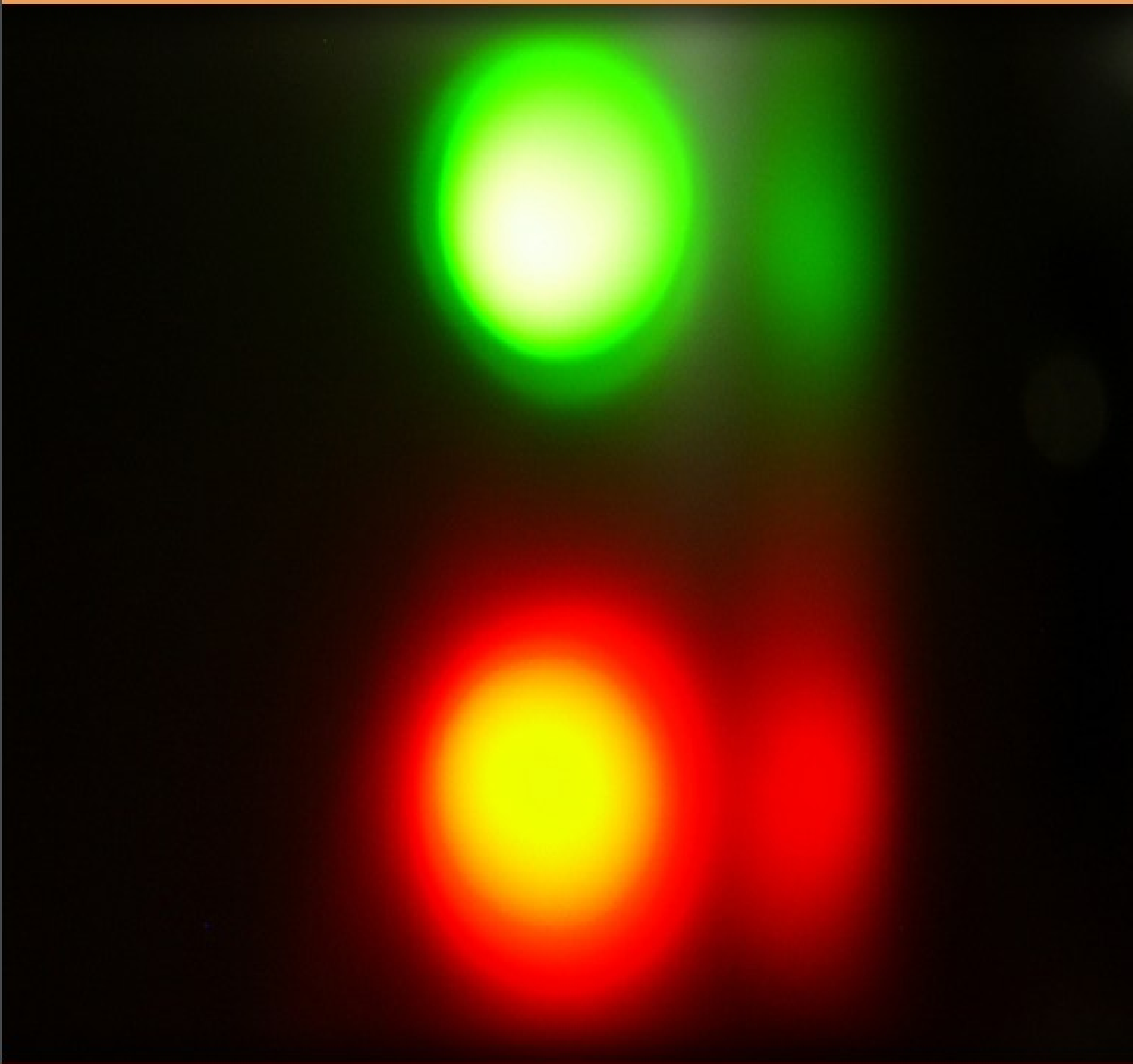
An Introduction to Nonlinearity in Control Systems

Derek Atherton



Derek Atherton

An Introduction to Nonlinearity in Control Systems

An Introduction to Nonlinearity in Control Systems

ISBN 978-87-7681-790-9

Contents

Preface

The book is intended to provide an introduction to the effects of nonlinear elements in feedback control systems. A central topic is the use of the Describing Function (DF) method since in combination with simulation it provides an excellent approach for the practicing engineer and follows on logically from a first course in classical control, such as the companion volume in this series. Some of the basic material on the topic can be found in my earlier book which is frequently referenced throughout the text.

