Understanding Complex Systems

Springer : COMPLEXITY

Federico Beffa

Weakly Nonlinear Systems

With Applications in Communications Systems

OPEN ACCESS

Springer

Springer Complexity

Springer Complexity is an interdisciplinary program publishing the best research and academic-level teaching on both fundamental and applied aspects of complex systems—cutting across all traditional disciplines of the natural and life sciences, engineering, economics, medicine, neuroscience, social and computer science.

Complex Systems are systems that comprise many interacting parts with the ability to generate a new quality of macroscopic collective behavior the manifestations of which are the spontaneous formation of distinctive temporal, spatial or functional structures. Models of such systems can be successfully mapped onto quite diverse "real-life" situations like the climate, the coherent emission of light from lasers, chemical reaction-diffusion systems, biological cellular networks, the dynamics of stock markets and of the internet, earthquake statistics and prediction, freeway traffic, the human brain, or the formation of opinions in social systems, to name just some of the popular applications.

Although their scope and methodologies overlap somewhat, one can distinguish the following main concepts and tools: self-organization, nonlinear dynamics, synergetics, turbulence, dynamical systems, catastrophes, instabilities, stochastic processes, chaos, graphs and networks, cellular automata, adaptive systems, genetic algorithms and computational intelligence.

The three major book publication platforms of the Springer Complexity program are the monograph series "Understanding Complex Systems" focusing on the various applications of complexity, the "Springer Series in Synergetics", which is devoted to the quantitative theoretical and methodological foundations, and the "Springer Briefs in Complexity" which are concise and topical working reports, case studies, surveys, essays and lecture notes of relevance to the field. In addition to the books in these two core series, the program also incorporates individual titles ranging from textbooks to major reference works.

Indexed by SCOPUS, INSPEC, zbMATH, SCImago.

Series Editors

Understanding Complex Systems

Federico Beffa

Weakly Nonlinear Systems

With Applications in Communications Systems

Springer

Federico Beffa
Federico Beffa Engineering
Alto Malcantone, Switzerland

ISSN 1860-0832 ISSN 1860-0840 (electronic)
Understanding Complex Systems
ISBN 978-3-031-40680-5 ISBN 978-3-031-40681-2 (eBook)
https://doi.org/10.1007/978-3-031-40681-2

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Paper in this product is recyclable.

There is nothing more practical than a good theory.

Foreword

The field of circuit design for communications is a rapidly evolving area of research, with new technologies and applications emerging all the time. One of the challenges in this field is the need to accurately model and analyze the behavior of nonlinear systems. This book offers an approach to this problem, that is both more rigorous and intuitive than previously published.

The book is organized into two main parts: signals and systems. The first part introduces the theory of distributions and covers topics such as basic properties of distributions, convolution, Fourier and Laplace transforms, and summable distributions. The second part focuses on the application of distributions in the study of convolution equations and their solutions.

In contrast to the few existing treatments, the approach taken highlights the algebraic structure underlying weakly nonlinear systems and is based on distributions, rather than functions. The use of distributions leads naturally to the convolution algebras of Linear Time-Invariant (LTI) systems and the ones suitable for weakly nonlinear systems emerges as simple extensions to higher order distributions, without having to resort to ad hoc operators. The main advantages of the approach include a new justification for the validity of the Volterra series; with a much-simplified notation, free of multiple integrals. The net result being a conceptual simplification and the ability to solve the associated nonlinear differential equations in a purely algebraic way.

Throughout the book, the author provides clear explanations of the key concepts and techniques, with numerous examples drawn from the area of circuits for wireless communications. As well as being of practical use to practitioners, these should help the reader to gain a deeper understanding of the material covered in the earlier part of the book. Of particular interest, to those interested in modern high-frequency circuit design, analysis of the classic amplifier cascode (common gate) shows the origin of nonlinear phenomenon, such as intermodulation, that can be accurately quantified from very simple models of the underlying transistors without having to resort to simulation. A similar analysis of local feedback in common source amplifiers (degeneration) shows similar rich behavior. As well as being theoretically interesting, this provides practical methods to the optimal design of such circuits.

This book is primarily intended for graduate students in engineering, who are interested in the theory of nonlinear systems and its application. However, it is likely that researchers in mathematics and physics will also find the material useful.

I'm sure that this book will serve as a valuable resource for anyone working with nonlinear dynamical systems who wish to learn more about this interesting and relevant topic.

Surrey, England Jon Strange
April 2023

Preface

When I started working in industry I was immediately confronted with systems whose performance was invariably limited by noise and nonlinearities. For noise, there is a well-developed theory that can be used to guide the design and suggest ways to improve the system. For distortion, I was not aware of good theories and the investigations and developments were done entirely by numerical simulations. It's not that I didn't have interest in nonlinear systems. But rather than that, despite the fact that I graduated from a good university, the only courses offered on nonlinear systems were in the field of control theory and almost entirely devoted to stability questions. No course taught widely applicable methods to calculate the response of nonlinear systems.

While working on trying to improve the linearity of a system, I read a paper making use of nonlinear transfer functions. Although I had heard the name "Volterra Series", I did not know what it was, nor did anyone I knew. Stimulated by that paper I looked for a book, but didn't find any in print. So, I looked in the second-hand market and found an out-of-print book mentioned in the paper. The book was very focused on practical applications, and didn't say much about the broader theory. In any case, having learnt how to use nonlinear transfer functions, I started using them, and they immediately illuminated the reasons for some effects that I saw in simulations and that I didn't fully grasp. Since then, I kept using that method very fruitfully. I also used nonlinear transfer functions in design reviews to try to pass on to colleagues the intuition that it gave me about nonlinear effects.

After some time, some colleagues started to see the usefulness, and I was asked to prepare an internal tutorial on the subject. In preparing it, I realized how limited my theoretical understanding on the subject was and developed a desire for a much deeper understanding: I was hooked. I started studying it more by myself as well as through other (old) books, reports, and papers. While all studies that I saw did introduce some ad-hoc operators, I tried to develop a formulation using standard ones. When I realized that pairing convolution with tensor product would do, it became self-evident that the Volterra series could be seen as a generalization of the Taylor series. Convolution took front and center, but to make the mathematics solid, I had to resort to Schwartz's distributions. Differential equations became convolution equations and the Volterra series became a generalized formal power series.

This book is a summary of my investigations in which I tried to develop the theory in a new and, I believe, simpler form. I included many examples. Some of them are short and serve to illustrate some points just discussed. Others are (condensed versions of) real-world applications where I try to illustrate the power of the theory.

The book was written primarily for engineers. As is common in electrical engineering, I use the symbol j to denote the imaginary unit of complex numbers. This is to avoid confusing it with currents which are commonly denoted by i. All plots in the book were generated with the open-source CAS system `maxima` [1] which I also used to check many calculations and to compute all numerical solutions of differential equations. All numerical simulations of nonlinear networks were performed with the open-source circuit simulator `Xyce` [2].

I would like to take the opportunity to thank some people that in a direct or indirect way have contributed to this book. First of all, I would like to thank my wife Alessandra for her constant encouragement and support during this project. I would also like to thank many colleagues at Analog Devices and Mediatek who shared their insights with me and stimulated me to go deeper. In particular, I would like to thank Jon Strange who, among other things, did put a lot of trust in me and involved me in many stimulating and future-looking projects. They were the stimulus that ultimately lead me to write this book.

Alto Malcantone, Switzerland Federico Beffa
February 2023

Contents

Part I Signals

Part II Systems

Chapter 1
Introduction

Nonlinear systems are everywhere, yet most engineering curricula devote very little time, if any, to them. The reason is twofold: first of all, there is no general theory describing arbitrary nonlinear systems. Second, the theory of linear systems is effective in facilitating the design of many real-world systems. In fact, for sufficiently small input signals, the behaviour of most nonlinear systems can be approximated by a linear model. As a consequence the majority of engineered systems are designed based on linear system theory and their usability is limited in one way or another by the deviation of the real system from the assumed linear behaviour. This book is intended to give engineers a powerful tool to model, understand and reduce the impact of mild deviations from linear behaviour and thereby design better systems.

This chapter tries to develop some intuition for what we call weakly-nonlinear systems. It also tries to give an idea of the theory to which this book is devoted and to its applicability. The chapter is not meant to introduce in a precise way any concept. In fact the exposition is rather informal. A proper systematic development of the theory will start with the next chapter.

1.1 Nonlinear Phenomena

The range of phenomena exhibited by nonlinear systems is much richer than the one of linear systems. To understand the applicability of the presented theory it's useful to have an idea of the main ones that may appear. In the following we give a bird's-eye view of them with qualitative descriptions.

1.1.1 Multiple Equilibrium Points

Most dynamical systems can be described by a system of differential equations that can be written in the form

F. Beffa, *Weakly Nonlinear Systems*, Understanding Complex Systems,
https://doi.org/10.1007/978-3-031-40681-2_1

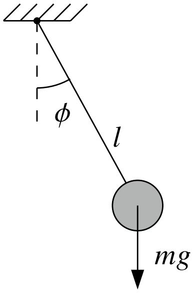


Fig. 1.1 Pendumum in Earth's gravitational field

$$\frac{\mathrm{d}}{\mathrm{d}t}u = f(u, x, t)\,,$$

with $u \in \mathbb{R}^n$ the state of the system and $x \in \mathbb{R}^m$ the driving or input signal. For simplicity in this chapter we limit ourselves to *autonomous* systems. These are systems described by the simpler equation

$$\frac{\mathrm{d}}{\mathrm{d}t}u = f(u)\,. \tag{1.1}$$

In the case of first and second order systems one can obtain a good qualitative understanding by examining the *phase portrait* of the system. This is a graphical representation of a family of state trajectories $t \mapsto u(t)$ for various initial conditions u_0 in the plane spanned by the components u_1 and u_2 of u that in this context is called the *state* or *phase space*. Note that the phase portrait can be sketched without having to solve the equation by considering the vector field defined by f in the state plane. With it, it's easy to estimate the trajectory of the state u for every initial condition u_0.

Of special interest are the zeros of the vector function f. That's because in those states the derivative with respect to time of the state vector u vanishes. In other words, the zeros of f are the equilibrium points of the system. *A nonlinear function f in general has several equilibrium points* and that's a first fundamental difference from linear systems which always have only one equilibrium point.

If f is well-behaved,[1] for every initial state u_0, the system (1.1) has a unique solution. This means that the trajectories in the phase plane do not intersect. Therefore, the trajectories can only begin or end at equilibrium points, at infinity or on limit cycles (see below).

As an example consider the ideal friction-less pendulum shown in Fig. 1.1 and described by the differential equation

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\phi + \omega_0^2 \sin(\phi) = 0\,, \qquad \omega_0 = \sqrt{\frac{g}{l}}$$

[1] We will make this statement precise in a later chapter.

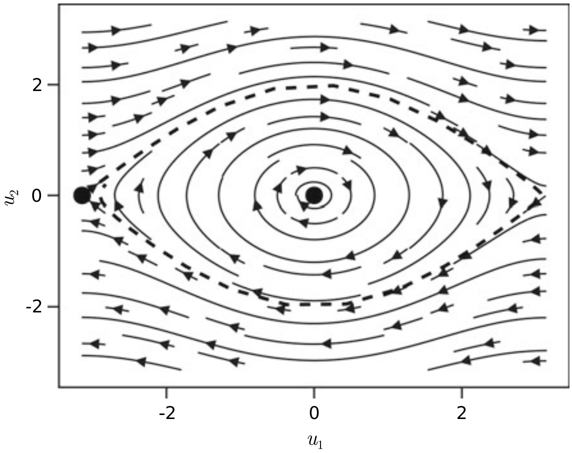


Fig. 1.2 Phase portrait of an ideal pendulum with $\omega_0 = 1$

with ϕ the angle from the vertical, g the gravitational acceleration and l the length of the arm. The equation can be rewritten as

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} u_2 \\ -\omega_0^2 \sin(u_1) \end{pmatrix}$$

where we have set $u_1 = \phi$, $u_2 = \mathrm{d}\phi/\mathrm{d}t$. The phase portrait of this system is clearly periodic along the u_1 axis. We can therefore limit the study to the range $u_1 = [-\pi, \pi)$.[2] In this range the system has two equilibrium points: $u_{0a} = (0, 0)$ and $u_{0b} = (\pi, 0)$.

The phase portrait of this system is depicted in Fig. 1.2 with the equilibrium points shown as black dots. The dashed lines connect the two equilibrium points and separates the phase plane in two distinct regions in which the system has different behaviour. The boundary between the two regions (the surface constituted by the dashed lines) is called the *separatrix*. Trajectories surrounding the equilibrium point u_{0a} are closed curves that represent oscillations. The trajectories above and below the separatrix represent the pendulum perpetually rotating around the pivot.

This shows a second fundamental difference from linear systems. *Nonlinear systems can exhibit different behaviour and characteristics in different regions of the phase space.*

[2] Given that physically the angle ϕ and $\phi + 2\pi$ describe the same location, we should think of the phase space as a cylinder rather than a plane.

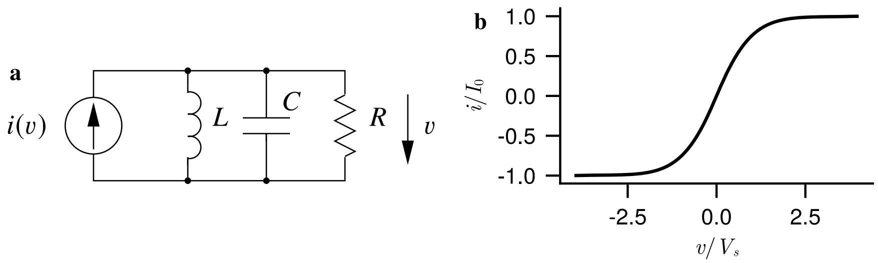


Fig. 1.3 **a** Electrical RLC oscillator with nonlinear feedback **b** Voltage-controlled current-source characteristic

1.1.2 Limit Cycles

A further phenomenon present in some nonlinear systems that doesn't exist in linear ones is that of the *limit cycles*. These are periodic solutions of the equations at specific signal levels. As a simple example, consider the oscillator shown in Fig. 1.3a. It consists of a passive RLC resonator and a nonlinear saturating voltage-controlled current source (VCCS) with characteristic

$$i(v) = I_0 \tanh\left(\frac{v}{V_s}\right)$$

and plotted in Fig. 1.3b. The system is described by

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}v + \frac{\omega_o}{q}\Big[1 - G_m(v)R\Big]\frac{\mathrm{d}}{\mathrm{d}t}v + \omega_0^2 v = 0. \tag{1.2}$$

with

$$\omega_0 = \frac{1}{\sqrt{LC}} \qquad\qquad q = \frac{R}{\omega_0 L}$$

and the nonlinear transconductance

$$G_m(v) = \frac{\mathrm{d}}{\mathrm{d}v}i(v) = \frac{I_0}{V_s}\mathrm{sech}^2\left(\frac{v}{V_s}\right).$$

If the maximum of $|v(t)|$ over a full period $\mathcal{T} = 2\pi/\omega_0$ remains small compared to V_s then the value of $\mathrm{sech}(v(t)/V_s)$ remains very nearly 1 over a full cycle. Therefore, under this assumption, if $G_m(0)R > 1$ the coefficient of the first order derivative of v in (1.2) is negative and the $(0, 0)$ equilibrium point of the equation is unstable. Differently from this, if the maximum of $|v(t)|$ over a period is much larger than V_s then the value of $\mathrm{sech}(v(t)/V_s)$ approaches 0 for most part of a period. Hence, in this regime of operation the system is governed by an equation corresponding to

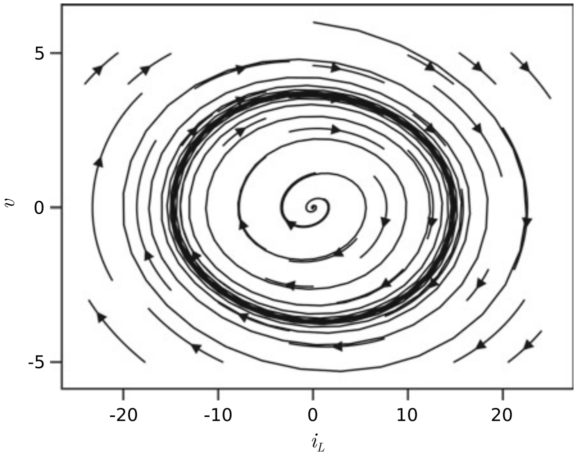


Fig. 1.4 Phase portrait of an electrical oscillator. $\omega_0 = 1$ rad/s, $q = 4$, $R = 1\ \Omega$, $V_s = 1$ V, $I_0 = 3V_s/R$

the one of a damped oscillator. Between these two extreme cases there is a periodic trajectory, a limit cycle. On this trajectory the energy dissipated during one cycle by the resistor R is perfectly balanced by the energy injected in the resonator by the controlled source. This behaviour of the system is clearly discernible in the phase portrait shown in Fig. 1.4 in which we chose the current flowing through the inductor (downwards) i_L and v as state variables.

This example has a stable limit cycle, but there are systems with unstable limit cycles: any infinitesimally small deviation from the perfectly periodic trajectory leads to a trajectory diverging from the limit cycle. Limit cycles can also be stable on one side and unstable on the other one.

1.1.3 Bifurcations

All practical systems depend upon some parameters. For example the oscillator of the previous section depends on the value of the resistor R, and it is interesting to study how the value of that parameter affects the behaviour of the system. In particular the number and type of equilibrium points of a system may depend on the value of some parameter. This is in fact the case for our oscillator: For $G_m(0)R < 1$ the system has a single stable equilibrium point, while for $G_m(0)R > 1$ that equilibrium point becomes unstable and a limit cycle makes its appearance. Parameter values at which the character of the system behaviour changes are called *critical* or *bifurcation points.*

As a second example, consider a system described by the differential equation

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}u = -\lambda u + u^3.$$

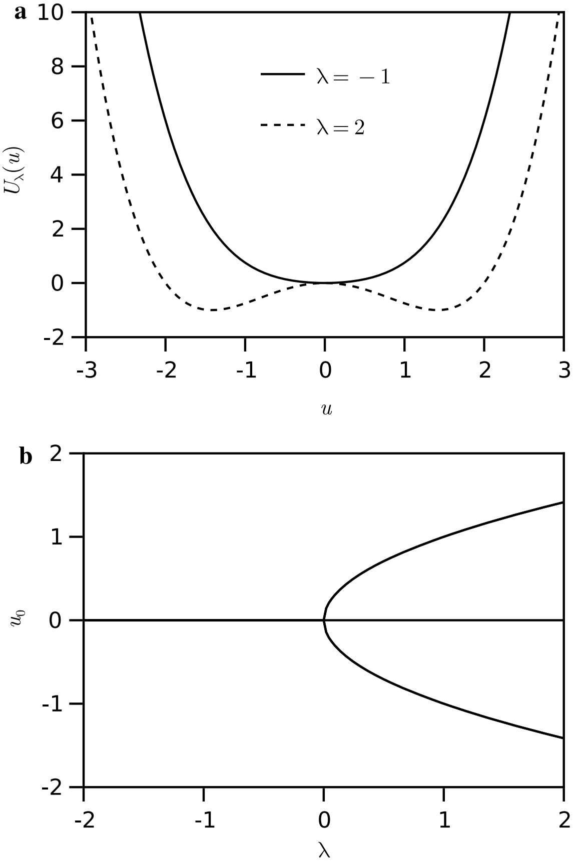


Fig. 1.5 **a** Pitchfork bifurcation potential **b** Pitchfork bifurcation equilibrium points

The system can be interpreted as having a potential energy

$$U_\lambda(u) = \frac{u^4}{4} - \lambda \frac{u^2}{2}.$$

For $\lambda < 0$ the potential energy has a single minimum at $u = 0$, while for $\lambda > 0$ it has two minima as shown in Fig. 1.5a. In the latter case $u = 0$ is an unstable equilibrium point and two new stable equilibrium points at $\pm\sqrt{\lambda}$ do appear. If we draw the equilibrium points of the system as a function of λ one obtain the so-called *pitchfork* shown in Fig. 1.5b.

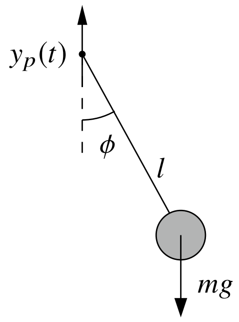


Fig. 1.6 Driven pendumum in Earth's gravitational field

1.1.4 Chaos

Consider the driven pendulum shown in Fig. 1.6. It is similar to the one of Sect. 1.1.1, with the difference that now the pivot moves in time in the vertical direction as described by the function y_p. This movement introduces a driving term in the differential equation that then becomes

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\phi + \omega_0^2 \sin\phi = -\frac{\sin\phi}{l}\frac{\mathrm{d}^2}{\mathrm{d}t^2}y_p(t).$$

Lets assume that the drive is periodic $y_p(t) = A\cos(\omega t)$. Figure 1.7 shows the time evolution of ϕ for two almost identical initial conditions. The upper curve was computed with the pendulum starting with $\frac{\mathrm{d}}{\mathrm{d}t}\phi(0) = 0\,\mathrm{rad/s}$ and at an angle of $\phi(0) = 1\,\mathrm{rad}$. The lower curve was computed with almost identical initial conditions $\frac{\mathrm{d}}{\mathrm{d}t}\phi(0) = 0\,\mathrm{rad/s}$ and $\phi(0) = 1 + 10^{-10}/l\,\mathrm{rad}$. The lower curve was thus started with a displacement corresponding to approximately an atom diameter from the upper one. The evolution of the two is initially very similar. However, after some time they become completely different and uncorrelated. This extreme sensitivity to initial conditions is the characteristic defining *chaotic systems* and makes long term predictions essentially impossible. In those systems the initial difference between adjacent trajectories grows on average exponentially [3].

In this simple case the phenomenon is intuitively understandable. When the pendulum reaches a position very close to the vertical, an infinitesimal difference in velocity can determine if it makes a full turn or if it goes back.

Note that if the initial oscillation is sufficiently small, the force exercised by the vertical drive is almost orthogonal to the direction in which the mass is free to move. For this reason, small oscillations are not pushed to large swings and do not show chaotic behaviour. There are therefore regions of the phase space exhibiting chaotic behaviour and regions not exhibiting it. The areas of these regions depend of course on the amplitude A of the drive. Small values of A lead to large areas in which the system behaves predictably and only small areas displaying chaotic behaviour.

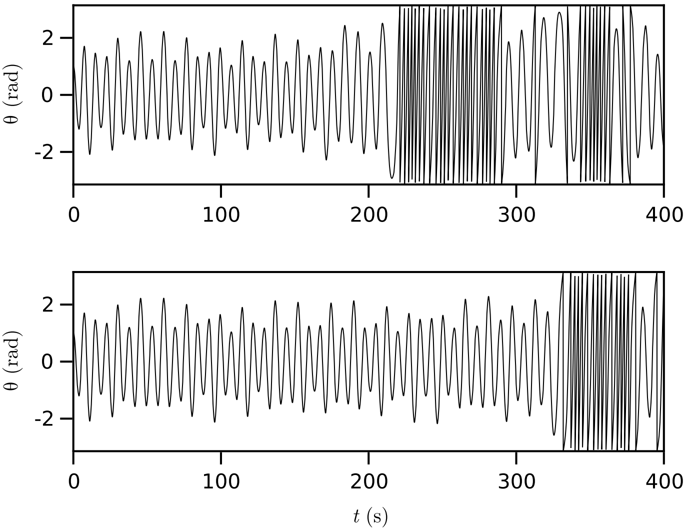


Fig. 1.7 Time evolution of the driven pendulum with $\omega_0 = 1$ rad/s, $\omega = 2\omega_0$, $g = 9,8$ m/s^2, $l = 9.8$ m, $A = l/10$. The upper curve was computed with initial conditions $\phi(0) = 1$ rad, $\frac{\mathrm{d}}{\mathrm{d}t}\phi(0) = 0$ rad/s, the lower one with initial conditions $\phi(0) = 1 + 10^{-10}/l$ rad, $\frac{\mathrm{d}}{\mathrm{d}t}\phi(0) = 0$ rad/s

1.2 Weakly-Nonlinear Systems

Chaos, bifurcations and other phenomena of nonlinear systems are fascinating, important and sometimes fundamental to the problem at hand. However, the vast majority of engineered system operate around stable equilibrium points by design. From an engineering point of view a quantitative theory to study the behaviour of nonlinear systems in the proximity of stable equilibrium points is therefore very important.

Inspection of the presented phase portraits suggest that in the neighbourhood of equilibrium points the behaviour of nonlinear systems is not too different from the one of linear systems, the deviation increasing with increasing distance of the state from the equilibrium points. In fact this statement can be made more precise. Consider a time invariant, single-input single-output (SISO) system with input x whose state dynamics is governed by the system of first order differential equations

$$\frac{\mathrm{d}}{\mathrm{d}t}u(t) = f(u(t), x(t)), \qquad f : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$$

and its output y by the algebraic equation

$$y(t) = g(u(t), x(t)), \qquad g : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}.$$

If around the equilibrium point $u_0 = 0$[3] and $x = 0$ the functions f and g are differentiable, then, using a Taylor expansion, the system behaviour can be approximated by the linear equations

$$\frac{\mathrm{d}}{\mathrm{d}t}u(t) \approx Au(t) + Bx(t) \qquad A \in \mathbb{R}^{n\times n}\,, \quad B \in \mathbb{R}^{n\times 1}$$

and

$$y(t) \approx Cu(t) + Dx(t)\,, \qquad C \in \mathbb{R}^{1\times n}\,, \quad D \in \mathbb{R}\,.$$

The response of the system to the input signal can then be expressed by a convolution integral between the impulse response h of the system and the input signal

$$y(t) = h(\tau) * x(t)\,. \tag{1.3}$$

Note that in this chapter by stable equilibrium point we mean one for which all eigenvalues of the linearized state equation are negative.

The linear systems theory is very useful. However, for many practical applications this idealisation is too crude and doesn't capture effects that limit the usability of a vast array of systems. The theory presented in this book enables one to solve the system equations when f and g are approximated by a higher order polynomial or even by power series. The theory therefore is able to give a more faithful description of the behaviour of many real systems. In particular, it allows probing into effects outside the reach of linear systems theory.

Consider first a *memory-less* system, that is, a system whose output $y(t)$ at time t depends only on the value of its input signal $x(t)$ at time t and not on any of its past (or future) values. Such a system can be represented by a function g mapping for every value of t the value $x(t)$ to $y(t)$

$$y(t) = g(x(t))\,.$$

Let's assume that for a zero input signal the output is zero and that g can be expanded in a Taylor series. Then we can write

$$y(t) = \sum_{k=1}^{\infty} g_k x^k(t)\,.$$

Using the Dirac δ distribution this expression can be written in the different form

$$y(t) = \sum_{k=1}^{\infty} g_k\, \delta(\tau_1, \ldots, \tau_k) * x^{\otimes k}(t)\,.$$

[3] By a change of variables it's always possible to move the equilibrium point of interest to the origin.

That's because, under assumptions to be made precise later, the δ distribution is the unit of the convolution product

$$\delta(\tau) * x(t) = x(t)\,.$$

The response of a *linear* memory-less system can therefore be written as

$$y(t) = g_1\delta(\tau) * x(t)$$

which shows a striking similarity with (1.3), the response of a linear dynamical system with impulse response h. In fact $g_1\delta$ is the impulse response of a linear memory-less system, and it vanishes everywhere except at the origin as expected.

From these considerations it's natural to hypothesise that the response of a class of nonlinear systems, around a stable equilibrium point, can be represented by a series of the form

$$y(t) = \sum_{k=1}^{\infty} h_k(\tau_1, \ldots, \tau_k) * x^{\otimes k}(t)\,.$$

This is in fact true, and it is the *Volterra series* representation of the system with h_k its kth order impulse response. This representation is valid only for sufficiently small input signals not pushing the state of the system beyond a separatrix. This limitation of the Volterra series should not surprise. In fact power series, which are a subset of the Volterra series, in general also have a finite convergence radius.

The similarity between power series and the Volterra series doesn't end here. By introducing suitable definitions, we can represent the cascade of nonlinear systems represented by their respective Volterra series in a similar way as the composition of power series.

We call systems that can be represented by a Volterra series *weakly-nonlinear* systems. A feature of weakly-nonlinear systems shared with linear ones is the fact that the differential equations describing the system have to be solved only once to obtain the impulse responses. The response of the system to a large set of different input signals can then be computed directly from them. The impulse responses therefore completely characterise weakly-nonlinear systems. As with linear systems, weakly-nonlinear ones have a frequency domain representation in terms of nonlinear transfer functions.

The theory can be extended to cover time-varying systems. In this case the impulse responses (or nonlinear transfer functions) have an explicit dependence on time

$$h_k(t, \tau_1, \ldots, \tau_k)\,.$$

1.3 Distributions

The Dirac δ distribution plays a key role in highlighting the relationship between power and Volterra series. An ad-hoc use of the δ distribution however easily leads to problems.

Consider for example the *Heaviside unit step* function (or *unit step* function)

$$\mathsf{1}_+(t) := \begin{cases} 0 & t < 0 \\ 1 & t \geq 0 \end{cases} \tag{1.4}$$

and the theorem stating that the Laplace transform of the derivative of a function f continuous for $t > 0$ is

$$sF(s) - f(0+)$$

with F the Laplace transform of f and $f(0+)$ the right-hand side limit to 0 of the function. A careless application of this theorem to $\mathsf{1}_+$ gives

$$\mathcal{L}\left\{\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{1}_+\right\} = s\frac{1}{s} - 1 = 0$$

where we have used the fact that $\mathcal{L}\{\mathsf{1}_+\} = 1/s$. However, we will show that the derivative of $\mathsf{1}_+$ is the δ impulse whose Laplace transform is 1. The error lies in the fact that δ is not a function, but rather a Schwartz's distribution, or *distribution* for short. The above theorem, in the stated form, is therefore not applicable.

Distributions are the proper setting for studying linear and weakly-nonlinear systems. In this setting the convolution product comes to play a central role. In fact, distributions allow defining *convolution algebras* with δ playing the role of the unit. The Laplace transform then not only maps convolution products into multiplications, but it also maps the unit of the convolution algebra into the unit of multiplication. In addition, in this setting, the derivative of a distribution f can be represented as the convolution of the distribution with the derivative of the unit

$$\frac{\mathrm{d}}{\mathrm{d}t}f = \frac{\mathrm{d}}{\mathrm{d}t}\delta * f\,.$$

Differential equations can therefore be transformed into convolution equations to obtain a complete time-domain mirror image of the Laplace domain algebraic equations. Distributions enable a uniform representation in terms of convolution products of ubiquitous and embarrassingly simple linear systems such as inductors, which a theory based on functions is unable to do

$$v = L\frac{\mathrm{d}}{\mathrm{d}t}\delta * i\,.$$

Here we see the current i as the input and the voltage v as the output of the system.

While we have been implicitly assuming causal systems and signals vanishing for $t < 0$, there are other convolution algebras. One of them is the convolution algebra of periodic distributions intimately related to the Fourier series, where the δ distribution plays a central role as well.

1.4 Numerical Simulations

Learning a theory requires some investment of time. The question is: is it worth it in a world full of computers and where numerical methods able to solve most nonlinear equations are readily available? In our view the answer is definitely a resounding yes. Numerical simulations and theory are not in competition, but rather they complement each other.

The theory is able to reveal the origin of the various effects at play and to clarify how each parameter affects the performance of a system. However, to do so we must use relatively simple models and therefore, most of the time, obtain approximate results.

Simulations on the other hand can be used to obtain accurate answers taking into account all details. However, the results are presented as tables of numbers (or curves) valid only for a specific set of values of the parameters. We can of course run many simulations and sweep parameters, but complex simulations are not fast and this poses practical limits. In addition, inferring the relationship between parameters and a specific effect from simulation results only is often challenging. We can say that a good theory based engineering model is like a (slightly distorted) picture, while numerical simulations are like dots of a halftone image. A good model is worth thousands of simulations.

Most of the time the difficulty in engineering problems lies in finding the simplest model able to correctly characterise the effects of interest. During the phase of model development, numerical simulators can be extremely useful by using them as ideal laboratories in which to validate hypotheses. In these virtual laboratories it's easy to change the laws of physics and suppress or decouple phenomena in a way that's impossible in the real world. Experiments conducted in these virtual laboratories can therefore be an invaluable guide in the development of a model. Once a good model has been found, it will rapidly guide the development of the system. Numerical simulations will then serve further for final tuning and verification.

1.5 Historical Notes

Around 1887 Vito Volterra developed the concept of *functionals* as an extension of functions of multiple variables to ones with an uncountable infinite number [4, 5]. Let $f(x_1, \dots, x_k)$ be a real valued function of the k real variables x_1 to x_k. The latter can be interpreted as the values of a function $x_.$ evaluated at the discrete points 1 to

k. He therefore conceived a functional as a function of another function x defined over a continuous finite interval. He then proposed the series now bearing his name as an extension of the Taylor series from functions to functionals.

In 1910 the mathematician M. Fréchet published a more detailed analysis of the conditions under which a functional can be expanded into a Volterra series [6]. This work is regarded by most as the foundation demonstrating the validity of Volterra's series expansion.

Volterra was well known and was invited to present his works in several countries, including the United States. During the second World War there was great pressure to develop anti-aircraft systems and N. Wiener of the Massachusetts Institute of Technology (MIT) found that by using Volterra's series he could analyse the response of a nonlinear device to noise [7]. His report was initially restricted. Its release after the war sparked interest in the engineering community at MIT and elsewhere. Several studies followed in the 50s to the 70s applying Volterra's series to nonlinear engineering problems, with [8–10] among the most significant ones. Wiener himself remained interested in the subject and developed his own variant of the theory based on Browning motion and leading to what's now called the *Wiener series* [11]. At the beginning of the 80s some books summarised the Volterra and Wiener theories [12, 13]. At around the same time, with the raise of desktop computers, engineering efforts started more and more to embrace numerical methods.

During the first decades of the 20th century there were two mathematical methods used by engineers and physicists that kept mathematicians occupied. The first was to find a solid mathematical justification for the *operational calculus* popularised by O. Heaviside. The second was the search for a solid mathematical interpretation for the δ distribution and its derivatives extensively used by P. A. M. Dirac in his landmark treatise on quantum mechanics which first appeared in 1930 [14].

The former was solved in two ways:

(i) With the help of the Laplace- and Fourier-transforms, highlighting the frequency domain aspects, and
(ii) by Mikusinski [15] using purely algebraic methods.

The second was solved by L. Schwartz by introducing new mathematical objects called *distributions* [16]. These are a special class of functionals with particularly attractive properties such at the fact that they are indefinitely differentiable. In addition, differentiation of distributions is a linear operation making series always differentiable term by term. With distributions Schwartz not only did put the δ "function" and its derivatives on a solid ground, but he also introduced convolution algebras and unified the two justifications for the operational calculus.

A deep understanding of distributions requires familiarity with advanced concepts of topological vector spaces [16, 17], which is probably why they are rarely introduced to engineers. However, the elementary part of the theory can be developed without recurring to particularly deep mathematical concepts and is of great practical value in physics and engineering problems. The aim of this book is to introduce distributions to engineers and use them to view the Volterra series and, more generally, weakly-nonlinear systems from a point of view different from the traditional one. The advantages are, among others, a conceptual simplification, a simpler notation freeing expressions from multiple integrals and an exposition of the theory of weakly-nonlinear systems as a natural extension of the linear one.

Part I
Signals

Chapter 2
Distributions

We investigate the mathematical description of signals that are commonly used in the analysis of technical problems. From a mathematical point of view it would be useful to limit the signals of interest to the set of continuous functions. However, this has several disadvantages. For example, suppose we are interested in the transient response of, say, a series RC low-pass-filter (LPF) to an input step voltage. If we describe the input signal with a continuous function, then the details of the calculations depend on the chosen description of the input signal rise transient. This however tends to mask the fact that, if the LPF time constant is much larger than the input signal rise-time, the output response is essentially independent of the input signal transient shape. For this reason, in these situations, it is much more convenient to use an *idealized* input unit step function such as the Heaviside unit step function

$$\mathsf{1}_+(t) = \begin{cases} 0 & t < 0 \\ 1 & t \geq 0 \end{cases}$$

which is not continuous at $t = 0$.

Consider further the LPF example. If we write the differential equation using the current as unknown, then we need the derivative of the driving signal. However, the derivative of the function $\mathsf{1}_+$ does not exist at $t = 0$ and is zero at every other point. We are therefore led to introduce the so-called Dirac impulse δ which however is not even a function, but rather a *generalized function* or *distribution*.

It follows from these considerations that a correct description of commonly used signals belongs to the theory of *distributions*. Distributions have many useful properties. The key one being that they can be differentiated any number of times. The main contributor to the development of this theory was Schwartz [16].

F. Beffa, *Weakly Nonlinear Systems*, Understanding Complex Systems,
https://doi.org/10.1007/978-3-031-40681-2_2

2.1 Test Functions

The key idea in the theory of distributions, is not to direct attention to the value of a function at every point of its domain, but instead to "measure" the behavior of a function when acting on a class of particularly well-behaved functions. In this section we introduce one such class of functions, the class of *test functions*.

Let $k = (k_1, \ldots, k_n)$ be an n-tuple of non-negative integers called a *multi-index*. The *differential operator* of order $|k|$ is defined by

$$D^k := D_1^{k_1} \cdot \cdots \cdot D_n^{k_n}\,, \quad D_i := \frac{\partial}{\partial \tau_i} \tag{2.1}$$

with

$$|k| := k_1 + \cdots + k_n \tag{2.2}$$

the length of the multi-index k and $\tau \in \mathbb{R}^n$. For functions of a single variable we also use the following shorter notation for the kth order derivative

$$f^{(k)} := \frac{\mathrm{d}^k f}{\mathrm{d}\tau^k} \tag{2.3}$$

where in this case k is of course a single non-negative integer.

Given an open set $U \subset \mathbb{R}^n$, the set of all k-times continuously differentiable functions $f : U \to \mathbb{C}$ is denoted by $C^k(U)$ or simply C^k.

Definition 2.1 (*Test function*) A function $\phi : \mathbb{R}^n \to \mathbb{C}$ is called a *test function* if it is indefinitely differentiable and has compact support, that is, if $\phi \in C^\infty$ and $\phi(\tau_1, \ldots, \tau_n) = 0$ outside a compact set K. The vector space of all such functions is denoted by $\mathcal{D}$.

To define a continuity criterion for distributions we need to define a topology that can be encoded in the form of a *convergence principle* in $\mathcal{D}$.

Definition 2.2 (*Convergence of test functions*) A sequence of functions $\phi_m \in \mathcal{D}$, $m \in \mathbb{N}$ is said to converge to $\phi \in \mathcal{D}$, in symbols

$$\phi_m \underset{\mathcal{D}}{\longrightarrow} \phi \quad \text{or} \quad \lim_{m\to\infty} \phi_m = \phi\,,$$

if the following two conditions are met:

1. There exist a compact set K such that it includes the support of all ϕ_m and of ϕ.
2. For every n-tuple k the sequence of functions $D^k\phi_m$ converges uniformly toward $D^k\phi$.

These conditions ensure that the limiting function (i) has compact support and (ii) that it is indefinitely differentiable, in other words, that the limiting function is also a test function.

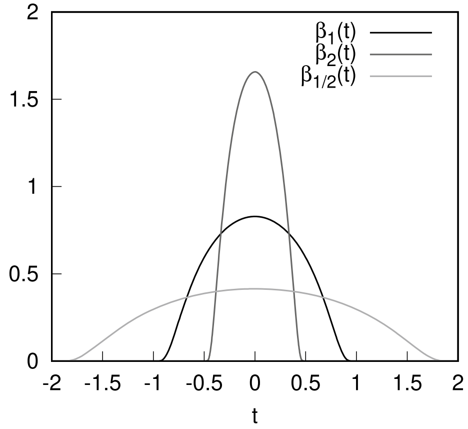


Fig. 2.1 Example test function

Example 2.1: Test function

Consider the following functions (see Fig. 2.1)

$$\beta_\nu(t) := \begin{cases} \frac{\nu}{B}\mathrm{e}^{\frac{-1}{1-(\nu t)^2}} & \text{for } |\nu t| < 1 \\ 0 & \text{for } |\nu t| \geq 1 \end{cases} \tag{2.4}$$

$$B := \int_{-1}^{1} \mathrm{e}^{\frac{-1}{1-t^2}}\,\mathrm{d}t$$

For each value of $\nu > 0$ and for $|\nu t| < 1$ the function β_ν is the composition of a rational function with no singularities and the exponential function. Since the latter two functions are indefinitely differentiable and the composition of indefinitely differentiable functions is also indefinitely differentiable, it follows that, in this range, β_ν is indefinitely differentiable.

To establish that β_ν is a test function we have further to show that

$$\lim_{|\nu t|\uparrow 1} D^k\beta_\nu(t) = 0 \tag{2.5}$$

for all values of k. This can be done by induction: assume that the kth order derivative is the product of β_ν and a polynomial in the two variables $\tau_1 = 1/(1-\nu t)$ and $\tau_2 = 1/(1+\nu t)$

$$D^k\beta_\nu(t) = p_k\left(\frac{1}{1-\nu t}, \frac{1}{1+\nu t}\right)\beta_\nu(t)\ . \tag{2.6}$$

This is clearly the case for $k = 0$. We show that this is then true for $k+1$:

$$\begin{aligned}
&D^{k+1}\beta_\nu(t)\\
&= \quad \Bigg[-\frac{\nu\,(D_2 p_k)\left(\frac{1}{1-\nu t},\frac{1}{1+\nu t}\right)}{(\nu\, t+1)^2}+\frac{\nu\, p_k\left(\frac{1}{1-\nu t},\frac{1}{1+\nu t}\right)}{2\;(\nu\, t+1)^2}\\
&\qquad +\frac{\nu\,(D_1 p_k)\left(\frac{1}{1-\nu t},\frac{1}{1+\nu t}\right)}{(\nu\, t-1)^2}-\frac{\nu\, p_k\left(\frac{1}{1-\nu t},\frac{1}{1+\nu t}\right)}{2\;(\nu\, t-1)^2}\Bigg]\,\beta_\nu(t)\\
&=: \quad p_{k+1}\left(\frac{1}{1-\nu t},\frac{1}{1+\nu t}\right)\beta_\nu(t)\;. \qquad (2.7)
\end{aligned}$$

If we express the limit as νt tends to 1 in terms of τ_1, we see that it is the limit of the product of a polynomial and a decreasing exponential which converges to 0

$$\lim_{\nu t\uparrow 1} D^k\beta_\nu(t)=\lim_{\tau_1\to\infty} p_k(\tau_1,\frac{\tau_1}{2\tau_1-1})\,\frac{\nu}{B}\,e^{-\frac{\tau_1^2}{2\tau_1-1}}=0\;. \qquad (2.8)$$

Similarly, the limit towards –1 can be expressed in terms of τ_2 with the same result. Hence $\beta_\nu\in C^\infty$.

While β_ν is a test function for each value of ν, the sequence (β_m), $m\in\mathbb{N}$ doesn't converge in $\mathcal{D}$. For $t\neq 0$, $m\to\infty$ the value of $\beta_m(t)$ converges toward zero, while the value of the functions at $t=0$ grows without bounds. The limiting function is therefore not continuous.

The sequence $(\beta_{1/m})$ also doesn't converge in $\mathcal{D}$. As $m\to\infty$ the support of the functions grows without bounds. It is therefore not possible to find a compact set K containing the support of all members of the sequence as well as that of the limiting function.

An example of a converging sequence is β_m/m^2 which converges toward the zero function.

Example 2.2: Regularisation

Consider an impulse of finite duration (see Fig. 2.2a)

$$\mathsf{1}_k(t):=\mathsf{1}_+(t)-\mathsf{1}_+(t-k)=\begin{cases}1 & 0\le t<k\\ 0 & \text{otherwise}\end{cases}$$

This function is clearly not continuous at $t=0$ and $t=k$. These jump discontinuities can be removed by convolving $\mathsf{1}_k$ with the function β_ν of the previous example

$$\mathsf{1}_k*\beta_\nu(t)=\int_{-\infty}^{\infty}\mathsf{1}_k(\tau)\,\beta_\nu(t-\tau)\,d\tau=\int_0^k\beta_\nu(t-\tau)\,d\tau\;. \qquad (2.9)$$

We say that the so obtained function is the *regularised* of $\mathsf{1}_k$ by β_ν (see Fig. 2.2a).

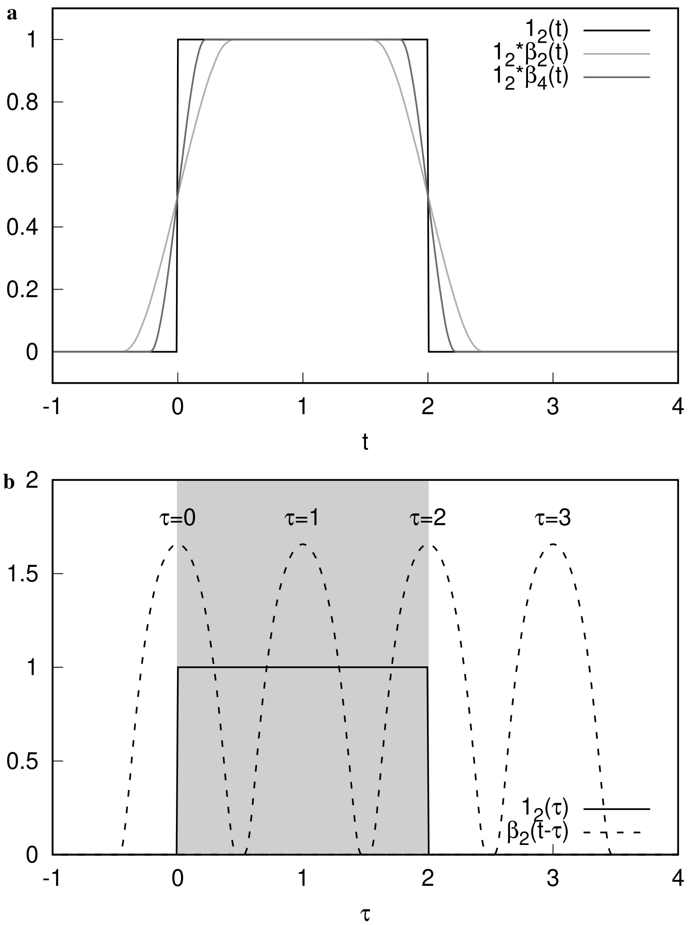


Fig. 2.2 **a** Regularized of the discontinuous function $\mathbb{1}_2(.)$ **b** Construction of the regularized of $\mathbb{1}_2(.)$

Observe that $\mathbb{1}_k * \beta_\nu$ is just a definite integral of β_ν and is therefore indefinitely differentiable. If the support of β_ν lies completely within the integration range, then, given the chosen normalization constant for β_ν, the value of $\mathbb{1}_k * \beta_\nu$ is 1. If the support of β_ν doesn't intersect the integration range, then the value of $\mathbb{1}_k * \beta_\nu$ is 0. For the remaining values of the independent variable t, $0 < \mathbb{1}_k * \beta_\nu(t) < 1$ (see Fig. 2.2b)

$$\mathbb{1}_k * \beta_\nu(t) = \begin{cases} 1 & 1/\nu \leq t \leq k - 1/\nu \\ 0 & t \leq -1/\nu \text{ or } t \geq k + 1/\nu \\ > 0 \text{ and } < 1 \text{ otherwise.} \end{cases} \tag{2.10}$$

We have thus established that $\mathbb{1}_k * \beta_\nu \in \mathcal{D}$.

From this example we see that for any open interval U and any closed interval $K \subset U$ we can construct a real valued test function ϕ with $0 \le \phi(t) \le 1$, a value of 1 within K and a value of 0 outside of U. This is a useful property that we will exploit later.

A similar construction can be made for test functions of more than one variable. For later reference we define a real valued test function with values between 0 and 1 that we call α such that

$$\alpha : \mathbb{R}^n \to [0, 1], \quad \tau \mapsto \begin{cases} 1 & |\tau| \le 1 \\ 0 & |\tau| \ge 2\,. \end{cases} \tag{2.11}$$

2.2 Distributions

A key aspect of the theory of distributions is the fact that it makes continuous functions differentiable any number of times. To see how this goes, remember from calculus that by partial integration we can transfer the operation of differentiation from one function to another one. Thus, if we pair the function of interest f with a function ϕ differentiable everywhere, then we can relate the derivative of f with a well-defined expression

$$\int_{-\infty}^{\infty} Df(\tau)\,\phi(\tau)\,d\tau = f(\tau)\,\phi(\tau)|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} f(\tau)\,D\phi(\tau)\,d\tau\;. \tag{2.12}$$

To make the expression independent of the limits of integration, the first term on the right-hand side should disappear. This can be achieved, for example, by choosing a function ϕ with compact support. In addition, to be able to assign a meaning to the derivative of any order, the function ϕ should be indefinitely differentiable. Note that these are precisely the properties of *test functions*.

An additional requirement is that of the assignment being unique. For example, the right-hand side expression should be identically zero only if $Df = 0$ (almost everywhere). Suppose that f has compact support. If the support of $D\phi$ doesn't overlap with the one of f then the right-hand expression is also zero and the assignment is not unique. To avoid this situation we are forced to pair the function f with *every* test function $\phi \in \mathcal{D}$.

A distribution is a generalization of these ideas and is defined as follows.

Definition 2.3 (*Distribution*) A *distribution* is defined as a *linear, continuous* function on the set of test functions

$$T : \mathcal{D}(\mathbb{R}^n) \to \mathbb{C}\,, \quad \phi \mapsto \langle T, \phi\rangle \tag{2.13}$$

This means that a distribution T has the following properties:

1. $\langle T, \phi_1 + \phi_2 \rangle = \langle T, \phi_1 \rangle + \langle T, \phi_2 \rangle$ for all $\phi_1,\ \phi_2 \in \mathcal{D}$.
2. $\langle T, c\,\phi \rangle = c\,\langle T, \phi \rangle$ for all $\phi \in \mathcal{D}$ and $c \in \mathbb{C}$.
3. From $\phi_k \underset{\mathcal{D}}{\longrightarrow} \phi$ it follows that $\langle T, \phi_k \rangle \to \langle T, \phi \rangle$, where the latter is the normal convergence of complex numbers.

Since distributions are linear by definition, the condition of continuity can be expressed in a slightly different, but equivalent way:

3'. From $\phi_k \underset{\mathcal{D}}{\longrightarrow} 0$ it follows that $\langle T, \phi_k \rangle \to 0$.

Two distributions T_1 and T_2 are equal if $\langle T_1, \phi \rangle = \langle T_2, \phi \rangle$ for every test function $\phi \in \mathcal{D}$. A distribution is called *real* if it evaluates to a real number when applied to any real valued test function.

The set of all distributions forms a vector space denoted by $\mathcal{D}'$, where addition of two distributions T_1 and T_2 and multiplication with a complex constant c are defined by

$$\begin{aligned}\langle T_1 + T_2, \phi \rangle &:= \langle T_1, \phi \rangle + \langle T_2, \phi \rangle \\ \langle cT, \phi \rangle &:= c\langle T, \phi \rangle = \langle T, c\,\phi \rangle .\end{aligned}$$

A mapping assigning a number to every element of a vector space is called a *functional*. Distributions are therefore functionals on test functions.

Example 2.3: Functions as distributions

Consider a continuous function $f \in C(\mathbb{R}^n)$. We can associate with it a distribution T_f by the procedure outlined at the beginning of the section

$$\langle T_f, \phi \rangle = \int_{\mathbb{R}^n} f(\tau)\,\phi(\tau)\,d^n\tau\,. \tag{2.14}$$

Linearity is clear from the properties of integrals. To see that it is continuous, consider a sequence of test functions converging to zero $\phi_m \underset{\mathcal{D}}{\longrightarrow} 0$. Then

$$\int_{\mathbb{R}^n} f(\tau)\,\phi_m(\tau)\,d^n\tau \le \sup_{t\in K} |\phi_m(t)| \int_K |f(\tau)|\,d^n\tau \ \longrightarrow\ 0$$

with K a compact set including the support of all ϕ_m.

Consider now two continuous functions f_1 and f_2. If $\langle T_{f_1}, \phi \rangle = \langle T_{f_2}, \phi \rangle$ for every $\phi \in \mathcal{D}$, then, by the properties of integrals of continuous functions, it follows that $f_1 = f_2$. We thus have an injective mapping from continuous functions to distributions. We can therefore *identify* continuous functions with their corresponding distributions and write $\langle f, \phi \rangle$ instead of $\langle T_f, \phi \rangle$.

The theory of distributions requires the use of Lebesgue integrals as opposed to Riemann ones, as Lebesgue's integration theory is more powerful and allows integrating a broader set of functions. Of course, when both integrals do exist, they coincide. A key concept in the Lebesgue theory of integration is that of the *measure*. For our purposes we can think of the Lebesgue measure as a volume and a set of *zero measure* in $\mathbb{R}^n$ as a (sufficiently regular) subspace of dimension $k < n$. A point on the real line $\mathbb{R}$, a line on a plane and a surface in $\mathbb{R}^3$ are all examples of sets of zero measure. The union of a *denumerable* family of sets of zero measure is itself a set of zero measure. Therefore, the set of rational numbers on the real line $\mathbb{R}$ has zero measure. Two locally integrable functions differing only on a set of zero measure are said to be equal *almost everywhere*.

Example 2.4: Locally integrable functions

Consider a *locally integrable* function $f \in \mathcal{L}^1_{\text{loc}}(\mathbb{R}^n)$, a function that is Lebesgue integrable over every compact set $K \subset \mathbb{R}^n$. As in the previous example we can associate it with a distribution through the integral (2.14). In this case however the mapping is not injective. Any two locally integrable functions f_1 and f_2 differing only in a set of measure zero produce the same value $\langle f_1, \phi\rangle = \langle f_2, \phi\rangle$ for every $\phi \in \mathcal{D}$. That means that they map to the same distribution.

In physical and engineering applications the values of a function in a set of zero measure is often unimportant. It is therefore natural to consider the *equivalence class* of all functions differing at most on a set of zero measure. In this way we obtain again an injective mapping, but now from the equivalence class of locally integrable functions differing at most on a set of zero measure (equal almost everywhere) to distributions, and we can again identify without ambiguity the former with the latter. To avoid overloading the notation we write a representative for the equivalence class, that is, we write $\langle f, \phi\rangle$ where f is a representative.

All distributions that can be represented by locally integrable functions through (2.14) are called *regular distributions*. However, not all distributions are regular and distributions that aren't regular are called *singular distributions*. Nonetheless, regular distributions are *dense* in $\mathcal{D}'$. That is, in a similar way as real numbers arise as a limiting process from rational ones, any distribution can be represented as a limit of regular distributions, where the convergence of distributions is defined as follows.

Definition 2.4 (*Convergence of distributions*) A sequence of distributions $(T_m)_{m\in\mathbb{N}}$ is said to converge to the distribution T, if the sequence of numbers $\langle T_m, \phi\rangle$ converges to the number $\langle T, \phi\rangle$ for every $\phi \in \mathcal{D}$. In symbols

$$T_m \xrightarrow[\mathcal{D}']{} T \quad \text{or} \quad \lim_{m\to\infty} T_m = T$$

if

$$\langle T_m, \phi \rangle \longrightarrow \langle T, \phi \rangle \quad \text{for every } \phi \in \mathcal{D}\,.$$

It is not obvious that the limit T is in fact a distribution, that is, linear and continuous. However, this is indeed the case. The space $\mathcal{D}'$ is thus closed under convergence. A proof can be found in [18].

This definition is based on a discrete parameter m traversing the natural numbers. If the parameter traverses a continuous set of values the situation is similar and can be reduced to the discrete case. Consider the sequence of distributions T_ν depending on the continuous parameter $\nu \in \mathbb{R}$. For each value of ν and each test function ϕ, the functional $\langle T_\nu, \phi \rangle$ evaluates to a number. For each test function ϕ the set of distributions T_ν therefore defines a function of ν

$$\zeta(\nu) = \langle T_\nu, \phi \rangle.$$

Lets define a sequence $(\nu_m)_{m\in\mathbb{N}}$ of values converging toward infinity. If for every such sequence and every test function

$$\lim_{k\to\infty} \zeta(\nu_k) = \lim_{k\to\infty} \langle T_{\nu_k}, \phi \rangle = \langle T, \phi \rangle$$

then

$$\lim_{\nu\to\infty} \langle T_\nu, \phi \rangle = \langle T, \phi \rangle.$$

Similarly for a continuous parameter converging toward a finite limit η.

Example 2.5: Dirac delta distribution

Consider the functions β_m of Example 2.1. The regular distributions associated with these functions form a sequence converging to a singular distribution. We have

$$\begin{gathered}
\lim\nolimits_{m\to\infty} \langle \beta_m, \phi \rangle \\
= \lim\nolimits_{m\to\infty} \int_{-1/m}^{1/m} \beta_m(\tau)\, \phi(\tau)\, d\tau \\
= \lim\nolimits_{m\to\infty} \left\{ \int_{-1/m}^{1/m} \beta_m(\tau)\, \phi(0)\, d\tau + \int_{-1/m}^{1/m} \beta_m(\tau)\, [\phi(\tau) - \phi(0)]\, d\tau \right\}
\end{gathered}$$

Since test functions are continuous and differentiable we can use the mean value theorem to express ϕ as

$$\phi(\tau) = \phi(0) + D\phi(\lambda)\, \tau$$

for some $\lambda \in (0, \tau)$. With this we can see that the second term converges to zero

$$\begin{aligned}\int_{-1/m}^{1/m} \beta_m(\tau)\,[\phi(\tau)-\phi(0)]\,d\tau \ &\le \int_{-1/m}^{1/m} |\beta_m(\tau)|\,|\phi(\tau)-\phi(0)|\,d\tau \\ &\le \sup_{\lambda\in(-\frac{1}{m},\frac{1}{m})} \frac{|D\phi(\lambda)|}{m} \int_{-1/m}^{1/m} |\beta_m(\tau)|\,d\tau \\ &\longrightarrow 0\end{aligned}$$

We therefore obtain

$$\begin{aligned}\lim_{m\to\infty}\langle\beta_m,\phi\rangle &= \phi(0)\ \lim_{m\to\infty}\int_{-1/m}^{1/m}\beta_m(\tau)\,d\tau \\ &= \phi(0)\ \lim_{m\to\infty} 1 \\ &= \phi(0).\end{aligned}$$

The sequence β_m thus converge to the Dirac delta distribution δ which is defined by

$$\langle\delta,\phi\rangle := \phi(0). \tag{2.15}$$

Besides the sequence β_m there are many other regular distribution sequences converging to δ. For example, with the same procedure used above, it is simple to show that the sequence defined by the following functions does also converge to δ

$$f_m(t) = \begin{cases} m/2 & |t| \le 1/m \\ 0 & |t| > 1/m \end{cases}$$

Note that the notation used in many technical texts to define the Dirac delta distribution is not mathematically correct and only has a symbolic value

$$\int_{-\infty}^{\infty}\delta(\tau)\,\phi(\tau)\,d\tau = \phi(0).$$

This notation imply the existence of a *function* with a value of zero everywhere but at $\tau = 0$ where its value is infinite. However, the value of the Lebesgue integral of such a function is zero since a single point of the real line has zero measure. This notation is however useful as it helps to remember several properties that we will see shortly.

Example 2.6: Cauchy principal value

The function $f(\tau) = 1/\tau$ is not locally integrable. For this reason we can't associate with it a regular distribution through (2.14). A way around this is to use the *Cauchy principal value* of the integral to define the following singular distribution

$$\begin{aligned}\left\langle \mathrm{pv}\,\frac{1}{\tau}, \phi \right\rangle &:= \mathrm{pv} \int_{-\infty}^{\infty} \frac{\phi(\tau)}{\tau}\, d\tau \\ &= \lim_{\epsilon \downarrow 0} \left\{ \int_{-\infty}^{-\epsilon} \frac{\phi(\tau)}{\tau}\, d\tau + \int_{\epsilon}^{\infty} \frac{\phi(\tau)}{\tau}\, d\tau \right\} \end{aligned} \tag{2.16}$$

Integrating by parts the first integral we obtain

$$\begin{aligned}\int_{-\infty}^{-\epsilon} \frac{\phi(\tau)}{\tau}\, d\tau &= \left(\phi(\tau)\, \ln|\tau|\right)\Big|_{-\infty}^{-\epsilon} - \int_{-\infty}^{-\epsilon} \ln|\tau|\, D\phi(\tau)\, d\tau \\ &= \phi(-\epsilon)\, \ln|\epsilon| - \int_{-\infty}^{-\epsilon} \ln|\tau|\, D\phi(\tau)\, d\tau\end{aligned}$$

and similarly for the second integral

$$\int_{\epsilon}^{\infty} \frac{\phi(\tau)}{\tau}\, d\tau = -\phi(\epsilon)\, \ln(\epsilon) - \int_{\epsilon}^{\infty} \ln(\tau)\, D\phi(\tau)\, d\tau\,.$$

We see that the first term of both integrals do diverge as ϵ goes to 0. However, using the mean value theorem, we note that there are values $\lambda_1 \in (0, \epsilon)$ and $\lambda_2 \in (-\epsilon, 0)$ such that

$$\begin{aligned}\phi(\epsilon) &= \phi(0) + \epsilon D\phi(\lambda_1) \\ \phi(-\epsilon) &= \phi(0) - \epsilon D\phi(\lambda_2)\end{aligned}$$

With $M = -\left(D\phi(\lambda_1) + D\phi(\lambda_2)\right)$, the limit of the sum of the diverging parts therefore do cancel

$$\lim_{\epsilon \downarrow 0} M \epsilon\, \ln|\epsilon| = 0$$

and we finally obtain

$$\langle \mathrm{pv}\,\frac{1}{\tau}, \phi \rangle = -\int_{-\infty}^{\infty} \ln|\tau|\, D\phi(\tau)\, d\tau\,.$$

This last integral is well defined as $\ln|\tau|$ is locally integrable and therefore defines a well defined regular distribution. We will meet this distribution again in the context of the Fourier transform of distributions.

2.3 Basic Properties

There are some useful operations that we can perform on locally integrable functions that can be carried over to distributions. A common operation is to shift a function f by an amount τ to obtain $t \mapsto f(t-\tau)$. If we apply the change of variable $\lambda = t - \tau$ to the regular distribution associated with the shifted function we obtain

$$\int_{-\infty}^{\infty} f(t-\tau)\,\phi(t)\,dt = \int_{-\infty}^{\infty} f(\lambda)\,\phi(\lambda+\tau)\,d\lambda.$$

By generalizing this result we define the operation of *shifting a distribution* by

$$\langle T(t-\tau), \phi(t)\rangle := \langle T(t), \phi(t+\tau)\rangle. \tag{2.17}$$

With this definition we can for example denote a Dirac pulse at time τ by $\delta(t-\tau)$

$$\langle \delta(t-\tau), \phi(t)\rangle := \phi(\tau).$$

! Notation

Note that a distribution T isn't a function of the variable t. In spite of this it is useful to write $T(t)$ to indicate the symbol used for the independent variable of the testing function (this will be useful when we'll introduce operations such as the convolution) and as a convenient notation to indicate some operations such as shifting. In no way this is meant to imply the existence of a function or that the distribution is regular.

Another useful operation is multiplication of the independent variable of a function by a constant a. By generalizing what happens with regular distributions, we define *multiplication of the independent variable by a constant a* for any distribution in $\mathcal{D}'(\mathbb{R}^n)$ by

$$\langle T(a\,t), \phi(t)\rangle := \left\langle T(t), \frac{1}{|a|^n}\phi\left(\frac{t}{a}\right)\right\rangle. \tag{2.18}$$

This operation is closely related to the concepts of even and odd distributions.

Definition 2.5 (*Even and odd distributions*) An *even* distribution T is defined as a distribution for which, for every test function ϕ

$$\langle T(t), \phi(-t)\rangle = \langle T(t), \phi(t)\rangle. \tag{2.19}$$

Similarly, an *odd* distribution satisfies

$$\langle T(t), \phi(-t)\rangle = -\langle T(t), \phi(t)\rangle \tag{2.20}$$

for every test function ϕ.

A further useful operation is *multiplication of a distribution with an indefinitely differentiable function* γ. First note that multiplication of a test function ϕ with an indefinitely differentiable function results in another test function. For this reason we can again generalize the behavior of regular distributions and define

$$\langle \gamma\, T, \phi\rangle := \langle T, \gamma\, \phi\rangle\,. \tag{2.21}$$

2.4 Differentiation of Distributions

At the beginning of Sect. 2.2 we mentioned that one of the distinguishing features of distributions is the fact that they can be differentiated any number of times. We also argued that, for regular distributions, partial integration leads to an expression which can be considered as the definition of the derivative of regular distributions

$$\begin{aligned}\langle f^{(1)}, \phi\rangle &= \int_{-\infty}^{\infty} f^{(1)}(\tau)\, \phi(\tau)\, d\tau \\ &= -\int_{-\infty}^{\infty} f(\tau)\, \phi^{(1)}(\tau)\, d\tau \\ &= \langle f, -\phi^{(1)}\rangle.\end{aligned} \tag{2.22}$$

In fact, this definition can be extended to singular distributions and to distributions of several variables, that is to arbitrary distributions.

Definition 2.6 The first order partial derivative of a distribution T on $\mathcal{D}(\mathbb{R}^n)$ is defined by

$$\langle D_i T, \phi\rangle := \langle T, -D_i\phi\rangle \quad i = 1, \ldots, n\,. \tag{2.23}$$

Since the derivative of a test function $D_i\phi$ is still a test function, it follows that the derivative of a distribution is always a distribution and that distributions can be differentiated an arbitrary number of times.

With k an n-tuple of non-negative integers, the derivative of order $|k|$ follows from the above definition

$$\langle D^k T, \phi\rangle = (-1)^{|k|}\langle T, D^k\phi\rangle\,. \tag{2.24}$$

The order of differentiation is irrelevant since test functions have continuous partial derivatives of all orders and hence

$$\langle D_i D_j T, \phi\rangle = \langle T, D_j D_i\phi\rangle = \langle T, D_i D_j\phi\rangle = \langle D_j D_i T, \phi\rangle\,.$$

The rule for the differentiation of the product of a distribution T and an indefinitely differentiable function γ is the same as the rule of differentiation for the product of two functions

$$\begin{aligned}\langle D_i(\gamma T), \phi\rangle &= -\langle \gamma T, D_i\phi\rangle = -\langle T, \gamma\, D_i\phi\rangle \\ &= -\langle T, D_i(\gamma\,\phi)\rangle + \langle T, D_i\gamma\,\phi\rangle \\ &= \langle D_i T, \gamma\,\phi\rangle + \langle D_i\gamma\, T, \phi\rangle \\ &= \langle \gamma\, D_i T, \phi\rangle + \langle D_i\gamma\, T, \phi\rangle\end{aligned}$$

or

$$D_i(\gamma T) = \gamma\, D_i T + (D_i\gamma)\, T\,. \tag{2.25}$$

Two important properties of distributional differentiation follow immediately from the definition. The first is that differentiation is a linear operation: given two distributions T_1 and T_2 and two numbers c_1 and c_2

$$D^k(c_1\, T_1 + c_2\, T_2) = c_1\, D^k T_1 + c_2\, D^k T_2\,. \tag{2.26}$$

The second is continuity: given a sequence of distributions $(T_m)_{m\in\mathbb{N}}$ converging toward a distribution T, the sequence of corresponding partial derivatives $(D^k T_m)_{m\in\mathbb{N}}$ converges to $D^k T$

$$\begin{aligned}\lim_{m\to\infty} \langle D^k T_m, \phi\rangle &= \lim_{m\to\infty} (-1)^{|k|}\langle T_m, D^k\phi\rangle \\ &= (-1)^{|k|}\langle T, D^k\phi\rangle \\ &= \langle D^k T, \phi\rangle\,.\end{aligned} \tag{2.27}$$

In other words, *the operations of limit-taking and differentiation can always be exchanged.* In particular this means that if a sequence of partial sums $S_m = \sum_{i=0}^{m-1} T_i$ converges to a series $S = \sum_{i=0}^{\infty} T_i$, then the *series can be differentiated term by term.*

! Notation

To distinguish a regular distribution defined by the usual derivative of a function f from the derivative in the sense of distributions of the regular distribution defined by f, we are going to always denote the former by $T_{f^{(k)}}$ or $T_{D^k f}$. The later will be denoted interchangeably by $f^{(k)}$, $D^k f$, $T_f^{(k)}$ or $D^k T_f$.

Example 2.7: Derivative of δ

The first order derivative of the Dirac delta distribution is

$$\langle \delta^{(1)}, \phi\rangle = -\langle \delta, \phi^{(1)}\rangle = -\phi^{(1)}(0)\,.$$

The kth order one is

$$\langle \delta^{(k)}, \phi \rangle = (-1)^k \phi^{(k)}(0) .$$

This example shows that, in general, to calculate the value $\langle T, \phi \rangle$ of a distribution T when applied to a test function ϕ, it's not enough to know the values of ϕ over $\text{supp}(T)$. We need to know the values of ϕ over a *neighborhood* of $\text{supp}(T)$.

Example 2.8: Derivative of $\mathsf{1}_+$

The derivative of the Heaviside unit step $\mathsf{1}_+$ as a function is zero everywhere but at $t = 0$ (a set of zero measure) where it is undefined. The *function* $\mathsf{1}_+^{(1)}$ is therefore locally integrable, and we can define the regular distribution $T_{\mathsf{1}_+^{(1)}}$ which evaluates to zero for every test function ϕ.

Differently from this, the derivative of $\mathsf{1}_+$ as a distribution is defined everywhere and, applying the definition, we find

$$\langle \mathsf{1}_+^{(1)}, \phi \rangle = -\langle \mathsf{1}_+, \phi^{(1)} \rangle = -\int_0^\infty \phi^{(1)}(\tau)\, d\tau = -\phi(\tau)|_0^\infty = \phi(0) = \langle \delta, \phi \rangle ,$$

that is

$$\mathsf{1}_+^{(1)} = \delta .$$

Example 2.9: Function versus distributional derivative

Consider a function $f : \mathbb{R} \to \mathbb{C}$ continuously differentiable k times everywhere but at $t = 0$, where it has a discontinuity such that both limits

$$\lim_{t \downarrow 0} f^{(i)}(t) \quad \text{and} \quad \lim_{t \uparrow 0} f^{(i)}(t)$$

exist for all $i \leq k$. Let's denote the difference between these limits by

$$\alpha_i = \lim_{t \downarrow 0} f^{(i)}(t) - \lim_{t \uparrow 0} f^{(i)}(t) .$$

Then we may represent the function f as

$$f(t) = f_{c,0}(t) + \alpha_0 \mathsf{1}_+(t)$$

with $f_{c,0}$ a continuous function. It is easy to see that for $t \neq 0$, $f_{c,0}^{(1)}(t) = f^{(1)}(t)$. Thus, using the results of Example 2.8 the first order derivative of f is

$$T_f^{(1)} = T_{f^{(1)}} + \alpha_0 \delta\,.$$

To compute the second-order derivative we can use the same procedure. We decompose the function $f^{(1)}$ into a continuous function $f_{c,1}$ and a step

$$f^{(1)}(t) = f_{c,1}(t) + \alpha_1 \mathsf{1}_+(t)\,.$$

Differentiating term by term we therefore obtain

$$\begin{aligned} T_f^{(2)} &= T_{f^{(1)}}^{(1)} + \alpha_0 \delta^{(1)} \\ &= T_{f_{c,1}}^{(1)} + \alpha_1 T_{\mathsf{1}_+}^{(1)} + \alpha_0 \delta^{(1)} \\ &= T_{f_{c,1}^{(1)}} + \alpha_1 \delta + \alpha_0 \delta^{(1)} \\ &= T_{f^{(2)}} + \alpha_1 \delta + \alpha_0 \delta^{(1)} \end{aligned}$$

The kth order derivative can be obtained by iterating this procedure

$$T_f^{(k)} = T_{f^{(k)}} + \alpha_0 \delta^{(k-1)} + \alpha_1 \delta^{(k-2)} + \cdots + \alpha_{k-1} \delta\,. \tag{2.28}$$

Example 2.10: Logarithm derivative

In Example 2.6 we showed that the Cauchy principal value of $1/\tau$ is a distribution

$$\langle \mathrm{pv}\,\frac{1}{\tau}, \phi\rangle = -\int_{-\infty}^{\infty} \ln|\tau|\, D\phi(\tau)\, d\tau.$$

We now recognize this result as saying

$$\mathrm{pv}\,\frac{1}{\tau} = D \ln|\tau|\,.$$

Example 2.11: Limit to ∞ of trigonometric functions

Consider the following parameterized distribution

$$f_\omega(t) = -\frac{\cos \omega t}{\omega} \qquad \omega > 0.$$

As ω tends to infinity, it converges to

$$\begin{aligned}
\lim_{\omega\to\infty} |\langle f_\omega, \phi\rangle| &= \lim_{\omega\to\infty} \left| \int_{\mathrm{supp}(\phi)} -\frac{\cos\omega t}{\omega}\,\phi(t)\,dt \right| \\
&\le \lim_{\omega\to\infty} \int_{\mathrm{supp}(\phi)} \frac{|\phi(t)|}{\omega}\,dt \\
&\le \lim_{\omega\to\infty} \frac{\sup|\phi(t)|}{\omega}\,K \\
&= 0
\end{aligned}$$

with $K = \mathrm{supp}(\phi)$. Since distributions can always be differentiated and for distributions arising from continuous functions the derivative as a distribution coincides with the derivative as a function, we have the following result

$$\begin{aligned}
\lim_{\omega\to\infty} \langle \sin\omega t, \phi\rangle &= \lim_{\omega\to\infty} \langle -D\frac{\cos\omega t}{\omega}, \phi\rangle \\
&= \lim_{\omega\to\infty} \langle \frac{\cos\omega t}{\omega}, D\phi\rangle \\
&= 0
\end{aligned}$$

or

$$\lim_{\omega\to\infty} \sin\omega t = 0. \tag{2.29}$$

Similarly one obtains

$$\lim_{\omega\to\infty} \cos\omega t = 0 \quad \text{and} \tag{2.30}$$

$$\lim_{\omega\to\infty} e^{\jmath\omega t} = 0. \tag{2.31}$$

Note that these limits do not exist for the corresponding functions.

2.5 Distributions with Compact Support

The property of multiplication of a distribution with an indefinitely differentiable function suggests another interesting generalization. Let's start again with a regular distribution f. Then, if we write out explicitly the integral of the distribution $\gamma\, f$

$$\langle \gamma\, f, \phi\rangle = \int_{-\infty}^{\infty} f(\tau)\,\gamma(\tau)\,\phi(\tau)\,d\tau$$

we see that in principle we could group the functions differently and write $\langle \phi\, f, \gamma \rangle$. This is however not a distribution in $\mathcal{D}'$ since γ is not a test function as its support is not compact. In spite of this, the number $\langle \phi\, f, \gamma \rangle$ is the same as $\langle \gamma\, f, \phi \rangle$ for every test function ϕ and every function $\gamma \in C^{\infty}$. A moment's reflection reveals that what makes these two expressions have the same value for every value of γ is the fact that, as a function, $\phi\, f$ has *compact support*.

To generalize this observation to arbitrary distributions we must first define what the support of a distribution T is.

A distribution is said to *vanish* on an open set $U \in \mathbb{R}^n$ if $\langle T, \phi \rangle = 0$ for all test functions ϕ with $\operatorname{supp}(\phi) \subset U$, where $\operatorname{supp}(\phi)$ is the support of the test function ϕ.

Definition 2.7 (*Support of a distribution*) The *support of a distribution T* is the complement of the largest open set U on which the distribution vanishes and is denoted by $\operatorname{supp}(T)$.

The set of all distributions with compact support is denoted by $\mathcal{E}'$ and forms a vector subspace of $\mathcal{D}'$, that is $\mathcal{E}' \subset \mathcal{D}'$.

Example 2.12: Support of δ

The value of the Dirac delta distribution δ applied to any test function ϕ with $\operatorname{supp}(\phi) \in U = (-\infty, 0) \cup (0, \infty)$ is zero. That is, δ vanishes on U. Its support is $\operatorname{supp}(\delta) = \mathbb{R} \setminus U = \{0\}$ and is therefore compact.

With the notion of the support of a distribution we can generalize our observation that $\langle \phi\, f, \gamma \rangle = \langle \gamma\, f, \phi \rangle$ by saying that distributions with compact support $T \in \mathcal{E}'$ can be extended to *continuous, linear* functionals L on indefinitely differentiable functions with arbitrary support. In this context the vector space of all indefinitely differentiable functions is denoted by $\mathcal{E}$ and, to give a meaning to the continuity of the functionals, it is equipped with the following convergence criteria.

Definition 2.8 (*Convergence in $\mathcal{E}$*) A sequence $(\gamma_m)_{n\in\mathbb{N}} \in \mathcal{E}$ is said to converge to γ if for every compact subset K of $\mathbb{R}^n$ and every n-tuple k, the set of functions $D^k\gamma_m$ converges uniformly to $D^k\gamma$

$$\sup_{x\in K} \left| D^k\gamma_m - D^k\gamma \right| \to 0, \qquad m \to \infty.$$

Assume that the support of T is the compact set K and let α be a test function equal to 1 in a neighborhood U of K. Then for every function $\gamma \in \mathcal{E}$ and for every point $\tau \in U$, $\alpha(\tau)\,\gamma(\tau) = \gamma(\tau)$. Therefore, there is a functional L such that

$$\langle L, \gamma \rangle = \langle T, \alpha\gamma \rangle \qquad \gamma \in \mathcal{E}. \tag{2.32}$$

That it is independent of the choice of α is easily verified: Suppose that α_1 and α_2 are two test functions equal to 1 in a neighborhood of K. Then in the smallest of these neighborhoods $\alpha_1 - \alpha_2 = 0$ and $\langle T, \alpha_1\,\gamma \rangle = \langle T, \alpha_2\,\gamma \rangle$.

The functional thus defined is *unique* since, for every sequence of test functions α_m equal to 1 for $|\tau| < m$ we have: on the one hand, by continuity of L

$$\lim_{m\to\infty} \langle L, \alpha_m\, \gamma\rangle = \langle L, \gamma\rangle$$

and on the other hand, for sufficiently large m

$$\langle T, \alpha_m\, \gamma\rangle = \langle L, \gamma\rangle\,.$$

Therefore every distribution with compact support T defines a unique continuous, linear functional L on $\mathcal{E}$.

The converse is also true: Every continuous, linear functional L restricted to $\mathcal{D} \subset \mathcal{E}$ defines a distribution T with compact support. For, if this was not the case and the support of T was not compact, then we could find a sequence of test functions $\phi_m \in \mathcal{D}$ with support in the complement of $|\tau| < m$, such that $\langle T, \phi_m\rangle = 1$ for all m. However, since in $\mathcal{E}$ $\lim_{m\to\infty} \phi_m = 0$, by continuity of L

$$\lim_{m\to\infty} \langle L, \phi_m\rangle = 0\,.$$

Therefore if $\langle L, \phi\rangle = \langle T, \phi\rangle$ for all $\phi \in \mathcal{D}$, then the support of T must be compact.

There are other vector sub-spaces of $\mathcal{D}'$ which can be extended to larger sets of functions than $\mathcal{D}$. We will encounter another one in the context of the Fourier transform. The set of test functions $\mathcal{D}$ are the common set on which all distributions are defined.

2.5.1 Single-Point Support

We now investigate distributions satisfying the following equation

$$t^k\, T = 0\,, \tag{2.33}$$

that is, distributions for which, for every $k \geq 1$ and test function ϕ

$$\langle t^k\, T, \phi\rangle = \langle T, t^k\, \phi\rangle = 0\,.$$

For simplicity we limit ourselves to the one dimensional case.

First observe that on the open set $U = (-\infty, 0) \cup (0, \infty)$ the function $t \mapsto t^k$ doesn't assume the zero value. For this reason, to satisfy the equation, T must vanish on U, or, stated in other words, the support of T must be the origin: $\text{supp}(T) = \{0\}$.

Since the support of T is compact (a single point) and, for any test function ϕ, the value of $\langle T, \phi \rangle$ is determined by the values of ϕ in a neighborhood (however small) of $\operatorname{supp}(T)$, we can expand ϕ using Taylor's formula with remainder [19]. For our purposes it is convenient to express the remainder in integral form which can be obtained by integrating by parts multiple times

$$\begin{aligned}
\phi(t) &= \phi(0) + \int_0^t \phi^{(1)}(\tau)\, d\tau \\
&= \phi(0) - (t-\tau)\phi^{(1)}(\tau)\Big|_0^t + \int_0^t (t-\tau)\,\phi^{(2)}(\tau)\, d\tau \\
&= \phi(0) + t\,\phi^{(1)}(0) - \frac{(t-\tau)^2}{2}\phi^{(2)}(\tau)\Big|_0^t + \int_0^t \frac{(t-\tau)^2}{2}\,\phi^{(3)}(\tau)\, d\tau \\
&= \cdots \\
&= \sum_{m=0}^{k-1} \frac{\phi^{(m)}(0)}{m!}\, t^m + \int_0^t \frac{(t-\tau)^{k-1}}{(k-1)!}\,\phi^{(k)}(\tau)\, d\tau.
\end{aligned}$$

By performing the substitution $\tau = t\,\lambda$ the remainder can be transformed in the following form

$$t^k \int_0^1 \frac{(1-\lambda)^{k-1}}{(k-1)!}\,\phi^{(k)}(t\,\lambda)\, d\lambda = \frac{t^k}{(k-1)!}\,\psi(t)$$

which makes it apparent that it is proportional to the product of t^k and an indefinitely differentiable function $\psi \in \mathcal{E}$. Note that, differently from ϕ, no addend has compact support. This poses no problem since T, having itself compact support, can be extended uniquely to a distribution on $\mathcal{E}$ (see Sect. 2.5).

