General systems
or
Extending linear theory to nonlinear systems

S. T. Glad
Department of Electrical Engineering
Linköping University
S-581 83 Linköping, Sweden

September 19, 2012

Chapter 1

Solving the system equations

1.1 An algorithm

Consider a nonlinear system

$$\dot{x} = f(t, x, u) \tag{1.1}$$

where x is an n-vector. If we want a solution for a particular time function $u(t)$ we might as well include u in the time variability and study the system

$$\dot{x} = f(t, x) \tag{1.2}$$

Suppose we want to solve for a certain initial condition x_0. We then have

$$\dot{x} = f(t, x), \quad x(t_0) = x_0 \tag{1.3}$$

It turns out that it is more convenient to analyze the equivalent integral equation

$$x(t) = x_0 + \int_{t_0}^{t} f\big(\tau, x(\tau)\big) d\tau \tag{1.4}$$

This equation has a form that immediately suggests a method of successive approximations to get a solution. Let $x_j(t)$ be the approximation at the j:th iteration. Then the next iterate is defined by

$$x_{j+1}(t) = x_0 + \int_{t_0}^{t} f\big(\tau, x_j(\tau)\big) d\tau \tag{1.5}$$

The natural initialization of the iteration is

$$x_0(t) = x_0 \tag{1.6}$$

Consider therefore t- and x-values satisfying

$$t_0 \le t \le t_1, \quad |x - x_0| \le b \tag{1.7}$$

Assume that for all t and x satisfying (1.7) it is true that

$$|f(t,x)| \le M \tag{1.8}$$
$$|f(t,x_1) - f(t,x_2)| \le \Lambda |x_1 - x_2| \tag{1.9}$$

In (1.8) and (1.9) the vertical bars denote the Euclidian vector norm. The inequality (1.9) is usually called a *Lipschitz condition* on f.

The basic existence and uniqueness facts are given by the following theorem.

Theorem 1.1 *A differential equation (1.3), where f is continuous and satisfies (1.8), (1.9) in (1.7), has a unique solution on the interval $t_0 \le t \le t_0 + a$ if $a > 0$ is small enough. The iteration defined by (1.5), (1.6) converges to that solution.*

Proof.

a) We show that

$$|x_n(t) - x_0| \le b; \quad t_0 \le t \le t_0 + a$$

for all n, provided a is chosen small enough. Obviously this is true for $n = 0$. Suppose it is known for all integers up to n. Then

$$|x_{n+1}(t) - x_0| \le \int_{t_0}^{t} |f(\tau, x_n(\tau))| d\tau \le M \int_{t_0}^{t} d\tau \le aM \le b$$

provided $a \le b/M$.

b) We estimate the distance between the iterates. Having shown a) we know that we can apply (1.8), (1.9) to all the x_n. We will use the notation

$$||v|| = \max_{t_0 \le t \le t_0 + a} |v(t)| \tag{1.10}$$

Consider the difference between two iterates

$$|x_{n+1}(t) - x_n(t)| \le \int_{t_0}^{t} |f(\tau, x_n(\tau)) - f(\tau, x_{n-1}(\tau))| \quad d\tau \le$$
$$\le \Lambda \int_{t_0}^{t} |x_n(\tau) - x_{n-1}(\tau)| \quad d\tau \le a\Lambda ||x_n - x_{n-1}|| = \theta ||x_n - x_{n-1}|| \tag{1.11}$$

We choose a small enough to have $\theta = a\Lambda < 1$.

c) We show that the iterations converge to something. Using the estimate of b) repeatedly we get

$$||x_{n+1} - x_n|| \le \theta ||x_n - x_{n-1}|| \le \ldots \le \theta^n ||x_1 - x_0||$$

If $m > n$ then

$$||x_m - x_n|| \le ||x_m - x_{m-1}|| + \cdots + ||x_{n+1} - x_n|| \le (\theta^{m-1} + \ldots + \theta^n) ||x_1 - x_0|| \le$$
$$\le \frac{\theta^n}{1-\theta} ||x_1 - x_0|| \tag{1.12}$$

This expression converges to zero as n goes to infinity and $\{x_n\}$ is thus a Cauchy sequence. In particular, $x_n(t)$, for fixed t, is a Cauchy sequence of real numbers.

It then has to converge to some value $x(t)$. Since this holds for all t in the chosen interval, we have shown that

$$x_n(t) \to x(t), \quad t_0 \le t \le t_0 + a,$$

for some function $x(t)$.

d) Show that x is continuous and satisfies (1.4). Since

$$|x(t+h)-x(t)| \le |x(t+h)-x_n(t+h)|+|x_n(t+h)-x_n(t)|+|x_n(t)-x(t)| \le \\ \le 2||x-x_n|| + |x_n(t+h)-x_n(t)| \quad (1.13)$$

and each x_n is continuous, it follows that x is a continuous function.

Consider

$$|x_n(t)-x_0-\int_{t_0}^{t} f\big(\tau,x(\tau)\big)d\tau| \le \int_{t_0}^{t} |f(\tau,x_{n-1}(\tau))-f(\tau,x(\tau))| \quad d\tau \le \theta||x_{n-1}-x||$$

It follows that

$$x_n(t) \to x_0 + \int_{t_0}^{t} f\big(\tau,x(\tau)\big)d\tau$$

as $n \to \infty$. As $x_n \to x$ it follows that x satisfies (1.4).

e) Show that x is a unique solution. Suppose there are two solutions x and z. Then using the same reasoning as in step b),

$$||x-z|| \le \theta||x-z||$$

Since $\theta < 1$, this implies that $||x-z|| = 0$ and consequently that $x = z$. □

Remark 1.1 *If f is continuous but does not satisfy the Lipschitz condition (1.9), then one can still prove existence but the solution is not necessarily unique, as shown by the differential equation*

$$\dot{x} = \sqrt{x}, \quad x(0) = 0$$

which has the solutions

$$x = 0, \quad x = \frac{t^2}{4}$$

Remark 1.2 *Theorem 1.1 guarantees only local existence, since a has to be chosen small enough. In general there is no guarantee that a solution exists over an arbitrarily large time interval, as shown by the differential equation.*

$$\dot{x} = x^2, \quad x(0) = 1$$

The solution is

$$x = \frac{1}{1-t}$$

which only exists for $t < 1$.

1.2 Writing a nonlinear system as a (bi)linear one

There is an interesting way of representing a nonlinear system as an infinite dimensional linear one. Consider the nonlinear system

$$\dot{x} = -x + x^2 \tag{1.14}$$

The obvious linear approximation is of course

$$\dot{x}_1 = -x_1 \tag{1.15}$$

where x_1 approximates x. We can make the representation exact by writing

$$\dot{x}_1 = -x_1 + x_2 \tag{1.16}$$

where $x_2 = x^2$. Suppose we regard x_2 as a new variable and compute its derivative. We get

$$\dot{x}_2 = 2x\dot{x} = -2x^2 + 2x^3 = -2x_2 + 2x_3$$

where we have introduced $x_3 = x^3$. Continuing, introducing $x_4 = x^4$, $x_5 = x^5$ etc, we get

$$\begin{aligned}\dot{x}_3 &= 3x^2\dot{x} = -3x_3 + 3x_4\\ \dot{x}_4 &= 4x^3\dot{x} = -4x_4 + 4x_5\\ &\vdots\end{aligned}$$

With a matrix notation this becomes the linear system

$$\frac{d}{dt}\begin{bmatrix}x_1\\x_2\\x_3\\x_4\\\vdots\end{bmatrix} = \begin{bmatrix}-1 & 1 & 0 & 0 & 0 & 0 & \cdots\\ 0 & -2 & -2 & 0 & 0 & 0 & \cdots\\ 0 & 0 & -3 & 3 & 0 & 0 & \cdots\\ 0 & 0 & 0 & -4 & 4 & 0 & \cdots\\ \vdots & \vdots & & & & & \end{bmatrix}\begin{bmatrix}x_1\\x_2\\x_3\\x_4\\\vdots\end{bmatrix} \tag{1.17}$$

with an infinite dimensional state vector. If (1.17) is initialized with $x_1(0) = x(0)$, $x_2(0) = x(0)^2$, $x_3(0) = x(0)^3$ etc. we ought to get the same solution as for the nonlinear system (1.14), provided the infinite dimensional calculations implied by (1.17) make sense. The linear system (1.17) is called a *Carleman linearization* of (1.14). Of course it is possible to look at truncated versions of the Carleman linearization, e.g. the system

$$\frac{d}{dt}\begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix} = \begin{bmatrix}-1 & 1 & 0\\ 0 & -2 & -2\\ 0 & 0 & -3\end{bmatrix}\begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix}$$

This system should be a better approximation of (1.14) than the straightforward linear approximation (1.15).

Next consider the situation with a control signal present

$$\dot{x} = -x + x^2 + u \tag{1.18}$$

The straightforward linear approximation now becomes

$$\dot{x}_1 = -x_1 + u \tag{1.19}$$

Introducing $x_2 = x^2$, $x_3 = x^3$ etc. gives

$$\dot{x}_2 = 2x\dot{x} = -2x^2 + 2x^3 + 2xu = -2x_2 + 2x_3 + 2x_1u$$

This no longer a linear system, due to the term $2x_1u$. If we continue the calculations of $\dot{x}_3$, $\dot{x}_4$,.. we get the infinite system

$$\frac{d}{dt}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ \vdots \end{bmatrix} = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 & 0 & \cdots \\ 0 & -2 & -2 & 0 & 0 & 0 & \cdots \\ 0 & 0 & -3 & 3 & 0 & 0 & \cdots \\ 0 & 0 & 0 & -4 & 4 & 0 & \cdots \\ \vdots & \vdots & & & & & \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ \vdots \end{bmatrix} + u \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ 2 & 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 3 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 4 & 0 & 0 & 0 & \cdots \\ \vdots & \vdots & & & & & \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ \vdots \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ \vdots \end{bmatrix} u \tag{1.20}$$

This is an infinite dimensional *bilinear* system, sometimes called the *Carleman bilinearization*. The general form of an n-dimensional bilinear system with a scalar input is

$$\dot{x} = Ax + u\ Dx + Bu \tag{1.21}$$

where x, B are n-vectors and A, D are $n \times n$-matrices. Of course it is possible to construct truncated versions of (1.20), e.g.

$$\frac{d}{dt}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -1 & 1 & 0 \\ 0 & -2 & -2 \\ 0 & 0 & -3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + u \begin{bmatrix} 0 & 0 & 0 \\ 2 & 0 & 0 \\ 0 & 3 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} u \tag{1.22}$$

To get a Carleman bilinearization of a system with more than one state variable is in principle straightforward. If we have two states x_1 and x_2 we have to consider time derivatives of x_1^2, x_1x_2 and x_2^2 to get a second order bilinearization. For a third order bilinearization we need derivatives of x_1^3, $x_1^2x_2$, $x_1x_2^2$, x_2^3 and so on. Let us consider an example.

Example 1.1 Consider a very simplified model for velocity control of an aircraft. If the velocity is x_1 and the mass normalized to 1, then

$$\dot{x}_1 = x_2 - f(x_1) \tag{1.23}$$

where x_2 is the engine thrust and $f(x_1)$ is the aerodynamic drag. A simplified engine model is just a time constant from pilot command u to engine thrust:

$$\dot{x}_2 = -x_2 + u \tag{1.24}$$

Together (1.23) and (1.24) form a model of the aircraft velocity control. Now assume that x_1 and x_2 are deviations from a nominal velocity and thrust, and approximate f with $f(x_1) = x_1 + x_1^2$, giving the model

$$\begin{aligned} \dot{x}_1 &= -x_1 - x_1^2 + x_2 \\ \dot{x}_2 &= -x_2 + u \end{aligned} \tag{1.25}$$

Introducing $z_1 = x_1$, $z_2 = x_2$, $z_3 = x_1^2$, $z_4 = x_1 x_2$ and $z_5 = x_2^2$ we have

$$\begin{aligned} \dot{z}_3 &= 2x_1\dot{x}_1 = -2x_1^2 - 2x_1^3 + 2x_1x_2 = -2z_3 + 2z_4 - 2x_1^3 \\ \dot{z}_4 &= \dot{x}_1x_2 + x_1\dot{x}_2 = -2x_1x_2 - x_1^2x_2 + x_2^2 + x_1u = -2z_4 + z_5 + z_1u - x_1^2x_2 \\ \dot{z}_5 &= -2x_2^2 + 2x_2u = -2z_5 + 2z_2u \end{aligned}$$

Neglecting third order terms we get the following truncated Carleman bilinearization.

$$\dot{z} = \begin{bmatrix} -1 & 1 & -1 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & -2 & 2 & 0 \\ 0 & 0 & 0 & -2 & 1 \\ 0 & 0 & 0 & 0 & -2 \end{bmatrix} z + u \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \end{bmatrix} x + \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} u$$

□

The generalization of Carleman bilinearizations to nonlinear systems of the form

$$\dot{x} = f(x) + g(x)u, \quad y = h(x)$$

is straightforward. Let $x_{(j)}$ denote the vector of all homogeneous degree j monomials in x, i. e.

$$\begin{aligned} x_{(1)} &= x \\ x_{(2)} &= (x_1^2 \;\; x_1x_2 \;\; \dots \;\; x_1x_n \;\; x_2^2 \;\; x_2x_3 \dots \;\; x_n^2)^T \\ x_{(3)} &= (x_1^3 \;\; x_1^2x_2 \;\; x_1^2x_3 \;\; \dots x_2^3 \;\; x_2^2x_3 \;\; \dots \;\; x_n^3)^T \\ &\vdots \end{aligned}$$

Proposition 1.1 *Consider a system*

$$\dot{x} = f(x) + u \;\; g(x), \quad y = h(x) \tag{1.26}$$

where u and y are scalars. Assume that $f(0) = 0$ and $h(0) = 0$, i. e. the origin is an equilibrium corresponding to $u = 0$ and $y = 0$. Also assume that f and g are analytic, i. e. they can be expanded into convergent power series:

$$\begin{aligned} f(x) &= F_1x + F_2x_{(2)} + F_3x_{(3)} + \cdots \\ g(x) &= g(0) + G_1x + G_2x_{(2)} + G_3x_{(3)} + \cdots \\ h(x) &= H_1x + H_2x_{(2)} + H_3x_{(3)} + \cdots \end{aligned}$$

where the F_i, G_i and H_i are matrices of suitable dimensions. Then the Carleman bilinearization of the system has the form

$$\dot{z} = \begin{bmatrix} F_1 & F_2 & F_3 & \dots \\ 0 & A_{21} & A_{23} & \dots \\ 0 & 0 & A_{33} & \dots \\ 0 & 0 & 0 & \ddots \\ \vdots & \vdots & & \end{bmatrix} z + \begin{bmatrix} G_1 & G_2 & G_3 & \dots \\ B_{20} & B_{21} & B_{22} & \dots \\ 0 & B_{30} & B_{31} & \dots \\ 0 & 0 & B_{40} & \dots \\ \vdots & \vdots & \vdots & \ddots \\ 0 & 0 & \dots & \end{bmatrix} z\, u + \begin{bmatrix} g(0) \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} u \quad (1.27)$$

where the elements of the j:th block row are obtained by differentiating $x_{(j)}$. The output equation is

$$y = \begin{bmatrix} H_1 & H_2 & H_3 & \dots \end{bmatrix} z \quad (1.28)$$

where z represents the vector

$$z = \begin{bmatrix} x^T & x_{(2)}^T & x_{(3)}^T & \cdots \end{bmatrix}^T \quad (1.29)$$

Proof. The expression for y follows immediately from the series expansion of h. Likewise the first row of (1.27) follows from the expansions of f and h. Now consider the time derivative of a degree k monomial

$$\frac{d}{dt}\left(x_1^{j_1} \cdots x_n^{j_n}\right) = j_1 x_1^{j_1-1} \cdots x_n^{j_n} \dot{x}_1 + \cdots + j_n x_1^{j_1} \cdots x_n^{j_n-1} \dot{x}_n$$

The right hand side of this expression is a sum of degree $k-1$ monomials multiplied by rows of $f(x) + ug(x)$. Terms not containing u will then be of degree k or higher while terms containing u will be of degree $k-1$ or higher. □

1.3 Exercises

1.1 Apply the iteration (1.5) to the differential equation

$$\dot{x} = 1 + x^2, \quad x(0) = 0$$

What do the iterations converge to?