The first chapter provides an introduction to nonlinearity from the basic definition to a discussion of the possible effects it can have on a system and the different behaviour that might be found, in particular when it occurs within a cascade system of elements or in a feedback loop. The final part of the chapter gives a brief overview of the contents of the book.

Phase plane methods for second order systems are covered in the second chapter. Many systems, particularly electromechanical ones, can be approximated by second order models so the concept can be particularly useful in practice. The second order linear system is first considered as surprisingly it is rarely covered in linear control system texts. The method has the big advantage in that the effects of more than one nonlinear element may be considered. The study is supported by several simulations done in Simulink including one where sliding motion takes place.

The third chapter is the first of three devoted to the study of feedback loops using DF methods. Although the method is an approximate technique its value, and limitations, are supported by a large number of examples containing analytical results and simulations, including the estimation of limit cycles and loop stability. Many early papers on DFs showing how theories could be used to predict specific phenomena were supported by simulations done on analogue computers. Here results from digital simulations using Simulink are presented and this allows much more control of initial conditions to show how different modes may exist dependent on the initial conditions. In particular, Chapter 5 contains some more advanced work including some new results on jump resonance, so some readers may wish to omit this chapter on a first reading.

A relay is a unique nonlinear element in that its output does not depend upon the input at all times but is determined by when the input passes through the relay switching levels. It is this feature which allows the exact determination of limit cycles and their stability in a feedback loop. The basic theory is presented and some simple examples covered. It is then shown in section 6.6 how the approach can be used with computational support to analyse quite complicated periodic modes in relay systems. Among the topics covered for the first time in a textbook are the evaluation of limit cycles with multiple pulses per cycle, as found in a satellite attitude control system, the determination of a limit cycle with sliding and other more advanced aspects which some readers may wish to omit.

A practical method developed in recent years for finding suitable parameters, or tuning, for a controller based on the so called loop cycling method of Ziegler and Nichols has used relay produced limit cycles. Chapter 7 covers these ideas using both approximate analysis based on the DF and exact analysis based on the relay methods of the previous chapter.

Chapter 8 covers the topic of absolute stability namely trying to obtain necessary and sufficient conditions for the stability of a feedback loop with a single nonlinear element. This problems has exercised the minds of theorists for nearly a century but a solution seems no nearer! Several necessary but not sufficient results are presented, which dependent on ones viewpoint may be regarded as 'conservative' or 'robust'. The former is usually the case when one has a mathematically defined nonlinearity and the latter may be used because the result gives stability for any nonlinearity with certain properties, for example, lying within a sector.

The final chapter, chapter 9, discusses quite briefly various methods which can be used for the design of nonlinear systems. The intent has been to provide sufficient information on the methods and their possible advantages and disadvantages. Several of them have complete books written on the topic and more detail could not have been given without the coverage of more specialised mathematics.

Finally my thanks to the University of Sussex for the use of an office and access to the computing facilities during my retirement; to my good friend Keith Godfrey and his student Roland for the computational input on jump resonance and their companionship at the races and to my wife, Constance, for her love and support.

Derek P Atherton
University of Sussex
Brighton
May 2011.