With this expansion we can express the value of $\langle T, \phi \rangle$ as a finite sum. Taking into account (2.33)

$$\begin{aligned}
\langle T, \phi \rangle &= \sum_{m=0}^{k-1} \frac{\phi^{(m)}(0)}{m!}\,\langle T, t^m \rangle + \frac{1}{(k-1)!}\langle T, t^k\,\psi(t)\rangle \\
&= \sum_{m=0}^{k-1} \frac{\phi^{(m)}(0)}{m!}\,\langle T, t^m \rangle \\
&= \sum_{m=0}^{k-1} c_m\,\langle \delta^{(m)}, \phi \rangle
\end{aligned}$$

or

$$T = \sum_{m=0}^{k-1} c_m \delta^{(m)} \tag{2.34}$$

with

$$c_m = (-1)^m \, \frac{\langle T, t^m \rangle}{m!} \,.$$

We have therefore established that the homogeneous equation

$$t^k \, T = 0$$

has an infinity of non-trivial solutions, each being a weighted sum of the Dirac delta distribution δ and its derivatives up to order $k - 1$. In addition, this shows that δ and its derivatives are the only distributions with the support consisting of a single point.

Example 2.13: Solutions of $t\,T = 1$

We want to find all solutions of the equation

$$t \, T = 1 \,.$$

If T would be a function then the equation would have no solution at $t = 0$ and $1/t$ at all other points. From this we guess that the solution as a distribution could be $T = \text{pv}\,1/t$. Indeed, this distribution satisfies the equation

$$\begin{aligned}
\left\langle t \, \text{pv}\frac{1}{t}, \phi \right\rangle &= \left\langle \text{pv}\frac{1}{t}, t\,\phi \right\rangle \\
&= \lim_{\epsilon \downarrow 0} \int_{t \geq \epsilon} \frac{1}{t} \, t\, \phi(t)\, dt \\
&= \int_{-\infty}^{\infty} \phi(t)\, dt \\
&= \langle 1, \phi \rangle \,.
\end{aligned}$$

However, this is not the only solution as the homogeneous equation has non-trivial solutions given by (2.34) (k = 1). The equation is therefore satisfied by all distributions of the form

$$T = \text{pv}\frac{1}{t} + c\,\delta(t)$$

with c an arbitrary constant.

Table 2.1 Main properties and operations on distributions

$\langle T, \sum_m c_m\, \phi_m\rangle$	$=$	$\sum_m c_m \langle T, \phi_m\rangle$
$\langle \sum_m c_m\, T_m, \phi\rangle$	$=$	$\sum_m c_m \langle T_m, \phi\rangle$
$\langle T(t-\tau), \phi(t)\rangle$	$=$	$\langle T(t), \phi(t+\tau)\rangle$
$\langle T(a\,t), \phi(t)\rangle$	$=$	$\langle T(t), \frac{1}{\lvert a\rvert^n}\phi\left(\frac{t}{a}\right)\rangle$
$\langle \gamma\, T, \phi\rangle$	$=$	$\langle T, \gamma\, \phi\rangle$
$\langle D^k T, \phi\rangle$	$=$	$(-1)^{\lvert k\rvert}\langle T, D^k\phi\rangle$
$\lim_{m\to\infty}\langle D^k T_m, \phi\rangle$	$=$	$\langle D^k T, \phi\rangle$
$\langle D^k(\sum_m c_m\, T_m), \phi\rangle$	$=$	$\sum_m c_m \langle D^k T_m, \phi\rangle$
$\langle \sum_{m=0}^{k-1} t^k D^m \delta, \phi\rangle$	$=$	0
$\langle S\otimes T, \phi\rangle$	$=$	$\langle S, \langle T, \phi\rangle\rangle = \langle T, \langle S, \phi\rangle\rangle$
$\langle S * T, \phi\rangle$	$=$	$\langle S(\tau)\otimes T(\lambda), \phi(\tau+\lambda)\rangle$

The properties of distributions discussed in this chapter and some that will be discussed in the following ones are summarised in Table 2.1.

Chapter 3
Convolution of Distributions

The convolution product plays a central role in the description of linear and weakly-nonlinear systems. In this chapter we develop its theory based on distributions. In addition, the chapter introduces the tensor product which will also come to play an important role in the description of weakly-nonlinear systems.

3.1 Tensor Product

The general notion of convolution of distributions is defined in terms of the *tensor product*. We therefore start by defining this product and, as before, we start by considering regular distributions.

The tensor product is a bilinear operation that can be used to generate a vector space out of other vector spaces. If f is a function on $\mathbb{R}^m$ and g a function on $\mathbb{R}^n$, then the tensor product of f and g is defined as the function

$$f \otimes g : \mathbb{R}^{m+n} \to \mathbb{C} \quad (\tau, \lambda) \mapsto f(\tau)g(\lambda)\,.$$

The tensor product of two locally integrable functions is itself locally integrable. Therefore, if we now assume f and g to be locally integrable, we can try to build the tensor product of the regular distributions T_f and T_g based on the tensor product $f \otimes g$. If we use as test function the tensor product of two suitable test functions ξ and ψ then we obtain

$$\langle T_f \otimes T_g, \xi \otimes \psi \rangle = \langle T_f, \xi \rangle \langle T_g, \psi \rangle$$

which is well-defined. However, for an arbitrary test function $\phi \in \mathcal{D}(\mathbb{R}^{m+n})$ it is not immediately apparent that the result is a distribution. Taking $m = n = 1$ for simplicity, we have

F. Beffa, *Weakly Nonlinear Systems*, Understanding Complex Systems,
https://doi.org/10.1007/978-3-031-40681-2_3

$$\begin{aligned}\langle T_f \otimes T_g, \phi\rangle &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\tau)\, g(\lambda)\, \phi(\tau,\lambda)\, d\lambda\, d\tau \\ &= \int_{-\infty}^{\infty} f(\tau) \int_{-\infty}^{\infty} g(\lambda)\, \phi(\tau,\lambda)\, d\lambda\, d\tau\,.\end{aligned}$$

The inner integral evaluates to a number for every value of the variable τ, that is, it is a complex valued function of τ that we call $\zeta(\tau)$. Furthermore, the variable τ only appears as an argument of the test function ϕ. Therefore ζ must have compact support. In addition, when computing its derivative, differentiation can be moved under the integral and ζ is therefore indefinitely differentiable. In other words ζ is a test function. We therefore have

$$\langle T_f \otimes T_g, \phi\rangle = \langle T_f, \langle T_g, \phi\rangle\rangle = \langle T_g, \langle T_f, \phi\rangle\rangle$$

where the last equality comes from the fact that we could reverse the order of integration without changing the result. This last property is referred to as *Fubini's theorem.*

The above arguments can be generalized to arbitrary distributions. That the inner functional is a function of τ and that has compact support is clear. The fact that it can be differentiated comes from the continuity and linearity of distributions

$$\begin{aligned}D\,\zeta(\tau) &= \lim_{\epsilon\to 0}\frac{\zeta(\tau+\epsilon)-\zeta(\tau)}{\epsilon} \\ &= \lim_{\epsilon\to 0}\frac{\langle T(\lambda), \phi(\tau+\epsilon,\lambda))\rangle - \langle T(\lambda), \phi(\tau,\lambda)\rangle}{\epsilon} \\ &= \lim_{\epsilon\to 0}\langle T(\lambda), \frac{\phi(\tau+\epsilon,\lambda)-\phi(\tau,\lambda)}{\epsilon}\rangle \\ &= \langle T(\lambda), D_1\phi(\tau,\lambda)\rangle\,. \end{aligned} \tag{3.1}$$

With this we see that ζ can be differentiated an arbitrary number of times and is thus a test function. We therefore obtain the following general definition for the tensor product of distributions.

Definition 3.1 (*Tensor product*) Given two distributions $S \in \mathcal{D}'(\mathbb{R}^m)$ and $T \in \mathcal{D}'(\mathbb{R}^n)$ the tensor product $S \otimes T$ is the distribution in $\mathcal{D}'(\mathbb{R}^{m+n})$ defined by

$$\langle S \otimes T, \phi\rangle := \langle S, \langle T, \phi\rangle\rangle = \langle T, \langle S, \phi\rangle\rangle\,. \tag{3.2}$$

It's easy to see that the tensor product of distributions is bilinear

$$\begin{aligned}(S+T)\otimes U &= S\otimes U + T\otimes U \\ S\otimes(T+U) &= S\otimes T + S\otimes U\,,\end{aligned} \tag{3.3}$$

and associative

$$(S \otimes T) \otimes U = S \otimes (T \otimes U)\,. \tag{3.4}$$

As a useful abbreviation of notation we define the tensor power by

$$\begin{aligned} T^{\otimes k} &:= \underbrace{T \otimes \ldots \otimes T}_{k \text{ times}}\,, \qquad k > 0 \\ T^{\otimes 0} &:= 1 \in \mathbb{C}\,. \end{aligned} \tag{3.5}$$

Example 3.1: Higher Dimensional Dirac Pulse

The tensor product of two Dirac pulses is

$$\begin{aligned} \langle \delta \otimes \delta(\tau, \lambda), \phi(\tau, \lambda)\rangle &= \langle \delta(\tau), \langle \delta(\lambda), \phi(\tau, \lambda)\rangle\rangle = \langle \delta(\tau), \phi(\tau, 0)\rangle \\ &= \phi(0, 0) \\ &=: \langle \delta(\tau, \lambda), \phi(\tau, \lambda)\rangle\,. \end{aligned}$$

3.2 Convolution of Distributions

We now come to the main objective of this section: the convolution of distributions. Remember that the convolution of integrable functions $f, g \in L^1$ is defined as follows

$$f * g\,(t) := \int_{-\infty}^{\infty} f(\tau)\, g(t - \tau)\, d\tau\,.$$

To obtain a distribution we may write

$$\begin{aligned} \langle f * g, \phi\rangle &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\tau)\, g(t - \tau)\, d\tau\, \phi(t)\, dt \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\tau)\, g(\lambda)\, \phi(\lambda + \tau)\, d\tau\, d\lambda \end{aligned}$$

which can be represented as the following tensor product

$$\langle f * g, \phi\rangle = \langle f(\tau) \otimes g(\lambda), \phi(\lambda + \tau)\rangle\,.$$

However, while indefinitely differentiable, the function $\psi(\tau, \lambda) = \phi(\lambda + \tau)$ is not a test function because its support is not compact. In fact, $\psi(\tau, \lambda)$ assumes the same

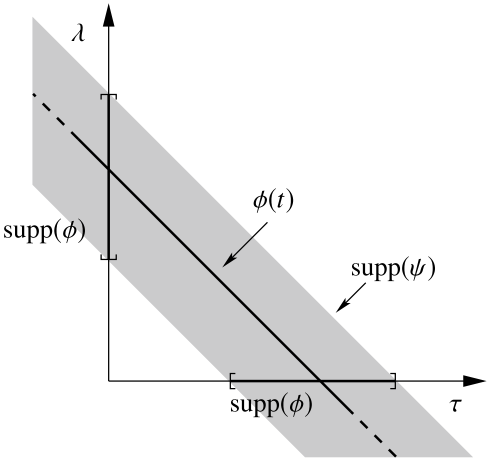


Fig. 3.1 Support of $\psi(\tau, \lambda) = \phi(\tau + \lambda)$

value $\phi(t)$ for every point on the diagonal line $t = \lambda + \tau$ of the (τ, λ)-plane (see Fig. 3.1). In spite of this, on account of our assumption that f and g are integrable functions (and not merely locally integrable), the above integral is well-defined. We therefore conclude that, similarly to the case of functions, *the convolution of distributions only exists for a subset of distributions with additional characteristics.*

Definition 3.2 (*Convolution*) Given two distributions S and T in $\mathcal{D}'(\mathbb{R}^n)$, if for every test function $\phi \in \mathcal{D}(\mathbb{R}^n)$ the tensor product $S \otimes T$ can be extended to functions of the form $\psi(\tau, \lambda) = \phi(\tau + \lambda)$, then the convolution product $S * T$ is defined by

$$\langle S * T, \phi \rangle := \langle S(\tau) \otimes T(\lambda), \phi(\tau + \lambda) \rangle \tag{3.6}$$

and is commutative

$$S * T = T * S \, . \tag{3.7}$$

A *sufficient condition for the existence of the convolution* is as follows: if the intersection of the support of $S \otimes T$, that is $\mathrm{supp}(S \otimes T) = \mathrm{supp}(S) \times \mathrm{supp}(T)$ and the support of $\psi(\tau, \lambda) = \phi(\tau + \lambda)$ is bounded, then $S * T$ is well defined. In other words, if for $\tau \in \mathrm{supp}(S)$ and $\lambda \in \mathrm{supp}(T)$ the sum $\tau + \lambda$ can only remain bounded if both τ and λ remain bounded, then the convolution product $S * T$ is well defined.

Note that this condition is sufficient but *not* necessary as shown for instance by the introductory example with integrable functions $f, g \in L^1$. In fact the convolution $f * g$ of integrable functions does always exist and is itself an integrable function

$$\begin{aligned} |\langle f * g, \phi \rangle| &= \left| \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(\tau)\, g(\lambda)\, \phi(\lambda + \tau)\, d\tau\, d\lambda \right| \\ &\leq \sup |\phi| \int_{-\infty}^{\infty} |f(\tau)|\, d\tau \int_{-\infty}^{\infty} |g(\lambda)|\, d\lambda \, . \end{aligned}$$

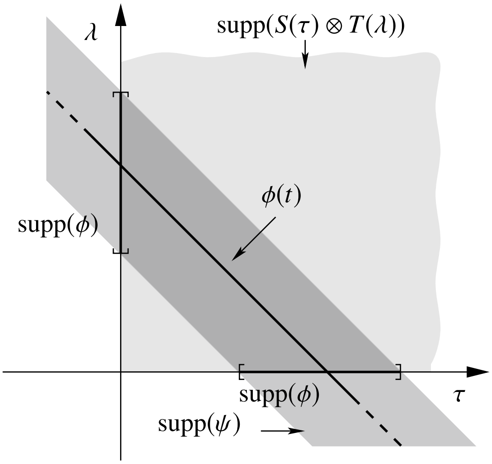


Fig. 3.2 Support of $S(\tau) \otimes T(\lambda)$ and of $\psi(\tau, \lambda) = \phi(\tau + \lambda)$

Example 3.2: One Sided Distributions

A subset of the real line $U \in \mathbb{R}$ is said to be *bounded on the left* if there is a real constant b such that $U \subset (b, \infty)$. Similarly, a subset U is called *bounded on the right* if there is a constant b such that $U \subset (-\infty, b)$.

Distributions whose support is bounded on the left (right) are called *right-sided (left-sided) distributions*. The set of all such distributions forms a vector space denoted by $\mathcal{D}'_R$ ($\mathcal{D}'_L$). Of particular interest for our purposes are right-sided distributions T with supp$(T) \in [0, \infty)$. We denote the space of all such distributions by $\mathcal{D}'_+$.

Figure 3.2 shows the support of $S(\tau) \otimes T(\lambda)$ and of $\psi(\tau, \lambda) = \phi(\tau + \lambda)$ for two distributions S and T in $\mathcal{D}'_+$. It is clear that, for any test function ϕ, their overlap is always bounded. Therefore, the convolution of right-sided or left-sided distributions is always well defined. Not so the convolution of a left-sided distribution with a right-sided one.

Example 3.3: Convolution with δ

Let T be any distribution in $\mathcal{D}'(\mathbb{R}^n)$ and δ the n dimensional Dirac pulse (see Example 3.1), then

$$\begin{aligned}\langle T * \delta, \phi\rangle &= \langle T(\tau) \otimes \delta(\lambda), \phi(\tau + \lambda)\rangle \\ &= \langle T(\tau), \langle \delta(\lambda), \phi(\tau + \lambda)\rangle\rangle \\ &= \langle T, \phi\rangle\end{aligned}$$

or

$$T * \delta = T\,. \tag{3.8}$$

Thus, δ is a *unit* of convolution.

Similarly, for any $|k|$th order derivative of the Dirac pulse

$$\begin{aligned}\langle T * D^k\delta, \phi\rangle &= \langle T(\tau)\otimes D^k\delta(\lambda), \phi(\tau+\lambda)\rangle\\ &= \langle T(\tau), \langle D^k\delta(\lambda), \phi(\tau+\lambda)\rangle\rangle\\ &= \langle T(\tau), \langle \delta(\lambda), (-1)^{|k|}D_\lambda^k\phi(\tau+\lambda)\rangle\rangle\\ &= \langle T(\tau), \langle \delta(\lambda), (-1)^{|k|}D_\tau^k\phi(\tau+\lambda)\rangle\rangle\\ &= \langle T(\tau), (-1)^{|k|}D^k\phi(\tau)\rangle\\ &= \langle D^kT(\tau), \phi(\tau)\rangle\end{aligned}$$

or

$$T * D^k\delta = D^kT \tag{3.9}$$

where D_λ^k and D_τ^k mean differentiation with respect to the variable λ and τ, respectively; and we made use of the fact that $D_\lambda^k\phi(\tau+\lambda) = D_\tau^k\phi(\tau+\lambda)$.

In our notation we use $T(t)$ to indicate a distribution to be associated with a test function whose independent variable is indicated by the symbol t. With this notation it seems natural to write $S(t) * T(t)$ to denote the convolution of two distributions. However, when we build the convolution of two shifted distributions this leads to confusion as $S(t-a) * T(t-a)$ does not represent the distribution $S * T$ shifted by a. To give a precise meaning to such expressions we introduce the *shifting operator* defined by

$$\langle \tau_a T, \phi(t)\rangle := \langle T(t), \phi(t+a)\rangle\,. \tag{3.10}$$

With it we fix the following notation

$$(S * T)(t-a) := \tau_a(S * T) \tag{3.11}$$

$$S(t-a) * T(t-b) := \tau_a S * \tau_b T\,. \tag{3.12}$$

The convolution product has several useful properties. The first one that we want to discuss is *distributivity*. If all appearing convolutions are well defined, then

$$\begin{aligned}\langle (S+T)*U,\phi\rangle &= \langle (S(\tau)+T(\tau))\otimes U(\lambda),\phi(\tau+\lambda)\rangle\\ &= \langle S(\tau)+T(\tau),\langle U(\lambda),\phi(\tau+\lambda)\rangle\rangle\\ &= \langle S(\tau),\langle U(\lambda),\phi(\tau+\lambda)\rangle\rangle +\\ &\quad\ \langle T(\tau),\langle U(\lambda),\phi(\tau+\lambda)\rangle\rangle\\ &= \langle S*U,\phi\rangle + \langle T*U,\phi\rangle\\ &= \langle S*U+T*U,\phi\rangle \end{aligned} \tag{3.13}$$

and similarly

$$S*(T+U) = S*T+S*U\,. \tag{3.14}$$

Further, *differentiation* of a convolution product is equivalent to differentiation of one of the products

$$\begin{aligned}\langle D_i(S*T),\phi\rangle &= -\langle S*T, D_i\phi\rangle\\ &= -\langle S(\tau)\otimes T(\lambda), D_{\tau,i}\phi(\tau+\lambda)\rangle\\ &= -\langle S(\tau),\langle T(\lambda), D_{\tau,i}\phi(\tau+\lambda)\rangle\rangle\\ &= \langle D_iS(\tau),\langle T(\lambda),\phi(\tau+\lambda)\rangle\rangle\\ &= \langle (D_iS)*T,\phi\rangle\end{aligned}$$

where $D_{\tau,i}$ is the partial differential operator with respect to the ith component of the variable τ. Since $D_{\tau,i}\phi(\tau+\lambda) = D_{\lambda,i}\phi(\tau+\lambda)$ differentiation can also be moved to the second factor so that

$$D_i(S*T) = (D_iS)*T = S*(D_iT)\,. \tag{3.15}$$

In a similar way one shows that the operation of *shifting* a convolution product can also be moved to one of the factors

$$(S*T)(\tau-a) = S(\tau-a)*T(\tau) = S(\tau)*T(\tau-a)\,. \tag{3.16}$$

Example 3.4: Convolution with δ

Consider two Dirac pulses and an arbitrary distribution T in $\mathcal{D}'(\mathbb{R}^n)$. By the shifting property of convolution we have

$$\delta(\tau-a)*\delta(\tau-b) = \delta(\tau-a-b) \tag{3.17}$$

$$T(\tau)*\delta(\tau-a) = T(\tau-a)\,. \tag{3.18}$$

The convolution of a distribution T with an indefinitely differentiable function γ is an indefinitely differentiable function. For, by keeping in mind that ϕ has compact support and therefore, as a distribution can be uniquely extended to functions in $\mathcal{E}$, we have

$$\begin{aligned}\langle T * \gamma, \phi\rangle &= \langle T(\tau), \langle\gamma(\lambda), \phi(\tau+\lambda)\rangle\rangle\\ &= \langle T(\tau), \langle\gamma(\lambda-\tau), \phi(\lambda)\rangle\rangle\\ &= \langle T(\tau), \langle\phi(\lambda), \gamma(\lambda-\tau)\rangle\rangle\\ &= \langle\phi(\lambda), \langle T(\tau), \gamma(\lambda-\tau)\rangle\rangle\,.\end{aligned}$$

Then, by arguments similar to the ones that led to the definition of the tensor product, one deduces that the inner distribution is an indefinitely differentiable function that we call ζ. We can therefore proceed further

$$\begin{aligned}\langle T * \gamma, \phi\rangle &= \langle\phi(\lambda), \zeta(\lambda)\rangle\\ &= \langle\zeta, \phi\rangle\end{aligned}$$

and obtain as claimed that $T * \gamma = \zeta$.

The convolution product is a *continuous* operation in the following sense. If T is a fixed convolution, $(S_m)_{m\in\mathbb{N}}$ a sequence of distributions converging in $\mathcal{D}'$ to S and all involved convolutions are well defined, then

$$\begin{aligned}\lim_{m\to\infty}\langle S_m * T, \phi\rangle &= \lim_{m\to\infty}\langle S_m(\tau), \langle T(\lambda), \phi(\tau+\lambda)\rangle\rangle\\ &= \langle S(\tau), \langle T(\lambda), \phi(\tau+\lambda)\rangle\rangle\\ &= \langle S * T, \phi\rangle\end{aligned}$$

or

$$\lim_{m\to\infty} S_m * T = S * T\,. \tag{3.19}$$

In particular we saw in Example 2.5 that δ can be represented as the limit of a sequence of test functions β_m and in Example 3.3 that δ is a unit of convolution. With continuity of convolutions we therefore deduce that *each distribution is the limit of a sequence of indefinitely differentiable functions* of the form $T * \phi$ with ϕ a test function. We saw an instance of this in Example 2.2.

The last property that we want to discuss in this section is *associativity*. In general the convolution of three or more distributions is not associative as is easily verified with simple examples.

Example 3.5: Convolution may not be Associative

Let's denote by 1 and 0 the constant functions evaluating to one and zero, respectively. Then

$$\begin{aligned}1 * (\delta^{(1)} * \mathsf{1}_+) &= 1 * \delta = 1\\ (1 * \delta^{(1)}) * \mathsf{1}_+ &= 0 * \mathsf{1}_+ = 0\\ (\delta^{(1)} * \mathsf{1}_+) * 1 &= \delta * 1 = 1\\ \delta^{(1)} * (\mathsf{1}_+ * 1) &= \text{undefined}\,.\end{aligned}$$

We can *guarantee associativity* by imposing a restriction similar to the one for the existence of the convolution of two distributions. Let's write the convolution of three distributions in terms of the tensor product

$$\langle S * T * U, \phi \rangle = \langle S(\tau) \otimes T(\lambda) \otimes U(\kappa), \phi(\tau + \lambda + \kappa) \rangle\,. \tag{3.20}$$

If the intersection of the support of $S(\tau) \otimes T(\lambda) \otimes U(\kappa)$ and the support of $\phi(\tau + \lambda + \kappa)$ is bounded, then, by the properties of the tensor product, the convolution is guaranteed to be associative. It's easily verified that the following is a *sufficient condition*: if all, but possibly one distribution have compact support, then the convolution product is associative.

Example 3.6: One-Sided Distributions

Consider three distributions $S(\tau)$, $T(\lambda)$ and $U(\kappa)$ in $\mathcal{D}'_+$. Then τ, λ and κ are ≥ 0. If the value of $\tau + \lambda + \kappa$ is bounded then there is a constant c for which $\tau + \lambda + \kappa < c$. It follows that τ is bounded by $\tau < c - (\lambda + \kappa)$ and similarly for the other variables. The convolution of distributions in $\mathcal{D}'_+$ is therefore always associative. This is also true for distributions in $\mathcal{D}'_R$ and $\mathcal{D}'_L$.

In addition, it's easily seen that $\mathcal{D}'_+$ is closed under convolution. That is, a convolution between distributions in $\mathcal{D}'_+$ results in another distribution in $\mathcal{D}'_+$.

The discussed properties of the convolution product are summarized in Table 3.1.

3.3 Approximation of Distributions

In this section we show how the convolution product can be used to obtain approximations of arbitrary distributions.

We saw that if T is a distribution in $\mathcal{D}'$ and ϕ is a test function in $\mathcal{D}$ then the convolution product $T * \phi$ is an indefinitely differentiable function. Its support is not necessarily bounded. However, let α be the test function defined by (2.11) and set $\alpha_m(\tau) = \alpha(\tau/m)$. Then for every $m \in \mathbb{N}$ the product $\alpha_m \cdot (T * \phi)$ is an indefinitely differentiable function with compact support and hence a test function.

Table 3.1 Properties of the convolution product

$S * T$	=	$T * S$
$S * (T + U)$	=	$S * T + S * U$
$(S * T) * U$	=	$S * (T * U)$
$D^k(S * T)$	=	$(D^k S) * T = S * (D^k T)$
$(S * T)(\tau - a)$	=	$S(\tau - a) * T(\tau) = S(\tau) * T(\tau - a)$
$\lim_{m\to\infty} S_m * T$	=	$S * T$
$T(\tau) * \delta(\tau - a)$	=	$T(\tau - a)$
$T * D^k\delta$	=	$D^k T$

Let (β_m) be a sequence of test functions converging to the δ distribution. Then, with the continuity of convolution, we see that in $\mathcal{D}'$

$$\lim_{m\to\infty} \alpha_m \cdot (T * \beta_m) = T \, . \tag{3.21}$$

This shows that *every distribution in $\mathcal{D}'$ is the limit of a sequence of test functions in $\mathcal{D}$*. In other words, $\mathcal{D}$ is a dense sub-vector space of $\mathcal{D}'$. Every distribution can thus be approximated to an arbitrary accuracy by a test function in $\mathcal{D}$.

Next we construct another dense sub-vector space of $\mathcal{D}'$. For simplicity we only treat the one dimensional case and for brevity we write κ_m for $\alpha_m \cdot (T * \beta_m)$. As we just discussed κ_m is a test function for every $m \in \mathbb{N}$. Let ϕ be another arbitrary test function. Then, for every m, we can find constants a and b such that the interval $[a, b]$ includes both, the support of κ_m as well as the one of ϕ. If we construct the finite sum of δ distributions weighted by k_m

$$S_{n,m} = \frac{b-a}{n} \sum_{j=1}^{n} \kappa_m(a + j\frac{b-a}{n})\delta(t - a - j\frac{b-a}{n})$$

and apply it to ϕ we obtain

$$\langle S_{n,m}, \phi \rangle = \frac{b-a}{n} \sum_{j=1}^{n} \kappa_m(a + j\frac{b-a}{n})\phi(a + j\frac{b-a}{n}) \, .$$

In the limit as n tends to infinity we obtain

$$\lim_{n\to\infty} \langle S_{n,m}, \phi \rangle = \int_a^b \kappa_m(\tau)\phi(\tau)\, d\tau \, .$$

By the choice of the interval $[a, b]$ we can extend it to the whole of $\mathbb{R}$ without changing the value of the integral. Hence, by letting m tend to infinity we finally obtain

$$\lim_{m\to\infty}\int_a^b \kappa_m(\tau)\phi(\tau)\,d\tau = \lim_{m\to\infty}\langle\kappa_m,\phi\rangle = \langle T,\phi\rangle\,.$$

We thus see that every distribution $T \in \mathcal{D}'$ is the limit of a finite sum of weighted Dirac pulses $S_n := S_{n,n}$. That is, *finite sums of weighted δ distributions form a dense sub-vector space of $\mathcal{D}'$*.

Note that a regular spacing between the δ distributions is not necessary and was chosen purely for convenience. In general any distribution can be approximated by a finite sum of the following form

$$T_n = \sum_{j=1}^{n} a_{n,j}\,\delta(t-\tau_{n,j}) \tag{3.22}$$

with $a_{n,j} \in \mathbb{C}$ and $\tau_{n,j} \in \mathbb{R}$.

3.4 Convolution of Periodic Distributions

In this section we investigate periodic distributions and their convolution. One way to define periodic distributions is to define them in a similar way as periodic functions.

Definition 3.3 (*Periodic distribution I*) A periodic distribution T is a distribution for which there exist a positive number $\mathcal{T}$ such that for all test functions ϕ

$$\langle T(\tau),\phi(\tau)\rangle = \langle T(\tau+\mathcal{T}),\phi(\tau)\rangle. \tag{3.23}$$

The smallest such number $\mathcal{T}$ is called the *fundamental period* of the distribution.

Periodic distributions have unbounded support. For this reason the convolution of two periodic distributions as defined by (3.6) does not exist. By exploiting their periodicity it is however possible to find an alternative definition for periodic distributions that allows for a well defined convolution product.

Consider a regular distribution arising from a $\mathcal{T}$-periodic function f. By exploiting its periodicity we find that

$$\begin{aligned}\langle f, \phi \rangle &= \int_{-\infty}^{\infty} f(t)\,\phi(t)\,dt \\ &= \sum_{m=-\infty}^{\infty} \int_{a+m\mathcal{T}}^{a+(m+1)\mathcal{T}} f(t)\,\phi(t)\,dt \\ &= \int_{a}^{a+\mathcal{T}} f(t) \sum_{m=-\infty}^{\infty} \phi(t - m\mathcal{T})\,dt \\ &= \int_{a}^{a+\mathcal{T}} f(t)\,\Phi(t)\,dt\end{aligned}$$

with a a constant,

$$\Phi(t) = \sum_{m=-\infty}^{\infty} \phi(t - m\mathcal{T}) \tag{3.24}$$

and where the exchange of summation and integration is justified by the fact that for every value of t the sum is finite. The function Φ is $\mathcal{T}$-periodic and indefinitely differentiable.

By introducing the identity

$$f(t) \equiv f^{\circ}([t]) \qquad (t \in \mathbb{R}) \tag{3.25}$$

with $[t]$ the equivalence class of real numbers modulo $\mathcal{T}$, we effectively and uniquely define a function f°. By writing $[t]$ as $\mathcal{T}/(2\pi)[\varphi]$ and noting that $[\varphi]$ is an equivalence class modulo 2π, we can think of f° as a function defined on a circle of radius $\mathcal{T}/(2\pi)$ at the origin of a plane, with $[\varphi]$ the polar angle. With this interpretation, the equivalence class $[t]$ is seen to represent the distance along the arc of the circle $\mathbb{T}$ from the reference $[0]$. In the following, to simplify notation, we are going to write a representative for an equivalence class.

Conversely, given a function f°, the identity (3.25) uniquely defines a periodic function f (see Fig. 3.3). The last integral above is therefore identical to the integral of $f^{\circ}\,\Phi^{\circ}$ on the circle $\mathbb{T}$

$$\int_{a}^{a+\mathcal{T}} f(t)\,\Phi(t)\,dt = \int_{\mathbb{T}} f^{\circ}(p)\,\Phi^{\circ}(p)\,dp.$$

We have thus obtained that to every regular periodic distribution f there corresponds a continuous linear functional f° on indefinitely differentiable functions Φ° on the circle $\mathbb{T}$. The set of all the latter functions is denoted by $\mathcal{D}(\mathbb{T})$. This space is isomorphic to the vector sub-space of $\mathcal{E}$ consisting of all indefinitely differen-

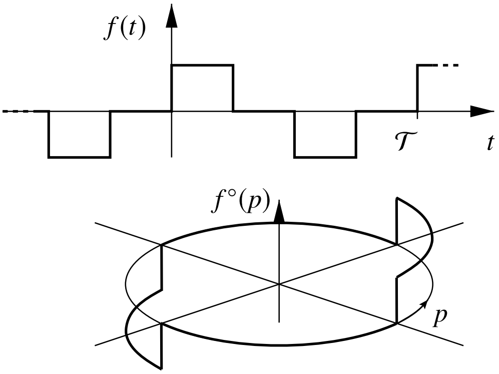


Fig. 3.3 Periodic function versus function on $\mathbb{T}$

tiable $\mathcal{T}$-periodic functions Φ and from which it inherits the following definition of convergence.

Definition 3.4 (*Convergence in* $\mathcal{D}(\mathbb{T})$) A sequence of functions $\Phi_m^{\circ} \in \mathcal{D}(\mathbb{T})$ is said to converge to $\Phi^{\circ} \in \mathcal{D}(\mathbb{T})$ if, for every natural number k, the functions $D^k \Phi_m^{\circ}$ converge uniformly to $D^k \Phi^{\circ}$.

In Sect. 3.2 we saw that every distribution T can be generated as the limit of a sequence of regular distributions f_m. Applying this result

$$\langle T, \phi \rangle = \lim_{m \to \infty} \langle f_m, \phi \rangle = \lim_{m \to \infty} \int_{\mathbb{T}} f_m^{\circ}(p)\, \Phi^{\circ}(p)\, dp = \langle T^{\circ}, \Phi^{\circ} \rangle$$

we see that not only regular, but every periodic distribution can be equivalently represented by a continuous, linear functional on $\mathcal{D}(\mathbb{T})$.

To see the converse, that is that every continuous, linear functional on $\mathcal{D}(\mathbb{T})$ represents a distribution on $\mathcal{D}(\mathbb{R})$, we have to show that every indefinitely differentiable $\mathcal{T}$-periodic function Φ can be generated by some test function ϕ as in (3.24). To this end we introduce the so-called *unitary functions*. These are test functions for which there is a number $\mathcal{T}$ such that

$$\sum_{m=-\infty}^{\infty} \xi(t - m\mathcal{T}) = 1. \tag{3.26}$$

Note that, here again, the sum is finite for every bounded range of t. We can find several such functions. The following example satisfies (3.26) for $\mathcal{T} = 1$ (see Fig. 3.4)

$$\xi_1(t) = \begin{cases} \int\limits_{|t|}^{1} e^{\frac{-1}{\tau(1-\tau)}}\, d\tau \Bigg/ \int\limits_{0}^{1} e^{\frac{-1}{\tau(1-\tau)}}\, d\tau & |t| < 1 \\ 0 & |t| \geq 1 \end{cases}$$

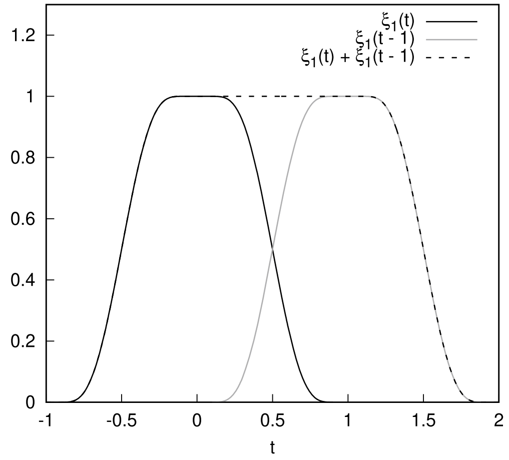


Fig. 3.4 Example unitary test function

With it we can construct unitary functions for arbitrary periods $\mathcal{T}$ by $\xi_{\mathcal{T}}(t) = \xi_1(t/\mathcal{T})$.

Now, given any $\mathcal{T}$-periodic test function Φ and an unitary function $\xi_{\mathcal{T}}$ we have

$$\begin{aligned}\Phi(t) &= \Phi(t) \sum_{m=-\infty}^{\infty} \xi_{\mathcal{T}}(t - m\mathcal{T}) \\ &= \sum_{m=-\infty}^{\infty} \xi_{\mathcal{T}}(t - m\mathcal{T})\, \Phi(t - m\mathcal{T}). \end{aligned} \tag{3.27}$$

Since $\xi_{\mathcal{T}}$ is a test function, so is $\xi_{\mathcal{T}}\, \Phi$. We have thus established that every indefinitely differentiable periodic function Φ can be represented as the sum of a test function ϕ. We conclude that periodic distributions are in one-to-one correspondence with continuous, linear functionals on $\mathcal{D}(\mathbb{T})$. It is therefore natural to call these functionals distributions on $\mathbb{T}$. They form a vector space that is denoted by $\mathcal{D}'(\mathbb{T})$.

We now have a second way to *define* periodic distributions.

Definition 3.5 (*Periodic distribution II*) A periodic distribution T is defined by

$$\langle T, \phi \rangle := \langle T^\circ, \Phi^\circ \rangle \tag{3.28}$$

with T° a distribution in $\mathcal{D}'(\mathbb{T})$ and where ϕ and Φ° are related by Eqs. (3.24) and (3.25).

This definition is compatible with the first one since replacing $\phi(t)$ by $\phi(t + \mathcal{T})$ doesn't change Φ°.

Example 3.7: Dirac comb $\delta_{\mathcal{T}}$

Consider the distribution in $\mathcal{D}'(\mathbb{T})$ defined by a Dirac pulse δ° with support consisting of the point at arc-length $p = 0$. Its value on a test function in $\mathcal{D}(\mathbb{T})$ is

$$\langle \delta^\circ(p), \Phi^\circ(p) \rangle = \Phi^\circ(0)\,.$$

The corresponding periodic distribution in $\mathcal{D}'(\mathbb{R})$ is

$$\delta_{\mathcal{T}}(t) := \sum_{m=-\infty}^{\infty} \delta(t + m\mathcal{T})$$

and evaluates to the same value as δ°

$$\begin{aligned} \langle *, \sum_{m=-\infty}^{\infty} \delta(t + m\mathcal{T}) \rangle \phi(t) &= \sum_{m=-\infty}^{\infty} \langle \delta(t + m\mathcal{T}), \xi_{\mathcal{T}}(t)\, \Phi(t) \rangle \\ &= \sum_{m=-\infty}^{\infty} \xi_{\mathcal{T}}(-m\mathcal{T})\, \Phi(-m\mathcal{T}) \\ &= \Phi^\circ(0) \sum_{m=-\infty}^{\infty} \xi_{\mathcal{T}}(-m\mathcal{T}) \\ &= \Phi^\circ(0)\,. \end{aligned}$$

Since the support of distributions in $\mathcal{D}'(\mathbb{T})$ is bounded, with the second definition, the convolution of periodic distributions is always well-defined and associative

$$\begin{aligned} \langle S * T, \phi \rangle &= \langle S(\tau) \otimes T(\lambda), \phi(\tau + \lambda) \rangle \\ &= \langle S^\circ(\tau), \langle T^\circ(\lambda), \Phi^\circ(\tau + \lambda) \rangle \rangle \\ &= \langle S^\circ * T^\circ, \Phi^\circ \rangle. \end{aligned}$$

In addition it's easily verified by replacing $\phi(t)$ by $\phi(t + \mathcal{T})$ that the resulting distribution is also $\mathcal{T}$-periodic. In other words, $\mathcal{D}'(\mathbb{T})$ is closed under convolution.

Chapter 4
Fourier Transform of Distributions

The Fourier transform is a major tool in the analysis of signals and systems. We will see that its extension to distributions will make the derivation of many results simpler and more direct than when working with functions.

4.1 Test Functions of Fast Descent

Consider a Lebesgue integrable function f. Its Fourier transform is defined by

$$\mathcal{F}\{f\}(\omega) := \hat{f}(\omega) := \int_{-\infty}^{\infty} f(t)\, \mathrm{e}^{-\jmath\omega t}\, dt \tag{4.1}$$

which is a continuous function of ω. If f is such that $\hat{f}$ is also integrable, then

$$f(t) = \mathcal{F}^{-1}\{\hat{f}\}(t) := \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\omega)\, \mathrm{e}^{\jmath\omega t}\, d\omega \tag{4.2}$$

almost everywhere, with $\mathcal{F}^{-1}\{\hat{f}\}$ the inverse Fourier transform. $\mathcal{F}^{-1}\{\hat{f}\}$ may differ from f at the points where f is not continuous.

The Fourier transform of the regular distribution f is thus

$$\begin{aligned}\left\langle \hat{f}(\omega), \phi(\omega) \right\rangle &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(t)\, \mathrm{e}^{-\jmath\omega t}\, dt\, \phi(\omega)\, d\omega \\ &= \int_{-\infty}^{\infty} f(t) \int_{-\infty}^{\infty} \phi(\omega)\, \mathrm{e}^{-\jmath\omega t}\, d\omega\, dt\end{aligned}$$

F. Beffa, *Weakly Nonlinear Systems*, Understanding Complex Systems,
https://doi.org/10.1007/978-3-031-40681-2_4

$$= \int_{-\infty}^{\infty} f(t)\hat{\phi}(t)\,dt\,. \tag{4.3}$$

The last integral looks like a distribution in a form suitable to be generalized to arbitrary distributions. However, the support of $\hat{\phi}$ is not compact. This is a manifestation of the uncertainty principle of the Fourier transform and can readily be seen by

$$\begin{aligned}\hat{\phi}(\omega) &= \int_{-\infty}^{\infty} \phi(t)\,\mathrm{e}^{-\jmath\omega t}\,dt = \int_{-a}^{b} \sum_{m=0}^{\infty} \frac{(-\jmath\omega)^m}{m!}\,t^m\,\phi(t)\,dt \\ &= \sum_{m=0}^{\infty} \frac{(-\jmath\omega)^m}{m!} \int_{-a}^{b} \phi(t)\,t^m\,dt \end{aligned} \tag{4.4}$$

with a and b constants such that the interval $[a, b]$ includes the support of ϕ.

To obtain a definition of the Fourier transform suitable for arbitrary distributions we have to replace the space of test functions $\mathcal{D}$ with a space closed under Fourier transformation. Suitable characteristics for the functions in this space can be inferred from the above expression for $\hat{\phi}$. First, given the uncertainty principle, the space has to be extended to functions of unbounded support (and therefore, the last step, the exchange of summation and integration, may not be valid). Then, if all summands have to remain finite, the limits $\lim_{t\to\pm\infty} \phi(t)\,t^m$ have to converge to zero for all values of m. Finally, to preserve arbitrary differentiability, the above characteristics must be satisfied by all derivatives of ϕ. These are the characteristics of the so-called *Schwartz space* $\mathcal{S}$ of which we give the general definition.

Definition 4.1 (*Schwartz space* $\mathcal{S}(\mathbb{R}^n)$) The Schwartz space $\mathcal{S}(\mathbb{R}^n)$ is the vector space of indefinitely differentiable functions $\phi : \mathbb{R}^n \to \mathbb{C}$ that, together with all their derivatives, decrease more rapidly than any power of $1/|\tau|$ as $|\tau| \to \pm\infty$. That is, for any n-tuples $m, k \in \mathbb{N}^n$ and $\tau \in \mathbb{R}^n$

$$\lim_{|\tau|\to\pm\infty} |\tau^m\, D^k\phi(\tau)| = 0\,. \tag{4.5}$$

Functions ϕ in the Schwartz space are called test functions of *rapid descent,* or *Schwartz functions.*

To see that the Fourier transform of a function $\phi \in \mathcal{S}(\mathbb{R})$ is indeed another function in the same space, consider the kth derivative of $\hat{\phi}$. By integrating by parts we find

$$\begin{aligned}
D^k\hat{\phi}(\omega) &= \int_{-\infty}^{\infty} (-\jmath t)^k\,\phi(t)\,\mathrm{e}^{-\jmath\omega t}\,dt\\
&= \frac{1}{\jmath\omega}\int_{-\infty}^{\infty} \mathrm{e}^{-\jmath\omega t} D\left[(-\jmath t)^k\,\phi(t)\right] dt
\end{aligned}$$

and by iterating m times

$$\begin{aligned}
\left|(\jmath\omega)^m\,D^k\hat{\phi}(\omega)\right| &= \left|\int_{-\infty}^{\infty} \mathrm{e}^{-\jmath\omega t} D^m\left[(-\jmath t)^k\,\phi(t)\right] dt\right|\\
&\le \int_{-\infty}^{\infty} \left|D^m\left[t^k\,\phi(t)\right]\right| dt\,.
\end{aligned}$$

Since this is valid for arbitrary k and m it shows that $\hat{\phi}$ is in fact a function in the Schwartz space. In addition, given that the Fourier transform, and its inverse are almost symmetric, a similar calculation shows that the inverse Fourier transform of a Schwartz function $\hat{\phi}$ is a function $\phi \in \mathcal{S}$. That is, *the Fourier transform is a bijection from the space $\mathcal{S}$ into itself.*

Example 4.1: Gauss Function

An important example of a function of rapid descent is the Gauss function

$$\phi(t) = \frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-t^2/(2\sigma^2)}\,.$$

It's widely known that its Fourier transform is

$$\hat{\phi}(\omega) = \mathrm{e}^{-\omega^2\sigma^2/2}\,.$$

One of the defining characteristics of distributions is their continuity. To talk about continuity we introduce a convergence principle (topology) similar to the ones we defined for $\mathcal{D}$ and $\mathcal{E}$.

Definition 4.2 (*Convergence in* $\mathcal{S}(\mathbb{R}^n)$) A sequence of functions $\phi_m \in \mathcal{S}(\mathbb{R}^n)$ is said to converge in $\mathcal{S}(\mathbb{R}^n)$ to a function $\phi \in \mathcal{S}(\mathbb{R}^n)$, if for each n-tuples $k, p \in \mathbb{N}^n$ and $\tau \in \mathbb{R}^n$ the sequence $|\tau|^p\,D^k\phi_m(\tau)$ converges uniformly to $|\tau|^p\,D^k\phi(\tau)$, that is if

$$\lim_{m\to\infty} |\tau|^p\,D^k\phi_m(\tau) = |\tau|^p\,D^k\phi(\tau)\,.$$

4.2 Fourier Transform of Tempered Distributions

For (4.3) to be an expression suitable for the definition of the Fourier transform for an arbitrary distribution, we must verify its linearity and continuity. The former is clear. To show the latter we have to verify that, if a sequence of test functions $\phi_m \in \mathcal{S}$ converges to zero, so does the sequence of their Fourier transforms $\hat{\phi}_m$. That this is the case is shown by the following upper bound

$$\begin{aligned}
|\hat{\phi}_m(\omega)| &= \left|\int_{-\infty}^{\infty} \phi_m(t)\,\mathrm{e}^{-\jmath\omega t}\,dt\right| \le \int_{|t|<1} |\phi_m(t)|\,dt + \int_{|t|\ge 1} |\phi_m(t)|\,dt \\
&\le 2\sup_{|t|<1} |\phi_m(t)| + \int_{|t|\ge 1} |\frac{\phi_m(t)\,t^2}{t^2}|\,dt \\
&\le 2\sup_{|t|<1} |\phi_m(t)| + \sup_{|t|\ge 1} |\phi_m(t)\,t^2| \int_{|t|\ge 1} \frac{1}{|t^2|}\,dt \\
&= 2\sup_{|t|<1} |\phi_m(t)| + 2\sup_{|t|\ge 1} |\phi_m(t)\,t^2|\,.
\end{aligned}$$

We have a good candidate for the definition of the Fourier transform for an arbitrary distribution. However, since the space $\mathcal{S}$ is larger than $\mathcal{D}$, the Fourier transform can only be defined for the following subset of distributions.

Definition 4.3 (*Tempered distributions*) *Tempered distributions* (also called distributions of *slow growth*) are distributions that can be extended to continuous, linear functionals on the Schwartz space $\mathcal{S}$.

The space of all continuous, linear functionals on $\mathcal{S}$ is denoted by $\mathcal{S}'$ and, since the Schwartz space $\mathcal{S}$ is a subspace of $\mathcal{E}$, we have the following inclusion: $\mathcal{E}' \subset \mathcal{S}' \subset \mathcal{D}'$. Consequently, from Sect. 2.5, we conclude that, if a distribution $T \in \mathcal{D}'$ can be extended to a continuous, linear functional on $\mathcal{S}$, then this extension is unique (and the other way around). $\mathcal{S}'$ can therefore be *identified* with tempered distributions.

Example 4.2: Slowly Increasing Function

Consider a locally integrable function f satisfying

$$|f(t)| \le C\,|t|^m \qquad \text{as} \quad |t| \to \infty$$

for some constant C and some natural number m. Then, f is a tempered distribution, since

$$\begin{aligned}
|\langle f, \phi\rangle| &\le \int_{|t|<1} |f(t)\,\phi(t)|\,dt + \int_{|t|\ge 1} |f(t)\,\phi(t)|\,dt \\
&\le \sup_{|t|<1} |\phi(t)| \int_{|t|<1} |f(t)|dt + \sup_{|t|\ge 1} \left(|t^{m+2}|\,|\phi(t)|\right) \int_{|t|\ge 1} \frac{C}{|t|^2} dt
\end{aligned}$$

is bounded for every $\phi \in \mathcal{S}$.

Example 4.3: Distributions in $\mathcal{E}'$

Distribution with bounded support are defined on all indefinitely differentiable function, independently of their asymptotic behavior as $t \to \infty$. For this reason the Fourier transform of distributions in $\mathcal{E}'$ is always well-defined.

Example 4.4: Multiplication with Polynomial

If T is a tempered distribution and p a polynomial, then $p\,T$ is a tempered distribution. $p\,T$ is in fact defined as

$$\langle p\,T, \phi\rangle = \langle T, p\,\phi\rangle$$

and it's easy to see that $p\,\phi \in \mathcal{S}$.

We can finally define the Fourier transform for tempered distributions.

Definition 4.4 (*Fourier transform on $\mathcal{S}'$*) The Fourier transform of a tempered distribution T and its inverse, are defined by

$$\langle \mathcal{F}\{T\}, \phi\rangle := \langle T, \mathcal{F}\{\phi\}\rangle \tag{4.6}$$

$$\left\langle \mathcal{F}^{-1}\{T\}, \phi\right\rangle := \left\langle T, \mathcal{F}^{-1}\{\phi\}\right\rangle \tag{4.7}$$

for every function $\phi \in \mathcal{S}$.

Clearly, the Fourier transform of a tempered distribution is a tempered distribution. Note that, given the properties of Schwartz functions, for a tempered distribution it's always the case that

$$\mathcal{F}^{-1}\{\mathcal{F}\{T\}\} = \mathcal{F}\left\{\mathcal{F}^{-1}\{T\}\right\} = T\,.$$

In addition the Fourier transform and its inverse satisfy the following *symmetry* relation

$$\begin{aligned} \langle \mathcal{F}\{T\}, \phi\rangle &= \langle T, \mathcal{F}\{\phi\}\rangle = \left\langle T(\omega), \int_{-\infty}^{\infty} \phi(t)\mathrm{e}^{-\jmath\omega t}dt \right\rangle \\ &= \left\langle T(\omega), 2\pi\mathcal{F}^{-1}\{\phi\}(-\omega)\right\rangle \\ &= \left\langle 2\pi\,\mathcal{F}^{-1}\{T(-\omega)\}, \phi\right\rangle\,. \end{aligned}$$

If in this expression we replace T by its Fourier transform and denote it by $\hat{T}$, then this symmetry relation can also be expressed as

$$\mathcal{F}\left\{\hat{T}(t)\right\} = 2\pi\, T(-\omega)\,. \tag{4.8}$$

As with functions, we will often use the convention of denoting by $\hat{T}$ the Fourier transform of a tempered distribution T.

Example 4.5: Fourier Transform and δ

The Fourier transform of the delta distribution δ is

$$\left\langle\hat{\delta},\phi\right\rangle = \left\langle\delta,\hat{\phi}\right\rangle = \hat{\phi}(0) = \int_{-\infty}^{\infty}\phi(t)dt = \langle 1,\phi\rangle$$

or

$$\hat{\delta} = 1\,.$$

Conversely, the Fourier transform of the constant function 1 is

$$\left\langle\hat{1},\phi\right\rangle = \left\langle 1,\hat{\phi}\right\rangle = \int_{-\infty}^{\infty}\hat{\phi}(\omega)\,d\omega = 2\pi\,\left\langle\delta,\mathcal{F}^{-1}\{\hat{\phi}\}\right\rangle = 2\pi\,\,\langle\delta,\phi\rangle$$

or

$$\hat{1} = 2\pi\,\delta\,.$$

This expression is often found in the technical literature symbolically written as

$$\delta(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\mathrm{e}^{-\jmath\omega t}\,d\omega = \frac{1}{2\pi}\int_{-\infty}^{\infty}\mathrm{e}^{\jmath\omega t}\,d\omega\,.$$

The Fourier transform of the derivative of δ is

$$\langle\mathcal{F}\{D\delta\}\,,\phi\rangle = \left\langle D\delta,\hat{\phi}\right\rangle = -\left\langle\delta,-\jmath\omega\hat{\phi}\right\rangle = \jmath\omega\left\langle\hat{\delta},\phi\right\rangle = \langle\jmath\omega,\phi\rangle$$

and, by iterating this procedure, for the higher order derivatives we find

$$\mathcal{F}\left\{D^k\delta\right\} = (\jmath\omega)^k\,.$$

Example 4.6: Complex Tones

The Fourier transform of a complex tone is

$$\begin{aligned}\left\langle \mathcal{F}\left\{ \mathrm{e}^{\jmath\omega_c t}\right\}, \phi\right\rangle &= \left\langle \mathrm{e}^{\jmath\omega_c t}, \hat{\phi}\right\rangle = \int\limits_{-\infty}^{\infty} \mathrm{e}^{\jmath\omega_c t}\hat{\phi}(t)\,dt = 2\pi\;\phi(\omega_c)\\ &= 2\pi\;\langle\delta(\omega-\omega_c), \phi\rangle\end{aligned}$$

or

$$\mathcal{F}\left\{\mathrm{e}^{\jmath\omega_c t}\right\} = 2\pi\;\delta(\omega-\omega_c)\,.$$

Similarly, the Fourier transform of a shifted Dirac pulse is found to be

$$\mathcal{F}\left\{\delta(t-\tau_o)\right\} = \mathrm{e}^{-\jmath\omega\tau_0}\,.$$

Example 4.7: Dirac comb

An equally spaced sequence of Dirac pulses is a tempered distribution called a Dirac comb with period $\mathcal{T}$

$$\delta_{\mathcal{T}}(t) := \sum_{m=-\infty}^{\infty} \delta(t-m\mathcal{T})\,.$$

The linearity and continuity of distributions permit to calculate its Fourier transform term by term and, by using previous results, we obtain

$$\mathcal{F}\left\{\delta_{\mathcal{T}}\right\} = \sum_{m=-\infty}^{\infty} \mathrm{e}^{\jmath\omega m\mathcal{T}}\,.$$

This distribution is formally the limit

$$\lim_{K,M\to\infty} \left\langle s_{P,K}(\omega) + s_{N,M}(\omega) - 1, \phi(\omega)\right\rangle$$

with

$$s_{P,K}(\omega) := \sum_{m=0}^{K-1} \mathrm{e}^{\jmath\omega m\mathcal{T}}$$

$$s_{N,M}(\omega) := \sum_{m=0}^{M-1} \mathrm{e}^{-\jmath\omega m\mathcal{T}}.$$

For values of $\omega \neq k\,2\pi/\mathcal{T}$, $k \in \mathbb{Z}$ the partial sums can be represented by

$$\begin{aligned} s_{P,K}(\omega) &= \frac{1-\mathrm{e}^{\jmath\omega K\mathcal{T}}}{1-\mathrm{e}^{\jmath\omega\mathcal{T}}} = \frac{1}{1-\mathrm{e}^{\jmath\omega\mathcal{T}}} - \frac{\mathrm{e}^{\jmath\omega K\mathcal{T}}}{1-\mathrm{e}^{\jmath\omega\mathcal{T}}} \\ s_{N,M}(\omega) &= \frac{1-\mathrm{e}^{-\jmath\omega M\mathcal{T}}}{1-\mathrm{e}^{-\jmath\omega\mathcal{T}}} = \frac{1}{1-\mathrm{e}^{-\jmath\omega\mathcal{T}}} - \frac{\mathrm{e}^{-\jmath\omega M\mathcal{T}}}{1-\mathrm{e}^{-\jmath\omega\mathcal{T}}}\,. \end{aligned}$$

The sum of the first terms is easily seen to equal 1 and, with the results of Example 2.11, the limit of the second ones do vanish. The support of $\mathcal{F}\{\delta_{\mathcal{T}}\}$ therefore consists in the set of points $\omega = k\,2\pi/\mathcal{T}$, $k \in \mathbb{Z}$. Consequently, when applied to any test function $\phi \in \mathcal{S}$, its value must be a weighted sum of the values of the test function at these points. Since, replacing $\phi(t)$ by $\phi(t+\mathcal{T})$ doesn't change the result, we can also deduce that the weighting factor must be the same for all terms. We thus have

$$\langle \mathcal{F}\{\delta_{\mathcal{T}}\}, \phi\rangle = \sum_{m=-\infty}^{\infty} C\phi(m\omega_c) = C\left\langle \delta_{\omega_c}, \phi\right\rangle$$

with C a constant and $\omega_c = 2\pi/\mathcal{T}$. The value of the constant can be found by inserting any Schwartz function. A convenient choice is the one of Example 4.1 with $\sigma = \sqrt{2\pi}/T$. With it, on the one hand we have

$$\langle \mathcal{F}\{\delta_{\mathcal{T}}\}, \phi\rangle = C\left\langle \delta_{\omega_c} \frac{\mathcal{T}}{2\pi}\mathrm{e}^{-(tT)^2/(4\pi)}, =\right\rangle C\frac{\mathcal{T}}{2\pi}\sum_{m=-\infty}^{\infty} \mathrm{e}^{-m^2\pi}$$

and on the other hand

$$\langle \mathcal{F}\{\delta_{\mathcal{T}}\}, \phi\rangle = \left\langle \delta_{\mathcal{T}}, \hat{\phi}\right\rangle = \left\langle \delta_{\mathcal{T}}, \mathrm{e}^{-(t/T)^2\pi}\right\rangle = \sum_{m=-\infty}^{\infty} \mathrm{e}^{-m^2\pi}$$

so that $C = 2\pi/T$. We have thus established the following important result

$$\mathcal{F}\{\delta_{\mathcal{T}}\} = \omega_c\,\delta_{\omega_c}\,. \tag{4.9}$$

The Fourier transforms of the δ and related distributions are summarized in Table 4.1.

A useful property of the Fourier transform is that it *preserves parity*. This means that the Fourier transform of an odd tempered distribution T is odd

Table 4.1 Fourier transformation of the δ and related distributions

T	$\mathcal{F}\{T\}$
δ	1
1	$2\pi\delta$
$D^k\delta$	$(\jmath\omega)^k$
$(-\jmath t)^k$	$2\pi\, D^k\delta$
$\delta(t-\tau_c)$	$\mathrm{e}^{-\jmath\omega\tau_c}$
$\mathrm{e}^{\jmath\omega_c t}$	$2\pi\delta(\omega-\omega_c)$
$\delta_{\mathcal{T}}$	$\frac{2\pi}{\mathcal{T}}\delta_{2\pi/\mathcal{T}}$

$$\begin{aligned}\langle\mathcal{F}\{T\},\phi(-t)\rangle &= \langle T,\mathcal{F}\{\phi(-t)\}\rangle = \left\langle T,\int_{\mathbb{R}}\phi(-t)\,\mathrm{e}^{-\jmath\omega t}dt\right\rangle\\ &= \left\langle T,\int_{\mathbb{R}}\phi(t)\,\mathrm{e}^{\jmath\omega t}dt\right\rangle\\ &= \left\langle T,\hat{\phi}(-\omega)\right\rangle = -\left\langle T,\hat{\phi}(\omega)\right\rangle\\ &= -\langle\mathcal{F}\{T\},\phi(t)\rangle\end{aligned}$$

and, similarly, the Fourier transform of an even tempered distribution is even.

We conclude this section with an important property of the Fourier transform of real distributions. Let T be a real distribution, ϕ a real valued Schwartz function and let denote complex conjugation by an over bar. Then

$$\begin{aligned}\left\langle\overline{\hat{T}},\phi\right\rangle = \overline{\left\langle\hat{T},\phi\right\rangle} = \overline{\left\langle T,\hat{\phi}\right\rangle} = \left\langle T,\overline{\hat{\phi}}\right\rangle\\ = \left\langle T,\hat{\phi}(-\omega)\right\rangle = \left\langle\hat{T}(-\omega),\phi\right\rangle\end{aligned}$$

or

$$\overline{\hat{T}}(\omega) = \hat{T}(-\omega) \tag{4.10}$$

as for real functions.

4.3 Distributions with Bounded Support

The Fourier transform of distributions with bounded support can be expressed in a simpler, useful form that we explore in this section. To this end, consider first the convolution between a distribution of bounded support T and the regular constant distribution 1

$$\begin{aligned}\langle 1 * T, \phi\rangle &= \langle T, \langle 1, \phi\rangle\rangle = \left\langle T(\tau), \int_{\mathbb{R}} \phi(\tau+\lambda)\, d\lambda\right\rangle \\ &= \langle 1, \langle T, \phi\rangle\rangle = \int_{\mathbb{R}} \langle T(\tau), \phi(\tau+\lambda)\rangle\, d\lambda\end{aligned}$$

and note the two equivalent integral representations. With these equalities we can then proceed to represent the Fourier transform of T by

$$\begin{aligned}\left\langle \hat{T}, \phi\right\rangle &= \left\langle T(t), \int_{\mathbb{R}} \phi(\omega) \mathrm{e}^{-\jmath t\omega}\, d\omega\right\rangle \\ &= \int_{\mathbb{R}} \left\langle T(t), \phi(\omega)\mathrm{e}^{-\jmath t\omega}\right\rangle d\omega \\ &= \int_{\mathbb{R}} \left\langle T(t), \mathrm{e}^{-\jmath\omega t}\right\rangle \phi(\omega)\, d\omega \\ &= \left\langle \left\langle T(t), \mathrm{e}^{-\jmath\omega t}\right\rangle, \phi(\omega)\right\rangle\end{aligned}$$

with

$$\hat{T}(\omega) = \left\langle T(t), \mathrm{e}^{-\jmath\omega t}\right\rangle \tag{4.11}$$

an indefinitely differentiable *function* of slow growth. In a similar way we obtain

$$\mathcal{F}^{-1}\{T\} = \frac{1}{2\pi}\left\langle T(\omega), \mathrm{e}^{\jmath\omega t}\right\rangle . \tag{4.12}$$

4.4 Fourier Transform and Convolution

Consider a tempered distribution S and a distribution with bounded support T. Their convolution is well-defined if, for every test function $\phi \in \mathcal{S}$

$$\langle S * T, \phi\rangle = \langle S(\tau), \langle T(\lambda), \phi(\tau+\lambda)\rangle\rangle = \langle T(\lambda), \langle S(\tau), \phi(\tau+\lambda)\rangle\rangle\ .$$

In the first case,

$$\langle T(\lambda), \phi(\tau+\lambda)\rangle$$

is a function $\zeta(\tau) \in \mathcal{S}$ and the outer functional is therefore well-defined. In the second case,

$$\langle S(\lambda), \phi(\tau+\lambda)\rangle$$

is an indefinitely differentiable function $\gamma(\tau) \in \mathcal{E}$. Consequently, the outer functional is again well-defined. Equality of the two expressions is guaranteed by the uniqueness

Table 4.2 Properties of the Fourier transformation

T	$\mathcal{F}\{T\}$
$\sum_m a_m T_m$	$\sum_m a_m \hat{T}_m$
$S * T$	$\hat{S}\,\hat{T}$
$S\,T$	$\frac{1}{2\pi}\hat{S} * \hat{T}$
$T(t-\tau)$	$\hat{T}e^{-\jmath\omega\tau}$
$Te^{\jmath\omega_c t}$	$T(\omega-\omega_c)$
$D^k T$	$(\jmath\omega)^k\,\hat{T}$
$(-\jmath t)^k\,T$	$D^k\hat{T}$
$\hat{T}$	$2\pi\,T(-\omega)$

of the extension of the distributions T and S to $\mathcal{E}'$ and $\mathcal{S}'$ respectively (and the other way around).

We now show that the Fourier transform of the convolution of S and T is the product of their Fourier transforms $\hat{S}$ and $\hat{T}$:

$$\begin{aligned}
\langle \mathcal{F}\{S * T\}, \phi\rangle &= \langle S * T, \mathcal{F}\{\phi\}\rangle = \langle S, \langle T, \mathcal{F}\{\phi\}\rangle\rangle \\
&= \left\langle S(\omega), \left\langle T(\lambda), \int_{\mathbb{R}} \phi(t)\mathrm{e}^{-\jmath(\omega+\lambda)t}\,dt\right\rangle\right\rangle \\
&= \left\langle S(\omega), \int_{\mathbb{R}} \phi(t)\left\langle T(\lambda), \mathrm{e}^{-\jmath\lambda t}\right\rangle \mathrm{e}^{-\jmath\omega t}\,dt\right\rangle \\
&= \left\langle \mathcal{F}\{S(\omega)\}, \phi(t)\left\langle T(\lambda), \mathrm{e}^{-\jmath\lambda t}\right\rangle\right\rangle \\
&= \left\langle \mathcal{F}\{S(\omega)\}\left\langle T(\lambda), \mathrm{e}^{-\jmath\lambda t}\right\rangle, \phi(t)\right\rangle \\
&= \left\langle \hat{S}\,\hat{T}, \phi\right\rangle
\end{aligned}$$

or

$$\mathcal{F}\{S * T\} = \hat{S}\,\hat{T}\,. \tag{4.13}$$

The product is well-defined since $\hat{T}$ is an indefinitely differentiable function of slow growth. A similar result is readily obtained for the inverse Fourier transform

$$\mathcal{F}^{-1}\{S * T\} = 2\pi\,\mathcal{F}^{-1}\{S\}\mathcal{F}^{-1}\{T\}\,. \tag{4.14}$$

These are central results and arguably the most important properties of the Fourier transformation. It can be shown that this relation is valid in other cases as well. For example, in the case of locally integrable functions which are slowly increasing [20].

With these properties, the previously obtained Fourier transforms for the Dirac δ distribution and the properties of the convolution product we immediately obtain the properties listed in Table 4.2. In particular, it's noteworthy the fact that *the Dirac δ distribution acting as a unit with respect to the convolution product is related to the fact that its Fourier transform is 1.*

Example 4.8: Fourier Transform of pv $1/t$

In Example 2.13 we saw that the equation

$$t\,T = 1$$

has solutions

$$T = \mathrm{pv}\,\frac{1}{t} + C\delta.$$

with C a constant. By noting that pv $1/t$ is odd, while δ is even, we can find the Fourier transform of the former. First observe that

$$\mathcal{F}\left\{(-\jmath t)\,T\right\} = D\hat{T}$$

and hence, by transforming both sides of the equation we have that

$$\jmath\, D\hat{T} = 2\pi\,\delta\,.$$

Since the Fourier transform preserves parity, we have to look for an odd solution of this equation, and we find

$$\hat{T}(\omega) = -\jmath\pi\,\mathrm{sign}(\omega) \qquad \text{or} \qquad \mathcal{F}\left\{\mathrm{pv}\,\frac{\jmath}{\pi\,t}\right\}(\omega) = \mathrm{sign}(\omega)\,.$$

With this result and the symmetry of the Fourier transform (Eq. (4.8)) we also find

$$\mathcal{F}\left\{\mathrm{sign}\right\}(\omega) = 2\pi\,\mathrm{pv}\,\frac{\jmath}{\pi\,(-\omega)} = \mathrm{pv}\,\frac{2}{\jmath\omega}\,.$$

Example 4.9: Fourier transform of 1_+

The Heaviside step function 1_+ can be written as

$$1_+(t) = \frac{1}{2}\left[1 + \mathrm{sign}(t)\right]\,.$$

Its Fourier transform is therefore

$$\mathcal{F}\left\{1_+(t)\right\} = \frac{1}{2}\left[2\pi\,\delta(\omega) + \mathrm{pv}\,\frac{2}{\jmath\omega}\right] = \pi\,\delta + \mathrm{pv}\,\frac{1}{\jmath\omega}\,.$$

From the symmetry of the Fourier transform we also obtain

$$\mathcal{F}\left\{\pi\,\delta + \mathrm{pv}\,\frac{1}{\jmath t}\right\} = 2\pi\,\mathbb{1}_{+}(-\omega)$$

or

$$\mathcal{F}\left\{\frac{1}{2}\delta + \mathrm{pv}\,\frac{\jmath}{2\pi\,t}\right\} = \mathbb{1}_{+}(\omega)\,.$$

4.5 Periodic Distributions

In this section we investigate the Fourier transform of periodic distributions. Consider first a regular distribution arising from a locally integrable periodic function f. If we introduce a function $f_{\sqcap}$

$$f_{\sqcap}(t) = \begin{cases} f(t) & a \le t < a + \mathcal{T} \\ 0 & \text{otherwise} \end{cases}$$

with a a constant, then f can be expressed as a convolution product

$$f(t) = f_{\sqcap}(t) * \delta_{\mathcal{T}}\,. \tag{4.15}$$

The Fourier transform of f can therefore be written as the product of the transforms of $f_{\sqcap}$ and $\delta_{\mathcal{T}}$, which is well-defined since $f_{\sqcap}$ has compact support

$$\mathcal{F}\{f\} = \left\langle f_{\sqcap}, \mathrm{e}^{-\jmath\omega t}\right\rangle \omega_c \delta_{\omega_c} = \frac{2\pi}{\mathcal{T}} \sum_{m=-\infty}^{\infty} \left\langle f_{\sqcap}, \mathrm{e}^{-\jmath m\omega_c t}\right\rangle \delta(\omega - m\omega_c)\,.$$

From this we see that the Fourier transform of f consists of a train of equally spaced Dirac pulses, each weighted by a numerical coefficient, and that this set of weighting numbers fully characterize it.

If we now represent f as the inverse Fourier transform of $\mathcal{F}\{f\}$ and make use of the results of Example 4.6, we obtain a trigonometric series

$$\begin{aligned} f(t) &= \frac{2\pi}{\mathcal{T}} \sum_{m=-\infty}^{\infty} \left\langle f_{\sqcap}, \mathrm{e}^{-\jmath m\omega_c t}\right\rangle \mathcal{F}^{-1}\{\delta(\omega - m\omega_c)\} \\ &= \frac{1}{\mathcal{T}} \sum_{m=-\infty}^{\infty} \left\langle f_{\sqcap}, \mathrm{e}^{-\jmath m\omega_c t}\right\rangle \mathrm{e}^{\jmath m\omega_c t}\,. \end{aligned} \tag{4.16}$$

called the Fourier series of f. The coefficients are values obtained by evaluating $f_{\sqcap}$ on indefinitely differentiable periodic functions which are members of $\mathcal{D}(\mathbb{T})$ and the values are identical to the ones obtained by evaluating the distribution $f^\circ \in \mathcal{D}'(\mathbb{T})$ corresponding to f (see Sect. 3.4) on the same functions. Consequently, the above trigonometric series is both a representation of a periodic distribution in $\mathcal{D}'(\mathbb{R})$ as well as that of a distribution in $\mathcal{D}'(\mathbb{T})$.

These arguments can be extended to general periodic distributions without any difficulty so that we have the following general definition of the Fourier series of a periodic distribution.

Definition 4.5 (*Fourier Series*) The *Fourier series* of a distribution $T^\circ \in \mathcal{D}'(\mathbb{T})$, or a periodic distribution $T \in \mathcal{D}'(\mathbb{R})$, is the trigonometric series

$$\sum_{m=-\infty}^{\infty} c_m \mathrm{e}^{\jmath m\omega_c p} \tag{4.17}$$

with coefficients

$$c_m = \frac{1}{\mathcal{T}} \left\langle T^\circ, \mathrm{e}^{-\jmath m\omega_c p} \right\rangle . \tag{4.18}$$

The coefficients are called the Fourier coefficients of the series.

The Fourier series is the only trigonometric series that converges to the distribution T° in $\mathcal{D}'(\mathbb{T})$. In fact, if for any $\Phi \in \mathcal{D}(\mathbb{T})$ the series

$$\sum_{m=-\infty}^{\infty} d_m \left\langle \mathrm{e}^{\jmath m\omega_c p}, \Phi \right\rangle$$

does converge, then by putting $\Phi = \mathrm{e}^{-\jmath m\omega_c p}$ and using the orthogonality of trigonometric functions we find that

$$\left\langle T^\circ, \mathrm{e}^{-\jmath m\omega_c p} \right\rangle = \mathcal{T} d_m$$

which shows that the coefficients d_m correspond to the Fourier coefficients of T°.

As every distribution, the Fourier series of a distribution can be differentiated term by term. Therefore, if we designate by $c_m(T^\circ)$ the mth Fourier coefficient of the distribution T°, we have that

$$c_m(D^k T^\circ) = (\jmath m\omega_c)^k c_m(T^\circ) . \tag{4.19}$$

A natural question to ask is: How do we know if a certain trigonometric series converges to a periodic distribution? To answer this question first note that the series of numbers

$$\sum_{m=1}^{\infty} \frac{1}{m^2}$$

is absolutely convergent. Therefore, if the magnitude of the coefficients $|c_m|$, as $m \to \infty$, are bounded above by $C/|m|^2$, with C a constant, then the series converges to a continuous function f and hence to a distribution. But distributions are always differentiable term by term an arbitrary number of times. Using (4.19) we therefore conclude that, if the magnitude of the coefficients of the series, as $m \to \infty$ are bound by $C|m|^k$ for some number $k \geq 0$ and a constant C, then the series converges to a distribution.

We derived the Fourier series starting from the Fourier transform and its property that converts convolution into a product. We therefore expect a similar property for the Fourier series. Consider the convolution of two distributions S° and T° with the same period $\mathcal{T}$. The Fourier coefficients of the resulting series are

$$\begin{aligned} c_m(S^\circ * T^\circ) &= \frac{1}{\mathcal{T}} \left\langle S^\circ * T^\circ, \mathrm{e}^{-\jmath m\omega_c t} \right\rangle \\ &= \frac{1}{\mathcal{T}} \left\langle S^\circ(t) \otimes T^\circ(\lambda), \mathrm{e}^{-\jmath m\omega_c (t+\lambda)} \right\rangle \\ &= \frac{1}{\mathcal{T}} \left\langle S^\circ(t), \mathrm{e}^{-\jmath m\omega_c t} \right\rangle \left\langle T^\circ(t), \mathrm{e}^{-\jmath m\omega_c \lambda} \right\rangle \\ &= \mathcal{T}\, c_m(S^\circ)\, c_m(T^\circ)\,. \end{aligned} \tag{4.20}$$

Consequently the Fourier series of the convolution of S° and T° is

$$S^\circ * T^\circ = \mathcal{T} \sum_{m=-\infty}^{\infty} c_m(S^\circ)\, c_m(T^\circ) \mathrm{e}^{\jmath m\omega_c t} \tag{4.21}$$

and, indeed we see that the Fourier series representation of periodic distributions transforms convolutions into products.

Example 4.10: Fourier series of $\delta_{\mathcal{T}}$

The mth Fourier coefficient of the Dirac comb $\delta_{\mathcal{T}}$ is

$$c_m(\delta_{\mathcal{T}}) = \frac{1}{\mathcal{T}} \left\langle \delta^\circ, \mathrm{e}^{-\jmath m\omega_c t} \right\rangle = \frac{1}{\mathcal{T}}$$

with $\omega_c = 2\pi/\mathcal{T}$. Hence, its Fourier series is

$$\delta_{\mathcal{T}} = \sum_{m=-\infty}^{\infty} \frac{1}{\mathcal{T}}\, \mathrm{e}^{\jmath m\omega_c t}\,.$$

If we now compute the convolution of $\delta_{\mathcal{T}}$ with another $\mathcal{T}$-periodic distribution T with Fourier coefficients $c_m(T)$, from (4.21) we see that, as expected, $\delta_{\mathcal{T}}$ act as a unit

$$c_m(T * \delta_{\mathcal{T}}) = \mathcal{T} c_m(\delta_{\mathcal{T}})\, c_m(T) = c_m(T)\,.$$

A periodic distribution can be represented as a convolution product between the Dirac comb $\delta_{\mathcal{T}}$ and a distribution different from the one of (4.15). For example, with $\xi_{\mathcal{T}}$ any unitary function we have

$$\begin{aligned} T &= T \sum_{m=-\infty}^{\infty} \xi_{\mathcal{T}}(t - m\mathcal{T}) \\ &= \sum_{m=-\infty}^{\infty} T(t - m\mathcal{T})\, \xi_{\mathcal{T}}(t - m\mathcal{T}) \\ &=: \sum_{m=-\infty}^{\infty} S(t - m\mathcal{T}) = S * \delta_{\mathcal{T}} \end{aligned} \tag{4.22}$$

which defines a distribution S whose support is finite and larger than a single period of T. Using this representation we can express the Fourier coefficients and the Fourier transform of T in terms of the one of S as

$$\hat{T} = \omega_c\, \hat{S}\, \delta_{\omega_c} = \frac{2\pi}{\mathcal{T}} \sum_{m=-\infty}^{\infty} \hat{S}(\omega)\, \delta(\omega - m\omega_c) \tag{4.23}$$

$$c_m(T) = \frac{\hat{S}(m\omega_c)}{\mathcal{T}}\,. \tag{4.24}$$

For this reason, if in some calculation we obtain the Fourier transform of a signal in this form, with $\hat{S}$ the transform of a known non-periodic distribution, then we can immediately write T in terms of S as in (4.22).

We close this section with a property that is the counterpart of (4.10) for the Fourier coefficients of a real periodic distribution

$$\bar{c}_m = c_{-m}\,. \tag{4.25}$$

4.6 Extension to Several Variables

The Fourier transform can be extended to functions of several variables by transforming each variable individually. That is, if f is an integrable function on $\mathbb{R}^n$, then we can apply the one-dimensional Fourier transform to each variable individually, keeping the other ones constant. After performing this operation with respect to each variable in turns, we obtain the following expression which defines of the n-dimensional Fourier transform

$$\begin{aligned}\hat{f}(\omega_1,\dots,\omega_n) &:= \mathcal{F}\{f\}(\omega_1,\dots,\omega_n)\\ &:= \int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty} f(\tau_1,\dots,\tau_n)\,\mathrm{e}^{-\jmath(\omega_1\tau_1+\dots+\omega_n\tau_n)}\,d\tau_1\dots d\tau_n\,.\end{aligned}$$

To shorten the notation we will write

$$\hat{f}(\omega) = \mathcal{F}\{f\}(\omega) = \int_{\mathbb{R}^n} f(\tau)\,\mathrm{e}^{-\jmath(\omega,\tau)}\,d^n\tau \tag{4.26}$$

with $\tau, \omega \in \mathbb{R}^n$ and

$$(\omega,\tau) := \sum_{m=1}^{n} \omega_m\tau_m\,.$$

The n-dimensional inverse Fourier transform can be derived with the same procedure, and we obtain the following definition

$$\mathcal{F}^{-1}\{f\}(\tau) := \frac{1}{(2\pi)^n}\int_{\mathbb{R}^n} f(\omega)\,\mathrm{e}^{\jmath(\omega,\tau)}\,d^n\omega\,. \tag{4.27}$$

With these definitions it's easy to see that our definition of Fourier transform for tempered distributions remains valid for $n > 1$ as well. All properties carry over in similar form. For example, looking back at the derivation of the symmetry relation given by (4.8), we see that in the n-dimensional case it becomes

$$\mathcal{F}\left\{\hat{T}(\tau)\right\} = (2\pi)^n\,T(-\omega)\,. \tag{4.28}$$

The only difference from the one dimensional case is the fact that the factor of 2π becomes $(2\pi)^n$. This happens to all properties involving factors of 2π.

The most important convolution property (4.13) remains unchanged, as can easily be verified by inspecting the derivation for the one-dimensional case.

Before proceeding, it's convenient to extend the multi-index notation that up to now we only used in conjunction with the differential operator. Let a be an n-tuple in $\mathbb{C}^n$ and k a multi-index that we allow to include negative numbers $(k_1,\dots,k_n) \in \mathbb{Z}^n$. Then we can define

$$\begin{aligned} a^k &:= a_1^{k_1}\dots a_n^{k_n} && \text{(exponentiation)}\\ k\,a &:= (k_1a_1,\dots,k_na_n) && \text{(direct product)}\\ \textstyle\sum_{k=l_l}^{l_u} f_k &:= \sum_{k_1=l_1}^{u_1}\dots\sum_{k_n=l_n}^{u_n} f_{k_1,\dots,k_n} && \text{(summation)}\end{aligned}$$

with f_k some function parameterized by the multi-index k and l, u lower resp. upper multi-indices. If in a summation we write integer numbers instead of l and u, we intend multi-indices equal to that number in every position.

We can introduce a multi-index notation for the factorial as well. However, this only makes sense for tuples of natural numbers $k \in \mathbb{N}^n$

$$k! := \prod_{i=1}^{n} (k_i)!.$$

Example 4.11: Fourier transform of δ

In Example 3.1 we saw that the δ distribution in $\mathcal{D}'(\mathbb{R}^n)$ is the tensor product of one dimensional δ's. Hence, with $\tau, \lambda \in \mathbb{R}^n$ and the results of Example 4.5 the Fourier transform of the n-dimensional shifted δ becomes

$$\mathcal{F}\{\delta(\tau - \lambda)\} = \mathrm{e}^{-\jmath(\omega,\tau)}$$

and it's partial derivative

$$\mathcal{F}\{D_i\delta(\tau)\} = (\jmath\omega_i) \qquad i = 1, \ldots, n\,.$$

Using the multi-index notation, the higher order partial derivatives can be conveniently expressed as

$$\mathcal{F}\left\{D^k\delta(\tau)\right\} = (\jmath\omega)^k\,.$$

Using the n-dimensional symmetry relation given by (4.28) we also immediately find

$$\mathcal{F}\left\{\mathrm{e}^{\jmath(\omega_c,\tau)}\right\} = (2\pi)^n\,\delta(\omega - \omega_c)$$

and

$$\mathcal{F}\left\{(-\jmath\tau)^k\right\} = (2\pi)^n D^k\delta(\omega)\,.$$

As in the one-dimensional case, the other properties of the n-dimensional Fourier transform are immediate consequences of the convolution property and the convolution and transform of the δ distribution.