1.2 Compute the second order Carleman bilinearization of the system

$$\dot{x}_1 = x_2$$
$$\dot{x}_2 = -x_2 + x_2^2 + u$$

1.3 Compute the second order Carleman bilinearization of the system

$$\dot{x}_1 = x_2^2$$
$$\dot{x}_2 = u$$

Chapter 2

Observability

In dynamical systems there are often physical variables that are not directly measured. In many situations it is important to know if their values can be computed from the measurements. Consider for example the system

$$\dot{x}_1 = x_2^2, \quad \dot{x}_2 = 1, \quad y = x_1 \tag{2.1}$$

The variable x_1 is directly measured. By differentiating we get

$$\dot{y} = x_2^2$$

This does not completely determine x_2 unless we know its sign a priori. However, by differentiating once more we get

$$\ddot{y} = 2x_2$$

which determines x_2 precisely. The example illustrates the standard method of analyzing observability of nonlinear systems — compute enough derivatives of the output and try to determine the state from them. Note that this method only determines whether it is possible to compute the state in principle. To do the computation in practice, when the output is always to some extent corrupted by noise, might require different methods. To proceed it is necessary to define observability more precisely.

2.1 Definition of observability.

Let the system description be

$$\dot{x} = f(x, u), \quad y = h(x) \tag{2.2}$$

where x is an n-vector, u an m-vector, y a p-vector and where f and h are infinitely differentiable functions. Let the solution of the differential equation with the initial state x_0 and the input u be denoted $\pi(t; x_0, u)$. Two points in the state space, x_1 and x_2, are said to be *indistinguishable* if they give rise to the same output, i.e.

$$h\big(\pi(t; x_1, u)\big) = h\big(\pi(t; x_2, u)\big)$$

for all $t \geq 0$ and *for all inputs* u. The set of all points that are indistinguishable from x is denoted $I(x)$. The following definition of observability is now natural.

Definition 2.1 *The system (2.2) is* observable at x_0 *if* $I(x_0) = \{x_0\}$. *It is called* observable *if this is true for all points* x_0.

The disadvantage with this definition is shown by the following example.

Example 2.1 Consider the scalar system

$$\dot{x} = 1, \quad y = \begin{cases} 0 & \text{if } x \leq 0 \\ x^2 & \text{if } x > 0 \end{cases}$$

The points $x_1 = -10^{10}$ and $x_2 = -1 - 10^{10}$ are clearly distinguishable, because, for $t > 10^{10}$ the outputs will be different. Up to that time, however, they will be exactly the same. □

To avoid situations like this one, where it is necessary to wait for a very long time to distinguish different states, a more demanding concept of observability is introduced.

Definition 2.2 *Let* U *be an open set. Two points* x_1 *and* x_2 *which both belong to* U *are said to be* U-indistinguishable *if they give the same outputs in all cases where both trajectories lie entirely in* U*, i.e.*

$$h\big(\pi(t; x_1, u)\big) = h\big(\pi(t; x_2, u)\big), \quad t \in [t_0, t_1]$$

as soon as

$$\pi(t; x_1, u) \in U, \quad \pi(t; x_2, u) \in U, \quad t \in [t_0, t_1]$$

The set of all points that are U-indistinguishable from x_0 is denoted $I_U(x_0)$. The system (2.2) is *locally observable at* x_0 if $I_U(x_0) = \{x_0\}$ for every open neighborhood U of x_0. If this is true at every point x_0, the system is said to be *locally observable*.

Note that local observability is a tougher requirement than just observability. Essentially local observability implies that it is possible to determine x from y instantaneously, which is what one wants in observers and filters. In one way local observability might be an unnecessarily strict condition however. In many cases x is approximately known before measurements are made, and then it is only necessary to use y to distinguish between states that are close to each other. This leads to one further definition.

Definition 2.3 *The system (2.2) is* locally weakly observable at x_0 *if there exists an open neighborhood* U *of* x_0 *such that for every neighborhood* V *of* x_0 *with* $V \subset U$, $I_V(x_0) = \{x_0\}$. *If this is true for all points* x_0, *the system is* locally weakly observable.

The physical interpretation of local weak observability is that the state x can be instantaneously distinguished from other nearby states by a look at the output y. It turns out that this is the most useful concept for nonlinear systems. One reason for this is that there exist simple tests, as we will see.

2.2 Testing observability.

Consider the system (2.2) and differentiate the output. This gives

$$\dot{y} = h_x(x)\dot{x} = h_x(x)f(x,u)$$

where h_x denotes the derivative:

$$h_x = \left[\tfrac{\partial h}{\partial x_1}, \ldots, \tfrac{\partial h}{\partial x_n}\right]$$

Define the function $h^{(1)}(x,u) = h_x(x)f(x,u)$ and differentiate once more

$$\ddot{y} = h_x^{(1)}(x,u)f(x,u) + h_u^{(1)}(x,u)\dot{u}$$

Defining $h^{(2)}(x,u,\dot{u}) = h_x^{(1)}(x,u)f(x,u) + h_u^{(1)}(x,u)\dot{u}$ and again differentiating gives

$$y^{3)} = h_x^{(2)}(x,u,\dot{u})f(x,u) + h_u^{(2)}(x,u,\dot{u})\dot{u} + h_{\dot{u}}^{(2)}(x,u,\dot{u})\ddot{u}$$

Proceeding in this fashion gives a system of equations.

$$\begin{aligned} \dot{y} &= h(x) \\ \dot{y} &= h^{(1)}(x,u) \\ \ddot{y} &= h^{(2)}(x,u,\dot{u}) \\ &\vdots \\ y^{(N)} &= h^{(N)}(x,u,\dot{u},\ldots,u^{(N-1)}) \end{aligned} \tag{2.3}$$

where the $h^{(i)}$ are recursively defined by

$$h^{(i+1)} = h_x^{(i)} f + h_u^{(i)}\dot{u} + \cdots + h_{u^{(i-1)}}^{(i)} u^{(i)}, \quad h^{(0)} = h \tag{2.4}$$

In principle (2.3) gives a test for observability: If it is possible to find an N such that (2.3) can be solved for x (with u, y and their derivatives regarded as known) then the system is locally observable. If we can show that (2.3) can be solved locally, then we get local weak observability.

Since it is difficult to analyze nonlinear systems of equations, one often looks at the linearized version of (2.3), i.e. at the Jacobian

$$J(x,u,\ldots,u^{(N-1)}) = \begin{bmatrix} h_x(x) \\ h_x^{(2)}(x,u,\dot{u}) \\ \vdots \\ h_x^{(N)}(x,u,\dot{u},\ldots,u^{(N-1)}) \end{bmatrix} \tag{2.5}$$

This gives a basic observability test.

Theorem 2.1 *Suppose there is a choice of N and u such that $J(x_0,u,\ldots,u^{(N-1)})$ has full rank. Then the system is locally weakly observable at x_0.*

Proof. Consider definition 2.3. If there is some point $\bar{x}$ which is indistinguishable from x_0, then we must have

$$\begin{aligned} y = h(x_0) &= h(\bar{x}) \\ \dot{y} = h^{(1)}(x_0, u) &= h^{(1)}(\bar{x}, u) \\ &\vdots \\ y^{(N)} = h^{(N)}(x_0, u, \dot{u}, \ldots, u^{(N-1)}) &= h^{(N)}(\bar{x}, u, \dot{u}, \ldots, u^{(N-1)}) \end{aligned}$$

since the same output function $y(t)$ is generated from both points. The nonlinear system of equations, whose Jacobian is J, thus has two solutions. Since J has full rank, the implicit function theorem shows that this is impossible if the set U of definition 2.3 is chosen small enough. □

Example 2.2 Consider the system

$$\dot{x}_1 = x_2, \quad \dot{x}_2 = 0, \quad y = x_1^2$$

We get

$$h^{(0)} = x_1^2, \quad h^{(1)} = 2x_1x_2, \quad h^{(2)} = 2x_2^2,$$

and note that

$$h^{(k)} = 0, \quad k > 2$$

There is thus no point in computing J for N greater than 2. For $N = 2$ we get

$$J = \begin{bmatrix} 2x_1 & 0 \\ 2x_2 & 2x_1 \\ 0 & 4x_2 \end{bmatrix}$$

This matrix has full rank except for $x_1 = x_2 = 0$. The system is thus locally weakly observable at every point except possibly the origin. In general it would not be possible to draw any further conclusion. In this particular case the system is in fact not locally weakly observable at the origin. We can verify this by actually solving the differential equation for the initial condition $x_1(0) = x_{10}$, $x_2(0) = x_{20}$, getting

$$y(t) = (x_{10} + x_{20}t)^2$$

We see that every point x_{10}, x_{20} is indistinguishable from its mirror point $-x_{10}, -x_{20}$ (reflection in the origin). Since every neighborhood of the origin contains pairs of points that are indistinguishable in this way, the system is not locally weakly observable at the origin. We also see that the system is not globally observable. □

2.3 Testing observability using Lie derivatives.

There is a variation of the observability test represented by Theorem 2.1 which uses Lie derivatives. The Lie derivative of the scalar function $\phi(x)$ in the direction given by the n-vector $f(x)$ is defined by

$$(L_f\phi)(x) = \phi_x(x)f(x)$$

Now consider the system (2.2), let the control be constant, $u = u_1$ and define $f_1(x) = f(x, u_1)$. Let $x(0) = x_0$. Then the time derivatives of the output can be written in terms of the Lie derivative:

$$\begin{aligned} \dot{y} &= h_x \dot{x} = h_x f_1 = L_{f_1} h \\ \ddot{y} &= (L_{f_1} h)_x f_1 = L_{f_1}^2 h \\ &\vdots \\ y^{(k)} &= (L_{f_1}^k h) \end{aligned} \tag{2.6}$$

If the control is changed between different constant values, a slightly more involved formula is the result.

Proposition 2.1 *If the system is initialized at x_0 and the control signal is u_1 for t_1 units of time, u_2 for t_2 units of time,...,u_k for t_k units of time, then*

$$\left(\frac{\partial^k}{\partial t_1 \cdots \partial t_k} y(t_1 + t_2 + \cdots + t_k) \right) \Big|_{t_1 = \cdots = t_k = 0} = \left(L_{f_1} L_{f_2} \ldots L_{f_k} h \right)(x_0)$$

Proof. If $k = 1$ this is the first row of (2.6). Let $k = 2$. For a general differentiable function of two variables, $\phi(t_1, t_2)$ it follows from definitions that

$$\frac{\partial^2 \phi}{\partial t_1 \partial t_2} \Big|_{t_1 = t_2 = 0} = \frac{\partial}{\partial t_1} \left(\frac{\partial \phi}{\partial t_2} \Big|_{t_2 = 0} \right) \Big|_{t_1 = 0}$$

Using this fact for the function

$$y(t_1 + t_2) = h(\pi_2(t_2, \pi_1(t_1, x_0)))$$

(where π_1 and π_2 are the solutions corresponding to u_1 and u_2 respectively) we compute first

$$\frac{\partial}{\partial t_2} h(\pi_2(t_2, \pi(t_1, x_0))) \Big|_{t_2 = 0} = (L_{f_2} h)(\pi_1(t_1, x_0))$$

using (2.6). Using again (2.6), with h replaced by $L_{f_2} h(\pi_1(t_1, x_0))$, we get

$$\frac{\partial}{\partial t_1} L_{f_2} h(\pi_1(t_1, x_0)) \Big|_{t_1 = 0} = (L_{f_1} L_{f_2} h)(x_0)$$

This quantity is then equal to

$$\frac{\partial^2}{\partial t_1 \partial t_2} y(t_1 + t_2) \Big|_{t_1 = t_2 = 0}$$

For a general k the proposition is proved by repeated use of the same argument. The extension to a vector valued y is straightforward. □

It is now natural to introduce some notation for the functions that are generated by successive Lie differentiation.

Definition 2.4 *Let $\mathcal{G}$ denote the collection of all functions of the form*

$$h, \; L_{f_1} h, \; L_{f_2} h, \; \ldots, \; L_{f_1} L_{f_2} h, \; \ldots, \; L_{f_1} L_{f_2} \cdots L_{f_k} h, \; \ldots$$

with the f_i corresponding to all possible choices of constant controls u_i.

The connection between this set of functions and observability is given by the following fact.

Proposition 2.2 *Let x_1 and x_2 be two points in the open set U. If they are U-indistinguishable then not only is $h(x_1) = h(x_2)$ but also*

$$\phi(x_1) = \phi(x_2) \quad \textit{for all} \quad \phi \in \mathcal{G}$$

Proof. Let the control signal be chosen as in Proposition 2.1 and let y_1 and y_2 be the outputs with the initial conditions x_1 and x_2 respectively. Since

$$y_1(t_1 + t_2 + \cdots + t_k) \equiv y_2(t_1 + t_2 + \cdots + t_k)$$

successive derivatives with respect to t_i are also equal. Then it follows from Proposition 2.1 that

$$\left(L_{f_1} L_{f_2} \ldots L_{f_k} h\right)(x_1) = \left(L_{f_1} L_{f_2} \ldots L_{f_k} h\right)(x_2)$$

□

As in analyzing (2.3) it is easier to study the Jacobians of the functions in $\mathcal{G}$.

Definition 2.5 *The system (2.2) satisfies the* observability rank condition *at x_0 if among all the row vectors of the form ϕ_x, where ϕ is any element in $\mathcal{G}$, there are n linearly independent elements.*

Finally we are ready to state a criterion for local weak observability in terms of repeated Lie derivatives.

Theorem 2.2 *If the system (2.2) satisfies the observability rank condition at x_0, then it is locally weakly observable at x_0.*

Proof.. Choose n functions ϕ_1,..,ϕ_n in $\mathcal{G}$ such that their derivatives are linearly independent. Form the function

$$\Phi = \begin{pmatrix} \phi_1 \\ \vdots \\ \phi_n \end{pmatrix}$$

Then the Jacobian of Φ is nonsingular at x_0. From the implicit function theorem it then follows that there is a neighborhood V of x_0 such that Φ restricted to V is one to one. In particular there can not be two different points in V such that $\Phi(x_1) = \Phi(x_2)$. Proposition 2.2 then shows that there are no indistinguishable points in V. □

2.4 Exercises

2.1 How is (2.3) and Theorem 2.1 changed if f and h are time-varying? Specialize to a linear time-varying system and show that this gives a proof of Theorem 9.10 in Rugh.

2.2 At what points are the systems below locally weakly observable? Are they observable,? Are they locally observable?
a.

$$\begin{aligned}\dot{x} &= u\\ y &= x^2\end{aligned}$$

b.

$$\begin{aligned}\dot{x} &= u\\ y &= \sin x\end{aligned}$$

c.

$$\begin{aligned}\dot{x}_1 &= x_2\\ \dot{x}_2 &= 0\\ y &= x_1^3\end{aligned}$$

d.

$$\dot{x} = \begin{bmatrix}-2 & 1\\ 1 & -2\end{bmatrix} x + u \begin{bmatrix}0 & 0\\ 1 & 0\end{bmatrix} x$$
$$y = \begin{bmatrix}1 & 1\end{bmatrix} x$$

2.3 Consider the aircraft speed dynamics (Example 1.1) problem.

$$\begin{aligned}\dot{x}_1 &= x_2 - f(x_1)\\ \dot{x}_2 &= -x_2 + u\\ y &= x_1\end{aligned}$$

Is the system locally weakly observable? What happens if instead the thrust x_2 is measured?

2.4 A ship that moves with a constant speed in a straight line is observed with a radar that measures distance only. Let x_1 and x_2 be the position of the ship in rectangular coordinates, v its speed through water, and θ its heading angle (a known constant). If y is the radar measurement, the dynamical equations are

$$\begin{aligned}\dot{x}_1 &= v\cos\theta\\ \dot{x}_2 &= v\sin\theta\\ \dot{v} &= 0\\ y &= \sqrt{x^2 + y^2}\end{aligned}$$

a. Is the system locally weakly observable?
b. What happens if the heading angle is constant but unknown?

Chapter 3

Controllability.

For linear systems it is well known that the concept of controllability plays a key role in the understanding of many phenomena. One of the great advances in nonlinear systems in recent years is the development of a theory for controllability and reachability. We will begin by giving a simple example where controllability is important.

