in competition, it is natural to seek a comparison between their control capabilities. Towards this end we introduce the following concepts.

*Definition 1.1:* An event $(t_0, x_0)$ in system (1.1) is *strongly max-min controllable at time* $T$ ($T \geqslant t_0$) if for each announced evader control $v$ on $[t_0, T]$ there exists a pursuer control $u$ on $[t_0, T]$ such that $x(T) = \phi(T, t_0, x_0, u, v) = 0$. The event $(t_0, x_0)$ is *strongly max-min*

controllable if it is strongly max-min controllable for some $T \in [t_0, \infty)$.

*Definition 1.2:* An event $(t_0, x_0)$ in system (1.1) is *weakly max-min controllable* if for each announced evader control $v$ on $[t_0, \infty)$, there exists a time $\tilde{t} = \tilde{t}(v) \geqslant t_0$ and a pursuer control $u$ on $[t_0, \tilde{t}]$ such that $x(\tilde{t}) = \phi(\tilde{t}, t_0, x_0, u, v) = 0$.

Clearly strong max-min controllability of an event implies weak max-min controllability. That the converse is also true is not immediately evident since it is not clear that when weak max-min controllability holds there exists any one time $T$ at which capture (i.e., $x(T) = 0$) can be imposed by the pursuer in face of any evader control. This, however, is indeed the case as is shown in [1], and the two concepts of max-min controllability are actually equivalent. Henceforth, we will simply speak about *max-min controllability* referring to the simpler Definition 1.1.

It should be observed that max-min controllability generalizes the concept of controllability in linear control systems as expounded by Kalman (see, e.g., [2], [3]). While the existing "one player" controllability theory will be brought to bear on our development of the two player case, certain significant difficulties and interesting differences arise, as will be pointed out below.

Our results on max-min controllability will be employed in solving the following *restricted end-point max-min control problem*, denoted $(P)$.

$(P)$: We are given a linear dual control system (1.1) with $x_0 \neq 0$. The evader announces a control function $v$, and the pursuer (if he has the capability) responds with a control function $u$ such that $x(T) = 0$, where $T > t_0$ is a prespecified time. The players' control choices are to be made in accordance with the optimization of the payoff functional

$$P(u, v) \triangleq \int_{t_0}^T \left[ \|u(t)\|^2 - \|v(t)\|^2 \right] dt \quad (1.2)$$

where it is understood that the evader is the maximizing player while the pursuer is the minimizer.