A periodic function on $\mathbb{R}^n$ is a function that is periodic in each independent variable individually, that is, such that there are positive numbers $\mathcal{T}_i$ for $i = 1, \ldots, n$, called the period of the ith independent variable, so that

$$f(\tau_1, \ldots, \tau_i + \mathcal{T}_i, \ldots, \tau_n) = f(\tau_1, \ldots, \tau_i, \ldots, \tau_n)\,.$$

This extension of the concept of a periodic function to higher dimensions permits us to widen the definition of periodic distributions (3.23) on test function of higher dimensions $\mathcal{D}(\mathbb{R}^n)$ in a straightforward way

$$\begin{aligned}&\langle T(\tau_1, \ldots, \tau_i + \mathcal{T}_i, \ldots, \tau_n), \phi(\tau_1, \ldots, \tau_i, \ldots, \tau_n)\rangle \\ &\quad = \langle T(\tau_1, \ldots, \tau_i, \ldots, \tau_n), \phi(\tau_1, \ldots, \tau_i, \ldots, \tau_n)\rangle\ .\end{aligned}$$

From this follows without any difficulty an extension of the second Definition 3.5 as well.

The n-dimensional Fourier series of a periodic distribution $T \in \mathcal{D}'(\mathbb{R}^n)$ is

$$\sum_{k=-\infty}^{\infty} c_k(T)\, \mathrm{e}^{\jmath(k\omega_c, \tau)}$$

with k an n-dimensional multi-index, ω_c the n-tuple $(2\pi/\mathcal{T}_1, \ldots, 2\pi/\mathcal{T}_n)$ and $c_k(T)$ the Fourier coefficients

$$c_k(T) = \frac{1}{\mathcal{T}_1 \cdot \cdots \cdot \mathcal{T}_n} \left\langle T^\circ, \mathrm{e}^{-\jmath(k\omega_c, \tau)} \right\rangle\ .$$

Chapter 5
Laplace Transform of Distributions

5.1 Definition

The classic Laplace transform is closely related to the Fourier one and has similar properties. In a way it can be seen as a modification of the latter in such a way that it can handle exponentially growing functions. To achieve this the transformable functions are required to be bounded on the left.[1] Concretely, the classic one-dimensional Laplace transform is defined on so-called *original functions* $f : \mathbb{R} \to \mathbb{C}$ with the following properties:

1. $\mathrm{f(t) = 0} \quad \text{for} \quad \mathrm{t < 0}$.
2. There is a real number σ_0 such that $f(t)\mathrm{e}^{-\sigma_0 t}$ is absolutely integrable over $\mathbb{R}$.

The greatest lower bound σ_f satisfying the last property is called the *abscissa of convergence* of f.

The Laplace transform of an original function f is a function of the complex variable $s := \sigma + \jmath\omega$ $(\sigma, \omega \in \mathbb{R})$ defined by

$$F(s) := \mathcal{L}\{f\}(s) := \int_0^\infty f(t)\mathrm{e}^{-st}dt \qquad \Re\{s\} > \sigma_f\,. \tag{5.1}$$

We adopt the common convention of denoting the Laplace transform of a function f with the same letter, but capitalized. That is, for this example $F = \mathcal{L}\{f\}$.

We want to extend this definition to distributions with support contained in $[0, \infty)$, that is to right-sided distributions $T \in \mathcal{D}'_+$. To this end note that the integrand in the above definition is the product of an original function f with an indefinitely differentiable function with unbounded support. If we multiply the latter by any indefinitely differentiable function $\gamma(t)$ with support bounded on the left and equal to 1 on a neighborhood of $[0, \infty)$, then we obtain the product of two functions

[1] There are left-sided and two-sided Laplace transforms as well, but we are not going to discuss them.

F. Beffa, *Weakly Nonlinear Systems*, Understanding Complex Systems,
https://doi.org/10.1007/978-3-031-40681-2_5

with support bounded on the left without changing the value of the integral. Then, since f is an original function, for any $\sigma_0 > \sigma_f$ the product $f(t)\mathrm{e}^{-\sigma_0 t}$ is a (regular) tempered distribution and $\gamma(t)\mathrm{e}^{-st}\mathrm{e}^{\sigma_0 t}$, for $\Re\{s\} > \sigma_0$, a test function of fast descent. We have thus obtained a way to define the Laplace transform for a restricted class of distributions.

Definition 5.1 (*Laplace transformable*) A distribution $T \in \mathcal{D}'_+$ is said to be *Laplace transformable* if there exists a constant $\sigma_0 \in \mathbb{R}$ such that

$$T(t)\,\mathrm{e}^{-\sigma_0 t}$$

is a distribution in $\mathcal{S}'(\mathbb{R})$. The greatest lower bound σ_T is called the *abscissa of convergence* of T.

Definition 5.2 (*Laplace transform*) The Laplace transform of a Laplace transformable distribution T is defined by

$$\mathcal{L}\{T\} := \langle T(t)\,\mathrm{e}^{-\sigma_0 t}, \gamma(t)\mathrm{e}^{-(s-\sigma_0)t}\rangle \quad \text{for} \quad \Re\{s\} > \sigma_0 > \sigma_T$$

with γ any indefinitely differentiable function with support bounded on the left and equal to 1 on a neighborhood of the support of T. It is commonly abbreviated by

$$\langle T(t), \mathrm{e}^{-st}\rangle \quad \text{for} \quad \Re\{s\} > \sigma_T\,.$$

The right-half plane $\Re\{s\} > \sigma_T$ is called *region of convergence* (ROC).

If T is Laplace transformable, its transform is a well-defined number for any value of the complex parameter s with $\Re\{s\} > \sigma_T$. In other words, it is a function of s. Since s only appears as a parameter of the test function of fast descent $\gamma(t)\mathrm{e}^{-(s-\sigma_0)t}$, with the continuity and linearity of distributions it is easy to see that

$$D_s\langle T(t), \mathrm{e}^{-st}\rangle = \langle T(t), -t\mathrm{e}^{-st}\rangle = \langle -tT(t), \mathrm{e}^{-st}\rangle\,.$$

In addition, since e^{-st} is an entire analytic function, *in the right half-plane* $\Re\{s\} > \sigma_T$, $\mathcal{L}\{T\}$ *is a holomorphic function.* This is an important result as it allows to use many results from complex analysis.

Higher order derivatives are obtained by iterating the above result so that we have

$$D_s^k\mathcal{L}\{T\}(s) = \mathcal{L}\left\{(-t)^k T\right\}(s)\,. \tag{5.2}$$

Note that the abscissa of convergence of the derivatives is the same as the one of T.

Example 5.1: Laplace Transform of δ

The Laplace transform of δ, of $\delta(t-a)$ and of $D^k\delta$ are

$$\begin{aligned}\langle \delta, \mathrm{e}^{-st}\rangle &= 1\\ \langle \delta(t-a), \mathrm{e}^{-st}\rangle &= \mathrm{e}^{-sa}\\ \langle D^k\delta, \mathrm{e}^{-st}\rangle &= \langle \delta, (-1)^k D^k \mathrm{e}^{-st}\rangle = s^k\,.\end{aligned}$$

In all cases the region of convergence is the entire complex plane $\mathbb{C}$.

Example 5.2: Laplace Transform of $\mathsf{1}_+(t)\, t^k/k!\, \mathrm{e}^{at}$

Let a be a complex number. The Laplace transform of the regular distribution $\mathsf{1}_+(t)\,\mathrm{e}^{at}$ is

$$\langle \mathrm{e}^{at}, \mathrm{e}^{-st}\rangle = \int_0^\infty \mathrm{e}^{-(s-a)t}\,dt = \frac{1}{s-a}$$

with abscissa of convergence $\sigma_{\exp(at)} = a$.

From this and (5.2), the Laplace transform of

$$\mathsf{1}_+(t)\,\frac{t^k}{k!}\,\mathrm{e}^{at}$$

is readily found to be

$$\mathcal{L}\left\{\mathsf{1}_+(t)\,\frac{t^k}{k!}\,\mathrm{e}^{at}\right\} = \frac{(-1)^k}{k!} D^k \frac{1}{s-a} = \frac{1}{(s-a)^{k+1}}\,.$$

5.2 Properties

The Laplace transform is a linear operation: given two Laplace transformable distribution S and T with abscissa of convergence σ_S resp. σ_T, the transform of their weighted sum is

$$\mathcal{L}\{c_1 S + c_1 T\} = c_1\mathcal{L}\{S\} + c_2\mathcal{L}\{T\} \quad \text{for} \quad \Re\{s\} > \max(\sigma_S, \sigma_T)\,.$$

Let a be a complex number. Then, from the definition, the Laplace transform of $\mathrm{e}^{at}T(t)$

$$\mathcal{L}\left\{\mathrm{e}^{-at}T\right\}(s) = \langle \mathrm{e}^{-at}T(t), \mathrm{e}^{-st}\rangle = \langle T(t), \mathrm{e}^{-(s+a)t}\rangle = \mathcal{L}\{T\}(s+a)$$

with region of convergence $\Re\{s\} > \sigma_T - \Re\{a\}$.

We saw in Example 3.6 that the convolution of distributions in $\mathcal{D}'_+$ is always well-defined. Therefore, the convolution of two Laplace transformable distributions S and T is well-defined. Then, for $\Re\{s\} > \sigma_0 = \max(\sigma_S, \sigma_T)$, the transform of their convolution product is by definition

$$\begin{aligned}&\mathcal{L}\{S * T\}(s)\\&= \langle (S * T)\mathrm{e}^{-\sigma_0 t}, \gamma(t)\mathrm{e}^{-(s-\sigma_0)t}\rangle\\&= \langle S(t) \otimes T(\lambda)\mathrm{e}^{-\sigma_0\lambda}, \gamma(t+\lambda)\mathrm{e}^{-(s-\sigma_0)(t+\lambda)}\rangle\,.\end{aligned}$$

By noting that, over a neighborhood of the support of $S \otimes T$, $\gamma(t+\lambda) = 1 = \gamma(t)\gamma(\lambda)$ we can proceed further and obtain

$$\begin{aligned}&\mathcal{L}\{S * T\}(s)\\&= \langle S(t)\mathrm{e}^{-\sigma_0 t} \otimes T(\lambda)\mathrm{e}^{-\sigma_0\lambda}, \gamma(t)\mathrm{e}^{-(s-\sigma_0)t}\gamma(\lambda)\mathrm{e}^{-(s-\sigma_0)\lambda}\rangle\\&= \langle S(t)\mathrm{e}^{-\sigma_0 t}, \langle T(\lambda)\mathrm{e}^{-\sigma_0\lambda}, \gamma(\lambda)\mathrm{e}^{-(s-\sigma_0)\lambda}\rangle\gamma(t)\mathrm{e}^{-(s-\sigma_0)t}\rangle\\&= \langle S(t)\mathrm{e}^{-\sigma_0 t}, \gamma(t)\mathrm{e}^{-(s-\sigma_0)t}\rangle\langle T(\lambda)\mathrm{e}^{-\sigma_0\lambda}, \gamma(\lambda)\mathrm{e}^{-(s-\sigma_0)\lambda}\rangle\\&= \mathcal{L}\{S\}(s)\mathcal{L}\{T\}(s)\end{aligned}$$

which is well-defined since, in the specified ROC, the Laplace transforms of S and T are holomorphic functions. We thus see that, as the Fourier transform, the Laplace transform converts convolutions into products

$$\mathcal{L}\{S * T\} = \mathcal{L}\{S\}\mathcal{L}\{T\} \quad \text{for} \quad \Re\{s\} > \max(\sigma_S, \sigma_T)\,. \tag{5.3}$$

A key advantage over the Fourier transform is that here the multiplication is between functions that are holomorphic in the specified open right-half plane.

In a similar way as we did for the Fourier transform, we can use this property to derive several additional properties in a straightforward way. Specifically, using the properties of the convolution product and the Laplace transform of the Dirac δ and related distributions (Example 5.1), we immediately obtain the properties in Table 5.1 that we have not yet discussed.

Example 5.3: Convolution of exp's

Let k and l be natural numbers and a a complex constant. We want to calculate the following convolution product

$$\mathsf{1}_+(t)\frac{t^k}{k!}\,\mathrm{e}^{at} * \mathsf{1}_+(t)\frac{t^l}{l!}\,\mathrm{e}^{at}\,.$$

From the convolution property of the Laplace transform and the results of Example 5.2 the transform of the above convolution product is

Table 5.1 Properties of the Laplace transformation

T	$\mathcal{L}\{T\}$	ROC
$\sum_m a_m T_m$	$\sum_m a_m \mathcal{L}\{T\}_m$	$\Re\{s\} > \sup_m(\sigma_m)$
$S * T$	$\mathcal{L}\{S\}\ \mathcal{L}\{T\}$	$\Re\{s\} > \max(\sigma_S, \sigma_T)$
$T(t-\tau)$	$e^{-s\tau}\mathcal{L}\{T\}$	$\Re\{s\} > \sigma_T$
$e^{-at}T$	$T(s+a)$	$\Re\{s\} > \sigma_T - \Re\{a\}$
$D^k T$	$s^k\,\mathcal{L}\{T\}$	$\Re\{s\} > \sigma_T$
$(-t)^k\,T$	$D^k\mathcal{L}\{T\}$	$\Re\{s\} > \sigma_T$

$$\frac{1}{(s-a)^k}\,\frac{1}{(s-a)^l} = \frac{1}{(s-a)^{k+l}}\,.$$

With it, we find the desired result as

$$1_+(t)\frac{t^{k+l}}{(k+l)!}\,\mathrm{e}^{at}\,.$$

5.3 Inverse Laplace Transform

The Laplace transform isn't only similar to the Fourier one, it can also be formally related to it. Consider first an original function f. By writing its Laplace transform as

$$\int_0^\infty f(t)\mathrm{e}^{-\sigma t}\mathrm{e}^{-\jmath\omega t}\,dt$$

we see that, for every value of $\sigma > \sigma_f$, it can be interpreted as the Fourier transform of the function $f(t)\mathrm{e}^{-\sigma t}$.

This relation between the two transforms can be extended to distributions. Consider a Laplace transformable distribution T. We have established that, for $\Re\{s\} > \sigma_T$, its Laplace transform is a holomorphic function of s. In addition, by definition, for every value $\sigma_0 > \sigma_T$, $T(t)\,\mathrm{e}^{-\sigma_0 t}$ is a distribution of slow growth and for $\sigma > \sigma_0$, $\gamma(t)\mathrm{e}^{-(\sigma-\sigma_0)t}\mathrm{e}^{-\jmath\omega t}$ is a test function of fast descent for every value of ω. We conclude that the Laplace transform of T considered as a function of ω for fixed σ must be a regular distribution of slow growth. Hence, the following integral is well-defined

$$\int_{\mathbb{R}} \langle T(t)\,\mathrm{e}^{-\sigma_0 t}, \gamma(t)\mathrm{e}^{-(\sigma-\sigma_0)t}\mathrm{e}^{-\jmath\omega t}\rangle\phi(\omega)\,d\omega$$

for any $\phi(\omega) \in \mathcal{S}$. This integral can be recognized as the tensor product $1(\omega) \otimes T(t)\,\mathrm{e}^{-\sigma_0 t}$ and, using Fubini's theorem, it can be rearranged to become

$$\langle T(t)\,\mathrm{e}^{-\sigma_0 t}, \gamma(t)\mathrm{e}^{-(\sigma-\sigma_0)t}\int_{\mathbb{R}} \mathrm{e}^{-\jmath\omega t}\phi(\omega)\,d\omega\rangle$$
$$= \langle T(t)\,\mathrm{e}^{-\sigma t}, \hat{\phi}(t)\rangle = \langle \mathcal{F}\{T(t)\,\mathrm{e}^{-\sigma t}\}, \phi(\omega)\rangle\,.$$

We thus obtain the claimed relation between Laplace and Fourier transforms

$$\mathcal{L}\{T\} = \mathcal{F}\{\mathrm{e}^{-\sigma t}T\} \quad \text{for} \quad \sigma > \sigma_T \tag{5.4}$$

which gives us a formal way to invert the Laplace transform.

A first consequence of this relation is that, given the Laplace transform of a distribution T *with abscissa of convergence* $\sigma_T < 0$, we can immediately find its Fourier transform by setting $s = \jmath\omega$

$$\hat{T}(\omega) = \mathcal{L}\{T\}\,(\jmath\omega)\,.$$

Another important consequence of (5.4) is that, if $\mathcal{L}\{T\} = 0$ on a vertical line with $\Re\{s\} > \sigma_T$, then $T = 0$. In fact, with the above result we have that

$$0 = \langle \mathcal{L}\{T\}\,, \phi\rangle = \langle \mathcal{F}\{\mathrm{e}^{-\sigma t}T\}, \phi\rangle = \langle \mathrm{e}^{-\sigma t}T, \hat{\phi}\rangle$$

from which we conclude that $\mathrm{e}^{-\sigma t}T$ and hence T must vanish. In addition, with $T = S - U$ this implies that, if $\mathcal{L}\{S\} = \mathcal{L}\{U\}$ on a vertical line of the region of convergence, then $S = U$. In other words, if a function, holomorphic in an open right-half plane, is the Laplace transform of a distribution in $\mathcal{D}'_+$, then it is the transform of a *unique* distribution.

The next logical question to ask is: which holomorphic functions are transforms of a distribution? To answer this question, consider first an holomorphic function F bounded by

$$|F(s)| \le \frac{C}{|s|^2} \quad \text{for} \quad \Re\{s\} \ge \sigma_0 > 0$$

with C a constant. Then

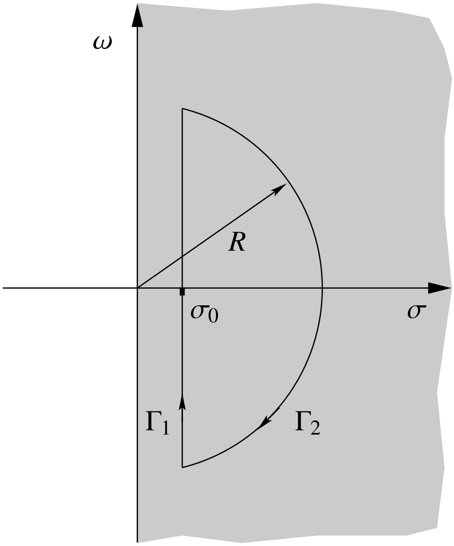
$$\int_{-\infty}^{\infty} \left|F(\sigma_0 + \jmath\omega)\mathrm{e}^{\jmath\omega t}\right|\,d\omega \le \int_{-\infty}^{\infty} \frac{C}{\sigma_0^2 + \omega^2}\,d\omega < \infty$$

and the inverse Fourier integral

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} F(\sigma_0 + \jmath\omega)\mathrm{e}^{\jmath\omega t}\,d\omega$$

exists and defines a continuous function that we may write as $\mathrm{e}^{-\sigma_0 t}f(t)$. The thus defined function f is therefore

Fig. 5.1 Integration path for $t < 0$



$$\begin{aligned} f(t) &= \frac{e^{\sigma_0 t}}{2\pi} \int_{-\infty}^{\infty} F(\sigma_0 + \jmath\omega) e^{\jmath\omega t}\, d\omega \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\sigma_0 + \jmath\omega) e^{(\sigma_0 + \jmath\omega)t}\, d\omega \\ &= \frac{1}{2\pi\jmath} \int_{\sigma_0 - \jmath\infty}^{\sigma_0 + \jmath\infty} F(s)\, e^{st}\, ds \end{aligned} \tag{5.5}$$

and corresponds to the integral of an holomorphic function along the vertical line defined by $\Re\{s\} = \sigma_0$. If we write the variable s in its polar representation $s = Re^{\jmath\varphi}$ it's easy to verify that in the right-half plane and for $t < 0$

$$|F(s)e^{st}| \le \frac{C}{R^2}\,.$$

Therefore, if we close the integration path of the above integral by first making the line finite, then closing it along the half-circle shown in Fig. 5.1 and then taking the limit $R \to \infty$, the value of the integral remains unchanged. In fact

$$\lim_{R\to\infty} \left| \frac{1}{2\pi\jmath} \int_{\Gamma_2} F(s)\, e^{st}\, ds \right| \le \lim_{R\to\infty} \frac{C}{2\pi\, R^2}\, R\pi = 0\,.$$

Having closed the integration path we can now use Cauchy's theorem and conclude that for $t < 0$, $f(t) = 0$.

Cauchy's theorem can also be used to show that the value of the integral is the same along any vertical line with $\Re\{s\} > 0$. To show this, we integrate along two vertical segments and close the path with horizontal ones. Since the contribution of the horizontal paths vanishes as we extend the length of the vertical ones toward infinity, we conclude that the value of the integral along the two vertical lines must be the same.

We have thus established that the function f doesn't depend on the value of σ_0, is continuous and vanishes for $t < 0$. These characteristics make f an original function and hence a Laplace transformable distribution in $\mathcal{D}'_+$. Furthermore, from the definition of f and (5.4) we see that F is its Laplace transform.

Now consider the more general function $G(s) = s^k F(s - \sigma_G)$ with $\sigma_G \in \mathbb{R}$, $k \in \mathbb{N}$ and F as before. From the properties of the Laplace transform we know that it is the transform of the distribution $g = D^k(\mathrm{e}^{\sigma_G t} f)$ which is also clearly in $\mathcal{D}'_+$. We therefore conclude that every function G that, for some $\sigma_G \in \mathbb{R}$, is holomorphic in the open right-half plane $\Re\{s\} > \sigma_G$ and is bounded above by a polynomial P

$$|G(s)| \leq P(|s|) \qquad \Re\{s\} > \sigma_G \tag{5.6}$$

is the Laplace transform of a distribution in $\mathcal{D}'_+$.

Without going into the details we also mention that the converse is also true. That is, every Laplace transformable distribution $T \in \mathcal{D}'_+$ is a derivative of some regular distribution associated with a continuous original function [16].

With the transforms of Examples 5.1 and 5.2 we can find the inverse Laplace transform of any rational function of s by partial fraction expansion. Note also that (5.5) corresponds to the classic inverse Laplace transform for functions.

Example 5.4: Laplace versus Fourier Transform of $\mathsf{1}_+$

In this example we calculate the Laplace transform of the Heaviside step function $\mathsf{1}_+$. While it's easy to obtain it directly from the definition, we calculate it from our previous results $D\mathsf{1}_+ = \delta$ (Example 2.8) and $\mathcal{L}\{D\delta\} = s$ (Example 5.1). This is to compare it with the methods used in Examples 4.8 and 4.9 to obtain its Fourier transform.

By using

$$D\mathsf{1}_+ = D\delta * \mathsf{1}_+$$

and the convolution property of the Laplace transform, we obtain the following equation for the Laplace transform of $\mathsf{1}_+$

$$s\,\mathcal{L}\{\mathsf{1}_+\} = 1\,.$$

Then, since all Laplace transforms are holomorphic functions in an open right-half plane and only the zero distribution has zero as its transform, we can conclude that

$$\mathcal{L}\{\mathsf{1}_+\} = \frac{1}{s} \qquad \Re\{s\} > 0\,.$$

As an extra step we want to obtain the Fourier transform of $\mathsf{1}_+$ from its Laplace transform. The abscissa of convergence of $\mathcal{L}\{\mathsf{1}_+\}$ is not smaller than zero. Therefore, we can't obtain it by simply setting $s = \jmath\omega$. However, in cases like this, where the abscissa of convergence is zero, given the continuity of distributions, it's still possible to obtain the Fourier transform as a limit, so that

$$\langle\mathcal{F}\{\mathsf{1}_+\}, \phi\rangle = \lim_{\Re\{s\}\downarrow 0} \int_\Gamma \frac{\phi(\Im\{s\})}{s} \frac{ds}{\jmath}$$

with Γ a vertical line in the ROC@. The limit is only problematic around the origin, where we can integrate along a small half-circle of radius ϵ

$$\begin{aligned}
&\lim_{\Re\{s\}\downarrow 0} \frac{1}{\jmath} \int_\Gamma \frac{\phi(\Im\{s\})}{s}\, ds \\
&= \lim_{\epsilon\downarrow 0} \frac{1}{\jmath} \left\{\int_{|\omega|>\epsilon} \frac{\phi(\omega)}{\jmath\omega}\, \jmath\, d\omega + \int_{-\pi/2}^{\pi/2} \frac{\phi(\epsilon\sin(\varphi))}{\epsilon e^{\jmath\varphi}}\, \jmath\epsilon\, e^{\jmath\varphi}\, d\varphi\right\} \\
&= \lim_{\epsilon\downarrow 0} \left\{\int_{|\omega|>\epsilon} \frac{\phi(\omega)}{\jmath\omega}\, d\omega\right\} + \int_{-\pi/2}^{\pi/2} \phi(0)\, d\varphi \\
&= \langle \mathrm{pv}\, \frac{1}{\jmath\omega} + \pi\delta, \phi\rangle\,.
\end{aligned}$$

Example 5.5: Exploiting Continuity

To simplify the notation we assume that all appearing functions (more correctly, regular distributions) disappear for $t < 0$. For example, we write t^k for $\mathsf{1}_+(t)\, t^k$.

From the results of Example 5.2, by setting $a = 0$, we can note a dualism between positive and negative powers

$$\mathcal{L}\left\{\frac{t^k}{k!}\right\} = \frac{1}{s^{k+1}} \qquad \Re\{s\} > 0\,.$$

If we sum the first N powers of t we obtain

$$\mathcal{L}\left\{\sum_{k=0}^{N-1} \frac{t^k}{k!}\right\} = \sum_{k=0}^{N-1} \frac{1}{s^{k+1}} = \frac{1}{s} \sum_{k=0}^{N-1} \frac{1}{s^k}\,.$$

Using the continuity of distributions we can let N tend to infinity. The original distribution converges to the exponential function, while the transform becomes a geometric series (plus a factor)

$$\mathcal{L}\left\{e^t\right\} = \frac{1}{s}\sum_{k=0}^{\infty}\frac{1}{s^k}$$

that converges for $|s| > 1$. This means that there is a right-half plane $\Re\{s\} > 1$ where the series converges and can be summed to obtain the expected result

$$\mathcal{L}\left\{e^t\right\} = \frac{1}{s}\,\frac{1}{1-1/s} = \frac{1}{s-1} \qquad \Re\{s\} > 1\,.$$

Note that the last expression can also be expressed as a geometric series of positive powers of s

$$\frac{1}{s-1} = -\sum_{k=0}^{\infty} s^k\,.$$

However, this series only converges if $|s| < 1$. Consequently, there is no right-half plane where the series is holomorphic and hence it isn't the Laplace transform of a distribution.

5.4 Extension to Several Variables

The Laplace transform can be extended to functions of several variables in the same way as we did for the Fourier transform. Let τ be an n-tuple in $\mathbb{R}^n$. A multi-variable original function $f : \mathbb{R}^n \to \mathbb{C}$ is a function that is an original function with respect to each variable independently, that is

1. $f(\tau) = 0$ if any $\tau_i < 0, i = 1, \dots, n$.
2. There is a $\sigma_0 \in \mathbb{R}^n$ such that $f(\tau)e^{-(\sigma_0,\tau)}$ is absolutely integrable over $\mathbb{R}^n$.

The classic Laplace transform is then defined as

$$\mathcal{L}\{f\}(s) = \int_{\mathbb{R}^n_+} f(\tau)\,e^{-(s,\tau)}\,d^n\tau \tag{5.7}$$

with $s \in \mathbb{C}^n$ and $\mathbb{R}^n_+$ the n-dimensional Cartesian product of the half-line $[0, \infty)$.

With this definition we see that the definition of the Laplace transform for distributions extends to higher dimensions essentially without modification. By interpreting k as a multi-index and all variables as n-dimensional ones, all properties of Table 5.1 remain valid. The only exception is the classic inverse Laplace transform integral (5.5). As we have seen, this integral is based on the inverse Fourier transform and the factor $2\pi\jmath$ has therefore to be replaced by $(2\pi\jmath)^n$.

Chapter 6
Summable Distributions

In this chapter we study a class of distributions, summable distributions, that can be extended on smooth bounded functions. These distributions have properties that are well suited to describe some classes of systems. However, the material is more technical than the rest of the book and can be skipped without loss of continuity.

6.1 Definition and Canonical Extension

One can define summable distributions in several equivalent ways [16]. The following one is the most suitable for our purposes.

Definition 6.1 (*Summable distributions*) A summable distribution T is a distribution that can be represented as a finite sum of derivatives (in the sense of distribution) of functions $f_k \in L^1$

$$T = \sum_{|k| \leq m} D^k f_k$$

with k an n-tuple in $\mathbb{N}^n$ and $m \in \mathbb{N}$.

We denote the vector space of summable distributions by $\mathcal{D}'_{L^1}$.

An important property of summable distributions is the fact that they can be extended to continuous linear functionals on $\mathcal{B}$, the set of indefinitely differentiable functions that, together with all their derivatives, are bounded

$$\mathcal{B} := \left\{ \phi \in C^\infty \,|\, D^k \phi \text{ is bounded}, \; k \in \mathbb{N}^n \right\}.$$

As usual, to talk about continuity, we need to define a convergence criterion (topology).

F. Beffa, *Weakly Nonlinear Systems*, Understanding Complex Systems,
https://doi.org/10.1007/978-3-031-40681-2_6

Definition 6.2 (*Convergence in $\mathcal{B}$*) A sequence of functions (η_j) in $\mathcal{B}$ converges to zero if the sequence as well as all sequences of the derivatives $(D^k\eta_j)$ converge uniformly to zero as j tends to infinity. That is, given the norms

$$p_m(\eta) := \sum_{|k|\leq m} \sup_{\tau\in\mathbb{R}^n} |D^k\eta(\tau)|,$$

the sequence (η_j) converges to zero if, as j tends to infinity, the sequence of numbers $(p_m(\eta_j))$ converges to zero for all $m \in \mathbb{N}$.

The application of a summable distribution T to a function $\eta \in \mathcal{B}$ is well defined since

$$\left|\langle D^k f_k, \eta\rangle\right| \leq \sup_{\tau\in\mathbb{R}^n} |D^k\eta(\tau)| \int_{\mathbb{R}^n} |f_k(\tau)|\, d^n\tau < \infty\,.$$

This shows that it's possible to extend a summable distribution to $\mathcal{B}$. However, the extension in general is not unique. The reason being that the set of test functions $\mathcal{D}$ is not dense in $\mathcal{B}$. That is to say that it's not possible to approximate to arbitrary accuracy any function in $\mathcal{B}$ with functions from $\mathcal{D}$.

Example 6.1: Constant Function

Consider the constant function $\mathbf{1} : t \mapsto 1$, the function $\alpha \in \mathcal{D}$ defined by (2.11), and functions $\alpha_j \in \mathcal{D}$ defined by $\alpha_j(t) = \alpha(t/j),\ j \in \mathbb{N}$. The product $\alpha_j\mathbf{1}$ is clearly a test function satisfying $\alpha_j(2j) = 0$ for all values of j. From this we see that

$$p_0(\alpha_j\mathbf{1} - \mathbf{1}) = \sup_{t\in\mathbb{R}} |\alpha_j(t)\mathbf{1}(t) - \mathbf{1}(t)| = 1$$

no matter how large j is.

Example 6.2: Average Functional

Consider the functions $\mathbf{1}, \alpha_j$ from Example 6.1 and the following functional L on $\mathcal{B}$

$$L(\eta) := \lim_{C\to\infty} \frac{1}{C} \int_{-C/2}^{C/2} \eta(\tau)\, d\tau\,.$$

It is easily seen that L is linear and continuous. Its value on the constant function $\mathbf{1}$ is 1 while its value on the test functions $\alpha_j\mathbf{1}$ is zero for all values of j

$$|L(\alpha_j\mathbf{1})| \leq \lim_{C\to\infty} \frac{1}{C} \int_{-2j}^{2j} |\alpha_j(\tau)|\, d\tau \leq \lim_{C\to\infty} \frac{4j}{C} = 0.$$

The functional L is therefore a valid extension to $\mathcal{B}$ of the zero distribution as is the zero functional on $\mathcal{B}$.

We can define a unique, canonical extension of a summable distribution T by requiring an additional condition on the extension [19]. A suitable condition can be obtained from the properties of Lebesgue integrals. Consider again the test functions $\alpha_j \in \mathcal{D}$ from Example 6.1 and a function $\eta \in \mathcal{B}$. If we apply T to $\alpha_j \eta$ we obtain

$$\begin{aligned}\left\langle T, \alpha_j \eta \right\rangle &= \sum_{|k|\leq m} (-1)^{|k|} \int_{\mathbb{R}^n} f_k(\tau)\, D^k(\alpha_j(\tau)\eta(\tau))\, d^n\tau \\ &= \sum_{|k|\leq m} (-1)^{|k|} \left(\int_{|\tau|\leq j} f_k(\tau)\, D^k\eta(\tau)\, d^n\tau + \int_{|\tau|>j} f_k(\tau)\, D^k(\alpha_j(\tau)\eta(\tau))\, d^n\tau \right) .\end{aligned}$$

The integral of an L^1 function can be approximated up to an arbitrary $\epsilon > 0$ by an integral over a suitably chosen compact subset K of $\mathbb{R}^n$. Therefore we can find a large enough N such that for $j > N$

$$\left\langle T, \alpha_j \eta \right\rangle = \epsilon + \sum_{|k|\leq m} (-1)^{|k|} \int_{|\tau|\leq j} f_k(\tau) D^k\eta(\tau)\, d^n\tau \, .$$

Thus, in the limit as j tends to infinity we obtain a well-defined continuous linear functional on $\mathcal{B}$.

The important observation from this derivation is the fact that, to find an extension to $\mathcal{B}$ of a summable distribution, it is not necessary to require uniform convergence on the whole of $\mathbb{R}^n$. An extension can be obtained by requiring uniform convergence on every compact subset $K \subset \mathbb{R}^n$. More precisely, by requiring the convergence criterion that we defined for the space $\mathcal{E}$. From this observation we define the following property.

Definition 6.3 (*Bounded convergence property*) A continuous linear functional on $\mathcal{B}$ has the *bounded convergence property* if, given any sequence (η_j) of functions $\eta_j \in \mathcal{B}$ with $p_m(\eta_j) < \infty$ for all m, and converging to zero in the space $\mathcal{E}$ as $j \to \infty$, then

$$\left\langle T, \eta_j \right\rangle \to 0\,, \qquad j \to \infty\,.$$

The sequence $(\alpha_j \eta)$ does converge to η in $\mathcal{E}$. Hence, by continuity, there is a unique extension of T to $\mathcal{B}$ with the bounded convergence property

$$\lim_{j\to\infty} \left\langle T, \alpha_j \eta \right\rangle = \langle T, \eta \rangle \, . \tag{6.1}$$

In particular this shows that this extension does not depend on the particular representation of T in terms of derivatives of integrable functions.

The converse is also true. The restriction to $\mathcal{D}$ of any continuous linear functional on $\mathcal{B}$ with the bounded convergence property defines a unique summable distribution. Thus, there is a one to one correspondence between summable distributions and continuous linear functionals on $\mathcal{B}$ with the bounded convergence property.

Definition 6.4 (*Canonical extension*) The *canonical* extension to $\mathcal{B}$ of a summable distribution is the unique extension to a continuous linear functional on $\mathcal{B}$ with the bounded convergence property.

In the following, whenever we use the extension of a summable distribution it will always be assumed to be the canonical one.

While our previous definition of differentiation carries over to summable distributions without problems, this is not the case for multiplication. In general the product of a bounded function $\eta \in \mathcal{B}$ with an unbounded one $\gamma \in \mathcal{E}$ is not bounded and therefore not in $\mathcal{B}$. Differently from this, the product of two bounded functions $\eta, \zeta \in \mathcal{B}$ is always in $\mathcal{B}$. Therefore, *for summable distributions $T \in \mathcal{D}'_{L^1}$ multiplication has to be restricted to functions in $\mathcal{B}$*

$$\langle \eta T, \zeta \rangle = \langle T, \eta\zeta \rangle \ .$$

6.2 Convolution of Summable Distributions

In Sect. 3.2 we defined the convolution product between two distributions $S, T \in \mathcal{D}'$ by

$$\langle S * T, \phi \rangle = \langle S(\tau) \otimes T(\lambda), \phi(\tau + \lambda) \rangle$$

and saw that in general, if the support of both S and T is unbounded, it may not exist. In this section we show that if S and T are summable, then their convolution product is well-defined despite the fact that their support is unbounded.

Consider the application of a summable distribution T to a function $\tau \mapsto \eta(\lambda + \tau) \in \mathcal{B}$ with λ a parameter. Following the same arguments as in Sect. 3.1, given the linearity and continuity of T, we deduce that it is a continuous and indefinitely differentiable function ζ belonging to $\mathcal{B}$

$$\zeta(\lambda) = \langle T(\tau), \eta(\lambda + \tau) \rangle \ .$$

For this reason the convolution of two summable distributions S and T is always well-defined

$$\langle S * T, \eta \rangle = \langle S(\lambda), \langle T(\tau), \eta(\lambda + \tau) \rangle \rangle = \langle S, \zeta \rangle$$

and commutative.

Next we investigate the convolution of a summable distribution T with a function in $\mathcal{B}$. Consider the application of T to the function $\tau \mapsto \eta(\lambda - \tau) \in \mathcal{B}$ parameterised by λ. As we just saw, it is a function that we call again ζ and that is clearly locally integrable. Hence, it defines a distribution in $\mathcal{D}'$ and with $\phi \in \mathcal{D}$ we can write

$$\begin{aligned}\langle\langle T(\tau), \eta(\lambda-\tau)\rangle, \phi(\lambda)\rangle &= \langle\zeta, \phi\rangle = \langle\phi, \zeta\rangle = \langle\phi(\lambda), \langle T(\tau), \eta(\lambda-\tau)\rangle\rangle \\ &= \langle\phi(\lambda) \otimes T(\tau), \eta(\lambda-\tau)\rangle \\ &= \left\langle T(\tau), \int_{\mathbb{R}^n} \phi(\lambda)\eta(\lambda-\tau)\, d^n\lambda\right\rangle \\ &= \left\langle T(\tau), \int_{\mathbb{R}^n} \phi(\xi+\tau)\eta(\xi)\, d^n\xi\right\rangle \\ &= \langle T(\tau) \otimes \eta(\xi), \phi(\xi+\tau)\rangle = \langle T * \eta, \phi\rangle \ .\end{aligned}$$

or

$$\langle T(\tau), \eta(\lambda-\tau)\rangle = (T * \eta)(\lambda) \tag{6.2}$$

This shows that a summable distribution can be regularised by a function in $\mathcal{B}$ and that the resulting regularised is also a function in $\mathcal{B}$.

6.3 Fourier Transform of Summable Distributions

The functions $\tau \mapsto e^{-\jmath(\omega,\tau)}$ with $\omega \in \mathbb{R}^n$ belong to $\mathcal{B}$. For this reason the Fourier transform of a summable distribution T can be expressed in a simple way. Let $\phi \in \mathcal{D}$, then

$$\begin{aligned}\langle\mathcal{F}\{T\}, \phi\rangle = \langle T(\tau), \mathcal{F}\{\phi\}(\tau)\rangle &= \left\langle T(\tau), \int_{\mathbb{R}^n} \phi(\omega)e^{-\jmath(\tau,\omega)}\, d^n\omega\right\rangle \\ &= \left\langle T(\tau), \left\langle\phi(\omega), e^{-\jmath(\tau,\omega)}\right\rangle\right\rangle = \left\langle T(\tau) \otimes \phi(\omega), e^{-\jmath(\tau,\omega)}\right\rangle \\ &= \left\langle\phi(\omega), \left\langle T(\tau), e^{-\jmath(\tau,\omega)}\right\rangle\right\rangle = \left\langle\left\langle T(\tau), e^{-\jmath(\tau,\omega)}\right\rangle, \phi(\omega)\right\rangle\end{aligned}$$

or

$$\mathcal{F}\{T\}(\omega) = \left\langle T(\tau), e^{-\jmath(\tau,\omega)}\right\rangle \ . \tag{6.3}$$

$\mathcal{F}\{T\}$ is thus a continuous function. Moreover it has at most polynomial growth, for, by representing T as a sum of integrable functions and the properties of the Fourier transform, for some $m \in \mathbb{N}$ we have

$$\begin{aligned}|\mathcal{F}\{T\}(\omega)| &= \left|\sum_{|k|\leq m} (\jmath\omega)^k \mathcal{F}\{f_k\}(\omega)\right| \\ &\leq \sum_{|k|\leq m} |\omega|^k \int_{\mathbb{R}^n} |f_k(\tau)|\, d^n\tau \leq C(1+|\omega|)^m\end{aligned}$$

with C a constant. Thus, *the Fourier transformed of a summable distribution is a function of slow growth.*

The converse is not in general true, but we can find a class of functions for which it is. This is the set $\mathcal{O}_M$, the set of functions of slow growth that are indefinitely differentiable.

To see that this is the case, consider the Fourier transformed $\hat{T}$ of some tempered distribution T and assume that $\hat{T} \in \mathcal{O}_M$. If $\phi \in \mathcal{D} \subset \mathcal{S}$ then its Fourier transformed $\hat{\phi}$ as well as $\hat{\phi}\hat{T}$ are in $\mathcal{S}$. Therefore we see that

$$\phi * T = \mathcal{F}^{-1}\{\hat{\phi}\hat{T}\} \in \mathcal{S} \subset L^1$$

is a summable distribution and we can apply it to a function $\eta \in \mathcal{B}$ to obtain

$$\langle \phi * T, \eta \rangle = \langle \phi(\lambda) \otimes T(\tau), \eta(\lambda + \tau) \rangle = \langle T(\tau), \langle \phi(\lambda), \eta(\lambda + \tau) \rangle \rangle \ .$$

Since $\phi \in \mathcal{D}$ and $\eta \in \mathcal{B}$ are arbitrary and $\langle \phi(\lambda), \eta(\lambda + \tau) \rangle \in \mathcal{B}$ we deduce that T is a summable distribution.

We conclude this section by showing that the property of the Fourier transform of transforming convolution products into ordinary products is valid for arbitrary summable distributions. Let S, T be summable distributions, then using (6.3) and the property of the exponential function $\mathrm{e}^{-\jmath(\tau+\lambda,\omega)} = \mathrm{e}^{-\jmath(\tau,\omega)}\mathrm{e}^{-\jmath(\lambda,\omega)}$ one readily obtain that

$$\mathcal{F}\{S * T\} = \mathcal{F}\{S\}\mathcal{F}\{T\}\,. \tag{6.4}$$

The product is well defined as $\mathcal{F}\{S\}$ and $\mathcal{F}\{T\}$ are both functions.

Part II
Systems

Chapter 7
Convolution Equations

The objective of this chapter is to show that the solution of ordinary differential equations, if based on distributions as opposed to functions, can be obtained by (mostly) algebraic methods. These methods are rigorous forms of the so-called Heaviside's operational or symbolic calculus. The close relationship to the integral transforms that convert convolution into the ordinary multiplication is also shown.

! Notation

With this chapter we stop using uppercase letters such as T to denote distributions. Instead, we start using lowercase letters such as the ones typically used to denote functions, for example f. We also adopt the convention of denoting the Laplace transform of a distribution, say f, with the same letter, but changed to uppercase, e.g. , $F = \mathcal{L}\{f\}$. When we need to distinguish between the ordinary and the distributional differential operator, we will in general denote the former by $\frac{\mathrm{d}}{\mathrm{d}t}$ and continue to denote the latter by D.

7.1 Convolution Algebra

An *algebra* $\mathcal{A}$ is a vector space together with an associative product $\odot$ such that multiplication of any two vectors produces another vector in $\mathcal{A}$ and such that for any constants a, b and any vectors $f, g, h \in \mathcal{A}$ the following distributivity laws are valid

$$(af + bg) \odot h = a(f \odot h) + b(g \odot h) \tag{7.1}$$

$$f \odot (ag + bh) = a(f \odot g) + b(f \odot h)\,. \tag{7.2}$$

F. Beffa, *Weakly Nonlinear Systems*, Understanding Complex Systems,
https://doi.org/10.1007/978-3-031-40681-2_7

The convolution product seems like an adequate product to make an algebra out of distributions. Unfortunately, as we saw, the convolution product is not defined for arbitrary distributions. The solution is to restrict the set of distributions to a vector subspace of $\mathcal{D}'$ on which the convolution is well-defined.

Definition 7.1 (*Convolution algebra*) A convolution algebra $\mathcal{A}'$ is a vector subspace of $\mathcal{D}'$ with the following properties:

- The convolution product is associative.
- $\mathcal{A}'$ with the convolution product forms an algebra.
- δ is in $\mathcal{A}'$.

A convolution algebra is thus an algebra with a unit and for which the product is always commutative. We also note that the triple $(\mathcal{A}', +, *)$ forms a commutative *ring*.

We have already met three examples of convolution algebras: (i) the set of right-sided distributions $\mathcal{D}'_+$, (ii) the set of periodic distributions and (iii) the set of distributions with compact support $\mathcal{E}'$.

7.2 Convolution Equations

In this section we study convolution equations. We will see that they provide a framework for studying a broad class of systems that is the time-domain counterpart of one based on the Laplace transform.

A *convolution equation* is an equation of the form

$$g * y = x \tag{7.3}$$

with g and x given distributions and y a distribution to be determined. In this section we assume g, x and y to be elements of a convolution algebra $\mathcal{A}'$. Suppose that g has an *inverse* in $\mathcal{A}'$, that is, there is an element denoted by $g^{*-1} \in \mathcal{A}'$ such that

$$g * g^{*-1} = g^{*-1} * g = \delta\,.$$

Then $g^{*-1} * x$ is a solution of the equation for any x, since

$$y = g^{*-1} * g * y = g^{*-1} * x\,.$$

Note that if there is an inverse g^{*-1} then it must be unique, since if g_1^{*-1} is another inverse we have

$$g * (g^{*-1} - g_1^{*-1}) = (g * g^{*-1}) - (g * g_1^{*-1}) = 0$$

Conversely, suppose that (7.3) has a solution for any right-hand x. Then it has a solution for $x = \delta$ and the solution is by definition the inverse of g. Consequently, we can say that, if g has an inverse in $\mathcal{A}'$, then the equation has a unique solution for any right-hand side x and the solution is

$$y = g^{*-1} * x\,. \tag{7.4}$$

Therefore, knowledge of g^{*-1} permits to find the solution of (7.3) for any right-hand side x. For this reason g^{*-1} is called the *elementary* or *fundamental solution* of the convolution equation.

Note that if g has an inverse g^{*-1}, but it's not an element of the convolution algebra $\mathcal{A}'$, then the expression $g^{*-1} * x$ may not exist and $g^{*-1} * g * y$ may not be associative (see Example 3.5). Hence, (7.4) can not be proved to be equivalent to (7.3).

Suppose that g_1 and g_2 are two elements of the convolution algebra $\mathcal{A}'$ having inverses g_1^{*-1} and g_2^{*-1}, respectively. Then their convolution product $g_1 * g_2$ has an inverse as well and it is given by

$$(g_1 * g_2)^{*-1} = g_1^{*-1} * g_2^{*-1} \tag{7.5}$$

for

$$\begin{aligned}(g_1 * g_2)^{*-1} * (g_1 * g_2) &= \delta\\ &= g_1 * g_1^{*-1} * g_2 * g_2^{*-1}\\ &= (g_1^{*-1} * g_2^{*-1}) * (g_1 * g_2)\,.\end{aligned}$$

From this we see that, if in (7.3) g can be represented as the convolution product of m invertible elements $g_i, i = 1, \ldots, m$, then the solution of the equation can be expressed as the convolution product of their inverses

$$y = g_1^{*-1} * \ldots * g_m^{*-1} * x\,. \tag{7.6}$$

In every algebra with a unit, one can perform a partial fraction expansion and every convolution algebra has a unit by definition. Therefore, every convolution product of inverses can be represented as a sum of inverses.

Example 7.1: Partial Fraction Expansion

Consider the following convolution product

$$(D\delta + a\delta)^{*-1} * (D\delta - b\delta)^{*-1}$$

with a and b different constants. Its partial fraction expansion has the form

$$c_a(D\delta + a\delta)^{*-1} + c_b(D\delta - b\delta)^{*-1}$$

with c_a and c_b constants to be determined. If we take the convolution of both expressions with

$$(D\delta + a\delta) * (D\delta - b\delta)$$

we obtain the following equation

$$\delta = c_a(D\delta - b\delta) + c_b(D\delta + a\delta)\,.$$

Equating the coefficients of δ and $D\delta$ we obtain two equations for c_a and c_b whose solution is

$$c_b = -c_a = \frac{1}{a+b}\,.$$

A (convolution) algebra is said to be free from *zero divisors* if

$$g_1 * g_2 = 0$$

implies that either $g_1 = 0$ or $g_2 = 0$. In this case the algebra is called an *integral domain* and a convolution equation with common factors on both sides of the equation can be simplified. For example, assuming $f \neq 0$, the equation

$$f * g * y = f * x$$

can be simplified to

$$g * y = x\,.$$

In fact, the original equation can be written as

$$f * (g * y - x) = 0$$

and since f is different from zero, we can deduce the simplified form.

We will see that the convolution algebra of right-sided distributions $\mathcal{D}'_+$ is an integral domain. The algebra of periodic distributions $\mathcal{D}'(\mathbb{T})$ is not.

7.3 Initial Value Problems

In this section we want to apply the results of the previous section to study initial value problems. In particular let L denote the linear differential operator with constant coefficients of order m

$$L = D^m + a_{m-1}D^{m-1} + \cdots + a_1 D + a_0$$

where for convenience we have set $a_m = 1$. We are interested in the solution of the differential equation

$$Ly(t) = x(t) \tag{7.7}$$

for $t \geq 0$ with initial conditions

$$(D^k y)(0) = y_k \qquad k = 0, \ldots, m-1 \tag{7.8}$$

x and y functions and differentiation intended in the usual sense of differentiation of functions.

As a first step in translating this problem into the language of distributions, we note that the convolution algebra $\mathcal{D}'_+$ is well suited for the study of initial value problems. Every element of the algebra can be thought of as being in a zero state for $t < 0$ and representing some excitation or state evolution for $t \geq 0$. The functions x and y can be associated with distributions of $\mathcal{D}'_+$ by extending them to negative values of t where we assign them the value of zero. To make this explicit it's usual to show them multiplied by the unit step $\mathsf{1}_+$.

The second step is to perform differentiation in the sense of distributions. With the results of Example 2.9, for the first derivative of $\mathsf{1}_+ y$ we have

$$D(\mathsf{1}_+(t)y(t)) = \mathsf{1}_+(t)Dy(t) + y_0\delta$$

and similarly, for the higher order derivatives

$$\begin{aligned} D^2(\mathsf{1}_+(t)y(t)) &= \mathsf{1}_+(t)D^2y(t) + y_0 D\delta + y_1\delta \\ &\cdots \\ D^m(\mathsf{1}_+(t)y(t)) &= \mathsf{1}_+(t)D^m y(t) + y_0 D^{m-1}\delta + \cdots + y_{m-1}\delta\,. \end{aligned}$$

Note that in all these expressions the first term on the right-hand side is the conventional derivative of the function y (multiplied by $\mathsf{1}_+$). Putting these results in the differential equation we obtain an equivalent equation for the distribution $\mathsf{1}_+ y$

$$\begin{aligned} L(\mathsf{1}_+ y) &= \mathsf{1}_+ Ly + \sum_{k=0}^{m-1} \sigma_k D^k\delta \\ &= \mathsf{1}_+ x + \sum_{k=0}^{m-1} \sigma_k D^k\delta \end{aligned}$$

with

$$\begin{aligned} \sigma_k &= a_{1+k}y_0 + a_{2+k}y_1 + \cdots + y_{m-k-1} \\ &= \sum_{i=0}^{m-1-k} a_{i+1+k}y_i\,, \qquad k = 0, \ldots, m-1 \end{aligned} \tag{7.9}$$

and $a_m = 1$.

The last step required to translate the initial value problem into a convolution equation is to use the fact that the kth order derivative of a distribution can be expressed as the convolution product with $D^k\delta$ so that

$$L(\mathsf{1}_+ y) = L\delta * \mathsf{1}_+ y\,.$$

The initial value problem defined by (7.7) and (7.8) is therefore equivalent to the following convolution equation of distributions

$$L\delta * \mathsf{1}_+ y = \mathsf{1}_+ x + \sum_{k=0}^{m-1} \sigma_k D^k \delta\,. \tag{7.10}$$

With the results of the previous section, if the distribution $L\delta$ has an inverse in $\mathcal{D}'_+$ (the elementary solution of the equation), the solution of the equation for arbitrary right-hand side $\mathsf{1}_+ x$ and initial conditions is given by

$$\mathsf{1}_+ y = (L\delta)^{*-1} * \mathsf{1}_+ x + \sum_{k=0}^{m-1} \sigma_k D^k \left[(L\delta)^{*-1}\right]\,. \tag{7.11}$$

It's worth highlighting two important points. The first one is the fact that the differential equation (7.7) is not a full description of the problem. To fully specify the problem it has to be accompanied by the initial conditions expressed by Eq. (7.8). Differently from this, the convolution Eq. (7.10) is a *full* description of the problem.

The second point that we want to highlight is the fact that (7.11) is a *global* solution of the problem, that is, the solution is specified for all times. Differently from this, the classical solution of the original initial value problem is a function only valid for $t \geq 0$.

Next we show that the inverse $(L\delta)^{*-1}$ exists. To this end note that if we insert it in (7.10) and set $x = 0$ as well as $\sigma_0 = 1$ and $\sigma_k = 0$, $k = 1, \ldots, m-1$ we obtain the equation *defining* the inverse

$$L\delta * (L\delta)^{*-1} = \delta\,.$$

The inverse of $L\delta$ is thus the distribution $\mathsf{1}_+ e$ with e the function which is the solution of the homogeneous equation

$$Le(t) = 0$$

with initial conditions

$$D^{m-1}e(t) = 1 \qquad \text{and} \qquad D^k e(t) = 0\,, \quad k = 0, \ldots, m-2\,.$$

Example 7.2: Fundamental Solution

Consider the differential operator

$$L = D + a.$$

The solution of the homogeneous differential equation $Le(t) = 0$ with initial condition $e(0) = 1$ is

$$e(t) = e^{-at}\,.$$

The inverse of $L\delta$ in the convolution algebra $\mathcal{D}'_+$ is therefore

$$(L\delta)^{*-1} = (D\delta + a\delta)^{*-1} = 1_+(t)e^{-at}\,.$$

This is easily verified by inserting it into the convolution equation for the operator L

$$\begin{aligned} L\delta * (L\delta)^{*-1} &= (D\delta + a\delta) * 1_+(t)e^{-at} = D(1_+(t)e^{-at}) + a1_+(t)e^{-at} \\ &= -a1_+(t)e^{-at} + \delta + a1_+(t)e^{-at} = \delta\,. \end{aligned}$$

In a similar way we find

$$(D\delta + a)^{-m} = 1_+(t)\frac{t^{m-1}}{(m-1)!}e^{-at}$$

with m a positive natural number.

Let's focus for a moment on the distribution $L\delta$ and observe that it looks like a polynomial P with $D\delta$ playing the role of the independent variable

$$L\delta = D^m\delta + a_{m-1}D^{m-1}\delta + \cdots + a_1 D\delta + a_0\delta\,.$$

Any polynomial can be represented as a product of factors

$$P(z) = (z - z_1)(z - z_2)\cdots(z - z_m)$$

with z_j the zeros that may or may not be distinct. From this and remembering that

$$D^k\delta * D^i\delta = D^{k+i}\delta$$

we deduce that the distribution $L\delta$ can be factored in a similar way. If we denote by f^{*k} the convolution product of $k \geq 0$ distributions equal to f with $f^{*0} = \delta$ and group common factors, then $L\delta$ can be represented as

$$L\delta = (D\delta - z_1\delta)^{*l_1} * (D\delta - z_2\delta)^{*l_2} * \cdots * (D\delta - z_n\delta)^{*l_n}$$

with l_j the multiplicity of the jth factor. The inverse $(L\delta)^{*-1}$ can then also be factored

$$(L\delta)^{*-1} = (D\delta - z_1\delta)^{*-l_1} * (D\delta - z_2\delta)^{*-l_2} * \cdots * (D\delta - z_n\delta)^{*-l_n}$$

f^{*-k} denoting the inverse of f^{*k}. With this factorization the elementary solution can either be directly expressed as a convolution product

$$\mathsf{1}_+(t)\, e(t) = \mathsf{1}_+(t)\frac{t^{l_1-1}}{(l_1-1)!}e^{z_1 t} * \cdots * \mathsf{1}_+(t)\frac{t^{l_n-1}}{(l_n-1)!}e^{z_n t}$$

or, by first performing a partial fraction expansion, can be expressed as a sum of convolution-free known distributions.

To show the relation to the Laplace method, we Laplace transform Eq. (7.10). The Laplace transform of the distribution $L\delta$ becomes a true polynomial in the variable s and the convolution product becomes the conventional multiplication so that the convolution equation becomes an algebraic equation

$$\begin{aligned} P(s)\, Y(s) &= X(s) + \sum_{k=0}^{m-1} \sigma_k s^k \\ P(s) &= (s^m + a_{m-1}s^{m-1} + \cdots + a_1 s + a_0) \\ &= (s - z_1)^{l_1}(s - z_2)^{l_2} \cdot \cdots \cdot (s - z_n)^{l_n}\,. \end{aligned}$$

The Laplace transformed of the inverse $(L\delta)^{*-1}$ is the reciprocal of $P(s)$ and corresponds to the Laplace transform of the elementary solution e

$$E(s) = \frac{1}{P(s)}.$$

With it the solution of the convolution equation can be written as

$$Y(s) = E(s)X(s) + E(s)\sum_{k=0}^{m-1} \sigma_k s^k\,.$$

The solution y of the original equation is then found by inverse Laplace transforming Y. In most cases this is most conveniently accomplished by partial fraction expansion.

This shows the parallelism between convolution equations in $\mathcal{D}'_+$ on one side and the Laplace transform method on the other one. In particular the distribution $D\delta$ is the time-domain counterpart of the variable s, the convolution product the counterpart of the ordinary multiplication and δ the one of the multiplicative unit element 1.

Example 7.3

Consider the differential equation

$$\left[D^2 + (a-b)D - ab\right] y(t) = x(t)$$

with initial conditions $(Dy)(0) = y(0) = 0$ and assume that a and b are different constants. The corresponding convolution equation

$$(D\delta + a\delta) * (D\delta - b\delta) * y = x$$

has as elementary solution the convolution product

$$e = (D\delta + a\delta)^{*-1} * (D\delta - b\delta)^{*-1}$$

with partial fraction expansion (see Example 7.1)

$$e = \frac{1}{a+b}\left[-(D\delta + a\delta)^{*-1} + (D\delta - b\delta)^{*-1}\right].$$

The inverse elements appearing in e were calculated in Example 7.2. Using those results we can express the elementary solution of the equation as

$$e(t) = \frac{1}{a+b}\left[-\mathsf{1}_+(t)\, e^{-at} + \mathsf{1}_+(t)\, e^{bt}\right].$$

If we Laplace transform the equation, the procedure is completely parallel. The Laplace transformed of the elementary solution is

$$E(s) = \frac{1}{a+b}\left[\frac{-1}{s+a} + \frac{1}{s-b}\right]$$

and by inversion we obtain the same distribution e.

We have seen that the initial value problem described by (7.7) and (7.8) can equivalently be described by the convolution (7.10). While the differential equation only has a meaning if x is a continuous function with isolated jump discontinuities, the convolution equation remain well-defined if $\mathsf{1}_+ x$ is replaced by any distribution in $\mathcal{D}'_+$. In particular, we can consider more general convolution equations of the form

$$L\delta * y = N\delta * x + \sum_{k=0}^{m-1} \sigma_k D^k \delta$$

with

$$N = b_n D^n + b_{n-1} D^{n-1} + \cdots + b_0 \,,$$

x any distribution in $\mathcal{D}'_+$ and where it's understood that the solution y must belong to the convolution algebra of distributions in $\mathcal{D}'_+$. As before, the solution of the equation is found by convolving with the convolutional inverse of $L\delta$

$$y = (L\delta)^{*-1} * N\delta * x + \sum_{k=0}^{m-1} \sigma_k (L\delta)^{*-1} * D^k\delta \,.$$

We want to establish if it's possible to replace the second summand on the right-hand side, representing the initial conditions, by a suitably selected input signal composed by a weighted sum of a Dirac pulse and it's derivatives, such that, in the complement of $t = 0$, the solution y remains unchanged. To this end its convenient to consider the Laplace transformed of y

$$Y(s) = \frac{Z(s)}{P(s)} X(s) + \frac{\sum_{k=0}^{m-1} \sigma_k s^k}{P(s)}$$

with $Z = \mathcal{L}\{N\delta\}$ a polynomial of degree n and the other symbols having the same meaning as before. The Laplace transform of the sought for input signal is a polynomial

$$X(s) = x_q s^q + \cdots + x_0$$

and it must be selected in such a way as to satisfy the equality

$$\frac{Z(s)}{P(s)} X(s) = \frac{\sum_{k=0}^{m-1} \sigma_k s^k}{P(s)} + W(s)$$

with $W(s)$ another polynomial. This polynomial corresponds also to a weighted sum of Dirac pulses and its derivatives, and hence only changes y at $t = 0$, which we allow.

The conditions for the existence of such an input signal $X(s)$ can be determined with the help of the division theorem of polynomials. It states that, given polynomials $Q(s)$ and $P(s) \neq 0$, there are unique polynomials $R(s)$ and $W(s)$ satisfying

$$Q(s) = P(s)W(s) + R(s)$$

with the degree of $R(s)$ being lower than the one of $P(s)$ [21]. From this theorem we deduce that, provided $Z(s)$ and $P(s)$ are *relatively prime*, that is, assuming that they have no common factors, we can select $X(s)$ so that the rest of the division of $Z(s)X(s)$ by $P(s)$ corresponds to $\sum_{k=0}^{m-1} \sigma_k s^k$. To achieve this we need m degrees of freedom, one for each σ_k. In other words, the input polynomial $X(s)$ must have degree $m-1$. Then we can choose the coefficients of $X(s)$ in such a way as to obtain the desired values for the rest of the division.

If $Z(s)$ and $P(s)$ have a common factor $K(s)$ then

$$\begin{aligned}\frac{Z(s)X(s)}{P(s)} &= \frac{K(s)Z'(s)X(s)}{K(s)P'(s)} = \frac{K(s)}{K(s)}\left(\frac{R(s)}{P'(s)} + W(s)\right)\\ &= \frac{K(s)R(s)}{P(s)} + W(s)\end{aligned}$$

and we see that the rest of the division $K(s)R(s)$ has a constrained form that can't be made to match $\sum_{k=0}^{m-1} \sigma_k s^k$ for arbitrary σ_ks.

We have therefore established that, in a convolution equation derived from an initial value problem, the terms representing the initial conditions can be replaced by a distribution x composed by a weighted sum of a Dirac pulse and its derivatives if and only if $Z(s)$ and $P(s)$ have no common factors. If we perform this substitution, in the complement of $t=0$, the solution of the equation y remains unchanged.

Example 7.4: Replacing Initial Conditions

Consider the initial value problem

$$\begin{aligned}\left(D^2 + a_1 D + a_0\right) y &= (b_1 D + b_0)\, x\\ (Dy)(0) = y_1, \qquad y(0) &= y_0\,.\end{aligned}$$

The corresponding convolution equation is

$$(D\delta^2 + a_1 D\delta + a_0\delta) * y = (b_1 D\delta + b_0) * x + y_0 D\delta + (a_1 y_0 + y_1)\delta\,.$$

Our objective is to replace the initial conditions by an input signal composed by a Dirac pulse and its derivatives so that in the complement of $t=0$ the solution y of the convolution equation with this input signal is identical to the solution of the equation with initial conditions and no input signal.

Expressed in the Laplace domain the problem is thus to find the coefficients of the polynomial

$$X(s) = x_1 s + x_0$$

such that

$$\frac{Z(s)X(s)}{P(s)} = \frac{R(s)}{P(s)} + W(s)$$

with

$$Z(s) = b_1 s + b_0, \qquad P(s) = s^2 + a_1 s + a_0, \qquad R(s) = y_0 s + a_1 y_0 + y_1$$

and $W(s)$ an arbitrary polynomial of degree lower than 2. By performing the polynomial division of the left-hand side of the equation we obtain

$$\frac{s(-a_1b_1x_1+b_0x_1+b_1x_0)-a_0b_1x_1+b_0x_0}{s^2+a_1s+a_0}+b_1x_1\,.$$

Thus $W(s)=b_1x_1$ and, by comparing coefficients of this expression with the right-hand side of the equation, the coefficients of $X(s)$ are found to be

$$\begin{aligned}x_0&=\frac{(a_1b_1-b_0)y_1+[(a_1^2-a_0)b_1-a_1b_0]y_0}{a_0b_1^2-a_1b_0b_1+b_0^2}\\ x_1&=\frac{b_1y_1+(a_1b_1-b_0)y_0}{a_0b_1^2-a_1b_0b_1+b_0^2}\,.\end{aligned}$$

This solution is well-defined except when the denominator, which is the same for both x_1 and x_0, becomes zero. This happens when

$$a_1=\frac{a_0b_1}{b_0}+\frac{b_0}{b_1}.$$

In this case the polynomial $Z(s)$ becomes a factor of $P(s)$

$$s^2+(\frac{a_0b_1}{b_0}+\frac{b_0}{b_1})2+a_0=(b_1s+b_0)(\frac{1}{b_1}s+\frac{a_0}{b_0})$$

in accordance with our general treatment of the problem.

Before concluding this section we show the important fact mentioned before that *the convolution algebra $\mathcal{D}'_+$ has no zero divisors*. To see this, consider a test function ϕ that is real-valued and positive everywhere on its support, for example β_ν from Example 2.1. We call such a function a positive test function. In Chap. 3 we saw that every distribution can be represented as the limit of a sequence of indefinitely differentiable functions. Let (g_m) and (y_m) be such sequences converging to g and y respectively and, for simplicity, assume that all functions are real-valued. Then, for every m there exists an open interval U contained in the support of g_m where, for every positive test function ζ with support in U, its value always has the same sign, for example positive

$$\langle g_m,\zeta\rangle>0\,.$$

We can make a similar construction for y_i as well. In addition, we can introduce a parameter λ such that

$$\lambda\mapsto\langle y_i(\tau),\phi(\tau+\lambda)\rangle$$

is a positive (or negative) test functions of λ with support in U. Then, assuming again a positive sign,

$$\langle g_m * y_i, \phi\rangle = \langle g_m(\lambda), \langle y_i(\tau), \phi(\tau + \lambda)\rangle\rangle$$

must be positive for every m and i and, with the continuity of distributions and convolution, so must be the limit. Consequently, if $g * y$ vanish for every test function then either g or y must be the zero distribution.

7.4 Integro-Differential Equations

Some initial value problems are naturally formulated as ordinary integro-differential equations

$$\begin{aligned}&D^m y(t) + a_{m-1} D^{m-1} y(t) + \cdots + a_1 D y(t) + a_0 y(t)\\&+ a_{-1}\int_0^t y(\tau_1)\,d\tau_1 + \cdots + a_{-n}\int_0^t \cdots \int_0^{\tau_{n-1}} y(\tau_n)\,d\tau_n \cdots d\tau_1\\&= x(t)\end{aligned}$$

with initial conditions

$$(D^k y)(0) = y_k \qquad k = 0, \ldots, m-1\,. \tag{7.12}$$

We still need initial conditions, but this time only m of them as the remaining information is included in the integrals.

These problems can be converted into convolution equations in the convolution algebra $\mathcal{D}'_+$ in a similar way as we discussed before. The new terms are the ones that are expressed as integrals and these can be written as convolution products

$$\begin{aligned}\int_0^t y(\tau_1)\,d\tau_1 &= 1_+(t) * 1_+(t)y(t)\\&\cdots\\\int_0^t \cdots \int_0^{\tau_{n-1}} y(\tau_n)\,d\tau_n \cdots d\tau_1 &= 1_+^{*n}(t) * 1_+(t)y(t)\,.\end{aligned}$$

The corresponding convolution equation is therefore

$$\begin{aligned}&\big(D^m\delta + a_{m-1}D^{m-1}\delta + \cdots + a_1 D\delta + a_0\delta\\&\quad + a_{-1}1_+ + \cdots + a_{-n}1_+^{*n}\big) * 1_+(t)y(t)\\&= 1_+(t)x(t) + \sum_{k=0}^{m-1} \sigma_k D^k\delta\end{aligned}$$

with $\sigma_k, k = 0, \dots, m-1$ as defined in (7.9).

As we have seen, the convolution algebra $\mathcal{D}'_+$ is an integral domain. For this reason we can multiply both sides of the equation with a non-zero distribution without changing the result. If we choose $D^n\delta$ as the distribution and make use of the fact that $\mathsf{1}_+(t)$ is the inverse of $D\delta$

$$D\delta * \mathsf{1}_+ = \delta$$

the equation becomes

$$\begin{aligned}&\left(D^{m+n}\delta + a_{m-1}D^{m-1+n}\delta + \cdots a_{-n}\delta\right) * \mathsf{1}_+(t)y(t)\\ &= D^n\delta * \mathsf{1}_+(t)x(t) + \sum_{k=0}^{m-1}\sigma_k D^{k+n}\delta\,.\end{aligned}$$

This is the type of convolution equation that we discussed in Sect. 7.3 and is solved by the same method. The solution of integro-differential equations thus requires no new technique.

The procedure of transforming the convolution equation that we just discussed is similar to the standard procedure used to convert an integro-differential equation into a differential equation by differentiating the equation. The key difference is that, while the former handles initial conditions automatically, the latter method requires extraction of additional conditions from the original equation.

7.5 Periodic Solutions

One is often interested in periodic solutions of differential equations. These solutions are most conveniently found with the help of the convolution algebra of periodic distributions.

Consider again the convolution equation obtained from the differential operator L of Sect. 7.3 where now the unit element of the algebra is the Dirac comb $\delta_{\mathcal{T}}$

$$L\delta_{\mathcal{T}} * y = x\,.$$

In Sect. 4.5 we established two important properties of the Fourier series:

1. The first one being that the Fourier coefficients of the convolution product of two Fourier series is equal to the product of the coefficients of the individual series times the period (Eq. (4.20)).
2. The second one being the fact that differentiation corresponds to multiplication of the kth Fourier coefficient by the factor $\jmath k\omega_c$ with $\omega_c = 2\pi/\mathcal{T}$. For this reason the kth Fourier coefficient of the distribution $L\delta_{\mathcal{T}}$ is proportional to a polynomial P evaluated at $\jmath k\omega_c$

$$\begin{aligned} c_k(L\delta_{\mathcal{T}}) &= \left[(\jmath k\omega_c)^m + a_{m-1}(\jmath k\omega_c)^{m-1} + \cdots + a_1(\jmath k\omega_c) + a_0\right]\frac{1}{\mathcal{T}} \\ &= P(\jmath k\omega_c)\frac{1}{\mathcal{T}}\,. \end{aligned}$$

By representing both x and y by their respective Fourier series and using these two properties, we can transform the above convolution equation into algebraic equations for the Fourier coefficients. Let's denote by c_k the kth Fourier coefficient of x and by d_k the one of y. Then the equation becomes

$$(\jmath k\omega_c - z_1)^{l_1}(\jmath k\omega_c - z_2)^{l_2} \cdot \cdots \cdot (\jmath k\omega_c - z_n)^{l_n}\, d_k = c_k$$

where, as before, we have expressed the polynomial P by its zero factors. To solve the equation we have to distinguish three cases:

1. If one or more zeros z_j of the polynomial equals $\jmath k\omega_c$ for some integer k and the coefficient c_k of x is different from zero, then the equation has no solution.
2. If one or more zeros z_j of the polynomial equals $\jmath k\omega_c$ for some integer k and the coefficient c_k is zero, then the equation has an infinity of solutions. In fact in this case d_k can be any number. Note also that if z_j is equal to $\jmath k\omega_c$ then the convolution product
$$L\delta_{\mathcal{T}} * e^{\jmath k\omega_c t} = 0$$
vanishes, which means that the convolution algebra of periodic distributions has zero divisors.
3. If no zero z_j equals $\jmath k\omega_c$ for any value of k then the equation has the unique solution given by the Fourier series with coefficients
$$d_k = \frac{c_k}{P(\jmath k\omega_c)}\,.$$

Example 7.5: Cont. of Example 7.3

We look for a periodic solution of the convolution equation of Example 7.3

$$(D\delta_{\mathcal{T}} + a\delta_{\mathcal{T}}) * (D\delta_{\mathcal{T}} - b\delta_{\mathcal{T}}) * y = x$$

assuming that the real part of a and b are both positive. In particular, we are interested in the elementary solution e of the equation. By setting $x = \delta_{\mathcal{T}}$ and expanding it by its Fourier series we obtain the following equation for the kth Fourier coefficient of e

$$e_k = \frac{1}{\mathcal{T}} \cdot \frac{1}{(\jmath k\omega_c + a)(\jmath k\omega_c - b)}\,.$$

By performing a partial fraction expansion and with the help of (4.24), we recognize them as the coefficients of the Fourier series of the distribution

$$e(t) = g(t) * \delta_{\mathcal{T}}$$

with

$$g(t) = \frac{-1}{a+b}\left[\mathsf{1}_+(t)\, e^{-at} + \mathsf{1}_+(-t)\, e^{bt}\right].$$

In fact the Fourier transform of g is

$$\begin{aligned} \hat{g}(\omega) &= \frac{1}{a+b}\left[\frac{-1}{\jmath\omega + a} + \frac{1}{\jmath\omega - b}\right] \\ &= \frac{1}{(\jmath\omega + a)(\jmath\omega - b)}. \end{aligned}$$

Note that g is a distribution of slow growth. The elementary solution of the equation in the algebra of periodic distributions is therefore the sum of periodically shifted tempered solutions of the differential equation.

Suppose now that we are interested in the solution for $x(t) = Ae^{\jmath\omega_c t}$. The only Fourier coefficient of x different from zero is $c_1 = A$. The Fourier coefficients of y are then also all zero except for

$$d_1 = \mathcal{T}\, c_1\, e_1 = A\, \hat{g}(\omega_c).$$

In this case the solution y of the equation is therefore

$$y(t) = A\, \hat{g}(\omega_c)\, e^{\jmath\omega_c t}.$$

7.6 General Convolution Equations

7.6.1 General Solutions

In this section we consider generic convolution equations of the form

$$g * y = x$$

with g, y and x generic distributions in $\mathcal{D}'$. Here the situation is different from when working in a convolution algebra. First the convolution between g and y may not exit. To guarantee its existence g must have compact support. This includes many important cases, for example, all linear differential operators with constant coefficients.

Second, g may not have an inverse. For example, if $g \in \mathcal{D}$ we have seen in Sect. 3.2 that $g * y \in \mathcal{E}$ and so can't equal δ for any $y \in \mathcal{D}'$. If it does, then the equation has an elementary solution, but it only serves to find solutions for x having compact support, otherwise the last convolution in

$$y = g^{*-1} * g * y = g^{*-1} * x$$

may not make sense.

Further, the homogeneous equation

$$g * y = 0$$

may have solutions different from $y = 0$. For this reason there may be an infinity of elementary solution, two of them differing by a solution of the homogeneous equation.

Despite these facts, general convolution equations have many practical applications.

Example 7.6: Electrostatics

Let ρ denote the electric charge density and u the electrostatic potential, both functions of the position in space. In empty space the two quantities are related by Poisson's equation

$$\Delta u(x) = -\frac{\rho(x)}{\epsilon_0}$$

with Δ the Laplace operator, $x \in \mathbb{R}^3$ the vector specifying position and ϵ_0 the permittivity of free space. This equation can be written as a convolution equation

$$\Delta\delta * u = -\frac{\rho(x)}{\epsilon_0}\,.$$

One can show that the inverse of $\Delta\delta$ is

$$-\frac{1}{4\pi\,|x|}\,.$$

If the charge density ρ is distributed over a finite region $\Omega \subset \mathbb{R}^3$ then the generated potential is

$$u(x) = \frac{1}{4\pi\epsilon_0\,|x|} * \rho(x)\,.$$

The homogeneous equation has solutions different from the trivial one: the so-called harmonic functions.

7.6.2 Tempered Solutions

If x is tempered and one is interested in tempered solutions of the equation then the convolution equation has a sense not only for g of compact support, but for the larger class of distributions of rapid descent $\mathcal{O}'_C$ [16]. This case is particularly important because one can then use the Fourier transform which may make it easier to find a solution.

In the following we briefly consider the one dimensional case where g is a linear differential operator with constant coefficients L and the convolution equation has the form

$$L\delta * y = x.$$

In this case there always is at least an elementary solution. If we Fourier transform both sides of the equation we find the equivalent equation

$$P\,\hat{y} = \hat{x}$$

with P a polynomial (and thus in $\mathcal{O}_M$).

If P has no zeros, then the only solution of the homogeneous equation is the trivial one and the inverse of P is a function of slow growth $1/P \in \mathcal{O}_M$. The only elementary solution of the equation is therefore the summable distribution

$$e = \mathcal{F}^{-1}\{\frac{1}{P}\}.$$

If P has a zero at ω_p then the homogeneous equation has nontrivial solutions. In particular, we saw Sect. 2.5.1 that if the multiplicity of the zero is k then the sums

$$\sum_{m=0}^{k-1} c_m\, D^m \delta(\omega - \omega_p)$$

with c_m constants, are all solutions of the Fourier transformed homogeneous equation $P\,\hat{y} = 0$. The solutions of the original homogeneous equation are found by inverse Fourier transformation to be

$$\sum_{m=0}^{k-1} \frac{c_m}{2\pi}\,(-\jmath t)^m\, e^{\jmath\omega_p t}\,.$$

The equation has therefore an infinity of elementary solutions. In addition, since $1/P \notin \mathcal{O}_M$, the solutions are not summable distributions.

Note that the equation may have non-tempered solutions that are not captured by Fourier transform techniques.

Example 7.7: Cont. of Example 7.3

We look for a tempered solution of the convolution equation of Example 7.3

$$(D\delta + a\delta) * (D\delta - b\delta) * y = x$$

assuming that the real part of a and b are both positive. A tempered elementary solution is easily found by solving the Fourier transformed the equation

$$\hat{e}(\omega) = \frac{1}{P(\omega)} = \frac{1}{(\jmath\omega + a)(\jmath\omega - b)}\,.$$

and determining its inverse

$$e(t) = \frac{-1}{a+b}\left[1_+(t)\,e^{-at} + 1_+(-t)\,e^{bt}\right]\,.$$

Note that despite the similarity between $\hat{e}(\omega)$ and $E(s)$ of Example 7.3 the tempered elementary solution is different from the solution found in the convolution algebra $\mathcal{D}'_+$.

Since $P(\omega)$ has no zeros, e is the only elementary tempered solution of the equation. Other solutions obtained by adding any linear combination of the solutions of the homogeneous equation (e^{-at} and e^{bt}) growth exponentially as t tends either to ∞ or to $-\infty$ and are therefore not tempered distributions.

7.7 Systems of Convolution Equations

One often has to solve a set of n simultaneous equations in n unknown distributions $y_1, \ldots, y_n$

$$\begin{aligned}
g_{11} * y_1 + g_{12} * y_2 + \cdots + g_{1n} * y_n &= x_1\\
g_{21} * y_1 + g_{22} * y_2 + \cdots + g_{2n} * y_n &= x_2\\
\vdots \qquad\qquad\qquad &\ \vdots\ \vdots\\
g_{n1} * y_1 + g_{n2} * y_2 + \cdots + g_{nn} * y_n &= x_n
\end{aligned}$$

with g_{jm} coefficients distributions, $x_1, \ldots, x_n$ right-hand side distributions and where all distributions belong to a distribution algebra $\mathcal{A}'$. This system of equations can conveniently be written in matrix form

$$G * Y = X \tag{7.13}$$

with G the $n \times n$ matrix with elements g_{jm} and Y, X *vector valued distributions* (column matrices) with elements y_j and x_j respectively. The space of vector valued distributions is denoted by $\mathcal{D}'(\mathbb{R}^m, \mathbb{C}^n)$ and application of a test function $\phi \in \mathcal{D}(\mathbb{R}^m)$ to a vector X is defined as the application of ϕ to each component individually

$$\langle X, \phi \rangle := \begin{bmatrix} \langle x_1, \phi \rangle \\ \vdots \\ \langle x_n, \phi \rangle \end{bmatrix} .$$

The determinant of the matrix G is defined as usual, with the convolution product replacing the standard product. It is a convolution belonging to the convolution algebra $\mathcal{A}'$. For example, the determinant of a 2×2 matrix G is

$$\det \begin{bmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{bmatrix} = g_{11} * g_{22} - g_{21} * g_{12} .$$

Suppose that the matrix G has an inverse G^{*-1}

$$G * G^{*-1} = \delta I$$

where δI is the identity matrix with the unit of $\mathcal{A}'$ on the diagonal and 0 everywhere else. If we compute the determinant of both sides of this equation we obtain

$$\det(G * G^{*-1}) = \det(G) * \det(G^{*-1}) = \det(\delta I) = \delta$$

from which we deduce that, if the matrix G has an inverse then $\det(G)$ has an inverse in $\mathcal{A}'$. Conversely, if $\det(G)$ has an inverse, then we can compute the inverse of G by

$$G^{*-1} = \det(G)^{*-1} * \tilde{G}^T$$

with $\tilde{G}$ the matrix of cofactors and $\tilde{G}^T$ its transpose.

We conclude that (7.13) has a solution for arbitrary right-hand side X if and only if $\det(G)$ has an inverse in $\mathcal{A}'$. The solution is given by

$$Y = G^{*-1} * X . \tag{7.14}$$

One shows in a similar way as for a single equation (see Sect. 7.2) that G^{*-1} and hence the solution of the equation is unique.

Chapter 8
Linear Time Invariant Systems

We assume the reader to have familiarity with linear time-invariant (LTI) systems. In this chapter we merely summarise the main results of this theory. We are going to call the quantities that are considered the input, the output and some characterization of the system *signals*. This should evoke a meaningful interpretation in most of the systems that we are going to discuss. Mathematically they are distributions.