Example 3.1 Consider a simple model of the motion of a four-wheeled vehicle as shown in figure 3.1 Let u_1 be the angular velocity with which the forward wheels are turned and let u_2 be the speed of the vehicle. If the motion is slow, inertial effects can be neglected, and u_1 and u_2 can be regarded as inputs. The model is then

$$\frac{d}{dt}\begin{pmatrix}\theta\\ \phi\\ \xi\\ \eta\end{pmatrix} = u_1\begin{pmatrix}1\\ 0\\ 0\\ 0\end{pmatrix} + u_2\begin{pmatrix}0\\ \sin\theta\\ \cos(\theta+\phi)\\ \sin(\theta+\phi)\end{pmatrix} \tag{3.1}$$

We note that the model has the form

$$\dot{x} = u_1 g_1(x) + u_2 g_2(x) \tag{3.2}$$

We know from experience that a vehicle like this is completely controllable: It is possible to get it into any position with any orientation, by a suitable choice of u_1 and u_2. Suppose however that we were given the task of writing a computer program that could compute the u_1 and u_2 that would take the vehicle from an arbitrary position and orientation to another arbitrary position and orientation. This is an example of a *motion planning* problem. It is not completely trivial even in a simple case like this. □

3.1 The basic ideas of controllability

Suppose we have a system described by

$$\dot{x} = f(x,u) \tag{3.3}$$

and that we are considering a number of different constant control signals: $u = u_1$, $u = u_2$ etc. Let us use the notation

$$f_j(x) = f(x,u_j)$$

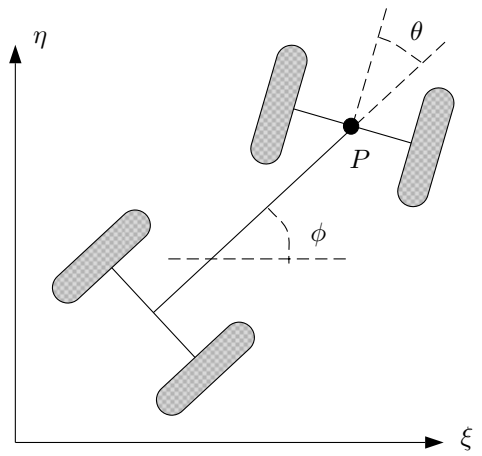


Figure 3.1: Vehicle geometry

Suppose we start at the point x_0 at $t = 0$. In what directions is it possible to move? Suppose the control u_j is chosen. Then a Taylor expansion gives

$$x(t) = x_0 + t\dot{x}(0) + O(t^2) = x_0 + tf_j(x_0) + O(t^2)$$

As expected we see that we can move in all the directions $f_j(x_0)$. Now consider what happens when the control u_1 is applied during a time interval of length h_1, followed by u_2 during h_2 time units. We get

$$x(h_1) = x_0 + h_1 f_1(x_0) + O(h_1^2)$$

$$x(h_2) = x(h_1) + h_2 f_2(x(h_1)) + O(h_2^2)$$

Since

$$f_2(x(h_1)) = f_2(x_0) + O(|x(h_1) - x_0|) = f_2(x_0) + O(h_1)$$

we get

$$x(h_2) = x_0 + h_1 f_1(x_0) + h_2 f_2(x_0) + O(h^2)$$

where $h = \max(h_1, h_2)$. Generalizing this derivation we can show that it is possible to move from x_0 in all directions of the form

$$h_1 f_1(x_0) + h_2 f_2(x_0) + \cdots + h_m f_m(x_0) \tag{3.4}$$

where h_1,..,h_m are *positive* numbers (since it is in general not possible to go backwards in time). Does (3.4) give all possible directions? To investigate that we have to consider higher order Taylor expansions. Consider

$$\dot{x} = f_j(x), \quad x(0) = z$$

The second order Taylor expansion gives

$$x(t) = z + tf_j(z) + \frac{t^2}{2} f_{j,x}(z) f_j(z) + O(t^3)$$

Now suppose that the initial point z has the Taylor expansion

$$z = x_0 + td_1 + t^2 d_2 + O(t^3)$$

Then, since

$$f_j(z) = f_j(x_0) + tf_{j,x}(x_0)d_1 + O(t^2)$$
$$f_{j,x}(z) = f_{j,x}(x_0) + O(t)$$

we get

$$x(t) = x_0 + t(d_1 + f_j(x_0)) + t^2(d_2 + f_{j,x}(x_0)d_1 + \frac{1}{2}f_{j,x}(x_0)f_j(x_0)) + O(t^3) \quad (3.5)$$

We can now use this formula to check the following scenario: Suppose that the controls u_1, u_2, u_3 and u_4 are used after each other, each for h units of time. Suppose also that it is possible to choose u_3 and u_4 in such a way that

$$f_3(x) = -f_1(x), \quad f_4(x) = -f_2(x)$$

We then get successively, using (3.5) (all quantities are evaluated at x_0)

$$x(h) = x_0 + hf_1 + h^2\frac{1}{2}f_{1,x}f_1 + O(h^3)$$

$$x(2h) = x_0 + h(f_1 + f_2) + h^2(\frac{1}{2}f_{1,x}f_1 + f_{2,x}f_1 + \frac{1}{2}f_{2,x}f_2) + O(h^3)$$

$$x(3h) = x_0 + h(f_1 + f_2 + f_3) + h^2(\frac{1}{2}f_{1,x}f_1 + f_{2,x}f_1 + \frac{1}{2}f_{2,x}f_2 + f_{3,x}(f_1 + f_2) + \frac{1}{2}f_{3,x}f_3) + O(h^3)$$

Using $f_3(x) = -f_1(x)$ gives

$$x(3h) = x_0 + hf_2 + h^2(f_{2,x}f_1 + \frac{1}{2}f_{2,x}f_2 - f_{1,x}f_2) + O(h^3)$$

Finally

$$x(4h) = x_0 + h(f_2 + f_4) + h^2(f_{2,x}f_1 + \frac{1}{2}f_{2,x}f_2 - f_{1,x}f_2 + f_{4,x}f_2 + \frac{1}{2}f_{4,x}f_4) + O(h^3)$$

Using $f_4(x) = -f_2(x)$ then gives

$$x(4h) = x_0 + h^2(f_{2,x}f_1 - f_{1,x}f_2) + O(h^3)$$

The expression which comes up above motivates the following definition

Definition 3.1 The *Lie bracket* of the vector fields $f(x)$ and $g(x)$ is

$$[f, g](x) = g_x(x)f(x) - f_x(x)g(x)$$

Note that the Lie bracket is itself a new vector field. We can now formulate the following result of our investigation.

Proposition 3.1 *A movement along $f_1(x)$, then $f_2(x)$, then $-f_1(x)$ and finally along $-f_2(x)$, each for h units of time, results in the position*

$$x(4h) = x_0 + h^2[f_1, f_2](x_0) + O(h^3)$$

The proposition shows that it is indeed possible to move along directions different from the f_j themselves.

Example 3.2 Let us continue Example 3.1. Introducing

$$f_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

and

$$f_2 = \begin{pmatrix} 0 \\ \sin\theta \\ \cos(\theta+\phi) \\ \sin(\theta+\phi) \end{pmatrix}$$

corresponding to the control signals $u = (1\ 0)^T$ and $u = (0\ 1)^T$ respectively, we get, with $x_0 = 0$

$$f_1(x_0) = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad f_2(x_0) = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}$$

Now consider the Lie Bracket

$$[f_1, f_2] = f_{2,x} f_1 - f_{1,x} f_2 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ -\sin\theta & 0 & 0 & 0 \\ -\sin(\theta+\phi) & -\sin(\theta+\phi) & 0 & 0 \\ \cos(\theta+\phi) & \cos(\theta+\phi) & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} =$$

$$= \begin{pmatrix} 0 \\ \cos\theta \\ -\sin(\theta+\phi) \\ \cos(\theta+\phi) \end{pmatrix}$$

In particular

$$[f_1, f_2](x_0) = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \end{pmatrix} \tag{3.6}$$

We see that we have found a movement which is linearly independent of the previous ones. From Proposition 3.1 we see that the Lie bracket $[f_1, f_2]$ corresponds to a small turning of the front wheels, followed by a small forward movement, followed by a turning back of the front wheels, followed by a small backwards movement. Our intuition tells us that this should result in a small turn (increase of ϕ) and a small increase in the y coordinate. This it precisely the result in (3.6). We could now consider more complicated movements. What happens if a small movement forwards is followed by the maneuver just described, followed by a small movement backwards, followed by the reverse of the maneuver? Applying Proposition 3.1 twice, we see that we should consider

$$[f_2, [f_1, f_2]] = \begin{pmatrix} 0 \\ 0 \\ \sin\phi \\ -\cos\phi \end{pmatrix}$$

In particular we have

$$[f_2,[f_1,f_2]](x_0)=\begin{pmatrix}0\\0\\0\\-1\end{pmatrix}$$

We have thus discovered a set of movements that leaves everything unchanged, except that the vehicle moves to the right. □

We have seen from the example that the Lie bracket is useful for studying the motion of nonlinear systems. Let us note some computational rules for Lie brackets. They are all proved by straightforward applications of the definitions.

$$[a,a] = 0 \tag{3.7}$$
$$[a,b] = -[b,a] \tag{3.8}$$
$$[a+b,c] = [a,c]+[b,c] \tag{3.9}$$
$$[a,[b,c]]+[b,[c,a]]+[c,[a,b]] = 0, \quad \text{(Jacobi identy)} \tag{3.10}$$

(A vector field that also has an operation satisfying (3.7) – (3.10) is called a *Lie algebra*.)

The Lie bracket has an interesting property under coordinate transformations. Suppose we have a differential equation

$$\dot{x}=f(x)$$

and that we change variables to

$$z=T(x)$$

where T is an infinitely differentiable transformation that also has an infinitely differentiable inverse (such a transformation is called a diffeomorphism):

$$x=S(z), \quad x=S(T(x))$$

In the z variables the differential equation is satisfied by

$$\dot{z}=T_x(x)\dot{x}=T_x(x)f(x)=T_x(S(z))f(S(z))=\tilde{f}(z)$$

If we have a second differential equation

$$\dot{x}=g(x)$$

it is in the same way transformed into

$$\dot{z}=\tilde{g}(z), \quad \tilde{g}(z)=T_x(S(z))g(S(z))$$

Now consider the Lie bracket $[\tilde{f},\tilde{g}]$. From the definition we have

$$[\tilde{f},\tilde{g}]=\tilde{g}_z\tilde{f}-\tilde{f}_z\tilde{g}=(T_xg)_zT_xf-(T_xf)_zT_xg=$$

$$=T_xg_xS_xT_xf-T_xg_xS_xT_xg+\begin{pmatrix}\vdots\\g^TT_{ixx}f\\\vdots\end{pmatrix}-\begin{pmatrix}\vdots\\f^TT_{ixx}g\\\vdots\end{pmatrix}=T_x[f,g]$$

We see that the Lie bracket of two vector fields undergoes the same linear transformation as the vector fields themselves, when the coordinate system is changed. We have thus proved the following proposition.

Proposition 3.2 *Consider a set of vector fields f_j, forming the right hand side of differential equations. If the coordinate system is changed using a diffeomorphism, then the f_j and their Lie brackets*

$$f_i, \ldots, [f_i, f_j], \ldots, [f_i, [f_j, f_k]], \ldots, [\ldots, [f_i, f_j], \ldots]$$

are all transformed by the same nonsingular linear transformation. Tests of linear dependence or independence thus give the same result in any coordinate system.

3.2 Controllability of general systems.

Using the ideas of the previous section we will now discuss controllability of systems of the form

$$\dot{x} = f(x, u) \tag{3.11}$$

where x is an n-vector and f is assumed to be infinitely differentiable. As in the previous section we look at vectors

$$f_j(x) = f(x, u_j)$$

correspond to a number of constant control signals u_j.

We make the following definitions.

Definition 3.2 *$A_U(x_0)$ is the* reachable set *from x_0, while remaining in the set U, i.e. all points x_f for which there exists a time interval $0 \leq t \leq t_f$ and a control u such that $x(0) = x_0$, $x(t_f) = x_f$ and $x(t), 0 \leq t \leq t_f$ lies in U.*

Definition 3.3 *The system (3.11) is said to be* controllable *if $A_{R^n}(x)$ is R^n for any x.*

Definition 3.4 *The system (3.11) is said to be* locally accessible *at x if $A_U(x)$ has a nonempty interior for any neighborhood U of x.*

If a system is locally accessible it means that a sphere or cube of dimension n is contained in $A_U(x)$ so that the reachable set has "full dimension".

Definition 3.5 *The system (3.11) is said to be symmetric, if for every u there is a $\bar{u}$ such that $f(x, \bar{u}) = -f(x, u)$.*

Example 3.3 A system of the form

$$\dot{x} = u_1 g_1(x) + \cdots + u_m g_m(x)$$

is symmetric (just change signs of the u_j). Descriptions of this type are typical in motion planning problems, see Examples 3.1 and 3.2. □

Remark 3.1 *Systems of the form*

$$\dot{x} = f(x) + u_1 g_1(x) + \cdots + u_m g_m(x)$$

typical for control applications, are in general not *symmetric, due to the presence of the* drift term *$f(x)$.*

The significance of symmetric systems from the controllability point of view is related to the observation that in equation (3.4) only positive values of the h_i are allowed. Changing the sign of h_i is equivalent to the replacement of f_i with $-f_i$. For a symmetric system this is always possible, so that the sign restriction on the h_i effectively dissappears. For a symmetric system it is thus possible to move in all directions that are linear combinations of the vectors f_i. For a symmetric system Proposition 3.1 can always be used, since $-f_1$ and $-f_2$ are available if f_1 and f_2 are. It follows that it is possible to control the system *as if* the right hand side of (3.11) contained not only $f_i = f(x, u_i)$ but also all vectors of the form $[f_i, f_j]$. But then Proposition 3.1 can be applied again to show that vectors of the form $[f_i, [f_j, f_k]]$, $[[f_i, f_j], [f_k, f_l]]$ have to be considered. One is led to the following definition.

Definition 3.6 The *Lie algebra* generated by $\{f_i = f(x, u_i)\}$ consists of all vectors that can be generated by taking linear combinations and successive Lie brackets of the vectors f_i. It is denoted $\{f_i\}_{LA}$.

Intuitively it is possible to move in any direction if this Lie algebra has enough elements. This leads to the following definition

Definition 3.7 The system (3.11) is said to satisfy the *controllability rank condition* at x_0 if there are n linearly independent elements in $\{f_i\}_{LA}(x_0)$, where $f_i(x_0) = f(x_0, u_i)$ for all possible choices of u_i.

Remark 3.2 *Using Jacobi's identity one can show that each element of* $\{f_i\}_{LA}(x_0)$ *can be written as a linear combination of iterated Lie brackets of the form*

$$[f_j, [f_{j-1}, [\ldots, [f_2, f_1]\ldots]]]$$

Example 3.4 In Example 3.2 we showed that

$$f_1(x_0) = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad f_2(x_0) = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \quad [f_1, f_2](x_0) = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}, \quad [f_2, [f_1, f_2]](x_0) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ -1 \end{pmatrix}$$

Clearly the controllability rank condition is satisfied at the origin. □

Using the controllability rank condition it is easy to formulate the main controllability results.

Theorem 3.1 *Let the controllability rank condition be satisfied at* x_0 *for the system (3.11). Then the system is locally accessible at* x_0.

Proof. (sketch) Take an f_i which is nonzero at x_0. (If all the f_i are zero at x_0, it is easy to see that the controllability rank condition can not be satisfied.) Let $\pi(t, x_0)$ be the solution of $\dot{x} = f_i$ starting at x_0. If $n = 1$ we are finished, since the set $\pi(t, x_0)$ for $t > 0$ clearly has a nonempty interior. If $n > 1$ consider solutions of $\dot{x} = f_k$, $k \neq i$, with $x(0) = \pi(t_1, x_0)$ for t_1-values close to 0. There must be some f_k such that f_k is not tangent to the curve $\pi(t_1, x_0)$ for some small t_1-value. (Assume this is not the case. Introduce a coordinate system where the curve $\pi(t, x_0)$ is the first coordinate axis. In that coordinate system all the f_i would then have the form $(*0\ldots0)^T$ along the first coordinate axis. All their Lie

brackets would then also have that form, contradicting the controllability rank condition.) The set $\gamma(t, \pi(t_1, x_0))$ where γ denotes the solution of $\dot{x} = f_k$ is then a two-dimensional set, parameterized by t and t_1, having nonempty interior. If $n > 2$ we can continue the construction to higher dimensions. □

If the system is symmetric we get a stronger result.