1 Introduction

1.1 What is nonlinearity?

In order to analyse the behaviour of engineering systems mathematical models are required for the various components. It is common practice to try and obtain linear models as a rich mathematical theory exists for linear systems. These models will always be approximate, although possibly quite accurate for defined ranges of system variables, but inevitably nonlinear effects will eventually be found for large excursions of system variables. Linear systems have the important property that they satisfy the superposition principle. This leads to many important advantages in methods for their analysis. For example, in circuit theory when an RLC circuit has both d.c and a.c input voltages, the voltage or current elsewhere in the circuit can be found by summing the results of separate analyses for the d.c. and a.c. inputs taken individually, also if the magnitude of the a.c. voltage is doubled then the a.c. voltages and currents elsewhere in the circuit will be doubled.

Thus, mathematically a linear system with input $x(t)$ and output $y(t)$ satisfies the property that the output for an input $ax_1(t) + bx_2(t)$ is $ay_1(t) + by_2(t)$, if $y_1(t)$ and $y_2(t)$ are the outputs in response to the inputs $x_1(t)$ and $x_2(t)$, respectively, and a and b are constants. A further point is that when the a.c. voltage is sinusoidal, which is normally assumed in using the term a.c., then all the other voltages and currents in the circuit are sinusoidal with the same frequency as the input.

A nonlinear system is defined as one which does not satisfy the superposition property. The simplest form of nonlinear system is the static nonlinearity where the output depends only on the current value of input but in a nonlinear manner, for example

$$y(t) = cx(t) + dx^3(t) \tag{1.1}$$

where the output is the summation of a linear and a nonlinear, cubic, term. This description of nonlinearity is given by a simple mathematical expression. Practical nonlinearities which occur in control engineering, however, can often not be so easily described; so that approximating them may require more complex mathematical models, such as a high order power series or a linear segmented approximation.

More commonly a nonlinear differential equation, for example

$$d^2y(t)/d(t)^2 + c(dy(t)/dt)^3 + y(t) = x((t) \tag{1.2}$$

will describe the behavior. From an engineering viewpoint it may be desirable to think of this equation in terms of a block diagram, as shown in Figure 1.1, consisting of linear dynamic elements and a static nonlinearity, in this case a cubic, which with input $dy(t)/dt$ gives an output $c(dy(t)/dt)^3$. Unfortunately there is no general approach to solving nonlinear, unlike linear, differential equations. The major point about nonlinear systems, however, is that their response is amplitude dependent so that if a particular form of response, or some measure of it, occurs for one input magnitude it may not result for some other input magnitude.

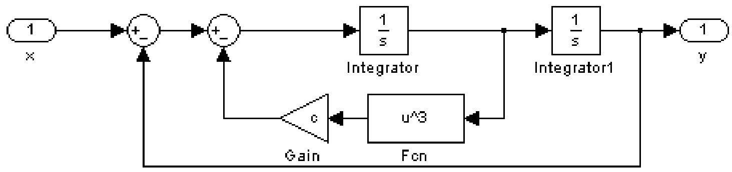


Figure 1.1 Block diagram in Simulink for equation 1.2

A further and very important point, is that unlike a linear operation, a nonlinear operation on a sinusoid of frequency, *f*, will not produce an output at frequency *f*, alone. For example, if such a sinusoid is applied to the static nonlinearity of equation (1.1) it is easy to show from substituting $x(t) = a\cos\omega t$ where $\omega = 2\pi f$ that there are outputs at frequencies *f* and 3*f* of magnitudes $ca + 3da^3/4$ and $da^3/4$, respectively. Perhaps the most interesting aspect of nonlinear systems is that, as will be shown later, they exhibit several forms of unique behaviour which are not possible in linear systems.

1.2 Forms of nonlinearity

All practical systems are nonlinear and in this section a brief overview is given of some nonlinear effects that often occur. In initial designs it may be possible to approximate the nonlinear effects by linear models but invariably it will be necessary to finally check their effects on the system performance either in simulation or/and the real hardware. Today's digital simulation languages are very good but to use them efficiently for investigating the effects of nonlinearity, or to assist in the design of a nonlinear system, requires a knowledge of the supporting theoretical methods presented in this book.