8.1 Basic Definitions

The meaning of time-invariant is very intuitive: suppose that we apply the input signal $x(t)$ to a system represented by an operator $\mathcal{H}$ and observe the signal

$$y(t) = \mathcal{H}[x(t)]$$

as its output (Fig. 8.1). The system is said to be *time-invariant* if by applying the delayed input signal $x(t-\tau)$ we observe the same output signal as before, except for a delay in time by an amount τ, that is, if

$$y(t-\tau) = \mathcal{H}[x(t-\tau)]\,. \tag{8.1}$$

The concept of linearity is subtler. A defining property of a linear system is the validity of the *superposition principle*: if $y_1(t)$ is the response of the system to the input $x_1(t)$ and $y_2(t)$ the one to $x_2(t)$, then the response to a linear combination of these inputs is

$$\begin{aligned} y(t) &= \mathcal{H}[c_1x_1(t) + c_2x_2(t)] = c_1\mathcal{H}[x_1(t)] + c_2\mathcal{H}[x_2(t)] \\ &= c_1y_1(t) + c_2y_2(t) \end{aligned} \tag{8.2}$$

with c_1 and c_2 constants. However, if we limit the definition of a linear system to this property, then we admit pathological systems as the following one.

F. Beffa, *Weakly Nonlinear Systems*, Understanding Complex Systems,
https://doi.org/10.1007/978-3-031-40681-2_8

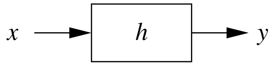


Fig. 8.1 Representation of a single-input single-output LTI system $\mathcal{H}$

Example 8.1: A Discontinuous System [22]

Consider a system accepting as input a piece-wise continuous function with at most a finite number of isolated jump discontinuities. The system response consists in the sum of the input signal jumps from $-\infty$ to the present time t.

The system satisfies (8.2). However, the behaviour is rather peculiar. If we apply, say, a rectangular input then the output is also rectangular. But, if we approximate to any degree of accuracy the rectangular input with a continuous function, then the output is always zero.

To exclude systems with such a bizarre behavior, we require linear systems to be *continuous*: if as $m \in \mathbb{N}$ tends to ∞ the sequence of input signals $x_m(t)$ converges (in the sense of distributions) to the signal $x(t)$, then the system response $y_m(t)$ corresponding to input $x_m(t)$ converges to the response $y(t)$ corresponding to $x(t)$.

Suppose that we apply an impulse $\delta(t)$ to the input of the system $\mathcal{H}$ and observe the signal $h(t)$ at its output. Then, by linearity, if we apply a finite number of pulses the output must be

$$\mathcal{H}[\sum_{j=1}^{n} a_j\,\delta(t-\tau_j)] = \sum_{j=1}^{n} a_j\,h(t-\tau_j) = h(t) * \sum_{j=1}^{n} a_j\,\delta(t-\tau_j)\,.$$

In Sect. 3.3 we saw that every distribution can be represented as the limit of a finite series of Dirac impulses. From this and the linearity of convolution (Eq. (3.19)) we obtain that, in the limit as n tends to infinity, if the input converges to the signal $x(t)$ the output of the system converges to

$$y(t) = h(t) * x(t)\,.$$

We therefore define

Definition 8.1 (*LTI System*) A single-input, single-output (SISO), linear time-invariant (LTI) system is a system that, when driven by an input signal $x(t)$ produces the output

$$y(t) = h(t) * x(t) \tag{8.3}$$

with $h(t)$ the *impulse response* of the system.

A system is called *real* if, when driven by a real distribution, its response is a real distribution. In other words, if its impulse response is a real distribution.

While we have been talking about signals depending on time, we can abstract from that and talk about signals depending on a generic n dimensional independent variable $\lambda \in \mathbb{R}^n$. In this case, instead of time-invariance, it makes more sense to adapt (8.1) to

$$y(\lambda - \tau) = \mathcal{H}[x(\lambda - \tau)]$$

and talk about *translation invariance*. A single-input single-output, linear translation-invariant system is then still described by a convolution product similar to (8.3) where however the independent variable t is replaced by the abstract n dimensional variable λ. We are going to call a system of this type an LTI system as well.

8.2 Causality

Assume for simplicity that h and x are integrable functions of time. The response of a system characterized by h when driven by the input x can then be written in integral form

$$y(t) = \int_{-\infty}^{\infty} h(\tau)\, x(t - \tau)\, d\tau = \int_{-\infty}^{\infty} h(t - \tau)\, x(\tau)\, d\tau\ .$$

Suppose now that the input vanishes for $t < 0$. Then from

$$y(t) = \int_{0}^{\infty} h(t - \tau)\, x(\tau)\, d\tau$$

we see that in general the system may produce a nonzero response $y(t)$ for $t < 0$, that is, before the input signal $x(t)$ has been applied.

If a system is *causal*, that is, if its output at time t_0 can only depend on values of the input signal at times $t \leq t_0$, then its impulse response $h(t)$ must vanish for $t < 0$. In other words h must be a right-sided distribution in $\mathcal{D}'_{+}$.

Note that in our interpretation of signals as being functions of time, non-causal systems are not physically implementable and appear to be meaning-less. However, non-causal systems are sometimes useful in theoretical studies. In addition, in many situations the theory of LTI systems can be applied to systems where the quantities of interest (the input and output) are not functions of time (see Example 7.6).

8.3 Stability

An important aspect of a system is its stability. Let $x(t)$ be a *bounded function*, that is, satisfying

$$\|x\|_\infty := \sup_{t\in\mathbb{R}} |x(t)| < \infty\,.$$

The response of a system characterized by the impulse response $h(t)$ to such an input signal is

$$y(t) = h(t) * x(t)\,.$$

The output $y(t)$ is well-defined if

$$\langle h * x, \phi\rangle < \infty$$

for every test function $\phi \in \mathcal{D}$ and for every sequence (ϕ_m) converging to zero in $\mathcal{D}$

$$\lim_{m\to\infty} \langle h * x, \phi_m\rangle = 0\,.$$

In this case we say that the system is *bounded-input bounded-output (BIBO) stable*.

For a system to be BIBO stable

$$\langle h(t) * x(t), \phi(t)\rangle = \langle h(t), \int_{\mathbb{R}} x(\tau)\phi(t+\tau)\,d\tau\rangle$$

must have a meaning. Observe that the inner integral is an indefinitely differentiable bounded function. For the convolution to have a meaning the impulse response of the system must therefore be extensible to a continuous linear form on $\mathcal{B}$. As we saw in Sect. 6.1 this is only the case if h is a summable distribution. Thus, *for a system to be BIBO stable, its impulse response must be a summable distribution.*

We mention without going into details that the definition of a BIBO stable system can be extended to input signals that are so-called bounded distributions and usually denoted by $\mathcal{B}'$ or $\mathcal{D}'_{L^\infty}$ [16].

The series connection, or cascade of two stable systems results in a stable system. This is so because the convolution of summable distributions is always well-defined and is itself a summable distribution. In addition, for linear systems the order of the connection is irrelevant as, if h_A and h_B are the impulse responses of the two systems

$$h_A * h_B = h_B * h_A\,.$$

8.4 Transfer Function

8.4.1 Stable Systems

If a system is stable then its impulse response h can be Fourier transformed and the transformed $\hat{h}$ is a continuous function of slow growth called the *frequency response* of the system. If the input signal x is also a summable distribution then it can also be Fourier transformed and the Fourier transform of the output signal can be represented by the product

$$\hat{y}(\omega) = \hat{h}(\omega)\hat{x}(\omega)\,. \tag{8.4}$$

If the input signal x is $\mathcal{T}$-periodic, then the system can be analysed in the convolution algebra of periodic distributions. To do so the impulse response h is converted in a periodic distribution by convolving it with the unit of the convolution algebra of periodic distributions $\delta_{\mathcal{T}}$

$$h_{\mathcal{T}} := h * \delta_{\mathcal{T}}\,.$$

Provided that $h_{\mathcal{T}}$ is well-defined, which for stable systems is always the case, then the output of the system can be represented by

$$y = h_{\mathcal{T}} * x\,.$$

Note that while the convolution used to define $h_{\mathcal{T}}$ is the convolution in $\mathcal{D}'(\mathbb{R})$, the latter is the convolution in $\mathcal{D}'(\mathbb{T})$. As discussed in Sect. 7.5, the equation is most conveniently solved with the help of the Fourier series. If we denote by $c_m(y)$, $c_m(h_{\mathcal{T}})$ and $c_m(x)$ the mth Fourier coefficient of y, $h_{\mathcal{T}}$ and x respectively, then the equation is solved if

$$c_m(y) = \mathcal{T} c_m(h_{\mathcal{T}}) c_m(x)$$

for every $m \in \mathbb{Z}$. From (4.24) we know that

$$c_m(h_{\mathcal{T}}) = \frac{\hat{h}(m\omega_c)}{\mathcal{T}}$$

with $\omega_c = 2\pi/\mathcal{T}$. Therefore, by knowing the Fourier transform of the impulse response we can immediately obtain the Fourier coefficients of the output signal by

$$c_m(y) = \hat{h}(m\omega_c) c_m(x)\,. \tag{8.5}$$

In particular, if the input is the complex tone $\mathrm{e}^{\jmath\omega_c t}$, the output is also a complex tone at the exact same frequency

$$y(t) = \hat{h}(\omega_c)\mathrm{e}^{\jmath\omega_c t}\,.$$

If the input of the system is the sum of two (or more) periodic signals x_A and x_B with incommensurate frequencies ω_A and ω_B, that is, if the ratio of the two frequencies ω_A/ω_B is an irrational number, then the input signal is not periodic, but *almost periodic*. Due to the linearity and continuity of the system, the response can still be calculated by the above technique for each input separately and the result combined

$$y(t) = \sum_{m=-\infty}^{\infty} \hat{h}(m\omega_A)c_m(x_A)\mathrm{e}^{\jmath m\omega_A t} + \hat{h}(m\omega_B)c_m(x_B)\mathrm{e}^{\jmath m\omega_B t}\,.$$

8.4.2 Causal Systems

If the system is causal, that is, if its impulse response h is a distribution in $\mathcal{D}'_+$, and one is interested in the system response for right-sided input signals $x \in \mathcal{D}'_+$, then the system response y can be calculated in the convolution algebra $\mathcal{D}'_+$. In particular, if h and x are Laplace transformable then the Laplace transformed of the output signal can be calculated by

$$Y(s) = H(s)X(s)\,. \tag{8.6}$$

The Laplace transformed $H(s)$ of the impulse response h is called the system *transfer function*.

If the system is BIBO stable, then the ROC of $H(s)$ includes the imaginary axis $s = \jmath\omega$. In this case the Fourier transformed of h is immediately obtained from the transfer function by

$$\hat{h}(\omega) = H(\jmath\omega)\,. \tag{8.7}$$

Note that if the system is not BIBO stable then this relation is not valid even if the Fourier transform of h does exits. See Example 5.4 for a simple example where the system corresponds to an ideal integrator.

In the following we are going to denote distributions belonging to $\mathcal{D}'_+ \cap \mathcal{D}'_{L^1}$ by $\mathcal{D}'_{L^1+}$.

8.5 Rational Transfer Functions

Consider a causal system described by a rational transfer function

$$H(s) = \frac{N(s)}{P(s)} = \frac{b_n s^n + b_{n-1}s^{n-1} + \cdots + b_0}{s^m + a_{m-1}s^{m-1} + \cdots + a_0}\,.$$

Given the Laplace transform $X(s)$ of the input signal x, the Laplace transformed of the output is

$$Y(s) = \frac{N(s)}{P(s)} X(s) \,.$$

If we multiply both sides of this equation by $P(s)$ we obtain

$$P(s)Y(s) = N(s)X(s)$$

and by inverse Laplace transforming the equation we obtain the convolution equation

$$\begin{aligned}&\big(D^n\delta + a_{n-1}D^{n-1}\delta + \cdots + a_0\delta\big) * y \\ &\quad = \big(b_m D^m\delta + b_{m-1}D^{m-1}\delta + \cdots + b_0\delta\big) * x \,.\end{aligned}$$

With the results of Sect. 7.3 we see that this equation corresponds to the initial value problem described by the linear differential equation with constant coefficients

$$Ly(t) = x_a(t)$$

with

$$\begin{aligned}L &= D^m + a_{m-1}D^{m-1} + \cdots + a_0, \\ x_a(t) &= (b_n D^n + b_{n-1}D^{n-1} + \cdots + b_0)x(t)\end{aligned}$$

and zero initial conditions

$$(D^k y)(0) = 0, \quad k = 0, \cdots, m-1 \,.$$

For this reason $y(t) = h(t) * x(t)$ is called the *zero state response* of the system.

It is obvious that the procedure can be reversed. We have therefore established a one-to-one correspondence between systems described by a rational transfer function and systems described by a linear differential equation with constant coefficients and zero initial conditions.

If the transfer function H of the system is *minimal*, that is, if its numerator and its denominator are relatively prime polynomials, then, in the complement of $t = 0$, it is possible to recreate the same output that would be produced by solving the corresponding initial value problem with *non-zero initial conditions*. This is achieved by driving the system with an input signal consisting of a weighted sum of a Dirac pulse and its derivatives

$$x = x_{m-1}D^{m-1}\delta + \cdots + x_0\delta$$

and by suitably selecting the weighting coefficients $x_0, \dots, x_{m-1}$ as described in Sect. 7.3 (see Example 7.4). Such a system is said to have *order* m and to be *observable* and *controllable* (see Sect. 8.6).

If $H(s)$ is a proper rational transfer function, that is if $m < n$, then it can be expanded into a sum of partial fractions of the form

$$\frac{c_{jk_j}}{(s - p_j)^{k_j}}, \qquad k_j = 1, \dots, l_j$$

with p_j the jth zero of $P(s)$, l_j its multiplicity and c_{jk_j} constants. From Example 7.2 and the properties of the Laplace transform we therefore see that the impulse response h is the sum of products of polynomials and exponential functions. In particular, we see that the system is stable if the real part of the poles of $H(s)$ are negative

$$\Re\{p_j\} < 0\,.$$

If n is not smaller than m then $H(s)$ can be decomposed into the sum of a polynomial and a proper rational function. The impulse response h is then the sum of the above polynomial-exponential functions and a weighted sum of Dirac impulses and its derivatives.

8.6 System State

In this section we review the concept of the state of a system. To this end consider the initial value problem described by the system of n differential equations

$$\frac{d}{dt}u = Au + x, \qquad u(0) = u_0 \in \mathbb{C}^n$$

with $A \in \mathbb{C}^{n\times n}$ an $n \times n$ matrix and u and x n dimensional vectors of complex valued functions of time. As before we can translate this initial value problem in the language of distributions by replacing the (conventional) derivative with the distributional one and work in the convolution algebra of right sided distributions

$$Du = Au + u_o\delta + x\,.$$

If we rearrange the equation and convolve each term with $I\mathsf{1}_+$ we obtain the equivalent equation

$$(I\delta - A\mathsf{1}_+) * u = I\mathsf{1}_+ * (u_0\delta + x)\,. \tag{8.8}$$

This form shows that the equation can be solved by left convolving both sides of the equation with the inverse of $(I\delta - A\mathsf{1}_+)$. Observing the analogy with the geometric series, provided it converges, the latter can be represented by the following series, where the standard product of the geometric series has been replaced by the convolution product

$$(I\delta - A1_+)^{*-1} = I\delta + A1_+ + (A1_+)^{*2} + \cdots .$$

The iterated convolutions are easily evaluated

$$(A1_+)^{*n} = A^n 1_+^{*n} = A^n \frac{t^{n-1}}{(n-1)!}$$

and using the identity

$$1_+^{*n} = 1_+^{*n} * \delta = 1_+^{*n} * 1_+ * D\delta = 1_+^{*n+1} * D\delta$$

we obtain

$$(I\delta - A1_+)^{*-1} = I\delta + \sum_{n=1}^{\infty} A^n \frac{t^{n-1}}{(n-1)!} = \sum_{n=0}^{\infty} A^n \frac{t^n}{n!} 1_+ * D\delta .$$

The last series can be expressed with the help of the *exponential matrix* defined by

$$\mathrm{e}^{At} := \sum_{n=0}^{\infty} A^n \frac{t^n}{n!} \tag{8.9}$$

which converges for every value of t

$$(I\delta - A1_+)^{*-1} = 1_+\mathrm{e}^{At} * D\delta . \tag{8.10}$$

Having established the convergence of the series, using the linearity and continuity of convolution one readily sees that indeed it defines the desired inverse

$$(I\delta - A1_+) * [I\delta + A1_+ + (A1_+)^{*2} + \cdots] = I\delta .$$

The solution of the equation is therefore given by

$$u = 1_+\mathrm{e}^{At} * I(D\delta * 1_+) * (u_0\delta + x) = 1_+\mathrm{e}^{At}u_0 + 1_+\mathrm{e}^{At} * x . \tag{8.11}$$

The exponential matrix has several useful properties that are immediately verified using its defining series

$$\begin{aligned} \mathrm{e}^{At}\mathrm{e}^{A\tau} &= \mathrm{e}^{A(t+\tau)} & \qquad \mathrm{e}^{A0} &= I \\ (\mathrm{e}^{At})^{-1} &= \mathrm{e}^{-At} & \qquad D\mathrm{e}^{At} &= A\mathrm{e}^{At} = \mathrm{e}^{At}A . \end{aligned}$$

Note however that in general

$$e^A e^B \neq e^{A+B}.$$

This is only valid if A and B commute, that is $AB = BA$.

Consider now the *state space representation* of a SISO LTI system

$$Du = Au + u_o\delta + Bx, \qquad A \in \mathbb{C}^{n\times n}, B \in \mathbb{C}^{n\times 1} \tag{8.12}$$

$$y = Cu + Dx \qquad C \in \mathbb{C}^{1\times n}, D \in \mathbb{C} \tag{8.13}$$

where now x represents the input signal of the system and y its output. The vector u is called the *state* of the system and (8.11) shows that it's value u_0 at a given point in time t_0 is the minimum amount of information required that together with the input signal at times $t \geq t_o$ allows determining the system behaviour at all future times $t > t_0$. In other words, the system state u_0 at time t_0 summarises the effect on the system of all past values of the input signal and of previous states.

8.6.1 Controllability

It's interesting to ask if it's possible to design the input signal in such a way that the system can be set in an arbitrary state u_0 in finite time. That is, can we design the input signal such that for $t > t_0$ the state vector equals $u(t) = e^{At}u_0$?

The problem is most easily analysed using impulsive inputs, starting from the zero state. From the above results we know that the system state dependence on the input signal x is given by

$$u = 1_+ e^{At} B * x\,.$$

Suppose that for an n dimensional system we use an input signal consisting of a weighted sum of a Dirac impulse and its derivatives up to order $n-1$

$$x = x_0\delta + x_1 D\delta + \cdots + x_{n-1}D^{n-1}\delta\,.$$

Since the system is linear, we can analyse the contribution of each term individually

$$\begin{aligned}
1_+ e^{At} B * x_0\delta &= 1_+ e^{At} Bx_0 \\
1_+ e^{At} B * x_1 D\delta &= D(1_+ e^{At} Bx_1) = 1_+ e^{At} ABx_1 + \delta Bx_1 \\
&\cdots \\
1_+ e^{At} B * x_{n-1} D^{n-1}\delta &= D^{n-1}(1_+ e^{At} Bx_{n-1}) = 1_+ e^{At} A^{n-1} Bx_{n-1} + \cdots
\end{aligned}$$

The terms replaced by dots on the last line are constituted by a weighted sum of a Dirac impulse and its derivative which are zero for $t > 0$. Putting all terms together we obtain for $t > 0$

$$1_+ \mathrm{e}^{At} * x = 1_+ \mathrm{e}^{At} \begin{bmatrix} B & AB & \dots & A^{n-1}B \end{bmatrix} \cdot \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_{n-1} \end{bmatrix}$$

From this we conclude that we can use a suitably designed input signal x to mimic the effect of an arbitrary initial state u_0 if and only if the matrix

$$C := \begin{bmatrix} B & AB & \dots & A^{n-1}B \end{bmatrix} \tag{8.14}$$

is invertible, in which case the weighting factors are

$$\begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_{n-1} \end{bmatrix} = C^{-1} u_0 \,.$$

The matrix C is called *controllability matrix.*

While the state of a system plays an important theoretical and conceptual role, in practice, when dealing with controllable systems we can always start from the zero state and drive the system in any desirable state. Things are completely different for non-controllable systems. As discussed in Sect. 8.6.3, these are systems possessing sub-systems that are not influenced by the input signal. In those systems the initial state may play an important role.

8.6.2 Observability

Another interesting question is whether it's possible to reconstruct the initial state of a system at time t_0 from the observation of its output at times $t > t_0$ assuming that A, B, C, D and the input signal x are known. From linearity and knowledge of the input signal we can assume x to be zero. (Alternatively we could compute the part of the output signal due to the input signal—the zero state response of the system—and subtract it from the observed output.) The question is then if we can calculate u_o from the observation of

$$y = C 1_+ \mathrm{e}^{At} u_0.$$

Suppose that the system is n dimensional. Then if we compute the first $n - 1$ derivatives of the output signal we obtain

$$Dy = C1_+ e^{At} A u_0 + C\delta u_0$$
$$\dots$$
$$D^{n-1} y = C1_+ e^{At} A^{n-1} u_0 + \cdots$$

where in the last equation we have represented by dots a weighted sum of a Dirac pulse and its derivatives as before. Thus, the observation of the output signal and of its first $n - 1$ derivatives at times $t > 0$ allows setting up the following system of equations

$$\lim_{t \to 0+} \begin{bmatrix} y(t) \\ Dy(t) \\ \vdots \\ D^{n-1} y(t) \end{bmatrix} = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix} \cdot u_0 \,.$$

This system of equations can only be solved for u_0 if the matrix

$$O := \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix} \tag{8.15}$$

is not singular. The matrix O is called the *observability matrix.*

8.6.3 Jordan Normal form

The simplest way to understand the structure of a system that is either not controllable, or not observable is by considering the system in Jordan normal form.

Consider a system in the state space representation

$$\begin{aligned} Du &= Au + Bx, \qquad && A \in \mathbb{C}^{n \times n},\, B \in \mathbb{C}^{n \times 1} \\ y &= Cu && C \in \mathbb{C}^{1 \times n} \,. \end{aligned}$$

In linear algebra is shown that, by choosing a suitable basis, every linear operator can be represented by a matrix of the following block form, called the *Jordan normal form*

$$A = \begin{bmatrix} J_1 & & & 0 \\ & J_2 & & \\ & & \ddots & \\ 0 & & & J_r \end{bmatrix}$$

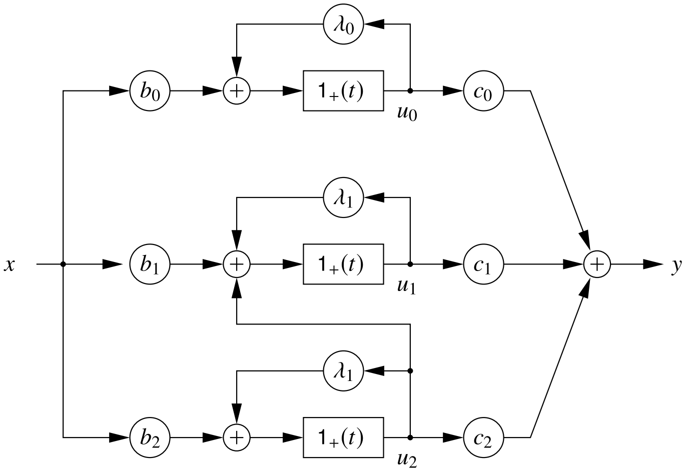


Fig. 8.2 Jordan normal form representation of a system

with

$$J_i = \begin{bmatrix} \lambda_i & 1 & & 0 \\ & \lambda_i & 1 & \\ & & \ddots & \ddots \\ 0 & & & \lambda_i \end{bmatrix}$$

the *elementary Jordan matrix*. The diagonal elements of J_i correspond all to the ith eigenvalue λ_i of A. If n_i correspond to the algebraic multiplicity of eigenvector λ_i and ν_i to its geometric multiplicity, then there are ν_i Jordan blocks J_i corresponding to eigenvalue λ_i. Thus, the total number of Jordan blocks corresponds to the number of independent eigenvectors of A. The Jordan normal form of a linear operator is unique up to permutations of the blocks.

A matrix for which the geometric multiplicity equals the algebraic multiplicity for each eigenvalue is called *semi simple*. In this case each block J_i is a 1×1 matrix and the Jordan normal form reduces to diagonal form.

A system in Jordan normal form can be interpreted as the parallel connection of independent sub-systems, each represented by a Jordan block J_i. Figure 8.2 shows the block diagram for a system with a simple eigenvalue λ_0 and a double eivenvalue λ_1 with $\nu_1 = 1$. From the figure it's easy to see that if $b_0 = 0$ then the state variable u_0 can't be excited by the input signal x. The same is true for u_2 if $b_2 = 0$. In either case the system is not controllable. One can check that these are the two conditions under which the determinant of the matrix C vanishes.

In a similar way the figure shows that if $c_0 = 0$ there is no path from u_0 to the output of the system and for $c_1 = 0$ there is no path from u_1. These are the two cases under which the system is not observable and correspond to the two conditions under which the determinant of the matrix O vanishes.

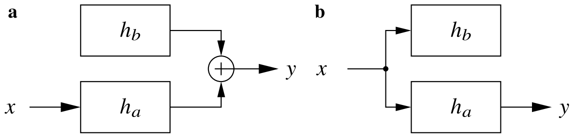


Fig. 8.3 **a** Not controllable system. **b** Not observable system

From these considerations we conclude that a non-observable system includes a sub-system whose output does not reach the global system output as schematically depicted in Fig. 8.3b. A non-controllable system includes a sub-system that is not reached by the input signal as schematically depicted in Fig. 8.3a.

Example 8.2: Jordan Block

Consider the system described by the following state-space representation

$$\begin{aligned} Du &= Au + Bx \\ y &= Cu \end{aligned}$$

with

$$A = \begin{bmatrix} \omega_{3dB} & 1 \\ 0 & \omega_{3dB} \end{bmatrix}, \qquad B = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}, \qquad C = \begin{bmatrix} c_0 \; c_1 \end{bmatrix}.$$

We want to compute an explicit expression for the exponential matrix e^{tA} allowing us to compute the response of the system to an arbitrary input signal x.

The matrix
$$A = \begin{bmatrix} \omega_{3dB} & 1 \\ 0 & \omega_{3dB} \end{bmatrix}$$

is an elementary Jordan matrix and can't be transformed in a diagonal matrix by a similarity transformation. In fact, as can be seen from the characteristic polynomial

$$\det(A - \lambda I) = (\omega_{3dB} - \lambda)^2,$$

the matrix has a single eigenvalue $\lambda = \omega_{3dB}$ with an algebraic multiplicity of 2 and the eigenspace belonging to this eigenvalue has dimension 1

$$\big(A - \omega_{3dB} I\big)v = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} v = 0 \qquad \Longrightarrow \qquad v = \alpha \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \alpha \in \mathbb{C}.$$

The matrix A can however be written as the sum of a diagonal matrix A_d and a particularly simple matrix A_c

$$A = A_d + A_c = \begin{bmatrix} \omega_{3dB} & 0 \\ 0 & \omega_{3dB} \end{bmatrix} + \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}.$$

Observe that the matrices A_d and A_c do commute. For this reason we can use the following property of the exponential matrix

$$\mathrm{e}^{t(A_d+A_c)} = \mathrm{e}^{tA_d}\mathrm{e}^{tA_c} \,.$$

Since A_d is diagonal, the first exponential matrix e^{tA_d} is easily calculated to be

$$\mathrm{e}^{tA_d} = \mathrm{e}^{\omega_{3dB}t} I \,.$$

The second exponential matrix e^{tA_c} is easily calculated from the series defining the exponential matrix by noting that the square of the matrix A_c vanishes

$$\mathrm{e}^{tA_c} = I + tA_c \,.$$

Putting these results together we obtain

$$\mathrm{e}^{tA} = \mathrm{e}^{\omega_{3dB}t}\left(\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} + \begin{bmatrix}0 & t\\ 0 & 0\end{bmatrix}\right) = \mathrm{e}^{\omega_{3dB}t}\begin{bmatrix}1 & t\\ 0 & 1\end{bmatrix} .$$

The above method can be used to calculate the exponential of any elementary Jordan matrix with the only modification that for an $n \times n$ matrix A it is the nth power of the matrix A_c that vanishes.

In the following we are always going to assume that the systems under consideration are controllable and observable.

Chapter 9
Weakly Nonlinear Time Invariant Systems

9.1 Introduction

As outlined in Chap. 1, the behavior of nonlinear systems is substantially richer than the one of linear systems. To deal with them there is a set of techniques, each one best suited to analyse particular aspects or particular classes of nonlinear systems. We target systems that are stable about an equilibrium point and that depend continuously on the input signal.

Before analysing in more details this class of systems, we give a short overview, mostly by way of examples, of systems described by nonlinear ordinary differential equations of the form

$$Dy = f(t, y), \qquad f : I \times X \to \mathbb{R}^n \tag{9.1a}$$

with $I \subset \mathbb{R}$, $X \subset \mathbb{R}^n$ and initial conditions

$$y(0) = y_0 \in X\,. \tag{9.1b}$$

We limit ourselves to the aspects that are helpful in better framing the concept of weakly nonlinear systems.

A first important difference compared to systems described by linear differential equations with constant coefficients is the fact that a solution may not exist for all $t > 0$ or may not be unique.

Example 9.1: IVP with many Solutions

Consider the following initial value problem (IVP)

$$Dy = \sqrt{|y|} \qquad y(0) = y_0\,.$$

F. Beffa, *Weakly Nonlinear Systems*, Understanding Complex Systems,
https://doi.org/10.1007/978-3-031-40681-2_9

If $y_0 > 0$ then the equation can be solved by the method of separation of the variables, and we obtain the unique solution

$$y(t) = \frac{1}{4}(t + 2\sqrt{y_o})^2 , \qquad t \geq 0 .$$

If $y_0 = 0$ then $y(t) = 0$ is a solution. However, it is not the only one. For any constant $c > 0$ the function

$$y_c(t) = \frac{1_+(t-c)}{4}(t-c)^2 , \qquad t \geq 0$$

is also a solution as one easily verifies by inserting it in the equation.

For $y_0 < 0$ we can again use the method of the separation of the variables to find the solution

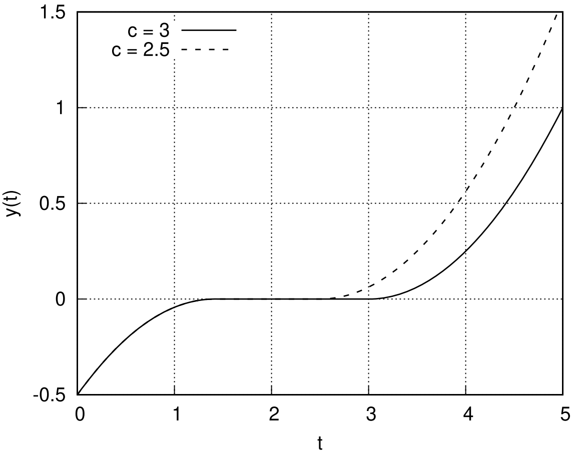
$$y(t) = -\frac{1}{4}(2\sqrt{|y_0|} - t)^2 .$$

However, due to the fact that at $y = 0$ the function $1/\sqrt{|y|}$ is not continuous (not even defined) this solution is only valid as long as $y(t) < 0$. When $y(t)$ reaches zero the equation can again be satisfied by multiple solutions

$$y_c(t) = \begin{cases} -\frac{1}{4}(2\sqrt{|y_0|} - t)^2 & t \in [0, 2\sqrt{|y_0|}) \\ 0 & t \in [2\sqrt{|y_0|}, c) \\ \frac{1}{4}(t-c)^2 & t \in [c, \infty) . \end{cases}$$

Therefore, for some initial conditions the equation has uncountably many solutions (Fig. 9.1).

Fig. 9.1 Two solutions of the initial value problem of Example 9.1 with $y_0 = -0.5$



From the above example we see that continuity of f is not enough to guarantee the existence of a unique solution of the initial value problem (9.1a). To guarantee uniqueness of a solution the function $f(t, y)$ must be more regular with respect to y.

Let $I \subset \mathbb{R}$ and $X \subset \mathbb{R}^n$. A function $f \in C(I \times X, \mathbb{R}^n)$ is called *locally Lipschitz continuous* in x if every point $(t_0, y_0) \in I \times X$ has a neighborhood $U \times V$ such that, for some constant $M > 0$

$$\|f(t, y) - f(t, x)\| \leq M\|y - x\|, \qquad t \in U, \quad x, y \in V.$$

If the function $f(t, y)$ in (9.1b) is continuous in t and locally Lipschitz continuous in y then Picard-Lindelöf's theorem guarantees the existence and uniqueness of the solution of the initial value problem (9.1a) [23].

If the function f doesn't depend explicitly on time, then the system is time invariant and the system equation becomes

$$Dy = f(y), \qquad f : X \to \mathbb{R}^n, \quad X \subset \mathbb{R}^n. \tag{9.2}$$

A solution of the equation for which $Dy = 0$ is called an *equilibrium point* of the system. When one investigates the stability of an equilibrium point y_e one can always assume it to be at the origin. In fact, by the change of variable $u = y - y_e$ one can always transform the system differential equation in one whose equilibrium point of interest is $u_e = 0$

$$Du = D(u + y_e) = f(u + y_e) =: g(u).$$

An equilibrium point is *stable* if for each $c > 0$ one can find an $\epsilon > 0$ such that

$$\|y(t_0)\| < \epsilon \quad \Longrightarrow \quad \|y(t)\| < c, \quad t \geq t_0.$$

It is *asymptotically stable* if it is stable and in addition ϵ can be chosen such that

$$\|y(t_0)\| < \epsilon \quad \Longrightarrow \quad \lim_{t\to\infty} \|y(t)\| = 0.$$

The set of all points $y(t_0)$ such that $\|y(t)\|$ converges to zero as t tends to infinity is called the *domain of attraction* of the equilibrium point. If an equilibrium point is not stable it is called *unstable*.

As already highlighted in Chap. 1, an important difference of time invariant nonlinear systems compared to LTI ones is the possibility of the existence of *multiple isolated equilibrium points.*

Example 9.2

Consider the system described by the following differential equation

$$Dy = -ay + cy^2$$

with a and c positive constants. From

$$0 = -ay + cy^2 = cy(y - a/c)$$

we see that the system has two equilibrium points:

$$y(t) = 0 \qquad \text{and} \qquad y(t) = a/c\,.$$

We are interested in the dynamic of the system starting from the initial condition $y(0) = y_0$ assuming that y_0 doesn't coincide with an equilibrium point. Since the function $f(y) = -ay + cy^2$ is locally Lipschitz continuous, there is a unique solution and this solution doesn't intersect the equilibrium points. The initial value problem can therefore be solved by separating the variables and integrating

$$\int_{y_0}^{y} \frac{dy}{cy(y - a/c)} = \int_{0}^{t} dt\,.$$

The solution is found to be

$$y(t) = y_0 \frac{\mathrm{e}^{-at}}{1 - y_0 \frac{c}{a}(1 - \mathrm{e}^{-at})}\,.$$

If y_0 is negative or $0 < y_0 c/a < 1$ the solution converges toward zero which therefore is an asymptotically stable equilibrium point (see Fig. 9.2). If $y_0 c/a > 1$ the solution diverges and reaches infinity in the finite time

$$t_\infty = \frac{1}{a} \ln\left(\frac{1}{1 - \frac{a}{y_0 c}}\right)\,.$$

From the above example we see that a nonlinear system can have multiple equilibrium points some of which can be stable, and some unstable. For a system to remain stable around a stable equilibrium point the initial condition may have to remain within a limited region around that point. Also, divergence from initial conditions near unstable equilibrium points can diverge faster than exponentially and reach infinity in finite time (finite escape time).

One of the most useful tools in the study of the stability of equilibrium points is the Lyapunov stability theory [24]. In particular Lyapunov's linearization (or indirect) method, states that

- If the linear approximation of the system about an equilibrium point is asymptotically stable then, in a neighborhood U of the equilibrium point, the (nonlinear)

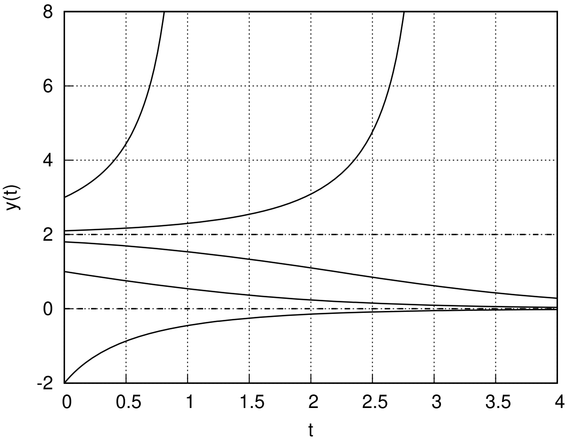


Fig. 9.2 Solutions for various initial conditions of the initial value problem of Example 9.2 with $a = 1$ and $c = 1/2$

system is asymptotically stable. The largest neighborhood U is the domain of attraction of the equilibrium point.

- If the linear approximation of the system about an equilibrium point is unstable, then the (nonlinear) system is unstable.

If the linear approximation of the system is neither asymptotically stable nor unstable then this method is inconclusive and one must turn to other methods, for example, Lyapunov's direct method [24].

Example 9.3

Consider the initial value problem described by the differential equation

$$Dy = cy^3$$

with c a constant; and the initial condition

$$y(0) = y_0 \, .$$

The only equilibrium point of the equation is the zero solution $y_e(t) = 0$. As it's immediately seen, the linearized equation is stable, but not asymptotically stable about the equilibrium point.

The nonlinear equation can be solved by the method of the separation of the variables

$$\int_{y_0}^{y} \frac{dy}{y^3} = c \int_{0}^{t} dt \, .$$

Performing the integrations and solving for y we find

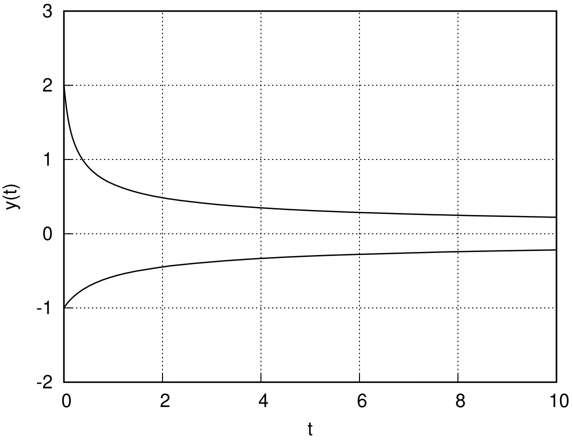


Fig. 9.3 Solutions for various initial conditions of the initial value problem of Example 9.3 with $c = -1$

$$y(t) = \frac{y_0}{\sqrt{1 - 2cy_0^2 t}} \,.$$

If $c > 0$ the solution diverges and reaches infinity at

$$t_\infty = \frac{1}{2cy_0^2} \,.$$

If $c < 0$ the equation is asymptotically stable for any value of the initial value y_0 (see Fig. 9.3).

Differently from what the above examples may suggest, most nonlinear differential equations can't be solved analytically. Therefore we are interested in methods to find approximate solutions around asymptotically stable equilibrium points in the spirit of a perturbation theory. Weakly nonlinear systems are a class of systems for which such a method exists and the solution is obtained in the form of a functional series.

Informally weakly nonlinear systems can be described as systems operated around an asymptotically stable equilibrium point and whose response depends continuously on the input signal x. They include systems described by a differential equation of the form

$$Dy = Cx + f(y) \,,$$
$$f : Y \to \mathbb{R}^n, \quad C : X \to \mathbb{R}^n \,, \quad X \subset \mathbb{R} \,, \quad Y \subset \mathbb{R}^n$$

with C a linear function and f a function that, within the excursion range of interest of y, can be approximated to any desired accuracy by a Taylor expansion. Note that polynomials are locally Lipschitz continuous. For this reason weakly nonlinear systems are well-behaved and produce a well-defined and unique output response.

9.2 Graded Algebra of Test Functions

In the previous section we illustrated some aspects of weakly nonlinear systems based on examples of systems described by nonlinear differential equations. We now look for a description based on distributions. We'll see that this allows reducing the problem of solving some classes of nonlinear differential equations to an essentially algebraic problem. However, before discussing systems, we need some preparation that we provide in this and the next section.

Let $V_k, k \in \mathbb{N}$ be vector spaces on $\mathbb{C}$ such that $V_k \cap V_j = \{0\}$ for $k \neq j$. The *direct sum*

$$V := \bigoplus_{k=0}^{\infty} V_k := \bigoplus_{k \geq 0} V_k \tag{9.3}$$

is the vector space whose elements are the sequences (x_k) in $\bigcup_{k=0}^{\infty} V_k$ with $x_k \in V_k$ and $x_k = 0$ for fast every k. That is, the set of all finite sequences with $x_k \in V_k$. The vector space structure of V is defined by the following addition and multiplication with scalars

$$(x_k) + c(y_k) := (x_k + cy_k), \qquad (x_k), (y_k) \in V\,, \quad c \in \mathbb{C}\,. \tag{9.4}$$

Each V_k is evidently a sub-vector space of V.

If furthermore V is provided with a multiplication

$$V \times V \to V, \qquad (x, y) \mapsto x \odot y$$

such that it forms an algebra and in addition

$$V_k \odot V_j \subset V_{k+j}\,, \qquad k, j \in \mathbb{N}$$

then it is called a *graded algebra*.

Let $V_k = \mathcal{D}(\mathbb{R}^k)$ be the vector space of test functions on $\mathbb{R}^k$ with $V_0 = \mathbb{C}$. Then

$$\mathcal{D}_\oplus := \bigoplus_{k \geq 0} \mathcal{D}(\mathbb{R}^k)$$

with the tensor product as multiplication

$$\phi \otimes \psi(\tau_1, \ldots, \tau_k, \tau_{k+1}, \ldots, \tau_{k+j}) := \phi(\tau_1, \ldots, \tau_k)\psi(\tau_{k+1}, \ldots, \tau_{k+j})$$

is a graded algebra that we call the *graded algebra of test functions*. We write elements of $\mathcal{D}_\oplus$ as sums with indices denoting the grade of the element

$$\phi = \sum_{j=0}^{N} \phi_j\,, \qquad \phi_j \in \mathcal{D}(\mathbb{R}^j)\,, \quad N \in \mathbb{N}\,.$$

In the graded algebra of test functions we define the following convergence criterion. A sequence (ϕ_m), $\phi_m \in \mathcal{D}_\oplus$ with

$$\phi_m = \sum_{j=0}^{N_m} \phi_{j,m}\,, \qquad \phi_{j,m} \in \mathcal{D}(\mathbb{R}^j)$$

converges to zero if

1. There exist compact sets $K_j \subset \mathbb{R}^j$, $j = 1, \dots, N$ with $N = \max_{m\in\mathbb{N}}(N_m)$ such that for each j and m
$$\mathrm{supp}(\phi_{j,m}) \subset K_j\,.$$
2. For every $j > 0$ and every j-tuple $k \in \mathbb{N}^j$ the sequence $(D^k\phi_{j,m})_{m\in\mathbb{N}}$ converges uniformly to zero. For $j = 0$ the sequence of numbers $(\phi_{0,m})_{m\in\mathbb{N}}$ converges to zero.

9.3 Direct Product of Distributions

The *direct product* V of vector spaces V_k on $\mathbb{C}$ is the vector space whose elements are the sequences (x_k) with $x_k \in V_k$, $k \in \mathbb{N}$. The vector space structure is defined as for the direct sum by (9.4). It is denoted by

$$V := \prod_{k\geq 0} V_k := \prod_{k=0}^{\infty} V_k\,. \tag{9.5}$$

The key difference from the direct sum is that, in a direct product, the sequence does not have to be finite.

Let $V_k = \mathcal{D}'(\mathbb{R}^k)$, with $V_0 = \mathbb{C}$. Then the direct product

$$\mathcal{D}'_\oplus := \prod_{k\geq 0} \mathcal{D}'(\mathbb{R}^k)$$

is the set of linear continuous functionals on $\mathcal{D}_\oplus$ defined by

$$h : \mathcal{D}_\oplus \to \mathbb{C}, \qquad \phi \mapsto \langle h, \phi\rangle := \sum_{j=0}^{\infty} \langle h_j, \phi_j\rangle \tag{9.6}$$

with

$$\phi = \sum_{j=0}^{\infty} \phi_j\,, \qquad h = \sum_{j=0}^{\infty} h_j\,, \qquad \phi_j \in \mathcal{D}(\mathbb{R}^j), \quad h_j \in \mathcal{D}'(\mathbb{R}^j)\,.$$

Since ϕ only has a finite number of terms different from zero, $\langle h, \phi\rangle$ is well-defined. As for $k \neq j$, $\mathcal{D}'(\mathbb{R}^k) \cap \mathcal{D}'(\mathbb{R}^j) = \{0\}$, here and in the following we denote elements of $\mathcal{D}'_{\oplus}$ by sums in a similar way as we do for elements of $\mathcal{D}_{\oplus}$.

Continuity in $\mathcal{D}'_{\oplus}$ is defined by the convergence that we defined for $\mathcal{D}_{\oplus}$ and follows from the continuity of distributions. Since $\mathcal{D}'_{\oplus}$ is a vector space, it's enough to verify continuity at the origin. Let $h \in \mathcal{D}'_{\oplus}$ and $\phi \in \mathcal{D}_{\oplus}$, then there exists an $N \in \mathbb{N}$ such that

$$|\langle h, \phi\rangle| \leq \sum_{j=0}^{N} |\langle h_j, \phi_j\rangle| \leq (N+1) \sup_{j\in\{0,\dots,N\}} |\langle h_j, \phi_j\rangle|$$

and according to our definition of convergence, when ϕ converges to zero, so does $\sup_j |\langle h_j, \phi_j\rangle|$ and hence $\langle h, \phi\rangle$.

In Sect. 3.1 we have introduced the tensor product of distributions and have seen that it is well defined between any pair of distributions. With it we can define a product $g \cdot h$ between elements g and h of $\mathcal{D}'_{\oplus}$. It's kth component is defined by

$$(gh)_k := (g \cdot h)_k := \sum_{j=0}^{k} g_j \otimes h_{k-j}\,, \qquad k \in \mathbb{N} \tag{9.7}$$

with g_j and h_j the jth components of g and h respectively. With this product $(\mathcal{D}'_{\oplus}, +, \cdot)$ becomes an algebra. As is common practice, we will often denote $g \cdot h$ simply by gh. Being based on an associative operation (the tensor product) the product that we just defined is associative.

Note the close similarity between the algebra of formal power series and the one that we have defined for $\mathcal{D}'_{\oplus}$. In both cases addition is defined component wise and the product has the form of a convolution.

9.4 Symmetric Distributions

Let S_k denote the set of all permutations of $\{1, \dots, k\}$. A distribution $h_k \in \mathcal{D}'(\mathbb{R}^k)$ is *symmetric* if

$$\langle h_k, \phi(\tau_{\sigma(1)}, \dots, \tau_{\sigma(k)})\rangle = \langle h_k, \phi(\tau_1, \dots, \tau_k)\rangle \tag{9.8}$$

for all permutations $\sigma \in \mathsf{S}_\mathsf{k}$ and every $\phi \in \mathcal{D}(\mathbb{R}^k)$. Symmetric distributions are fully characterized by symmetric test functions for

$$\langle h_k, \phi(\tau_1, \dots, \tau_k)\rangle = \left\langle h_k, \frac{1}{k!} \sum_{\sigma \in \mathsf{S_k}} \phi(\tau_{\sigma(1)}, \dots, \tau_{\sigma(k)})\right\rangle$$

and the sum of test functions on the right-hand side is a symmetric test function. The sum of symmetric distributions is a symmetric distribution. Therefore, they form a vector subspace of distributions that we denote by $\mathcal{D}'_{\text{sym}}(\mathbb{R}^k)$. Similarly, we denote the vector subspace of all symmetric test functions on $\mathbb{R}^k$ by $\mathcal{D}_{\text{sym}}(\mathbb{R}^k)$, the one of the direct sum of symmetric test functions by $\mathcal{D}_{\oplus,\text{sym}}(\mathbb{R}^k)$ and the one of the direct product of symmetric distributions by $\mathcal{D}'_{\oplus,\text{sym}}(\mathbb{R}^k)$.

A symmetric distribution can be constructed from an arbitrary distribution $f \in \mathcal{D}'(\mathbb{R}^k)$ by averaging over all permutations of the independent variables

$$[f]_{\text{sym}} := \frac{1}{k!} \sum_{\sigma \in \mathsf{S_k}} f(\tau_{\sigma(1)}, \dots, \tau_{\sigma(k)}) \tag{9.9}$$

with

$$\langle f(\tau_{\sigma(1)}, \dots, \tau_{\sigma(k)}), \phi(\tau_1, \dots, \tau_k)\rangle := \langle f(\tau_1, \dots, \tau_k), \phi(\tau_{\sigma(1)}, \dots, \tau_{\sigma(k)})\rangle\,.$$

Such an operation is called *symmetrisation*.

The tensor product is a bi-linear operation. Therefore, the power of an element of $\mathcal{D}'_\oplus$ composed by a finite number of distributions $f_j \in \mathcal{D}'(\mathbb{R}^{n_j})$, $n_j \geq 1$, $j = 1, \dots, m, m \geq 2$ can be expressed as a sum of tensor products

$$\left(\sum_{j=1}^{m} f_j\right)^k = \sum_{j_1=1}^{m} \cdots \sum_{j_k=1}^{m} f_{j_1} \otimes \cdots \otimes f_{j_k}, \quad k \in \mathbb{N}$$

with the sum ranging over all possible combinations of the indexes $j_1, \dots, j_k$. If the distributions $f_1, \dots, f_m$ are symmetric then one can reorder the indexes $j_1, \dots, j_k$ by any permutation σ without changing the value of the sum. Hence, the tensor products on the right-hand side can be replaced by symmetrized products

$$\sum_{j_1=1}^{m} \cdots \sum_{j_k=1}^{m} f_{j_1} \otimes \cdots \otimes f_{j_k} = \sum_{j_1=1}^{m} \cdots \sum_{j_k=1}^{m} \left[f_{j_1} \otimes \cdots \otimes f_{j_k}\right]_{\text{sym}}\,.$$

The tensor product of symmetric distributions inside the symmetrisation operator act as a commutative operator. For this reason the sum includes summands that are equal and, by grouping them, we obtain an expression that is similar to the multinomial formula [21]

$$\left(\sum_{j=1}^{m} f_j\right)^k = \sum_{|\alpha|=k} \frac{k!}{\alpha!}\left[f^{\otimes\alpha}\right]_{\text{sym}}\,, \qquad f=(f_1,\dots,f_m) \tag{9.10}$$

with α an m-tuple in $\mathbb{N}^m$,

$$f^{\otimes\alpha} := f_1^{\otimes\alpha_1}\otimes\cdots\otimes f_m^{\otimes\alpha_m} \tag{9.11}$$

and where we made use of the multi-index notation introduced in Sect. 4.6.

In general the product that we defined on $\mathcal{D}'_{\oplus}$ applied to two elements of $\mathcal{D}'_{\oplus,\text{sym}}$ does not result in an element of $\mathcal{D}'_{\oplus,\text{sym}}$. This can be remedied by symmetrizing the product

$$(gh)_k := (g\cdot h)_k := \sum_{j=0}^{k}\left[g_j\otimes h_{k-j}\right]_{\text{sym}}\,, \qquad g,h\in\mathcal{D}'_{\oplus,\text{sym}}\,. \tag{9.12}$$

Unless explicitly stated otherwise, when working in $\mathcal{D}'_{\oplus,\text{sym}}$ we will always assume the use of this symmetrized product.

The last property of symmetric distributions that we want to mention is the fact that, in a convolution algebra, the inverse of a symmetric distribution is symmetric, for

$$\begin{aligned}\delta(\tau_1,\tau_2) &= f(\tau_1,\tau_2) * f^{*-1}(\tau_1,\tau_2)\\ &= f(\tau_2,\tau_1) * f^{*-1}(\tau_1,\tau_2)\\ &= f(\tau_1,\tau_2) * f^{*-1}(\tau_2,\tau_1)\,.\end{aligned}$$

9.5 Weakly Nonlinear Systems

We are looking for a representation, in the spirit of a perturbation theory, of a class of nonlinear systems including the ones described by differential equations of the form

$$Ly = x + \sum_{k=2}^{K} c_k y^k \tag{9.13}$$

with $x\in\mathcal{D}'(\mathbb{R})$ a given input signal, L a linear differential operator with constant coefficients

$$L = D^m + a_{m-1}D^{m-1} + \cdots + a_1 D + a_0$$

and where we assume that the linearized system is stable.

In Chap. 7 we saw that, in the language of distributions, a linear differential equation with constant coefficients becomes a convolution equation. If we want to apply the results obtained for convolution equations, we need to give a meaning to the nonlinear terms appearing in the above equation.

In general, it's not possible to define a multiplication valid for arbitrary distributions. Therefore, the terms y^k, $k > 1$ can't be assumed to belong to $\mathcal{D}'(\mathbb{R})$. To work around this problem we can assume y to belong to a direct product of distributions, $y = (y_0, y_1, y_2, \dots)$, and use the product defined on that space. Since the product between functions with values in $\mathbb{C}$ is commutative $f \cdot g = g \cdot f$, we require y to belong to the direct product of symmetric distributions $\mathcal{D}'_{\oplus,\mathrm{sym}}$. Then, if y_1 is the solution of the linearized equation its powers become tensor powers

$$(y_1)^k = y_1^{\otimes k}\,.$$

If y_1 is a regular distribution, that is a locally integrable *function*, then we can recover the meaning of the powers in the differential equation by evaluating $y_1^{\otimes k}$ on the diagonal

$$y_1^{\otimes k}(t, \dots, t) = y_1^k(t)\,.$$

The same remains true if we replace y_1 by a sum of distributions.

To complete the interpretation of the differential equation in the language of distributions it remains to be clarified what is the effect of the one dimensional differential operator D appearing in (9.13) on the components $y_k \in \mathcal{D}'_{\mathrm{sym}}(\mathbb{R}^k)$ of y. To this end, suppose y_k to be a regular distribution. Then it is a locally integrable function

$$y_k : \tau \mapsto y_k(\tau_1, \dots, \tau_k)\,, \qquad \tau \in \mathbb{R}^k$$

and we can associate with it a function of the single variable t by defining an operation that we call "evaluating on the diagonal"

$$\mathrm{ev_d}(y_k) := t \mapsto y_k(t, \dots, t)\,, \qquad t \in \mathbb{R}\,.$$

If we assume this function to be differentiable, then the derivative with respect to t is well-defined

$$D\mathrm{ev_d}(y_k)(t) = D_1 y_k(t, \dots, t) + \cdots + D_k y_k(t, \dots, t)$$

and, as a distribution, can be represented by

$$Dy_k := \left(\sum_{j=0}^{k-1} \delta^{\otimes j} \otimes D\delta \otimes \delta^{\otimes k-1-j}\right) * y_k\,. \tag{9.14}$$

This last expression is symmetric and is valid for arbitrary distributions. Therefore, we can take it as the definition of the effect of the differential operator D on distributions $y_k \in \mathcal{D}'_{\mathrm{sym}}(\mathbb{R}^k)$. For $y \in \mathcal{D}'_{\oplus,\mathrm{sym}}$ and any $\phi \in \mathcal{D}_{\otimes}$, $\langle y, \phi\rangle$ only has a finite number of terms different from zero. For this reason the effect of D on y can be defined as acting on each component individually.

For $y \in \mathcal{D}'_{\oplus,\text{sym}}$ to be a solution of (9.13) in a convolution algebra, the equation must be satisfied by each component y_k of y individually. If y has to be compatible with our assumption of the system being described around the zero equilibrium point, then the 0th component y_0 must always be zero

$$y_0 = 0\,.$$

In analogy with the theory of formal power series we call distributions $y \in \mathcal{D}'_{\oplus}$ with $y_0 = 0$ *nonunits* [25].

For $k = 1$ the only terms belonging to $\mathcal{D}'(\mathbb{R})$ appearing in the equation are y_1 and x. Hence, y_1 is the solution of the linearized equation and, as discussed in Sect. 8.1, can be represented by

$$y_1 = h_1 * x\,.$$

For $k = 2$ we have

$$L\delta * y_2 = c_2\, y_1^{\otimes 2} \qquad \delta \in \mathcal{D}'(\mathbb{R}^2)$$

and we see that, for the computation of y_2, the tensor power of y_1 plays the role of an input signal applied to a linear system. Assuming that $L\delta$ has an inverse, we obtain

$$y_2 = c_2\,(L\delta)^{*-1} * y_1^{\otimes 2}\,.$$

The above expression can be further manipulated by noting that

$$\begin{aligned}
&\langle (a(\tau_1) \otimes b(\tau_2)) * (f(\tau_1) \otimes g(\tau_2))\,, \phi(\tau_1, \tau_2)\rangle \\
&\quad = \langle (a(\tau_1) \otimes b(\tau_2)) \otimes (f(\lambda_1) \otimes g(\lambda_2))\,, \phi(\tau_1 + \lambda_1, \tau_2 + \lambda_2)\rangle \\
&\quad = \langle (a(\tau_1) * f(\tau_1)) \otimes (b(\tau_2) * g(\tau_2))\,, \phi(\tau_1, \tau_2)\rangle
\end{aligned}$$

or

$$(a \otimes b) * (f \otimes g) = (a * f) \otimes (b * g)\,. \tag{9.15}$$

With this expression and the solution found for y_1 we can express y_2 as

$$y_2 = h_2 * x^{\otimes 2}\,, \qquad h_2 := c_2\,(L\delta)^{*-1} * h_1^{\otimes 2}$$

where raising to a tensor power is assumed to have higher priority than convolution.

From this it is not difficult to see that every component y_k can be expressed as the convolution of a distribution h_k specific to the problem and the input signal x raised to the tensor power of k

$$y_k = h_k * x^{\otimes k}\,.$$

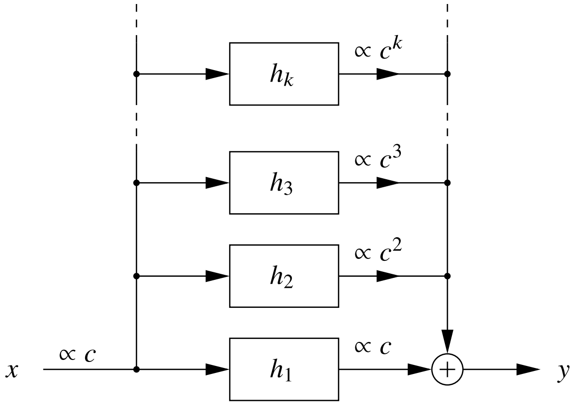


Fig. 9.4 Block diagram representation of a time-invariant weakly nonlinear system. If the input signal is proportional to the constant $c \in \mathbb{C}$, the output of the block characterized by the kth order impulse response h_k is proportional to c^k

We are therefore led to define a *weakly nonlinear* (or *analytic*) time-invariant (WNTI) system as a system $\mathcal{H}$ whose behavior around the zero equilibrium point can be described by an element h of $\mathcal{D}'_{\oplus,\text{sym}}$ such that, when driven by the input signal x, its output is given by

$$y = h[x] := \sum_{k=1}^{\infty} h_k * x^{\otimes k}\,, \qquad h_k \in \mathcal{A}'_k\,, \quad x \in \mathcal{A}'_1 \tag{9.16}$$

with $\mathcal{A}'_1$ a convolution algebra in $\mathcal{D}'(\mathbb{R})$ and $\mathcal{A}'_k$ a convolution algebra in $\mathcal{D}'_{\text{sym}}(\mathbb{R}^k)$ compatible with $\mathcal{A}_1$ and the tensor product. This means that, if $x \in \mathcal{A}'_1$, then $x^{\otimes k}$ must be an element of $\mathcal{A}'_k$. We denote such a set of convolution algebras by $\mathcal{A}'_{\oplus,\text{sym}}$. The distribution h_k is called the *kth order impulse response* (or *kernel*) of the system. A block diagram representation of a weakly nonlinear system is shown in Fig. 9.4. Note that, if the input signal is multiplied by a constant $c \in \mathbb{C}$, y_k is scaled by a factor of c^k

$$y_k = h_k * (c\,x)^{\otimes k} = c^k \left(h_k * x^{\otimes k}\right)\,.$$

The interpretation of the output of our definition of a weakly nonlinear system requires some comment as it doesn't always represent a quantity that can be interpreted as a signal depending on time. Under the assumption that all involved distributions belong to a convolution algebra, then one can distinguish the following cases

- If the impulse responses h_k as well as the input signal x are regular distributions and the convolutions $h_k * x^{\otimes k}$ are well-defined (see Sect. 3.2) then all output components y_k are locally integrable functions. In this case we can evaluate the y_k on the diagonal

$$\mathrm{ev_d}(y_k)(t) = \mathrm{ev_d}(h_k * x^{\otimes k})(t) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h_k(\tau_1, \ldots, \tau_k) x(t - \tau_1) \cdot \cdots \cdot x(t - \tau_k) d\tau_1 \cdots d\tau_k \tag{9.17}$$

and obtain an interpretation for the y_k as signals of time.

If the input signal is scaled by the constant c, then, at each time t, the output $\mathrm{ev_d}(y)(t)$ is seen to be a power series in c

$$\mathrm{ev_d}(y)(t) := \sum_{k=1}^{\infty} c^k \mathrm{ev_d}(h_k * x^{\otimes k})(t)\,.$$

If this series has a convergence radius greater than zero valid at all times, then $\mathrm{ev_d}(y)$ represents a well-defined function of time and we have a clear procedure to interpret the output of the system.

- If some or all of the impulse responses h_k are not regular, there is still a class of input signals for which all y_k are regular distributions. (Remember that the convolution of any distribution with a test function is an indefinitely differentiable function.) The system restricted to this class of input signals may still be evaluated on the diagonal to obtain a function of time $\mathrm{ev_d}(y)$ as in the previous case.
- If for no input signal (different from zero) there is a constant $c > 0$ such that $\mathrm{ev_d}(y)(t)$ remains finite at all times then the system can't be represented using an element of $\mathcal{D}'_{\oplus,\mathrm{sym}}$.

Example 9.4: Polynomial System

In this example we consider a class of systems whose impulse responses are not regular.

Suppose that the output of a system $\mathcal{H}$ is represented by a nonlinear function f of the input signal x and that the function f can be adequately approximated by a Taylor polynomial around the origin

$$y = f(x) \approx \sum_{k=1}^{K} \frac{f^{(k)}(0)}{k!} x^k\,, \qquad f(0) = 0\,, \quad K > 1\,. \tag{9.18}$$

It is readily seen that such a system can be represented by the impulse responses

$$h_k = \frac{f^{(k)}(0)}{k!} \delta^{\otimes k}\,, \qquad k = 1, \ldots, K\,.$$

The response of the system to the input signal x as represented by these impulse responses is

$$y = h[x] = \sum_{k=1}^{K} \frac{f^{(k)}(0)}{k!} \delta^{\otimes k} * x^{\otimes k} \, .$$

If the input signal is not a regular distribution, for example if it is a Dirac pulse, then neither the initial representation given by (9.18), nor the evaluation on the diagonal $\mathrm{ev_d}(h[\delta])$ do have a meaning. In spite of this, the impulse responses and their outputs y_k are mathematically well-defined.

If the class of input signals is restricted to regular distributions then the output obtained from the representation in terms of impulse responses by evaluating on the diagonal $\mathrm{ev_d}(h[x])$ agrees with the original one.

If f is analytic, then it can be represented by a power series ($K \to \infty$). In this case the output of the system is only well defined if the magnitude of the input signal $|x(t)|$ remains smaller than the convergence radius of the series at all times.

Let h_k be the kth order impulse responses of the weakly nonlinear system $\mathcal{H}$ and x its input signal. In Sect. 3.3 we saw that an arbitrary distribution can be approximated to any desired accuracy by a finite sum of Dirac pulses. Hence, x can be approximated by

$$x \approx \sum_{j=1}^{M} a_j \, \delta(t - \lambda_j) \, , \qquad a_j \in \mathbb{C}, \quad \lambda_j \in \mathbb{R}$$

and the output of h_k by

$$\begin{aligned} y_k &\approx \sum_{j_1=1}^{M} \cdots \sum_{j_k=1}^{M} h_k * a_{j_1} \delta(\tau_1 - \lambda_{j_1}) \otimes \cdots \otimes a_{j_k} \delta(\tau_k - \lambda_{j_k}) \\ &= \sum_{j_1=1}^{M} \cdots \sum_{j_k=1}^{M} a_{j_1} \cdot \cdots \cdot a_{j_k} h_k(\tau_1 - \lambda_{j_1}, \ldots, \tau_k - \lambda j_k) \, . \end{aligned}$$

This expression suggests the interpretation for h_k as that portion of the system defining how the response of the system depends on the combination of k simultaneous points in time of the input signal.

In addition, if we compare the expression representing the output at time t of the (causal) impulse response h_k

$$\mathrm{ev_d}(h_k * x^{\otimes k})(t)$$

with the one of a polynomial system (see Example 9.4)

$$\mathrm{ev_d}(c_k \delta^{\otimes k} * x^{\otimes k})(t) = c_k \, x^k(t)$$

we see that, the output at time t of the latter only depends on the kth power of the current value of the input signal. In contrast to this, the output at time t of the former depends on all combinations of products of k (past) values of the input signal. The impulse responses h_k can thus be interpreted as the memory of the system. The given representation of weakly nonlinear systems can be seen as a generalization of the Taylor approximation method for memory-less systems to systems with memory. It is called the Volterra functional series in honor of V. Volterra who first proposed it [5].

9.6 Nonlinear Transfer Functions

All impulse responses h_k of a causal weakly nonlinear system must vanish if any argument τ_j is less than zero. This is most easily seen if we consider the case where the impulse responses as well as the input signal x are regular distributions, for then

$$\mathrm{ev}_{\mathrm{d}}(y_k)(t) = \int\limits_{-\infty}^{\infty} \cdots \int\limits_{-\infty}^{\infty} h_k(t-\tau_1,\ldots,t-\tau_k)x(\tau_1)\cdot\cdots\cdot x(\tau_k)d\tau_1\cdots d\tau_k\,.$$

As every distribution is the limit of smooth functions, this must then be true for arbitrary distributions. The impulse responses of all orders of causal systems are therefore right-sided distributions.

The Laplace transform of the k-order impulse response h_k is called the *nonlinear transfer function of order k*

$$H_k(s_1,\ldots,s_k) = \langle h_k(\tau_1,\ldots,\tau_k), \mathrm{e}^{-s_1\tau_1-\cdots-s_k\tau_k}\rangle\,. \tag{9.19}$$

Due to the symmetry of h_k, it is a symmetric function of the variables s_1 to s_k

$$H_k(s_1,\ldots,s_k) = H_k(s_{\sigma(1)},\ldots,s_{\sigma(k)})\,, \qquad \sigma \in \mathsf{S}_\mathsf{k}\,. \tag{9.20}$$

As the Laplace transform converts convolution products into ordinary products, the Laplace transform of $y_k = h_k * x^{\otimes k}$ is

$$Y_k(s1,\ldots,s_k) = H_k(s_1,\ldots,s_k)X(s_1)\cdot\cdots\cdot X(s_k)\,.$$

Just as with LTI systems, the many useful properties of the Laplace transform makes it a very valuable tool for solving convolution equations describing weakly nonlinear systems. In particular, on top of converting convolution products into ordinary multiplications, in their region of convergence, the Laplace transformed of distributions are holomorphic functions.

Consider a system described by a differential equation with constant coefficients of the type considered before

$$\begin{aligned} Ly &= Nx + \textstyle\sum_{j=2}^{J} c_j y^j \\ L &= D^n + a_{n-1} D^{n-1} + \cdots + a_0 \\ N &= b_m D^m + b_{m-1} D^{m-1} + \cdots + b_0 . \end{aligned} \tag{9.21}$$

The part of the corresponding convolution equation relevant for the calculation of $y_k, k > 1$, is

$$(L\delta^{\otimes k}) * y_k = \sum_{j=2}^{k} c_j (y_1 + \cdots + y_{k-1})^j \,.$$

As the Laplace transform of $D\delta^{\otimes k}$ is

$$\mathcal{L}\{D\delta^{\otimes k}\}(s_1, \ldots, s_k) = s_1 + \cdots + s_k$$

(see (9.14)), the Laplace transform of $L\delta^{\otimes k}$ is a polynomial in $s_1 + \cdots + s_k$

$$P(s_1 + \cdots + s_k) = (s_1 + \cdots + s_k)^n + a_{n-1}(s_1 + \cdots + s_k)^{n-1} + \cdots + a_0 \,.$$

Note that the coefficients of this polynomial are the same for all k, including $k = 1$. The only difference between the various values of k is in the argument. If we factor it, we see that the denominator of H_k adds to the denominator of the lower order transfer functions H_j, $j = 1, \ldots, k-1$ terms of the form

$$(s_1 + \cdots + s_k - p_j)^{l_j}$$

with p_j the jth pole and l_j its multiplicity. If we assume H_k to be a proper rational function, then its partial fraction expansion will include terms of the form

$$\frac{F(s_1, \ldots, s_{k-1})}{(s_1 + \cdots + s_k - p_j)^{l_j}}$$

and similar ones where some of the variables $s_1, \ldots, s_{k-1}$ may be missing. If by the calculation of the inverse Laplace transform we start by inverse transforming with respect to s_k we obtain the expression

$$F(s_1, \ldots, s_{k-1})\, \tau_k^{l_j - 1}\, \mathrm{e}^{[p_j - (s_1 + \cdots + s_{k-1})]\tau_k}\, 1_+(\tau_k)\,.$$

By using the shifting property of the Laplace transform and denoting by f the inverse transform of F, the complete inverse transform of the above expression is

$$f(\tau_1 - \tau_k, \ldots, \tau_{k-1} - \tau_k) \otimes \tau_k^{l_j - 1}\, \mathrm{e}^{p_j \tau_k}\, 1_+(\tau_k)\,.$$

If H_k is not a proper rational function, then it can be decomposed into a polynomial and a proper rational function. The inverse Laplace transform of the polynomial part results in Dirac pulses and its derivatives.

This shows that, if the system under consideration can be described by a differential equation with constant coefficients of the indicated type, then, similarly to the first order impulse response h_1, the higher order impulse responses are sums of Dirac pulses, their derivatives and products of polynomials and exponential functions in the variables $\tau_1, \dots, \tau_k$. In addition, it also shows that, if the linear transfer function $H_1(s_1)$ has all its poles in the left-hand side of the complex plane, then not only does the regular part of h_1 (that is discarding the Dirac pulses and its derivatives) decay exponentially as its argument tends to infinity, but so also do the regular part of all higher order impulse responses h_k. In particular, we see that all impulse responses are summable distributions

$$h_k \in \mathcal{D}'_{L^1+}(\mathbb{R}^k)\,, \qquad k = 1, 2, \dots .$$

In the following, unless explicitly stated otherwise, we are always going to assume the systems to be of this type.

Example 9.5

We revisit Example 9.2 and find an approximate solution of the initial value problem

$$Dy = -ay + cy^2\,, \qquad y(0) = y_0\,, \qquad a, c > 0\,,$$

valid around its zero equilibrium point.

As we saw, in translating an initial value problem into the language of distributions, the initial conditions become part of the equation which, in this case, comes to be

$$(D + a)y = y_0\delta + cy^2\,.$$

We can think of this equation as an equation describing a system driven by the input signal $x = y_0\delta$. The solution of the equation y is an element of $\mathcal{D}'_{\oplus,\mathrm{sym}}$ and has the form

$$y = \sum_{k=1}^{\infty} h_k * x^{\otimes k}\,.$$

The system is therefore fully characterized if we find the impulse responses h_k. The solution of the original problem is then found by multiplying each impulse response h_k by y_0^k

$$y_k = h_k y_0^k\,.$$

To find the impulse responses we apply the input signal $x = \delta$ and insert $y = h$ into the equation. The equation is solved if it is satisfied by each component h_k of h individually. The component h_k can be computed from the equation and the impulse responses of lower order h_j, $j = 1, \dots, k - 1$.

To find h_1 we retain only terms of the equation belonging to $\mathcal{D}'(\mathbb{R})$

$$(D\delta + a\delta) * h_1 = \delta\,.$$

If we Laplace transform the equation we obtain

$$(s_1 + a)H_1(s_1) = 1$$

from which we immediately obtain the first order transfer function

$$H_1(s_1) = \frac{1}{s_1 + a}$$

and, by inverse Laplace transformation, the first order impulse response

$$h_1(\tau_1) = 1_+(\tau_1)\,\mathrm{e}^{-a\tau_1}\,.$$

The second order impulse response h_2 is found by retaining in the equation only terms belonging to $\mathcal{D}'(\mathbb{R}^2)$

$$(D + a)\delta^{\otimes 2} * h_2 = c\,h_1^{\otimes 2}\,.$$

From the Laplace transformed equation

$$(s_1 + s_2 + a)H_2(s_1, s_2) = c\,H_1(s_1)H_1(s_2)$$

we immediately obtain the second order nonlinear transfer function

$$H_2(s_1, s_2) = \frac{c\,H_1(s_1)H_1(s_2)}{s_1 + s_2 + a}\,.$$

Note that it’s often convenient to write higher-order transfer functions in terms of the first-order one. In this example

$$H_2(s_1, s_2) = c\,H_1(s_1 + s_2)H_1(s_1)H_1(s_2)\,.$$

To obtain the second order impulse response we can inverse Laplace transform, first with respect to one Laplace variable, then with respect to the other one, and finally by symmetrizing the result. We first inverse transform with respect to s_2 the expression

$$H_1(s_1 + s_2)H_1(s_2) = \frac{1}{[s_2 + (s_1 + a)](s_2 + a)}\,.$$

Assuming $s_1 \neq 0$[1] and expanding in partial fractions we find

[1] The obtained expression is a continuous function of s_1 which we extend by continuity to $s_1 = 0$.

$$\frac{1}{s_1}\left(1-\mathrm{e}^{-s_1\,\tau_2}\right)\mathrm{e}^{-a\,\tau_2}\,1_+(\tau_2)\,.$$

We then combine this expression with the other factors of H_2

$$\frac{c}{(s_1+a)s_1}\left(1-\mathrm{e}^{-s_1\,\tau_2}\right)\mathrm{e}^{-a\,\tau_2}\,1_+(\tau_2)$$

and inverse transform with respect to s_1. This can be done by expanding in partial fractions the first factor

$$\mathcal{L}^{-1}\left\{\frac{c}{(s_1+a)s_1}\right\}(\tau_1)=\frac{c}{a}\left(1-\mathrm{e}^{-a\tau_1}\right)1_+(\tau_1)$$

and by using the shifting property of the Laplace transform to find

$$\frac{c}{a}\left[\left(1-\mathrm{e}^{-a\tau_1}\right)1_+(\tau_1)-\left(1-\mathrm{e}^{-a(\tau_1-\tau_2)}\right)1_+(\tau_1-\tau_2)\right]\mathrm{e}^{-a\tau_2}\,1_+(\tau_2)\,.$$

Note that this expression is not symmetric and that if we had first inverse transformed with respect to s_1 and then to s_2, we would have obtained an expression with τ_1 and τ_2 exchanged.

The second-order impulse response is obtained from the above expression by symmetrisation

$$h_2(\tau_1,\tau_2)=\left[\frac{c}{a}\left[\left(1-\mathrm{e}^{-a\tau_1}\right)-\left(1-\mathrm{e}^{-a(\tau_1-\tau_2)}\right)1_+(\tau_1-\tau_2)\right]\mathrm{e}^{-a\tau_2}\right]_{\mathrm{sym}}$$

where we have suppressed the explicit Heavyside step functions with the understanding that the expression is zero if $\tau_1<0$ or $\tau_2<0$. As h_2 is a regular distribution, it can be evaluated on the diagonal and we obtain

$$\mathrm{ev}_{\mathrm{d}}(h_2)(t)=\frac{c}{a}\left(\mathrm{e}^{-at}-\mathrm{e}^{-2at}\right)\,.$$

The third order impulse response h_3 is found by retaining only elements belonging to $\mathcal{D}'(\mathbb{R}^3)$ in the equation. As a first step we write

$$(D+a)\delta^{\otimes 3}*h_3=c\,(h_1+h_2)^2$$

for no other term can produce distributions belonging to $\mathcal{D}'(\mathbb{R}^3)$. The right hand side can be expanded with the help of (9.10) and, retaining only the terms of interest, we obtain

$$(D+a)\delta^{\otimes 3}*h_3=2c\,[h_1\otimes h_2]_{\mathrm{sym}}\,.$$

The Laplace transformed equation is

$$(s_1 + s_2 + s_3 + a)H_3(s_1, s_2, s_3) = 2c\,[H_1(s_1)H_2(s_2, s_3)]_{\text{sym}}$$

and with it the third order nonlinear transfer function is readily obtained

$$H_3(s_1, s_2, s_3) = 2c\,H_1(s_1 + s_2 + s_3)\,[H_1(s_1)H_2(s_2, s_3)]_{\text{sym}}\ .$$

By expressing H_2 in terms of H_1 we can write H_3 in terms of H_1 alone

$$\begin{aligned} H_3(s_1, s_2, s_3) = \frac{2}{3}c^2\,H_1(s_1 + s_2 + s_3)H_1(s_1)H_1(s_2)H_1(s_3)& \\ \cdot\,[H_1(s_1 + s_2) + H_1(s_1 + s_3) + H_1(s_2 + s_3)]&\ . \end{aligned}$$

The computation of the third order impulse response proceeds along the same lines as the computation of h_2. After some algebraic manipulations and exploiting the properties of the Laplace transform we obtain a rather long expression whose evaluation on the diagonal is

$$\mathrm{ev}_{\mathrm{d}}(h_3)(t) = \left(\frac{c}{a}\right)^2 \left(\mathrm{e}^{-at} - 2\mathrm{e}^{-2at} + \mathrm{e}^{-3at}\right)\ .$$

At this point it's interesting to compare the first three elements of the approximate solution that we computed here with the exact solution that we calculated in Example 9.2 and that we reproduce here for convenience

$$y(t) = y_0 \frac{\mathrm{e}^{-at}}{1 - y_0\frac{c}{a}(1 - \mathrm{e}^{-at})}\ .$$

If $|y_0 c/a| < 1$ the exact solution can be expanded in a geometric power series

$$\begin{aligned} y(t) &= y_0\mathrm{e}^{-at} \sum_{j=0}^{\infty} \left[\frac{y_0 c}{a}(1 - \mathrm{e}^{-at})\right]^j \\ &= y_0\mathrm{e}^{-at} + y_0^2\frac{c}{a}\left(\mathrm{e}^{-at} - \mathrm{e}^{-2at}\right) \\ &\quad + y_0^3\left(\frac{c}{a}\right)^2\left(\mathrm{e}^{-at} - 2\mathrm{e}^{-2at} + \mathrm{e}^{-3at}\right) + \cdots \\ &= \mathrm{ev}_{\mathrm{d}}(h_1 y_0 + h_2 y_0^2 + h_3 y_0^3)(t) + \cdots \end{aligned}$$

and see that the lowest order terms correspond to the calculated response components y_1, y_2 and y_3. Note also that the convergence radius of the power series derived from the exact solution corresponds to the radius of the largest open ball, centered at the origin and contained in the domain of attraction of the equilibrium point

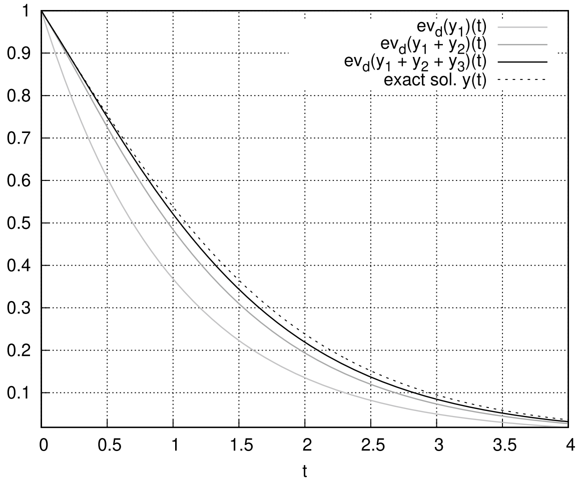


Fig. 9.5 Comparison of the approximate solutions of the logistic differential equation in terms of y_1, y_2 and y_3 against the exact solution $y(t)$ for $a = 1, c = 1/2, y_0 = 1$

$$\mathbb{B}(0, a/c) := \left\{ y_0 \in \mathbb{R} \mid y_0 < \frac{a}{c} \right\} .$$

Figure 9.5 compares the exact solution of the initial value problem with the approximation given by $\mathrm{ev_d}(y_1 + y_2 + y_3)$ for $a = 1, c = 1/2, y_0 = 1$.

While for this particular example it was easier to compute the exact solution than to calculate the approximation, the latter allows us to obtain the output of the system described by the differential equation

$$Dy + ay = x + cy^2 \tag{9.22}$$

for any input signals $x \in \mathcal{D}'_+(\mathbb{R})$ maintaining the system withing the region of attraction of the equilibrium point

$$\mathrm{ev_d}(y)(t) \approx \mathrm{ev_d}(y_1 + y_2 + y_3)(t)$$

with

$$y_1(t) = \int_0^t h_1(t - \tau_1)x(\tau_1)d\tau_1$$

$$\mathrm{ev_d}(y_2)(t) = \int_0^t \int_0^t h_2(t - \tau_1, t - \tau_2)x(\tau_1)x(\tau_2)d\tau_1 d\tau_2$$

$$\mathrm{ev_d}(y_3)(t) = \int_0^t \int_0^t \int_0^t h_3(t - \tau_1, t - \tau_2, t - \tau_3)x(\tau_1)x(\tau_2)x(\tau_3)d\tau_1 d\tau_2 d\tau_3 .$$

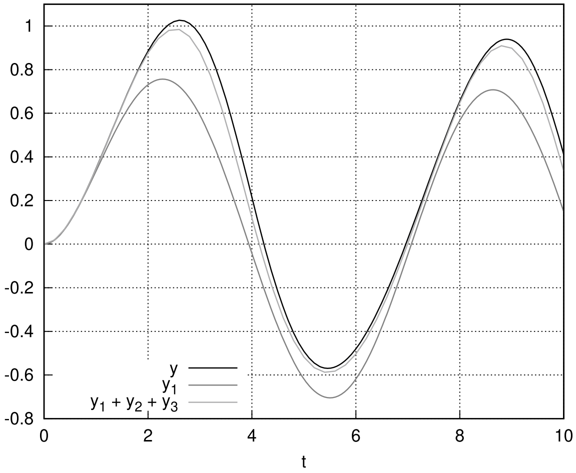


Fig. 9.6 Comparison of the approximate solutions of (9.22) with a sinusoidal input $x(t) = 1_+(t)\sin(t)$ in terms of the three lowest order response components y_1, y_2 and y_3 against the solution $y(t)$ obtained by numerical integration for $a = 1, c = 1/2$

Here and in many problems, this amounts to limiting the magnitude of the input signal to sufficiently small values. Figure 9.6 show the approximate solution for a sinusoidal input $x(t) = 1_+(t)\sin(t)$ and compares it to the solution obtained by numerical integration of the differential equation for $a = 1, c = 1/2$.