Theorem 3.2 *If a symmetric system of the form (3.11), satisfies the controllability rank condition at x_0, then the reachable set $A_U(x_0)$ contains a full neighborhood of x_0 for every neighborhood U of x_0.*

Proof. (sketch) From Theorem 3.1 we know that there exists an $\epsilon > 0$ and a point x_1 such that the full n-dimensional sphere with radius ϵ, centered at x_1 can be reached from x_0. Since the system is symmetric, there is a control signal $\bar{u}$ on some time interval $[0, t_1]$, that reverses the motion and carries the state from x_1 back to x_0. Now keep $\bar{u}$ fixed and consider the differential equation $\dot{x} = f(x, \bar{u})$ for starting points in the ϵ-sphere around x_1. It follows from the fundamental theorems on differential that the points in the sphere will be carried onto a neighborhood of x_0. □

Example 3.5 Consider the system

$$\begin{array}{rcl} \dot{x}_1 & = & u \\ \dot{x}_2 & = & x_1^2 \end{array} \tag{3.12}$$

If we define

$$f_1 = f(x, 0) = \begin{pmatrix} 0 \\ x_1^2 \end{pmatrix}, \quad f_2 = f(x, 1) = \begin{pmatrix} 1 \\ x_1^2 \end{pmatrix}$$

then it is clear that f_1 and f_2 are linearly independent at all points where $x_1 \neq 0$. Computing some Lie brackets one gets

$$[f_1, f_2] = \begin{pmatrix} 0 \\ -2x_1 \end{pmatrix}, \quad [[f_1, f_2], f_2] = \begin{pmatrix} 0 \\ 2 \end{pmatrix}$$

Since $[[f_1, f_2], f_2]$ and f_2 are linearly independent everywhere, the system satisfies the controllability rank condition at all points. According to Theorem 3.1 the system then has the accessibility property at all points. In this case the system is clearly not controllable, since the x_2-variable can not be decreased. □

3.3 Control affine systems

Let us specialize to control systems of the form

$$\dot{x} = f(x) + g(x)u \tag{3.13}$$

where x is an n-vector and u an m-vector. Let us write the system in the form

$$\dot{x} = f(x) + \sum_{i=1}^{m} u_i g_i(x) \tag{3.14}$$

to emphasize that the control variables can be seen as coefficients of the vector fields g_i. From Theorem 3.1 we immediately get

Theorem 3.3 *Consider the control affine system (3.14). Let the Lie algebra generated by f,g_1,...,g_m at x_0 have full rank. Then the system is locally accessible at x_0.*

Proof. Follows from Theorem 3.1. □

In our discussion of accessibility we have not included a discussion of the time needed to reach a point. Sometimes the effect of the drift vector field f will be to make certain points reachable only at certain times.

Example 3.6 Consider the system

$$\dot{x} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u$$

The set of points that can be reached at time t from the origin is the vertical line $x_1 = t$. The system is locally accessible, since our definition allows us to take the union of these sets for all $t > 0$ when calculating $A_U(0)$. However, it is clear that our ability to control the system is severely restricted by the fact that a given point can not necessarily be reached at a given time. □

Motivated by this example it is natural to define the following.

Definition 3.8 *Let $A_U(x_0, T)$ be the points that can be reached from x_0 in precisely T units of time, with trajectories staying in U, i.e. all points x_f for which there exists a control u such that $x(0) = x_0$, $x(T) = x_f$ and $x(t), 0 \leq t \leq T$ lies in U.*

Definition 3.9 *The system (3.14) is* locally strongly accessible *from x_0 if for any neighborhood U of x_0 the set $A_U(x_0, T)$ contains an non-empty opens subset for any sufficiently small $T > 0$.*

To get strong accessibility, the criterion has to be modified somewhat.

Theorem 3.4 *For the system (3.14 consider the following set of Lie brackets*

$$[h_j, [h_{j-1}, [\ldots [h_1, g_i] \ldots]]], \quad i = 1, \ldots, m \tag{3.15}$$

where the h_j are taken from the set $f, g_1, \ldots, g_m$. If the span of (3.15) at x_0 has full rank, then the system is locally strongly accessible from x_0.

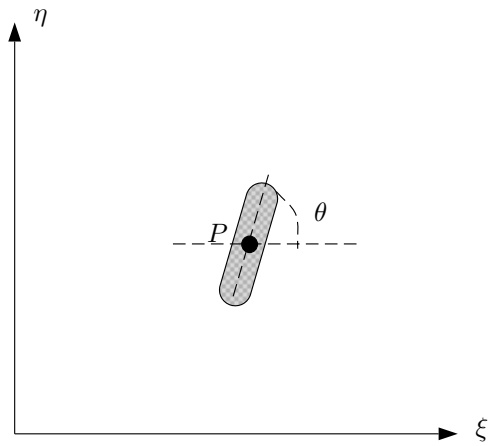
Proof. (Sketch) Introduce the extra state variable x_{n+1} satisfying $\dot{x}_{n+1} = 1$, $x(0) = 0$ so that $x_{n+1} = t$. The result can now be obtained by applying Theorem 3.3 to the system

$$\begin{bmatrix} \dot{x} \\ \dot{x}_{n+1} \end{bmatrix} = \begin{bmatrix} f(x) \\ 1 \end{bmatrix} + \sum_{i=1}^{m} u_i \begin{bmatrix} g_i(x) \\ 0 \end{bmatrix}$$

□

3.4 Exercises.

3.1 Consider the following one-wheeled vehicle ("uni-cycle").



Suppose that the vehicle can balance on its single wheel and that it is possible to turn the wheel with the speed u_1 and to move forward with speed u_2. The dynamics is then

$$\begin{aligned}\dot{\theta} &= u_1\\ \dot{\xi} &= u_2\cos\theta\\ \dot{\eta} &= u_2\sin\theta\end{aligned}$$

What can be said about controllability/accessibility if u_1, u_2 are control signals?

3.2 What can be said about the controllability/accessibility of the aircraft model of Example 1.1?

3.3 Check the controllability/accessibility of the system

$$\dot{x}_1 = u_1x_3 + u_2 \tag{3.16}$$
$$\dot{x}_2 = u_1x_1 \tag{3.17}$$
$$\dot{x}_3 = u_1x_2 \tag{3.18}$$

3.4 Consider a rigid body with angular velocities x_i. Assume that there is just one control signal, which is the torque along an axis with coordinates (b_1, b_2, b_3). Then the system description is

$$\dot{x}_1 = a_1x_2x_3 + b_1u \tag{3.19}$$
$$\dot{x}_2 = a_2x_1x_3 + b_2u \tag{3.20}$$
$$\dot{x}_3 = a_3x_1x_2 + b_3u \tag{3.21}$$

where the a_i are given by the moments of inertia. What can be said about controllability and reachability, starting from $x = 0$?

3.5 Consider the system

$$\dot{x} = f(x) + g(x)u, \quad f(0) = 0$$

Let the linearization of the system at $x = 0$ be controllable. Show that this implies local strong accessibility. (Actually one can show that a full neighborhood of the origin can be reached in this case.)

3.6 Prove the statement of Remark 3.2.

Chapter 4

Input-output descriptions – Volterra series

For linear systems there are explicit characterizations of the output as a function of the input like

$$y(t) = \int_0^t h(t,\tau)u(\tau)d\tau$$

where h is the impulse response, or

$$Y(s) = G(s)U(s)$$

where Y, U are Laplace transformed quantities and G is the transfer function. Is it possible to do something similar for nonlinear systems?

To get a feeling for the problem, let us for a moment discuss a single-input-single-output discrete time system. If the system is initialized at $t = 0$, then we can write

$$y(t) = F(t, u(0), u(1), \ldots, u(t)), \quad t = 0, 1, 2, \ldots$$

for some function F. If F is sufficiently smooth we can make a Taylor expansion

$$y(t) = y_0(t) + \sum_{j=0}^{t} g_1(t,j)u(j) + \sum_{j=0}^{t}\sum_{k=0}^{t} g_2(t,j,k)u(j)u(k) + \cdots$$

where

$$y_0(t) = F(t, 0, \ldots, 0), \quad g_1(t,j) = \partial F/\partial u(j), \quad g_2(t,j,k) = \frac{1}{2}\frac{\partial^2 F}{\partial u(j)\partial u(k)}$$

Sums of discrete time variables usually correspond to integrals of continuous time variables. It is then a reasonable guess that we should be able to obtain a description of the form

$$\begin{aligned} y(t) = y_0(t) + \int_{\infty}^{\infty} h_1(t,\sigma)u(\sigma)d\sigma + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h_2(t,\sigma_1,\sigma_2)u(\sigma_1)u(\sigma_2)d\sigma_1 d\sigma_2 + \cdots \\ \cdots + \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} h_n(t,\sigma_1,\ldots,\sigma_n)u(\sigma_1)\cdots u(\sigma_n)d\sigma_1 \ldots d\sigma_n + \cdots \end{aligned} \quad (4.1)$$

for some class of nonlinear systems. We will assume that the variables are defined so that $u(t) = 0$ corresponds to $y(t) = 0$ which means that $y_0(t) = 0$. Also we usually consider the time-invariant case where the functions h_i depend only on on the difference between the time variables:

$$y(t) = \int_{\infty}^{\infty} h_1(t-\sigma)u(\sigma)d\sigma + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h_2(t-\sigma_1, t-\sigma_2)u(\sigma_1)u(\sigma_2)d\sigma_1 d\sigma_2 + \cdots$$
$$\cdots + \int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty} h_n(t-\sigma_1,\ldots,t-\sigma_n)u(\sigma_1)\cdots u(\sigma_n)d\sigma_1\ldots d\sigma_n + \cdots \quad (4.2)$$

With a simple variable change this can also be written

$$y(t) = \int_{\infty}^{\infty} h_1(\sigma)u(t-\sigma)d\sigma + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h_2(\sigma_1, \sigma_2)u(t-\sigma_1)u(t-\sigma_2)d\sigma_1 d\sigma_2 + \cdots$$
$$\cdots + \int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty} h_n(\sigma_1,\ldots,\sigma_n)u(t-\sigma_1)\cdots u(t-\sigma_n)d\sigma_1\ldots d\sigma_n + \cdots \quad (4.3)$$

A description like (4.1), (4.2) or (4.3) is called a *Volterra series* for the system. The functions h_n are called *kernels*. If only the n:th kernel is nonzero the system is said to be *homogeneous* of degree n. In this chapter we will show how a Volterra series can be computed for a fairly general nonlinear system. We will also look at frequency response representations of the kernels by looking at the multivariable Laplace transform:

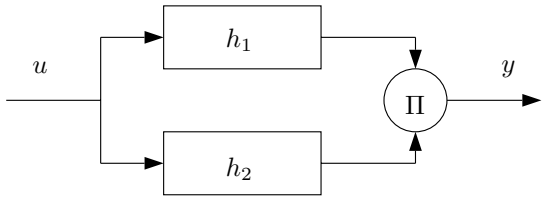
$$H(s_1,\ldots,s_n) = \int_0^{\infty}\ldots\int_0^{\infty} h(\sigma_1,\ldots,\sigma_n)e^{-s_1\sigma_1}\cdots e^{-s_n\sigma_n}d\sigma_1\ldots d\sigma_n \quad (4.4)$$

These function are sometimes referred to as higher order transfer functions.

4.1 Some simple Volterra series

For some simple systems the Volterra series can be calculated directly.

Example 4.1 Consider a multiplicative parallel connection of two linear systems with impulse responses h_1 and h_2 as shown below



The output signal y is given by

$$y(t) = \int_{-\infty}^{\infty} h_1(\sigma_1)u(t-\sigma_1)d\sigma_1 \int_{-\infty}^{\infty} h_2(\sigma_2)u(t-\sigma_2)d\sigma_2 =$$
$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(\sigma_1,\sigma_2)u(t-\sigma_1)u(t-\sigma_2)d\sigma_1 d\sigma_2 \quad (4.5)$$

where $h(\sigma_1, \sigma_2) = h_1(\sigma_1)h_2(\sigma_2)$. which is of the form (4.3) with only one term in the series. The system is thus homogeneous of degree 2. □

Example 4.2 Consider the special case of the previous example where the linear systems are given by

$$h_1(\sigma) = \Delta(\sigma)e^{-\sigma}, \quad h_2(\sigma) = \Delta(\sigma)e^{-2\sigma}, \quad \text{where}$$

$$\Delta(\sigma) = \begin{cases} 1, & \text{if } \sigma \geq 0 \\ 0, & \text{if } \sigma < 0 \end{cases}$$

then

$$h(\sigma_1, \sigma_2) = e^{-(\sigma_1 + 2\sigma_2)}\Delta(\sigma_1)\Delta(\sigma_2)$$

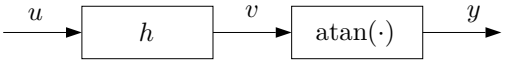
(The Δ:s give impulse responses that are causal) Taking the Laplace transform gives the transfer function

$$H(s_1, s_2) = \frac{1}{(s_1 + 1)(s_2 + 2)}$$

□

Now consider an example which has an infinite series.

Example 4.3 Below is a system of Wiener type, i.e. a linear system followed by a static nonlinearity.



Expanding the arctangent into its Taylor series gives

$$y(t) = v(t) - \frac{1}{3}v(t)^3 + \frac{1}{5}v(t)^5 - \cdots = \sum_{k \text{ odd}} \frac{(-1)^{\frac{k-1}{2}}}{k} \; v(t)^k$$

Substituting

$$v(t) = \int_{-\infty}^{\infty} h(\sigma)u(t - \sigma)d\sigma$$

gives

$$y(t) = \sum_{k \text{ odd}} \frac{(-1)^{\frac{k-1}{2}}}{k} \left(\int_{-\infty}^{\infty} h(\sigma)u(t - \sigma)d\sigma \right)^k$$

Rewriting the integrals as multiple integrals in a manner analogous to (4.5) gives

$$y(t) = \sum_{k \text{ odd}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{(-1)^{\frac{k-1}{2}}}{k} h(\sigma_1) \cdots h(\sigma_k)u(t-\sigma_1) \ldots u(t-\sigma_k)d\sigma_1 \ldots d\sigma_k$$

□

As a special case we look at

Example 4.4 Consider the system of the previous example with

$$h(t) = \Delta(t)e^{-t}$$

Then, for odd n, the n:th kernel is given by

$$h_n(\sigma_1, \ldots, \sigma_n) = (-1)^{\frac{n-1}{2}} \frac{1}{n} \Delta(\sigma_1) \cdots \Delta(\sigma_n) \cdot e^{-(\sigma_1 + \cdots + \sigma_n)}$$

while the corresponding transfer function is

$$H_n(s_1, \ldots, s_n) = (-1)^{\frac{n-1}{2}} \frac{1}{n(s_1 + 1) \cdots (s_n + 1)}$$

□

4.2 The Volterra series of a bilinear system

Volterra series are also easily calculated for bilinear systems. A single-input-single-output bilinear system is given by

$$\begin{aligned} \dot{x} &= Ax + (Dx + b)u \\ y &= cx \\ x(0) &= 0 \end{aligned} \tag{4.6}$$

where x is an n-vector, y and u are scalars, and A, D, b and c are matrices of appropriate dimensions. Note that we assume an initial value of x that is zero. The kernels of the Volterra series for this system have an appealing and simple structure.

Theorem 4.1 *The bilinear system (4.6) has a Volterra series (4.3), where the n:th order kernel is given by*

$$h_n(\sigma_1, \ldots, \sigma_n) = \begin{cases} ce^{A\sigma_1} De^{A(\sigma_2-\sigma_1)} D \cdots De^{A(\sigma_n-\sigma_{n-1})} b & \text{if } 0 \le \sigma_1 \le \cdots \le \sigma_n \\ 0 & \text{otherwise} \end{cases} \tag{4.7}$$

Proof. We use the change of variables $x = e^{At}z$ in (4.6), which gives

$$Ax + Dx\, u + bu = \dot{x} = e^{At}\dot{z} + Ae^{At}z$$

Solving for $\dot{z}$ gives

$$\dot{z} = u \underbrace{e^{-At}De^{At}}_{\bar{D}(t)} z + \underbrace{e^{-At}b}_{\bar{b}(t)} u$$

Using the iteration (1.5) gives

$$z_{j+1}(t) = \int_0^t (\bar{b}(\tau) + \bar{D}(\tau) z_j(\tau)) u(\tau) d\tau$$

Iterating gives

$$\begin{aligned} z_1(t) &= \int_0^t \bar{b}(\sigma_1) u(\sigma_1) d\sigma_1 \\ z_2(t) &= \int_0^t \bar{b}(\sigma_1) u(\sigma_1) d\sigma_1 + \int_0^t \bar{D}(\sigma_1) u(\sigma_1) \int_0^{\sigma_1} \bar{b}(\sigma_2) u(\sigma_2) d\sigma_1 d\sigma_2 \\ &\vdots \\ z_n(t) &= \int_0^t \bar{b}(\sigma_1) u(\sigma_1) d\sigma_1 + \cdots \\ &\quad \cdots + \int_0^t \bar{D}(\sigma_1) u(\sigma_1) \ldots \int_0^{\sigma_{n-1}} \bar{b}(\sigma_n) u(\sigma_n) d\sigma_1 \cdots d\sigma_n \\ &\vdots \end{aligned}$$

Using that

$$y(t) = cx(t) = ce^{At}z(t)$$

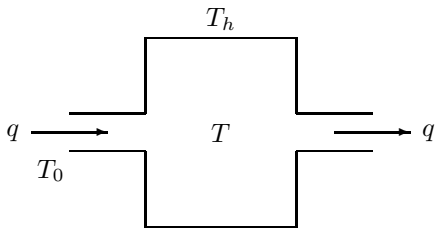
and substituting the expressions for $\bar{b}$ and $\bar{D}$ gives

$$y(t) = \int_0^t ce^{A(t-\sigma_1)}bu(\sigma_1)d\sigma_1 + \int_0^t \int_0^{\sigma_1} ce^{A(t-\sigma_1)}De^{A(\sigma_1-\sigma_2)}bu(\sigma_1)u(\sigma_2)d\sigma_1 d\sigma_2 \cdots$$
$$\cdots + \int_0^t \int_0^{\sigma_1} \cdots \int_0^{\sigma_{n-1}} ce^{A(t-\sigma_1)}De^{A(\sigma_1-\sigma_2)}D \cdots$$
$$\cdots De^{A(\sigma_{n-1}-\sigma_n)}bu(\sigma_1)\cdots u(\sigma_n)d\sigma_1 \ldots d\sigma_n + \cdots$$

Performing the change of variables $t - \sigma_i \to \sigma_i$ completes the proof. □

A kernel of the form (4.7) is called a *triangular kernel* since it is zero outside a triangular region.