In typical control engineering problems nonlinearity may occur in the dynamics of the plant to be controlled or in the components used to implement the control. In the latter case, for example, a valve actuator may have a dead zone due to friction effects and will certainly saturate for large inputs, so this may be referred to as an inherent nonlinearity, because it exists although one might possibly prefer this not to be the case. Alternatively one may have intentional nonlinearities which have been purposely designed into the system to improve the system specifications, either for technical or economic reasons. A good example of this is the on-off control used in many temperature control systems, where the objective is to have the temperature oscillate about the required value.

Identifying the precise form of a nonlinearity may not be easy and like all modeling exercises the golden rule is to be aware of the approximations in a nonlinear model and the conditions for its validity. It might be argued that linear systems theory is not applicable to practical control engineering problems because they are always nonlinear. This is an overstatement, of course, but a valid reminder. All systems have actuator saturation and in some cases it might occur for relatively low error signals, for example in rotary position control it is not unusual for a step input of say 10°, or even less, to produce the maximum motor drive torque. It simply is the result of good economical design as to produce linear operation to a higher torque level would require a larger and more expensive motor. Valves used to control fluid or gas flow, apart from having the nonlinear effects mentioned above, can have a slightly different behavior when opening compared with closing due to the unidirectional pressure of the fluid.

Friction always occurs in mechanical systems and is very difficult to model, with many quite sophisticated models having been presented in the literature. The simplest is to assume the three components, illustrated in Figure 1.2, of stiction, an abbreviation for static friction, Coulomb friction and viscous friction. As its name implies stiction is assumed to exist only at zero differential speed between the two contact surfaces. Coulomb friction with a value less than stiction is assumed to be constant at all speeds, and viscous friction is a linear effect being directly proportional to speed. In practice there is often a term proportional to a higher power of speed, and this is also the situation for many shaft loads, for example a fan for which the drive torque typically increases as a power of speed.

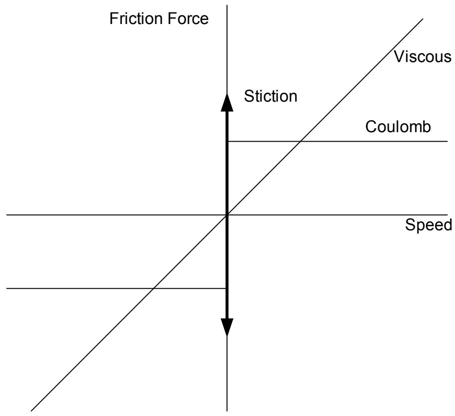


Figure 1.2 Three basic friction components

A mathematical expression sometimes used to approximate friction is

$$n(v) = (a - be^{-c|v|} + d\,|\,v\,|)\,\mathrm{sgn}(v) \tag{1.3}$$

Here the parameters *a, b, c* and *d* are chosen to provide the desired shape against speed which initially decreases from the value at zero speed and then starts to increase.

Many mechanical loads are driven through gearing rather than directly. Although geared drives, like all areas of technology, have improved through the years they always have some small backlash. This may be avoided by using anti-backlash gears, which are only available for low torques. Backlash, which is a very complicated phenomenon, involving impacts between surfaces is often modeled in a very simplistic manner. For example, the simple approach used in some digital simulation languages, such as Simulink, the simulation component of the well-known software package, Matlab, consists of an input-output position characteristic of two parallel straight lines with possible horizontal movement between them.

This makes two major assumptions, first that the load shaft friction is high enough for contact to be maintained with the drive side of the backlash when the drive slows down to rest. Secondly when the drive reverses the backlash is crossed and the new drive side of the gear 'picks up' the load instantaneously with no loss of energy in the impact and both then move at the drive shaft speed. Clearly both these assumptions are never true in practice but no checks exist in Simulink to determine how good they are or, indeed, when they are completely invalid. Better facilities for modeling phenomena such as backlash can be found in simulation languages such as 20-Sim or Dynast, which do not use the block concept of Simulink. A good discussion of friction and backlash can be found in reference1.1.