This example shows how by representing the solution of a nonlinear differential equation describing a weakly nonlinear system by a sequence of distributions $y \in \mathcal{D}'_{\oplus,\text{sym}}$ we have reduced the problem of solving a nonlinear differential equation to an essentially algebraic problem. While some expressions are rather long, they can be manipulated rather easily by modern computer algebra systems (CAS).

Example 9.6

We revisit Example 9.3 and try to find an approximate solution in $\mathcal{D}'_{\oplus,\text{sym}}$ of the initial value problem

$$Dy = cy^3\,, \qquad y(0) = y_0\,, \qquad c < 0$$

valid around its zero equilibrium point. Note that the linearized equation is stable, but not asymptotically stable.

As before we calculate the impulse responses by setting $y_0 = 1$. The solution for an arbitrary y_0 is then found by multiplying the kth order impulse response h_k by y_0^k.

The first order impulse response h_1 is found by writing the convolution equation corresponding to the above initial value problem and retaining only terms of first order

$$D\delta * h_1 = \delta\,.$$

By Laplace transforming the equation, the first order transfer function $H_1(s_1)$ is found to be

$$H_1(s_1) = \frac{1}{s_1}\,.$$

From it, the first order impulse response is

$$h_1(\tau_1) = \mathsf{1}_+(\tau_1)\,.$$

The equation doesn't have second order nonlinearities. Therefore the second order impulse response and the second order transfer function are both zero

$$h_2(\tau_1,\tau_2) = 0\,, \qquad H_2(s_1,s_2) = 0\,.$$

The third order impulse response is found by retaining all third order terms in the convolution equation

$$D\delta * h_3 = c\, h_1^{\otimes 3}\,.$$

By Laplace transforming the equation we find for the third order transfer function

$$H_3(s_1,s_2,s_3) = \frac{c}{(s_1+s_2+s_3)s_1s_2s_3}\,.$$

From this, the third order impulse response is obtained by inverse Laplace transforming with respect to one variable at a time and by symmetrizing the result

$$\begin{aligned} h_3(\tau_1,\tau_2,\tau_3) = c\Big[&\tau_3\mathsf{1}_+(\tau_3) + (\tau_2-\tau_3)\mathsf{1}_+(\tau_3-\tau_2) \\ &+\mathsf{1}_+(\tau_2-\tau_1)\big[(\tau_1-\tau_3)\mathsf{1}_+(\tau_3-\tau_1) + (\tau_3-\tau_2)\mathsf{1}_+(\tau_3-\tau_2)\big]\Big]_{\text{sym}}\,. \end{aligned}$$

From the above results we could conclude that, to third order, the approximate solution of the initial value problem is

$$\mathrm{ev}_{\mathrm{d}}(y)(t) = y_0\mathsf{1}_+(t) + cy_0^3\mathsf{1}_+(t)t + \cdots\,.$$

This is however only valid for sufficiently small values of t. The reason is best seen by comparing the above expression with the exact solution of the initial value problem that we obtained in Example 9.3 and that we repeat here for convenience

$$y(t) = \frac{y_0}{\sqrt{1-2cy_0^2t}}\,.$$

The Taylor expansion around zero of the function

$$x \mapsto \frac{1}{\sqrt{1-x}}$$

is

$$1+\frac{1}{2}x+\frac{1\cdot 3}{2\cdot 4}x^2+\frac{1\cdot 3\cdot 5}{2\cdot 4\cdot 6}x^3+\frac{1\cdot 3\cdot 5\cdot 7}{2\cdot 4\cdot 6\cdot 8}x^4+\cdots$$

and has a convergence radius of 1. Therefore, as long as $|2cy_0^2t|<1$, the exact solution can be represented by the power series

$$y(t)=y_0\left[1+cy_0^2t+\frac{3}{2}(cy_0^2t)^2+\frac{5}{2}(cy_0^2t)^3+\frac{35}{8}(cy_0^2t)^4+\cdots\right]$$

whose first two terms coincide with $y_0h_1(t)$ and $\mathrm{ev}_\mathrm{d}(y_0^3h_3)(t)$ respectively. However, as t increases, the higher order terms become more and more important and, when $|2cy_0^2t|=1$, the Taylor expansion stops being a valid representation of the exact solution of the initial value problem.

The last example shows that, in general, the solution of a nonlinear differential equation in terms of an element of $\mathcal{D}'_{\oplus,\mathrm{sym}}$ is only meaningful around an equilibrium point for which *the linearized equation is asymptotically stable*. The reason being that, if this is not the case then the response of the system to any part of the input signal can persist indefinitely in time without ever decreasing to negligible levels. Since this is true for the response of any order, the output $\mathrm{ev}_\mathrm{d}(y)$ can not in general be represented by a power series. We can say that *systems that are representable by a Volterra series are those whose output does not depend on the too distant past.*

In the case in which the linearized system is asymptotically stable all impulse responses are summable distributions. Their Fourier transforms are therefore continuous functions that can be obtained from the nonlinear transfer functions H_k by

$$\hat{h}_k(\omega_1,\ldots,\omega_k)=H_k(\jmath\omega_1,\ldots,\jmath\omega_k)\,.$$

As the nonlinear transfer functions are rational functions, the Fourier transforms $\hat{h}_k$ are indefinitely differentiable and of polynomial growth, so they belong to $\mathcal{O}_M$.

9.7 Periodic Input Signals

In this section we investigate the response of weakly nonlinear systems to periodic input signals. Given a periodic input signal x, every tensor power $x^{\otimes k}$ is evidently also a (higher dimensional) periodic distribution. Therefore, every component y_k of the system response y can be calculated in the convolution algebra of periodic distributions and represented by a Fourier series.

Let x be a $\mathcal{T}$-periodic input signal with Fourier coefficients

$$c_m(x) = \frac{1}{\mathcal{T}} \langle x, \mathrm{e}^{-\jmath m \frac{2\pi}{\mathcal{T}} t} \rangle \,, \qquad m \in \mathbb{Z} \,.$$

Further, let $m = (m_1, \dots, m_k) \in \mathbb{Z}^k$ be a multi-index and $\omega_c = 2\pi/\mathcal{T}$, then the Fourier coefficients of the kth tensor power of x are

$$\begin{aligned}
c_m(x^{\otimes k}) &= \frac{1}{\mathcal{T}^k} \langle x^{\otimes k}, \mathrm{e}^{-\jmath \omega_c (m,\tau)} \rangle \\
&= \frac{1}{\mathcal{T}} \langle x, \mathrm{e}^{-\jmath m_1 \omega_c \tau_1} \rangle \cdots \frac{1}{\mathcal{T}} \langle x, \mathrm{e}^{-\jmath m_k \omega_c \tau_k} \rangle \\
&= c_{m_1}(x) \cdots c_{m_k}(x) \,.
\end{aligned}$$

With this expression and a straightforward generalization of Eqs. (4.21) and (4.24) to higher dimensional distributions, the Fourier coefficients of y_k are readily seen to be

$$c_m(y_k) = \hat{h}_k(m_1 \omega_c, \dots, m_k \omega_c)\, c_{m_1}(x) \cdots c_{m_k}(x) \tag{9.23}$$

with $\hat{h}_k$ the Fourier transform of the kth order impulse response of the system.

9.8 Multi-tone Input Signals

In some applications, for example in the study of interference and distortion in communication systems, one is often interested in the response of a system to input signals consisting of sinusoidal tones. If the frequencies of the tones are commensurate, that is, if their ratios are rational numbers, then one can find a common period and the input signal is periodic. The system response can thus be obtained by using the results of the previous section. However, for multi-tone input signals the results are often more directly interpretable by using a different indexing scheme for the tones composing the output components y_k [13].

9.8.1 General Case

Let's consider a system driven by an input consisting of N complex tones

$$x(t) = \sum_{n=1}^{N} A_n \chi_n(t), \quad \chi_n(t) := \mathrm{e}^{\jmath \omega_n t}, \quad A_n := |A_n| \mathrm{e}^{\jmath \varphi_n}$$

initially assumed to have commensurate angular frequencies $\omega_1, \ldots, \omega_N$. Our objective is to calculate the system response of order k

$$y_k = h_k * x^{\otimes k}\,.$$

Consider first the tensor power $x^{\otimes k}$. It can be expanded with the help of (9.10)

$$\begin{aligned} x^{\otimes k} &= \left(\sum_{n=1}^{N} A_n \chi_n\right)^{\otimes k} \\ &= \sum_{|m|=k} \frac{k!}{m!} A_1^{m_1} \cdots A_N^{m_N} \cdot \left[\chi_1^{\otimes m_1} \otimes \chi_N^{\otimes m_N}\right]_{\text{sym}} \end{aligned}$$

with m the multi-index $m = (m_1, \ldots, m_N)$ whose elements range from 0 to N. Observe that this expression is the Fourier series representation of $x^{\otimes k}$. With it and (9.23) the Fourier series representation of y_k is thus found to be

$$\begin{aligned} y_k &= \sum_{|m|=k} \frac{k!}{m!} A_1^{m_1} \cdots A_N^{m_N}\, \hat{h}_{k,m} \cdot \left[\chi_1^{\otimes m_1} \otimes \cdots \otimes \chi_N^{\otimes m_N}\right]_{\text{sym}} \\ \hat{h}_{k,m} &:= \hat{h}_k(\underbrace{\omega_1, \ldots, \omega_1}_{m_1}, \ldots, \underbrace{\omega_N, \ldots, \omega_N}_{m_N}) \end{aligned} \tag{9.24}$$

with $\hat{h}_k$ the Fourier transform of the impulse response of order k. As this sum is finite and only composed by indefinitely differentiable functions, it is itself an indefinitely differentiable function that can be evaluated on the diagonal

$$y_k(t) := \text{ev}_{\text{d}}(y_k)(t) = \sum_{|m|=k} y_{k,m}(t) \tag{9.25}$$

$$y_{k,m}(t) := \frac{k!}{m!} A_1^{m_1} \cdots A_N^{m_N}\, \hat{h}_{k,m}\, \text{e}^{\jmath \omega_m t} \tag{9.26}$$

$$\omega_m := \sum_{n=1}^{N} m_n\, \omega_n = m_1\omega_1 + \cdots + m_N\omega_N\,. \tag{9.27}$$

The kth order response of the system is therefore a sum composed by

$$\frac{(N-1+k)!}{(N-1)!k!} \tag{9.28}$$

complex tones, each one uniquely determined by a specific multi-index m. In this context the multi-index m is also called a *frequency mix* and $|m|$ its *order*.

These results show several important properties of weakly nonlinear systems.

- In contrast to linear systems, weakly nonlinear systems generate tones at frequencies not present at its input.
- In general, tones at a specific frequency are generated by frequency mixes of various orders.
- To fully characterize $\hat{h}_k$ (and hence h_k) one needs k input tones.

At the beginning of this section we assumed the input frequencies to be commensurate. If this is not the case then the input signal is not periodic, but *almost periodic*. For such signals one can still define a Fourier series [16, Sect. VI.9] and the obtained results remains valid.

9.8.2 Real Case

In this section we specialize the above results to the case of a real system driven by an input consisting of N sinusoidal signals

$$x(t) = \sum_{n=1}^{N} |A_n| \cos(\omega_n t + \varphi_n)$$

and where we assume $\omega_1, \ldots, \omega_N > 0$. To re-use previous results it's convenient to represent the input signal in terms of complex exponentials and use separate indexes for positive and negative angular frequencies

$$x(t) = \frac{1}{2} \sum_{n=1}^{N} A_n \chi_n(t) + A_{-n} \chi_{-n}(t), \qquad \chi_n(t) := \mathrm{e}^{\jmath \omega_n t}$$

$$A_n := |A_n| \mathrm{e}^{\jmath \varphi_n}, \qquad A_{-n} := \overline{A}_n = |A_n| \mathrm{e}^{-\jmath \varphi_n}, \qquad \omega_{-n} := -\omega_n \,.$$

The quantity A_n is called the *phasor* of the sinusoidal signal

$$|A_n| \cos(\omega_n t + \varphi_n) \,.$$

With this notation and using the multi-index $m = (m_{-N}, \ldots, m_{-1}, m_1, \ldots, m_N)$ the output component y_k is easily calculated with the help of (9.24)–(9.27)

$$\begin{aligned}
y_k(t) &= \sum_{|m|=k} y_{k,m}(t) \\
y_{k,m}(t) &= \frac{1}{2^k} \frac{k!}{m!} A_{-N}^{m_{-N}} \cdots A_{-1}^{m_{-1}} A_1^{m_1} \cdots A_N^{m_N} \, \hat{h}_{k,m} \, \mathrm{e}^{\jmath \omega_m t} \\
\hat{h}_{k,m} &= \hat{h}_k(\underbrace{\omega_{-N}, \ldots, \omega_{-N}}_{m_{-N}}, \ldots, \underbrace{\omega_N, \ldots, \omega_N}_{m_N})
\end{aligned}$$

$$\omega_m = \sum_{\substack{n=-N\\n\neq 0}}^{N} m_n\,\omega_n = (m_1 - m_{-1})\omega_1 + \cdots + (m_N - m_{-N})\omega_N\,.$$

To N sinusoidal input tones there correspond $2N$ complex tones. Therefore, in the real case, the sum is composed by

$$\frac{(2N-1+k)!}{(2N-1)!k!} \tag{9.29}$$

frequency mixes.

In the real case there is some extra structure that we can exploit. Consider a specific frequency mix $m = (m_{-N}, \ldots, m_{-1}, m_1, \ldots, m_N)$. From the above expression, it's apparent that the multi-index

$$\mathrm{rv}(m) := (m_N, \ldots, m_1, m_{-1}, \ldots, m_{-N}) \tag{9.30}$$

obtained from m by reversing the order of the entries does also appear in the Fourier series of y_k. If $m \neq \mathrm{rv}(m)$ then from $k!/\mathrm{rv}(m)! = k!/m!$, $\omega_{\mathrm{rv}(m)} = -\omega_m$, $A_{\mathrm{rv}(m)} = \overline{A}_m$ and $\hat{h}_{k,\mathrm{rv}(m)} = \overline{\hat{h}}_{k,m}$ we deduce that the sum of $y_{k,m}(t)$ and $y_{k,\mathrm{rv}(m)}(t)$ is a sinusoidal signal

$$\begin{aligned} y^c_{k,m}(t) &:= y_{k,m}(t) + y_{k,\mathrm{rv}(m)}(t) = 2\Re\{y_{k,m}\} \\ &= \frac{1}{2^{k-1}}\frac{k!}{m!}|A_1|^{m_1+m_{-1}} \cdots |A_N|^{m_N+m_{-N}}\,|\hat{h}_{k,m}| \\ &\quad \cdot \cos(\omega_m t + \varphi_m + \psi_{k,m}) \end{aligned} \tag{9.31}$$

with

$$\hat{h}_{k,m} = |\hat{h}_{k,m}|\,\mathrm{e}^{\jmath\psi_{k,m}} \tag{9.32}$$

$$\varphi_m = \sum_{\substack{n=-N\\n\neq 0}}^{N} m_n\,\varphi_n = (m_1 - m_{-1})\varphi_1 + \cdots + (m_N - m_{-N})\varphi_N\,. \tag{9.33}$$

If $m = \mathrm{rv}(m)$ then the multi-index $\mathrm{rv}(m)$ is not distinct from m and the Fourier series component described by $\mathrm{rv}(m)$ coincides with the one described by m. In this case $\omega_m = 0$ and, as the system is assumed to be real, $\hat{h}_{k,m}$ must be real. The response $y_{k,m}$ therefore becomes

$$y_{k,m}(t) = \frac{1}{2^k}\frac{k!}{m!}|A_1|^{2m_1} \cdots |A_N|^{2m_N}\,\hat{h}_{k,m}\,, \qquad m = \mathrm{rv}(m) \tag{9.34}$$

Note that m and $\mathrm{rv}(m)$ can only be equal for even values of k. Also, note that there can be multi-indexes m resulting in $\omega_m = 0$ for which $m = \mathrm{rv}(m)$ doesn't hold.

Example 9.7

Consider again the system described by the differential equation

$$Dy + ay = x + cy^2 \qquad a, c > 0\,.$$

that we analysed in Example 9.5. Here we are interested in the steady state response of the system when driven by the input signal

$$x(t) = |A|\sin(\omega_1 t) = |A|\cos(\omega_1 t - \pi/2).$$

In our previous analysis of this system we calculated the first three nonlinear transfer functions H_1, H_2 and H_3. Using those results, the output components y_1, y_2 and y_3 are immediately obtained from (9.31) and (9.34) without having to calculate any inverse Laplace transform.

Concretely, as the input signal consists of a single sinusoidal tone, the frequency mixes are composed by two entries $m = (m_{-1}, m_1)$. The output of first order y_1 is obtained from the above equations by setting $k = 1$ and by summing over all multi-indexes satisfying the constraint $|m| = m_{-1} + m_1 = 1$. There are only two such multi-indexes: $(0, 1)$ and $\text{rv}((0, 1)) = (1, 0)$. The first order output of the system is therefore given by

$$y_1(t) = \Re\left\{H_1(\jmath\omega_1) A e^{\jmath\omega_1 t}\right\}$$

with $A = |A|e^{-\jmath\pi/2}$.

The second order response of the system y_2 is obtained by setting $k = 2$ and summing over all multi-indexes under the constraint $|m| = 2$. There are three of them: $(2, 0)$, $(0, 2)$ and $(1, 1)$. The first one is the reverse of the second one. Therefore, the contribution of these two is obtained from (9.31)

$$y^c_{2,(0,2)}(t) = \frac{1}{2}\Re\left\{H_2(\jmath\omega_1, \jmath\omega_1) A^2 e^{\jmath 2\omega_1 t}\right\}\,.$$

Since the remaining multi-index is equal to its reverse $(1, 1) = \text{rv}((1, 1))$, its contribution is the constant given by (9.34)

$$y_{2,(1,1)} = |A|^2 H_2(-\jmath\omega_1, \jmath\omega_1)\,.$$

The response of second order is thus

$$y_2(t) = y^c_{2,(0,2)}(t) + y_{2,(1,1)}\,.$$

The third order response of the system y_3 is obtained by setting $k = 3$ and summing over all multi-indexes for which $|m| = 3$. There are four of them: $(3, 0)$, $(2, 1)$, $(1, 2)$ and $(0, 3)$. Two of them are the reverse of the other two. For this reason the response of third order of the system y_3 is given by

$$y_3(t) = y^c_{3,(0,3)}(t) + y^c_{3,(1,2)}(t)$$

with

$$\begin{aligned} y^c_{3,(0,3)}(t) &= \frac{1}{4}\Re\left\{H_3(\jmath\omega_1, \jmath\omega_1, \jmath\omega_1)A^3 \mathrm{e}^{\jmath 3\omega_1 t}\right\} \\ y^c_{3,(1,2)}(t) &= \frac{3}{4}\Re\left\{H_3(-\jmath\omega_1, \jmath\omega_1, \jmath\omega_1)|A|^2\, A\mathrm{e}^{\jmath\omega_1 t}\right\} . \end{aligned}$$

Example 9.8: Two Tones Input

Suppose that we would like to implement a causal real LTI system. However, due to unavoidable limitations of physical components, the implementation behaves as a real weakly nonlinear system characterized by the nonlinear transfer functions H_k (see Fig. 9.4). We are interested in its output when driven by an input signal consisting in two sinusoidal tones

$$x(t) = |A_1| \cos(\omega_1 t + \varphi_1) + |A_2| \cos(\omega_2 t + \varphi_2) .$$

We think of the two tones as closely spaced in frequency and denote the difference of their angular frequencies by $\Delta\omega = \omega_2 - \omega_1$.

As the input is composed by two sinusoidal signals, the frequency mixes have four entries $m = (m_{-2}, m_{-1}, m_1, m_2)$. From (9.29) we calculate that there are 4, 10 and 20 frequency mixes of order one, two and three, respectively. They are listed in Table 9.1.

The first order output y_1 is the output that would be produced by a perfectly linear system. All other tones are undesired. In particular, while tones relatively distant in frequency from ω_1 and ω_2 are relatively easily suppressed with filters, tones close to them are much more difficult to filter out. The tones closest in frequency to ω_1 and ω_2 listed in Table 9.1 are the tones associated with the frequency mixes $(1, 0, 2, 0)$, $(0, 1, 0, 2)$ end their reverses

$$y^c_{3,(1,0,2,0)}(t) = \frac{3}{4}\Re\left\{\overline{A_2}\, A_1^2\, H_3(-\jmath\omega_2, \jmath\omega_1, \jmath\omega_1)\mathrm{e}^{\jmath(\omega_1 - \Delta\omega)t}\right\}$$

and

$$y^c_{3,(0,1,0,2)}(t) = \frac{3}{4}\Re\left\{\overline{A_1}\, A_2^2\, H_3(-\jmath\omega_1, \jmath\omega_2, \jmath\omega_2)\mathrm{e}^{\jmath(\omega_2 + \Delta\omega)t}\right\}$$

both produced by nonlinearities of third order.

The frequency mixes of fifth order are 56. Among them we can easily identify frequency mixes producing tones at every frequency generated by third order nonlinearities, in particular at $\omega_1 - \Delta\omega = 2\omega_1 - \omega_2$. To see this, start with a frequency mix m producing the frequency of interest and add the same number $l > 0$ to m_n and

Table 9.1 Frequency mixes generated by the first, second and third order nonlinearities of a weakly nonlinear system driven by two sinusoidal tones

order	$m = (m_{-2}, m_{-1}, m_1, m_2)$	ω_m
$k = 1$	$(1, 0, 0, 0)$	$-\omega_2 = -\omega_1 - \Delta\omega$
	$(0, 1, 0, 0)$	$-\omega_1 = -\omega_1$
	$(0, 0, 1, 0)$	$\omega_1 = \omega_1$
	$(0, 0, 0, 1)$	$\omega_2 = \omega_1 + \Delta\omega$
$k = 2$	$(2, 0, 0, 0)$	$-2\omega_2 = -2\omega_1 - 2\Delta\omega$
	$(0, 2, 0, 0)$	$-2\omega_1 = -2\omega_1$
	$(0, 0, 2, 0)$	$2\omega_1 = 2\omega_1$
	$(0, 0, 0, 2)$	$2\omega_2 = 2\omega_1 + 2\Delta\omega$
	$(1, 1, 0, 0)$	$-\omega_2 - \omega_1 = -2\omega_1 - \Delta\omega$
	$(1, 0, 1, 0)$	$-\omega_2 + \omega_1 = -\Delta\omega$
	$(1, 0, 0, 1)$	$0 = 0$
	$(0, 1, 1, 0)$	$0 = 0$
	$(0, 1, 0, 1)$	$-\omega_1 + \omega_2 = \Delta\omega$
	$(0, 0, 1, 1)$	$\omega_1 + \omega_2 = 2\omega_1 + \Delta\omega$
$k = 3$	$(3, 0, 0, 0)$	$-3\omega_2 = -3\omega_1 - 3\Delta\omega$
	$(0, 3, 0, 0)$	$-3\omega_1 = -3\omega_1$
	$(0, 0, 3, 0)$	$3\omega_1 = 3\omega_1$
	$(0, 0, 0, 3)$	$3\omega_2 = 3\omega_1 + 3\Delta\omega$
	$(2, 1, 0, 0)$	$-2\omega_2 - \omega_1 = -3\omega_1 - 2\Delta\omega$
	$(2, 0, 1, 0)$	$-2\omega_2 + \omega_1 = -\omega_1 - 2\Delta\omega$
	$(2, 0, 0, 1)$	$-\omega_2 = -\omega_1 - \Delta\omega$
	$(1, 2, 0, 0)$	$-\omega_2 - 2\omega_1 = -3\omega_1 - \Delta\omega$
	$(0, 2, 1, 0)$	$-\omega_1 = -\omega_1$
	$(0, 2, 0, 1)$	$-2\omega_1 + \omega_2 = -\omega_1 + \Delta\omega$
	$(1, 0, 2, 0)$	$-\omega_2 + 2\omega_1 = \omega_1 - \Delta\omega$
	$(0, 1, 2, 0)$	$\omega_1 = \omega_1$
	$(0, 0, 2, 1)$	$2\omega_1 + \omega_2 = 3\omega_1 + \Delta\omega$
	$(1, 0, 0, 2)$	$\omega_2 = \omega_1 + \Delta\omega$
	$(0, 1, 0, 2)$	$-\omega_1 + 2\omega_2 = \omega_1 + 2\Delta\omega$
	$(0, 0, 1, 2)$	$\omega_1 + 2\omega_2 = 3\omega_1 + 2\Delta\omega$
	$(1, 1, 1, 0)$	$-\omega_2 = -\omega_1 - \Delta\omega$
	$(1, 1, 0, 1)$	$-\omega_1 = -\omega_1$
	$(1, 0, 1, 1)$	$\omega_1 = \omega_1$
	$(0, 1, 1, 1)$	$\omega_2 = \omega_1 + \Delta\omega$

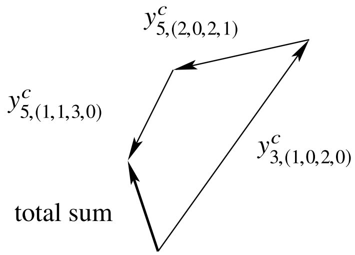


Fig. 9.7 Hypothetical phasor diagram for the response at $\omega_1 - \Delta\omega$ under the assumption that components of order higher than fifth are negligible

m_{-n} for any n ranging from 1 to N (the number of input sinusoidal tones, here 2)

$$m' = (m_{-N}, \ldots, m_{-n} + l, \ldots, m_n + l, \ldots, m_N)\,.$$

Then the order of the new frequency mix m' is $2l$ higher than the one of m and the angular frequencies ω_m and $\omega_{m'}$ associated with the two frequency mixes are identical (see (9.27)).

Using this construction starting from $(1, 0, 2, 0)$, we see that the fifth order mixes $(2, 0, 2, 1)$, $(1, 1, 3, 0)$ and their reverses produce tones at $\omega_1 - \Delta\omega$

$$\begin{aligned} y^c_{5,(2,0,2,1)}(t) &= \frac{15}{8}\Re\left\{\overline{A_2}^2\, A_1^2\, A_2 H_5(-\jmath\omega_2, -\jmath\omega_2, \jmath\omega_1, \jmath\omega_1, \jmath\omega_2)\mathrm{e}^{\jmath(\omega_1-\Delta\omega)t}\right\} \\ y^c_{5,(1,1,3,0)}(t) &= \frac{5}{4}\Re\left\{\overline{A_2}\,\overline{A_1}\, A_1^3\, H_5(-\jmath\omega_2, -\jmath\omega_1, \jmath\omega_1, \jmath\omega_1, \jmath\omega_1)\mathrm{e}^{\jmath(\omega_1-\Delta\omega)t}\right\}\,. \end{aligned}$$

The total response of the system at the frequency $\omega_1 - \Delta\omega$ is therefore a possibly infinite sum composed by the above mixes and higher order ones

$$y^c_{3,(1,0,2,0)} + y^c_{5,(2,0,2,1)} + y^c_{5,(1,1,3,0)} + \cdots\,.$$

This sum can be represented graphically by drawing the phasor of each summand as a vector in the complex plane and summing them by vector addition. Figure 9.7 shows the phasor diagram for the above sum under the assumption that summands of order higher than fifth can be neglected.

Observe that summands of different order depend differently on the amplitude of the input signals $|A_1|$ and $|A_2|$. For small input amplitudes the third order one is usually the dominant. As the amplitude of the input tones grows, higher order summands become first significant and then dominant. This means that both the magnitude as well as the phase of the output tone does change with the amplitude of the input signals. At some level of the input tones there may even be a canceling effect where the output tone becomes very small.

Among the 56 frequency mixes of fifth order there are several of them generating tones at new frequencies. In particular the closest in frequency to ω_1 and ω_2 (not generated by lower order mixes) are at $\omega_1 - 2\Delta\omega$ and $\omega_2 + 2\Delta\omega$. Similarly, higher

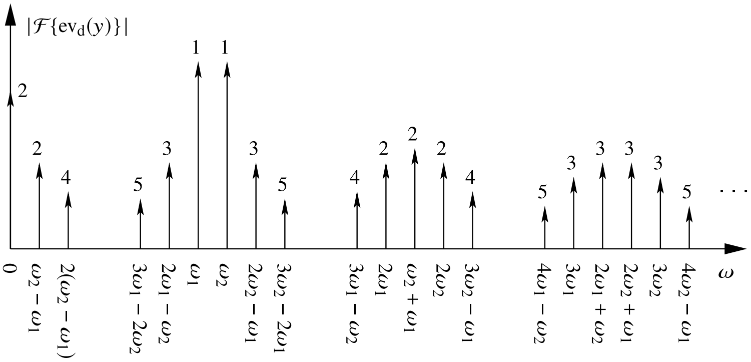


Fig. 9.8 Positive part of a typical magnitude output spectrum of a weakly nonlinear system driven by two sinusoidal input tones. The number q above each spectral line indicates the lowest order nonlinearity generating the line. The same line is also generated by every nonlinearity of order $q + 2l, l \in \mathbb{N}$. Only lines generated by fifth or lower order nonlinearities are shown

odd order frequency mixes introduce tones at new frequencies spaced by $\Delta\omega$ from the previous ones. Figure 9.8 illustrates a typical spectrum of the output signal. For simplicity of representation the figure only shows lines generated by fifth or lower order nonlinearities.

Chapter 10
Composition of Weakly Nonlinear Time Invariant Systems

10.1 Cascade of Noninteracting Systems

When building large systems, it's common to construct them by combining smaller subsystems. To gain the ability to investigate such systems, in this section we study the fundamental operation of *cascading* two systems, that is, of connecting the output of one system to the input of another one. In our treatment we are going to assume that this connection doesn't change the behavior of the involved systems. This is not always the case. Therefore, before applying what follows, we must carefully ponder this aspect.

Consider the cascade of the weakly nonlinear systems $\mathcal{G}$ and $\mathcal{H}$ as shown in Fig. 10.1. Both systems are characterised by their respective impulse responses g_k and h_k that we assume to belong to a convolution algebra $\mathcal{A}'_{\oplus,\mathrm{sym}}$. We are looking for an expression to represent

$$y = (h \circ g)[x] := h[g[x]]\,,$$

the composition of $\mathcal{H}$ after $\mathcal{G}$ that we denote by $\mathcal{H} \circ \mathcal{G}$.

Let's first consider the system $\mathcal{G}$. It's output z when driven by the one dimensional distribution x_1 is

$$z = \sum_{k=1}^{\infty} g_k * x_1^{\otimes k}\,.$$

If instead of representing the input signal by a one dimensional distribution x_1, we represent it by a sequence $x = (0, x_1, 0, \ldots) \in \mathcal{A}'_{\oplus,\mathrm{sym}}$ with all its components but x_1 equal to zero and use the product that we defined on $\mathcal{D}'_{\oplus,\mathrm{sym}}$, then we can express z in the equivalent form

$$z = \sum_{k=1}^{\infty} g_k * x^k\,.$$

F. Beffa, *Weakly Nonlinear Systems*, Understanding Complex Systems,
https://doi.org/10.1007/978-3-031-40681-2_10

Table 10.1 Lowest order impulse responses of the composite system $h \circ g$ in terms of the impulse responses of h and g

$$\begin{aligned}
(h \circ g)_1 &= h_1 * g_1\\
(h \circ g)_2 &= h_1 * g_2 + h_2 * g_1^{\otimes 2}\\
(h \circ g)_3 &= h_1 * g_3 + 2\, h_2 * [g_1 \otimes g_2]_{\text{sym}} + h_3 * g_1^{\otimes 3}\\
(h \circ g)_4 &= h_1 * g_4 + h_2 * \left[g_2^{\otimes 2}\right]_{\text{sym}} + 2\, h_2 * [g_1 \otimes g_3]_{\text{sym}}\\
&\quad + 3\, h_3 * \left[g_1^{\otimes 2} \otimes g_2\right]_{\text{sym}} + h_4 * g_1^{\otimes 4}\\
(h \circ g)_5 &= h_1 * g_5 + 2h_2 * [g_1 \otimes g_4]_{\text{sym}} + 2h_2 * [g_2 \otimes g_3]_{\text{sym}}\\
&\quad + 3h_3 * \left[g_1^{\otimes 2} \otimes g_3\right]_{\text{sym}} + 3h_3 * \left[g_1 \otimes g_2^{\otimes 2}\right]_{\text{sym}}\\
&\quad + 4h_4 * \left[g_1^{\otimes 3} \otimes g_2\right]_{\text{sym}} + h_5 * g_1^{\otimes 5}
\end{aligned}$$

The obtained expression is even more reminiscent of a power series than the original one and, more importantly, it is more amenable to generalisation. In fact, if we assume that this expression remains valid for arbitrary input signals belonging to $\mathcal{A}'_{\oplus,\text{sym}}$ then the same expression can be used to describe the output of $\mathcal{H}$ in terms of z

$$y = \sum_{k=1}^{\infty} h_k * z^k .$$

We can then define the *composition* of weakly nonlinear systems by

$$\begin{aligned}
(h \circ g)[x_1] := \sum_{k=1}^{\infty} (h \circ g)_k * x_1^{\otimes k} &:= \sum_{k=1}^{\infty} h_k * z^k\\
&= h_1 * (g_1 * x_1 + g_2 * x_1^{\otimes 2} + \cdots)\\
&\quad + h_2 * (g_1 * x_1 + g_2 * x_1^{\otimes 2} + \cdots)^2\\
&\quad + h_3 * (g_1 * x_1 + g_2 * x_1^{\otimes 2} + \cdots)^3\\
&\quad + \cdots
\end{aligned} \tag{10.1}$$

with $(h \circ g)_k$ denoting the kth order impulse response of the overall system and consisting of all terms of dimension k. Note that, for every value of k, there are only a finite number of them as the lowest tensor power of x_1 appearing in z^n is the nth one and thus

$$(z^n)_k = 0 \qquad \text{for} \quad n > k .$$

The first five components are listed in Table 10.1 for easy reference. Note, here as well, the analogy with power series and their composition [25].

The above definition by itself is not complete as the convolution between distributions of different dimensions is not defined. To complete the definition we have

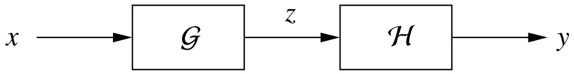


Fig. 10.1 Cascade of the system $\mathcal{G}$ with $\mathcal{H}$

Table 10.2 Convolutions between impulse responses of different order appearing in the composition of weakly nonlinear systems and their definition. They are grouped by the resulting order, from second to fifth. To simplify the notation the symmetrization operation is not explicitly shown

Convolution	Definition
$h_1 * g_2$	$[h_1(\tau_1) \otimes \delta(\tau_2 - \tau_1)] * g_2(\tau_1, \tau_2)$
$h_1 * g_3$	$[h_1(\tau_1) \otimes \delta(\tau_2 - \tau_1, \tau_3 - \tau_1)] * g_3(\tau_1, \tau_2, \tau_3)$
$h_2 * (g_1 \otimes g_2)$	$[h_2(\tau_1, \tau_2) \otimes \delta(\tau_3 - \tau_2)] * [g_1(\tau_1) \otimes g_2(\tau_2, \tau_3)]$
$h_1 * g_4$	$[h_1(\tau_1) \otimes \delta(\tau_2 - \tau_1, \tau_3 - \tau_1, \tau_4 - \tau_1)] * g_4(\tau_1, \tau_2, \tau_3, \tau_4)$
$h_2 * (g_2 \otimes g_2)$	$[h_2(\tau_1, \tau_3) \otimes \delta(\tau_2 - \tau_1, \tau_4 - \tau_3)] * [g_2(\tau_1, \tau_2) \otimes g_2(\tau_3, \tau_4)]$
$h_2 * (g_1 \otimes g_3)$	$[h_2(\tau_1, \tau_2) \otimes \delta(\tau_3 - \tau_2, \tau_4 - \tau_2)] * [g_1(\tau_1) \otimes g_3(\tau_2, \tau_3, \tau_4)]$
$h_3 * (g_1^{\otimes 2} \otimes g_2)$	$[h_3(\tau_1, \tau_2, \tau_3) \otimes \delta(\tau_4 - \tau_3)] * [g_1(\tau_1) \otimes g_1(\tau_2) \otimes g_2(\tau_3, \tau_4)]$
$h_1 * g_5$	$[h_1(\tau_1) \otimes \delta(\tau_2 - \tau_1, \tau_3 - \tau_1, \tau_4 - \tau_1, \tau_5 - \tau_1)] * \ g_4(\tau_1, \tau_2, \tau_3, \tau_4, \tau_5)$
$h_2 * (g_1 \otimes g_4)$	$[h_2(\tau_1, \tau_2) \otimes \delta(\tau_3 - \tau_2, \tau_4 - \tau_2, \tau_5 - \tau_2)] * \ [g_1(\tau_1) \otimes g_4(\tau_2, \tau_3, \tau_4, \tau_5)]$
$h_2 * (g_2 \otimes g_3)$	$[g_2(\tau_1, \tau_3) \otimes \delta(\tau_2 - \tau_1, \tau_4 - \tau_3, \tau_5 - \tau_3)] * \ [g_2(\tau_1, \tau_2) \otimes g_3(\tau_3, \tau_4, \tau_5)]$
$h_3 * (g_1^{\otimes 2} \otimes g_3)$	$[h_3(\tau_1, \tau_2, \tau_3) \otimes \delta(\tau_4 - \tau_3, \tau_5 - \tau_3)] * \ [g_1(\tau_1) \otimes g_1(\tau_2) \otimes g_3(\tau_3, \tau_4, \tau_5)]$
$h_3 * (g_1 \otimes g_2^{\otimes 2})$	$[h_3(\tau_1, \tau_2, \tau_4) \otimes \delta(\tau_3 - \tau_2, \tau_5 - \tau_4)] * \ [g_1(\tau_1) \otimes g_2(\tau_2, \tau_3) \otimes g_2(\tau_4, \tau_5)]$
$h_4 * (g_1^{\otimes 3} \otimes g_2)$	$[h_4(\tau_1, \tau_2, \tau_3, \tau_4) \otimes \delta(\tau_5 - \tau_4)] * \ [g_1(\tau_1) \otimes g_1(\tau_2) \otimes g_1(\tau_3) \otimes g_2(\tau_4, \tau_5)]$

thus to give a meaning to all undefined convolutions appearing in the expression for $(h \circ g)$.

Let's consider the convolutions appearing in $(h \circ g)_k$. The undefined ones are the ones involving h_l with $l < k$. The first thing to note is that, for every l, all of them are convolution products between h_l and a distribution that is the tensor product of l distributions. In addition, by definition, the sum of the dimensions of these l distributions must be k. The convolution products that we have to define have therefore all the form

$$h_l * \left[g_1^{\otimes \alpha_1} \otimes \cdots \otimes g_{k-l+1}^{\otimes \alpha_{k-l+1}} \right]_{\text{sym}}, \qquad \alpha_i \in \mathbb{N} \tag{10.2}$$

with

$$\sum_{i=1}^{k-l+1} i\,\alpha_i = k, \qquad \sum_{i=1}^{k-l+1} \alpha_i = l. \tag{10.3}$$

The simplest case is the one for $l = 1$

$$h_1 * g_k$$

which represents a nonlinearity of order k, g_k, followed by a linear system h_1. To find a suitable general definition for this convolution we let us guide by regular distributions belonging to L_1, evaluated on the diagonal. Setting $k = 2$ for simplicity we have

$$\begin{aligned}\mathrm{ev}_{\mathrm{d}}(z_2)(t) &= \mathrm{ev}_{\mathrm{d}}(g_2 * x_1^{\otimes 2})(t)\\ &= \int_0^\infty \int_0^\infty g_2(\lambda_1, \lambda_2) x_1(t - \lambda_1) x_1(t - \lambda_2) d\lambda_1 d\lambda_2 \,.\end{aligned}$$

Using this expression as the input of h_1 we obtain

$$\begin{aligned}\mathrm{ev}_{\mathrm{d}}(y_2)(t) &= \int_0^\infty h_1(\tau_1) \mathrm{ev}_{\mathrm{d}}(z_2)(t - \tau_1)\, d\tau_1\\ &= \int_0^\infty h_1(\tau_1) \int_0^\infty \int_0^\infty g_2(\lambda_1, \lambda_2)\\ &\quad \cdot x_1(t - \lambda_1 - \tau_1) x_1(t - \lambda_2 - \tau_1)\, d\lambda_1 d\lambda_2 d\tau_1\\ &= \int_0^\infty \int_0^\infty \int_0^\infty h_1(\tau_1) g_2(\lambda_1 - \tau_1, \lambda_2 - \tau_1)\, d\tau_1\\ &\quad \cdot x_1(t - \lambda_1) x_1(t - \lambda_2)\, d\lambda_1 d\lambda_2 \,.\end{aligned}$$

Note that the innermost integral in the last expression is a convolution integral between h_1 and g_2. It can be generalised to arbitrary distributions by building the tensor product of $h_1(\tau_1)$ with $\delta(\tau_2 - \tau_1)$, the Dirac delta distribution in τ_2 parameterised (shifted) by τ_1

$$\begin{aligned}&\langle [h_1(\tau_1) \otimes \delta(\tau_2 - \tau_1)] * g_2(\tau_1, \tau_2), \phi(\tau_1, \tau_2) \rangle\\ &\quad = \langle h_1(\tau_1) \otimes \delta(\tau_2 - \tau_1) \otimes g_2(\lambda_1, \lambda_2), \phi(\tau_1 + \lambda_1, \tau_2 + \lambda_2) \rangle\\ &\quad = \langle h_1(\tau_1) \otimes g_2(\lambda_1, \lambda_2), \langle \delta(\tau_2 - \tau_1), \phi(\tau_1 + \lambda_1, \tau_2 + \lambda_2) \rangle \rangle\\ &\quad = \langle h_1(\tau_1) \otimes g_2(\lambda_1, \lambda_2), \phi(\tau_1 + \lambda_1, \tau_1 + \lambda_2) \rangle \;.\end{aligned}$$

The above derivation generalises without any difficulty to the convolution between h_1 and the impulse response of order k of $\mathcal{G}$. Taking into account that impulse responses have to be symmetric, we thus define the convolution between h_1 and g_k by

$$(h_1 * g_k)(\tau_1, \dots, \tau_k) := [h_1(\tau_1) \otimes \delta(\tau_2 - \tau_1, \dots, \tau_k - \tau_1)]_{\mathrm{sym}} * g_k(\tau_1, \dots, \tau_k) \,.$$

In other words, to convolve h_1 with a distribution of dimension k we promote h_1 to a distribution of k dimensions by building the indicated tensor product and use the standard definition of convolution.

The Laplace transformed of $h_1 * g_k$ has a very simple representation and leads to an easy interpretation. With

$$\begin{aligned}&\langle h_1(\tau_1) \otimes \delta(\tau_2 - \tau_1, \ldots, \tau_k - \tau_1), \mathrm{e}^{-s_1\tau_1 - \cdots - s_k\tau_k}\rangle \\ &\quad = \langle h_1(\tau_1), \langle \delta(\tau_2 - \tau_1, \ldots, \tau_k - \tau_1), \mathrm{e}^{-s_1\tau_1 - \cdots - s_k\tau_k}\rangle\rangle \\ &\quad = \langle h_1(\tau_1), \mathrm{e}^{-(s_1 + \cdots + s_k)\tau_1}\rangle\end{aligned}$$

we find

$$\mathcal{L}\{h_1 * g_k\}(s_1, \ldots, s_k) = H_1(s_1 + \cdots + s_k)\, G_k(s_1, \ldots, s_k)\,.$$

Therefore, if the input signal x_1 consists of N tones, the nonlinear system component g_k generates new tones at frequencies that are linear combinations of k of the input frequencies at a time (see (9.25)). The linear system h_1 following it simply filters these newly generated tones as prescribed by its transfer function H_1, in accordance with expectation.

Consider next the next simplest undefined convolution

$$h_2 * [g_1 \otimes g_2]_{\mathrm{sym}}\ .$$

As for the previous case, we look for a way to promote h_2 to a distribution of dimension $k = 3$ so that we can use the standard definition of convolution. We do so by working with multi-tone input signals as this leads to easier interpretations.

Let $g_1 \otimes g_2$ be driven by 3 unit tones

$$x_1(t) = \mathrm{e}^{\jmath\omega_1 t} + \mathrm{e}^{\jmath\omega_2 t} + \mathrm{e}^{\jmath\omega_3 t}\,,$$

then its output is

$$\begin{aligned}&[g_1 \otimes g_2]_{\mathrm{sym}} * x_1^{\otimes 3} \\ &\quad = \sum_{n_1=1}^{3}\sum_{n_2=1}^{3}\sum_{n_3=1}^{3} G_1(\jmath\omega_{n_1})G_2(\jmath\omega_{n_2}, \jmath\omega_{n_3})\mathrm{e}^{\jmath(\omega_{n_1}\tau_1 + \omega_{n_2}\tau_2 + \omega_{n_3}\tau_3)} \\ &\quad = \sum_{n_1=1}^{3}\sum_{n_2=1}^{3}\sum_{n_3=1}^{3} G_1(\jmath\omega_{n_1})\mathrm{e}^{\jmath\omega_{n_1}\tau_1}\, G_2(\jmath\omega_{n_2}, \jmath\omega_{n_3})\mathrm{e}^{\jmath(\omega_{n_2}\tau_2 + \omega_{n_3}\tau_3)}\end{aligned}$$

with $G_1(s_1)G_2(s_2, s_3)$ the Laplace transform of $g_1 \otimes g_2$. This expression suggests that $g_1 \otimes g_2$ can be interpreted as the parallel combination of a linear system and a second order one. For each term of the sum, the tone at ω_{n_1} passes through the linear system g_1 while the other two pass through g_2. The output $\mathrm{ev}_{\mathrm{d}}(g_1 \otimes g_2)(t)$ can thus be considered as consisting of a sum of pairs of tones, one at ω_{n_1} and the other at $\omega_{n_2} + \omega_{n_3}$. This sum of tones couples constitute the input of h_2 which processes them and, for each tone couple generates the signal

$$H_2(\jmath\omega_{n_1}, \jmath\omega_{n_2} + \jmath\omega_{n_3})G_1(\jmath\omega_{n_1})G_2(\jmath\omega_{n_2}, \jmath\omega_{n_3})\mathrm{e}^{\jmath(\omega_{n_1}+\omega_{n_2}+\omega_{n_3})t} .$$

Given these considerations, we define the convolution between h_2 and $[g_1 \otimes g_2]_{\text{sym}}$ by

$$h_2 * [g_1 \otimes g_2]_{\text{sym}} := [(h_2(\tau_1, \tau_2) \otimes \delta(\tau_3 - \tau_2)) * (g_1(\tau_1) \otimes g_2(\tau_2, \tau_3))]_{\text{sym}} .$$

Its Laplace transform is

$$\mathcal{L}\{h_2 * [g_1 \otimes g_2]_{\text{sym}}\}(s_1, s_2, s_3) = [H_2(s_1, s_2 + s_3)G_1(s_1)G_2(s_2, s_3)]_{\text{sym}} .$$

The above considerations can be extended to the general case (10.2). The tensor product of l distributions

$$g_1^{\otimes\alpha_1} \otimes \cdots \otimes g_{k-l+1}^{\otimes\alpha_{k-l+1}} , \qquad \sum_{i=1}^{k-l+1} \alpha_i = l$$

can be thought of as a set of l parallel subsystems of order lower than k. The constraints (10.3) make sure that with k input tones, its output can be made to consists of l tones at linear combinations of the original input frequencies. These can then be passed as input to h_l.

The intended meaning of the generalised convolution expressed by (10.2) can thus be captured by promoting h_l to a k dimensional distribution obtained by building the tensor product of h_l and $k - l$ appropriately shifted δ distributions constructed as follows.

- The first independent variable of each of the l distributions

  $$g_{m_1}(\tau_1, \ldots, \tau_{m_1}) \otimes \cdots \otimes g_{m_j}(\tau_{n+1}, \ldots, \tau_{n+m_j}) \otimes \cdots \otimes g_{m_l}(\tau_{k-m_l+1}, \ldots, \tau_k)$$

  form the list of independent variables of h_l

  $$h_l(\tau_1, \ldots, \tau_{n+1}, \ldots, \tau_{k-m_l+1}) .$$

- For each additional variable of g_{m_j}, $m_j > 1$, we tensor-multiply h_l by a Dirac distribution in this same variable, shifted by the first one

  $$\delta(\tau_{n+2} - \tau_{n+1}) \otimes \cdots \otimes \delta(\tau_{n+m_j} - \tau_{n+1}) .$$

- The resulting k dimensional distribution has finally to be symmetrized.

Table 10.3 Convolutions between impulse responses of different order appearing in the composition of weakly nonlinear systems and their Laplace transforms. They are grouped by the resulting order, from second to fifth. To simplify the notation the symmetrization operation is not explicitly shown

Convolution	Laplace transform
$h_1 * g_2$	$H_1(s_1 + s_2)G_2(s_1, s_2)$
$h_1 * g_3$	$H_1(s_1 + s_2 + s_3)G_3(s_1, s_2, s_3)$
$h_2 * (g_1 \otimes g_2)$	$[H_2(s_1, s_2 + s_3)]G_1(s_1)G_2(s_2, s_3)$
$h_1 * g_4$	$H_1(s_1 + s_2 + s_3 + s_4)G_4(s_1, s_2, s_3, s_4)$
$h_2 * (g_2 \otimes g_2)$	$[H_2(s_1 + s_2, s_3 + s_4)]G_2(s_1, s_2)G_2(s_3, s_4)$
$h_2 * (g_1 \otimes g_3)$	$[H_2(s_1, s_2 + s_3 + s_4)]G_1(s_1)G_3(s_2, s_3, s_4)$
$h_3 * (g_1^{\otimes 2} \otimes g_2)$	$[H_3(s_1, s_2, s_3 + s_4)]G_1(s_1)G_1(s_2)G_2(s_3, s_4)$
$h_1 * g_5$	$H_1(s_1 + s_2 + s_3 + s_4 + s_5)G_4(s_1, s_2, s_3, s_4, s_5)$
$h_2 * (g_1 \otimes g_4)$	$[H_2(s_1, s_2 + s_3 + s_4 + s_5)]G_1(s_1)G_4(s_2, s_3, s_4, s_5)$
$h_2 * (g_2 \otimes g_3)$	$[H_2(s_1 + s_2, s_3 + s_4 + s_5)]G_2(s_1, s_2)G_3(s_3, s_4, s_5)$
$h_3 * (g_1^{\otimes 2} \otimes g_3)$	$[H_3(s_1, s_2, s_3 + s_4 + s_5)]G_1(s_1)G_1(s_2)G_3(s_3, s_4, s_5)$
$h_3 * (g_1 \otimes g_2^{\otimes 2})$	$[H_3(s_1, s_2 + s_3, s_4 + s_5)]G_1(s_1)G_2(s_2, s_3)G_2(s_4, s_5)$
$h_4 * (g_1^{\otimes 3} \otimes g_2)$	$[H_4(s_1, s_2, s_3, s_4 + s_5)]G_1(s_1)G_1(s_2)G_1(s_3)G_2(s_4, s_5)$

A few convolution examples are given in Table 10.2. The Laplace transformed of these examples are tabulated in Table 10.3. With this definition we have completed the description of how to compose weakly nonlinear systems.

Example 10.1: Third-Order Nonlinearity

The third order nonlinearity of $\mathcal{H} \circ \mathcal{G}$ is generated in three distinct ways: First, by the nonlinearity of third order of $\mathcal{H}$ applied to the output of the linear part of $\mathcal{G}$

$$H_3(s_1, s_2, s_3)G_1(s_1)G_1(s_2)G_1(s_2)X_1(s_1)X_1(s_2)X_1(s_3)\,,$$

second, by the nonlinearity of third order of $\mathcal{G}$ passing through the linear part of $\mathcal{H}$

$$H_1(s_1 + s_2 + s_3)G_3(s_1, s_2, s_2)X_1(s_1)X_1(s_2)X_1(s_3)$$

and third, by the second order nonlinearity of $\mathcal{H}$ applied to the output of first and second order of $\mathcal{G}$

$$2\ [H_2(s_1, s_2 + s_3)G_1(s_1)G_2(s_2, s_3)]_{\text{sym}}\ .$$

These mechanisms are represented graphically in Fig. 10.2. In particular one should note that, even if neither $\mathcal{G}$ nor $\mathcal{H}$ shows nonlinearities of third order, the combined system $\mathcal{H} \circ \mathcal{G}$ in general still has an impulse response of third order different from zero.

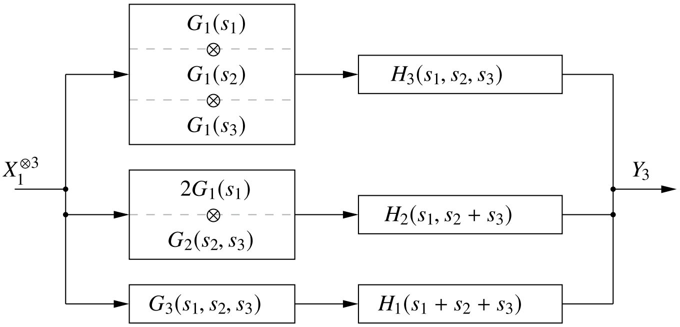


Fig. 10.2 Graphical representation of the third order nonlinearity generated by the composition $\mathcal{H} \circ \mathcal{G}$. Each path has to be understood as symmetrised

Example 10.2: Memory-Less Systems

Consider the convolution(10.2) with h_l a Dirac distribution of dimension $l < k$

$$\delta^{\otimes l} * \left[g_1^{\otimes \alpha_1} \otimes \cdots \otimes g_{k-l+1}^{\otimes \alpha_{k-l+1}} \right]_{\text{sym}} .$$

By definition, the lower dimensional distribution $\delta^{\otimes l}$ is promoted to a distribution of dimension k by building the tensor product with shifted Dirac distributions as explained. For simplicity, we denote the promoted distribution by h_k. Application of the convolution to a test function $\phi \in \mathcal{D}(\mathbb{R}^k)$ is defined by

$$\left\langle h_k(\tau) \otimes \left[g_1^{\otimes \alpha_1} \otimes \cdots \otimes g_{k-l+1}^{\otimes \alpha_{k-l+1}}(\lambda) \right]_{\text{sym}}, \phi(\tau + \lambda) \right\rangle$$
$$= \left\langle \left[g_1^{\otimes \alpha_1} \otimes \cdots \otimes g_{k-l+1}^{\otimes \alpha_{k-l+1}}(\lambda) \right]_{\text{sym}}, \langle h_k(\tau), \phi(\tau + \lambda) \rangle \right\rangle$$

with $\tau, \lambda \in \mathbb{R}^k$. The inner distribution is easily evaluated

$$\langle h_k(\tau), \phi(\tau + \lambda) \rangle = \phi(\lambda)$$

and from this we conclude that for any $l \leq k$

$$\delta^{\otimes l} * \left[g_1^{\otimes \alpha_1} \otimes \cdots \otimes g_{k-l+1}^{\otimes \alpha_{k-l+1}} \right]_{\text{sym}} = \left[g_1^{\otimes \alpha_1} \otimes \cdots \otimes g_{k-l+1}^{\otimes \alpha_{k-l+1}} \right]_{\text{sym}} .$$

With this result we see that the response of a memoryless weakly-nonlinear system can be written in the following equivalent forms

$$y = \sum_{k=1}^{\infty} c_k x^k = \sum_{k=1}^{\infty} c_k \delta^{\otimes l_k} * x^k \,, \qquad l_k \le k \,. \tag{10.4}$$

In general we will use $l_k = k$ so that, if the input signal is a one dimensional distribution x_1, we do not need to use the extended definition of convolution.

Our definition of convolution between distributions of different dimensions and our definition of the one dimensional differential operator operating on higher dimensional distributions (9.14) are compatible. In fact the former is a generalization of the latter. Consider the differential operator acting on the k dimensional Dirac distribution $\delta^{\otimes k}$. Application to a test function $\phi \in \mathcal{D}(\mathbb{R}^k)$ results in

$$\begin{aligned}\left\langle D\delta^{\otimes k}, \phi\right\rangle &= \left\langle \sum_{j=1}^{k} D_j \delta^{\otimes k}, \phi \right\rangle = -\left\langle \delta^{\otimes k}, \sum_{j=1}^{k} D_j \phi \right\rangle \\ &= -\sum_{j=1}^{k} D_j \phi(0, \dots, 0) \,.\end{aligned}$$

If we now consider ϕ as a function of the variable τ_1 only and D_{τ_1} the total differential operator, then we can write

$$\begin{aligned}-\sum_{j=1}^{k} D_j \phi(0, \dots, 0) &= -\left\langle \delta(\tau_1), D_{\tau_1} \phi(\tau_1, \dots, \tau_1)\right\rangle \\ &= \left\langle D_{\tau_1} \delta(\tau_1), \phi(\tau_1, \dots, \tau_1)\right\rangle \\ &= \left\langle (D\delta) * \delta^{\otimes k}, \phi\right\rangle\end{aligned}$$

which shows that our definition of the differential operator acting on a higher dimensional distribution is equal to the convolution of the one dimensional distribution $D\delta$ promoted by our definition of convolution to a k dimensional distribution

$$D\delta^{\otimes k} = (D\delta) * \delta^{\otimes k} \,. \tag{10.5}$$

This is also apparent from the Laplace transformed that in both cases are equal to

$$s_1 + \cdots + s_k \,.$$

The differential operator and the extended definition of convolution do satisfy (3.15). We show this by way of an example. Consider the convolution between h_2 and $g_1 \otimes g_2$. Suppose further that h_2 is the derivative in the sense of (9.14) of another distribution w_2

$$h_2 = D w_2 \,.$$

Applying the convolution product to a test function $\phi \in \mathcal{D}(\mathbb{R}^3)$ and using the extended definition of convolution we obtain

$$\begin{aligned}\langle h_2 * (g_1 \otimes g_2), \phi\rangle &= \langle Dw_2 * (g_1 \otimes g_2), \phi\rangle \\ &= \langle\{[Dw_2(\tau_1, \tau_2)] \otimes \delta(\tau_3 - \tau_2)\} \otimes [g_1(\lambda_1) \otimes g_2(\lambda_2, \lambda_3)], \\ &\quad \phi(\tau_1 + \lambda_1, \tau_2 + \lambda_2, \tau_3 + \lambda_3)\rangle \\ &= \langle[Dw_2(\tau_1, \tau_2)] \otimes [g_1(\lambda_1) \otimes g_2(\lambda_2, \lambda_3)], \\ &\quad \phi(\tau_1 + \lambda_1, \tau_2 + \lambda_2, \tau_2 + \lambda_3)\rangle\ .\end{aligned}$$

Further, using the definition of differentiation and noting that

$$D_{\tau_2}\phi(\tau_1 + \lambda_1, \tau_2 + \lambda_2, \tau_2 + \lambda_3) = (D_{\lambda_2} + D_{\lambda_3})\phi(\tau_1 + \lambda_1, \tau_2 + \lambda_2, \tau_2 + \lambda_3)$$

we obtain

$$\begin{aligned}&- \langle[w_2(\tau_1, \tau_2)] \otimes [g_1(\lambda_1) \otimes g_2(\lambda_2, \lambda_3)], \\ &\qquad (D_{\tau_1} + D_{\tau_2})\phi(\tau_1 + \lambda_1, \tau_2 + \lambda_2, \tau_2 + \lambda_3)\big\rangle \\ &= - \langle[w_2(\tau_1, \tau_2)] \otimes [g_1(\lambda_1) \otimes g_2(\lambda_2, \lambda_3)], \\ &\qquad (D_{\lambda_1} + D_{\lambda_2} + D_{\lambda_3})\phi(\tau_1 + \lambda_1, \tau_2 + \lambda_2, \tau_2 + \lambda_3)\big\rangle \\ &= \langle w_2(\tau_1, \tau_2) \otimes D[g_1(\lambda_1) \otimes g_2(\lambda_2, \lambda_3)], \\ &\qquad \phi(\tau_1 + \lambda_1, \tau_2 + \lambda_2, \tau_2 + \lambda_3)\rangle\end{aligned}$$

or, summarising

$$(Dw_2) * (g_1 \otimes g_2) = w_2 * D(g_1 \otimes g_2)\,. \tag{10.6}$$

10.2 Feedback

A powerful technique used in the design of all sorts of systems is feedback. In control systems design, this technique is used to stabilise and adjust the dynamics of a system to achieve a desired behaviour. It's also used to reduce the sensitivity of systems to poorly controlled parameters. Here we are interested in describing the nonlinearities of a system making use of feedback based on the ones of its constituting subsystems.

Consider the system shown in Fig. 10.3 composed by a forward subsystem $\mathcal{G}$ and a feedback subsystem $\mathcal{H}$. The input of $\mathcal{G}$ is the difference between the input signal x and the signal z, a signal obtained by sensing the output y and suitably processed by $\mathcal{H}$. The system is described by the following equations

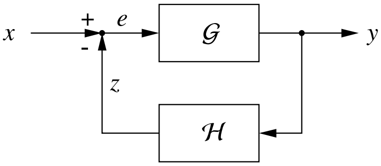


Fig. 10.3 Weakly nonlinear system with feedback

$$\begin{aligned} e &= x - z \\ z &= (h \circ g)[e] \\ y &= g[e] \,. \end{aligned}$$

Our objective is to obtain the impulse responses of the system based on the ones of $\mathcal{G}$ and $\mathcal{H}$. We denote the overall system by $\mathcal{W}$ and its impulse response of order k by w_k.

We start by computing the linear impulse response. The composition of linear systems is obtained by convolving their first order impulse responses. We can therefore write the equation

$$e_1 = \delta - z_1 = \delta - h_1 * g_1 * e_1$$

and, solving for e_1, we obtain

$$e_1 = (\delta + h_1 * g_1)^{*-1} \,.$$

With e_1 the calculation of the linear impulse response is immediate

$$w_1 = g_1 * e_1 = (\delta + h_1 * g_1)^{*-1} * g_1 \,.$$

Its Laplace transform is a classical result of linear system theory

$$W_1(s) = \frac{G_1(s)}{1 + H_1(s)G_1(s)} \,.$$

If in the frequency range of interest the magnitude of the linear loop gain is large $|H_1(\jmath\omega)G_1(\jmath\omega)| \gg 1$ then the linear response of the system is almost exclusively determined by the feedback network

$$W_1(\jmath\omega) \approx \frac{1}{H_1(\jmath\omega)} \,.$$

For completeness, we give the Laplace transform of e_1 as well

$$E_1(s_1) = \frac{1}{1 + H_1(s)G_1(s)} \,.$$

With it the first order transfer function can be written as

$$W_1(s_1) = G_1(s_1)E_1(s_1)\,.$$

Next we compute the impulse response of second order W_2. Using the generalised response of weakly nonlinear systems (10.1), the second order component of the output signal y and of the feedback signal z are given by

$$\begin{aligned} y_2 &= g_2 * e_1^{\otimes 2} + g_1 * e_2 \\ z_2 &= h_2 * y_1^{\otimes 2} + h_1 * y_2\,. \end{aligned}$$

Note that, since we used a Dirac impulse as input, the output components y_2 and y_1 correspond to the impulse responses w_2 and w_1 respectively. By substituting the first equation into the second, using the previous result for w_1 and taking into account the fact that the input signal is a one dimensional distribution, we obtain an equation in e_2

$$z_2 = -e_2 = h_2 * (g_1 * e_1)^{\otimes 2} + h_1 * (g_2 * e_1^{\otimes 2} + g_1 * e_2)\,.$$

whose solution is

$$e_2 = -(\delta^{\otimes 2} + h_1 * g_1)^{*-1} * (h_2 * w_1^{\otimes 2} + h_1 * g_2 * e_1^{\otimes 2})\,.$$

With e_2 and the previous results for e_1 and w_1 the second order impulse response is thus given by

$$w_2 = g_2 * e_1^{\otimes 2} + g_1 * e_2\,.$$

Its Laplace transform is

$$W_2(s_1, s_2) = G_2(s_1, s_2)E_1(s_1)E_1(s_2) + G_1(s_1 + s_2)E_2(s_1, s_2)$$

with

$$E_2(s_1, s_2) = \\ -\frac{H_2(s_1, s_2)W_1(s_1)W_1(s_2) + H_1(s_1 + s_2)G_2(s_1, s_2)E_1(s_1)E_1(s_2)}{1 + H_1(s_1 + s_2)G_1(s_1 + s_2)}\,.$$

Combining these expressions and using previous results, we can write W_2 in the following form

$$\begin{aligned} W_2(s_1, s_2) = &\Big\{E_1(s_1 + s_2)G_2(s_1, s_2) \\ &- W_1(s_1 + s_2)H_2(s_1, s_2)G_1(s_1)G_1(s_2)\Big\}E_1(s_1)E_1(s_2) \end{aligned} \tag{10.7}$$

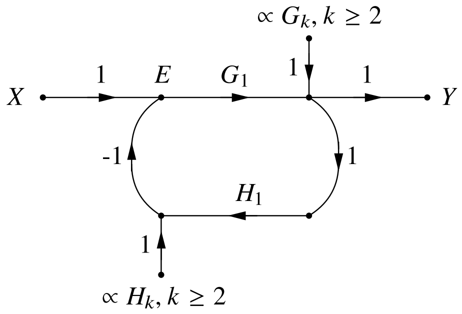


Fig. 10.4 Signal flow graph of a weakly nonlinear system with feedback

which is easily interpretable with the help of the signal flow graph (SFG) shown in Fig. 10.4 (see Appendix A).

The first term is composed by the transmission of the input signal—we think of it as composed by k tones — through the linear system to node E, the input of the nonlinear subsystem $\mathcal{G}$. This part of the signal flow is represented by the factor $E_1(s_1)E_1(s_2)$. The second order nonlinearity of $\mathcal{G}$ then generates a new tone as determined by $G_2(s_1, s_2)$. This newly generated tone is represented in the SFG by a source node because it is different from the input ones. The propagation of the new tone to the output of the system is accounted for by the last factor, $E_1(s_1 + s_2)$.

The second summand in (10.7) has a similar interpretation. The input signal first propagates through the linear system to the input of the other nonlinear subsystem $\mathcal{H}$. This part of the signal flow is represented by $G_1(s_1)E_1(s_1)G_1(s_2)E_1(s_2) = W_1(s_1)W_1(s_2)$. The second order nonlinearity of $\mathcal{H}$ then generates a new tone as determined and accounted for by the $H_2(s_1, s_2)$ factor. Finally, the new tone propagates to the output of the system, contributing the last factor, $-W_1(s_1 + s_2)$.

We now proceed with the calculation of the third order impulse response of the system. The procedure is similar to the one used for the computation of the second order one. From

$$\begin{aligned} y_3 &= g_3 * e_1^{\otimes 3} + 2g_2 * [e_1 \otimes e_2]_{\text{sym}} + g_1 * e_3 \\ z_3 &= h_3 * y_1^{\otimes 3} + 2h_2 * [y_1 \otimes y_2]_{\text{sym}} + h_1 * y_3 \end{aligned}$$

and the previous results we obtain an equation for e_3

$$\begin{aligned} z_3 = -e_3 = h_3 {*} (g_1 * e_1)^{\otimes 3} & \\ &+ 2h_2 * \left[(g_1 * e_1) \otimes (g_2 * e_1^{\otimes 2} + g_1 * e_2)\right]_{\text{sym}} \\ &\quad + h_1 * (g_3 * e_1^{\otimes 3} + 2g_2 * [e_1 \otimes e_2]_{\text{sym}} + g_1 * e_3) \end{aligned}$$

whose solution is

$$\begin{aligned}e_3 = -(\delta^{\otimes 3} + h_1 * g_1)^{*-1} * \Big\{ & h_3 * (g_1 * e_1)^{\otimes 3} \\ & + 2h_2 * \big[(g_1 * e_1) \otimes (g_2 * e_1^{\otimes 2} + g_1 * e_2)\big]_{\text{sym}} \\ & + h_1 * (g_3 * e_1^{\otimes 3} + 2g_2 * [e_1 \otimes e_2]_{\text{sym}})\Big\}.\end{aligned}$$

The third order impulse response is obtained by inserting this expression for e_3 and the previous ones for e_1 and e_2 into

$$w_3 = g_3 * e_1^{\otimes 3} + 2g_2 * [e_1 \otimes e_2]_{\text{sym}} + g_1 * e_3\,.$$

As we find the expressions more easily interpretable, we perform this calculation in the Laplace domain. The Laplace transform of the last expressions for w_3 and e_3 are

$$\begin{aligned}W_3(s_1, s_2, s_3) =& G_3(s_1, s_2, s_3)E_1(s_1)E_1(s_2)E_1(s_3) \\ & + 2\,[G_2(s_1, s_2 + s_3)E_1(s_1)E_2(s_2, s_3)]_{\text{sym}} \\ & + G_1(s_1 + s_2 + s_3)E_3(s_1, s_2, s_3)\end{aligned}$$

and

$$\begin{aligned}E_3(s_1, s_2, s_3) = & \frac{-1}{1 + H_1(s_1 + s_2 + s_3)G_1(s_1 + s_2 + s_3)} \\ & \Big\{ H_3(s_1, s_2, s_3)W_1(s_1)W_1(s_2)W_1(s_3) \\ & + 2\Big[H_2(s_1, s_2 + s_3)W_1(s_1)\big[G_2(s_2, s_3)E_1(s_2)E_1(s_3) \\ & + G_1(s_2 + s_3)E_2(s_2, s_3)\big]\Big]_{\text{sym}} \\ & + H_1(s_1 + s_2 + s_3)\Big[G_3(s_1, s_2, s_3)E_1(s_1)E_1(s_2)E_1(s_3) \\ & + 2\,[G_2(s_1, s_2 + s_3)E_1(s_1)E_2(s_2, s_3)]_{\text{sym}}\Big]\Big\}\end{aligned}$$

respectively. Combining these and previous results we can express W_3 as follows

$$\begin{aligned}W_3(s_1, s_2, s_3) = & E_1(s_1 + s_2 + s_3)G_3(s_1, s_2, s_3)E_1(s_1)E_1(s_2)E_1(s_3) \\ & -W_1(s_1 + s_2 + s_3)H_3(s_1, s_2, s_3)W_1(s_1)W_1(s_2)W_1(s_3) \\ & +2W_1(s_1 + s_2 + s_3)H_2(s_1, s_2 + s_3)W_1(s_1) \\ & \cdot\Big[W_1(s_2 + s_3)H_2(s_2, s_3)W_1(s_2)W_1(s_3) \\ & -E_1(s_2 + s_3)G_2(s_2, s_3)E_1(s_2)E_1(s_3)\Big]_{\text{sym}} \\ & -2E_1(s_1 + s_2 + s_3)G_2(s_1, s_2 + s_3)E_1(s_1)\end{aligned}$$

$$\cdot\Big[H_1(s_2+s_3)E_1(s_2+s_3)G_2(s_2,s_3)E_1(s_2)E_1(s_3)$$
$$+E_1(s_2+s_3)H_2(s_2,s_3)W_1(s_2)W_1(s_3)\Big]_{\text{sym}}. \qquad (10.8)$$

While this expression is rather long, it can be readily interpreted with the help of the SFG of Fig. 10.4. The first term is composed by the factor $E_1(s_1)E_1(s_2)E_1(s_3)$ representing the input signal propagating through the linear part of the system to the input of $\mathcal{G}$. The third order nonlinearity of $\mathcal{G}$ then generates a new tone as witnessed by $G_3(s_1, s_2, s_3)$. Finally, the newly generated tone propagates through the linear part of the system to the output, $E_1(s_1 + s_2 + s_3)$.

The second term has a similar structure and represents the contribution to the third order nonlinearity of $\mathcal{W}$ by the third order nonlinearity of $\mathcal{H}$.

The next summand represents the mixing of the second order nonlinear component of $\mathcal{H}$ with the input signal in the second order nonlinearity of $\mathcal{H}$ (again). Specifically, thinking of the input signal as composed by three tones, the factors $W_1(s_1)$, $W_1(s_2)$ and $W_1(s_3)$ represent the input tones propagating through the linear part of the system to the input of $\mathcal{H}$. There, the second and third tones pass through the second order nonlinearity of $\mathcal{H}$ generating a new second order tone, $H_2(s_2, s_3)$. This second order tone then propagates through the linear part of the system to the input of $\mathcal{H}$, $-W_1(s_2 + s_3)$. There the second order tone and the first input tone pass through the second order distortion of $\mathcal{H}$ together and generate a new third order tone as witnessed by $2H_2(s_1, s_2 + s_3)$. Finally, the third order tone propagates through the linear part of the system to the output, $-W_1(s_1 + s_2 + s_3)$.

The remaining summands have all a similar structure and interpretation as the one just described. They describe the first input tone mixing with a second order tone. The difference between them lies in which subsystem generates the second order tone and which one mixes the first tone with the second order one.

Higher order impulse responses and nonlinear transfer functions of $\mathcal{W}$ can be obtained in a similar way. While the expressions become long, they can easily be computed with the help of computer algebra systems (CAS) computer programs and, referring to the SFG in Fig. 10.4, can be interpreted without difficulty.

From the first three nonlinear transfer functions of the feedback based system $\mathcal{W}$ we can draw the following conclusions.

- The nonlinear transfer functions of the constituting subsystems play the role of controlled sources.
- The linear part of the subsystems plays a pivotal role. It describes the propagation around the system of all input and generated signals.
- The system $\mathcal{W}$ can have an impulse response of order k different from zero even if the impulse responses of order k of both subsystems $\mathcal{G}$ and $\mathcal{H}$ are zero. In particular, we saw how a third order nonlinearity can be generated by various combinations of the nonlinearities of second order of $\mathcal{G}$ and $\mathcal{H}$.
- The nonlinear terms generated exclusively by the forward subsystem $\mathcal{G}$ can be suppressed by making the magnitude of the loop gain $|H_1(s)G_1(s)|$ large (in a suitable portion of the spectrum). This is so because all such terms in the nonlinear

transfer function of order k are proportional to

$$E_1(s_1)\cdots E_1(s_k)E_1(s_1+\cdots+s_k)$$

and, as the loop gain is made large, E_1 becomes small.

- The nonlinear terms generated exclusively by the feedback subsystem $\mathcal{H}$ are not suppressed by making the magnitude of the loop gain $|H_1(s)G_1(s)|$ large. That's because none of these terms are proportional to the linear component of the error signal E_1. Instead, they are all proportional to

$$W_1(s_1)\cdots W_1(s_k)W_1(s_1+\cdots+s_k)$$

which doesn't necessarily become small as the loop gain is made large.

- Nonlinear terms generated by combinations of nonlinearities of $\mathcal{G}$ as well as of $\mathcal{H}$ include factors in E_1 and therefore do experience some level of suppression at large loop gains.

Example 10.3: Linear Feedback

As a special case we consider a system with linear feedback. This means that all transfer functions of $\mathcal{H}$ are zero, except for H_1. In this case the second and third order nonlinear transfer functions of the system are

$$\begin{aligned} W_2(s_1,s_2) &= \frac{G_2(s_1,s_2)E_1(s_1)E_1(s_2)}{1+H_1(s_1+s_2)G_1(s_1+s_2)} \\ &= E_1(s_1+s_2)G_2(s_1,s_2)E_1(s_1)E_1(s_2) \end{aligned}$$

and

$$\begin{aligned} W_3(s_1,s_2,s_3) = E_1(s_1+s_2+s_3)\Big\{&G_3(s_1,s_2,s_3) \\ &- 2\big[G_2(s_1,s_2+s_3)H_1(s_2+s_3)E_1(s_2+s_3)G_2(s_2,s_3)\big]_{\text{sym}}\Big\} \\ &\cdot E_1(s_1)E_1(s_2)E_1(s_3) \end{aligned}$$

respectively. Both of them are proportional to

$$E_1(s_1)\cdots E_1(s_k)E_1(s_1+\cdots+s_k)$$

and can therefore be suppressed by making the loop gain large.

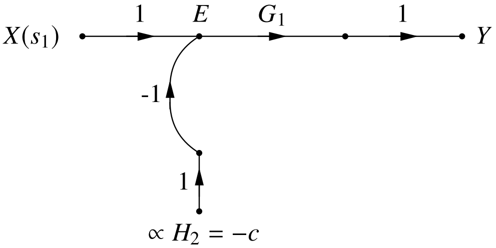


Fig. 10.5 Signal flow graph for the system of Example 10.4

Example 10.4

We revisit Example 9.5 again. Here however we replace the initial condition $y_0\delta$ by a generic input signal x so that the system equation becomes

$$(D\delta + a\delta) * y = x + cy^2.$$

Using (10.1) we can rewrite the equation in the following form

$$y = (D\delta + a\delta)^{*-1} * (x + cy^2)$$

which can be interpreted as describing a linear system with nonlinear feedback. The problem can therefore be recast as the problem of finding the nonlinear transfer functions of a system $\mathcal{W}$ constituted by the forward subsystem $\mathcal{G}$ with linear transfer function

$$G_1(s_1) = \frac{1}{s_1 + a}$$

and a feedback subsystem $\mathcal{H}$ described by the second order nonlinear transfer function

$$H_2(s_1, s_2) = -c$$

as shown in Fig. 10.5. Note that we have assumed negative feedback for consistency with our general treatment. This last expression is obtained by specialising the general expression cy^2 to an input signal having only a one dimensional component y_1

$$cy_1^2 = cy_1^{\otimes 2} = c\delta^{\otimes 2} * y_1^{\otimes 2}.$$

The obtained expression clearly describes a system whose only impulse response differing from zero is the second order one $h_2 = c\delta^{\otimes 2}$ (see also Example 10.2).

In this formulation of the problem the solution is found by inserting the above expressions for the transfer functions of the subsystems into Eqs. (10.7) and (10.8). The obtained expressions obviously agree with the ones obtained in Example 9.5 by calculation from the convolution equation.

10.3 Linearisation

Many systems are designed based on the theory of linear systems and the deviation from linear behavior in practical implementations is undesired. For this reason in practical implementations one often tries to minimise the responses of order higher than one. In this section we investigate the possibility of suppressing higher order responses by preceding the system in question $\mathcal{H}$ with another system $\mathcal{G}$ or by following it with a system $\mathcal{K}$.

We call a system $\mathcal{K}$ designed to suppress all nonlinear transfer functions of $\mathcal{K} \circ \mathcal{H}$ up to order k a *post-lineariser of order* k and a system $\mathcal{K}$ suppressing all responses of $\mathcal{K} \circ \mathcal{H}$ of order higher than one a *post-lineariser*. Similarly, we call a system $\mathcal{G}$ designed to suppress all nonlinear transfer functions of $\mathcal{H} \circ \mathcal{G}$ up to order k a *pre-lineariser of order* k and a system $\mathcal{G}$ suppressing all responses of $\mathcal{H} \circ \mathcal{G}$ of order higher than one a *pre-lineariser* or *pre-distorter*.

We first investigate post-linearisers. The first requirement is that the system $\mathcal{K}$ should not change the linear response of $\mathcal{H}$. This is only the case if the linear impulse response of $\mathcal{K}$ is a Dirac impulse

$$k_1 = \delta \,.$$

Next, we look for a condition to suppress the response of second order. Referring to Table 10.2 we see that the second order response of $\mathcal{K} \circ \mathcal{H}$ disappears if

$$(k \circ h)_2 = k_1 * h_2 + k_2 * h_1^{\otimes 2} = 0 \,.$$

Therefore, if h_1 has an inverse, we can make $(k \circ h)_2$ disappear by choosing

$$k_2 = -h_2 * (h_1^{\otimes 2})^{*-1} \,. \tag{10.9}$$

In the Laplace domain this is

$$K_2(s_1, s_2) = -\frac{H_2(s_1, s_2)}{H_1(s_1)H_1(s_2)} \,. \tag{10.10}$$

Next we look for a condition to suppress on top of $(k \circ h)_2$ also $(k \circ h)_3$. Referring again to Table 10.2 we find the following condition

$$(k \circ h)_3 = k_1 * h_3 + 2\,k_2 * [h_1 \otimes h_2]_{\text{sym}} + k_3 * h_1^{\otimes 3} = 0 \,.$$

As for the second order, this equation can be solved for k_3 only if h_1 has an inverse, in which case, using the previously obtained values for k_1 and k_2, we find

$$k_3 = \big(-h_3 + 2h_2 * (h_1^{\otimes 2})^{*-1} * [h_1 \otimes h_2]_{\text{sym}}\big) * (h_1^{\otimes 3})^{*-1} \tag{10.11}$$

with Laplace transform

$$K_3(s_1, s_2, s_3) = \frac{-H_3(s_1, s_2, s_3) + 2\left[\frac{H_2(s_1, s_2+s_3)}{H_1(s_2+s_3)} H_2(s_2, s_3)\right]_{\text{sym}}}{H_1(s_1)H_1(s_2)H_1(s_3)} . \tag{10.12}$$

This procedure can be extended to find the transfer functions of $\mathcal{K}$ up to order j such that they cancel the nonlinear responses of $\mathcal{K} \circ \mathcal{H}$ up to the jth order. The condition for the existence of k_j is always the same: the existence of the inverse of h_1. This is so because in each equation $(k \circ h)_j = 0$, k_j appears convolved with $h_1^{\otimes k}$. If we let j tend to infinity we obtain a post-lineariser suppressing all nonlinear responses of $\mathcal{H}$.