Example 4.5 Consider a heat exchanger model.



A fluid which initially has the temperature T_0 flows with the flow rate q through the heat exchanger, which is surrounded by a medium with temperature T_h. It is assumed that very good mixing takes place so that one can assume the same temperature T at every point in the heat exchanger. If the heat capacity of the fluid is c per unit volume and C for the whole heat exchanger, and if the heat transfer coefficient of the walls is κ, then a heat balance gives

$$\frac{d}{dt}(CT) = qcT_0 - qcT + \kappa(T_h - T)$$

Assume that the flow rate is controlled around a nominal flow q_0 so that

$$q = q_0 - u$$

Then, using the numerical values

$$c/C = 1, \ \kappa/C = 1, \ T_h = q_0 = -T_0 = 1$$

gives the model

$$\dot{T} = -2T + uT + u \tag{4.8}$$

where the temperature T is a state variable and the flow change u is the input. (Note that a positive u means a decrease in flow.)

Using (4.23) we get the following kernels

$$h_1(t_1) = e^{-2t_1}, \quad h_2(t_1, t_2) = e^{-2t_2}, \ldots, h_n(t_1, \ldots, t_n) = e^{-2t_n}, \ldots \tag{4.9}$$

□

Example 4.6 For the bilinear system

$$\dot{x} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} x\, u + \begin{pmatrix} 1 \\ 1 \end{pmatrix} u \tag{4.10}$$

$$y = (1 \quad 0)$$

one has

$$A = 0, \quad e^{At} = I$$

$$h_1(t_1) = ce^{At}b = cb = 1 \tag{4.11}$$

$$h_2(t_1, t_2) = ce^{At_1}De^{A(t_2-t_1)}b = cDb = 1 \tag{4.12}$$

$$h_n(t_1, \ldots, t_n) = cD^{n-1}b = 0, \quad n > 2 \tag{4.13}$$

□

The last two examples show that bilinear systems might have finite or infinite Volterra series.

4.3 Volterra series for control affine systems

Having obtained a formula for computing the Volterra series for a bilinear systems, we can use the idea of Carleman bilinearization from section 1.2 to handle a general control affine system.

$$\dot{x} = f(x) + g(x)u, \quad y = h(x), \quad x(0) = 0 \tag{4.14}$$

It is assumed that $f(0) = 0$, $h(0) = 0$. We can first compute an N:th truncation of the Carleman bilinearization (1.27):

$$\begin{aligned} \dot{z}_N &= \underbrace{\begin{bmatrix} F_1 & F_2 & F_3 & \ldots & F_N \\ 0 & A_{21} & A_{23} & \ldots & A_{2,N-1} \\ 0 & 0 & A_{33} & \ldots & A_{3,N-2} \\ 0 & 0 & 0 & \ddots & \\ \vdots & \vdots & & & \\ 0 & 0 & \ldots & 0 & A_{N,1} \end{bmatrix}}_{A_N} z_N + \underbrace{\begin{bmatrix} G_1 & G_2 & G_3 & \ldots & G_N \\ B_{20} & B_{21} & B_{22} & \ldots & B_{2,N-1} \\ 0 & B_{30} & B_{31} & \ldots & B_{3,N-2} \\ 0 & 0 & B_{40} & \ldots & \\ \vdots & \vdots & \vdots & \ddots & \\ \ldots & 0 & B_{N,0} & B_{N,1} & \end{bmatrix}}_{D_N} z_N\, u + \underbrace{\begin{bmatrix} g(0) \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}}_{b_N} u \\ y &= \underbrace{\begin{bmatrix} H_1 & H_2 & H_3 & \ldots & H_N \end{bmatrix}}_{c_N} z_N \end{aligned} \tag{4.15}$$

where z_N is a truncation of z containing components corresponding to monomials of degree N or lower. Using (4.23) the first N kernels can then be computed from

$$h_n(t_1, \ldots, t_n) = \begin{cases} c_N e^{A_N t_1} D_N e^{A_N(t_2-t_1)} D_N \cdots D_N e^{A_N(t_n-t_{n-1})} b_N & \text{if } 0 \le t_1 \le t_2 \le \ldots \le t_n \\ 0 & \text{otherwise} \end{cases}$$

$$n = 1, \ldots, N \tag{4.16}$$

The approximation achieved is shown by the following theorem.

Theorem 4.2 *Let the kernels h_i be calculated from (4.16). Then there is a positive number ϵ such that the output y of (4.14) is given by*

$$y(t) = \sum_{k=1}^{N} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} h_k(\sigma_1, \dots, \sigma_k) u(t-\sigma_1) \dots u(t-\sigma_k) d\sigma_1 \dots d\sigma_k + o(||u||^N) \tag{4.17}$$

for $||u|| \leq \epsilon$.

Proof. (Sketch) One can show that, starting with $x(0) = 0$, and for sufficiently small u, the solution of (4.14) satisfies $||x|| = O(||u||)$. The truncation of the Carleman bilinearization removes terms which are powers of x higher than N. Their size is $o(||x||^N) = o(||u||^N)$. The truncated Carleman bilinearization will in general have an infinite Volterra series. Truncating this series after N terms also give the error $o(||u||^N)$. □

4.4 Uniqueness of Volterra series

In Theorem 4.2 we obtained formulas for the Volterra kernels for a truncated Volterra series. There are actually many ways of calculating kernels and one might wonder if they always give the same result. It is then important to know the following fact

Theorem 4.3 *If a system has a Volterra expansion*

$$y(t) = \sum_{k=1}^{N} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} h_k^{(i)}(s_1, \dots, s_k) u(t-s_1) \dots u(t-s_k) ds_1 \dots ds_k + \\ + o(||u||^N) \tag{4.18}$$

for two sets of kernels $h_1^{(1)}, \dots, h_N^{(1)}$ and $h_2^{(2)}, \dots, h_N^{(2)}$, then for any k, $1 \leq k \leq N$ and any u

$$\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} h_k^{(1)}(s_1, \dots, s_k) u(t-s_1) \dots u(t-s_k) ds_1 \dots ds_k = \\ \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} h_k^{(2)}(s_1, \dots, s_k) u(t-s_1) \dots u(t-s_k) ds_1 \dots ds_k \tag{4.19}$$

Proof.(sketch) Subtracting the series and replacing u with ϵu, for some fixed u, we get

$$0 = \sum_{k=1}^{N} \epsilon^k \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \left(h_k^{(1)}(s_1, \dots, s_k) - h_k^{(2)}(s_1, \dots, s_k) \right) \cdot \\ \cdot u(t-s_1) \dots u(t-s_k) ds_1 \dots ds_k + o(\epsilon^N)$$

Letting ϵ tend to zero we see that we get a contradiction unless, for any u and any k, $1 \leq k \leq N$

$$\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \left(h_k^{(1)}(s_1, \dots, s_k) - h_k^{(2)}(s_1, \dots, s_k) \right) u(t-s_1) \dots u(t-s_k) ds_1 \dots ds_k = 0$$

□

Unfortunately it does not follow from (4.19) that $h_i^{(}1) = h_i^{(}2)$. Consider for example a second order kernel. By making the variable change $\sigma_1 \to \sigma_2$, $\sigma_2 \to \sigma_1$ in the integral one sees that

$$y(t) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(\sigma_1,\sigma_2)u(t-\sigma_1)u(t-\sigma_2)d\sigma_1 d\sigma_2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(\sigma_2,\sigma_1)u(t-\sigma_1)u(t-\sigma_2)d\sigma_1 d\sigma_2$$

This means that the kernels $h(\sigma_1,\sigma_2)$ and $h(\sigma_2,\sigma_1)$ give exactly the same input-output behaviour. This will then also be true for $\frac{1}{2}\big(h(\sigma_1,\sigma_2)+h(\sigma_2,\sigma_1)\big)$. In Example 4.2 we can thus choose between the expressions

$$h(\sigma_1,\sigma_2) = e^{-(\sigma_1+2\sigma_2)}\Delta(\sigma_1)\Delta(\sigma_2) \tag{4.20}$$

$$h(\sigma_1,\sigma_2) = e^{-(2\sigma_1+\sigma_2)}\Delta(\sigma_1)\Delta(\sigma_2) \tag{4.21}$$

$$h(\sigma_1,\sigma_2) = \frac{1}{2}e^{-\sigma_1-\sigma_2}(e^{-\sigma_1}+e^{-\sigma_2})\Delta(\sigma_1)\Delta(\sigma_2) \tag{4.22}$$

The situation is analogous for higher order kernels and we could equally well write the kernel (4.7) for a bilinear system as

$$h_n(t_1,\ldots,t_n) = \begin{cases} ce^{At_n}De^{A(t_{n-1}-t_n)}D\cdots De^{A(t_1-t_2)}b & \text{if } t_1 \geq t_2 \geq \ldots \geq t_n \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{4.23}$$

One way of removing this ambiguity is to use triangular kernels that are always nonzero over the same triangular region. Another way is to use the *symmetric kernel*

$$h_{sym}(\sigma_1,\sigma_2) = \frac{1}{2}\big(h(\sigma_1,\sigma_2)+h(\sigma_2,\sigma_1)\big)$$

which for an n:th order kernel becomes

$$h_{sym}(\sigma_1,\ldots,\sigma_n) = \frac{1}{n!}\sum_{\pi(\ )} h(\sigma_{\pi(1)},\ldots,\sigma_{\pi(n)}) \tag{4.24}$$

where the summation is over all permutations $\pi(\ )$ of the indices. Obviously the symmetric kernel will have the property that

$$h_{sym}(\sigma_1,\ldots,\sigma_n) = h_{sym}(\sigma_{\pi(1)},\ldots,\sigma_{\pi(n)})$$

where π is any permutation. The Laplace transform of a symmetric kernel is called the symmetric transfer function. Because of the linearity of the Laplace transform it can also be obtained from the formula

$$H_{sym}(s_1,\ldots,s_n) = \frac{1}{n!}\sum_{\pi(\ )} H(s_{\pi(1)},\ldots,s_{\pi(n)}) \tag{4.25}$$

where H is the transform of some non-symmetric kernel.

4.5 Response to simple input functions

When considering the response of an n:th order homogeneous system to various input signals, it is convenient to work with an operator, defined to work on n different input signals:

$$H_n[u_1, \ldots, u_n] = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} h_n(\sigma_1, \ldots, \sigma_n) u_1(t-\sigma_1) \ldots u_n(t-\sigma_n) d\sigma_1 \ldots d\sigma_n \tag{4.26}$$

The response of a Volterra series containing only the nth term

$$y(t) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} h_n((t_1, \ldots, t_n)) u(t-t_1) \cdots u(t-t_n) dt_1 \ldots dt_n$$

is then

$$y = H_n[u, \ldots, u]$$

Let us consider the operation of a second order operator $H_2[.,.]$ on an input which is a sum of some more elementary signals v_i.

$$u(t) = \sum_{i=1}^{p} \alpha_i v_i(t)$$

From the definitions we get immediately that

$$y = H_2[u,u] = H_2\left[\sum_{i=1}^{p} \alpha_i v_i, \sum_{j=1}^{p} \alpha_j v_j\right] = \sum_{i=1}^{p}\sum_{j=1}^{p} \alpha_i \alpha_j H_2[v_i, v_j]$$

This formula shows that $H_s[.,.]$ is a *bilinear* operator. It also shows that it is enough to know the action of the operator on pairs of the more elementary signals v_i.

This result is easily generalized to the general case.

Proposition 4.1 *Let $H_n[.,\ldots,.]$ be the operator corresponding to a degree n homogeneous Volterra system as defined in (4.26). If*

$$u(t) = \sum_{i=1}^{p} \alpha_i v_i(t)$$

then

$$y = H_n[u, \ldots, u] = \sum_{i_1=1}^{p} \cdots \sum_{i_n=1}^{p} \alpha_{i_1} \cdots \alpha_{i_n} H_n[v_{i_1}, \ldots, v_{i_n}]$$

The proposition shows that the operator $H_n[.,\ldots,.]$ is *multilinear*. We are now ready to consider some different input signals.

4.5.1 Response to impulses

For linear systems we know that the kernel has an interpretation as the response to an impulse. For a general homogeneous system we have

Proposition 4.2 *If the input of a degree n homogeneous system, with kernel $h_n(t_1, \ldots, t_n)$, is a unit impulse $\delta(t)$, then the output is $y(t) = h_n(t, \ldots, t)$.*

Proof.

$$y(t) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} h_n(\sigma_1, \ldots, \sigma_n)\delta(t-\sigma_1)\ldots\delta(t-\sigma_n)d\sigma_1 \ldots d\sigma_n = h_n(t, \ldots, t)$$

from the properties of the δ-functions. □

Example 4.7 The impulse response of the system in Example 4.2 is e^{-3t} for $t > 0$. □

We see that the the response to a single impulse only involves the "diagonal" part of the kernel. To see the "off-diagonal" parts, several impulses are needed.

Example 4.8 Consider the response of a second order homogeneous system with kernel $h_2(t_1, t_2)$ to the input

$$u(t) = \delta(t) + \delta(t + T)$$

Denoting the delta-functions u_1 and u_2 respectively we have

$$y = H_2[u_1 + u_2, u_1 + u_2] = H_2[u_1, u_1] + H_2[u_1, u_2] + H_2[u_2, u_1] + H_2[u_2, u_2]$$

where $H_2[,]$ is the bilinear operator corresponding to the kernel h_2. Using the properties of delta-functions we get

$$y(t) = h_2(t, t) + h_2(t, t + T) + h_2(t + T, t) + h_2(t + T, t + T)$$

□

From the example we see that it is possible (in principle) to determine the kernel from identification experiments with multiple impulses.

4.5.2 Response to exponentials

Consider an input of the form

$$u(t) = \sum_{k=1}^{p} \alpha_k e^{s_k t}$$

applied to a degree n homogeneous system, with kernel h_n, transfer function H_n and operator $H_n[., .., .]$. Using Proposition 4.1 gives

$$y = \sum_{k_1=1}^{p} \cdots \sum_{k_n=1}^{p} \alpha_{k_1} \cdots \alpha_{k_n} H_n[e^{s_{k_1} t}, \ldots, e^{s_{k_n} t}]$$

We have

$$H_n[e^{s_{k_1} t}, \ldots, e^{s_{k_n} t}] =$$

$$= e^{(s_{k_1} + \cdots + s_{k_n})t} \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} h_n(\sigma_1, \ldots, \sigma_n) e^{-s_{k_1}\sigma_1} \cdots - e^{s_{k_n}\sigma_n} d\sigma_1 \ldots d\sigma_n =$$

$$= H_n(s_{k_1}, \ldots, s_{k_n}) e^{(s_{k_1}+\cdots+s_{k_n})t}$$

so that finally we get

$$y(t) = \sum_{k_1=1}^{p} \cdots \sum_{k_n=1}^{p} \alpha_{k_1} \cdots \alpha_{k_n} H_n(s_{k_1}, \ldots, s_{k_n}) e^{(s_{k_1}+\cdots+s_{k_n})t}$$

To interpret the formula one usually wants to group together terms having the same exponential function. With a bit of combinatorics one can prove

Proposition 4.3 *If the input $\alpha_1 e^{s_1 t} + \cdots + \alpha_p e^{s_p t}$ is applied to a degree-n homogeneous system with the symmetric transfer function H_{sym} then the output is*

$$y(t) = \sum \alpha_1^{m_1} \cdots \alpha_p^{m_p} C_{m_1,\ldots,m_p}(s_1, \ldots, s_p) e^{(m_1 s_1 + \cdots + m_p s_p)t} \tag{4.27}$$

where the sum is taken over all positive m_j whose sum is n, and where the coefficients are given by

$$C_{m_1,\ldots,m_p}(s_1, \ldots, s_p) = \frac{n!}{m_1! \cdots m_p!} H_{sym}(\underbrace{s_1, \ldots, s_1}_{m_1}, \ldots, \underbrace{s_p, \ldots, s_p}_{m_p})$$

Corollary 4.1 *If $n = p$ then the coefficient of*

$$e^{(s_1+\cdots+s_n)t} \quad \textit{is} \quad n! H_{sym}(s_1, \ldots, s_n)$$

Proposition 4.3 and its corollary can be used to compute transfer functions by identification of the coefficients. The idea is shown by the following example.