Probably, the most widely used intentional nonlinearity is the relay. The on-off type, which can be described mathematically by the signum function, that is switches on if its input exceeds zero and off if it goes below zero, is widely used normally with some hysteresis between the switching levels. Use of this approach provides a control strategy where the controlled variable oscillates about the desired level. The switching mechanism varies significantly according to the application from electromechanical relays at low speed to fast electronic switches employing transistors or thyristors. A common usage of the relay is in the temperature control of buildings, where typically the switching is provided from a temperature sensor having a pool of mercury on a metal expansion coil. As the temperature drops the coil contracts and this causes a change in angle of the mercury capsule so that eventually the mercury moves and closes a contact. When the temperature increases the coil expands causing a change in angle for the mercury to flow and break the contact.

Electronic switching controllers are being used in many modern electric motor drive systems, for example, to regulate phase currents in stepping motors and switched reluctance motors and to control currents in vector control drives for induction motors. Relays with a dead zone, that is, three position relays giving positive, negative and a zero output are also used. When used in a position control system the zero output allows for a steady state position within the dead zone but this affects the resulting steady state control accuracy.

Hysteresis effects in magnetic materials sometimes have to be modeled. This is often done by assuming a hysteresis loop of the form of a B-H loop for a magnetic material typically obtained for a sinusoidal input. However the shape usually varies with the amplitude and frequency of the input and does not in fact remain constant with a random excitation. Provided the input is sinusoidal and the shape does remain reasonably constant then a nonlinear function of the form $n(x, \dot{x})$ may provide a reasonable model as the path taken around the hysteresis depends upon whether the input, x, is increasing or decreasing, i.e the derivative of x, $\dot{x}$.

1.3 Structure and Behaviour

From a control engineering viewpoint there are two major reasons why one needs to know about nonlinearity. Firstly with respect to obtaining mathematical models of devices, particularly if identification techniques are being used, and secondly for ensuring the design meets the desired specifications when the control system is nonlinear. To appreciate these aspects it is appropriate to discuss very briefly in the introduction a few aspects of behaviour due to the presence of nonlinearity. These are dependent on the structure of the nonlinear system and the relevance of this is explained by again considering a sinusoidal input.

Consider a nonlinearity $x + dx^3$, and a dynamic element with transfer function $K/s(s+b)$, where b and d are constants. If they are placed in cascade and a sinusoidal input applied the output will be a deterministic waveform containing two frequency components one at the same frequency as the input and the other at three times that frequency. Any cascade combination of linear and nonlinear elements will always produce a deterministic output for any given discrete input frequency spectrum, which in principle can be evaluated. New frequency components can only be created by the nonlinearities and the linear elements simple alter the relative magnitudes and phases of these components.

For example, if the above nonlinearity is placed before and after the linear transfer function and a sinusoid of frequency, *f*, is applied at the input, then the input to the second nonlinearity will consist of the fundamental, *f*, plus third harmonic, 3*f*, with magnitudes and phases dependent on both the input sinusoidal magnitude and frequency. These two frequencies applied to the second nonlinearity will produce an output containing the frequency components, *f, 3f, 5f, 7f,* and *9f*. One could define a frequency response for such a cascade structure of linear and nonlinear elements as the ratio of the output at the fundamental frequency, *f*, to the input sinusoid at this frequency. The result, as for a linear system, would be a magnitude and phase, which varies with, *f*, but because of the nonlinearities it would also vary with the amplitude of the input sinusoid. Thus an approximate frequency response model for the combination could be portrayed graphically by a set of frequency response plots for different input amplitudes, or gain and phase plots against amplitude for different frequencies.

For many problems encountered in control engineering this may prove to be a reasonably good model since many of the linear dynamic elements, like the one given, have low pass dynamics so that the frequency, *f*, will predominate at the output. With no linear dynamic elements in the combination then these latter plots would be the same for all frequencies and the approximate, first harmonic, or quasi-linearized, model would be gain and phase curves as a function of the input amplitude. This representation of a nonlinear element is known as a describing function (DF), which is the topic of several chapters. Sometimes the above model for the combination based on the fundamental frequency, *f*, only, is referred to as an amplitude and frequency dependent describing function.