The impulse responses of a pre-lineariser $\mathcal{G}$ can be obtained following a similar procedure. To preserve the response of $\mathcal{H}$, its linear response must be a Dirac impulse as for a post-lineariser

$$g_1 = \delta.$$

The second order response of $\mathcal{H} \circ \mathcal{G}$ disappears if

$$g_2 = -h_1^{*-1} * h_2 \tag{10.13}$$

or, expressed in the Laplace domain, if

$$G_2(s_1, s_2) = -\frac{H_2(s_1, s_2)}{H_1(s_1 + s_2)} . \tag{10.14}$$

The third order response of $\mathcal{H} \circ \mathcal{G}$ disappears if

$$g_3 = h_1^{*-1} * \left(-h_3 + 2h_2 * \left[\delta \otimes (h_1^{*-1} * h_2)\right]_{\text{sym}}\right) \tag{10.15}$$

whose Laplace transform is

$$G_3(s_1, s_2, s_3) = \frac{-H_3(s_1, s_2, s_3) + 2\left[H_2(s_1, s_2 + s_3)\frac{H_2(s_2, s_3)}{H_1(s_2+s_3)}\right]_{\text{sym}}}{H_1(s_1 + s_2 + s_3)} \tag{10.16}$$

and so on. Again, the prerequisite for the existence of these solutions is the existence of the inverse of h_1. Note also that in general the transfer functions of a pre-lineariser are different from the ones of a post-lineariser.

In summary, we can state that a *weakly nonlinear system can be linearised with a pre- or a post-lineariser only if its linear transfer function has a stable inverse* in the convolution algebra of interest. In the convolution algebra of right sided distributions this means the existence of a causal and stable inverse.

A generic linear system may not have an inverse. For example, if the linear impulse response h_1 is a right-sided, indefinitely differentiable function then $h_1 * w$ is an indefinitely differentiable function independently from the choice of w. This means that $h_1 * w = \delta$ has no solution and hence h_1 has no inverse.

A class of systems of special interest to us is the class of causal systems whose transfer functions are rational functions

$$H_1(s) = \frac{N(s)}{P(s)} .$$

For this class of systems $H_1(s)$ is stable and has a causal stable inverse if all poles *and zeros* of $H_1(s)$ are in the left-half of the complex plane.

Example 10.5: Memory-less System Linearisation

In this example we consider a third order memory-less system $\mathcal{H}$ with impulse responses

$$h_1 = a_1\delta \qquad h_2 = 0 \qquad h_3 = -a_3\delta^{\otimes 3} .$$

We would like to find a pre-lineariser $\mathcal{G}$ suppressing the responses of third order.

The linear impulse response of the system has an inverse

$$a_1\delta * \frac{1}{a_1}\delta = \delta.$$

Therefore it can be linearised using the results of this section. As $h_2 = 0$, the second-order impulse response of the pre-lineariser must also vanish

$$g_2 = 0 .$$

The third order impulse response of the pre-lineariser is obtained by applying (10.15) and we find

$$g_3 = \frac{1}{a_1}\delta * a_3\delta^{\otimes 3} = \frac{a_3}{a_1}\delta^{\otimes 3} .$$

Note that while the pre-lineariser suppresses responses of third order, it does introduce responses of higher order

$$\begin{aligned} h \circ g &= a_1\delta * (\delta + \frac{a_3}{a_1}\delta^{\otimes 3}) - a_3\delta^{\otimes 3} * (\delta + \frac{a_3}{a_1}\delta^{\otimes 3})^3 \\ &= a_1\delta - 3\frac{a_3^2}{a_1}\delta^{\otimes 5} - 3\frac{a_3^3}{a_1^2}\delta^{\otimes 7} - \frac{a_3^4}{a_1^3}\delta^{\otimes 9} . \end{aligned}$$

It's easy to see that to suppress the nonlinear responses up to order k the pre-lineariser must be of order k. To suppress them all a full pre-lineariser is needed.

10.4 System Manipulations

In this section we highlight some properties of weakly nonlinear systems that allow us to manipulate weakly nonlinear system composed by sub-systems in such a way as to obtain different interconnections of the sub-systems without changing the behavior of the overall system.

The first property that we discuss is the associativity of addition which comes from the fact that $\mathcal{D}'_{\oplus,\text{sym}}$ is a vector space. Thus if f, g and h are three weakly nonlinear systems driven by the same input signal x, the ways in which the outputs are summed is irrelevant

$$(f[x] + g[x]) + h[x] = f[x] + (g[x] + h[x]) = f[x] + g[x] + h[x]\,.$$

The same is true for the product of the output signals

$$(f[x] \cdot g[x]) \cdot h[x] = f[x] \cdot (g[x] \cdot h[x]) = f[x] \cdot g[x] \cdot h[x]\,.$$

This is the case because the product that we defined on $\mathcal{D}'_{\oplus,\text{sym}}$ is defined in terms of the tensor product and the latter is associative.

A second important property is commutativity. Addition is always commutative, therefore the order in which the signal appears as input to adders is irrelevant

$$f[x] + g[x] = g[x] + f[x]\,.$$

While the tensor product is not commutative the *symmetrised* tensor product is and with it the product in $\mathcal{D}'_{\oplus,\text{sym}}$

$$f[x] \cdot g[x] = g[x] \cdot f[x]\,.$$

Thus the order in which the signals appearing as input to multipliers is irrelevant as well. In fact, because it's cumbersome to draw symmetrised block diagrams, *we will generally draw unsymmetrised block diagrams and, if not stated explicitly, imply symmetrisation.*

A further equivalence of block diagrams comes from the distributivity of the product over addition

$$\begin{aligned}(f[x] + g[x]) \cdot h[x] &= f[x] \cdot h[x] + g[x] \cdot h[x]\,,\\ f[x] \cdot (g[x]) + h[x]) &= f[x] \cdot g[x] + f[x] \cdot h[x]\,.\end{aligned}$$

This property originates from the multi-linearity of the tensor product. A block diagram representation of the first equality is shown in Fig. 10.6.

Another equivalence is given by the equation

$$(g \circ f)[x] + (h \circ f)[x] = (g + h) \circ f[x]\,.$$

To prove the validity of this equation we prove its validity for terms of each order individually. To simplify the expressions, let's denote the sum of all lth tensor products resulting in a distribution of order k by

$$f_k^{(l)} := \sum_{\substack{|\alpha|=l \\ |\kappa\alpha|=k}} [f^{\otimes\alpha}]_{\text{sym}}$$

with α a multi-indexe in $\mathbb{N}^k$ and $\kappa = (1, 2, \ldots, k)$. With this notation the kth order impulse responses of the summands on the left-hand side can be written as

$$(g \circ f)_k = \sum_{l=1}^{k} g_l * f_k^{(l)}, \qquad (h \circ f)_k = \sum_{l=1}^{k} h_l * f_k^{(l)}.$$

The two can be combined using the distributivity of convolution (3.13) to obtain

$$\sum_{l=1}^{k} (g_l + h_l) * f_k^{(l)}$$

which is the kth order impulse response of the expression on the right-hand side.

The last useful property in manipulating block diagrams is the *right distributivity of composition*

$$(g \circ f)[x] \cdot (h \circ f)[x] = (g \cdot h) \circ f[x].$$

We prove again this equality by proving its validity for terms of each order individually. The impulse response of order k on the left-hand side is

$$\sum_{i+j=k} \sum_{l=1}^{i} g_l * f_i^{(l)} \sum_{m=1}^{j} h_m * f_j^{(m)}.$$

Hence, dropping symmetrisation operators for simplicity of notation

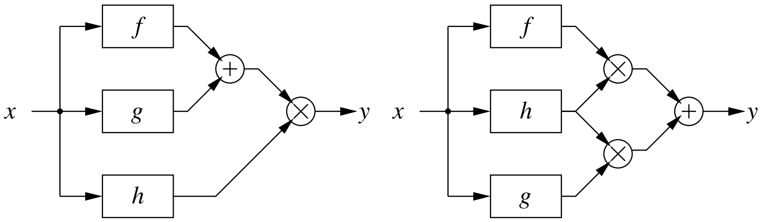


Fig. 10.6 Distributivity of WNTI systems

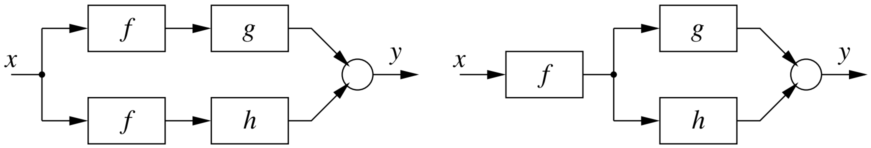


Fig. 10.7 Right distributivity of composition of WNTI systems. The empty circle represents either a sum or a product

$$\sum_{i+j=k} \sum_{l+m\le k} (g_l * f_i^{(l)}) \otimes (h_m * f_j^{(m)}) = \sum_{i+j=k} \sum_{l+m\le k} (g_l \otimes h_m) * (f_i^{(l)} \otimes f_j^{(m)})$$
$$= \sum_{s=1}^{k} \sum_{l+m=s} (g_l \otimes h_m) * f_k^{(s)}$$

which corresponds to the kth order impulse response of the right-hand side of the equation. A block diagram representation of the property is shown in Fig. 10.7.

10.5 Structure

A review of our development of the theory of weakly nonlinear systems up to this point reveals that weakly nonlinear systems arise out of stable linear systems and multipliers. In particular, multipliers are the only mean by which we can combine linear systems to produce systems of higher order.[1] In this section we investigate the overall structure of systems constructed this way.

Let's start by considering the most generic impulse response of second-order that can be constructed out of a single multiplier and linear systems h_A, h_B and h_C

$$h_2(\tau_1, \tau_2) = [h_C * (h_A \otimes h_B)]_{\text{sym}} \,.$$

The block diagram of a system whose only impulse response is h_2 is shown in Fig. 10.8a. We call a system whose only impulse response is h_i a *monomial system of ith order.*

In Sect. 3.3 we showed that every distribution can be approximated to arbitrary accuracy by a set of weighted Dirac impulses. We can thus approximate the linear system h_C by

$$h_C(\tau) \approx \sum_{n=0}^{N} c_n \delta(\tau - \lambda_n), \qquad c_n \in \mathbb{C}, \quad \lambda_n \in [0, \infty).$$

[1] In our formalism multiplication is represented by the tensor product. It is only at the end, when the output signal of interest is "evaluated on the diagoman" with the operator $\text{ev}_\text{d}()$ that the tensor product collapses to a multiplication.

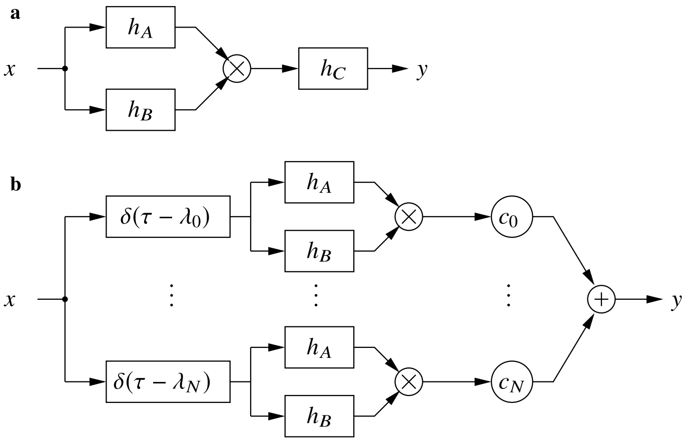


Fig. 10.8 **a** Block diagram of the most generic monomial system of second-order constructed with a single multiplier and linear systems h_A, h_B and h_C. **b** Approximation of the system in Fig. 10.8 a

where we assume the system to be causal. Using this approximation in h_2 we obtain

$$h_2(\tau_1, \tau_2) \approx \sum_{n=0}^{N} \big[c_n \delta(\tau_1 - \lambda_n) * \big(h_A(\tau_1) \otimes h_B(\tau_2)\big)\big]_{\text{sym}} .$$

The shifting property of convolution (3.16) extends to convolutions between distributions of different dimensions in a similar way as the differentiation rule (10.6). In particular for the one dimensional convolution f_1 and the ith dimensional one g_i we have

$$f_1(\tau_1 - \lambda) * g_i(\tau_1, \ldots, \tau_i) = f_1(\tau_1) * g_i(\tau_1 - \lambda, \ldots, \tau_i - \lambda).$$

Using this property the response of the system can be expressed as

$$\begin{aligned} y_2(\tau_1, \tau_2) &= h_2(\tau_1, \tau_2) * \big(x(\tau_1) \otimes x(\tau_2)\big) \\ &\approx \sum_{n=0}^{N} c_n \left[h_A(\tau_1) \otimes h_B(\tau_2)\right]_{\text{sym}} * \big(x(\tau_1 - \lambda_n) \otimes x(\tau_2 - \lambda_n)\big). \end{aligned}$$

This shows that all delays required to approximate h_C to any desired accuracy can be moved to delays of the input signal as illustrated in Fig. 10.8b.

If we use a similar approximation for h_A and h_B we obtain

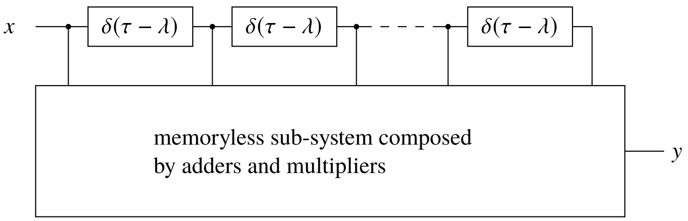


Fig. 10.9 Conceptual structure of a WNTI system

$$y_2(\tau_1, \tau_2) \approx \sum_{n_c=0}^{N_c} \sum_{n_a=0}^{N_b} \sum_{n_b=0}^{N_a} c_{n_c} \left[a_{n_a} b_{n_b}\right]_{\text{sym}} x(\tau_1 - \lambda(n_a + n_c)) \otimes x(\tau_2 - \lambda(n_b + n_c))$$

where we have assumed the use of equal and uniform delays for all sub-systems.

Monomial systems of higher order can be constructed in a similar way by combining linear systems and more multipliers. If we approximate all linear sub-systems as we did above for the second-order monomial system, it's easy to see that all delays can be moved to the input of the system. A system of order K is the sum of monomial sub-systems of order up to K. Therefore, *weakly nonlinear systems of finite order can be represented as composed by two sections*: An input *tapped delay line* sub-system that represents the memory of the system and a *memoryless* sub-system composed by adders and multipliers as illustrated in Fig. 10.9.

An estimate for the maximum delay necessary to faithfully represent a given system of order K can be obtained from the sampling theorem (see Example 12.5): If the maximum frequency component of the input signal is f_{max}, then the highest frequency at the output of the system is $K f_{\text{max}}$ and the delay must be bounded by

$$\lambda < \frac{1}{2\,K f_{\text{max}}}.$$

The number of taps depends on the amount of memory of the linear sub-systems to be approximated.

The system structure represented in Fig. 10.9 is not the most economical one. A comparison between Fig. 10.8a and b reveals that if one moves all the system memory to the input of the system then one needs a larger number of multipliers than by distributing the memory across sub-systems. This is entirely analogous to the trade-off in the implementation of discrete time filters as finite-impulse response (FIR) versus infinite-impulse response (IIR) filters.

Chapter 11
Weakly Nonlinear Time Invariant Circuits

The aim of this chapter is to show the utility of the theory that we developed. This is done by applying it to the analysis of nonlinear effects, that is of deviation from linear behaviour, in analog circuits. The vast majority of analog circuits are limited by noise on the bottom end of their dynamic range and by nonlinear effects on the upper end. While the analysis of noise is well understood by practising engineers, the analysis of nonlinear effects is much less so, and their minimisation poses great practical challenges. The applications presented in this chapter are therefore of practical utility.

The components serving as the building blocks of analog circuits can be represented by linear elements and controlled sources representing nonlinear behaviour. The total response of the circuit can be calculated from a hierarchy of electrical networks with the familiar small-signal linear network forming its core. The hierarchy of networks is constituted by the linear core driven by sources of increasing order. This can be seen as a specialisation to electrical networks of the signal-flow graph method that we saw in Sect. 10.2.

Analog electrical circuits are operated around a stable equilibrium point called the (quiescent) *operating point* of the circuit. The dynamic variables of interest in the theory of weakly nonlinear systems are the ones describing the deviation from the operating point (see Sect. 9.1). We call such variables *small-signal* (or *incremental*) variables. In the following, to distinguish the incremental part of a quantity from the total quantity, we will adopt the notational conventions summarised in Table 11.1.

In Sects. 11.2 and 11.3 of this chapter we develop equivalent circuits for electronic components allowing us to model arbitrary weakly nonlinear analog circuits. In the remaining sections we study concrete circuits used in many types of systems and in particular in communication systems. Before that, in the following section we review a few standard metrics used to characterise the nonlinear behaviour of weakly nonlinear analog circuits.

F. Beffa, *Weakly Nonlinear Systems*, Understanding Complex Systems,
https://doi.org/10.1007/978-3-031-40681-2_11

Table 11.1 Definition of symbols used for various quantities

Definition	Quantity	Subscript	Example
Total quantity	Lower-case	Upper-case	v_C
Operating point	Upper-case	Upper-case	V_C
Small-signal quantity	Lower-case	Lower-case	v_c
kth component of the small-signal quantity	Lower-case	Lower-case and index k	$v_{c.k}$
Laplace transform of the kth component of the small-signal quantity	Upper-case	Lower-case and idex k	$V_{c,k}$

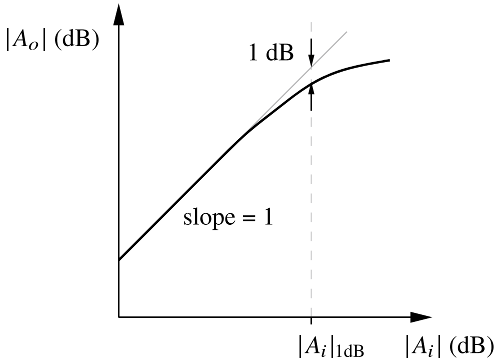
11.1 Metrics for Nonlinear Effects

It's common to distinguish between two classes of nonlinear effects. The first is characterised with input signals of large magnitude, the compression characteristics being the archetypal example. The second is characterised using small signals with intermodulation as the archetypal example. In the following we analyse these and related effects.

11.1.1 Gain Compression and Expansion

Gain compression and gain expansion refer to the change in the gain experienced by a signal passing through a weakly nonlinear system as the amplitude of the input signal changes. At sufficiently large input signal levels all electronic circuits exhibit saturation. However, at the onset of deviation of gain from the small signal value, we may observe a gradual gain reduction, referred to as gain compression; or some gain increase, referred to as gain expansion (see Fig. 11.1). Which of these effects occurs and at which signal level depends on the nonlinear characteristics of the system.

Fig. 11.1 Output signal magnitude versus input signal one of a typical weakly nonlinear system



Consider a weakly nonlinear system $\mathcal{H}$ driven by a sinusoidal signal

$$x(t) = |A_i| \cos(\omega_1 t + \varphi_1) = \Re\{A_i e^{\jmath \omega_1 t}\}.$$

As discussed in Sect. 9.8.2 its output is composed by tones at ω_1 and at integer multiples of it, the harmonics. Let's denote by y_{ω_1} the sum of all the terms at ω_1

$$\begin{aligned} y_{\omega_1}(t) &:= |A_o| \cos(\omega_1 t + \psi_1) = \Re\{A_o e^{\jmath \omega_1 t}\} \\ &= y^c_{1,(0,1)}(t) + y^c_{3,(1,2)}(t) + y^c_{5,(2,3)}(t) + \cdots \\ &= \Re\left\{\left[\hat{h}_{1,(0,1)} + \frac{3}{4}|A_i|^2 \hat{h}_{3,(1,2)} + \frac{5}{8}|A_i|^4 \hat{h}_{5,(2,3)} + \cdots\right] A_i e^{\jmath \omega_1 t}\right\}. \end{aligned}$$

From this expression we see that y_{ω_1} is proportional to the input signal. Therefore, in a similar way as we do with LTI systems, we can consider the ratio of the output signal phasor to the one of the input signal and obtain a sort of frequency response. However, differently from the frequency response of linear systems, the obtained ratio is a function of the input signal amplitude and is called the *describing function*

$$K(|A_i|, \omega_1) := \frac{A_o}{A_i} = \hat{h}_{1,(0,1)} + \frac{3}{4}|A_i|^2 \hat{h}_{3,(1,2)} + \frac{5}{8}|A_i|^4 \hat{h}_{5,(2,3)} + \cdots . \qquad (11.1)$$

Its magnitude is called the *gain* of the system

$$G(|A_i|, \omega_1) := |K(|A_i|, \omega_1)| = \frac{|A_o|}{|A_i|}.$$

At sufficiently small input signal levels, at the onset of nonlinear behaviour, the third order nonlinearity usually dominates and the describing function can be approximated by

$$K(|A_i|, \omega_1) \approx \hat{h}_{1,(0,1)} \cdot \left(1 + \frac{3}{4}|A_i|^2 \frac{\hat{h}_{3,(1,2)}}{\hat{h}_{1,(0,1)}}\right). \qquad (11.2)$$

Note that we have factored the linear frequency response to obtain an explicit factor representing the deviation of the system's behaviour from the one of a perfectly linear system. This factor can be visualised in the complex plane as the sum of the vector

$$\Xi = \Xi_r + \jmath\, \Xi_i := \frac{3}{4}|A_i|^2 \frac{\hat{h}_{3,(1,2)}}{\hat{h}_{1,(0,1)}}; \qquad \Xi_r, \Xi_i \in \mathbb{R}$$

and the unit vector 1 (see Fig. 11.2). If the angle of Ξ is around 0° then the two vectors point approximately in the same direction. Therefore, as the amplitude of the input signal grows, the magnitude of the output signal grows faster than linearly and the system exhibits gain expansion. If the angle of Ξ is around 180° then the two vectors point approximately in opposite directions and the system exhibits gain compression.

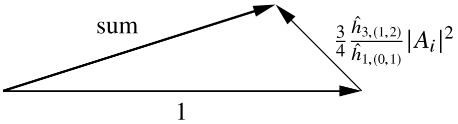

sum
$\frac{3}{4}\frac{\hat{h}_{3,(1,2)}}{\hat{h}_{1,(0,1)}}|A_i|^2$
1

Fig. 11.2 Visualisation of $K(|A_i|, \omega_1)/\hat{h}_{1,(0,1)}$ as the sum in the complex plane of Ξ and the unit vector

If the angle is around ±90° then the vectors are approximately perpendicular and the gain of the system is less sensitive to variations of the input signal (terms of order higher than third will become important). However, in this case it is the *angle* of the output signal that is sensitive to changes in the input signal magnitude. Such a system is said to exhibit *amplitude-modulation (AM) tophase-modulation (PM) conversion.*

Let's have a closer look at the gain of the system. The ratio of the system gain to the one of the system if it would be perfectly linear is called the gain compression/expansion ratio and denoted by GCER

$$\text{GCER} := \frac{G(|A_i|, \omega_1)}{\left|\hat{h}_{1,(0,1)}\right|}. \tag{11.3}$$

Using (11.2), at the onset of deviation from linear behaviour, it is given by

$$\begin{aligned}\text{GCER} &\approx \sqrt{(1+\Xi_r)^2 + \Xi_i^2} \\ &= (1+\Xi_r)\sqrt{1 + \frac{\Xi_i^2}{(1+\Xi_r)^2}}\,.\end{aligned}$$

If we expand the square root in a Taylor series

$$\text{GCER} \approx (1+\Xi_r)(1 + \frac{1}{2}\frac{\Xi_i^2}{(1+\Xi_r)^2} + \cdots)$$

we see that, to first order, the GCER can be estimated by

$$\text{GCER} \approx 1 + \frac{3}{4}\Re\left\{\frac{\hat{h}_{3,(1,2)}}{\hat{h}_{1,(0,1)}}\right\}|A_i|^2\,. \tag{11.4}$$

Given our small signal assumption, this expression should only be used to estimate gain compression or expansion up to ca. 1 dB.

A standard linearity metric used to test analog circuits is the 1 dB *compression point* which is the signal magnitude causing the system gain to decrease by 1 dB. Equation (11.4) allows estimating the magnitude of the input signal producing a given gain compression or expansion

$$|A_i| \approx \sqrt{\frac{4}{3}\left|\frac{(\text{GCER}-1)}{\Re\left\{\frac{\hat{h}_{3,(1,2)}}{\hat{h}_{1,(0,1)}}\right\}}\right|}\,. \tag{11.5}$$

If $\Re\{\hat{h}_{3,(1,2)}/\hat{h}_{1,(0,1)}\}$ is negative the 1 dB compression point can thus be estimated by

$$A_{1\text{dB}} := \frac{0.381}{\sqrt{\left|\Re\left\{\frac{\hat{h}_{3,(1,2)}}{\hat{h}_{1,(0,1)}}\right\}\right|}}\,. \tag{11.6}$$

For small input signals the phase change can also be calculated from the ratio $K(|A_o|, \omega_1)/\hat{h}_{1,(0,1)}$

$$\Delta\psi_1 = \arctan\frac{\Xi_i}{1+\Xi_r} \approx \frac{\Xi_i}{1+\Xi_r} \approx \Xi_i\,.$$

From this we can estimate the input signal magnitude producing a phase change of $\Delta\psi_1$ radiants by

$$|A_i| \approx \sqrt{\frac{4}{3}\left|\frac{\Delta\psi_1}{\Im\left\{\frac{\hat{h}_{3,(1,2)}}{\hat{h}_{1,(0,1)}}\right\}}\right|}\,. \tag{11.7}$$

11.1.2 Intermodulation

In Example 9.8 we analyzed the response of a weakly nonlinear system to a two tones input signal and found that it is composed by several tones at various frequencies. In the context of communication systems and analog circuit design all signal tones at a frequency that is not a multiple of one of the input frequencies are referred to as*intermodulation products*. An intermodulation product is said to be of order k and denoted by IMk if k is the lowest order nonlinearity able to produce it (see Fig. 9.8). For example, given input tones ω_1 and ω_2, the tones at $2\omega_1 - \omega_2$ and $2\omega_2 - \omega_1$ are intermodulation products of third order (IM3); the ones at $3\omega_1 - 2\omega_2$ and $3\omega_2 - 2\omega_1$ of fifth order (IM5).

As an example showing the importance of controlling and limiting the strength of intermodulation products, consider a communication receiver designed for a specific service. Most communication services divide the allocated frequency band in equally spaced channels. Suppose that we are interested in receiving a signal transmitted by a distant transmitter on channel j. Suppose further that the receiver also receives relatively strong interfering signals on channels $j+m$ and $j+2m$ destined to other users. If the receiver is not sufficiently linear then the two interfering signals will produce intermodulation products degrading and possibly completely masking the wanted signal. While the modulation of the involved signals plays a role, due to its

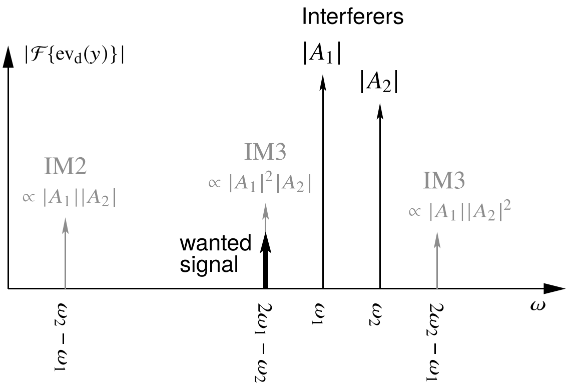


Fig. 11.3 Interfering signals causing the IM3 product to mask the wanted signal

simplicity, communication receivers are also invariably benchmarked and tested with tones as shown in Fig. 11.3.

Consider the two tones input signal

$$x(t) = |A_1|\cos(\omega_1 + \varphi_1) + |A_2|\cos(\omega_2 + \varphi_2) = \Re\{A_1 e^{\jmath\omega_1} + A_2 e^{\jmath\omega_2}\}$$

where we assume $\omega_2 > \omega_1 > 0$. The intermodulation product of order k characterised by the frequency mix m is

$$y^c_{k,m}(t) = \frac{1}{2^{k-1}}\frac{k!}{m!}\Re\{A_1^{m_1}\overline{A_1}^{m_{-1}} A_2^{m_2}\overline{A_2}^{m_{-2}}\hat{h}_{k,m}e^{\omega_m t}\}\,.$$

At relatively low input signal levels the strongest intermodulation products are the second and the third order ones with amplitudes

$$\begin{aligned}
A_{\mathrm{IM2}L} &:= \left|y^c_{2,(0,1,0,1)}(t)\right| = |A_1|\,|A_2|\,\left|\hat{h}_{2,(0,1,0,1)}\right| \\
A_{\mathrm{IM2}H} &:= \left|y^c_{2,(0,0,1,1)}(t)\right| = |A_1|\,|A_2|\,\left|\hat{h}_{2,(0,0,1,1)}\right| \\
A_{\mathrm{IM3}L} &:= \left|y^c_{3,(0,2,0,1)}(t)\right| = \frac{3}{4}\,|A_1|^2\,|A_2|\,\left|\hat{h}_{3,(0,2,0,1)}\right| \\
A_{\mathrm{IM3}H} &:= \left|y^c_{3,(0,1,0,2)}(t)\right| = \frac{3}{4}\,|A_1|\,|A_2|^2\,\left|\hat{h}_{3,(0,1,0,2)}\right|\,.
\end{aligned}$$

These expressions show that the IM2 products are proportional to the amplitudes of each of the two input tones while the IM3 products are proportional to the square of the magnitude of the closest tone and proportional to the magnitude of the more distant one (see Fig. 11.3).

The standard intermodulation test is performed with two tones of equal amplitude

$$|A_1| = |A_2| = A\,.$$

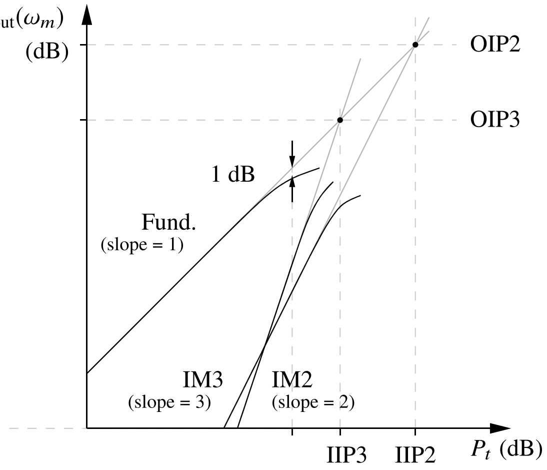


Fig. 11.4 Second and third order intermodulation intercept points

In this case the magnitude of the IM product of order k is proportional to A^k (remember that $|m| = k$)

$$\frac{1}{2^{k-1}} \frac{k!}{m!} A^k \left| \hat{h}_{k,m} \right| .$$

Thus knowing the IMk product level at one value of A is enough to compute its value at a different value of A. This is of course only true at sufficiently small input signals, when the contributions to the IMk product of nonlinearities of order higher than k can be neglected. Instead of specifying the IMk at a specific value of A it is common practice to specify the *intermodulation intercept point* of order k (IPk). This is the level, extrapolated from sufficiently small values of A, at which the IMk reaches the same magnitude as the (linear) output of the system at ω_m when driven by a single tone of magnitude A and frequency ω_m (see Fig. 11.4). The kth order intercept point is thus defined by the equation

$$\frac{1}{2^{k-1}} \frac{k!}{m!} A^k \left| \hat{h}_{k,m} \right| = A \left| \hat{h}_1(\omega_m) \right| .$$

Solving for the amplitude we find

$$A_{\text{IIP}k} := \sqrt[k-1]{\frac{2^{k-1} m!}{k!} \left| \frac{\hat{h}_1(\omega_m)}{\hat{h}_{k,m}} \right|} . \tag{11.8}$$

This quantity is also called the *input referred IPk* and denoted by IIPk. Sometimes it is more convenient to refer this quantity to the output of the circuit in which case it is called the *output referred IPk* and denoted by OIPk. Its value is found by multiplying the IIPk by the linear gain at ω_m

$$A_{\text{OIP}k} := \left|\hat{h}_1(\omega_m)\right| A_{\text{IIP}k}\,. \tag{11.9}$$

The second and third order intercept points are the most important ones and can be estimated by

$$A_{\text{IIP2}} = \left|\frac{\hat{h}_1(\omega_m)}{\hat{h}_{2,m}}\right| \tag{11.10}$$

$$A_{\text{IIP3}} = \sqrt{\frac{4}{3}\left|\frac{\hat{h}_1(\omega_m)}{\hat{h}_{3,m}}\right|}\,. \tag{11.11}$$

Expressed in decibels the IPk assumes a particularly simple form. To that end, let's first rewrite the output referred IP k as

$$\begin{aligned} A_{\text{OIP}k} &= \left|\hat{h}_1(\omega_m)\right| \sqrt[k-1]{\frac{2^{k-1}m!}{k!}\left|\frac{A^k}{A^k}\frac{\hat{h}_1(\omega_m)}{\hat{h}_{k,m}}\right|} \\ &= A\left|\hat{h}_1(\omega_m)\right| \sqrt[k-1]{\frac{A\left|\hat{h}_1(\omega_m)\right|}{A^k\frac{k!}{2^{k-1}m!}\left|\hat{h}_{k,m}\right|}}\,. \end{aligned}$$

Then note that

$$\left(A\left|\hat{h}_1(\omega_m)\right|\right)^2$$

is the output power of the fundamental tone normalised to a load of 1/2 Ω. Similarly,

$$\left(A^k\frac{k!}{2^{k-1}m!}\left|\hat{h}_{k,m}\right|\right)^2$$

is the one of the IMk product. Thus, if for a fixed and sufficiently small value of A we denote by P_o the output power of the fundamental expressed in dB relative to some reference power and by $P_{\text{IM}k}$ the one of the IMk product relative to the same reference level, then we can express the OIPk by

$$\text{OIP}k = P_o + \frac{P_o - P_{\text{IM}k}}{k-1}\,. \tag{11.12}$$

Similarly, by denoting the normalised power of an input tone by P_t, the IIPk can be expressed by

$$\text{IIP}k = P_t + \frac{P_o - P_{\text{IM}k}}{k-1}\,. \tag{11.13}$$

These relationships are easily checked geometrically for the IP2 and IP3 in Fig. 11.4.

In *memory-less* weakly nonlinear systems for which, for every k, $\hat{h}_{k,m}$ is a real number c_k independent of m, the IP3 and the input signal level producing a gain compression/expansion of GCER are both proportional to (see (11.5))

$$\sqrt{\left|\frac{c_1}{c_3}\right|}\,.$$

Therefore, in this type of systems, these two quantities are proportional to each other

$$20\log\left(\frac{A_{1\text{dB}}}{A_{\text{IIP3}}}\right) = 20\log\sqrt{|\text{GCER}-1|}\,.$$

For a memory-less system exhibiting gain compression, the difference between the IP3 and the 1 dB compression point is

$$20\log\left(\frac{A_{1\text{dB}}}{A_{\text{IIP3}}}\right) = 20\log\sqrt{1-10^{-1/20}} \approx -9.6\text{ dB}\,.$$

11.1.3 Desensitisation

The response of an LTI system to a signal is unaffected by the presence of a second signal. As long as we have a way of distinguishing the two signals, for example by separating them in frequency, we can ignore the presence of the second one. This is not the case in nonlinear systems where the response to one signal is affected by the presence of other ones. The effect is again most easily illustrated using a two tones input signal.

Let $\mathcal{H}$ be a weakly nonlinear system driven by the two tones input signal

$$x(t) = |A_1|\cos(\omega_1 + \varphi_1) + |A_2|\cos(\omega_2 + \varphi_2) = \Re\{A_1 e^{\jmath\omega_1} + A_2 e^{\jmath\omega_2}\}\,.$$

The first tone represents the signal of interest, while the second one is an undesired signal that is referred to as a *blocking signal* or a *jammer*. As discussed, the response of the system is composed by several tones at various frequencies, among which several at ω_1. As in Sect. 11.1.1, we denote by y_{ω_1} the sum of all terms at ω_1

$$\begin{aligned} y_{\omega_1}(t) &:= |A_o|\cos(\omega_1 t + \psi_1) = \Re\left\{A_o e^{\jmath\omega_1 t}\right\} \\ &= y^c_{1,(0,0,1,0)}(t) + y^c_{3,(0,1,2,0)}(t) + y^c_{3,(1,0,1,1)}(t) + \cdots \\ &= \Re\left\{\left[\hat{h}_{1,(0,0,1,0)} + \frac{3}{4}\,|A_1|^2\,\hat{h}_{3,(0,1,2,0)} + \frac{3}{2}\,|A_2|^2\,\hat{h}_{3,(1,0,1,1)} + \cdots\right]A_1 e^{\jmath\omega_1 t}\right\}. \end{aligned}$$

and find again an expression that is proportional to the phasor of the first input tone. At relatively small input signal levels the contributions of order higher than third can usually be neglected. In addition, we assume that the magnitude of the blocking signal is much larger than the one of the desired signal

$$|A_1| \ll |A_2| \; .$$

Under these assumptions y_{ω_1} can be simplified to

$$y_{\omega_1}(t) \approx \Re\left\{\left[\hat{h}_{1,(0,0,1,0)} + \frac{3}{2}\,|A_2|^2\,\hat{h}_{3,(1,0,1,1)}\right]A_1 e^{\jmath\omega_1 t}\right\}.$$

Following a procedure similar to the one that we used to analyse gain compression and expansion, we build the ratio of the output phasor to the one of the first tone

$$XM(|A_2|\,,\omega_1) := \frac{A_o}{A_1} = \hat{h}_{1,(0,0,1,0)} \cdot \left(1 + \frac{3}{2}\,|A_2|^2\,\frac{\hat{h}_{3,(1,0,1,1)}}{\hat{h}_{1,(0,0,1,0)}}\right)$$

to obtain a sort of frequency response. Similarly to the approximation of the describing function (11.2), it is the product of the linear frequency response of the system and a factor that characterises the deviation from linear behaviour. Differently from the describing function, however, this second factor depends on the amplitude of the *second* tone, the blocking signal.

The ratio $XM(|A_2|\,,\omega_1)/\hat{h}_{1,(0,0,1.0)}$ can again be visualised in the complex plane as the sum of the unit vector and the vector

$$\frac{3}{2}\,|A_2|^2\,\frac{\hat{h}_{3,(1,0,1,1)}}{\hat{h}_{1,(0,0,1,0)}}\,.$$

If the angle of the latter is close to 180° then the second tone will induce a reduction in the gain experienced by the first one. If the angle is close to 0° it will induce a gain expansion and, if the angle is close to ±90° it will induce mostly a change in the phase of the first tone. The change in gain can be characterised by the magnitude of the above ratio, the *desensitisation ratio*

$$DR := \left|\frac{XM(|A_2|\,,\omega_1)}{\hat{h}_{1,(0,0,1.0)}}\right| = \left|1 + \frac{3}{2}\,|A_2|^2\,\frac{\hat{h}_{3,(1,0,1,1)}}{\hat{h}_{1,(0,0,1,0)}}\right| \tag{11.14}$$

and to second order in $|A_2|$ can be estimated by

$$DR \approx 1 + \frac{3}{2}\,|A_2|^2\,\Re\left\{\frac{\hat{h}_{3,(1,0,1,1)}}{\hat{h}_{1,(0,0,1,0)}}\right\}. \tag{11.15}$$

From this expression we can estimate the magnitude of the blocker causing a certain wanted signal gain change

$$|A_2| \approx \sqrt{\frac{2}{3}\left|\frac{(DR-1)}{\Re\left\{\frac{\hat{h}_{3,(1,0,1,1)}}{\hat{h}_{1,(0,0,1,0)}}\right\}}\right|}. \tag{11.16}$$

If $\Re\{\hat{h}_{3,(1,0,1,1)}/\hat{h}_{1,(0,0,1,0)}\}$ is negative, a desensitisation of 1 dB is produced by a blocker at the 1 dB *blocking level*

$$A_{\mathrm{B1dB}} := \frac{0.269}{\sqrt{\left|\Re\left\{\frac{\hat{h}_{3,(1,0,1,1)}}{\hat{h}_{1,(0,0,1,0)}}\right\}\right|}}. \tag{11.17}$$

The change in phase of the first tone caused by the presence of the blocker can also be estimated from $XM(|A_2|, \omega_1)/\hat{h}_{1,(0,0,1.0)}$. To first order a phase change of $\Delta\psi_1$ radiants is produced by a blocker of magnitude

$$|A_2| \approx \sqrt{\frac{2}{3}\left|\frac{\Delta\psi_1}{\Im\left\{\frac{\hat{h}_{3,(1,0,1,1)}}{\hat{h}_{1,(0,0,1,0)}}\right\}}\right|}. \tag{11.18}$$

Note that if the blocker is modulated, then the modulation will be transferred from it to the wanted signal. For example, if the blocker is amplitude modulated (AM) and the angle of $\hat{h}_{3,(1,0,1,1)}/\hat{h}_{1,(0,0,1,0)}$ is close to either 180°or 0° then the gain experienced by the wanted signal is modulated and, as a result, its output amplitude will also be modulated. If the angle of $\hat{h}_{3,(1,0,1,1)}/\hat{h}_{1,(0,0,1,0)}$ is close to ±90° then an amplitude modulation of the blocker will produce a phase modulation of the wanted signal. This effect of transferring the modulation of one signal to another one is called *cross-modulation.*

11.2 Nonlinear Two-Terminal Elements

In this section we investigate two-terminal electrical components that can be characterised by two quantities x_E and y_E, related by an equation of the form

$$f(x_E, y_E) = 0$$

with f a function called the element x-y characteristic (see Fig. 11.5). If the equation can be expressed as a function of x_E, $y_E = \tilde{f}(x_E)$ then the element is called an$x$-controlled device. Similarly, if it can be expressed as a function of y_E, $x_E = \tilde{f}(y_E)$ then it is called a y-controlled device.

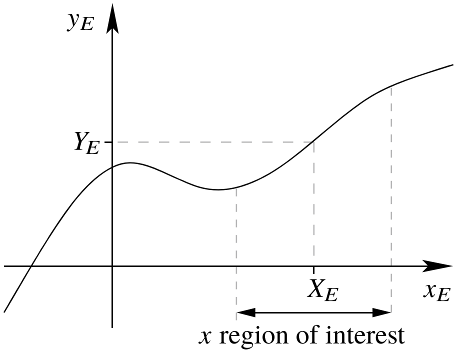


Fig. 11.5 Characteristic of an x-controlled two-terminal element

The devices that interest us are the ones that, in a region of interest around a quiescent operating point (X_E, Y_E), are either x- or y-controlled and whose function $\tilde{f}$ can be approximated to any desired accuracy by a power series

$$y_e = \sum_{k=1}^{\infty} \tilde{f}_k x_e^k \qquad (x\text{-controlled})$$

or

$$x_e = \sum_{k=1}^{\infty} \tilde{f}_k y_e^k \qquad (y\text{-controlled})$$

with

$$y_e = y_E - Y_E, \qquad x_e = x_E - X_E .$$

11.2.1 Nonlinear Resistors

A nonlinear resistor is a device characterised by the current i_R flowing through it, the voltage v_R across its terminals and by an i-v characteristic $f_R(i_R, v_R) = 0$. In the following we are going to represent a nonlinear resistor by the symbol shown in Fig. 11.6. A current controlled resistor can be characterised by a function

$$v_R = r(i_R)$$

which, by assumption, around the operating point (I_R, V_R), can be approximated by a power series

$$v = \sum_{k=1}^{\infty} r_k i^k , \qquad v = v_r = v_R - V_R , \quad i = i_r = i_R - I_R .$$

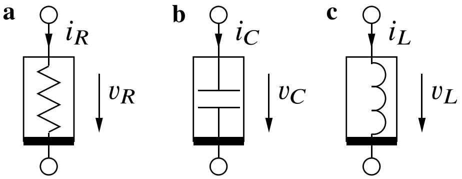


Fig. 11.6 Symbols used to represent nonlinear two-terminal devices. **a** resistor. **b** capacitor. **c** inductor

If we consider the current i and the voltage v as signals, or, more precisely, elements of $\mathcal{D}'_{\oplus,\text{sym}}$, then a nonlinear resistor can be regarded as a weakly nonlinear system and the components of v can be expressed in terms of the ones of i using (10.1) and Table 10.1

$$\begin{aligned} v_1 &= r_1 i_1 \\ v_2 &= r_1 i_2 + r_2 i_1^{\otimes 2} \\ v_3 &= r_1 i_3 + 2r_2 \left[i_1 \otimes i_2\right]_{\text{sym}} + r_3 i_1^{\otimes 3} \\ \dots \end{aligned} \tag{11.19}$$

From this representation we observe that each voltage component v_k is determined (i) by a term proportional to the kth current component i_k and (ii) by other terms proportional to current components of order lower than k. In an electric network, the former can be represented by a linear resistor of value r_1, the latter by a voltage source $\tilde{v}_{R,k}$ whose value is determined by the current components $i_n, n = 1, \dots, k-1$ (see Fig. 11.7a)

$$v_k = r_1 i_k + \tilde{v}_{R,k}(i_1, \dots, i_{k-1}) .$$

The various current and voltage components can therefore be calculated using a hierarchy of *linear* networks. First, we find the linear current i_1 using linearised components and the sources representing the system input. Once i_1 is found, $\tilde{v}_{R,2}$ can be determined. With it we can draw the second order network. It is obtained from the linearised network by removing the system input sources (since they are of first order), by adding the second order source $\tilde{v}_{R,2}$ and, if the case, the ones of other nonlinear components. With this network we compute i_2. Having found the first two components of i, the third order source $\tilde{v}_{R,3}$ can be calculated. We then proceed to draw the third order network which is again composed by the linearised network with the addition of independent sources of third order only. With it, we find i_3 and so on.

If the nonlinear resistor is voltage controlled, then its characteristic around the operating point can be described by a power series where the role of the independent variable is played by the voltage v

$$i = \sum_{k=1}^{\infty} g_k v^k .$$

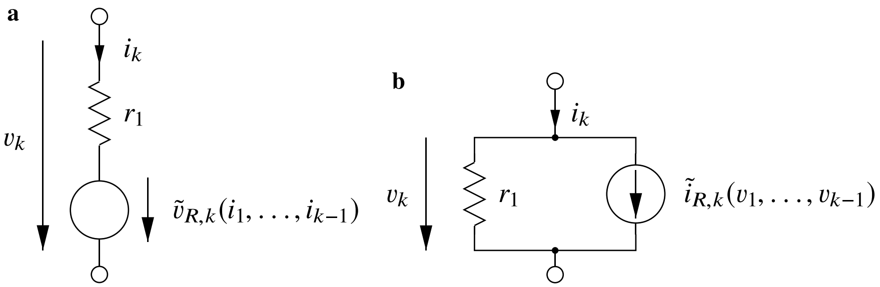

a
i_k
r_1
v_k
$\tilde{v}_{R,k}(i_1,\dots,i_{k-1})$
b
i_k
v_k
r_1
$\tilde{i}_{R,k}(v_1,\dots,v_{k-1})$

Fig. 11.7 **a** Weakly nonlinear resistor current-controlled equivalent model **b** Weakly nonlinear resistor voltage-controlled equivalent model

Proceeding as for the case of a current controlled nonlinear resistor, but with the roles of the signals i and v exchanged, we can express the first few components of the current i in terms of the ones of the voltage

$$\begin{aligned} i_1 &= g_1 v_1 \\ i_2 &= g_1 v_2 + g_2 v_1^{\otimes 2} \\ i_3 &= g_1 v_3 + 2g_2 \left[v_1 \otimes v_2\right]_{\text{sym}} + g_3 v_1^{\otimes 3} \\ &\dots \end{aligned} \tag{11.20}$$

As before, each current component i_k is the sum of a term linear in v_k and other terms only depending on components of v of order lower than k

$$i_k = g_1 v_k + \tilde{i}_{R,k}(v_1,\dots,v_{k-1})\,.$$

From this representation we deduce the equivalent circuit shown in Fig. 11.7b.

If a nonlinear resistor is voltage as well as current controlled, then we can choose the most convenient representation for the problem at hand. If one representation is known, then the other one can be obtained by power series inversion. For example, if we know the voltage-controlled representation, the current-controlled one is obtained by inserting the expression for the components given by (11.20) into (11.19) and by choosing the coefficients r_k so that the equations are satisfied. Specifically, r_1 is found by solving

$$i_1 = g_1 v_1 = g_1 r_1 i_1$$

which gives

$$r_1 = \frac{1}{g_1}\,.$$

r_2 is obtained by solving

$$i_2 = g_1 v_2 + g_2 v_1^{\otimes 2} = g_1(r_1 i_2 + r_2 i_1^{\otimes 2}) + g_2 r_1^2 i_1^{\otimes 2}\,.$$

Using the previously obtained value for r_1, the equation is satisfied if

$$g_1 r_2 + g_2 r_1^2 = 0$$

or

$$r_2 = -\frac{g_2}{g_1^3} \, .$$

r_3 is found in a similar way to be

$$r_3 = \frac{2g_2^2 - g_1 g_3}{g_1^5} \, .$$

Higher order coefficients are easily calculated using the same procedure.

11.2.2 Nonlinear Capacitors

A nonlinear capacitor is a two-terminal device whose voltage v_C across the terminals and the charge q_C stored in it are related by a q-v characteristic $f_C(q_C, v_C) = 0$. In the following, we are going to represent a nonlinear capacitor by the symbol shown in Fig. 11.6. A voltage controlled capacitor is a capacitor whose charge is a function of the voltage $q_C = \tilde{f}_C(v_C)$. Since the electric current is the time derivative of electric charge, if the voltage v_C is a differentiable function of time, the capacitor current is related to the voltage across its terminals by

$$i_C = \frac{\mathrm{d}\tilde{f}_C(v_C)}{\mathrm{d}v_C} \frac{\mathrm{d}v_C}{\mathrm{d}t} \, .$$

The slope of the q-v characteristic is called the *small signal* (or *incremental*) *capacitance* of the nonlinear capacitor

$$C(v_C) := \frac{\mathrm{d}\tilde{f}_C(v_C)}{\mathrm{d}v_C} \, .$$

As before, we assume it to be expandable in a power series around the operating point (Q_C, V_C)

$$c(v) := C(v + V_C) = \sum_{k=0}^{\infty} c_{k+1} v^k \, , \qquad v = v_C - V_C \, .$$

Using this expression in the equation for the current, we can express the latter as the following power series

$$i_C = \sum_{k=0}^{\infty} c_{k+1} v^k \frac{\mathrm{d}v}{\mathrm{d}t} = \sum_{k=1}^{\infty} \frac{c_k}{k} \frac{\mathrm{d}}{\mathrm{d}t} v^k \,.$$

This last expression can be extended to currents and voltages represented by elements of $\mathcal{D}'_{\oplus,\mathrm{sym}}$, so that we take it as defining the relationship between current and voltage of a voltage-controlled weakly-nonlinear capacitor

$$i = i_C = \sum_{k=1}^{\infty} \frac{c_k}{k} D v^k \,. \tag{11.21}$$

The first few components of the current expressed in terms of the components of the voltage are given by

$$\begin{aligned}
i_1 &= c_1 D v_1 \\
i_2 &= c_1 D v_2 + \frac{c_2}{2} D v_1^{\otimes 2} \\
i_3 &= c_1 D v_3 + c_2 D \left[v_1 \otimes v_2\right]_{\mathrm{sym}} + \frac{c_3}{3} D v_1^{\otimes 3} \\
&\dots
\end{aligned} \tag{11.22}$$

Each current component i_k is the sum of a term linear in Dv_k and others that only depend on the voltage components of order lower than k. In an electric network the kth component of the current can therefore be represented by a linear capacitor of value c_1 and a current source (see Fig. 11.8b)

$$i_k = c_1 D v_k + \tilde{i}_{C,k}(v_1, \dots, v_{k-1}) \,.$$

The various components are calculated with the same hierarchy of networks that we described for nonlinear resistors.

An initial charge q_0 on the capacitor can be represented as usual by a current pulse $q_0\delta$ applied across the capacitor in the linear convolution equation.

Note that the linearised current-voltage characteristic of a capacitor is not by itself an asymptotically stable differential equation. The nonlinear transfer function formalism is therefore only applicable when the nonlinear capacitor is embedded in a network whose linear approximation is asymptotically stable.

A charge controlled nonlinear capacitor is a capacitor whose voltage is a function of the charge $v_C = \varsigma(q_C)$. Expanding this function around the operating point (Q_C, V_C) we obtain

$$v = \sum_{k=1}^{\infty} \varsigma_k q^k \,, \qquad v = v_C - V_C \,, \quad q = q_C - Q_C \,.$$

The electric charge is the integral of the current. In the convolution algebra of right sided distributions this can be expressed by the convolution product between current and the Heaviside step function

$$q(t) = \int_0^t i(\tau)\,d\tau = 1_+(t) * i(t)\,.$$

Substituting this equation in the preceding series we obtain a relation between current and voltage

$$v = \sum_{k=1}^{\infty} \varsigma_k (1_+ * i)^k\,.$$

The first few voltage components expressed as a function of the current components are given by

$$\begin{aligned}
v_1 &= \varsigma_1 1_+ * i_1\\
v_2 &= \varsigma_1 1_+ * i_2 + \varsigma_2 (1_+ * i_1)^{\otimes 2}\\
v_3 &= \varsigma_1 1_+ * i_3 + \varsigma_2 2\left[(1_+ * i_1) \otimes (1_+ * i_2)\right]_{\text{sym}} + \varsigma_3 (1_+ * i_1)^{\otimes 3}\\
&\ldots
\end{aligned}$$

As for the previous cases we see that each voltage component v_k is composed by a term linear in the current i_k and other ones that only depend on current components of order lower than k

$$v_k = \varsigma_1 1_+ * i_k + \tilde{v}_{C,k}(i_1, \ldots, i_{k-1})\,.$$

This expression can be represented in an electric network by the equivalent circuit shown in Fig. 11.8a.

If a capacitor is voltage controlled as well as charge controlled, then one can use either representation and one can be converted in the other one. The following

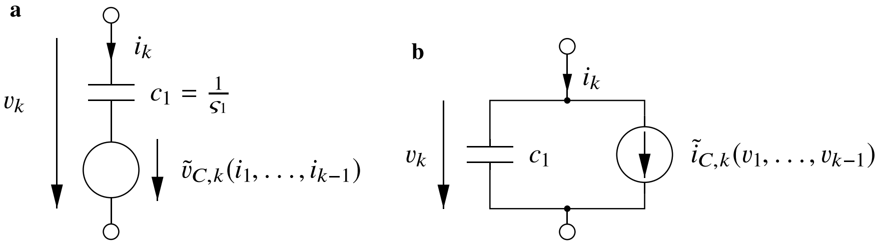

a
i_k
v_k
$c_1 = \frac{1}{\varsigma_1}$
$\tilde{v}_{C,k}(i_1, \ldots, i_{k-1})$
b
i_k
v_k
c_1
$\tilde{i}_{C,k}(v_1, \ldots, v_{k-1})$

Fig. 11.8 **a** Weakly nonlinear capacitor current-controlled equivalent model **b** Weakly nonlinear capacitor charge-controlled equivalent model equations give the first three coefficients of the charge controlled representation expressed in terms of the ones of the voltage controlled ones

$$\begin{aligned}\varsigma_1 &= \frac{1}{c_1}\\ \varsigma_2 &= -\frac{c_2}{2c_1^3}\\ \varsigma_3 &= \frac{c_2^2}{2c_1^5} - \frac{c_3}{3c_1^4}\,.\end{aligned}$$

They were obtained by the same inversion procedure that we used to relate the two representations of nonlinear resistors.

11.2.3 Nonlinear Inductors

A nonlinear inductor is a two-terminal device whose current i_L and magnetic flux ϕ_L are related by the ϕ-i characteristic $f_L(\phi_L, i_L) = 0$. In the following we are going to represent a nonlinear inductor by the symbol shown in Fig. 11.6. A current controlled inductor is an inductor whose flux is a function of the current $\phi_L = \tilde{f}_L(i_L)$. The voltage across the terminals of an inductor is the time derivative of the flux. Thus, if the current is a differentiable function of time, the voltage is

$$v_L = \frac{\mathrm{d}\tilde{f}_L(i_L)}{\mathrm{d}i_L}\frac{\mathrm{d}i_L}{\mathrm{d}t}\,.$$

The slope of the ϕ-i characteristic is called the *small signal* (or *incremental*) *Inductance* of the inductor

$$L(i_L) := \frac{\mathrm{d}\tilde{f}_L(i_L)}{\mathrm{d}i_L} \tag{11.23}$$

that we assume, around the quiescent operating point (Φ_L, I_L), to be expandable in a power series

$$l(i) := L(i + I_L) = \sum_{k=0}^{\infty} l_{k+1} i^k\,, \qquad i = i_L - I_L\,.$$

It is apparent that inductors and capacitors are "dual" of each other, with the roles of current and voltage exchanged. We can therefore adapt previous results and define the voltage and current relationship of a current controlled weakly nonlinear inductor by

$$v = v_C = \sum_{k=1}^{\infty} \frac{l_k}{k} D i^k\,. \tag{11.24}$$

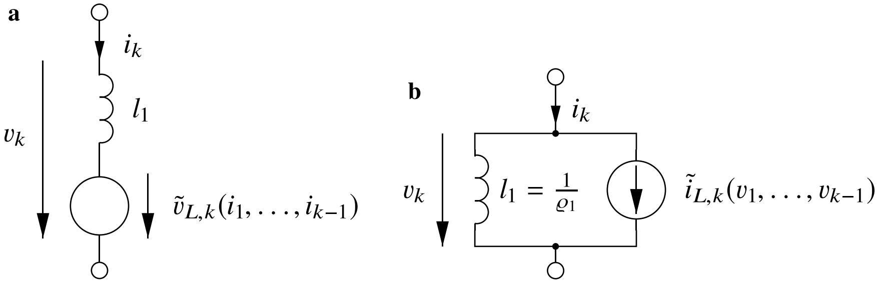

a
i_k
l_1
v_k
$\tilde{v}_{L,k}(i_1,\dots,i_{k-1})$
b
i_k
v_k
$l_1 = \frac{1}{\varrho_1}$
$\tilde{i}_{L,k}(v_1,\dots,v_{k-1})$

Fig. 11.9 **a** Weakly nonlinear inductor current-controlled equivalent model **b** Weakly nonlinear inductor flux-controlled equivalent model

The first few components of the voltage expressed in terms of the components of the current are given by

$$\begin{aligned} v_1 &= l_1 D i_1 \\ v_2 &= l_1 D i_2 + \frac{l_2}{2} D i_1^{\otimes 2} \\ v_3 &= l_1 D i_3 + l_2 D \left[i_1 \otimes i_2\right]_{\text{sym}} + \frac{l_3}{3} D i_1^{\otimes 3} \\ &\dots \end{aligned} \tag{11.25}$$

The component k has the form

$$v_k = l_1 D v_k + \tilde{v}_{L,k}(i_1,\dots,i_{k-1})$$

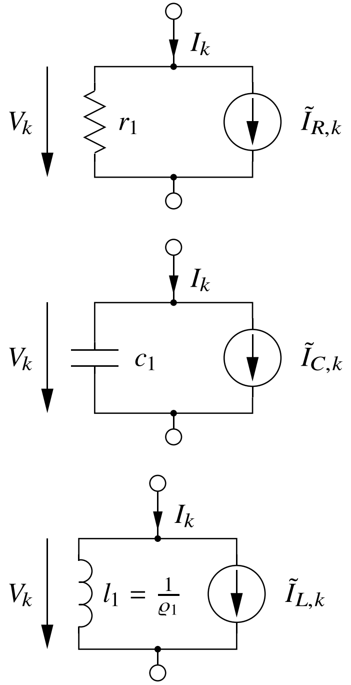
from which we read the equivalent circuit shown in Fig. 11.9a.

Similarly, the flux controlled representation of a weakly nonlinear inductor is

$$i = \sum_{k=1}^{\infty} \varrho_k (1_+ * v)^k ,$$

with the first few voltage components expressed as a function of the current components given by

$$\begin{aligned} i_1 &= \varrho_1 1_+ * v_1 \\ i_2 &= \varrho_1 1_+ * v_2 + \varrho_2 (1_+ * v_1)^{\otimes 2} \\ i_3 &= \varrho_1 1_+ * v_3 + \varrho_2 2 \left[(1_+ * v_1) \otimes (1_+ * v_2)\right]_{\text{sym}} + \varrho_3 (1_+ * v_1)^{\otimes 3} \\ &\dots \end{aligned}$$

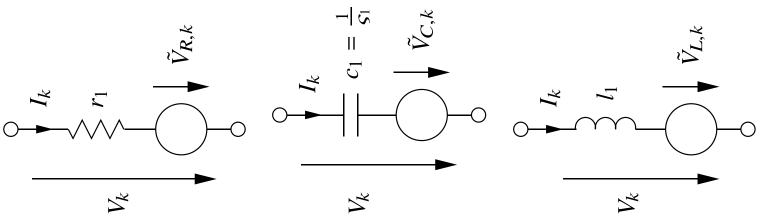


$$\tilde{I}_{R,2}(s_1, s_2) = g_2 V_1(s_1) V_1(s_2)$$
$$\tilde{I}_{R,3}(s_1, s_2, s_3) = 2g_2 \left[V_1(s_1) V_2(s_2, s_3)\right]_{\mathrm{sym}} + g_3 V_1(s_1) V_1(s_2) V_1(s_3)$$

$$\tilde{I}_{C,2}(s_1, s_2) = \frac{c_2}{2}(s_1 + s_2) V_1(s_1) V_1(s_2)$$
$$\tilde{I}_{C,3}(s_1, s_2, s_3) = (s_1 + s_2 + s_3) \left\{ c_2 \left[V_1(s_1) V_2(s_2, s_3)\right]_{\mathrm{sym}} + \frac{c_3}{3} V_1(s_1) V_1(s_2) V_1(s_3) \right\}$$

$$\tilde{I}_{L,2}(s_1, s_2) = \varrho_2 \frac{V_1(s_1)}{s_1} \frac{V_1(s_2)}{s_2}$$
$$\tilde{I}_{L,3}(s_1, s_2, s_3) = 2\varrho_2 \left[\frac{V_1(s_1) V_2(s_2, s_3)}{s_1(s_2 + s_3)} \right]_{\mathrm{sym}} + \varrho_3 \frac{V_1(s_1) V_1(s_2) V_1(s_3)}{s_1 s_2 s_3}$$

Fig. 11.10 Current source representation in the Laplace domain of nonlinearities of weakly nonlinear elements



$$\tilde{V}_{R,2}(s_1, s_2) = r_2 I_1(s_1) I_1(s_2)$$

$$\tilde{V}_{R,3}(s_1, s_2, s_3) = 2r_2 \left[I_1(s_1) I_2(s_2, s_3)\right]_{\text{sym}} + r_3 I_1(s_1) I_1(s_2) I_1(s_3)$$

$$\tilde{V}_{C,2}(s_1, s_2) = \varsigma_2 \frac{I_1(s_1)}{s_1} \frac{I_1(s_2)}{s_2}$$

$$\tilde{V}_{C,3}(s_1, s_2, s_3) = 2\varsigma_2 \left[\frac{I_1(s_1) I_2(s_2, s_3)}{s_1(s_2 + s_3)}\right]_{\text{sym}} + \varsigma_3 \frac{I_1(s_1) I_1(s_2) I_1(s_3)}{s_1 s_2 s_3}$$

$$\tilde{V}_{L,2}(s_1, s_2) = \frac{l_2}{2}(s_1 + s_2) I_1(s_1) I_1(s_2)$$

$$\tilde{V}_{L,3}(s_1, s_2, s_3) = (s_1 + s_2 + s_3) \left\{ l_2 \left[I_1(s_1) I_2(s_2, s_3)\right]_{\text{sym}} + \frac{l_3}{3} I_1(s_1) I_1(s_2) I_1(s_3) \right\}$$

Fig. 11.11 Voltage source representation in the Laplace domain of nonlinearities of weakly nonlinear elements

The component k has the form

$$i_k = \varrho_1 \mathbf{1}_+ * v_k + \tilde{i}_{L,k}(v_1, \ldots, v_{k-1}) \,.$$

which leads to the equivalent circuit shown in Fig. 11.9b.

As for nonlinear capacitors, the nonlinear impulse responses formalism can only be applied to circuits including nonlinear inductors when they are part of networks whose linear approximation is asymptotically stable.

11.3 Nonlinear Multi-port Elements

Weakly nonlinear multi-port and multi-terminal elements can be represented by two-terminal elements and controlled sources. Therefore, in this section we introduce weakly nonlinear controlled sources. With them, we will have at disposal all the necessary circuit elements necessary to model arbitrary weakly nonlinear electronic components.

A *controlled source* is a two terminal element whose voltage v_{CS} or current i_{CS} is controlled by a control voltage v_X or current i_X, a quantity in another part of the electric network of which it is part. There are four types of controlled sources:

- the voltage-controlled voltage source (VCVS), characterised by the equation $v_{CS} = \mu(v_X)$;
- the voltage-controlled current source (VCCS), characterised by the equation $i_{CS} = g_m(v_X)$;
- the current-controlled voltage source (CCVS), characterised by the equation $v_{CS} = r_m(i_X)$, and
- the current-controlled current source (CCCS), characterised by the equation $i_{CS} = \alpha(i_X)$.

As before, we assume that, around a quiescent operating point, the characterising function can be approximated to any desired accuracy by a power series. The incremental quantities can then be represented by elements of $\mathcal{D}'_{\oplus,\text{sym}}$. For example, we assume that a VCCS can be represented by

$$i = \sum_{k=1}^{\infty} g_{mk} v^k , \qquad i = i_{CS} - I_{CS} , \quad v = v_X - V_X .$$

The first three components of the current expressed in terms of the components of the voltage can be derived in the same way as we did for a voltage-controlled weakly-nonlinear resistor and are

$$\begin{aligned} i_1 &= g_{m1} v_1 \\ i_2 &= g_{m1} v_2 + g_{m2} v_1^{\otimes 2} \\ i_3 &= g_{m1} v_3 + 2 g_{m2} \left[v_1 \otimes v_2\right]_{\text{sym}} + g_{m3} v_1^{\otimes 3} \\ &\dots \end{aligned} \tag{11.26}$$

Note that each current component i_k is the sum of a term linear in the kth component of the incremental control voltage v_k and other terms that only depend on components of the voltage v of order lower than k

$$i_k = g_{mk} v_k + \tilde{i}_{CS,k}(v_1, \dots, v_{k-1}) .$$

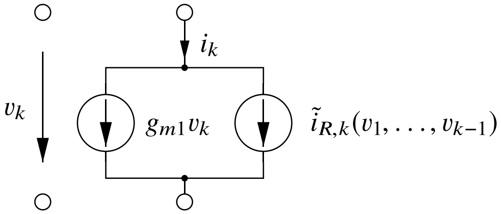


Fig. 11.12 Equivalent circuit of a weakly-nonlinear VCCS

In an electric network a VCCS can thus be represented by a *linear* VCCS and independent current sources only depending on control voltage components of order lower than k. As for the two terminal weakly-nonlinear elements considered in Sect. 11.2, the linear term of a VCCS plays a special role. For this reason the quantity g_{m1} has been given a name: it is called the *transconductance* of the source. A two-port representation of a VCCS is shown in Fig. 11.12.

The situation is entirely analogous for the other types of controlled sources. The coefficient of the linear term of a CCVS r_{m1} is called the *transresistance*, the one of a CCCS α_1 is called the *current transfer ratio* and the one of a VCVS μ_1 the *voltage transfer ratio*.

11.4 Low-Pass Filter with Nonlinear Capacitor

In this section we investigate the low-pass filter (LPF) shown in Fig. 11.13a. When implemented in an integrated circuit technology, a considerable fraction of the circuit area is often occupied by the capacitor. Given that the price of integrated circuits is determined to a large extent by occupied area, to reduce the cost of the circuit, it is desirable to use a capacitor type with a high capacitance per unit area. The highest capacitance per unit area available in CMOS technologies is offered by MOS capacitors which however have a rather nonlinear characteristic. For this reason we investigate the effects introduced in the circuit by the use of a nonlinear capacitor. In particular, we are interested in the upper linearity limit set by the nonlinear capacitor and therefore assume the operational amplifier (OpAmp) to be ideal.

11.4.1 Nonlinear Transfer Functions

Under the assumption of an ideal OpAmp, the circuit of Fig. 11.13a can be represented by the small-signal equivalent circuit shown in Fig. 11.13b. Since MOS capacitors are voltage controlled, we represent the nonlinear capacitor as a voltage controlled device. Then, using Kirchhoff's current law (KCL) the system equation is

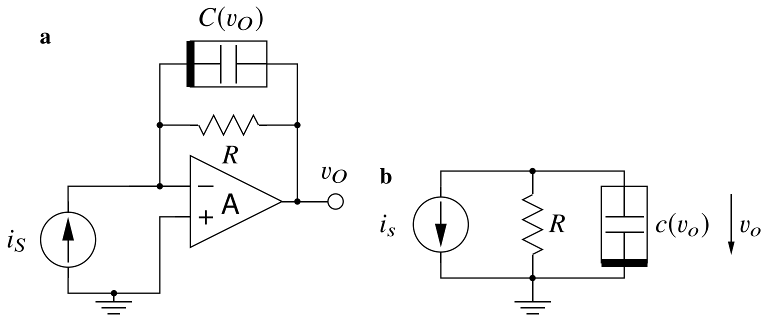


Fig. 11.13 **a** Active *RC* low-pass filter circuit **b** Active *RC* low-pass filter with ideal OpAmp model

$$\frac{v_o}{R} + c_1 D v_o = -i_s - \sum_{k=2}^{\infty} \frac{c_k}{k} D v_o^k \,. \tag{11.27}$$

To highlight that the nonlinear part of the capacitor characteristic act as a source, we have moved that part of the characteristic to the right-hand side of the equation together with the source i_s and collected all linear terms on the left-hand side.

We solve the equation in the Laplace domain using the equivalent circuits that we developed in Sects. 11.2 and 11.3. The first order output voltage component $V_{o.1}$ is obtained by replacing the nonlinear capacitor with an ideal capacitor with a value c_1 corresponding to the value of the nonlinear capacitor at the operating point

$$\frac{V_{o,1}(s_1)}{R} + c_1 s_1 V_{o,1}(s_1) = -I_s(s_1) \,.$$

Using a Dirac impulse as input signal, the first order transfer function is found to be

$$H_1(s_1) = .V_{o.1}\Big|_{I_s(s_1)=1} = \frac{-R}{1 + \frac{s_1}{\omega_{3\text{dB}}}}$$

with

$$\omega_{3\text{dB}} := \frac{1}{Rc_1}$$

the 3 dB cut-off frequency of the filter.

Having found the first order output component of the voltage, we can calculate the equivalent source representing the second order nonlinearity of the capacitor $\tilde{I}_{c,2}(s_1, s_2)$ (see Fig. 11.10). With a Dirac pulse as the first order input we find

$$\tilde{I}_{c,2}(s_1, s_2) = \frac{c_2}{2}(s_1 + s_2) H_1(s_1) H_1(s_2) \,.$$

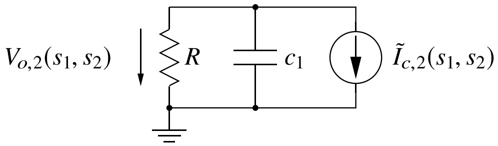


Fig. 11.14 LPF 2nd order equivalent circuit

The second order transfer function is found with the help of the second order equivalent circuit. It is obtained from the first order one by removing the current source $I_s(s_1)$, which is of first order, and by inserting the source $\tilde{I}_{c,2}(s_1, s_2)$ representing the second order nonlinearity of the capacitor. The second order equivalent circuit is shown in Fig. 11.14. Note how the current generated by second (and higher) order nonlinearity is injected into the input node of the filter. For this reason, as discussed in Sect. 10.2, feedback is unable to suppress it. Since the variables in this network are of second order, we have to use the definition of the derivative for second order distributions (Eq. (9.14)). The second order transfer function is thus found to be

$$\begin{aligned} H_2(s_1, s_2) &= \frac{-R}{1 + (s_1 + s_2)Rc_1} \tilde{I}_{c,2}(s_1, s_2) \\ &= \frac{c_2}{2}(s_1 + s_2) H_1(s_1 + s_2) H_1(s_1) H_1(s_2) \,. \end{aligned}$$

With the first two components of the output voltage we can compute the equivalent source representing the capacitor nonlinearity of third order (see Fig. 11.10)

$$\begin{aligned} &\tilde{I}_{c,3}(s_1, s_2, s_3) \\ &\quad = (s_1 + s_2 + s_3)\left\{c_2 \left[H_1(s_1) H_2(s_2, s_3)\right]_{\text{sym}} + \frac{c_3}{3} H_1(s_1) H_1(s_2) H_1(s_3)\right\} \end{aligned}$$

and with it the equivalent circuit of third order. Using the definition of the derivative for third order distributions, the third order transfer function is found to be

$$\begin{aligned} H_3(s_1, s_2, s_3) &= \frac{-R}{1 + (s_1 + s_2 + s_3)Rc_1} \tilde{I}_{c,3}(s_1, s_2, s_3) \\ &= (s_1 + s_2 + s_3) H_1(s_1 + s_2 + s_3) \\ &\quad \cdot \Big\{ \frac{c_2^2}{2} \left[H_1(s_1) H_1(s_2) H_1(s_3)(s_2 + s_3) H_1(s_2 + s_3)\right]_{\text{sym}} \\ &\quad + \frac{c_3}{3} H_1(s_1) H_1(s_2) H_1(s_3) \Big\} \,. \end{aligned}$$

11.4.2 Second Order Intermodulation

Having found the first three transfer functions of the filter, we can evaluate the impact of the nonlinearities in concrete situations. As a first situation, suppose that there is a strong modulated signal in the stop-band of the filter. If the even order nonlinearities generate strong IM products masking the wanted signal in the pass-band, then the filter is of little use. To have a first indication of the strength of this effect, we only consider the nonlinearity of second order and calculate the IP2. We model the modulated signal in the stop-band with two tones at ω_1 and ω_2. We further assume $\omega_2 > \omega_1$ and

$$\Delta\omega := \omega_2 - \omega_1 < \omega_{3\text{dB}}$$

so that one of the IM2 products falls in the pass-band of the filter. The IM2 of interest is characterized by the frequency mix $m = (0, 1, 0, 1)$ and thus by the frequency response

$$\begin{aligned} H_2(-\jmath\omega_1, \jmath\omega_2) &= \frac{c_2}{2}\jmath\Delta\omega H_1(\jmath\Delta\omega)H_1(-\jmath\omega_1)H_1(\jmath\omega_2) \\ &\approx -\frac{c_2}{2}\jmath\Delta\omega R\frac{R\,\omega_{3\text{dB}}}{-\jmath\omega_1}\frac{R\,\omega_{3\text{dB}}}{\jmath\omega_2} \approx -\frac{c_2}{2}\jmath\Delta\omega\frac{R^3\omega_{3\text{dB}}^2}{\omega_1^2}\,. \end{aligned}$$

With it the IIP2 and OIP2 are obtained from (11.10)

$$I_{\text{IIP2}} \approx \left|\frac{2}{c_2\Delta\omega R^2}\right|\left(\frac{\omega_1}{\omega_{3\text{dB}}}\right)^2 \tag{11.28}$$

$$V_{\text{OIP2}} \approx \left|\frac{2}{c_2\Delta\omega R}\right|\left(\frac{\omega_1}{\omega_{3\text{dB}}}\right)^2. \tag{11.29}$$

We have denoted the two intercept points by I_{IIP2} and V_{OIP2} to make it clear that the first characterizes the magnitude of the input current while the latter the magnitude of the output voltage.

These expressions reveal that the more the blocker is in the stop band, the lower the IM2. This makes intuitive sense as the voltage generated across the nonlinear capacitor by the interfering signal is the smaller, the lower the capacitor impedance. Since we have assumed a voltage-controlled capacitor, a small voltage will produce small intermodulation products. The above expressions also reveal that the IM2 is not homogeneous across the pass-band, but it is stronger when $\Delta\omega$ approaches the filter 3 dB cut-off frequency of the filter.

The IP2 can also be expressed in a slightly different form. If we replace one occurrence of $\omega_{3\text{dB}}$ by $1/(c_1 R)$ in the above expression we obtain

$$V_{\text{OIP2}} \approx 2\left|\frac{c_1}{c_2}\right|\frac{\omega_1^2}{\Delta\omega\,\omega_{3\text{dB}}}\,. \tag{11.30}$$

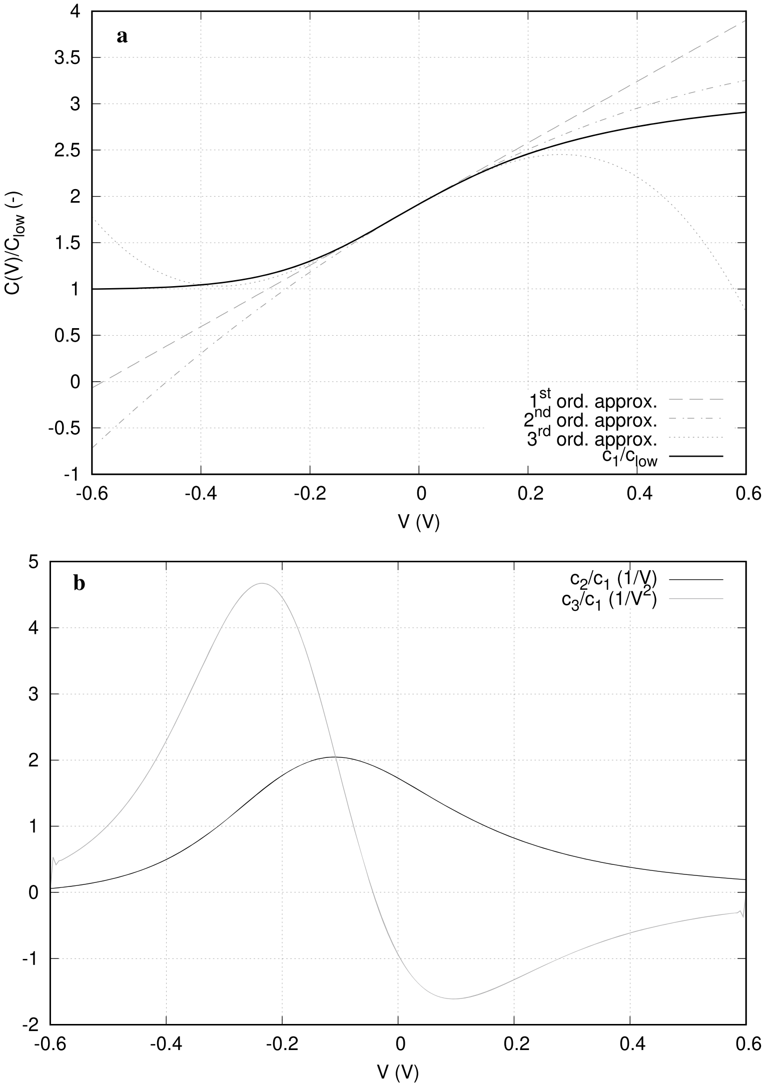


Fig. 11.15 **a** Typical characteristic of an n-type accumulation-mode MOS varactor with a channel length of 0.2 μm in a 40 nm CMOS technology **b** Small-signal model coefficients of an n-type accumulation-mode MOS varactor with a channel length of 0.2 μm in a 40 nm CMOS technology

This form highlights the value of the OIP2 as a function of the ratio of the linear capacitor coefficient to the coefficient of the second order nonlinearity.

Figure 11.15a shows the typical characteristic of an n-type accumulation-mode MOS varactor [26] with a channel length of 0.2 μm in a 40 nm CMOS technology. Figure 11.15b shows the small-signal model coefficients normalized to the linear

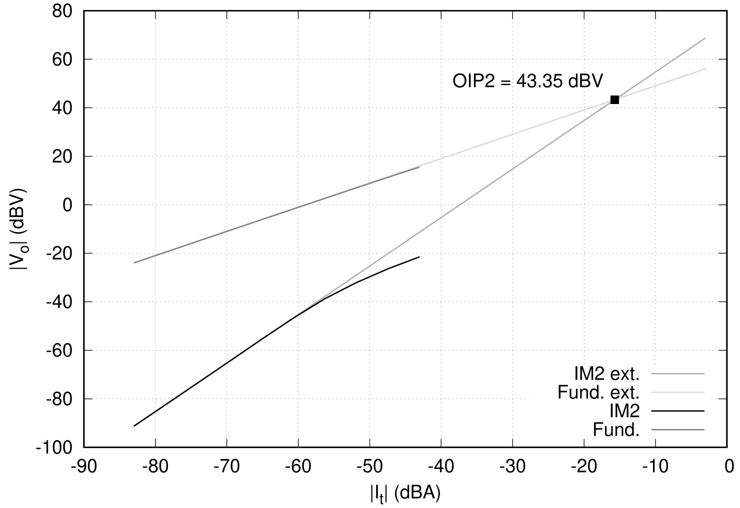


Fig. 11.16 Simulated IM2 of the LPF with the capacitor having the characteristic shown in Fig. 11.15a and driven by two tones of equal magnitude at 48.75 and 51.25 MHz respectively

capacitance c_1 as a function of the voltage across the capacitor. If we use such a capacitor biased at 0 V to implement a LPF with a cut-off frequency of 5 MHz to suppress a signal at 50 MHz modeled as two tones at 48.75 and 51.25 MHz respectively, (11.30) predicts that the filter will have an OIP2 of

$$\mathrm{OIP2} \approx 44 \text{ dBV} .$$

For comparison we simulated the LPF IP2 numerically. To obtain a cut-off frequency of 5 MHz we used a resistor of 1 kΩ and a nominal capacitance of 31.58 pF. The results of the simulation are shown in Fig. 11.16. The value of the IP2 agrees very well with the predicted value. The IM2 starts to deviate from the ideal slope of 2 at a level of the input tones of ca. –55 dBA. This means that at that level the contribution of higher order nonlinearities to the IM2 become important. A -55 dBA tone at 50 MHz passing through a linear LPF with a transfer function equal to $H_1(\jmath\omega)$ and the above component values produces an output tone with a magnitude of approximately

$$\sqrt{2}\, 10^{-55/20}\, R\, \frac{\omega_{3\mathrm{dB}}}{\omega} \approx 251 \text{ mV} .$$

The capacitor characteristic in Fig. 11.15a shows that a linear c-v approximation is only reasonably accurate up to this value. We thus see that a rough estimate of the range of validity of the approximation can be obtained by overlapping the approximation with the real characteristic. At larger positive and negative voltage levels the capacitor characteristic flattens out, and we can speculate that this is the reason for the slower increase of the IM2 at large signal levels.