Example 4.9 Consider the pendulum equation

$$\ddot{y} + 2\zeta\dot{y} + \sin y = u$$

and assume that the transfer functions up to order three are wanted. Using the input $\exp(s_1 t) + \exp(s_2 t)$ the output will be

$$y(t) = H_1(s_1) e^{s_1 t} + H_1(s_2) e^{s_2 t} + 2H_2(s_1, s_2) e^{(s_1+s_2)t} + \cdots$$

where of course H_1 is the usual transfer function. Substituting into the pendulum equation and using the series expansion of sin, the coefficient of $\exp((s_1 + s_2)t)$ of the left hand side is

$$2\big((s_1 + s_2)^2 + 2\zeta(s_1 + s_2) + 1\big) H_2(s_1, s_2)$$

Since there is no corresponding term in the right hand side, it follows that $H_2 = 0$.

Using the input

$$u(t) = e^{s_1 t} + e^{s_2 t} + e^{s_3 t}$$

the output will have the form

$$y(t) = H_1(s_1) e^{s_1 t} + H_1(s_2) e^{s_2 t} + H_1(s_3) e^{s_3 t} + 6H_3(s_1, s_2, s_3) e^{(s_1+s_2+s_3)t} + \cdots$$

Collecting all the coefficients of $exp((s_1 + s_2 + s_3)t)$ gives the equation

$$\big((s_1+s_2+s_3)^2 + 2\zeta(s_1+s_2+s_3)+1\big)6H_3(s_1,s_2,s_3) - H_1(s_1)H_1(s_2)H_1(s_3) = 0$$

or

$$H_3(s_1,s_2,s_3) = \frac{1}{6}H_1(s_1)H_1(s_2)H_1(s_3)H_1(s_1+s_2+s_3)$$

where

$$H_1(s) = \frac{1}{s^2+2\zeta s+1}$$

□

4.5.3 Response to sinusoidal inputs.

Using the results of the previous section it is easy to determine the response of a nonlinear system to sine or cosine signals. Consider the following example

Example 4.10 Let the input to a degree 2 homogeneous system with symmetric transfer function be

$$u(t) = A_0 + 2A_1\cos\omega t = A_1e^{-i\omega t} + A_0 + A_1e^{i\omega t}$$

Then the output is

$$y(t) = A_1^2H(i\omega,i\omega)e^{2i\omega t} + 2A_0A_1H(0,i\omega)e^{i\omega t} + A_0^2H(0,0) + 2A_1^2H(i\omega,-i\omega)+$$
$$+2A_0A_1H(0,-i\omega)e^{-i\omega t} + A_1^2H(-i\omega,-i\omega)e^{-2i\omega t}$$

or in real form

$$y(t) = A_0^2H(0,0) + 2A_1^2H(i\omega,-i\omega) + 4A_0A_1|H(0,i\omega)|\cos(\omega t + \arg H(0,i\omega))+$$
$$+2A_1^2|H(i\omega,i\omega)|\cos(2\omega t + \arg H(i\omega,i\omega))$$

□

This example clearly shows an important nonlinear phenomenon: different frequencies are mixed. The amplitude of the cosine with frequency ω depends not only on the input at that frequency but also on the constant component. Likewise the constant component in the output depends on the the cosine in the input. For a nonlinear system it is thus perfectly possible that high frequency noise gives a change in the dc-level of the output.

4.6 Exercises

4.1 Compute the first three terms of the Volterra series for the scalar system

$$\dot{x} = \left(1+x^2\right)u, \quad x(0) = 0$$

4.2 What are the kernels of

$$\dot{x} = \begin{pmatrix} -2 & 0 \\ 0 & -3 \end{pmatrix} x + \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} x\,u + \begin{pmatrix} 1 \\ 0 \end{pmatrix} u \tag{4.28}$$

4.3 Compute the step response of the system defined in the previous exercise using

1. a direct solution of the differential equation,
2. the Volterra series.

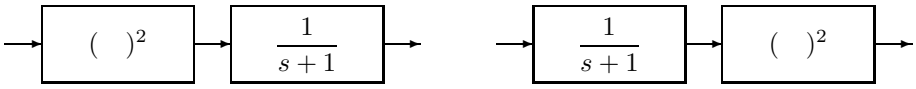
4.4 Compute the second order Carleman bilinearization and the Volterra series of the system

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= -x_2 + x_2^2 + u \\ y &= x_1 \end{aligned} \tag{4.29}$$

4.5 Compute the second order Carleman bilinearization and the Volterra series of the system

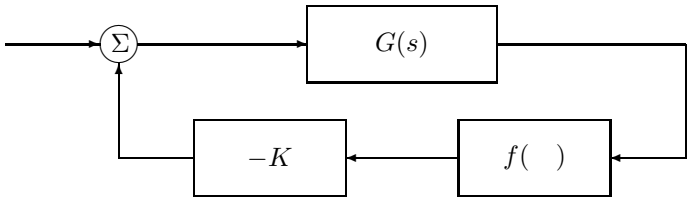
$$\begin{aligned} \dot{x}_1 &= x_2^2 \\ \dot{x}_2 &= u \\ y &= x_1 \end{aligned} \tag{4.30}$$

4.6 What are the transfer functions of the two systems given below?



4.7 What are the first nonzero higher order transfer functions from the reference r to the output y of the the feedback system shown below when

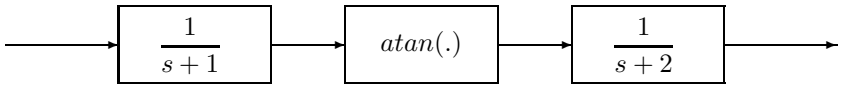
1. $f(y) = y^2$
2. $f(y) = y^3$



4.8 Let the signal $2A_1 \cos \omega_1 t + 2A_2 \cos \omega_2 t$ be the input to a degree-2 homogeneous system . What is the output?

4.9 Suppose that the input to the systems of exercise 4.6 is $2A_1 \cos \omega t + 2A_2 \cos 2\omega t$ where the low frequency cosine is the signal and the high frequency one is a disturbance. Discuss using the previous exercise which system is the better low pass filter.

4.10 Compute the kernels and transfer functions for the Volterra series of the system below



Chapter 5

Realization of input-output descriptions.

The purpose of the present chapter is to analyze the inverse problem of the previous chapter. In realization theory one assumes that a sequence of Volterra kernels or transfer functions is given and tries to find a corresponding state space description. A common reason for this is that simulation programs usually require the state space form. When constructing Volterra series one of the main methods is to use Carleman bilinearization to get a bilinear system and then use the formula for a bilinear system, (4.23). For the inverse problem it is then natural to search for a bilinear system corresponding to the given Volterra series.

5.1 A convenient form for transfer functions

Consider the formulas for a bilinear system (**??**), (4.23) and make the variable transformation

$$\begin{aligned}\tau_1 &= t_1 - t_2\\ \tau_2 &= t_2 - t_3\\ &\vdots\\ \tau_n &= t_n\end{aligned}$$

This gives

$$y(t) = \sum_{n=1}^{\infty} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} h_{reg}(t_1, \dots, t_n) u(t - t_1 - \dots - t_n) u(t - t_2 - \dots - t_n) \dots \\ \dots u(t - t_n) dt_1 \dots dt_n \quad (5.1)$$

where

$$h_{reg}(t_1, \dots, t_n) = \begin{cases} ce^{At_n} De^{At_{n-1}} D \cdots De^{At_1} b, & t_1 \geq 0, \dots, t_n \geq 0 \\ 0, & \text{otherwise} \end{cases} \quad (5.2)$$

When the Volterra series and its kernels are written in this way we call h_{reg} a *regular kernel*. From the formulas (4.23), (5.2) it follows that the regular and

triangular kernels are related through the formulas

$$h_{reg}(t_1,\dots,t_n) = h_{tri}(t_1+\cdots+t_n, t_2+\cdots+t_n,\dots,t_n) \tag{5.3}$$
$$h_{tri}(t_1,\dots,t_n) = h_{reg}(t_1-t_2, t_2-t_3,\dots,t_{n-1}-t_n,t_n) \tag{5.4}$$

The main advantage of the regular kernel is that the corresponding transfer functions become simple. Taking the multivariable Laplace transform

$$H_{reg}(s_1,\dots,s_n) = \int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty} h_{reg}(t_1,\dots,t_n)e^{-s_1t_1-\cdots-s_nt_n}\,dt_1\cdots dt_n$$

of (5.2) gives

$$H_{reg}(s_1,\dots,s_n) = c(s_nI-A)^{-1}D(s_{n-1}I-A)^{-1}D\cdots D(s_1I-A)^{-1}b \tag{5.5}$$

in analogy with the transfer function formula for linear systems. Laplace transformation of the relations (5.3), (5.4) gives the corresponding relation between the transfer functions

$$H_{reg}(s_1,\dots,s_n) = H_{tri}(s_1, s_2-s_1, s_3-s_2,\dots,s_n-s_{n-1}) \tag{5.6}$$
$$H_{tri}(s_1,\dots,s_n) = H_{reg}(s_1, s_1+s_2, s_1+s_2+s_3,\dots,s_1+\cdots+s_n) \tag{5.7}$$

Example 5.1 Consider the heat exchanger model (4.8).

$$\dot{T} = -2T + uT + u \tag{5.8}$$

Using (5.2) we get the following regular kernels

$$h_1(t_1) = e^{-2t_1},\quad h_2(t_1,t_2) = e^{-2(t_1+t_2)},\dots,h_n(t_1,\dots,t_n) = e^{-2(t_1+t_2+\cdots+t_n)},\dots \tag{5.9}$$

The regular transfer functions are then

$$H_1(s_1) = \frac{1}{s_1+2},\quad H_2(s_1,s_2) = \frac{1}{(s_1+2)(s_2+2)},\dots$$
$$\dots H_n(s_1,\dots,s_n) = \frac{1}{(s_1+2)\cdots(s_n+2)} \tag{5.10}$$

□

5.2 Realizations of finite Volterra series.

Suppose we are given a sequence of regular transfer functions

$$H_1(s_1),\ \ H_2(s_1,s_2),\dots,H_N(s_1,\dots,s_N)$$

The realization problem is to find a state space description of a system with a finite Volterra series, corresponding to these transfer functions. Let us use the notation

$$\hat{H}(s_1,\dots,s_N) = (H_1(s_1), H_2(s_1,s_2),,\dots,H_N(s_1,\dots,s_N)) \tag{5.11}$$

for the sequence of transfer functions.

We will try to find a bilinear system that has the required Volterra series, i.e. a bilinear system

$$\begin{aligned} \dot{x} &= Ax + u\,Dx + b\,u \\ y &= cx \end{aligned} \tag{5.12}$$

with the regular transfer functions satisfying

$$H_m(s_1,\ldots,s_m) = c(s_mI - A)^{-1}D\cdots D(s_1I - A)^{-1}b, \quad m = 1,2,\ldots \tag{5.13}$$

The problem is thus to find matrices c, A, D and b so that (5.13) is satisfied. This problem turns out to be easier to solve if the transfer functions are expanded into negative powers of s_i. Using the expansion

$$(sI - A)^{-1} = s^{-1}I + s^{-2}A + s^{-3}A^2 + \cdots$$

we can write (5.13) as

$$\begin{aligned} H_m(s_1,\ldots,s_m) &= c\sum_{j_n=0}^{\infty} A^{j_n}s_n^{-(j_n+1)}D\cdots D\sum_{j_1=0}^{\infty} A^{j_1}s_1^{-(j_1+1)}b = \\ &= \sum_{j_1=0}^{\infty}\cdots\sum_{j_n=0}^{\infty} cA^{j_n}D\cdots DA^{j_1}bs_1^{-(j_1+1)}\cdots s_n^{-(j_n+1)} \end{aligned} \tag{5.14}$$

Suppose now that each of the given transfer functions is expanded in the same way

$$H_n(s_1,\ldots,s_n) = \sum_{j_1=0}^{\infty}\cdots\sum_{j_n=0}^{\infty} h_{j_1,\ldots,j_n}s_1^{-(j_1+1)}\ldots s_n^{-(j_n+1)} \tag{5.15}$$

We see that we can formulate the realization problem in the following way.

Proposition 5.1 *Finding a bilinear system having a given finite, N:th order, Volterra series is equivalent to finding matrices A,D,b and c such that*

$$cA^{j_m}DA^{j_{m-1}}\cdots DA^{j_1}b = \begin{cases} h_{j_1,\ldots,j_m}, & \text{if } m \le N \\ 0, & \text{if } m > N \end{cases} \tag{5.16}$$

where $h_{j_1,\ldots,j_m}$ is given by (5.15).

In studying the realization problem certain operators turn out to be useful. The operator S transforms a sequence of transfer functions $\hat{H}$ into a new sequence according to the following rules.

$$S\hat{H} = (SH_1, SH_2, \ldots, SH_N) \tag{5.17}$$

$$SH_n(s_1,\ldots,s_n) = s_1H_n(s_1,\ldots,s_n) - [s_1H_n(s_1,\ldots,s_n)]_{s_1=\infty} \tag{5.18}$$

If H_n is expanded this takes the following form

$$SH_n(s_1,\ldots,s_n) = \sum_{j_1=0}^{\infty}\cdots\sum_{j_n=0}^{\infty} h_{j_1+1,\ldots,j_n}s_1^{-(j_1+1)}\ldots s_n^{-(j_n+1)} \tag{5.19}$$

The operator T shifts the transfer functions to the left:

$$T\hat{H} = (TH_2, TH_3, \ldots, TH_{N-1}, 0) \tag{5.20}$$

$$TH_n(s_1,\ldots,s_n) = [s_1 H_n(s_1,\ldots,s_n)]_{s_1=\infty, s_2=s_1,\ldots,s_n=s_{n-1}}, \quad n>1 \tag{5.21}$$

or equivalently for the series expansion

$$TH_n(s_1,\ldots,s_n) = \sum_{j_1=0}^{\infty}\cdots\sum_{j_{n-1}=0}^{\infty} h_{0,j_1,\ldots,j_{n-1}} s_1^{-(j_1+1)}\cdots s_{n-1}^{-(j_{n-1}+1)} \tag{5.22}$$

The evaluation operator E is defined by

$$E\hat{H} = EH_1(s_1) = [s_1 H(s_1)]_{s_1=\infty} \tag{5.23}$$

and is consequently an operator from a sequence of transfer functions to the real numbers. Finally L operates from the real numbers into the space of sequences of transfer functions:

$$Lr = \hat{H}(s_1,\ldots,s_N)r \tag{5.24}$$

for any real number r. To illustrate the use of these operators, consider

$$\hat{H} = (0, H_2(s_1,s_2))$$

We have

$$S^k\hat{H} = (0, S^k H_2(s_1,s_2))$$

where

$$S^k H_2(s_1,s_2) = S^k \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} h_{i,j} s_1^{-(i+1)} s_2^{-(j+1)} = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} h_{i+k,j} s_1^{-(i+1)} s_2^{-(j+1)}$$

Applying the T operator gives

$$T\hat{H} = (TS^k H_2(s_1,s_2), 0)$$

$$TS^k H_2(s_1,s_2) = \sum_{j=0}^{\infty} h_{k,j} s_1^{-(j+1)}$$

More applications of the S-operator give the result

$$S^m T S^k \hat{H} = (S^m T S^k H_2(s_1,s_2), 0)$$

with

$$S^m T S^k H_2(s_1,s_2) = \sum_{j=0}^{\infty} h_{k,j+m} s_1^{-(j+1)}$$

Finally the E-operator gives

$$ES^m T S^k \hat{H} = (ES^m T S^k H_2(s_1,s_2), 0) = h_{k,m}$$

Introducing the L-operator this can be written

$$ES^m T S^k L = h_{k,m}$$

where m and k are arbitrary nonnegative integers.