If alternatively we assume the system structure to consist of a feedback combination of the linear and nonlinear elements, with the transfer function $K/s(s+b)$ in the forward loop and the feedback loop containing $x + dx^3$ fed back through a negative gain then, a very different situation is possible for the response to a sinusoidal input. Dependent on the values of *b, d* and the amplitude and frequency of the input, some possibilities for the output are that it is (a) approximately sinusoidal with the same frequency as the input, similar to the aforementioned cascade connection; (b) approximately sinusoidal with a frequency related to that of the input; (c) a combination of primarily the input frequency and another frequency or (d) a waveform known as chaotic, which is not definable mathematically but completely repeatable for the same initial conditions.

These behaviours and others are unique to nonlinear feedback systems, aspects which make such systems extremely interesting. However, it has meant that no general analytical method is available for predicting their behavior. Several approaches will be considered in this book all of which will be restricted in their applicability or, put alternatively, the situations which they can address. Thus the importance of simulation studies for investigating nonlinear systems in association with analytical methods cannot be underestimated. Much of the support for the theoretical material presented in the early chapters, particularly in the 50s to 60s, was done using analogue simulation. Today simulations are done digitally and several are included using Simulink in the following chapters to illustrate the concepts and provide solutions for specific problems. Care has to be taken in simulating nonlinear systems particularly those with linear segmented characteristics because of the discontinuities. Some comments are made on the simulations where appropriate.

1.4 Overview of contents

The phase plane approach discussed in chapter 2 is very useful for step response and stability studies but is basically restricted to second order systems. However, many engineering systems, particularly in the mechatronics field, may be approximated by a second order differential equation so the results are still of value. It also provides a simple basis for understanding some of the more advanced topics, such as optimum control and sliding mode control, covered in chapter 9.

Stability of a feedback loop is of major importance so that it is not surprising that much early work was concerned with this topic. During the 1940's engineers in several countries developed what has become known as the describing function method where a nonlinearity is replaced by an amplitude dependent gain to a sinusoid, known as the describing function, DF. Chapter 3 introduces the DF, shows how its value can be calculated for various nonlinearities and includes a table of results. The next two chapters, 4 and 5, deal with applications of the describing function for estimating limit cycles and the stability of a nonlinear feedback loop. It is also shown how the describing function can be evaluated for other than a single sinusoidal input and applications of describing functions for a bias plus sinusoid and two sinusoids are given. These include areas such as limit cycle stability, jump resonance and subharmonic oscillations.

The relay is a special form of nonlinear element where the output is not continuously dependent on the input. This allows special results to be developed for relay systems a topic covered in chapter 6. Chapter 7 deals with a design approach that has received much attention in recent years, namely using the information contained in relay produced limit cycles to set the parameters of a controller, typically known as relay autotuning. The topic of absolute stability, namely the development of exact criteria for guaranteeing the stability of a feedback loop with a single nonlinearity and linear transfer function, is covered in chapter 8. Dependent on the viewpoint these results may be regarded as robust, as they prove stability for variations in the nonlinearity, or conservative if one is considering stability for a specific nonlinearity.

The coverage to this point, apart from chapter 7, has like many textbooks on linear control, been primarily concerned with introducing analytical tools. Chapter 9, however, focuses more on design and looks at how some of the ideas covered and other techniques may be used for nonlinear control system design. The coverage of these topics is quite brief, some necessarily so because of the additional theoretical concepts which would have to be introduced to go into them more deeply. Hopefully, sufficient information, together with the references, is given for the reader to understand the concepts involved, their possible relevance for particular applications and how they might be applied.

1.5 References

1.1 Friedland, B, 1996 Advanced control system design, Chapter 7 Prentice Hall.

2 The Phase Plane Method

2.1 Introduction

A significant amount of the research into nonlinear differential equations in the nineteenth and early twentieth centuries done by mathematicians and physicists was devoted to second order differential equations. There were two major reasons for this, namely that the dynamics of many problems of practical interest could be approximated by these equations and secondly the phase plane approach allowed a graphical examination of their solutions. The systems of interest in the late nineteenth and early twentieth centuries were found in fields such as celestial mechanics, nonlinear mechanical systems and electronic oscillations. This section will introduce the basic concepts of the phase plane approach and then give a brief overview of how the method has been further developed for use in control system analysis and design.