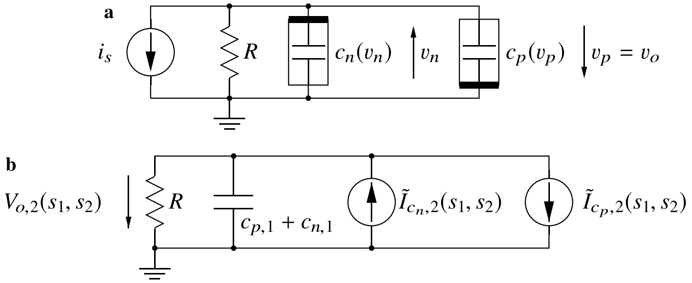


Fig. 11.17 **a** LPF with back-to-back capacitors ideal OpAmp model **b** LPF with back-to-back capacitors 2nd order equivalent circuit

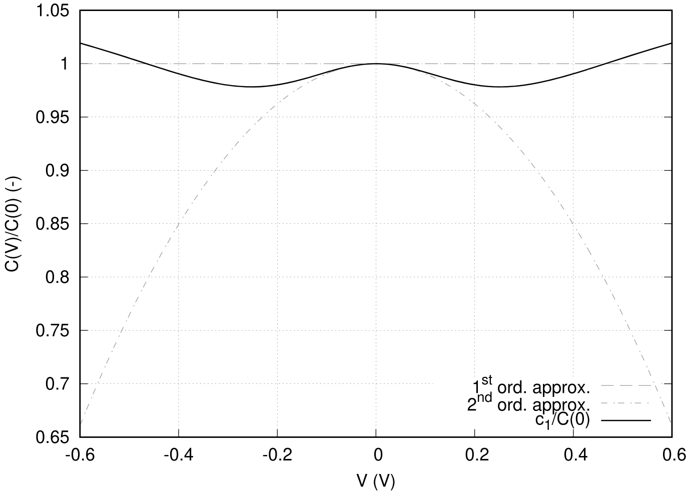


Fig. 11.18 Characteristic of two back-to-back n-type accumulation-mode MOS varactor each with the characteristic shown in Fig. 11.15a

For many applications, such as in communication receivers, this IP2 is insufficient. One way to improve it is by using two equal nonlinear capacitors connected back-to-back as shown in Fig. 11.17a, each providing half of the required capacitance. In this way, when v_o increases, the capacitance of one capacitor increases, while the one of the other capacitor decreases.

One way to analyze this circuit is to consider the combination of the two capacitors as a single nonlinear capacitor with the effective characteristic shown in Fig. 11.18. Figure 11.19 shows that, with identical devices, c_2 is identically zero. Hence, the IM2 is completely suppressed.

Another way to analyze the circuit is to consider each capacitor individually. If the linear transfer function has to remain the same as the one of the original circuit with a single capacitor, then we must have $c_{p,1} + c_{n,1} = c_1$. The second order network is

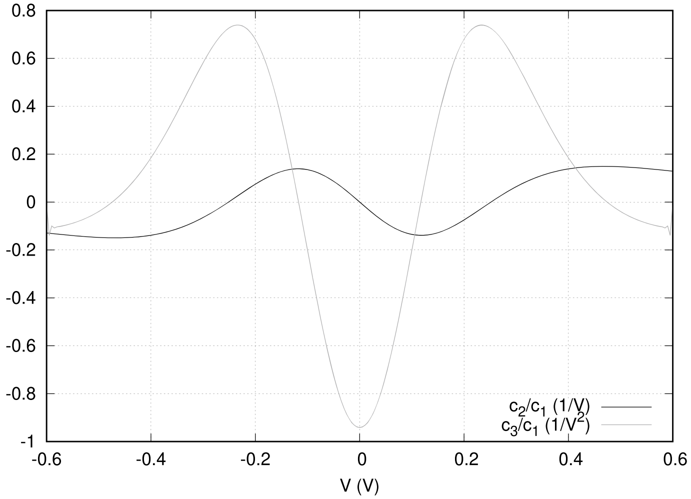


Fig. 11.19 Small-signal model coefficients of two back-to-back n-type accumulation-mode MOS varactor

therefore composed by the same linear components as before, but now it includes two sources, each representing the second-order nonlinearity of one of the two capacitors (see Fig. 11.17b). The one of c_p has the same reference direction as the one of the original circuit and has a value of

$$\tilde{I}_{c_p,2}(s_1, s_2) = \frac{c_{p,2}}{2}(s_1 + s_2)V_{o,1}(s_1)V_{o,1}(s_2)\,.$$

The one of c_n has the opposite reference direction and a value of

$$\begin{aligned}\tilde{I}_{c_n,2}(s_1, s_2) &= \frac{c_{n,2}}{2}(s_1 + s_2)\big(-V_{o,1}(s_1)\big)\big(-V_{o,1}(s_2)\big)\\ &= \frac{c_{n,2}}{2}(s_1 + s_2)V_{o,1}(s_1)V_{o,1}(s_2)\,.\end{aligned}$$

As the two negative signs coming from $v_n = -v_o$ cancel, the two currents flow in opposite directions and, if $c_{n,2} = c_{p,2}$ they cancel each other.

Note that this cancelling effect of even order responses is quite general. Given an arbitrary even order frequency response $\hat{h}_{k,m}$, the response of (even) order k to N input tones with phasors $A_1, \ldots, A_n$ is

$$y^c_{k,m}(t) = \Re\left\{\frac{1}{2^{k-1}}\frac{k!}{m!}A_{-N}^{m_{-N}} \cdots A_{-1}^{m_{-1}} A_1^{m_1} \cdots A_N^{m_N}\,\hat{h}_{k,m}\,\mathrm{e}^{\jmath\omega_m t}\right\}.$$

If the sign of all input tones is reversed, every phasor will be multiplied by $\mathrm{e}^{\jmath\pi}$. As k is assumed to be even and $|m| = k$ these factors will multiply to 1

$$\mathrm{e}^{\jmath\pi k} = 1; \qquad k \text{ even.}$$

For this reason the response remains unchanged, but with opposite reference direction. Therefore, *all even order harmonics and intermodulation products will be suppressed.*

In reality there are two limitations to the amount of canceling that is practically achievable. The first one is due to the fact that small unavoidable manufacturing imperfections make nominally identical devices slightly different. This effect is called *mismatch*. For this reason the coefficients of c_p will be slightly different from the one of c_n. Let's represent the small variations due to mismatch in the following way

$$\begin{aligned} c_{p,1} &= c_{\text{nom},1} + \Delta c_{p,1} & \qquad c_{n,1} &= c_{\text{nom},1} + \Delta c_{n,1} \\ c_{p,2} &= c_{\text{nom},2} + \Delta c_{p,2} & \qquad c_{n,2} &= c_{\text{nom},2} + \Delta c_{n,2} \end{aligned}$$

with

$$\begin{aligned} c_{\text{nom},1} &= \frac{c_1}{2} & \qquad \Delta c_{p,1}, \Delta c_{n,1} &\ll c_{\text{nom},1} \\ c_{\text{nom},2} &= \frac{c_2}{2} & \qquad \Delta c_{p,2}, \Delta c_{n,2} &\ll c_{\text{nom},2}\,. \end{aligned}$$

Then, the two current sources $\tilde{I}_{c_p,2}(s_1, s_2)$ and $\tilde{I}_{c_n,2}(s_1, s_2)$ can be represented by a single source with the same reference direction of the former and a value of

$$\tilde{I}_{c_p,2}(s_1, s_2) - \tilde{I}_{c_n,2}(s_1, s_2) = \frac{\Delta c_{p,2} - \Delta c_{n,2}}{2}(s_1 + s_2) V_{o,1}(s_1) V_{o,1}(s_2)\,.$$

The resulting network is similar to the one of the original circuit, the only difference being that the coefficient c_2 is replaced by $\Delta c_{p,2} - \Delta c_{n,2}$. The OIP2 is therefore

$$\begin{aligned} V_{\text{OIP2,B2B}} &\approx 2 \left| \frac{c_1 + \Delta c_{p,1} + \Delta c_{p,1}}{\Delta c_{p,2} - \Delta c_{n,2}} \right| \frac{\omega_1^2}{\Delta\omega\, \omega_{3\text{dB}}} \\ &\approx 2 \left| \frac{c_2}{\Delta c_{p,2} - \Delta c_{n,2}} \right| \left| \frac{c_1}{c_2} \right| \frac{\omega_1^2}{\Delta\omega\, \omega_{3\text{dB}}} \\ &= \left| \frac{c_2}{\Delta c_{p,2} - \Delta c_{n,2}} \right| V_{\text{OIP2}}\,. \end{aligned} \tag{11.31}$$

Compared to the original circuit the IP2 has been improved by the mismatch limited factor

$$\left| \frac{c_2}{\Delta c_{p,2} - \Delta c_{n,2}} \right|.$$

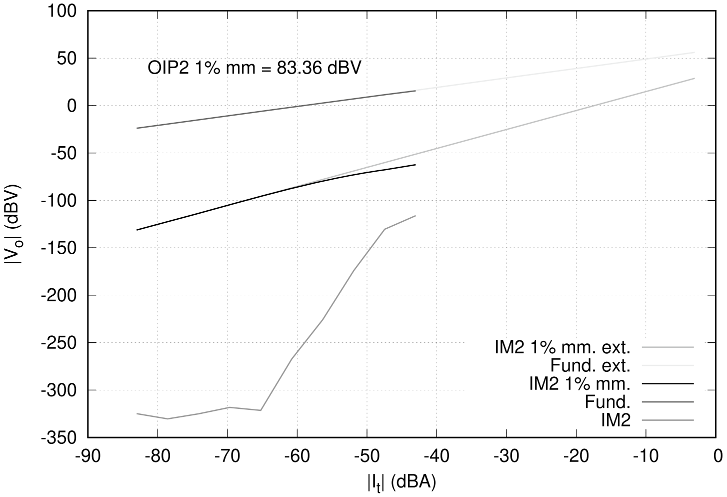


Fig. 11.20 Simulated IM2 of the LPF with the capacitor having the characteristic shown in Fig. 11.15a and driven by two tones of equal magnitude at 48.75 and 51.25 MHz respectively

Figure 11.20 shows the results of simulations with two identical nonlinear capacitors, and for the case where $\Delta c_{p,2}/c_{\mathrm{nom},2} = -\Delta c_{n,2}/c_{\mathrm{nom},2} = 0.01$. In the latter case we observe the expected improvement of

$$20\log\Big(\frac{c_2}{0.02c_2/2}\Big) = 20\log\Big(\frac{1}{0.01}\Big) = 40 \text{ dB} .$$

In the former, at input signal levels up to –65 dBA the value of the IM2 is limited by numerical noise. At larger input signal levels the simulation result is the product of the limited accuracy of the used numerical algorithm.

The second practical limitation is constituted by the fact that the terminals of real components are often coupled to other nodes of the circuit. This coupling can be modeled with parasitic components connected to the terminals. The parasitic components of the positive terminal are often different from the ones of the negative terminal. In addition, parasitic components are often nonlinear.

We conclude this section by noting that if the two tones are in the pass-band of the filter the OIP2 is

$$V_{\mathrm{OIP2}} \approx \left|\frac{2}{c_2 \Delta\omega R}\right| = 2\left|\frac{c_1}{c_2}\right| \frac{\omega_{3\mathrm{dB}}}{\Delta\omega} . \tag{11.32}$$

11.4.3 Third Order Intermodulation

In this section we investigate the situation where there are two interfering signals, one at ω_1 and a second one close to twice this frequency $\omega_2 = 2\omega_1 - \Delta\omega$ so that the lower side-band IM3 falls in the pass-band of the filter

$$2\omega_1 - \omega_2 = \Delta\omega < \omega_{3\text{dB}} \qquad \omega_1, \omega_2 > \omega_{3\text{dB}} > 0\,.$$

To characterize this situation we compute the IP3.

The IM3 of interest is obtained from the third order transfer function

$$\begin{aligned}H_3(s_1, s_2, s_3) = {} & (s_1 + s_2 + s_3) H_1(s_1 + s_2 + s_3) \\ & \cdot \left\{ \frac{c_2^2}{2} \left[H_1(s_1) H_1(s_2) H_1(s_3)(s_2 + s_3) H_1(s_2 + s_3) \right]_{\text{sym}} \right. \\ & \left. + \frac{c_3}{3} H_1(s_1) H_1(s_2) H_1(s_3) \right\}.\end{aligned}$$

evaluated at the frequency mix $m = (1, 0, 2, 0)$. Setting $s_1 = \jmath\omega_1, s_2 = \jmath\omega_1$ and $s_3 = -\jmath\omega_2$ the term enclosed in the symmetrization operator becomes

$$\begin{aligned}& H_1(\jmath\omega_1) H_1(\jmath\omega_1) H_1(-\jmath\omega_2) \\ & \quad \cdot \frac{1}{6}\big[2\jmath(2\omega_1) H_1(\jmath 2\omega_1) + 4\jmath(-\omega_1 + \Delta\omega) H_1(\jmath(-\omega_1 + \Delta\omega))\big]\,.\end{aligned}$$

If we assume $|\omega_1 - \Delta\omega| > \omega_{3\text{dB}}$ and use the approximation

$$\jmath\omega H_1(\jmath\omega) \approx \jmath\omega \frac{-R}{\jmath\omega c_1 R} = \frac{-1}{c_1}$$

we can simplify it to

$$H_1(\jmath\omega_1) H_1(\jmath\omega_1) H_1(-\jmath\omega_2) \frac{-1}{c_1}\,.$$

Using these results we obtain

$$\begin{aligned}& H_3(\jmath\omega_1, \jmath\omega_1, -\jmath\omega_2) \\ & \qquad \approx \jmath\Delta\omega H_1(\jmath\Delta\omega) H_1(\jmath\omega_1) H_1(\jmath\omega_1) H_1(-\jmath\omega_2)\Big[-\frac{c_2^2}{2c_1} + \frac{c_3}{3}\Big] \\ & \qquad \approx \jmath\Delta\omega(-R)\Big(\frac{-1}{\jmath\omega_1 c_1}\Big)^2 \Big(\frac{-1}{-2\jmath\omega_1 c_1}\Big)\Big[-\frac{c_2^2}{2c_1} + \frac{c_3}{3}\Big] \\ & \qquad = \frac{\Delta\omega R}{(\omega_1 c_1)^3}\Big[\frac{c_3}{6} - \frac{c_2^2}{4c_1}\Big]\,.\end{aligned}$$

The IIP3 and OIP3 are obtained by inserting this result in (11.11)

$$I_{\mathrm{IIP3}} \approx \sqrt{\frac{4}{3}\left|\frac{(\omega_1 c_1)^3}{\Delta\omega\left[\frac{c_3}{6} - \frac{c_2^2}{4c_1}\right]}\right|} \tag{11.33}$$

$$V_{\mathrm{OIP3}} \approx \sqrt{\frac{4}{3}\frac{\omega_1^3}{\Delta\omega\,\omega_{3\mathrm{dB}}^2}\left|\frac{1}{\frac{c_3}{6c_1} - \frac{1}{4}\left(\frac{c_2}{c_1}\right)^2}\right|}\,. \tag{11.34}$$

These expressions reveal that the IP3 depends not only on the third order coefficient c_3, but also from the second order one c_2. The reason is the fact that second order intermodulation products are fed back to the input of the nonlinear component, where, in combination with the fundamental tones, they pass again through the second order nonlinearity. This is the effect that was discussed in Sect. 10.2 with the help of the signal-flow graph of Fig. 10.4 and the reason for c_2 being squared. The expression for the OIP3 highlights the fact that it is the ratio of the coefficients c_2 and c_3 to the linear capacitance c_1 that matters. The expressions also reveal that the IM3 generated by second order and third order nonlinearities have either the same or opposite phase and that, if

$$\frac{c_3}{c_1} = \frac{3}{2}\left(\frac{c_2}{c_1}\right)^2, \tag{11.35}$$

the two cancel each other.

In the previous section we discussed the fact that using equal nonlinear capacitors connected back-to-back eliminates even order components from the response of the system. This is not the case for odd order nonlinearities. To see this, we can draw the third order equivalent network of the filter with back-to-back capacitors. The equivalent sources representing the third order nonlinearities of c_p and c_n are

$$\tilde{I}_{c_p,3}(s_1, s_2, s_3) = \frac{c_{p,3}}{3}(s_1 + s_2 + s_3)V_{o,1}(s_1)V_{o,1}(s_2)V_{o,1}(s_3)$$

and

$$\tilde{I}_{c_n,3}(s_1, s_2, s_3) = -\frac{c_{n,3}}{3}(s_1 + s_2 + s_3)V_{o,1}(s_1)V_{o,1}(s_2)V_{o,1}(s_3)$$

respectively, where we have considered that $V_{o,2}(s_2, s_3)$ is zero. For equal capacitors $c_{p,3} = c_{n,3} = c_3/2$, therefore, having the sources opposite reference directions, they combine to form a single source equivalent to the one of a single nonlinear capacitor with $c_2 = 0$.

As an example, we consider again a filter with a cur-off frequency of 5 MHz implemented with a nonlinear MOS capacitor having the characteristic shown in Fig. 11.15a and biased at 0 V. At this bias point the ratios c_2/c_1 and c_3/c_1 are 1.73 and –0.94 respectively. If the filter is driven by a tone at 15 MHz and a second one at 27.5 MHz (11.34) predicts an OIP3 of

$$\text{OPI3} \approx 16.0\ \text{dBV}\,.$$

Note that in this example it is the second order nonlinearity that dominates the IM3 as

$$\left|\frac{c_3}{6c_1}\right| \approx 0.16 < \frac{1}{4}\left(\frac{c_2}{c_1}\right)^2 \approx 0.75\,.$$

Thus, using back-to-back capacitors improves the OIP3 up to

$$\text{OIP3}_{\text{B2B}} \approx 23.6\ \text{dBV}\,.$$

For comparison, we simulated the filter with the full nonlinear capacitor characteristic of Fig. 11.15a. The obtained IM3 as a function of the input tones magnitude is shown in Fig. 11.21. The figure also shows the IM3 obtained using back-to-back capacitors. In both cases the obtained IP3 is in good agreement with the above calculations. The IM3 starts to depart from a straight line with a slope of three at a level of the input tones of ca. –67 dBA. This corresponds to an output fundamental magnitude of ca. 0.2 V for the tone at ω_1 and is close to the level at which the polynomial approximation starts to deviate significantly from the real characteristic of the capacitor.

Further, we verified the occurrence of canceling between the IM3 produced by the third order nonlinearity with the one produced by second order. Figure 11.22 shows the magnitude of the IM3 as a function of the bias voltage of the capacitor. The curve shows a clear notch at a bias voltage of ca. –0.19 V, the bias voltage at

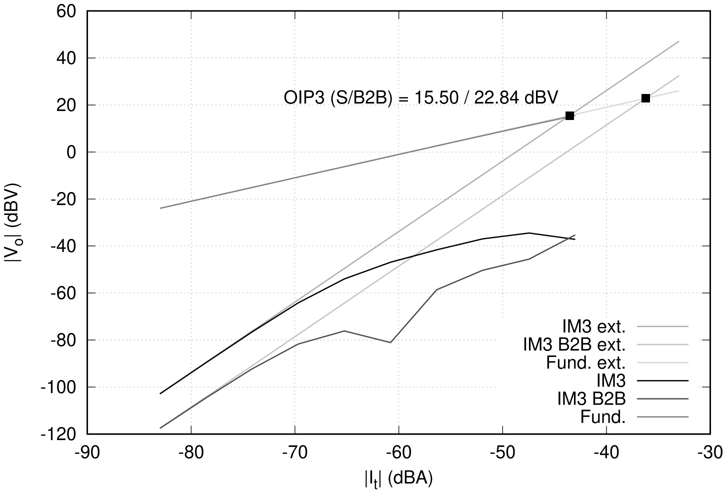


Fig. 11.21 Simulated IP3 of the LPF with the capacitor having the characteristic shown in Fig. 11.15a and with two equal back-to-back (B2B) capacitors. The two input tones were of equal magnitude at 15 and 27.5 MHz respectively

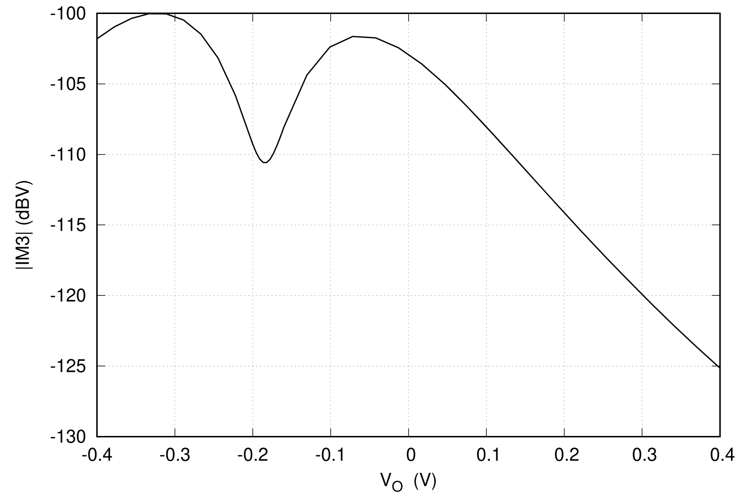


Fig. 11.22 Simulated IM3 of the LPF with the capacitor having the characteristic shown in Fig. 11.15a as a function of the capacitor bias voltage V_O. The filter was driven by two equal tones of magnitude 0.1 at 15 and 27.5 MHz respectively

which the coefficient ratios c_2/c_1 and c_3/c_1 satisfy the canceling condition expressed by (11.35). This notch disappears at large signal levels, where contributions to the IM3 from higher order nonlinearities become important. The curve also suggest that, to obtain the best linearity, one should use a large bias voltage bringing the MOS capacitor in strong inversion, where its capacitance becomes almost constant.

Before concluding this section we investigate the case in which the two tones are in the pass-band of the filter. In this case the term in the symmetrization operator in $H_3(\jmath\omega_1, \jmath\omega_1, -\jmath\omega_2)$ is

$$\begin{aligned} &H_1(\jmath\omega_1)H_1(\jmath\omega_1)H_1(-\jmath\omega_2) \\ &\qquad \cdot \frac{1}{6}\big[2\jmath(2\omega_1)H_1(\jmath 2\omega_1) + 4\jmath(-\omega_1 + \Delta\omega)(-R)\big] \\ &= H_1(\jmath\omega_1)H_1(\jmath\omega_1)H_1(-\jmath\omega_2) \cdot \frac{2}{3}\jmath\big[\omega_1 H_1(\jmath 2\omega_1) + (\omega_1 - \Delta\omega)R\big]. \end{aligned}$$

If we further assume that $2\omega_1$ also falls in the pass-band of the filter it simplifies to

$$-H_1(\jmath\omega_1)H_1(\jmath\omega_1)H_1(-\jmath\omega_2) \cdot \frac{2}{3}\jmath\Delta\omega R\,.$$

The third order nonlinear transfer function evaluated at the frequency mix $m = (1, 0, 2, 0)$ is therefore

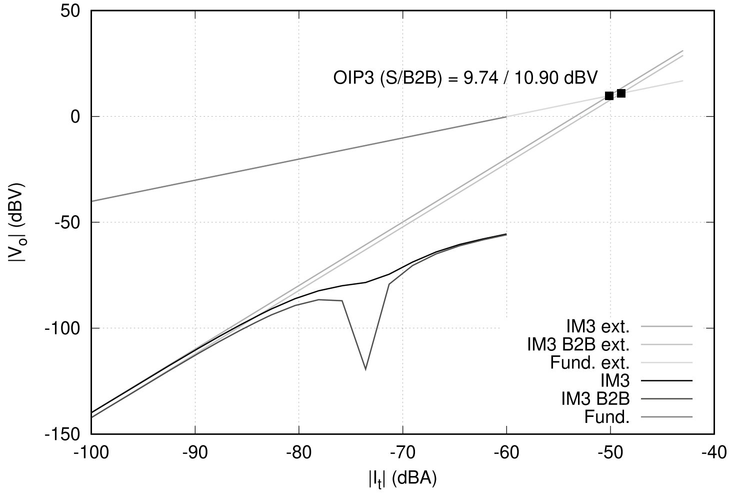


Fig. 11.23 Simulated IP3 of the LPF with the capacitor having the characteristic shown in Fig. 11.15a and with two equal back-to-back (B2B) capacitors. The two input tones were of equal magnitude at 1 and 1.1 MHz respectively

$$
\begin{aligned}
H_3(\jmath\omega_1, \jmath\omega_1, -\jmath\omega_2) &\approx \jmath\Delta\omega R^4\Big[\frac{c_3}{3} - \jmath\frac{c_2^2}{3}\Delta\omega R\Big] \\
&= \jmath\frac{\Delta\omega}{\omega_{3\text{dB}}}\frac{R^3}{3}\Big[\frac{c_3}{c_1} - \jmath\frac{c_2^2}{c_1}\Delta\omega R\Big].
\end{aligned}
$$

With this, the OIP3 is

$$
V_{\text{OIP3,IB}} \approx \sqrt{\frac{4\omega_{3\text{dB}}}{\Delta\omega}\frac{1}{\left|\frac{c_3}{c_1} - \jmath\left(\frac{c_2}{c_1}\right)^2\frac{\Delta\omega}{\omega_{3\text{dB}}}\right|}}\,. \tag{11.36}
$$

The results of a simulation with one tone at 1 MHz and the second one at 1.1 MHz is shown in Fig. 11.23. The results are again in good agreement with the OIP3 estimated with the help of the above equation which gives 10.1 and 10.7 dBV for a single capacitor and for back-to-back capacitors respectively.

11.4.4 Large Signal Effects

In this section we evaluate gain compression and amplitude-modulation to phase-modulation due to the nonlinear capacitor. The onset of both of these effects is governed by the third order transfer function evaluated at the frequency mix $m = (0, 1, 2, 0)$ relative to the linear transfer function at the fundamental

$$\begin{aligned}\frac{H_3(\jmath\omega_1, \jmath\omega_1, -\jmath\omega_1)}{H_1(\jmath\omega_1)} &\approx -\jmath\omega_1 R^3\Big[\frac{c_3}{3} - \jmath\frac{c_2^2}{3}\omega_1 R\Big] \\ &= -\frac{\omega_1}{\omega_{3\text{dB}}}\frac{R^2}{3}\Big[\Big(\frac{c_2}{c_1}\Big)^2\frac{\omega_1}{\omega_{3\text{dB}}} + \jmath\frac{c_3}{c_1}\Big].\end{aligned}$$

where we have assumed $2\omega_1 < \omega_{3\text{dB}}$. The phase of this expression determines the presence of gain compression or expansion and AM2PM.

As a concrete example, we consider again a low-pass filter with a cut-off frequency of 5 MHz, $R = 1\,\text{k}\Omega$, the nonlinear capacitor with the characteristic shown in Fig. 11.15a and driven by a sinusoidal tone at 1 MHz. In this case the term in the square bracket in the above expression, multiplied by minus one, evaluates to $-0.6 + \jmath 0.94$. As the real part is negative we expect some gain compression. However, the imaginary part has a larger magnitude which implies that AM2PM should be somewhat more pronounced. If we use (11.7) to estimate the amplitude of the input tone producing a phase change of 1° we obtain a value of –67.3 dBA which corresponds to an output swing of 0.61 mV. A look at Fig. 11.15a shows that at these levels a second order approximation of the capacitor characteristic is a very poor approximation of the real characteristic. For this reason we can't expect this estimate to be accurate.

A believable prediction can be made for levels where the approximation is good. For example, a phase change of 0.1° is predicted to happen at an input signal level of –77.3 dBA which corresponds to an output swing of 0.193 mV. Similarly, (11.5) predicts a 10 mdB gain compression at an input level of –77.2 dBA. These levels compare quite favorably with the values obtained by a numerical simulation and shown in Fig. 11.24. The simulation shows that these effects remain very small up to the large output swing of 1 V RMS which is close to the reliability limit supported by these devices.

11.5 Class-AC Common-Source Stage

In this section we analyse the common-source stage shown in Fig. 11.25a for use as an RF amplifier. In particular, we are interested in the distortion introduced by the nonlinear $i - v$ characteristic of the transistor and in the influence on distortion of the choice of gate bias voltage V_G. For simplicity in this section we neglect the C_{gd} capacitance. We will consider circuits with some form of local feedback in a later section.

The following is a simple large-signal MOSFET model presented in many textbooks [27, 28]

$$i_D = \begin{cases} 0 & v_{GS} - V_T \le 0 \\ K'\frac{W}{L}(v_{GS} - V_T - \frac{v_{DS}}{2})v_{DS}(1 + \lambda v_{DS}) & 0 < v_{DS} \le (v_{GS} - V_T) \\ \frac{K'}{2}\frac{W}{L}(v_{GS} - V_T)^2(1 + \lambda v_{DS}) & 0 \le (v_{GS} - V_T) \le v_{DS}\,. \end{cases}$$

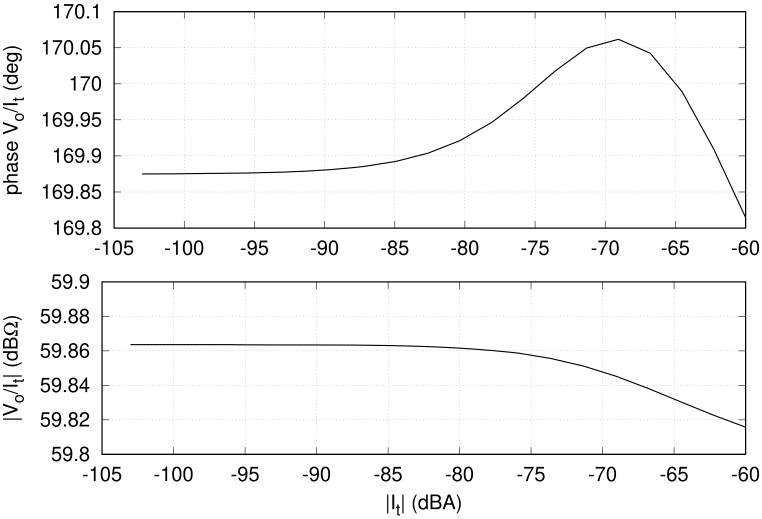


Fig. 11.24 Simulated AM2PM and gain compression of the LPF with a nonlinear capacitor having the characteristic shown in Fig. 11.15a and driven by a tone at 1 MHz

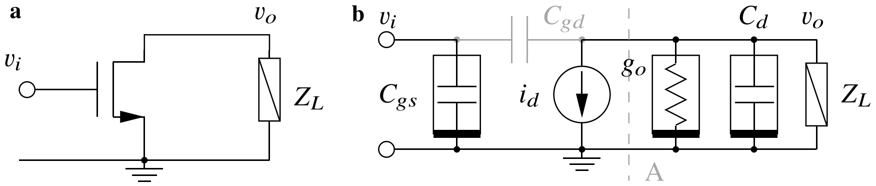


Fig. 11.25 **a** Common-source amplifier AC schematic **b** Common-source amplifier small-signal model

The second equation describes the so-called *linear region* of the characteristic. This is the region where the *overdrive voltage* $v_{GS} - V_T$ is sufficiently large to cause a conductive surface charge *channel* in the active area at the surface between source and drain of the transistor and v_{DS} is sufficiently small that the channel extends all along from the source to the drain terminal of the transistor. In this region the transistor behaves essentially as a nonlinear resistor.

The third equation describes the *saturation region* of the characteristic and is the one of interest for implementing amplifiers and most other analogue circuits. In this region $v_{GS} - V_T$ is sufficiently large to cause the formation of a conductive channel. However, v_{DS} is larger than the *saturation voltage* which means that the channel is present close to the source side of the transistor, but doesn't extend all along to the drain terminal. In this region the current through the transistor i_D is almost independent of the drain voltage and the transistor behaves to a good approximation as a voltage-controlled current source with v_{GS} the control voltage. The parameter λ takes into account the fact that the length of the channel does depend on the drain

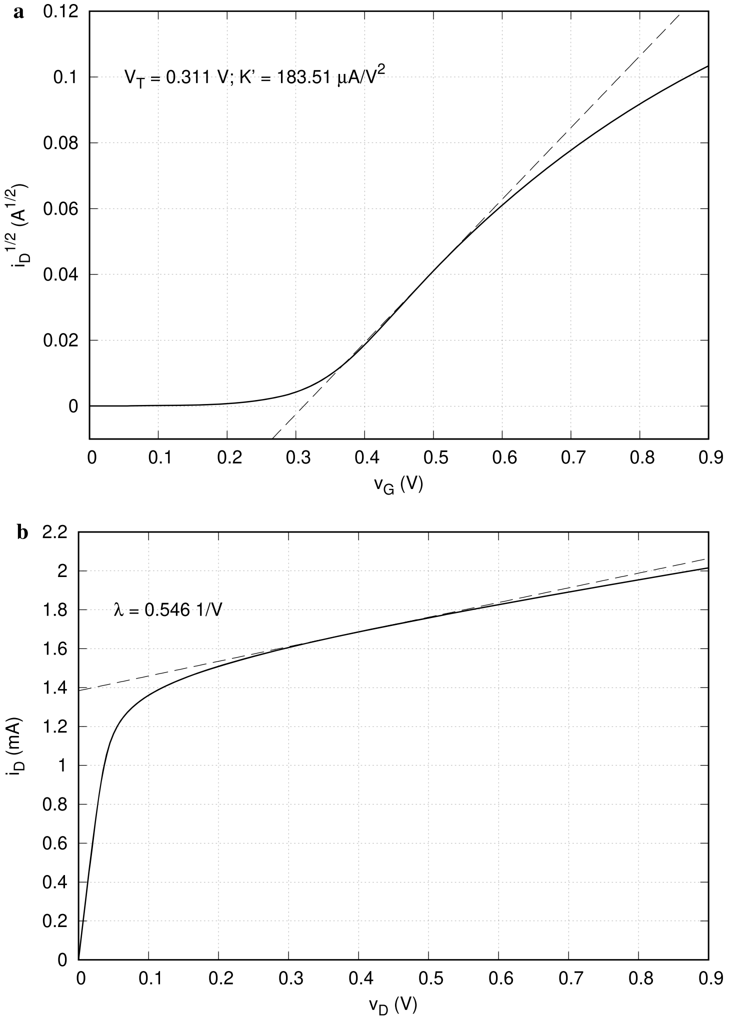


Fig. 11.26 **a** FinFET input side characteristic. $L = 22\,\text{nm}$, $n_{\text{fin}} = 10, n_f = 16, m = 1, W_{\text{eff}} = n_{\text{fin}} n_f m 71\,\text{nm}$ **b** FinFET output side characteristic. $L = 22\,\text{nm}$, $n_{\text{fin}} = 10, n_f = 16, m = 1, W_{\text{eff}} = n_{\text{fin}}\, n_f\, m\, 71\,\text{nm}$

voltage and makes i_D a weak function of the drain voltage [28]. In this simple model the saturation voltage is equal to the overdrive voltage $v_{GS} - V_T$.

The characteristic of real transistors depends on many effects not captured by this simple model. To enable the design of analogue circuits, very accurate transistor models have been developed and made available in circuit simulators. Unfortunately, most of those models depend on several dozens to hundreds of parameters making them unsuitable for analytical estimates. Figure 11.26a shows the characteristic of a

FinFET with a channel length $L = 22\,\text{nm}$ as given by the CMG-BSIM model [29] with parameters from [30]. Figure 11.26a shows $\sqrt{i_D}$ as a function of v_G with the source connected to ground and v_D at a fix potential of 0.4 V. It shows that between 0.35 and 0.65 V the deviation of the characteristic from a straight line as predicted by the above simple model is quite small. Figure 11.26b shows i_D as a function of v_D for a fix gate voltage of 0.5 V. Here as well, the simple model gives a fairly good approximation over an extended range of the characteristic. The pictures show the values of K', V_T and λ obtained by fitting the model to the curves.

Using the simple model the current i_D can be split in two parts

$$i_D = i_{D,a} + i_{D.b}$$

with

$$\begin{aligned} i_{D,a} &= \frac{K'}{2}\frac{W}{L}(v_{GS} - V_T)^2 \\ i_{D,b} &= g_O(i_{D,a})v_{DS} = i_{D,a}\lambda v_{DS}\,. \end{aligned}$$

The current $i_{D,a}$ can be interpreted as the output of an ideal voltage-controlled current-source, while the current $i_{D,b}$ can be interpreted as the current due to a nonlinear load resistance. Since an ideal current source is not affected by its load, from an analysis point of view, it is convenient to analyse the two parts separately and combine the effects with the results of Sect. 10.1. For this reason we lump the components to the right of line A in Fig. 11.25b into a nonlinear load. In this section we focus on $i_{D,a}$. A common nonlinear load will be considered in the next section. Similarly, for analysis purposes, the nonlinear C_{gs} capacitance can be considered part of the driving circuit. In the case of a resistive source we can reuse the results of the previous section with minor modifications. Often however, the distortion introduced by C_{ga} is small compared to the one introduced by the i-v characteristic. In the following we will simply write i_D for $i_{D,a}$.

While the above model can be used to obtain a relatively good approximation of the transconductance g_m of the transistor, it doesn't provide a good estimate of higher order distortion terms. Therefore, to analyse distortion we approximate the transistor characteristic around the operating point by a third order polynomial

$$i_d = g_m v_{gs} + g_2 v_{gs}^2 + g_3 v_{gs}^3$$

and extract the coefficients from simulation. Figure 11.27 compares first, second and third order polynomial approximations to the full characteristic at a bias level of V_G = 0.5 V and V_D = 0.4 V. At this bias level a third order approximation provides a good approximation up to a signal level of about 150 mV. Figure 11.28 shows the three coefficients g_m, g_2 and g_3 as a function of the gate bias voltage V_G simulated using CMG-BSIM models. While the simple model predicts a vanishing third order coefficient g_3 the picture shows that it disappears only at a single gate bias point. For small gate bias voltages the g_3 coefficient is positive, while for large values it's

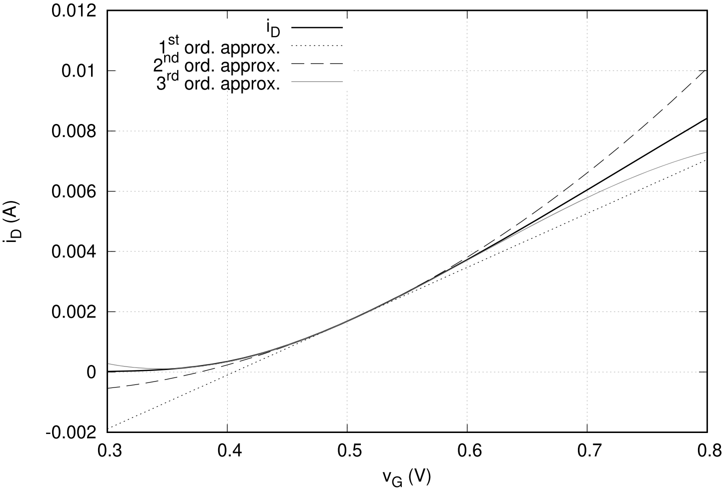


Fig. 11.27 Polynomial approximations of the transistor characteristic around $V_G = 0.5\,\text{V}$. $V_D = 0.4\,\text{V}$, same transistor size as in Fig. 11.26a

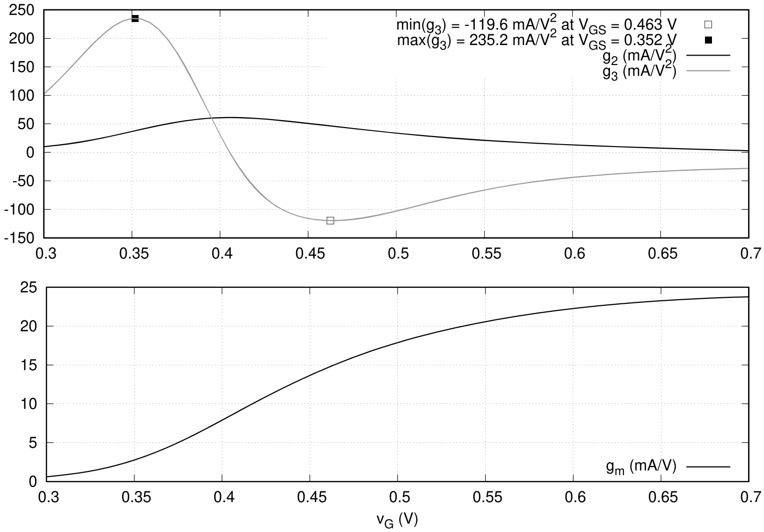


Fig. 11.28 First three coefficients of a polynomial approximation of the transistor characteristic as a function of the gate bias point. $V_D = 0.4\,\text{V}$, same transistor size as in Fig. 11.26a

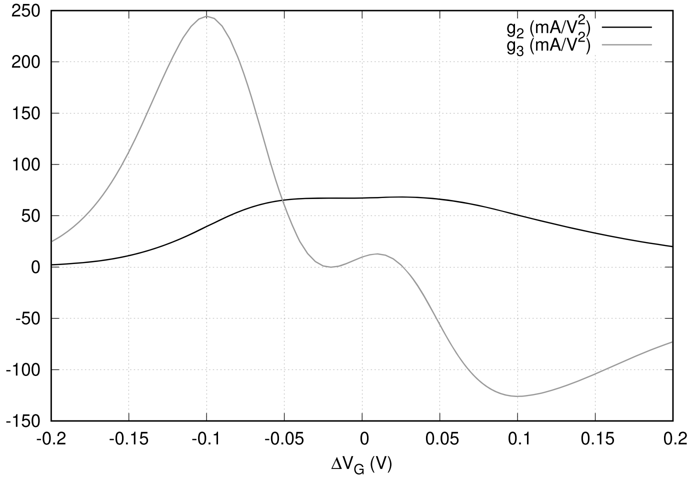


Fig. 11.29 Second and third order coefficients of a polynomial approximation of the Class-AC stage characteristic as a function of the gate bias deviation from the nominal gate bias point ΔV_G for $V_D = 0.4\,\text{V}$

negative. We may try to minimize third order distortion by biasing the transistor at the bias point at which g_3 is zero. However, this strategy doesn't lead to a robust design. In fact mismatch between the transistor and the bias devices introduces a statistical Gaussian bias error with a typical standard deviation of order [31]

$$\sigma_{V_T} \approx \frac{5\text{ mV} \cdot \mu m}{\sqrt{WL}}$$

where L and W are the length respectively the width of the active channel. A more fruitful approach is to use two transistors connected in parallel, but biased at different bias levels. One biased at the minimum of g_3 and the second at its maximum. The relative size of the two transistors is chosen in such a way as to make the sum of the g_3s cancel. In this way the deviation of the bias point of each transistor due to mismatch has a smaller impact on the value of g_3. The resulting effective g_3 of the transistor couple, a so called Class-AC stage, is shown in Fig. 11.29.

The IIP3 of the stage can be estimated from (11.11). For a single transistor biased at $V_G = 0.46\,\text{V}$ we read from Fig. 11.28 $g_m \approx 15\,\text{mS}$, $g_3 \approx -110\,\text{mA/V}^2$ giving an IIP3 of ca. –10.4 dBV. From Fig. 11.29 we see that a Class-AC stages reduces g_3 by ca. a factor of 10, while leaving g_m almost unchanged. From this data we estimate that the IIP3 should be ca. 10 dB higher or approximately –0.4 dBV. The results obtained by numerical simulation with the full transistor models are shown in Fig. 11.30. To suppress the effect of the nonlinear output conductance g_o the drain was held at 0.4 V using an ideal voltage source. The circuit was driven by a voltage source with a resistance of 50 Ω generating two tones of equal amplitude at $f_1 = 1.01\,\text{GHz}$ and $f_2 = 1.02\,\text{GHz}$. Note that the simulation does include the effect of a slightly

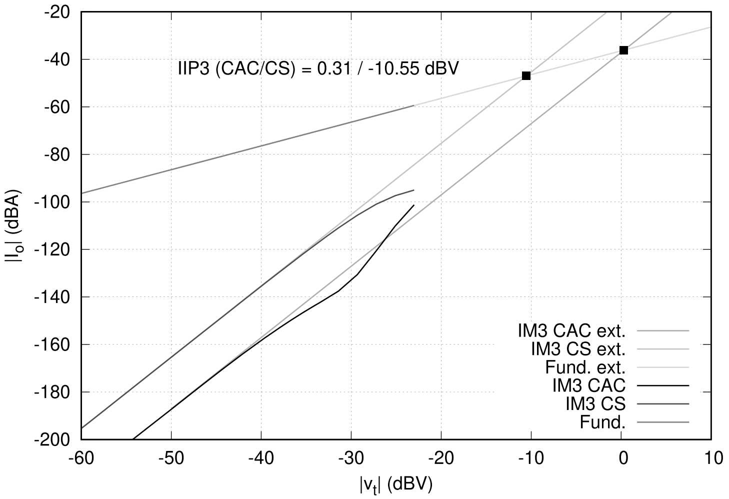


Fig. 11.30 Simulated IM3 of a Class-AC stage compared to the one of a simple common-source stage consisting of the Class-A device only. Class-A device: $L = 22\,\text{nm}, n_{\text{fin}} = 10, n_f = 16, m = 1$ biased at $V_G = 0.453\,\text{V}$. Class-C device: $L = 22\,\text{nm}$, $n_{\text{fin}} = 10$, $n_f = 8$, $m = 1$ biased at $V_G = 0.342\,\text{V}$. $V_D = 0.4\,\text{V}$. $|v_t|$ is the magnitude of each of the two input tones

nonlinear C_{gs} as well as the one of C_{gd}. The results are in good agreement with our estimates up to a level of about –25 dBV ($\approx$ 80 mV) per tone that translates in a peak input voltage of 160 mV. This is in line with expectation as beyond this level the third order approximation of the characteristic starts to break down as noted earlier.

The Class-AC stage reduces g_3, but doesn't reduce g_2. Therefore, if the second order transfer function of the preceding or following stage is also large, then the combined system will still produce third order distortion. If we call the first subsystem $\mathcal{G}$ and the second one $\mathcal{H}$ the combined third order impulse response is in fact (see Table 10.1)

$$(h \circ g)_3 = h_1 * g_3 + 2\,h_2 * [g_1 \otimes g_2]_{\text{sym}} + h_3 * g_1^{\otimes 3}$$

which doesn't disappear even if g_3 and h_3 are both zero. One approach to reduce g_2 (on top of g_3) is to use a complementary structure comprised of an nMOS Class-AC stage and a pMOS one as sketched in Fig. 11.31. Here we use common-gate stages (see the next section) to reduce the effects of C_{gd} and combine the currents through a transformer. For good results one needs large coupling between the primary and secondary of the transformer. In a monolithic implementation this is best achieved using equal coils stacked one on top of the other. We can also directly connect the drains of the two stages. In this case the bias currents of the two stages must coincide and a mean of controlling the DC drain voltage is necessary.

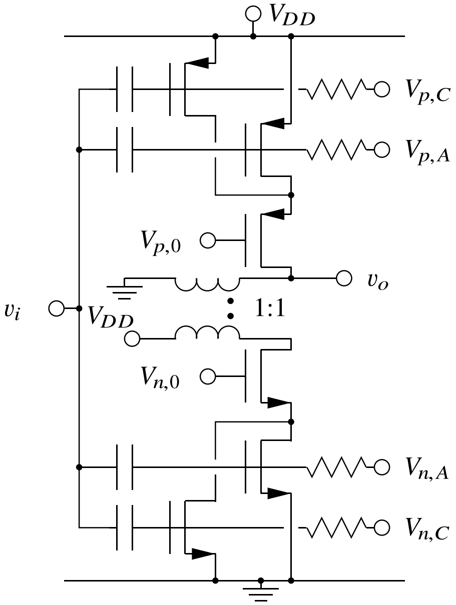


Fig. 11.31 Complementary Class-AC stage suitable for RF applications

11.6 Common-Gate Stage

In this section we investigate the linearity properties of the common-gate stage shown in Fig. 11.32a. We first consider the case in which the stage is driven by a source with internal resistance R_s and then specialise to the case in which the stage is used to form a *Cascode*. A basic variant of the Cascode stage suitable for use at RF frequencies is the combination of a common-source stage followed by a common-gate one. The combination of the two stages behaves as an improved common-source stage with much reduced C_{gd} and output conductance g_o [27]. In this section we will show that, under suitable conditions that we will work out, the addition of a common-gate stage does not degrade distortion either. Due to these very desirable benefits the Cascode stage is a widely used configuration.

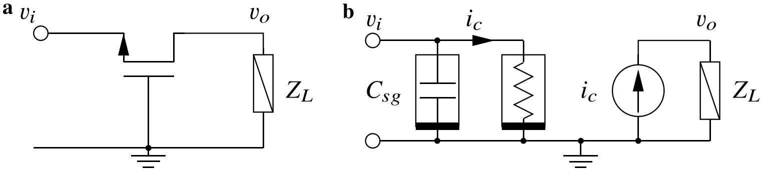


Fig. 11.32 **a** Common-gate stage AC schematic **b** Common-gate stage Small-signal model

Consider the small-signal model shown in Fig. 11.32b. The input voltage v_i corresponds to the source voltage. The input current is the current entering into the source terminal. The part of the input current that doesn't flow through C_{sg} is labeled i_c and represents the current that flows through the transistor active channel to the drain. The current leaving the drain terminal must therefore have the same value. This is represented by the output side current-controlled current source with unit gain. For simplicity, we neglect the distortion introduced by the nonlinear capacitance C_{sg} as well as the one introduced by the drain capacitance that in the figure was lumped together with the load Z_L. As before we characterise the linearity of the circuit by calculating the nonlinear terms present in the output current i_c.

11.6.1 Nonlinear Transfer Functions

According to the model presented in Sec. 11.5 (with $\lambda = 0$) the static characteristic of the transistor in saturation is given by

$$i_D = \frac{\beta}{2} v_{OD}^2$$

with $v_{OD} = v_{GS} - V_T$ the overdrive voltage and $\beta = K'W/L$. In the present situation it is more convenient to express the input voltage as a function of the current. This is easily achieved by inverting the equation. If we further separate the DC bias terms from the small signal quantities we obtain

$$v_{gs} = \sqrt{\frac{2(I_D + i_d)}{\beta}} - V_{OD}$$

which we approximate by a third order Taylor polynomial around the operating point

$$v_{gs} \approx V_{OD}\left[\frac{1}{2}\frac{i_d}{I_D} - \frac{1}{8}\left(\frac{i_d}{I_D}\right)^2 + \frac{1}{16}\left(\frac{i_d}{I_D}\right)^3\right].$$

Using the relations $v_i = -v_{gs}$ and $i_c = -i_d$ we obtain that the input characteristic corresponds to the one of a nonlinear resistor

$$v_i = r_1 i_c + r_2 i_c^2 + r_3 i_c^3 \tag{11.37}$$

with

$$r_1 = \frac{1}{g_m} = \frac{V_{OD}}{2I_D}, \qquad r_2 = \frac{V_{OD}}{8I_D^2}, \qquad r_3 = \frac{V_{OD}}{16I_D^3}. \tag{11.38}$$

Note that while the original expression giving i_D as a function v_{OD} doesn't include any third order term, the inverted expression gives a well-defined term of third order.

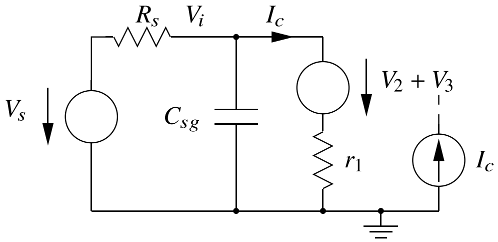


Fig. 11.33 Equivalent circuit for the calculation of the nonlinear transfer functions of the common-gate circuit

As a result, the latter is less sensitive to modeling inaccuracies than the former. We will therefore use the above values as estimates for $r_i, i = 1, \dots, 3$.

Using the above third order polynomial to model the source-gate nonlinear characteristic we obtain the equivalent circuit shown in Fig. 11.33 with V_2 and V_3 the second resp. third order equivalent nonlinear source as given in Table 11.11.

The first order transfer function is calculated by discarding the contribution of all sources of order different from one. This amounts to calculating the contribution due to the input source and using a Dirac impulse as input signal. Working in the Laplace domain, the Kirchhoff's voltage law gives

$$R_s(I_{c,1} + sC_{sg}r_1I_{c,1}) + r_1I_{c,1} = 1\,.$$

Solving for the first order component of I_c we find

$$I_{c,1}(s) = H_1(s) = \frac{1}{R_s + r_1 + sC_{sg}r_1R_s} = \frac{1}{R_s + r_1}\frac{1}{1 + \frac{s}{\omega_0}}$$

with $\omega_0 = (R_s + r_1)/(C_{sg}r_1R_s)$.

With $I_{c,1}$ and referring to Table 11.11 we can now compute the equivalent source of second order $V_2 = r_2I_{c,1}(s_1)I_{c,1}(s_2)$. The second order transfer function is the response to this source which is easily calculated to be

$$I_{c,2}(s_1, s_2) = H_2(s_1, s_2) = -\frac{1 + (s_1 + s_2)C_{sg}R_s}{R_s + r_1 + (s_1 + s_2)C_{sg}r_1R_s}r_2I_{c,1}(s_1)I_{c,1}(s_2)$$

or, expressed in terms of H_1

$$H_2(s_1, s_2) = -r_2[1 + (s_1 + s_2)C_{sg}R_s]H_1(s_1 + s_2)H_1(s_1)H_1(s_2)\,.$$

With $I_{c,2}$ we can compute the equivalent source of third order

$$V_3 = 2r_2\left[I_{c,1}(s_1)I_{c,2}(s_2, s_3)\right]_{\text{sym}} + r_3I_{c,1}(s_1)I_{c,1}(s_2)I_{c,1}(s_3)\,.$$

The third order transfer function is the response to V_3 which is calculated in a similar way as H_2

$$H_3(s_1, s_2, s_3) = -[1 + (s_1 + s_2 + s_3)C_{sg}R_s]H_1(s_1 + s_2 + s_3)V_3(s_1, s_2, s_3)\,.$$

11.6.2 Cascode

We now specialise to the case of a Cascode. Since the transfer functions are found by analysing a sequence of linear networks, we can use the Thévenin-Norton theorem [32] to transform the source into the parallel connection of an ideal current source and the internal resistor R_s as shown in Fig. 11.34. The resistor R_s corresponds now to the reciprocal of the output conductance g_o of the driving common-source stage. The latter is usually much larger than r_1, so it has little effect on the operation of the circuit. For this reason and to obtain easier to interpret expressions we calculate the transfer functions in the limit as R_s tends to infinity. Under this assumption and using the results of the previous section, the first, second and third order transfer functions from the ideal source I_s to the output current I_c are

$$H_{c1}(s) := \lim_{R_s \to \infty} H_1(s)R_s = \frac{1}{1 + \frac{s}{\omega_0}}\,, \tag{11.39}$$

$$\begin{aligned} H_{c2}(s_1, s_2) &:= \lim_{R_s \to \infty} H_2(s_1, s_2)R_s^2 \\ &= -r_2(s_1 + s_2)C_{sg}H_{c1}(s_1)H_{c1}(s_2)H_{c1}(s_1 + s_2) \end{aligned} \tag{11.40}$$

and

$$\begin{aligned} H_{c3}(s_1, s_2, s_3) &:= \lim_{R_s \to \infty} H_3(s_1, s_2, s_3)R_s^3 \\ &= (s_1 + s_2 + s_3)C_{sg}\left\{2r_2^2\left[H_{c1}(s_1 + s_2)(s_1 + s_2)C_{sg}\right]_{\text{sym}} - r_3\right\} \\ &\quad \cdot H_{c1}(s_1)H_{c1}(s_2)H_{c1}(s_3)H_{c1}(s_1 + s_2 + s_3) \end{aligned} \tag{11.41}$$

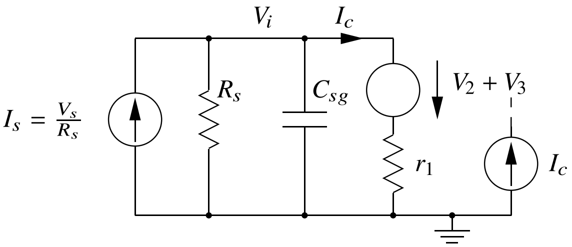


Fig. 11.34 Equivalent circuit for the calculation of the nonlinear transfer functions of the Cascode circuit

respectively, where now $\omega_0 = 1/(C_{sg}r_1)$. Note that the symmetrization in H_{c3} is intended over all three Laplace variables s_1, s_2 and s_3

$$\begin{aligned}\left[H_{c1}(s_1+s_2)(s_1+s_2)C_{sg}\right]_{\text{sym}} = \frac{1}{3}\Big\{ & H_{c1}(s_1+s_2)(s_1+s_2)C_{sg}\\ & + H_{c1}(s_1+s_3)(s_1+s_3)C_{sg}\\ & + H_{c1}(s_3+s_2)(s_3+s_2)C_{sg}\Big\}.\end{aligned}$$

Consider now the classic two-tones third order intermodulation test with one tone at ω_1 and the second one at $\omega_2 = \omega_1 + \Delta\omega$. In particular consider the IM3 tone characterised by $m = (1, 0, 2, 0)$. Assuming $|\Delta\omega| \ll |\omega_1|$ the above symmetrised expression can be approximated by

$$\left[H_{c1}(s_1+s_2)(s_1+s_2)C_{sg}\right]_{\text{sym}} \approx \frac{2}{3}\frac{\jmath\omega_1 C_{sg}}{1+\jmath\frac{2\omega_1}{\omega_0}}$$

and, with it, the third order transfer function by

$$\begin{aligned}& H_{c3}(\jmath\omega_1, \jmath\omega_1, -\jmath\omega_2)\\ & \quad\approx \jmath\omega_1 C_{sg}\left\{\frac{4}{3}r_2^2\jmath\omega_1 C_{sg}H_{c1}(2\jmath\omega_1) - r_3\right\}H_{c1}(\jmath\omega_1)H_{c1}(\jmath\omega_1)H_{c1}(-\jmath\omega_1)H_{c1}(\jmath\omega_1).\end{aligned}$$

If $\omega_1 \le \omega_0/5$ the value of $|H_{c1}(\jmath\omega_1)|$ can be approximated by 1 with an error of less than 2% and the magnitude of $|H_{c3}|$ becomes very nearly

$$\omega_1 C_{sg}\left|\frac{4}{3}r_2^2\jmath\omega_1 C_{sg}H_{c1}(2\jmath\omega_1) - r_3\right|.$$

Using (11.38) for the coefficients of the nonlinear characteristic of the transistor we thus obtain

$$|H_{c3}(\jmath\omega_1, \jmath\omega_1, -\jmath\omega_2)| \approx \frac{1}{8I_D^2}\frac{\omega_1 C_{sg}}{g_m}\left|\frac{2}{3}\frac{\jmath\omega_1 C_{sg}}{g_m}H_{c1}(2\jmath\omega_1) - 1\right|. \tag{11.42}$$

The magnitude of the IM3 tone normalised to the DC current I_D is therefore

$$\left|\frac{I_{c3,m}}{I_D}\right| \approx \frac{3}{32}\frac{\omega_1 C_{sg}}{g_m}\left|\frac{2}{3}\frac{\jmath\omega_1 C_{sg}}{g_m}H_{c1}(2\jmath\omega_1) - 1\right|\left|\frac{I_s}{I_D}\right|^3.$$

From this expression we can read several interesting aspects. First, both the second and the third order nonlinearities of the transistor characteristic contribute to the IM3 tone. This is visible from the appearance of r_3 as well as r_2 in the expression for H_{c3}. The contribution to an intermodulation product of third order by second-order nonlinearities is due to the presence of (local) feedback. This can be appreciated graphically by looking at Fig. 11.34. The second order source V_2 creates a current that circulates again through the input of the circuit. Therefore, the generated second order tones pass again through the second order distortion where they can mix with the input tones to produce frequency mixes of third order.

The contribution to the IM3 tone from second-order distortion is approximately orthogonal to the one from third order distortion. Therefore, it's not possible to size the transistor in such a way as to make the two cancel each other, not even at a specific frequency.

The IM3 is largely dominated by r_3 up to very high frequencies and for ω_1 up to ca. $\omega_0/10$ it is proportional to ω_1. The quantity g_m/C_{sg} corresponds (neglecting C_{gd}) to the angular frequency at which a common-source stage has unity current gain. It is called *transit frequency* and denoted by

$$\omega_T = \frac{g_m}{C_{sg}} \, . \tag{11.43}$$

It is one of the key parameters used to characterise the high-frequency capabilities of transistors. With it the magnitude of the IM3 up to ca. $\omega_1 \leq \omega_T/10$ can be approximated by

$$\left|\frac{I_{c3,m}}{I_D}\right| \approx \frac{3}{32}\frac{\omega_1}{\omega_T}\left|\frac{I_s}{I_D}\right|^3 .$$

This shows that for low distortion one needs fast transistors. Looking again at Fig. 11.34 we can appreciate that in the limit as ω_1/ω_T tends to zero (which means that C_{sg} tends to zero) the nonlinear sources become floating and can't generate any frequency mix current (remember that we also assume $R_s \to \infty$).

In general, distortion introduced by the input (common-source) stage of the Cascode configuration generates frequency mixes of second-order. These can mix with the fundamental tones in the second-order distortion of the output (common-gate) stage to produce other IM3 components. However, since $|H_{c2}(\jmath 2\omega_1, -\jmath\omega_2)|$ is also proportional to ω_1/ω_T this does not substantially change the situation.

For simplicity in our discussion we assumed $R_s \to \infty$. From the gained insight we can appreciate that at low frequencies it is a finite value of R_s which will limit IM3 and, the lower R_s, the higher the IM3. In general however, the common-gate stage of a Cascode is not the stage limiting low frequency linearity.

11.7 Degenerated Common-Source Stage

In this section we investigate the effect of local feedback on distortion and show that introduction of feedback may lead to degraded linearity. As a concrete example we analyse the degenerated common-source amplifier depicted in Fig. 11.35a. The impedance Z_e is called the *degeneration impedance*. Its presence reduces the gate-source voltage across the transistor by an amount proportional to the output current. In

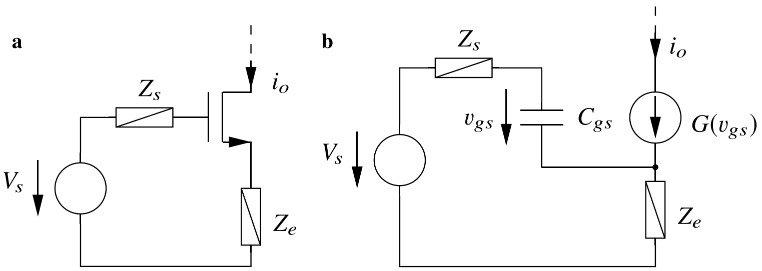


Fig. 11.35 **a** Degenerated common-source stage AC schematic **b** Degenerated common-source stage small-signal model

other words it introduces feedback around the transistor. The impedance Z_s represents a generic driving impedance.

One way to analyse the circuit is to model the transistor as a nonlinear voltage-controlled current-source characterised by a third (or higher) order polynomial

$$i_d = g_m v_{gs} + g_2 v_{gs}^2 + g_3 v_{gs}^3$$

and solve Kirchhoff's equations for v_{gs}. Having found the voltage components $v_{gs,1}, \ldots, v_{gs,k}$ up to some order of interest k, one then finds the output current components $i_{o,1}, \ldots, i_{o,k}$ by use of the polynomial approximating the transistor characteristic. Instead of using this method we show how the use of a *nullor* allows the problem to be solved in a more direct way, by permitting to directly obtain an equation for the output current i_o.

Nullators and Norators are pathological network elements. A *Nullator* is a two terminal element represented by the symbol shown in Fig. 11.36a and characterised by the *two* equations

$$V = 0 \qquad I = 0\,.$$

A *Norator* is a two terminal element represented by the symbol shown in Fig. 11.36b whose current and voltage are arbitrary and completely determined by the surrounding network. In other words it is characterised by *zero* equations. For a linear network to have a well-defined solution a Nullator must therefore always appear alongside a Norator. Such a pair is called a *Nullor* and can be used to model several elements such as controlled sources, OpAmps and transistors. In particular, we can use it to represent the inverted series (see Sect. 11.6.2)

$$v_{gs} = r_1 i_d + r_2 i_d^2 + r_3 i_d^3$$

of the transistor characteristic. A nullor based small-signal model of the degenerated common-source stage using this transistor characteristic representation is shown in Fig. 11.37. Note that the transistor characteristic is represented by a nonlinear resistor.

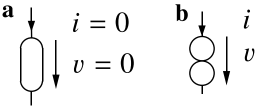


Fig. 11.36 **a** Symbol of the Nullator reminding the shape of the zero digit 0 **b** Symbol of a Norator reminding the infinity symbol ∞

11.7.1 Nonlinear Transfer Functions

From the model in Fig. 11.37 and Kirchhoff's laws we obtain the following system of equations relating the output current I_o to the input signal V_s

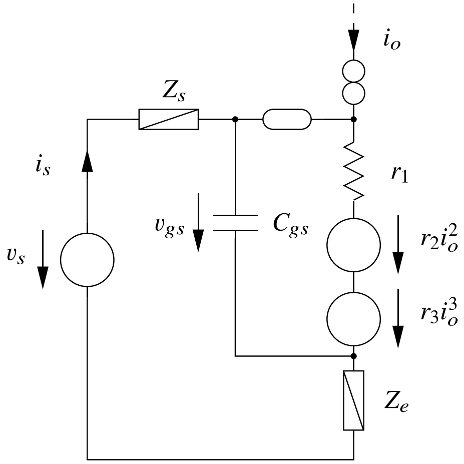
$$\begin{aligned} V_s &= (Z_s + \frac{1}{sC_{gs}})I_s + Z_e(I_s + I_o) \\ V_{gs} &= \frac{1}{sC_{cs}} I_s \\ V_{gs} &= r_1 I_o + r_2 I_o^2 + r_3 I_o^3 \,. \end{aligned}$$

After eliminating V_{gs} and I_s we obtain the single equation

$$\begin{aligned} V_s = &\big[r_1 + Z_e + (Z_s + Z_e) r_1 C_{gs} s\big] I_o \\ &+ \big[1 + (Z_s + Z_e) C_{gs} s\big](r_2 I_o^2 + r_3 I_o^3) \,. \end{aligned} \tag{11.44}$$

The first order transfer function is obtained by applying a Dirac impulse as input and discarding all terms of order higher than one in the equation. This is equivalent

Fig. 11.37 Nullor based small-signal model of a degenerated common-source stage



to removing the nonlinear sources from the equivalent circuit. Using the relation $r_1 = 1/g_m$ we obtain

$$H_1(s) = \frac{g_m}{L(s)} = \frac{g_m}{1 + g_m Z_e + sC_{gs}(Z_e + Z_s)} \,. \tag{11.45}$$

To compute the second order nonlinear transfer function we first insert the first order solution into the nonlinear terms and retain only second order ones. Alternatively we can use Fig. 11.11 to read the value of the second order nonlinear source for a nonlinear resistor. In both cases, after adjusting the representation of the differential operator by replacing the variable s by $s_1 + s_2$, we obtain

$$\begin{aligned} 0 = &\big[r_1 + Z_e + (Z_s + Z_e) r_1 C_{gs}(s_1 + s_2)\big] H_2(s_1, s_2) \\ &+ \big[1 + (Z_s + Z_e) C_{gs}(s_1 + s_2)\big] r_2 H_1(s_1) H_1(s_2) \,. \end{aligned}$$

Note that for brevity we didn't explicitly write the argument of impedances. Their value has of course to be evaluated at $s_1 + s_2$. The second order nonlinear transfer function is thus

$$H_2(s_1, s_2) = -r_2 H_1(s_1) H_1(s_2) H_1(s_1 + s_2)\big[1 + (Z_s + Z_e) C_{gs}(s_1 + s_2)\big] \,. \tag{11.46}$$

To find the third order nonlinear transfer function we proceed in a similar way and obtain

$$\begin{aligned} H_3(s_1, s_2, s_3) = -\ &\big\{2r_2\, [H_1(s_1) H_2(s_2, s_3)]_{\text{sym}} + r_3 H_1(s_1) H_1(s_2) H_1(s_3)\big\} \\ &H_1(s_1 + s_2 + s_3)\big[1 + (Z_s + Z_e) C_{gs}(s_1 + s_2 + s_3)\big] \,. \end{aligned} \tag{11.47}$$

11.7.2 Resistive Degeneration

We now specialise to the case of resistive degeneration $Z_e = R_e$ and a resistive driving impedance $Z_s = R_s$ and calculate the intermodulation products of third order when driven by two tones of equal amplitudes at frequency ω_1 and $\omega_1 + \Delta\omega$ respectively. As usual we assume $\Delta\omega \ll \omega_1$.

As a first step, to calculate H_3 for the mix $(1, 0, 2, 0)$ we evaluate

$$\begin{aligned} &[H_1(\jmath\omega_1) H_2(\jmath\omega_1, -\jmath(\omega_1 + \Delta\omega))]_{\text{sym}} \\ &\qquad \approx -r_2 \frac{g_m^4}{3\, L(\jmath\omega_1)^2 L(-\jmath\omega_1)} \left[2\frac{N(-\jmath\Delta\omega)}{L(-\jmath\Delta\omega)} + \frac{N(2\omega_1)}{L(2\omega_1)}\right] \end{aligned}$$

with

$$N(s) := 1 + (Z_s + Z_e) s C_{gs} \,.$$

Inserting this expression into H_3 we obtain

$$H_3(\jmath\omega_1, \jmath\omega_1, -\jmath(\omega_1+\Delta\omega)) \\ \approx \frac{g_m^4 N(\jmath\omega_1)}{L(\jmath\omega_1)^3 L(-\jmath\omega_1)} \left\{ \frac{2}{3} r_2^2 g_m \left[2\frac{N(-\jmath\Delta\omega)}{L(-\jmath\Delta\omega)} + \frac{N(2\omega_1)}{L(2\omega_1)} \right] - r_3 \right\}.$$

Since in Sect. 11.5 we characterised the transistor in terms of g_m, g_2 and g_3, we express the coefficients r_2 and r_3 in terms of them using the results of Sect. 11.2.1

$$r_2 = \frac{-g_2}{g_m^3} \qquad\qquad r_3 = \frac{2g_2^2 - g_m g_3}{g_m^5}.$$

Substituting these expressions leads finally to

$$H_3(\jmath\omega_1, \jmath\omega_1, -\jmath(\omega_1+\Delta\omega)) \\ \approx \frac{N(\jmath\omega_1)}{L(\jmath\omega_1)^3 L(-\jmath\omega_1)} \left\{ 2\frac{g_2^2}{g_m} \left[\frac{2}{3}\frac{N(-\jmath\Delta\omega)}{L(-\jmath\Delta\omega)} + \frac{1}{3}\frac{N(2\omega_1)}{L(2\omega_1)} - 1 \right] + g_3 \right\}. \tag{11.48}$$

We can now discuss the effect of a small amount of feedback introduced by a small resistor R_e on linearity. First note that, as expected, for $Z_e = 0\ \Omega$ the term in square brackets vanishes making the IM3 depend only on g_3. As R_e is increased the contribution of g_2 increases and at low to moderate frequencies there is some possibility of cancelling between the contribution due to g_3 and g_2. As R_e increases beyond this value, the second order contribution starts to dominate. At high frequencies only imperfect cancelling is possible due to shift in phase of the g_2 contribution.

Figure 11.38b shows the low to moderate frequency IIP3 of the Class-AC stage from Sect. 11.5. It shows that cancelling occurs for very small amounts of feedback and, as is typical for cancelling effects, the performance is very sensitive to small variations in component values. Due to the large value of g_2, a small to moderate amount of feedback with $g_m R_e$ in the range of 0.03–2.5 leads to an actual degradation in IIP3. Note that small values of Z_e may be introduced unintentionally by parasitic effects due to the interconnections between components.

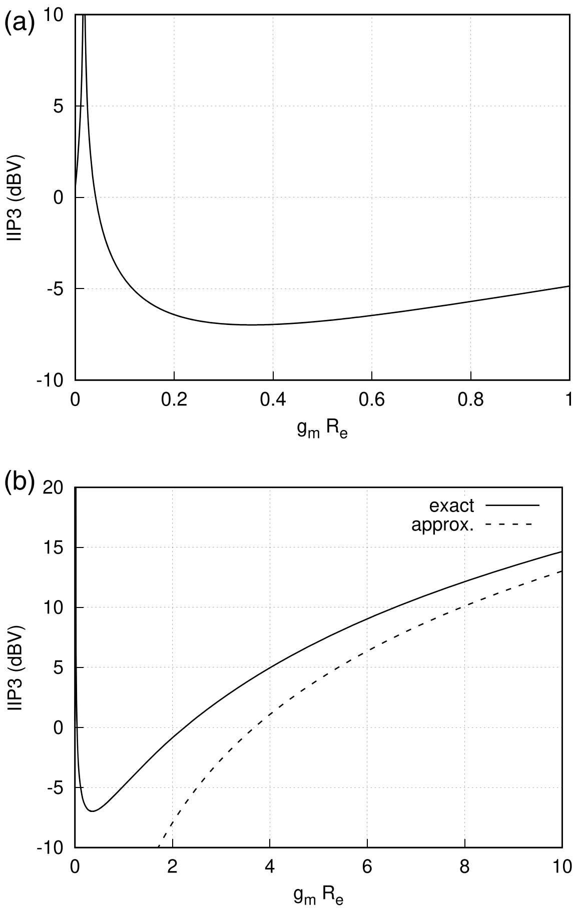
A linearity improvement can be obtained by using a large amount of feedback $g_m R_e \gg 1$. To simplify calculations let's assume $\left|\omega_1 C_{gs}(R_s + R_e)\right| \ll 1$, then

$$H_3(\jmath\omega_1, \jmath\omega_1, -\jmath(\omega_1+\Delta\omega)) \approx \frac{1}{L(\jmath\omega_1)^3 L(-\jmath\omega_1)} \left\{ -2\frac{g_2^2}{g_m} + g_3 \right\}$$

and

$$H_1(\jmath\omega_1) \approx \frac{1}{R_e}.$$

Fig. 11.38 **a** Weak feedback region of Fig. 11.38b **b** Moderate-frequency resistively degenerated common-source transconductance stage IIP3. The solid curves are calculated using (11.45), (11.47); the dashed line with (11.49). The transistor parameters are approximate values taken from Figs. 11.28 and 11.29: $g_m = 17\,\mathrm{mS}$, $g_2 = 70\,\mathrm{mS/V}$, $g_3 = 10\,\mathrm{mS/V^2}$, $R_s = 50\,\Omega$



Using (11.11) to calculate the IIP3 shows that under these conditions the latter does in fact increase with increasing R_e

$$\mathrm{IIP3} \approx \sqrt{\frac{4(g_m R_e)^3}{3\left|\frac{g_3}{g_m} - 2\left(\frac{g_2}{g_m}\right)^2\right|}}, \qquad g_m R_e \gg 1\,. \tag{11.49}$$

The reason for the improvement is a substantially reduced amplitude of the voltage V_{gs} controlling the nonlinear sources compared to the circuit input signal V_s. Linearity thus comes at the expenses of a much reduced signal transconductance which for RF circuits is often not acceptable.

11.7.3 Inductive Degeneration

A second type of degeneration widely used ar RF frequencies is the inductive one. This type of degeneration is often used in the input stage of low-noise RF amplifiers (LNAs), a basic small-signal model of which is shown in Fig. 11.39.

An important characteristic of RF amplifiers is the input impedance Z_i. In many situations it is required to be real and equal to the source impedance R_s, or some standard value. From our model a simple calculation shows that Z_i is given by

$$Z_i = R_i + \jmath X_i = g_m \frac{L_e}{C_{gs}} + s(L_s + L_e) + \frac{1}{sC_{gs}}\,.$$

A degeneration inductor L_e thus allows a real part to be introduced to the input impedance without using resistors. Avoiding resistors at the input of LNAs is necessary to avoid limiting the achievable sensitivity. The reactive part of the impedance can be cancelled over some frequency band by resonating it, in our example using the inductor L_s. The input network thus consists of a series resonator tuned at the center frequency of the band of interest.

In this section we analyse the linearity characteristics of this stage and in particular its IP3. The nonlinear transfer functions are readily obtained from our previous results by setting $Z_e = sL_e$ and $Z_s = R_s + sL_s$. Doing so, the first order transfer function becomes

$$H_1(s) = \frac{g_m}{1 + s(g_m L_e + C_{gs} R_s) + s^2 C_{gs}(L_e + L_s)}\,.$$

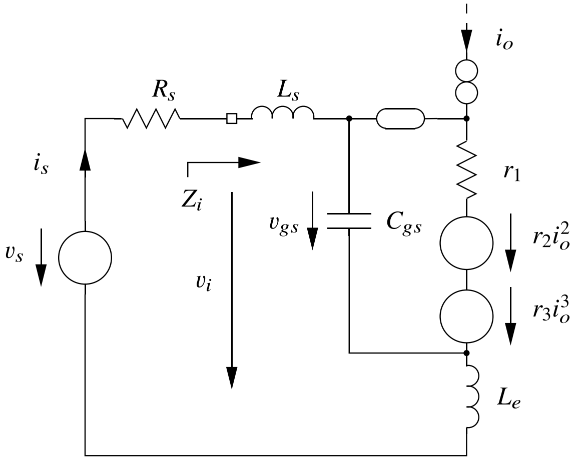


Fig. 11.39 Small-signal model of an inductively degenerated common-source stage

Note that $g_m L_e = C_{gs} R_i$. We can therefore write the denominator in the standard form

$$H_1(s) = \frac{g_m}{1 + \frac{s}{\omega_0}\frac{1}{q_t} + \left(\frac{s}{\omega_o}\right)^2} \tag{11.50}$$

with

$$\omega_0^2 = \frac{1}{C_{gs}(L_e + L_s)} \qquad q_i = \frac{1}{R_i \omega_0 C_{gs}}$$
$$\frac{1}{q_t} = \frac{1}{q_i} + \frac{1}{q_s} \qquad q_s = \frac{1}{R_s \omega_0 C_{gs}}\,.$$

The same parameters can also be used to put $N(s)$ in standard form

$$N(s) = 1 + \frac{s}{\omega_0}\frac{1}{q_s} + \left(\frac{s}{\omega_o}\right)^2.$$

The value of H_3 relevant for the two-tones IP3 test can then be obtained by substituting these expressions in (11.48). The resonance frequency of the input resonator is evidently set to the frequency of the input signal $\omega_0 = \omega_1$ so that

$$N(j\omega_1) = \frac{j}{q_s}, \qquad L(j\omega_1) = \frac{j}{q_t}, \qquad H_1(j\omega_1) = -jq_t gm\,.$$

and

$$H_3(j\omega_1, j\omega_1, -j(\omega_1 + \Delta\omega)) \approx \frac{-jq_t^4)}{q_s}\left\{2\frac{g_2^2}{g_m}\left[\frac{2}{3} + \frac{1}{3}\frac{N(2\omega_1)}{L(2\omega_1)} - 1\right] + g_3\right\}.$$

With these results we can compute the IIP3 using Eq. (11.11) as before

$$\text{IIP3} \approx \frac{2}{q_t}\sqrt{\frac{q_s}{q_t}\frac{1}{\left|2\left(\frac{g_2}{g_m}\right)^2\left[\frac{2}{3} + \frac{1}{3}\frac{N(2\omega_1)}{L(2\omega_1)} - 1\right] + \frac{g_3}{g_m}\right|}}\,. \tag{11.51}$$

In the common case in which the input resistance R_i is equal to the source impedance $q_s/q_t = 2$. The IP3 of the circuit is thus approximately inversely proportional to the quality factor of the input resonance. This is due to the fact that at resonance the magnitude of the voltage across the reactive components is roughly q_t times the one across the resistive part. In other words, the voltage V_{gs} controlling the nonlinear sources is amplified by a factor of ca. q_t compared to the input signal V_s. This very same characteristic is also the reason for the good noise characteristic of the circuit: the input network provides some voltage gain before the first noisy device.

The quality factor of the network also influences the relative contributions of g_2 and g_3 to distortion through the term

$$\frac{N(2\jmath\omega_1)}{L(2\jmath\omega_1)} = \frac{-3 + \jmath\frac{2}{q_s}}{-3 + \jmath\frac{2}{q_t}}.$$

For large quality factors $q_t, q_s \gg 1$ the ratio approaches 1 which makes the IP3 essentially independent of g_2. For small quality factor values the contribution due to g_2 is not negligible, especially if g_3/g_m is small compared to $(g_2/g_m)^2$ as is the case with Class-AC stages.

In practical implementations the component values are affected by manufacturing variations. For this reason and to avoid the need for tuning, the quality factor q_t is most often chosen to have a value smaller than 5.

11.8 Pseudo-Differential Circuits

The analog signal path of many RF and mixes-signal integrated circuits is *differential*. This means that the signal of interest is transmitted on two equal lines carrying the same signal, but with opposite polarities. The main objective is to make the system insensitive to noise affecting both lines equally. This can be, for example, noise due to the activity of digital circuits propagating through the common substrate of the IC. A *differential circuit* is one that is designed to process the difference between the two input terminals sensing the two lines carrying the signal and rejecting the common component. Formally, if v_i^+ and v_i^- are the two input voltages (relative to ground), the *differential-mode* voltage is defined as

$$v_d := v_i^+ - v_i^-$$

and the *common-mode* voltage as

$$v_c := \frac{v_i^+ + v_i^-}{2}\,.$$

Using this representation the two input voltages can be written as

$$v_i^+ = v_c + \frac{v_d}{2}\,, \qquad\qquad v_i^- = v_c - \frac{v_d}{2}\,.$$

The prototypical differential circuit is the *differential-pair* shown in Fig. 11.40. In the ideal drawn form the output currents are always $i_o^+ = i_o^- = I_0/2$ as long as $v_i^1 = v_i^-$. Any common-mode signal component is thus fully rejected.