We see from this simple example that the operators can be used to pick out individual coefficients of the transfer function expansion. The result is easily generalized.

Proposition 5.2 *The operators E,T, S and L are linear and satisfy the relation*

$$ES^{j_m}TS^{j_{m-1}}\cdots TS^{j_1}L = \begin{cases} h_{j_1,\ldots,j_m}, & \text{if } m \le N \\ 0, & \text{if } m > N \end{cases} \tag{5.25}$$

Proof. The linearity is obvious from the definition. The proof of the formula just involves repeated calculations of the type used to show the simple example above. □

We see that the operator formula (5.25) has exactly the same pattern as the matrix formula (5.16). If the linear operators in (5.25) were known to operate between finite dimensional spaces, then they could be represented by matrices. These matrices would then satisfy (5.16) and the realization problem would be solved. The space of all rational transfer functions is an infinite dimensional one however. What might save the situation is the possibility that only a finite dimensional subspace is involved in (5.25). To investigate this question, the following subspaces are introduced.

$$X_1 = span(\hat{H}, S\hat{H}, S^2\hat{H}, \ldots)$$

Letting TX_1 denote the image of X_1 under the operator T, define

$$X_2 = span(TX_1, STX_1, S^2TX_1, \ldots)$$

$$X_3 = span(TX_2, STX_2, S^2TX_2, \ldots)$$

and so on.

In this fashion subspaces $X_1, \ldots, X_N$ are created for the transfer function sequence

$$\hat{H} = (H_1, \ldots, H_N) \tag{5.26}$$

of order N. Define

$$X = span(X_1, \ldots, X_N) \tag{5.27}$$

It is clear that the repeated application of the operators in the left hand side of (5.25) will never lead outside X. This gives the following theorem

Theorem 5.1 A system described by regular transfer functions as in (5.26) is realizable by a bilinear system if and only if X, defined by (5.27) is finite dimensional. In that case the realization implied by (5.25), (5.16) is minimal (i.e. no bilinear system with a lower dimensional state space has the same transfer function).

Proof. Suppose X is finite dimensional. Pick a basis for X. Then the linear operators S, T, L and E can be represented as matrices A,D,b and c that will satisfy (5.16) and a bilinear realization has been found.

Conversely assume that a bilinear system (5.12) exists, having the transfer function $\hat{H}(s_1, \ldots, s_N)$ and a state space of dimension m. If $z \in R^m$, define the function

$$\psi(z) = \left(c(s_1I - A)^{-1}z, c(s_2I - A)^{-1}D(s_1I - A)^{-1}z, \ldots\right)$$

Direct calculations give

$$\psi(Ab) = \left(\sum_{j=0}^{\infty} cA^{j+1}bs_1^{-(j+1)}, \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} cA^jDA^{i+1}bs_1^{-(i+1)}s_2^{-(j+1)}, \ldots\right) =$$

$$= S\hat{H}(s_1, \ldots, s_N)$$

and

$$\psi(Db) = \left(\sum_{j=0}^{\infty} cA^j Dbs_1^{-(j+1)}, \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} cA^j DA^i Dbs_1^{-(i+1)} s_2^{-(j+1)}, \ldots \right) =$$

$$= T\hat{H}(s_1, \ldots, s_N)$$

From repeated calculations of this type it follows that

$$\psi(A^{j_k} D \ldots DA^{j_1} b) = S^{j_k} T \ldots TS^{j_1} H(s_1, \ldots, s_N)$$

which shows that all rational functions in X can be generated by ψ. Since ψ is a linear function defined on an m-dimensional space, it follows that the dimension of X can be at most m and in particular X is finite dimensional. This calculation also shows that any bilinear realization has a state space dimension higher than or equal to the dimension of X, so that the realization defined by the E, S ,T and L operators is minimal. □

Remark 5.1 *The minimality is only with respect to bilinear systems. It is quite possible that there exists a more general nonlinear description with a lower dimensional state space but the same transfer functions. See Exercise 5.2.*

We illustrate the realization procedure with two simple examples.

Example 5.2 Let $\hat{H}$ be given by

$$\hat{H}(s_1, \ldots, s_2) = \left(\frac{1}{s_1 + 1}, \frac{1}{(s_1 + 1)(s_2 + 2)} \right)$$

Then

$$S\frac{1}{s_1 + 1} = \frac{s_1}{s_1 + 1} - 1 = -\frac{1}{s_1 + 1}$$

and

$$S\frac{1}{(s_1 + 1)(s_2 + 2)} = \frac{s_1}{(s_1 + 1)(s_2 + 2)} - \frac{1}{s_2 + 2} = -\frac{1}{(s_1 + 1)(s_2 + 2)}$$

so that

$$S\hat{H} = -\hat{H}$$

Furthermore

$$T\hat{H} = \left(\left. \frac{s_1}{(s_1 + 1)(s_2 + 2)} \right|_{s_1=\infty, s_2=s_1}, 0 \right) = \left(\frac{1}{s_1 + 2}, 0 \right)$$

and

$$S(T\hat{H}) = \left(S\frac{1}{s_1 + 2}, 0 \right) = \left(-\frac{2}{s_1 + 2}, 0 \right) = -2\, T\hat{H}$$

Also

$$T(T\hat{H}) = (0, 0)$$

We see that, no matter how many times the S and T operators are applied, only transfer functions that are linear combinations of $\hat{H}$ and $T\hat{H}$ are generated. The space X is thus two-dimensional. Identifying the basis vectors

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

with $\hat{H}$ and $T\hat{H}$ respectively, we see that the matrices A (corresponding to S) and D (corresponding to T) have to satisfy

$$A\begin{pmatrix} 1 \\ 0 \end{pmatrix} = -\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad A\begin{pmatrix} 0 \\ 1 \end{pmatrix} = -2\begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

$$D\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad D\begin{pmatrix} 0 \\ 1 \end{pmatrix} = 0$$

This gives

$$A = \begin{pmatrix} -1 & 0 \\ 0 & -2 \end{pmatrix}, \quad D = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$$

Since we have

$$E\hat{H} = 1, \quad E(T\hat{H}) = 1, \quad Lr = \hat{H}$$

we get, representing these operators with c and b

$$b = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad c = \begin{pmatrix} 1 & 1 \end{pmatrix}$$

We see that the bilinear system

$$\dot{x} = \begin{pmatrix} -1 & 0 \\ 0 & -2 \end{pmatrix} x + u \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} x + \begin{pmatrix} 1 \\ 0 \end{pmatrix} u$$

$$y = \begin{pmatrix} 1 & 1 \end{pmatrix} x$$

has the given Volterra transfer functions. □

Example 5.3 Consider the transfer function

$$\hat{H} = (0, H_2) = \left(0, \frac{1}{(s_1^2 + 3s_1 + 2)(s_2 + 3)}\right)$$

An application of the S operator gives

$$SH_2 = \frac{s_1}{(s_1^2 + 3s_1 + 2)(s_2 + 3)} - 0$$

and

$$S^2 H_2 = \frac{s_1^2}{(s_1^2 + 3s_1 + 2)(s_2 + 3)} - \frac{1}{s_2 + 3} =$$

$$= \frac{-3s_1 - 2}{(s_1^2 + 3s_1 + 2)(s_2 + 3)} = -3SH_2 - 2H_2$$

so that

$$S^2 \hat{H} = -3S\,\hat{H} - 2\,\hat{H}$$

showing that X_1 is finite dimensional. Now

$$T\hat{H} = (0, 0)$$

while

$$TS\hat{H} = \left(\frac{1}{s_1 + 3}, 0\right)$$

Finally

$$STS\hat{H} = \left(\frac{s_1}{s_1 + 3} - 1, 0\right) = \left(\frac{-3}{s_1 + 3}, 0\right) = -3TS\hat{H}$$

From these calculations we see that no matter how many times we apply the S and T operators, only rational functions that are linear combinations of

$$\left(0, \frac{1}{(s_1^2 + 3s_1 + 2)(s_2 + 3)}\right), \quad \left(0, \frac{s_1}{(s_1^2 + 3s_1 + 2)(s_2 + 3)}\right), \quad \left(\frac{1}{s_1 + 3}, 0\right)$$

are produced, so that X is three dimensional. If these rational functions are used as basis vectors we get

$$A\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad A\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} -2 \\ -3 \\ 0 \end{pmatrix}, \quad A\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ -3 \end{pmatrix}$$

and

$$D\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = 0, \quad D\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \quad D\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = 0$$

showing that

$$A = \begin{pmatrix} 0 & -2 & 0 \\ 1 & -3 & 0 \\ 0 & 0 & -3 \end{pmatrix}, \quad D = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

Since H corresponds to the first basis element, one has

$$b^T = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}$$

and since

$$E\left(\frac{1}{s_1 + 3}, 0\right) = 1$$

while E operating on the other basis elements give zero, c is

$$c = \begin{pmatrix} 0 & 0 & 1 \end{pmatrix}$$

The desired bilinear system is

$$\dot{x} = \begin{pmatrix} 0 & -2 & 0 \\ 1 & -3 & 0 \\ 0 & 0 & -3 \end{pmatrix} x + \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} x\, u + \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} u$$

$$y = \begin{pmatrix} 0 & 0 & 1 \end{pmatrix} x$$

□

Working through some examples one realizes that the procedure always seems to work if the transfer functions are in factored form, with each factor depending on only one variable. Such rational functions deserve a special name.

Definition 5.1 *A rational function $H(s_1, \ldots, s_n)$ is called a* recognizable *function if it can be written in the form*

$$H(s_1, \ldots, s_n) = \frac{P(s_1, \ldots, s_n)}{Q_1(s_1) \cdots Q_n(s_n)}$$

It turns out that the realization problem is solvable precisely when we have such transfer functions.

Theorem 5.2 *An system having the regular transfer function*

$$\hat{H} = (H_1, \ldots, H_N)$$

is bilinear realizable if and only if each H_i is strictly proper and recognizable.

Proof. The only if part follows from (5.13). The if part follows from an investigation of what happens when the S-operator is applied to a rational function of one variable.

$$S\frac{b_1 s^{n-1} + \cdots + b_n}{s^n + a_1 s^{n-1} + \cdots + a_n} = \frac{(b_2 - a_1 b_1)s^{n-1} + \cdots + (-a_n b_1)}{s^n + a_1 s^{n-1} + \cdots + a_n}$$

The result of applying S to a degree n strictly proper rational function is thus a new rational function with the same denominator, but different numerator coefficients. Since there are only n numerator coefficients, the repeated application of the operator can generate at most an n-dimensional space of rational functions. A strictly proper recognizable transfer function can be written

$$\frac{P(s_1, \ldots, s_n)}{Q_1(s_1) \ldots Q_n(s_n)} =$$

$$= \frac{1}{Q_2(s_2) \ldots Q_n(s_n)} \left(P_1(s_2, \ldots, s_n)\frac{s_1^{m-1}}{Q_1(s_1)} + \cdots + P_m(s_2, \ldots, s_n)\frac{1}{Q_1(s_1)} \right)$$

If S is applied to this expression, it will effectively work only on expressions that are rational functions of s_1. The argument above then applies and shows that X_1 is finite dimensional. In the same way X_2 through X_N must be finite dimensional. □

5.3 Realization of infinite Volterra series.

The ideas of the previous section can in principle be extended to infinite Volterra systems, where $\hat{H}$ consists of an infinite sequence of regular transfer functions. However it is obvious that this infinite sequence must have a very special structure for the space X to be finite dimensional. A simple example of such a situation is the following.

Example 5.4 Consider the regular transfer function

$$\hat{H} = \left(\frac{1}{s_1 + a}, \frac{1}{(s_1 + a)(s_2 + a)}, \frac{1}{(s_1 + a)(s_2 + a)(s_3 + a)}, \dots \right)$$

Using the definitions of the operators

$$S\hat{H} = \left(\frac{-a}{s_1 + a}, \frac{-a}{(s_1 + a)(s_2 + a)}, \dots \right) = -a\hat{H}$$

$$T\hat{H} = \hat{H}$$

which shows that X has dimension one. The scalar realization is

$$\dot{x} = -ax + xu + u, \quad y = x$$

(For the special case $a = 2$ this is the heat exchanger of Example 5.1.) □

5.4 A stability result.

Using the realization theory, it is possible to give a stability criterion for homogeneous systems based on the poles of the transfer function.

Theorem 5.3 *Let a system be described by a strictly proper recognizable regular transfer function*

$$H(s_1, \dots, s_n) = \frac{P(s_1, \dots, s_n)}{Q_1(s_1) \dots Q_n(s_n)},$$

If the roots of the factors Q_1 through Q_n all lie strictly in the left half plane then the system is input output stable (in the sense that a bounded input produces a bounded output).

Proof. The transfer function can be realized as a bilinear system using the procedure given in section 5.2. If the natural choice of basis is made, the matrix A will have a block triangular structure where each block has eigenvalues corresponding to the roots of one of the factors Q_i. (This is illustrated in Example 5.3.) The Volterra kernel can now be written in the form (5.2) where the eigenvalues of A all lie strictly in the left half plane. Consequently one has

$$\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} |h(t_1, \dots, t_n)| dt_1 \dots dt_n \leq K$$

for some constant K. If $|u(t)| \leq C$, it follows that

$$|y(t)| \quad \leq \quad \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} |h_n((t_1, \dots, t_n))| C^n dt_1 \dots dt_n \quad \leq \quad KC^n$$

□

5.5 Exercises.

5.1 Suppose the realization procedure described above is applied to the linear system

$$\frac{b_1 s + b_2}{s^2 + a_1 s + a_2}$$

What is the result? What canonical form is it?

5.2 What is the minimal bilinear realization of the regular transfer function

$$\frac{2}{(s_1 + 1)(s_2 + 2)}$$

Try to think of a scalar nonlinear system that realizes this transfer function, showing that the minimal bilinear realization is not minimal in the class of real analytic systems. Hint: Compute the symmetric transfer function.

5.3 Suppose the transfer function of exercise 5.1 is to be realized . What choice of basis in the X-space gives the ordinary observable form (observer form in Kailath's terminology)?

5.4 Give a minimal bilinear realization of the regular transfer function

$$\frac{s_1 s_2 + 1}{(s_1 + 4)(s_1 + 3)(s_2 + 2)(s_2 + 1)}$$

5.5 Can the regular transfer function

$$\frac{1}{s_1 s_2 + 1}$$

be realized by a finite dimensional bilinear system?

5.6 Give a minimal bilinear realization for the transfer function

$$\hat{H}(s_1, s_2) = \left(\frac{1}{s_1 + a}, \frac{1}{(s_1 + b)(s_2 + c)}, 0, 0, \dots \right)$$

5.7 Consider the Volterra system

$$\hat{H} = \left(\frac{b_1}{s_1 + 1}, \frac{b_2}{(s_1 + 1)(s_2 + 1)}, \frac{b_3}{(s_1 + 1)(s_2 + 1)(s_3 + 1)}, \dots \right)$$

where the b-coefficients satisfy

$$R(s) = b_1 s^{-1} + b_2 s^{-2} + b_3 s^{-3} + \cdots$$

for some strictly proper rational function R. Show that the system is realizable by a finite dimensional bilinear system.