A significant amount of the early development in control theory from 1930 was driven by two areas, namely to achieve better control of industrial processes and to achieve better performance in fire control problems. The problems in the latter area received significantly more attention in the war years after 1939, where the requirement in many cases was related to the position control of radar antennas and guns in both stationary and moving situations. The dynamic equations representing many of these position control systems could be represented reasonably accurately by second order nonlinear differential equations. It was therefore not surprising that much of the early work on nonlinear control used the phase plane approach. Control engineers did make significant contributions to this field since, whereas the earlier work had typically assumed nonlinearities defined by continuous mathematical functions, for control system analysis it was often more appropriate to approximate intrinsic nonlinearity, such as friction, or intentionally introduced nonlinearity, such as a relay, by linear segmented characteristics. The approach is still useful today because of the fundamental insight it provides into aspects of nonlinear system behaviour; the fact there are still many control problems which can be approximated by second order dynamics, and also because more than one nonlinearity can be considered.

2.2 Basic Principles

The formulation used in early work on second order systems was to assume a representation in terms of the two first order equations

$$\begin{aligned} \dot{x}_1 &= P(x_1, x_2) \\ \dot{x}_2 &= Q(x_1, x_2) \end{aligned} \tag{2.1}$$

Equilibrium, or singular, points represent a stationary system for the dynamics and occur when $\dot{x}_1 = \dot{x}_2 = 0$

The slope of any solution curve, or trajectory, in the x_1 - x_2 state plane is

$$\frac{dx_2}{dx_1} = \frac{\dot{x}_2}{\dot{x}_1} = \frac{Q(x_1,x_2)}{P(x_1,x_2)} \tag{2.2}$$

Before considering the nonlinear case further it is important to first examine the linear case, which is rarely discussed in detail in texts on linear control. Most, usually, just restrict the contribution to considering the step response of a second order system.

2.2.1 The Linear Case

The equations for the linear situation may be written

$$\begin{aligned} \dot{x}_1 &= ax_1 + bx_2 \\ \dot{x}_2 &= cx_1 + dx_2 \end{aligned} \tag{2.3}$$

or in matrix form

$$\begin{aligned} \dot{x} &= Ax \\ where \quad A &= \begin{pmatrix} a & b \\ c & d \end{pmatrix}. \end{aligned} \tag{2.4}$$

Eliminating x_2 from equations (2.3) by differentiating the first equation and substituting for $\dot{x}_2$ from the second equation and x_2 from the first yields

$$\ddot{x}_1 - (a + d)\dot{x}_1 + (ad - bc)x_1 = 0$$

which after taking Laplace transforms has the characteristic equation

$$s^2 - (a + d)s + ad - bc = 0$$

which is the same as the eigenvalue equation of the matrix A, with λ typically replacing *s* when discussed in mathematical texts.

The roots, or eigenvalues, of the equation depend on the values of the elements of A. If the equation is written in the form

$$\lambda^2 + B\lambda + C = 0$$

then the two eigenvalues are given by

$$\lambda_1 = -B + (B^2 - 4C)^{1/2}/2 \text{ and } \lambda_2 = -B - (B^2 - 4C)^{1/2}/2$$

The solution to the differential equation is of the form

$$\begin{aligned} x_1(t) &= K_1 e^{\lambda_1 t} + K_2 e^{\lambda_2 t} \text{ for } \lambda_1 \neq \lambda_2 \\ x_1(t) &= K_1 e^{\lambda_1 t} + tK_2 e^{\lambda_2 t} \text{ for } \lambda_1 = \lambda_2 \end{aligned}$$

For the linear system there is only one singular point at the origin $x_1 = 0, x_2 = 0$ of the state plane and the behaviour of the motion near to the singular point depends on the eigenvalues.