Chapter 6

Canonical forms

6.1 Controller forms

Many design methods for nonlinear systems assume that the system has the following triangular form, where we assume u to be a scalar.

$$\begin{aligned}
\dot{x}_1 &= f_1(x_1, x_2) \\
\dot{x}_2 &= f_2(x_1, x_2, x_3) \\
&\vdots \\
\dot{x}_{n-1} &= f_{n-1}(x_1, \ldots, x_n) \\
\dot{x}_n &= f_n(x_1, \ldots, x_n) + g_n(x_1, \ldots, x_n)u
\end{aligned} \tag{6.1}$$

To ensure that the upper parts are connected to the lower parts, and that the control affects the system, it is assumed that

$$g_n \neq 0, \quad \frac{\partial f_j}{\partial x_{j+1}} \neq 0, \;\; j = 1, \ldots, n-1 \tag{6.2}$$

From the triangular form (6.1) it is possible to do exact linearization, Lyapunov based backstepping and many other design methods. Since this system form is useful it is natural to ask if it is possible to transform a system into this form. One can get a feeling for this by looking at successive Lie brackets. With the notation

$$f(x) = \begin{bmatrix} f_1(x_1, x_2) \\ \vdots \\ f_{n-1}(x_1, \ldots, x_n) \\ f_n(x_1, \ldots, x_n) \end{bmatrix}, \quad g(x) = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ g_n(x_1, \ldots, x_n \end{bmatrix}$$

we can calculate the successive Lie brackets

$$[f, g] = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \partial f_{n-1}/\partial x_n \cdot g_n \\ \times \end{bmatrix}, \quad [f, [f, g]] = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \partial f_{n-2}/\partial x_{n-1} \cdot g_n \\ \times \\ \times \end{bmatrix}, \ldots$$

Using the notation

$$(ad^0 f, g) = g, \quad (adf, g) = [f, g], \quad (ad^2 f, g) = [f, [f, g]], \ldots \tag{6.3}$$

and

$$D_k = \text{column vectors with zeros in the first } n-k \text{ positions} \tag{6.4}$$

we have that

$$D_k \text{ is spanned by } (ad^j f, g),\ j = 0, \ldots, k-1$$

Since the first $n-k$ positions of the vectors in D_k are zero, any Lie brackets among such vectors will also have zeros in those positions, i.e. lie in D_k. D_k is thus closed under Lie brackets. A set of vectors having this property is called *involutive*.

Let us now look at generalizations of (6.1) to multi-input systems. Let u be an m-vector and let x_1,...,x_n be vectors whose dimensions are ν_1,...,ν_n respectively. We assume that

$$\text{rank } \frac{\partial f_j}{\partial x_{j+1}} = \nu_j, \quad \text{rank } g(x) = \nu_n \tag{6.5}$$

and that these ranks are constant in some open subset U of the state space. Since

$$\text{rank } \frac{\partial f_j}{\partial x_{j+1}} \leq \dim\ x_{j+1} = \nu_{j+1}$$

it follows that

$$\nu_n \geq \nu_{n-1} \geq \cdots \geq \nu_1 \tag{6.6}$$

Let g_i be the i:th column of g. In analogy with the single input case we get

$$[f, g_i] = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \partial f_{n-1}/\partial x_n \cdot g_{n,i} \\ \times \end{bmatrix}, \quad [f, [f, g_i]] = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \partial f_{n-2}/\partial x_{n-1} \cdot g_{n,i} \\ \times \\ \times \end{bmatrix}, \ldots$$

where $\partial f_{n-1}/\partial x_n$, $\partial f_{n-2}/\partial x_{n-1}$ are now block matrices and $g_{n,i}$ is a ν_n-dimensional column vector. Defining now

$$D_k = \text{column vectors with zeros in the first } \nu_1 + \cdots + \nu_{n-k} \text{ positions} \tag{6.7}$$

we have that

$$D_k \text{is spanned by } (ad^j f, g_i),\ j = 0, \ldots, k-1,\ i = 1, \ldots, m$$

and that for each k,

$$(ad^j f, g_i),\ j = 0, \ldots, k-1,\ i = 1, \ldots, m$$

are involutive.

Let us now consider the possibility of taking a system

$$\dot{\bar{x}} = \bar{f}(\bar{x}) + \bar{g}(\bar{x})u \tag{6.8}$$

with state space dimension $\bar{n}$ and using a coordinate change $x = T(\bar{x})$ to get it into the form (6.1). We assume T to be a diffeomorphism (invertible, infinitely differentiable in both directions).

Theorem 6.1 *The system (6.8) can locally be transformed into the form (6.1), satisfying (6.5) with a diffeomorphism $x = T(\bar{x})$ if and only if*

$$(ad^j \bar{f}, \bar{g}_i), \;\; j = 0, \ldots, k-1, \;\; i = 1, \ldots, m \tag{6.9}$$

are involutive, span constant dimensional spaces and

$$(ad^j \bar{f}, \bar{g}_i), \;\; j = 0, \ldots, \bar{n}-1, \;\; i = 1, \ldots, m \tag{6.10}$$

has dimension $\bar{n}$.

Proof. From our calculations above we saw that the involutivity of (6.9) and the dimension of (6.10) have to be satisfied by the system (6.1). From Proposition 3.2 it follows that the Lie brackets of the system (6.8) must have the same properties.

The sufficiency of these conditions follows from a famous theorem by Frobenius, but the details are omitted here. □

The actual calculation of the coordinate change can be complicated and we postpone that discussion.

Instead we note that it is possible to proceed from the triangular form (6.1) to other standard forms. Suppose for simplicity that all x_i have the same dimension. Then we can introduce new coordinates successively by

$$z_1 = x_1, \quad z_2 = f_1(x_1, x_2)$$

Since $\partial f_1/\partial x_2)$ has full rank, this coordinate change is locally invertible. We get

$$\begin{aligned}
\dot{z}_1 &= z_2 \\
\dot{z}_2 &= \partial f_1/\partial x_1 f_1 + \partial f_1/\partial x_2 f_2 = \tilde{f}_2(x_1, x_2, x_3)
\end{aligned}$$

Introducing $z_3 = \tilde{f}_2(x_1, x_2, x_3)$ we get $\dot{z}_2 = z_3$. Proceeding in this fashion we get the coordinate change

$$\begin{aligned}
z_1 &= x_1 \\
z_2 &= f_1(x_1, x_2) \\
&\vdots \\
z_j &= \tilde{f}_{j-1}(x_1, \ldots, x_j) \\
&\vdots \\
z_n &= \tilde{f}_{n-1}(x_1, \ldots, x_n)
\end{aligned}$$

It is easy to see that the conditions $\partial f_{j-1}/\partial x_j \neq 0$ carry over into $\partial \tilde{f}_{j-1}/\partial x_j \neq 0$. The coordinate change is thus locally invertible. The system dynamics be-

comes

$$\begin{aligned}\dot{z}_1 &= z_2\\ \dot{z}_2 &= z_3\\ &\vdots\\ \dot{z}_{n-1} &= z_n\\ \dot{z}_n &= \tilde{f}_n(z) + \tilde{g}_n(z)u\end{aligned}$$

In the general case when the x_i do not have the same dimension the coordinate change is somewhat more involved. It still begins with $z_1 = x_1$. Now suppose that $\nu_2 > \nu_1$. Then divide x_2 into two components $\hat{x}_2$ and $\bar{x}_2$ (possibly after reordering the variables) so that $\partial f_1/\partial \hat{x}_2$ is nonsingular (this is always possible since $\partial f_1/\partial x_2$ has full rank). Then introduce the new variable as

$$z_2 = \begin{bmatrix}\hat{z}_2\\ \bar{z}_2\end{bmatrix} = \begin{bmatrix}f_1(x_1, x_2)\\ \bar{x}_2\end{bmatrix}$$

The coordinate change is still invertible, since $\partial f_1/\partial \hat{x}_2$ is nonsingular. Introducing new variables successively in this fashion gives the following description

$$\dot{z} = A_0 z + \begin{bmatrix}0\\ \vdots\\ 0\\ \hat{f}_n(z)\end{bmatrix} + \begin{bmatrix}0\\ \vdots\\ 0\\ \hat{g}_n(z)\end{bmatrix} u \tag{6.11}$$

where A_0 is a matrix of the form

$$A_0 = \begin{bmatrix}0 & E_1 & 0 & \dots & 0\\ 0 & 0 & E_2 & \dots & 0\\ \vdots & & & \ddots & \vdots\\ 0 & 0 & \dots & 0 & E_{n-1}\\ 0 & 0 & \dots & 0 & 0\end{bmatrix} \tag{6.12}$$

and each E_i consists of a ν_i-dimensional unit matrix and a $\nu_1 \times (\nu_2 - \nu_1)$ zero matrix as follows.

$$E_i = \begin{bmatrix}I_{\nu_i} & 0_{\nu_1 \times (\nu_2 - \nu_1)}\end{bmatrix} \tag{6.13}$$

By using state feedback we can perform the ultimate simplification, converting the system to a chain of integrators. Assume for simplicity that g is invertible (otherwise some redundant control signals can be removed). Use the feedback

$$u = \hat{g}_n(z)^{-1}(-f_n(z) + v) \tag{6.14}$$

the system becomes

$$\dot{z} = A_0 z + \begin{bmatrix}0\\ \vdots\\ 0\\ I_m\end{bmatrix} v \tag{6.15}$$

This is called a *Brunovsky canonical form*. Note that in particular the system is transformed into a linear system. Define the numbers

$$\rho_i = \text{ number of } \nu_k \text{ such that } \nu_k \geq i, \quad i = 1, \ldots, m \tag{6.16}$$

From the definition

$$\rho_1 \geq \rho_2 \geq \cdots \geq \rho_m \tag{6.17}$$

The ρ_i are called the *controllability indices* of the system. They can be interpreted as the length of the chains of integrators in A_0. This can be seen by permuting the variables. Let x_{ij} denote the j:th element of the vector x_i. Define

$$\zeta_1 = z_{11}, \ \zeta_2 = \dot{\zeta}_1 = z_{21}, \ \zeta_3 = \dot{\zeta}_2 = z_{31}, \ldots, \zeta_{\rho_1} = z_{n1}$$

If $\nu_1 > 1$, take

$$\zeta_{\rho_1+1} = z_{12}, \ \zeta_{\rho_1+2} = z_{22}, \ldots$$

otherwise take

$$\zeta_{\rho_1+1} = z_{22}, \ \zeta_{\rho_1+2} = z_{32}, \ldots$$

Continuing in this fashion changes (6.11) to

$$\begin{aligned}
\dot{\zeta}_1 &= \zeta_2 \\
\dot{\zeta}_2 &= \zeta_3 \\
&\vdots \\
\dot{\zeta}_{\rho_1} &= a_1(\zeta) + b_1(\zeta)u \\
\dot{\zeta}_{\rho_1+1} &= \zeta_{\rho_1+2} \\
\dot{\zeta}_{\rho_1+2} &= \zeta_{\rho_1+3} \\
&\vdots \\
\dot{\zeta}_{\rho_1+\rho_2} &= a_2(\zeta) + b_2(\zeta)u \\
&\vdots \\
\dot{\zeta}_{\rho_1+\cdots+\rho_n} &= a_{\nu_n}(\zeta) + b_{\nu_n}(\zeta)u
\end{aligned} \tag{6.18}$$

where a_i and b_i are the i:th rows of $\tilde{f}_n$ and $\tilde{g}_n$ respectively, with the variables permuted suitably.

6.2 Computing the coordinate change

Let us consider the problem of finding a coordinate transformation $z = T(x)$ going directly from (6.8) to the form (6.18). Assume that we have checked the conditions on the Lie brackets specified in Theorem 6.1. This also gives the numbers ν_i and ρ_i. Let us write

$$\zeta_1 = \phi_1(\bar{x})$$

where ϕ_1 is an unknown function to be determined. To simplify notation we will write x instead of $\bar{x}$ and f, g instead of $\bar{f}$, $\bar{g}$. We have

$$\dot{\zeta}_1 = L_f\phi_1 + \sum_{i=1}^{m} u_i L_{g_i}\phi_1$$

If $\rho_1 > 1$, then we want $\dot{\zeta}_1 = \zeta_2$. We then get the conditions

$$L_{g_i}\phi_1 = 0, \;\; i = 1, \ldots, m$$

and ζ_2 has to be defined as

$$\zeta_2 = (L_f\phi_1)(x)$$

If $\rho_1 > 2$ we differentiate and get

$$\dot{\zeta}_2 = L_f^2\phi_1 + \sum_{i=1}^{m} u_i L_{g_i} L_f \phi_1$$

which gives the conditions

$$L_{g_i} L_f \phi_1 = 0, \;\; i = 1, \ldots, m$$

and $\zeta_3 = (L_f^2\phi_1)(x)$. Using the formula

$$L_f L_g - L_g L_f = L_{[f,g]} \tag{6.19}$$

this can be rewritten

$$L_{[f,g_i]}\phi_1 = 0, \;\; i = 1, \ldots, m$$

Continuing in this fashion, and doing analogous calculations for

$$\zeta_{\rho_1+1} = \phi_2(x), \ldots \zeta_{\rho_1+\cdots+\rho_{m-1}+1} = \phi_m(x)$$

gives the following set of partial differential equations for ϕ_k, $k = 1, \ldots, m$.

$$\begin{aligned} L_{g_i}\phi_k &= 0, \quad i = 1, \ldots, m \\ L_{[f,g_i]}\phi_k &= 0, \quad i = 1, \ldots, m \\ &\vdots \\ L_{(ad^{\rho_k-2}f,g_i)}\phi_k &= 0, \quad i = 1, \ldots, m \end{aligned} \tag{6.20}$$

The functions b_i in (6.18) will be given by

$$L_{(ad^{\rho_k-1}f,g_i)}\phi_k, \quad k = 1, \ldots, m, \;\; i = 1, \ldots, m \tag{6.21}$$

The functions ϕ_k should also be chosen so that these quantities form a nonsingular matrix.

Example 6.1 Consider the following system of tanks ("the Lund tanks").

$$\begin{aligned} \dot{x}_1 &= \gamma u_1 + \sqrt{x_3} - \sqrt{x_1} \\ \dot{x}_2 &= \gamma u_2 + \sqrt{x_4} - \sqrt{x_2} \\ \dot{x}_3 &= (1-\gamma)u_2 - \sqrt{x_3} \\ \dot{x}_4 &= (1-\gamma)u_1 - \sqrt{x_4} \end{aligned}$$

where $0 \leq \gamma < 1$. We have that

$$g_1 = \begin{bmatrix} \gamma \\ 0 \\ 0 \\ 1-\gamma \end{bmatrix}, \quad g_2 = \begin{bmatrix} 0 \\ \gamma \\ 1-\gamma \\ 0 \end{bmatrix}, \quad [f, g_1] = \begin{bmatrix} \frac{\gamma}{2\sqrt{x_1}} \\ -\frac{1-\gamma}{2\sqrt{x_4}} \\ 0 \\ \frac{1-\gamma}{2\sqrt{x_4}} \end{bmatrix}, \quad [f, g_2] = \begin{bmatrix} -\frac{1-\gamma}{2\sqrt{x_3}} \\ \frac{\gamma}{2\sqrt{x_2}} \\ \frac{1-\gamma}{2\sqrt{x_3}} \\ 0 \end{bmatrix}$$

Clearly g_1 and g_2 span a two-dimensional space. Since they are constant, they are automatically involutive. Also g_1, g_2, $[f, g_1]$ and $[f, g_2]$ span the whole space. It follows that Theorem 6.1 is satisfied with $\nu_1 = 2$, $\nu_2 = 2$. Consequently the controllability indices are $\rho_1 = 2$, $\rho_2 = 2$. The conditions (6.20) become

$$L_{g_1}\phi_1 = 0, L_{g_2}\phi_1 = 0, \quad L_{g_1}\phi_2 = 0, \quad L_{g_2}\phi_2 = 0$$

Since g_1 and g_2 are constant, it is natural to try linear coordinate changes ϕ_1 and ϕ_2.

$$\phi_1(x) = v_1^T x, \quad \phi_2(x) = v_2^T x$$

The conditions are then

$$v_1^T g_1 = 0, \quad v_1^T g_2 = 0, \quad v_2^T g_1 = 0, \quad v_2^T g_2 = 0$$

For instance we can select

$$z_1 = (1-\gamma)x_1 - \gamma x_4, \quad z_3 = (1-\gamma)x_2 - \gamma x_3$$

It follows that

$$\begin{aligned}\dot{z}_1 &= (1-\gamma)(\sqrt{x_3} - \sqrt{x_1}) + \gamma\sqrt{x_4} = z_2\\ \dot{z}_3 &= (1-\gamma)(\sqrt{x_4} - \sqrt{x_2}) + \gamma\sqrt{x_3} = z_4\end{aligned}$$

and with this coordinate change the system is

$$\begin{aligned}\dot{z}_1 &= z_2\\ \dot{z}_2 &= f_1(z) + g_{11}(z)u_1 + g_{12}(z)u_2\\ \dot{z}_3 &= z_4\\ \dot{z}_4 &= f_2(z) + g_{21}(z)u_1 + g_{22}(z)u_2\end{aligned}$$

where f_1, f_2 and the g_{ij} are computed by differentiating z_2 and z_4. □

6.3 Exercises

6.1 Finish Example 6.1 by computing the f_i and the g_{ij}.

6.2 Compute the feedback that gives the Brunovsky canonical form for the preceding example.

6.3 Prove that

$$L_f L_g - L_g L_f = L_{[f,g]}$$