Differential circuits do also have disadvantages. A real current source is implemented with transistors and requires a certain voltage across its terminals to work

Fig. 11.40 Differential pair

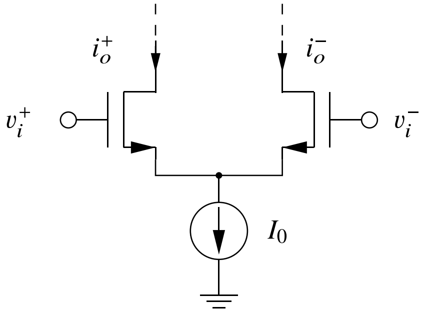
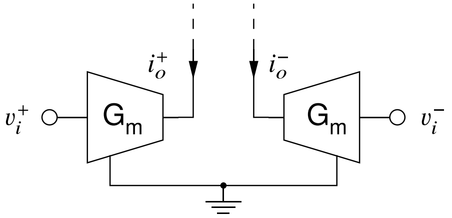


Fig. 11.41 Pseudo-differential transconductance



properly. This reduces the headroom left for signal processing and in modern processes operating at supplies voltages below 1.0 V poses severe challenges. In addition, a current source does not only generate a DC current, but it also generates noise, reducing the sensitivity of the circuit to small signals.

Pseudo-differential circuits are a class of circuits that alleviate some of these problems while retaining some of the benefits of differential circuits. They are circuits composed by two equal single-ended sub-circuits each connected to one of the two lines carrying the differential signal. An example pseudo-differential transconductance is shown in Fig. 11.41.

In pseudo-differential circuits the input common-mode signal component is not rejected, but, if the circuit is sufficiently linear, the common-mode input appears as a common-mode signal at the output and remains separable from the wanted differential signal which appears at the output in differential form. The objective of this section is to quantify the conversion between common-mode and differential-mode in weakly nonlinear circuits.

We first show that weakly nonlinear circuits driven by a purely differential input signal produce a mixture of differential- and common-mode output signals. Let's denote the input signals by x^+, x^-. the output signals by y^+, y^-, the relative common- and differential-mode components by the same letter with index c and d respectively; and the nonlinear transfer function of order k of the single-ended subsystems by h_k. By assumption the input signal is purely differential

$$x^+ = \frac{x_d}{2} \qquad\qquad x^- = -\frac{x_d}{2}\,.$$

The outputs of order k are therefore

$$y_k^+ = \frac{1}{2^k} h_k * x_d^{\otimes k}, \qquad y_k^- = \frac{(-1)^k}{2^k} h_k * x_d^{\otimes k}$$

from which we conclude that for k *even* the output is a common-mode signal, while for k *odd* it is differential.

Let's now consider the response of a weakly nonlinear circuit to a mixture of differential- and common-mode signals

$$x^+ = x_c + \frac{x_d}{2} \qquad x^- = x_c - \frac{x_d}{2}.$$

Let's first consider the second-order response of the two circuit halves. The positive and negative outputs are

$$\begin{aligned} y_2^+ &= h_2 * \left(x_c + \frac{x_d}{2}\right)^{\otimes 2} \\ &= h_2 * x_c^{\otimes 2} + \frac{1}{4} h_2 * x_d^{\otimes 2} + h_2 * [x_d \otimes x_c]_{\text{sym}} \end{aligned}$$

and

$$\begin{aligned} y_2^- &= h_2 * \left(x_c - \frac{x_d}{2}\right)^{\otimes 2} \\ &= h_2 * x_c^{\otimes 2} + \frac{1}{4} h_2 * x_d^{\otimes 2} - h_2 * [x_d \otimes x_c]_{\text{sym}} \end{aligned}$$

respectively. The second-order differential output signal component is therefore

$$y_{d,2} = 2h_2 * [x_d \otimes x_c]_{\text{sym}}$$

which includes the common-mode input signal. A similar calculation for the third order component gives

$$y_{d,3} = h_3 * \left(\frac{x_d^{\otimes 3}}{4} + 3\left[x_d \otimes x_c^{\otimes 2}\right]_{\text{sym}}\right)$$

which also includes a term depending on the input common-mode. One can generalise the calculations and show that *the differential- and common-mode input components are mixed by nonlinearities of all orders.*

Consider now the cascade of two pseudo-differential weakly nonlinear circuits driven by a purely differential signal. If the two subsystems are optimised independently to maximise IP3 without paying attention to even order distortion components, then, when the two subsystems are put together, one may obtain a lower than expected total IP3. This is because the first stage produces second (and higher even) order mixes as common-mode signals which are also fed as input to the second subsystem. The second (and higher order) distortion components of the latter will then mix differential- and common-mode to produce differential output signal components at the IM3 frequencies.

Chapter 12
Linear Time-Varying Systems

12.1 Linear Time-Varying Systems

In this chapter we consider linear time-varying (LTV) systems. These are systems whose behaviour depends on the particular moment in time at which they are used. The change with time may arise for example due to the sensitivity of system components to environmental changes. Examples of systems suffering from this type of sensitivity include wireless communication systems in which the communication channel between the transmitter- and the receiver-antennas is highly dependent on the environment in between and around the antennas. The variation in time may also be imposed intentionally by design to achieve functions that can't be realised with LTI-systems. This is the case for example in communication mixers whose function is to shift in frequency the spectrum of a signal.

In this section we introduce a definition of linear time-varying systems valid under the assumption that all signals are regular distributions. A generalisation will be given in Sect. 12.3. The assumption of linearity means that the *superposition principle* must hold. In addition, as for LTI-systems, we require that LTV-systems depend *continuously* on the input signal. We therefore define

Definition 12.1 (*LTV-system*) A single-input, single-output, linear time-varying system is a system that when driven by the input signal x produces a response y that can be expressed by

$$y(t) = h(t, \xi) *_t x(t) := \int_{-\infty}^{\infty} h(t, \xi) x(t - \xi) d\xi \,. \tag{12.1}$$

$h(t, \xi)$ is the *time-varying impulse response* of the system.

The meaning of the variable ξ is best illustrated by anticipating somewhat the results of Sect. 12.3 and apply as input signal a Dirac impulse at time t_0

$$y(t) = h(t, \xi) *_t \delta(t - t_0) = h(t, t - t_0) \,.$$

F. Beffa, *Weakly Nonlinear Systems*, Understanding Complex Systems,
https://doi.org/10.1007/978-3-031-40681-2_12

Thus ξ represents the time lapsed since the application of the input impulse.

For *causal* systems the output must vanish before the input is applied. This implies that the impulse response must vanish for negative values of ξ

$$h(t,\xi) = 0\,, \qquad \xi < 0\,. \tag{12.2}$$

Therefore, the response of a causal system driven by a regular distributions $x \in \mathcal{D}'_+$ is given by

$$y(t) = \int_0^t h(t,\xi)x(t-\xi)d\xi = \int_0^t h(t,t-\xi)x(\xi)d\xi\,.$$

12.2 Linear Ordinary Differential Equations

An important class of LTV-systems is the one of systems described by differential equations with variable coefficients of the form

$$L\left(t,\frac{\mathrm{d}}{\mathrm{d}t}\right)y(t) = N\left(t,\frac{\mathrm{d}}{\mathrm{d}t}\right)x(t)$$

with

$$\begin{aligned} L\left(t,\frac{\mathrm{d}}{\mathrm{d}t}\right) &= \frac{\mathrm{d}^m}{\mathrm{d}t^m} + a_{m-1}(t)\frac{\mathrm{d}^{m-1}}{\mathrm{d}t^{m-1}} + \cdots + a_0(t),\\ N\left(t,\frac{\mathrm{d}}{\mathrm{d}t}\right) &= b_n(t)\frac{\mathrm{d}^n}{\mathrm{d}t^n} + b_{n-1}(t)\frac{\mathrm{d}^{n-1}}{\mathrm{d}t^{n-1}} + \cdots + b_0(t) \end{aligned}$$

time-dependent differential operators. It's easy to verify that every such system with $n < m$ can be represented in a state-space representation with time dependent matrices

$$\frac{\mathrm{d}}{\mathrm{d}t}u = A(t)u + B(t)x \qquad A(.) \in C(\mathbb{R},\mathbb{C}^{n\times n}),\, B(.) \in C(\mathbb{R},\mathbb{C}^{n\times 1}) \tag{12.3}$$

$$y = C(t)u + D(t)x \qquad C \in C(\mathbb{R},\mathbb{C}^{1\times n}),\, D \in C(\mathbb{R},\mathbb{C}) \tag{12.4}$$

with u the system state. Given an input signal x, the system response y is fully determined if one can find a state u satisfying the first equation and suitable initial conditions. The study of the dynamics of the system can therefore be reduced to the study of a system of n differential equations of first order.

12.2.1 Fundamental Solution

Consider the initial value problem described by the system of n differential equations

$$\frac{\mathrm{d}}{\mathrm{d}t} y = A(t) y \tag{12.5}$$

and initial conditions

$$y(0) = y_0 \in \mathbb{C}^n \tag{12.6}$$

with $A(.)$ an $n \times n$ matrix of complex valued functions of time $a_{ij}(.)$. If the functions forming $A(.)$ are bounded and continuous, then the right-hand side of the equation is Lipschitz continuous and, as discussed in Sect. 9.1, the equation has a unique solution. By choosing the initial value equal to the unit vector $e_j \in \mathbb{C}^n$ pointing in direction j, for $j = 1, \dots, n$ we can thus obtain n independent solutions y_j of the equation. The matrix formed by the column vectors y_j

$$Y(t) := \big[y_1(t), \dots, y_n(t) \big] \tag{12.7}$$

is called *principal fundamental matrix* of the system and satisfies the matrix equation

$$\frac{\mathrm{d}}{\mathrm{d}t} Y = A(t) Y, \qquad Y(0) = I\,. \tag{12.8}$$

Knowing Y, the solution of the initial value problem is thus given by

$$y(t) = Y(t) y_0 \qquad t \geq 0\,.$$

In addition, since the columns of Y are independent at all times, $\det(Y(t)) \neq 0$ at all times. The inverse of Y, Y^{-1}, is thus well-defined as is the *evolution operator* (also called *state transition matrix*)

$$U(t, \tau) := Y(t) Y^{-1}(\tau)\,. \tag{12.9}$$

Note that the evolution operator satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t} U(t, \tau) = \Big(\frac{\mathrm{d}}{\mathrm{d}t} Y(t) \Big) Y^{-1}(\tau) = A(t) Y(t) Y^{-1}(\tau) = A(t) U(t, \tau)$$

and

$$U(\tau, \tau) = I$$

and is thus the principal fundamental matrix of the system *at time* τ. From (12.9) we also immediately obtain

$$U(t, \lambda) U(\lambda, \tau) = U(t, \tau)$$

and

$$U(\tau, t) = [U(t, \tau)]^{-1}\,.$$

The initial value problem described by (12.5) and initial conditions $y(t_0) = y_0$ can be translated in the language of distributions by extending the functions by zero for $t < t_0$ and by replacing the differential operator by the distributional one

$$Dy = A(t)y + y_0\delta(t - t_0) \tag{12.10}$$

as usual. Differently from the case where $A(.)$ is constant, this equation can not be written as a convolution equation. For this reason and since for arbitrary distributions multiplication is only well-defined with smooth functions, for the equation to be well-defined the functions $a_{ij}(.)$ must belong to $\mathcal{E}$. This may seem like a very serious limitation, but remember that any distribution can be approximated to arbitrary accuracy by such a function (see Sect. 3.3). In this case the *fundamental (or elementary) solution* of the equation relative to time τ is defined as the solution of the matrix equation

$$LE_\tau = I\delta(t - \tau) \tag{12.11}$$

with L the differential operator

$$L := L(t, D) := D - A(t)\,.$$

If $U(t, \tau)$ is the evolution operator of the original differential equation (12.5) then

$$D\mathsf{1}_+(t - \tau)U(t, \tau) = \delta(t - \tau)I + \mathsf{1}_+(t - \tau)A(t)U(t, \tau)$$

shows that

$$E_\tau(t) = \mathsf{1}_+(t - \tau)U(t, \tau) \tag{12.12}$$

is the fundamental solution relative to τ of the above distributional equation and

$$y(t) = E_{t_0}(t)y_0 \tag{12.13}$$

is the solution of (12.10).

12.2.2 Formal Solution

We now look for an explicit formal solution in $\mathcal{D}'_+(\mathbb{R}, \mathbb{C}^n)$ of the equation

$$Dy = A(t)y + y_0\delta + x \tag{12.14}$$

with $A(.)$ a matrix of functions in $\mathcal{E}$ as before. As a first step we rewrite the equation as an integral equation. To do this we write Dy as $D\delta * y$ and convolve both sides of the equation with $\mathsf{1}_+$ to obtain

$$y - \mathsf{1}_+ * \big(A(t)y\big) = y_0 \mathsf{1}_+ + \mathsf{1}_+ * x\,.$$

Thus, if x is a bounded regular distribution then the equation can be written as

$$y(t) - \int_0^t A(\tau)y(\tau)\mathrm{d}\tau = y_0 \mathsf{1}_+(t) + \int_0^t x(\tau)\mathrm{d}\tau\,. \tag{12.15}$$

Instead of solving this equation directly, we consider a more general integral equation and then specialise to this case.

A *Volterra integral equation of the second kind* is an equation of the form

$$y(t) = \int_0^t k(t,\tau)y(\tau)\,\mathrm{d}\tau + x(t)\,, \qquad t \geq 0 \tag{12.16}$$

with x a given regular distribution in $\mathcal{D}'_+(\mathbb{R}, \mathbb{C}^n)$, k an $n \times n$ matrix of continuous functions $[k_{ij}]$, $i, j = 1, \ldots, n$ and y the required unknown in $\mathcal{D}'_+(\mathbb{R}, \mathbb{C}^n)$. This equation can be solved by an algebraic method based on a group [33, 34].

Definition 12.2 (*Group*) A group is a pair $(\mathcal{G}, \bullet)$ consisting of a non-empty set of objects $\mathcal{G}$ and a binary operation $\bullet$, usually called the group multiplication, satisfying the following properties

1. $\bullet$ is associative: $(g_1 \bullet g_2) \bullet g_3 = g_1 \bullet (g_2 \bullet g_3)$.
2. $\bullet$ has an identity element e: $g \bullet e = e \bullet g = g$.
3. Every element $g \in \mathcal{G}$ has an inverse element $g^{-1} \in \mathcal{G}$:

$$g \bullet g^{-1} = g^{-1} \bullet g = e.$$

Note that the unit element is unique, since if e' is a second unit $e = e \bullet e' = e'$ shows that it must be equal to the first one. A group $\mathcal{G}$ *acts (from the left) on* a non-empty set X if there is a function

$$\mathcal{G} \times X \to X\,, \qquad (g, x) \mapsto g \cdot x$$

such that the following hold:

1. $e \cdot x = x$ for all $x \in X$.
2. $g_1 \cdot (g_2 \cdot x) = (g_1 \bullet g_2) \cdot x$ for all $g_1, g_2 \in \mathcal{G}$ and $x \in X$.

Let now $k(t, \tau)$ be an $n \times n$ matrix of functions $[k_{ij}]$, $i, j = 1, \ldots, n$ continuous in the two variables t, τ, with $0 \leq \tau \leq t$ and x a locally bounded, locally integrable function in $\mathcal{D}'_+(\mathbb{R}, \mathbb{C}^n)$. That x is *locally bounded* means that it is bounded on every finite interval. We define the operation of $I + k$ on x by

$$(I + k) \cdot x := x(t) + \int_0^t k(t, \tau)x(\tau)\,\mathrm{d}\tau\,.$$

The resulting function is again a locally bounded, locally integrable function in $\mathcal{D}'_+(\mathbb{R}, \mathbb{C}^n)$ as x and the elements $I + k$ can be made to form a group. A suitable group multiplication can be found by writing

$$\begin{aligned}(I + k_1) \cdot [(I + k_2) \cdot x] = x(t) &+ \int_0^t k_2(t, \tau)x(\tau)\,\mathrm{d}\tau \\ &+ \int_0^t k_1(t, \tau)x(\tau)\,\mathrm{d}\tau + \int_0^t k_1(t, \tau_1) \int_0^{\tau_1} k_2(\tau_1, \tau_2)x(\tau_2)\,\mathrm{d}\tau_2\,\mathrm{d}\tau_1\end{aligned}$$

and noting that

$$\begin{aligned}&\int_0^t k_1(t, \tau_1) \int_0^{\tau_1} k_2(\tau_1, \tau_2)x(\tau_2)\,\mathrm{d}\tau_2\,\mathrm{d}\tau_1 \\ &\qquad = \int_0^t \int_{\tau_2}^t k_1(t, \tau_1)k_2(\tau_1, \tau_2)\,\mathrm{d}\tau_1\, x(\tau_2)\,\mathrm{d}\tau_2\,.\end{aligned}$$

Since the inner integral on the right-hand side results in a matrix of continuous functions, we can define the group multiplication by

$$(I + k_1) \bullet (I + k_2) := I + k_1 + k_2 + k_1 \star k_2$$

with

$$k_1 \star k_2(t, \tau) := \int_\tau^t k_1(t, \lambda)k_2(\lambda, \tau)\,\mathrm{d}\lambda\,. \tag{12.17}$$

For convenience we also put

$$k \star x(t) := k \star x(t, 0) := \int_0^t k(t, \tau)x(\tau)\mathrm{d}\tau \tag{12.18}$$

so that we can write

$$(I+k)\cdot x = x + k \star x\,.$$

The unit of the group is readily seen to be I.

It remains to show that every element of the group $I+k$ has an inverse $(I+k)^{-1}$. From the similarity with the geometric series we infer that the inverse is given by

$$(I+k)^{-1} := I + \sum_{n=1}^{\infty}(-1)^n k^{\star n} \tag{12.19}$$

and show that this series converges in every interval $0 \le \tau \le t \le T$. By definition, for every locally bounded function $x = (x_1, \dots, x_n)$ and every finite interval $0 \le t \le T$ we can find an upper bound given by

$$p_T(x) := \max_{1\le i\le n}\Big\{\sup_{0\le t\le T}|x_i(t)|\Big\}$$

so that, given the linearity of k,

$$p_T(k\star x) \le p_T(k)\,p_T(x)\,T$$

with

$$p_T(k) := \max_i \left\{\sum_{j=1}^{n}\sup_{0\le\tau\le t\le T}|k_{ij}(t,\tau)|\right\}\,.$$

Thus

$$p_T(k\star k) \le p_T(k)^2\,T$$

and by induction

$$p_T(k^{\star n}) \le p_T(k)^n\,\frac{T^{n-1}}{(n-1)!}\,.$$

This upper bound is the nth term of a convergent series and implies the converges of (12.19) for every value of T. Having established convergence one immediately verifies that indeed

$$(I+k)\bullet(I-k+k^{\star 2}\mp\cdots) = I$$

and

$$(I-k+k^{\star 2}\mp\cdots)\bullet(I+k) = I\,.$$

With this group the Volterra equation (12.16) can be written as

$$(I-k)\cdot y = x\,.$$

and is solved by multiplying on the left with $(I-k)^{-1}$

$$y(t) = x(t) + w \star x(t) \tag{12.20}$$

$$w := \sum_{n=1}^{\infty} k^{\star n}. \tag{12.21}$$

The matrix function w is called the *resolvent kernel* of the equation.

The group can't be extended to a ring or an algebra with the natural addition as these would include the elements k. These elements pose two problems. First, the inverse of these elements are not necessarily functions. For example, the inverse of $(t-\tau)^{m-1}/(m-1)!$ is $D^m\delta$ and for singular distributions multiplication is only defined with functions in $\mathcal{E}$. Second, such a ring includes zero divisors. From now on we will generally drop the symbols $\bullet$ of group multiplication and $\cdot$ of group operation as is commonly done with multiplication symbols.

We now come back to the special case of (12.15) for which

$$k(t,\tau) = A(\tau)\,.$$

The solution is given by (12.20) with

$$\begin{aligned}
k^{\star n} \star \Bigl(y_0 1_+(t) + \int_0^t x(\tau)\,\mathrm{d}\tau\Bigr) \\
&= \int_0^t \int_0^{\tau_1} \cdots \int_0^{\tau_{n-1}} A(\tau_1)\cdots A(\tau_n)\,\mathrm{d}\tau_n \cdots \mathrm{d}\tau_1\, y_0 \\
&+ \int_0^t \int_0^{\tau_1} \cdots \int_0^{\tau_{n-1}} A(\tau_1)\cdots A(\tau_n) \int_0^{\tau_n} x(\lambda)\,\mathrm{d}\lambda\,\mathrm{d}\tau_n \cdots \mathrm{d}\tau_1\,.
\end{aligned}$$

These expressions can be written more compactly by introducing the notion of a *time-ordered product* of operators. We define $T\{A_1(\tau_1)\cdots A_n(\tau_n)\}$ as the product with factors arranged from left to right in order of decreasing times. For example

$$T\{A_1(\tau_1)A_2(\tau_2)\} = \begin{cases} A_1(\tau_1)A_2(\tau_2) & \tau_1 \geq \tau_2 \\ A_2(\tau_2)A_1(\tau_1) & \tau_1 < \tau_2\,. \end{cases}$$

With this meta-operator we can now write

$$T\Bigl\{\Bigl(\int_0^t A(\tau)\,\mathrm{d}\tau\Bigr)^2\Bigr\} = \int_0^t \int_0^t T\{A(\tau_1)A(\tau_2)\}\,\mathrm{d}\tau_2\,\mathrm{d}\tau_1$$

$$= \int_0^t \int_0^{\tau_1} A(\tau_1)A(\tau_2)\,\mathrm{d}\tau_2\,\mathrm{d}\tau_1 + \int_0^t \int_{\tau_1}^{t} A(\tau_2)A(\tau_1)\,\mathrm{d}\tau_2\,\mathrm{d}\tau_1$$

$$= \int_0^t \int_0^{\tau_1} A(\tau_1)A(\tau_2)\,\mathrm{d}\tau_2\,\mathrm{d}\tau_1 + \int_0^t \int_0^{\tau_2} A(\tau_2)A(\tau_1)\,\mathrm{d}\tau_1\,\mathrm{d}\tau_2$$

$$= 2\int_0^t \int_0^{\tau_1} A(\tau_1)A(\tau_2)\,\mathrm{d}\tau_2\,\mathrm{d}\tau_1\,.$$

and more generally

$$T\left\{\left(\int_0^t A(\tau)\,\mathrm{d}\tau\right)^n\right\} = n!\int_0^t \cdots \int_0^{\tau_{n-1}} A(\tau_1)\cdots A(\tau_n)\,\mathrm{d}\tau_n \cdots \mathrm{d}\tau_1 \tag{12.22}$$

because there are $n!$ possible orderings of the n times $\tau_1, \ldots, \tau_n$. Using these expressions we have

$$k^{\star n} \star y_0 1_+(t) = \frac{1}{n!}T\left\{\left(\int_0^t A(\tau)\,\mathrm{d}\tau\right)^n\right\} y_0 \tag{12.23}$$

and

$$\begin{aligned} k^{\star n}\star &\int_0^t x(\tau)\,\mathrm{d}\tau \\ &= \int_0^t \int_0^{\lambda_1} \cdots \int_0^{\lambda_{n-1}} A(\lambda_1)\cdots A(\lambda_n) \int_0^{\lambda_n} x(\tau)\,\mathrm{d}\tau\,\mathrm{d}\lambda_n \cdots \mathrm{d}\lambda_1 \\ &= \int_0^t \int_\tau^t \int_\tau^{\lambda_1} \cdots \int_\tau^{\lambda_{n-1}} A(\lambda_1)\cdots A(\lambda_n)\,\mathrm{d}\lambda_n \cdots \mathrm{d}\lambda_1\, x(\tau)\,\mathrm{d}\tau \\ &= \frac{1}{n!}T\left\{\left(\int_\tau^t A(\lambda)\,\mathrm{d}\lambda\right)^n\right\} \star x\,. \end{aligned} \tag{12.24}$$

The solution of (12.15) can thus be written in the simple form

$$y(t) = E_0(t)y_0 + E_\tau(t) \star x(t) \tag{12.25}$$

with

$$E_\tau(t) = 1_+(t-\tau)T\{e^{\int_\tau^t A(\lambda)\,d\lambda}\} \tag{12.26}$$

the fundamental solution of the equation relative to τ and where we have made explicit the fact that for $t < \tau$ it is zero.

In the special case in which $A(.)$ commutes with $\int_\tau^t A(\lambda)\,d\lambda$ the time ordering operator has no effect and the solution of the equation is a direct generalisation of the solution obtained using the method of separation of the variables for the scalar equation

$$E_\tau(t) = 1_+(t-\tau)e^{\int_\tau^t A(\lambda)\,d\lambda}\,.$$

In particular this is the case if $A(.)$ is constant, in which case the fundamental solution becomes

$$E_\tau(t) = 1_+(t-\tau)e^{A(t-\tau)}\,, \qquad A \in \mathbb{C}^{n\times n}$$

and the expression for the solution y becomes a convolution identical to (8.11).

For this particular case it is interesting to observe that, for a small-time increment Δt, the evolution from an initial state y_0 can be approximated (to first order) by

$$y(\Delta t) \approx (I + A\Delta t)\cdot y_0$$

so that, by iteration

$$y(n\Delta t) \approx (I + A\Delta t)^{\bullet n}\cdot y_0\,.$$

Now if we set $\Delta t = t/n$ we obtain that, in the limit as n tends to infinity

$$\lim_{n\to\infty}\left(I + A\frac{t}{n}\right)^{\bullet n} = e^{At}\,.$$

The fundamental solution of (12.14) given by (12.26) can also be interpreted as a matrix function of the two variables t and τ

$$W(t,\tau) := 1_+(t-\tau)T\left\{e^{\int_\tau^t A(\lambda)\,d\lambda}\right\}\,.$$

As every element of the matrix is locally integrable, it is also a regular distribution that can be applied to test functions $\phi \in \mathcal{D}(\mathbb{R}^2)$. In particular, we can choose test functions of the form $\psi(t)x_j(\tau)$ with $\psi, x_j \in \mathcal{D}(\mathbb{R})$, $j = 1,\dots n$ in which case we obtain

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} W(t,\tau)\psi(t)x(\tau)\,dt\,d\tau = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} W(t,\tau)\psi(t)\,dt\,x(\tau)\,d\tau$$

with $x = (x_1, \dots, x_n)$. The inner integral on the right-hand side evaluates to a matrix of indefinitely differentiable functions in $\mathcal{E}$ [16, 35]. For this reason and remembering that every distribution f is the limit of a sequence of indefinitely differentiable functions (for example $f_m = f * \beta_m$ with β_m the test functions of Example (2.4)) we can extend $W \star x$ by continuity to operate on vector valued distributions in $\mathcal{E}'(\mathbb{R}, \mathbb{C}^n)$ by defining it as the distribution satisfying the system of equations

$$\langle (W \star x)_i, \psi \rangle = \sum_{j=1}^{n} \Big\langle x_j, \int_{-\infty}^{\infty} w_{ij}(t, \tau)\psi(t)\, dt \Big\rangle, \qquad i = 1, \dots, n\,. \tag{12.27}$$

The thus extended linear map $W\star$ is a distribution valued continuous function

$$W\star : \mathcal{E}'(\mathbb{R}, \mathbb{C}^n) \to \mathcal{D}'(\mathbb{R}, \mathbb{C}^n)\,.$$

With this definition we obtain for example that the solution of the equation with an input signal

$$x = y_0 \delta(t - t_0), \qquad y_0 \in \mathbb{C}^n$$

is

$$\langle \big(W \star y_0 \delta(t - t_0)\big)_i, \psi \rangle = \sum_{j=1}^{n} y_{0,j} \int_{-\infty}^{\infty} w_{ij}(t, t_0)\psi(t)\, dt$$

or

$$W \star y_0 \delta(t - t_0) = W(t, t_0) y_0\,.$$

This shows that the matrix W plays a similar role as the fundamental solution $E_\tau(t)$ and is called the (two-sided) *fundamental kernel* (or *elementary kernel*) of the differential operator $D - A(t)$. It also shows that, as with LTI systems, the initial conditions can be absorbed in the input vector signal x.

Example 12.1: Oscillator with Increasing Resonance [33]

Consider an ideal oscillator with a resonance frequency increasing with the square root of time

$$D^2 y + \omega_0^2\, t\, y = x \tag{12.28}$$

to which we apply an input signal

$$x = y_0 D\delta + y_1 \delta$$

corresponding to initial conditions $y(0) = y_0$ and $y'(0) = y_1$.

The equation can be rewritten in the state-space form by defining the state

$$u = \begin{bmatrix} y \\ Dy \end{bmatrix}$$

to obtain

$$Du = A(t)u + B\delta, \qquad y = Cu$$

with

$$A(t) = \begin{bmatrix} 0 & 1 \\ -\omega_0^2 t & 0 \end{bmatrix}, \qquad B = \begin{bmatrix} y_0 \\ y_1 \end{bmatrix}, \qquad C = \begin{bmatrix} 1 & 0 \end{bmatrix}.$$

In essence we need to calculate $W(t, 0)$. Using (12.23) and remembering (12.22) we have, for n even

$$A(\tau)^{\star n} \star 1_+(t) = \begin{bmatrix} -\frac{t^{3n/2}\omega_0^n}{\prod_{k=1}^n a_k} & 0 \\ 0 & -\frac{t^{3n/2}\omega_0^n}{\prod_{k=1}^{n-1} b_k} \end{bmatrix}, \qquad n \text{ even}$$

and for n odd

$$A(\tau)^{\star n} \star 1_+(t) = \begin{bmatrix} 0 & \frac{t^{3(n-1)/2+1}\omega_0^{n-1}}{\prod_{k=1}^{n-1} b_k} \\ \frac{t^{3(n+1)/2-1}\omega_0^{n+1}}{\prod_{k=1}^n a_k} & 0 \end{bmatrix}, \qquad n \text{ odd}$$

where $(a_k)_{k\geq 1}$ and $(b_k)_{k\geq 1}$ are the following sequences of integers

$$\begin{aligned} (a_k)_{k\geq 1} &:= (2, 3, 5, 6, 8, 9, 11, 12, \dots) \\ (b_k)_{k\geq 1} &:= (3, 4, 6, 7, 9, 10, 12, 13, \dots). \end{aligned}$$

The fundamental kernel at $(t, 0)$ is thus

$$\begin{aligned} W(t, 0) &= I + \sum_{n=1}^{\infty} A(\tau)^{\star n} \star 1_+(t) \\ &= \begin{bmatrix} 1 - \frac{\omega_0^2 t^3}{6} + \frac{\omega_0^4 t^6}{180} \mp \cdots & t - \frac{\omega_0^2 t^4}{12} \pm \cdots \\ -\frac{\omega_0^2 t^2}{2} + \frac{\omega_0^4 t^5}{30} \mp \cdots & 1 - \frac{\omega_0^2 t^3}{3} + \frac{\omega_0^4 t^6}{72} \mp \cdots \end{bmatrix}. \end{aligned}$$

The series can be recognised as linear combinations of the Airy Ai and Bi functions and their derivatives Ai' and Bi'

$$W(t, 0) = \begin{bmatrix} w_0(t) & w_1(t) \end{bmatrix}$$

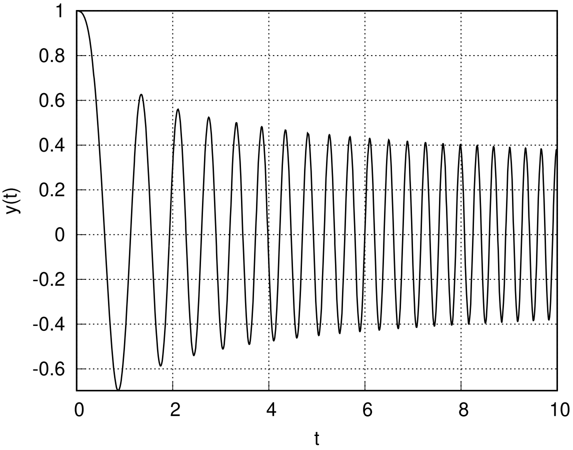


Fig. 12.1 Solutions of (12.28) for $y_0 = 1$, $y_1 = 0$ and $\omega_0 = 2\pi$

with

$$w_0(t) = \frac{3^{1/6}\Gamma(2/3)}{2}\begin{bmatrix} (\sqrt{3}\mathsf{Ai}(-t\omega_0^{2/3}) + \mathsf{Bi}(-t\omega_0^{2/3})) \\ -\omega_0^{2/3}(\sqrt{3}\mathsf{Ai}'(-t\omega_0^{2/3}) + \mathsf{Bi}'(-t\omega_0^{2/3})) \end{bmatrix}$$

and

$$w_1(t) = \frac{\Gamma(1/3)}{2\cdot 3^{2/3}}\begin{bmatrix} \frac{3\mathsf{Ai}(-t\omega_0^{2/3}) - \sqrt{3}\mathsf{Bi}(-t\omega_0^{2/3})}{\omega_0^{2/3}} \\ -3\mathsf{Ai}'(-t\omega_0^{2/3}) + \sqrt{3}\mathsf{Bi}'(-t\omega_0^{2/3}) \end{bmatrix}.$$

The signal of interest y is thus given by

$$y(t) = CW(t,0)B.$$

Specifically, for $y_0 = 1$ and $y_1 = 0$ (see Fig. 12.1)

$$y(t) = \frac{3^{1/6}\Gamma(2/3)}{2}\left(\sqrt{3}\mathsf{Ai}(-t\omega_0^{2/3}) + \mathsf{Bi}(-t\omega_0^{2/3})\right).$$

The full fundamental kernel $W(t,\tau)$ can be obtained using Eqs. (12.9) and (12.12) and computing the inverse of $W(t,0)$

$$W(t,\tau) = 1_+(t-\tau)W(t,0)[W(\tau,0)]^{-1}.$$

12.2.3 Perturbation Theory

The solution of (12.14) presented above is of great theoretical value. However, when it comes to solving practical problems it is in general very difficult to find a closed form for the fundamental kernel $W(t, \tau)$. In many situations the problem at hand looks similar to a solvable problem, but with additional terms. If those terms are small in comparison to the ones of the solvable problem then one can obtain a good approximation to the solution of the problem by the following perturbative method.

Suppose that the matrix $A(.)$ can be split in two parts: one that leads to a solvable problem and that we denote by $A_0(.)$ and one with relatively small elements, the perturbation term, that makes the equation unsolvable and that we denote by $\tilde{A}(.)$

$$Dy = [A_0(t) + \tilde{A}(t)]y + x\,.$$

Let $W_0(t, \tau)$ be the fundamental kernel of the solvable part of the equation, $Y_0(.)$ its principal fundamental matrix, that is, the solution of the matrix equation

$$DY_0(t) = A_0(t)Y_0(t)\,, \qquad Y_0(0) = I\,.$$

and let express y in terms of a new vector $\tilde{y}$ defined by

$$y = Y_0(t)\tilde{y}\,.$$

Then the equation becomes

$$DY_0(t)\tilde{y} + Y_0(t)D\tilde{y} = [A_0(t) + \tilde{A}(t)]Y_0(t)\tilde{y} + x$$

which reduces to

$$D\tilde{y} = Q(t)\tilde{y} + Y_0^{-1}(t)x$$

with

$$Q(t) := Y_0^{-1}(t)\tilde{A}(t)Y_0(t)\,.$$

This equation has the same form as the original one. Its solution is therefore given by

$$\tilde{y}(t) = \mathsf{1}_+(t-\tau)T\left\{\mathrm{e}^{\int_\tau^t Q(\lambda)\,\mathrm{d}\lambda}\right\} \star \left[Y_0^{-1}(t)x(t)\right].$$

The advantage that we gain is the fact that, if the elements of $\tilde{A}(.)$ are small, then the series expansion of this solution converges very quickly. Differently from this, to obtain a good approximation using the series of the original formulation of the problem requires a large number of terms (compare with Example 12.1).

If x is composed by regular distributions then the first terms of the solution of the equation are given by

$$\begin{aligned}y(t) &= Y_0(t)\int_0^t Y_0^{-1}(\lambda)x(\lambda)\,\mathrm{d}\lambda\\ &\quad + Y_0(t)\int_0^t\int_0^\tau Q(\tau)Y_0^{-1}(\lambda)x(\lambda)\,\mathrm{d}\lambda\,\mathrm{d}\tau + \cdots\\ &= \int_0^t W_0(t,\lambda)x(\lambda)\,\mathrm{d}\lambda + \int_0^t\int_0^\tau W_0(t,\tau)\tilde{A}(\tau)W_0(\tau,\lambda)x(\lambda)\,\mathrm{d}\lambda\,\mathrm{d}\tau + \cdots .\end{aligned}$$

The first term that we denote by y_0 is the solution of the unperturbed equation. In general, it is given by

$$y_0 = W_0 \star x .$$

The next term is the first order perturbation term that we denote by y_1. Note that it can be expressed as the action of the unperturbed system on an input signal x_1 constructed by multiplying y_0 by the perturbation $\tilde{A}$

$$y_1 = W_0 \star x_1, \qquad x_1(t) = \tilde{A}(t)y_0(t) .$$

Similarly, the nth order perturbation term can be represented as the action of the unperturbed system on an input signal obtained by multiplying the perturbation term of order $n-1$ by $\tilde{A}$

$$y_n = W_0 \star x_n, \qquad x_n(t) = \tilde{A}(t)y_{n-1}(t) .$$

The output of the system

$$y(t) = \sum_{n=0}^{\infty} y_n(t)$$

can thus be calculated iteratively starting from the response of the unperturbed system where each successive term is the result of multiplying the output of the previous term by $\tilde{A}$ and feeding it back as input of the unperturbed system. This reminds of a feedback system with the unperturbed system playing the role of the forward path and $\tilde{A}$ of the feedback one.

12.2.4 Non-smooth Coefficients

For several applications the requirement of differential operators with indefinitely differentiable coefficients is too restrictive. In those situations it's useful to work in the subspace of $\mathcal{D}'$ constituted by distributions that are m times differentiable and denoted by $\mathcal{D}'^m$. These distributions are said to be of *order m* and are the continuous linear functionals on the set of m times continuously differentiable functions with compact support $\mathcal{D}^m$. Convergence is defined in a similar way as for distributions in $\mathcal{D}'$.

Given a distribution $f \in \mathcal{D}'^m$ the product of f with an m times continuously differentiable function g is well-defined

$$\langle fg, \phi \rangle = \langle f, g\phi \rangle$$

since if $\phi \in \mathcal{D}^m$ then $g\phi$ is also in $\mathcal{D}^m$. Note that we can exchange the roles of f and g and still obtain a well-defined multiplication. Thus if f is an m times continuously differentiable function, it can be multiplied by a distribution of order m.

Example 12.2: Dirac Distribution

The Dirac distribution δ belongs to $\mathcal{D}'^0$ and its multiplication with continuous functions is well-defined as long as one restricts considerations to $\mathcal{D}'^m$.

12.3 Impulse Response Generalisation

In the previous section we saw that differential equations describing LTV systems aren't convolution equations. In spite of this we found that the solution of the equation can be written with the help of the operator $\star$ acting on a (matrix) function characterising the system (the fundamental kernel) and the input vector x. In particular, for a system described by the state-space representation (12.3)–(12.4) with $D(t) = 0$, the input-output characteristic is given by

$$\begin{aligned} y(t) &= C(t)\mathbb{1}_+(t-\tau)T\{\mathrm{e}^{\int_\tau^t A(\lambda)\,\mathrm{d}\lambda}\} \star B(t)x(t) \\ &= C(t)\mathbb{1}_+(t-\tau)T\{\mathrm{e}^{\int_\tau^t A(\lambda)\,\mathrm{d}\lambda}\}B(\tau) \star x(t)\,. \end{aligned}$$

This expression highlights how in LTV systems the operator $\star$ is the natural operator taking the place of convolution in LTI systems. However, because the Fourier- and Laplace-transform convert convolutions into products and because for many purposes the frequency domain characteristics of a system are more interesting than the time domain ones, in engineering circles it is common to express the input-output characteristic of LTV systems in terms of a convolution like operator as we did in Sect. 12.1. This is easily done by the change of variable

$$\xi = t - \tau$$

and by defining the time-varying impulse response $h(t, \xi)$ as a function of the variables t and ξ

$$y(t) = \int_0^t h(t, t-\xi)\, x(\xi)\, d\xi.$$

$$h(t, \xi) = C(t) 1_+(\xi) T \left\{ \mathrm{e}^{\int_{t-\xi}^{t} A(\lambda)\, \mathrm{d}\lambda} \right\} B(t-\xi)$$

where we have assumed x to be a regular distribution in $\mathcal{D}'_+$. Note that, while the above integral looks very similar to a convolution, it differs from the convolution that we defined in Sect. 3.2. A generalisation of the above convolution like operation for LTV systems is obtained by adapting (12.27) and defining it as the distribution satisfying the following equality

$$\langle h *_t x, \phi \rangle = \langle x(\xi), \langle h(t, t-\xi), \phi(t) \rangle \rangle \tag{12.29}$$

where we have generalised the inner integral of (12.27) to the application of a parameterised distribution to the test function ϕ. Since this operation shares several properties with convolution, the operator $*_t$ is called the *convolution product for time-varying systems*. In the technical literature it is most often called convolution and denoted by the same symbol as the one used for convolution. In the following we will also often simply call it convolution, but maintain the use of the symbol $*_t$ to make it clear that it is not the operation defined by (3.6).

In the special case in which the time-varying impulse response is the product of an indefinitely differentiable function f and a distribution g

$$h(t, \xi) = f(t) g(\xi)$$

the convolution product for time-varying systems can be expressed in terms of a proper convolution by

$$\langle h *_t x, \phi \rangle = \langle g * x, f\phi \rangle\,.$$

In the previous section we discussed the fact that, for systems described by a differential equation, the application $\langle h(t, t-\xi), \phi(t) \rangle$ appearing on the right-hand side of (12.29), regarded as a function of the parameter ξ, is a function belonging to $\mathcal{E}$. For this reason, for the equation to have a meaning, x must be restricted to distributions in $\mathcal{E}'$. However, if we define a function $\gamma \in \mathcal{E}$ bounded from the left with $\gamma(t) = 1$ in a neighbourhood of $[0, \infty)$ and assume h to be such that

$$\xi \mapsto \gamma(\xi) \langle h(t, t-\xi), \phi(t) \rangle$$

is a Schwartz function for every $\phi \in \mathcal{S}$, then (12.29) remains valid for right-sided tempered distributions

$$x \in \mathcal{S}' \cap \mathcal{D}'_+\,.$$

Note the similarity with the definition of the Laplace transform and the fact that, as for the Laplace transform, the value of the distribution does not depend on the choice of γ. For this reason and as is commonly done for the Laplace-transform, we will generally not write γ explicitly.

Before concluding this section we note some properties of the operator $*_t$. The first is that it is associative

$$(h_B *_t h_A) *_t x = h_B *_t (h_A *_t x)$$

with h_A and h_B the time-varying impulse responses of two systems. This is a direct consequence of the fact that $*_t$ is related to $\star$ by a simple variable transformation and by the definition of the latter (see Eqs. (12.17) and (12.18)).

A second important property, or rather the lack of it, is that, $*_t$ is not commutative. Therefore, differently from LTI systems, the order of LTV systems is important. As an example consider the cascade of a low-pass filter with a 3 dB cut-off frequency of ω_{3dB} and the frequency shifting system of Example 12.4 with $w_0 \gg \omega_{3dB}$. Suppose that the system is driven by a signal with a frequency falling in the pass-band of the LPF. Then if the signal passes first through the LPF and then into the frequency shifting system, the output will have a large magnitude. Differently from this, if the input signal first passes through the frequency translating system then the signal at the input of the LPF will lie in the stop-band of the latter and will appear much attenuated at its output.

12.4 Time-Varying Frequency Response

12.4.1 Definition

Consider a system described by the time-varying impulse response $h(t, \xi)$. Under the assumption that the input signal x is a right-sided tempered distribution the system response y can be written as

$$\begin{aligned}
\langle y(t), \phi(t)\rangle &= \left\langle \mathcal{F}^{-1}\{\mathcal{F}\{x\}\}, \int_{-\infty}^{\infty} h(t, t-\xi)\phi(t)\,dt \right\rangle \\
&= \left\langle \hat{x}(\omega), \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(t, t-\xi)\phi(t)\,dt\, \mathrm{e}^{\jmath\omega\xi}\,d\xi \right\rangle \\
&= \left\langle \hat{x}(\omega), \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(t, t-\xi)\mathrm{e}^{-\jmath\omega(t-\xi)}\,d\xi\, \mathrm{e}^{\jmath\omega t}\,\phi(t)\,dt \right\rangle
\end{aligned}$$

$$= \left\langle \hat{x}(\omega), \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{h}(t,\omega)\, \mathrm{e}^{\jmath\omega t}\, \phi(t)\, dt \right\rangle$$

$$= \left\langle \frac{1}{2\pi} \hat{h}(t,\omega)\, \mathrm{e}^{\jmath\omega t} \star \hat{x}(t), \phi(t) \right\rangle$$

or

$$y(t) = \frac{1}{2\pi} \hat{h}(t,\omega)\, \mathrm{e}^{\jmath\omega t} \star \hat{x}(t) \tag{12.30}$$

where $\hat{h}(t,\omega)$ is the Fourier transform with respect to ξ of $h(t,\xi)$ and is called the *time-varying frequency response* of the system

$$\hat{h}(t,\omega) := \int_{-\infty}^{\infty} h(t,\xi) \mathrm{e}^{-\jmath\omega\xi}\, d\xi\,. \tag{12.31}$$

In particular, for regular distributions we have

$$y(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{h}(t,\omega) \hat{x}(\omega)\, \mathrm{e}^{\jmath\omega t}\, d\omega\,.$$

It's easy to check that for real systems the time-varying frequency response at $-\omega$ is equal to the conjugate complex of the value at ω

$$\hat{h}(t,-\omega) = \overline{\hat{h}}(t,\omega)$$

for each value of t.

To obtain a physical interpretation for $\hat{h}(t,\omega)$ we apply a complex tone $\mathrm{e}^{\jmath\omega_0 t}$ as input signal. This is allowed because periodic distributions are isomorphic to distributions with compact support (see Sect. 3.4). With this input signal the output of the system is found with the help of (12.30) to be

$$y(t) = \hat{h}(t,\omega_0) \mathrm{e}^{\jmath\omega_0 t}$$

and suggests the interpretation for the time-varying frequency response $\hat{h}(t,\omega)$ as the complex envelope at ω_0 of the output signal (see Fig. 12.2).

If the output signal y is a tempered distribution it can be Fourier transformed. A useful expression relating the spectrum of y and the one of the input signal x can be obtained by expressing y with the help of (12.30)

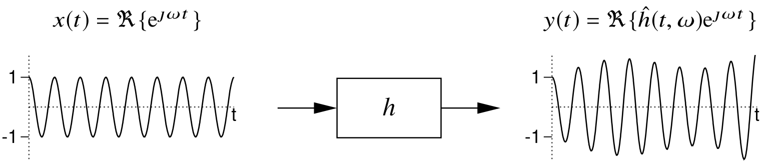

$x(t) = \Re\{e^{\jmath\omega t}\}$
1
-1
t
h
$y(t) = \Re\{\hat{h}(t,\omega)e^{\jmath\omega t}\}$
1
-1
t

Fig. 12.2 Illustrative representation of the response of a real LTV system to an input tone

$$\begin{aligned}
\langle \hat{y}, \phi \rangle &= \left\langle \mathcal{F}\{\frac{1}{2\pi}\hat{h}(t,\omega)\, e^{\jmath\omega t} \star \hat{x}\}, \phi(t) \right\rangle \\
&= \left\langle \hat{x}(\omega), \frac{1}{2\pi} \int\limits_{-\infty}^{\infty} \hat{h}(t,\omega)\, e^{\jmath\omega t} \hat{\phi}(t)\, dt \right\rangle \\
&= \left\langle \hat{x}(\omega), \frac{1}{2\pi} \int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} \hat{h}(t,\omega)\, e^{-\jmath(w-\omega)t}\, dt\, \phi(w)\, dw \right\rangle \\
&= \left\langle \hat{x}(\omega), \frac{1}{2\pi} \int\limits_{-\infty}^{\infty} \hat{\hat{h}}(w-\omega,\omega)\, \phi(w)\, dw \right\rangle \\
&= \left\langle \frac{1}{2\pi}\hat{\hat{h}}(w-\omega,\omega) \star \hat{x}(w), \phi(w) \right\rangle
\end{aligned}$$

or

$$\hat{y}(w) = \frac{1}{2\pi}\hat{\hat{h}}(w-\omega,\omega) \star \hat{x}(w) \tag{12.32}$$

with

$$\hat{\hat{h}}(w,\omega) := \int\limits_{-\infty}^{\infty} \hat{h}(t,\omega)\, e^{-\jmath wt}\, dt\,. \tag{12.33}$$

The function $\hat{\hat{h}}$ is the two-dimensional Fourier transform of the time-varying impulse response h and, in the context of communication systems, is called the *doppler-spread function.* Equation (12.32) shows that the input and output spectra of an LTV system are related by a convolution like operation. In particular, for regular distributions they are related by the following integral

$$\hat{y}(w) = \frac{1}{2\pi} \int\limits_{-\infty}^{\infty} \hat{\hat{h}}(w-\omega,\omega)\hat{x}(\omega)\, d\omega\,.$$

For tempered distributions the time-varying impulse response h, the time-varying frequency response $\hat{h}$ and the doppler-spread function $\hat{\hat{h}}$ are isomorphic to each other.

For this reason an LTV system with a tempered time-varying impulse response can be described by any of these functions.

Example 12.3

In this example we investigate the relationship between an LTI system to which we apply a right-sided input tone and an LTV system activated at $t = 0\,\mathrm{s}$ and driven by a tone, resulting in equal output signals.

Consider an LTI system described by the differential equation

$$Dy + ay = x\,, \qquad a > 0$$

to which we apply the signal $x(t) = 1_+(t)\mathrm{e}^{\jmath\omega t} \in \mathcal{D}'_+$. The response of the system can be calculated with the help of the Laplace transform. The transfer function of the system and the Laplace transformed of the input signals are

$$H(s) = \frac{1}{s + a}\,, \qquad \Re\{s\} > -a$$

and

$$X(s) = \frac{1}{s - \jmath\omega}\,, \qquad \Re\{s\} > 0$$

respectively. The system response is thus found by inverse Laplace transforming

$$Y(s) = H(s)X(s) = \frac{1}{(s + a)(s - \jmath\omega)}\,, \qquad \Re\{s\} > 0$$

which gives

$$y(t) = \frac{\mathrm{e}^{-at}}{a + \jmath\omega}\Big(\mathrm{e}^{(a+\jmath\omega)t} - 1\Big)\,.$$

We now re-interpret the system as a time-variable one consisting of the above LTI system and an ideal switch at its input. For $t < 0$ the input is disconnected from the system (switch open) which therefore produces the constant output signal $y(t) = 0$. At $t = 0$ the input signal is connected to the input of the LTI system by closing the switch. The full system is therefore described by the differential equation

$$Dy + ay = 1_+(t)x\,.$$

The input signal is now the complex tone $x(t) = \mathrm{e}^{\jmath\omega t}$.

To obtain the system response we first compute the time evolution operator $U(t, \tau)$ which is the solution of

$$Dy + ay = \delta(t - \tau)\,, \qquad t \ge \tau > 0$$

and given by

$$U(t,\tau) = \mathsf{1}_+(t)\mathsf{1}_+(t-\tau)\mathrm{e}^{-a(t-\tau)}\,.$$

With it the response of the system to the input $x(t) = \mathrm{e}^{\jmath\omega t}$ is calculated to be

$$\begin{aligned} y(t) = U(t,\tau) \star x(t) &= \int_0^t \mathrm{e}^{-a(t-\tau)}\mathrm{e}^{\jmath\omega\tau}\mathrm{d}\tau \\ &= \frac{\mathrm{e}^{-at}}{a+\jmath\omega}\Big(\mathrm{e}^{(a+\jmath\omega)t} - 1\Big) \end{aligned}$$

which of course agrees with the calculation through Laplace transform. However, with the new interpretation we see that the system posses a time-varying frequency response $\hat{h}(t,\omega)$. The easiest way to calculate it is through the relation $y(t) = \hat{h}(t,\omega)\mathrm{e}^{\jmath\omega t}$ and we obtain

$$\hat{h}(t,\omega) = \frac{\mathrm{e}^{-(a+\jmath\omega)t}}{a+\jmath\omega}\Big(\mathrm{e}^{(a+\jmath\omega)t} - 1\Big)\,.$$

This shows the relationship between $\hat{h}(t,\omega)$ and the LTI frequency response $H(\jmath\omega) = 1/(a+\jmath\omega)$. Differently from the latter, $\hat{h}(t,\omega)$ includes the full information about the variation in time of the system. In this particular example, about when the switch is closed.

Example 12.4: Frequency Translation

Consider a system described by the doppler-spread function

$$\hat{\hat{h}}(w,\omega) = 2\pi\delta(w - w_0)\,.$$

According to (12.32) the spectrum of the output signal is given by

$$\hat{y}(w) = \delta(w - w_0 - \omega) \star \hat{x}(w) = \hat{x}(w - w_0)\,.$$

Therefore, the effect of the system described by the above doppler-spread function is to shift in frequency the spectrum of the input signal by w_0. Such a device is referred to as a *mixer*.

The time-varying frequency response and the time-varying impulse response corresponding to this delay-spread function are easily calculated to be

$$\hat{h}(t,\omega) = \mathrm{e}^{\jmath w_0 t}$$

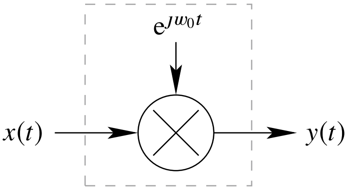


Fig. 12.3 Block diagram of a frequency-translating LTV system

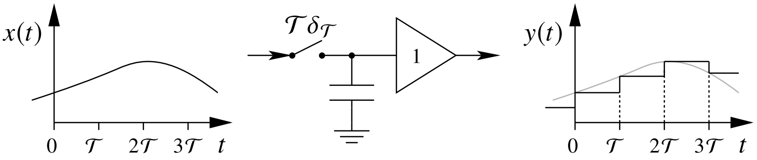


Fig. 12.4 Ideal sample and hold

and

$$h(t,\xi) = e^{\jmath w_0 t}\delta(\xi)$$

respectively. If we apply a complex tone $e^{\jmath\omega_0 t}$ as input signal we can calculate the output signal from the former and (12.30) as

$$y(t) = \frac{1}{2\pi} e^{\jmath(w_0+\omega)t} \star 2\pi\delta(t-\omega_0) = e^{\jmath(w_0+\omega_0)t}$$

or from the latter and (12.29) as

$$y(t) = e^{\jmath w_0 t}\delta(\xi) *_t e^{\jmath\omega_0 t} = e^{\jmath(w_0+\omega_0)t} .$$

In both cases the angular frequency of the input tone is shifted by w_0 as expected.

The time-varying impulse response shows clearly that the system is memory-less, that is, the value of the output signal at time t only depends on the input signal at time t. The effect of the system is to simply multiply the input signal by the complex tone $e^{\jmath w_0 t}$ as illustrated in Fig. 12.3.

Example 12.5: Sample and Hold

In this example we consider an ideal sample and hold: the output of the system is constructed by sampling the input signal x at regular intervals $\mathcal{T}$ and by holding the value of each sample constant for the duration of a period $\mathcal{T}$. Sample and hold blocks are used for example at the input of analog-to-digital converters (ADC) to give the converter enough time to compare the value of a sample with one or more reference signal levels. The operation of a sample and hold is illustrated in Fig. 12.4.

The ideal sample and hold is characterised by the following time-varying impulse response

$$h(t,\xi) = \delta_{\mathcal{T}}(t-\xi)1_{\mathcal{T}}(\xi) = \sum_{n=-\infty}^{\infty} \delta(t-\xi-n\mathcal{T})1_{\mathcal{T}}(\xi)$$

with

$$1_{\mathcal{T}}(\xi) = \begin{cases} 1 & 0 \leq \xi < \mathcal{T} \\ 0 & \text{otherwise}\,. \end{cases}$$

Note that in this case (12.29) doesn't make sense as h is a singular distribution and in the right-hand side expression x is not applied to a smooth function. To give a meaning to

$$y(t) = h(t,\xi) *_t x(t)$$

we have to restrict the input signal x to belong to $\mathcal{E}$. Then we can write

$$\begin{aligned} \langle y, \phi \rangle &= \langle h(t,\xi) *_t x(t), \phi(t) \rangle \\ &= \sum_{n=-\infty}^{\infty} \left\langle x(\xi)\delta(\xi - n\mathcal{T}), \int_{\xi}^{\xi+\mathcal{T}} \phi(t)\,dt \right\rangle \\ &= \sum_{n=-\infty}^{\infty} x(nT)\langle 1_{\mathcal{T}}(t-n\mathcal{T}), \phi(t) \rangle \end{aligned}$$

or

$$y(t) = \sum_{n=-\infty}^{\infty} x(nT)1_{\mathcal{T}}(t-n\mathcal{T})$$

and we obtain the desired system response. The system response can also be written as a (proper) convolution

$$y(t) = \mathcal{T}\delta_{\mathcal{T}}(t)x(t) * \frac{1}{\mathcal{T}}1_{\mathcal{T}}(t)\,.$$

From this expression, assuming x to be Fourier transformable, it's easy to compute the output spectrum. From (4.14) we read that the Fourier transform of $\mathcal{T}\delta_{\mathcal{T}}\,x$ is the convolution of the transforms of the factors divided by 2π

$$\mathcal{F}\{\mathcal{T}\delta_{\mathcal{T}}\,x\} = \delta_{\omega_s} * \hat{x}$$

with ω_s the sampling angular frequency $2\pi/\mathcal{T}$. Thus, the output spectrum is

$$\hat{y}(\omega) = [\delta_{\omega_s} * \hat{x}(\omega)]\frac{1}{\mathcal{T}}\hat{1}_{\mathcal{T}}(\omega)$$

with

$$\hat{1}_{\mathcal{T}}(\omega) = \mathcal{T}\frac{\sin \pi \frac{\omega}{\omega_s}}{\pi \frac{\omega}{\omega_s}}\mathrm{e}^{-\jmath\omega\frac{\mathcal{T}}{2}}\,.$$

This expression shows clearly the effects of sampling and of holding in the frequency domain. The operation of sampling is represented by the factor in square brackets. Its effect is to produce an infinite number of copies of the spectrum of the input signal shifted by multiples of ω_s

$$\delta_{\omega_s} * \hat{x}(\omega) = \sum_{n=-\infty}^{\infty} \hat{x}(\omega - n\omega_s)\,.$$

If the original signal has to be recovered from the samples then one must avoid (or reduce to negligible levels) overlapping between the copies. This amount to saying that the power of the input signal residing outside the frequency range $(-\omega_s/2, \omega_s/2)$ must be negligible. Or, in other words, the sampling frequency must be at least twice the frequency of the highest component of the input signal spectrum containing a non-negligible amount of power. This is the statement of the famous *sampling theorem*. If this condition is satisfied then the input signal can be recovered with the help of a low-pass-filter eliminating the copies with $n \neq 0$. When the copies of the input signal do overlap one says that sampling causes *aliasing*. Note that, if the spectrum of the input signal x only occupies a small fraction of the frequency range $(-\omega_s/2, \omega_s/2)$ then one may find a sampling frequency lower than ω_s not causing aliasing.

The effect of holding act as an LTI filter introducing a delay of $\mathcal{T}/2$. The filter has a low-pass characteristic with notches at multiples of ω_s. The effects of sampling and of holding on the spectrum on a signal are illustrated in Fig. 12.5.

The need to restrict x to being an indefinitely differentiable function may seem like excess of rigor. Note however that if x is not continuous at the sample instants $n\mathcal{T}$ then the problem is not "merely" a mathematical one, but any physical implementation will fail to work properly. This is so because if the input signal varies very rapidly compared to the actual speed of the physical sampling switch, then the value of the sample will be affected by many implementation details and in particular by noise. The result is a system producing unpredictable sample values.

From a mathematical point of view one may enlarge the type of allowed input signals to the class of continuous functions. Then the system response is mathematically well-defined, but it's not a distribution anymore. In fact, the value of a Dirac impulse is defined as the value of the test function at zero. If we multiply the test function with a continuous function, the value is still well-defined. However, we can't expect to be able to compute the derivatives of the output signal. Compare also with Sect. 12.2.4.

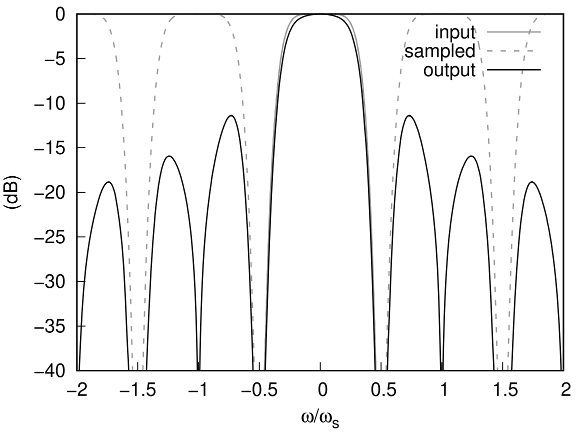


Fig. 12.5 Illustration of the effect of sampling and of a sample and hold on the spectrum of a signal

We started this section by performing a calculation leading to the definition of the time-varying frequency response of a system and a relation expressing the output of the system in terms of it. If we assume Laplace transformable, right-sided signals and redo a similar calculation replacing the Fourier transform by the Laplace one we obtain the *time-varying transfer function* of the system

$$H(t,s) = \int_0^\infty h(t,\xi)\mathrm{e}^{-s\xi}\,d\xi \qquad \Re\{s\} > \sigma\,. \tag{12.34}$$

With it the output of the system is given by

$$y(t) = \frac{1}{2\pi\jmath} H(t,s)\mathrm{e}^{st} \star X(s)\,. \tag{12.35}$$

12.4.2 Differential Equation

Consider again a linear time-varying system whose state u is described by the system of differential equations

$$Du = A(t)u + B(t)x$$

and assume that it is driven by a complex tone

$$x(t) = \mathrm{e}^{\jmath\omega t}\,.$$

From (12.30) we know that the components of the state u can be represented by

$$u_i(t) = \hat{u}_i(t,\omega)\mathrm{e}^{\jmath\omega t}\,, \qquad i = 1, \dots, n\,.$$

Inserting this representation for u and the complex tone for x in the equation we obtain

$$D\hat{u}(t,\omega)\mathrm{e}^{\jmath\omega t} = A(t)\hat{u}(t,\omega)\mathrm{e}^{\jmath\omega t} + B(t)\mathrm{e}^{\jmath\omega t}\,.$$

The left-hand side can be written as

$$D\hat{u}(t,\omega)\mathrm{e}^{\jmath\omega t} = \mathrm{e}^{\jmath\omega t}(\jmath\omega + D)\hat{u}(t,\omega)$$

so that we obtain an equation for $\hat{u}(t,\omega)$

$$(\jmath\omega + D)\hat{u}(t,\omega) = A(t)\hat{u}(t,\omega) + B(t)\,. \tag{12.36}$$

With $\hat{u}(t,\omega)$ we can directly obtain the time-varying frequency response of the system without having to first compute the fundamental kernel

$$\hat{h}(t,\omega) = C(t)\hat{u}(t,\omega) + D(t)\,.$$

In particular, if the system is described by a (possibly) higher-order differential equation

$$L(t,D)y = N(t,D)x$$

with

$$\begin{aligned} L(t,D) &= D^m + a_{m-1}(t)D^{m-1} + \cdots + a_0(t),\\ N(t,D) &= b_n(t)D^n + b_{n-1}(t)D^{n-1} + \cdots + b_0(t) \end{aligned}$$

we can directly obtain an equation for the time-varying frequency response of the system by replacing the differential operator D in L by the operator $\jmath\omega + D$ and in N by $\jmath\omega$ [36]

$$L(t, \jmath\omega + D)\hat{h}(t,\omega) = N(t, \jmath\omega)\,. \tag{12.37}$$

Note that this formulation in terms of distributions and distributional derivatives takes care of the initial conditions automatically. If one works with functions and the standard derivative then the initial conditions for the problem are obtained from

$$y(t) = \hat{h}(t,\omega)\mathrm{e}^{\jmath\omega t}\,.$$

Example 12.6

Consider a system that is switched off up to time $t = 0$ ($y(t) = 0, t < 0$) at which point it is turned on and is then described by the differential equation

$$Dy + ty = x\,.$$

We are interested in the time-varying frequency response of the system. We compute it in three different ways.

First we compute it via the time evolution operator U. For $t \geq \tau > 0$ it is found by solving the differential equation

$$Dy + ty = \delta(t - \tau)\,.$$

As can be verified by inserting it into the equation, it is given by

$$U(t, \tau) = \mathrm{e}^{-t^2/2+\tau^2/2}\,.$$

To obtain the time-varying frequency response we apply the input $x(t) = \mathrm{e}^{\jmath\omega t}$ and obtain

$$y(t) = \int_0^t U(t, \tau)x(\tau)\mathrm{d}\tau = \mathrm{e}^{-t^2/2} \int_0^t \mathrm{e}^{\tau^2/2+\jmath\omega\tau}\mathrm{d}\tau.$$

From this and

$$y(t) = \hat{h}(t, \omega)\mathrm{e}^{\jmath\omega t}$$

we deduce that

$$\hat{h}(t, \omega) = \mathrm{e}^{-t^2/2-\jmath\omega t} \int_0^t \mathrm{e}^{\tau^2/2+\jmath\omega\tau}\mathrm{d}\tau\,.$$

The time-varying frequency response of the system can also be obtained by Fourier transforming the time-varying impulse response. The latter is obtained from the time evolution operator using the variable substitution $\xi = t - \tau$

$$h(t, \xi) = 1_+(\xi)1_+(t)\mathrm{e}^{-t^2/2+(t-\xi)^2/2}$$

where we made explicit that for $t < 0$ the response of the system vanishes. The time-varying frequency response is thus

$$\begin{aligned}\hat{h}(t,\omega) &= \int_{-\infty}^{\infty} h(t,\xi)\mathrm{e}^{-\jmath\omega\xi}d\xi = \mathrm{e}^{-t^2/2}\int_0^t \mathrm{e}^{(t-\xi)^2/2}\mathrm{e}^{-\jmath\omega\xi}d\xi\\ &= \mathrm{e}^{-t^2/2}\int_0^t \mathrm{e}^{\tau^2/2}\mathrm{e}^{-\jmath\omega(t-\tau)}\mathrm{d}\tau = \mathrm{e}^{-t^2/2-\jmath\omega t}\int_0^t \mathrm{e}^{\tau^2/2+\jmath\omega\tau}\mathrm{d}\tau\end{aligned}$$

which matches the one obtained with the previous method.

A third method to compute the time-varying frequency response is by solving the corresponding differential equation

$$(D+\jmath\omega)\hat{h} + t\hat{h} = 1_{+}(t)\,.$$

The solution is

$$\hat{h}(t,\omega) = 1_{+}(t)\mathrm{e}^{-t^2/2-\jmath\omega t}\int_0^t \mathrm{e}^{\tau^2/2+\jmath\omega\tau}\mathrm{d}\tau\,.$$

as is verified by inserting it in the equation and where we made explicit that for $t < 0$ it is zero.

12.5 Linear Periodically Time-Varying Systems

12.5.1 Floquet Theory

In this section we consider in more details linear periodically time-varying (LPTV) systems. In particular, we study systems that can be described by a state-space representation with matrices $A(.)$, $B(.)$, $C(.)$ and $D(.)$ having periodic smooth functions as elements. These include systems described by differential equations with periodic, indefinitely differentiable coefficients.

Consider the differential equation

$$\dot{y} = A(t)y + B(t)x \tag{12.38}$$

with $A(.)$ an $n \times n$-matrix and $B(.)$ an $n \times 1$ one, both with $\mathcal{T}$-periodic indefinitely differentiable elements and where, for brevity, we denote by $\dot{y}$ the (distributional) derivative of y and similarly for other quantities. Let further $Y(.)$ be the principal fundamental matrix of the equation and

$$U(t,\tau) = Y(t)Y^{-1}(\tau)$$

the evolution operator. From the periodicity of $A(.)$ we obtain

$$\begin{aligned}\dot{U}(t+\mathcal{T},\mathcal{T}) &= \dot{Y}(t+\mathcal{T})Y^{-1}(\mathcal{T})\\ &= A(t+\mathcal{T})Y(t+\mathcal{T})Y^{-1}(\mathcal{T})\\ &= A(t)U(t+\mathcal{T},\mathcal{T})\end{aligned}$$

from which, with $U(t,t) = I$ and the uniqueness of the solution of the equation we deduce

$$U(t+\mathcal{T},\mathcal{T}) = U(t,0)$$

and

$$Y(t+\mathcal{T}) = Y(t)Y(\mathcal{T})\,.$$

Let now introduce

$$P(t) = Y(t)\mathrm{e}^{-tF}\,, \qquad F \in \mathbb{C}^{n\times n}$$

with F an $n \times n$ matrix with constant coefficients and define the variable transformation

$$y(t) = P(t)z(t)\,.$$

In terms of z the equation becomes

$$\dot{P}(t)z + P(t)\dot{z} = A(t)P(t)z + B(t)x$$

or

$$\dot{z} = P^{-1}(t)\big[A(t)P(t) - \dot{P}(t)\big]z + P^{-1}(t)B(t)x\,.$$

To simplify this equation we calculate the derivative of P

$$\begin{aligned}\dot{P}(t) &= \dot{Y}(t)\mathrm{e}^{-tF} - Y(t)\mathrm{e}^{-tF}F\\ &= A(t)Y(t)\mathrm{e}^{-tF} - Y(t)\mathrm{e}^{-tF}F\\ &= A(t)P(t) - P(t)F\,.\end{aligned}$$

Using this result in the previous expression we finally obtain

$$\dot{z} = Fz + P^{-1}(t)B(t)x\,.$$

This equation is similar to the original one, but with the important difference that the periodically time-varying matrix $A(.)$ of the original equation has been replaced by a *constant* matrix F. This shows that the evolution operator of any system of differential equations with $A(.)$ a $\mathcal{T}$-periodic smooth matrix can be represented in the form

$$U(t,\tau) = P(t)\mathrm{e}^{(t-\tau)F}P^{-1}(\tau)\,, \tag{12.39}$$

This is called the *Floquet representation* of the evolution operator.

Let $y_0 \in \mathbb{C}^n$, from the analysis of LTI-systems we know that $\mathrm{e}^{tF} y_0$ is a linear combination of functions of the form

$$p_i(t)\mathrm{e}^{\lambda_i t}$$

with λ_i an eigenvalue of F and p_i a polynomial of degree lower than the algebraic multiplicity of λ_i. The Floquet representation tells us that the solution of (12.38) is a linear combination of functions of the form

$$\tilde{p}_i(t)\mathrm{e}^{\lambda_i t}$$

where $\tilde{p}_i$ are again polynomials, but in this case with $\mathcal{T}$-periodic smooth coefficients.

Example 12.7

In this example we look for the solution of the equation

$$Dy = A(t)y + x$$

with

$$A(t) = \begin{bmatrix} \omega_{3dB} + \Delta\omega \cos \omega_m t & 1 \\ 0 & \omega_{3dB} + \Delta\omega \cos \omega_m t \end{bmatrix}.$$

In particular we are interested in the evolution operator of the equation as it allows us to calculate the solution for an arbitrary input x.

First observe that $A(.)$ can be written as a sum of two matrices

$$A(t) = \begin{bmatrix} \omega_{3dB} & 1 \\ 0 & \omega_{3dB} \end{bmatrix} + \begin{bmatrix} \Delta\omega \cos \omega_m t & 0 \\ 0 & \Delta\omega \cos \omega_m t \end{bmatrix},$$

the first of which is constant, and we denote it by F. To find the principal fundamental matrix we make the ansatz

$$Y(t) = P(t)\mathrm{e}^{tF}, \qquad P(t) = p(t)I$$

with p an indefinitely differentiable periodic function with period $2\pi/\omega_m$. Inserting this ansatz in the equation we find

$$\begin{aligned} DY &= D\big[p(t)I\mathrm{e}^{tF}\big] \\ &= \dot{p}(t)I\mathrm{e}^{tF} + p(t)IF\mathrm{e}^{tF} \\ &= \big[F + \frac{\dot{p}}{p}I\big]Y(t)\,. \end{aligned}$$

From this expression we see that it satisfies the equation if

$$\frac{\dot{p}}{p} = \Delta\omega \cos \omega_m t.$$

The function p can be calculated from this equation and the condition $Y(0) = I$ using the method of the separation of variables from which we obtain

$$p(t) = \mathrm{e}^{\frac{\Delta\omega}{\omega_m} \sin \omega_m t}.$$

The principal fundamental matrix is thus

$$Y(t) = \mathrm{e}^{\frac{\Delta\omega}{\omega_m} \sin \omega_m t} \mathrm{e}^{tF}.$$

With Y and using the results of Example 8.2 for e^{tF} the evolution operator is found to be

$$U(t,\tau) = \frac{\mathrm{e}^{\frac{\Delta\omega}{\omega_m} \sin \omega_m t}}{\mathrm{e}^{\frac{\Delta\omega}{\omega_m} \sin \omega_m \tau}} \mathrm{e}^{\omega_{3dB}(t-\tau)} \begin{bmatrix} 1 & t-\tau \\ 0 & 1 \end{bmatrix}.$$

12.5.2 Time-Varying Frequency Response

Consider a SISO linear periodically time-varying system described by the state-space representation

$$Du = A(t)u + B(t)x \tag{12.40}$$

$$y = C(t)u + D(t)x \tag{12.41}$$

with $A(.)$, $B(.)$, $C(.)$ and $D(.)$ indefinitely differentiable $\mathcal{T}$-periodic matrix functions. Thanks to linearity we can analyse the response of the system for $D(t) = 0$ and add the contribution of $D(t)x$ at the end.

In the previous section we established that the evolution operator of (12.40) can be expressed in the form

$$U(t,\tau) = P(t)\mathrm{e}^{(t-\tau)F} P^{-1}(\tau)$$

with $P(t)$ an invertible, indefinitely differentiable $\mathcal{T}$-periodic matrix function and F a constant matrix. Using this representation for the response of the system we obtain

$$y(t) = 1_+(t-\tau)C(t)P(t)\mathrm{e}^{(t-\tau)F} P^{-1}(\tau)B(\tau) \star x(t)$$

or, in terms of the time-varying impulse response

$$y(t) = h_C *_t x(t)$$

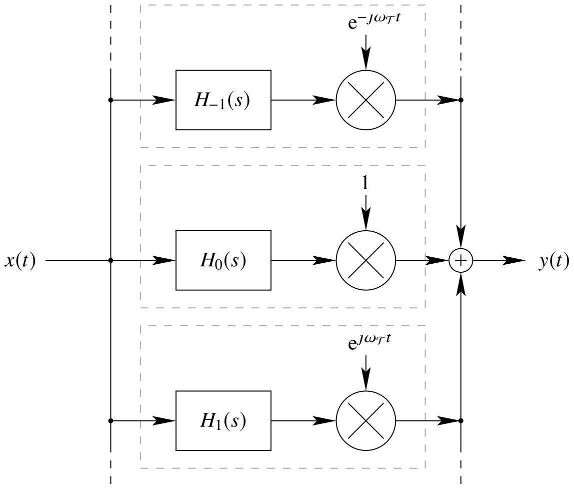

$e^{-\jmath\omega_{\mathcal{T}}t}$
$H_{-1}(s)$
1
$x(t)$
$H_0(s)$
$y(t)$
$e^{\jmath\omega_{\mathcal{T}}t}$
$H_1(s)$

Fig. 12.6 Representation of a stable LPTV system

with

$$h_C(t,\xi) = 1_+(\xi)C(t)P(t)e^{\xi F}P^{-1}(t-\xi)B(t-\xi)\,.$$

If we now add the contribution to the output from $D(t)x$ we finally find

$$y(t) = h *_t x(t)$$

with

$$h(t,\xi) = h_C(t,\xi) + D(t)\delta(\xi)\,.$$

The fist term h_C is a regular distribution growing at most exponentially with respect to ξ while the second has bounded support. The impulse response h is therefore Laplace transformable with respect to ξ. This implies that the system possess a time-varying transfer function $H(t,s)$. $H(t,s)$ is a function in the variables t and s and the above expression makes it clear that it is periodic in t. Therefore, with respect to t, we can expand it in a Fourier series

$$H(t,s) = \sum_{n=-\infty}^{\infty} H_n(s)e^{\jmath n\omega_{\mathcal{T}}t}$$

with $\omega_{\mathcal{T}} = 2\pi/\mathcal{T}$ and $H_n(s)$ functions of the variable s alone.

The last expression shows that LPTV systems can be regarded as the parallel connection of LTI subsystems with transfer functions H_n whose outputs are shifted in frequency by $n\omega_{\mathcal{T}}$ (see Fig. 12.6 and Example 12.4). This is best seen by applying a complex tone to a stable system. Thus, assume that all the eigenvalues of F have a negative real part, then the time-varying frequency response $\hat{h}(t,\omega)$ does also exist and is also a regular distribution that can be identified with a function in the variables t and ω. Proceeding as above we can write it as

$$\hat{h}(t,\omega) = \sum_{n=-\infty}^{\infty} \hat{h}_n(\omega) \mathrm{e}^{\jmath n \omega_{\mathcal{T}} t}$$

with $\hat{h}_n(\omega) = H_n(\jmath\omega)$. If we now apply a complex input tone $\mathrm{e}^{\jmath\omega_0 t}$ to the system and use (12.30) to calculate the system response we obtain

$$y(t) = \sum_{n=-\infty}^{\infty} \hat{h}_n(\omega_0) \mathrm{e}^{\jmath(n\omega_{\mathcal{T}}+\omega_0)t} \,.$$

The output is thus seen to consist of a sum of tones at $\omega_0 + n\omega_{\mathcal{T}}$, $n \in \mathbb{Z}$, each one weighted by $\hat{h}_n(\omega_0)$. It is readily seen that for a real system the following relation must hold

$$\hat{h}_{-n}(-\omega) = \overline{\hat{h}_n}(\omega) \,.$$

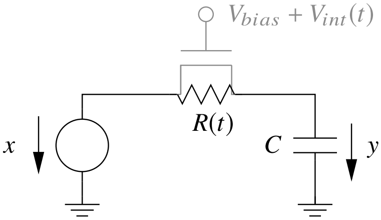
Example 12.8: LPTV LPF

In this example we examine a series RC low-pass filter (LPF) where, to reduce the physical area occupied by the circuit, the series resistor is implemented with a MOSFET. While this will produce some distortion, here we are interested in what happens if the gate bias voltage is disturbed by a periodic signal (see Fig. 12.7). This could happen for example if in a mixed-signal system (both analog and digital signals present) a line distributing the system clock is in proximity of the gate bias line, and the two are not properly isolated. The system is described by the following differential equation

$$\left[D + \omega_{3dB}(t)\right]y = \omega_{3dB}(t)x(t)\,, \qquad \omega_{3dB}(t) = \frac{1}{R(t)C}$$

with y the voltage across the capacitor, x the source voltage and $R(.)$ a periodic function. Given the periodicity of $R(.)$, $\omega_{3dB}(.)$ is also a periodic function with the same period that we assume to be smooth. $\omega_{3dB}(.)$ can therefore be expanded in a

Fig. 12.7 RC low-pass filter with a PTV resistor



Fourier series that, for simplicity of analysis, we assume to be given by

$$\omega_{3dB}(t) = \omega_0 + \Delta\omega \cos(\omega_m t)\,, \qquad \omega_0, \Delta\omega, \omega_m > 0$$

with $\Delta\omega \ll \omega_0$. We are interested in characterising the frequency response of the filter.

The equation describing the system separates into a differential equation with constant coefficients and a small perturbation term

$$\big(D + \omega_0\big)y + \Delta\omega \cos(\omega_m t)y = \omega_{3dB}(t)x(t)\,.$$

We can therefore solve the problem using the perturbation theory that we developed in Sect. 12.2.3. In addition, instead of solving for the time-varying impulse response and obtain the time-varying frequency response by Fourier transformation, it is convenient to solve directly the equation for the latter. Proceeding as in Sect. 12.4.2 we obtain

$$\big(D + \jmath\omega + \omega_0\big)\hat{h}(t,\omega) + \Delta\omega \cos(\omega_m t)\hat{h}(t,\omega) = \omega_{3dB}(t)$$

and we can identify $-\Delta\omega\cos(\omega_m t)$ with the perturbation term $\tilde{A}$ and $-(\jmath\omega + \omega_0)$ with the matrix A_0 of the unperturbed system.

We start by computing the time-varying frequency response of the unperturbed system that we denote by $\hat{h}_0(t,\omega)$ and which has to satisfy

$$\big(D + \jmath\omega + \omega_0\big)\hat{h}_0(t,\omega) = \omega_0 + \frac{\Delta\omega}{2}\big(\mathrm{e}^{\jmath\omega_m t} + \mathrm{e}^{-\jmath\omega_m t}\big)$$

where we have represented $\cos \omega_m t$ by complex tones. Note that the variation in $R(.)$ results in additional input tones to an otherwise time invariant system. The solution of the equation is readily calculated to be

$$\hat{h}_0(t,\omega) = H(\omega) + \frac{\Delta\omega}{2\omega_0}\big[H(\omega + \omega_m)\mathrm{e}^{\jmath\omega_m t} + H(\omega - \omega_m)\mathrm{e}^{-\jmath\omega_m t}\big]$$

with

$$H(\omega) = \frac{1}{1 + \jmath\frac{\omega}{\omega_0}}$$

the frequency response of the RC filter without disturbances (that is for $\omega_{3dB}(t) = \omega_0$). $H(\omega)/\omega_0$ plays the role of the fundamental kernel W_0 of Sect. 12.2.3. However, because A_0 is time invariant we can work in the convolution algebra of periodic distributions and instead of the fundamental kernel, the system can be characterised by the fundamental solution of the equation. In this example the kth Fourier coefficient of the fundamental solution of the equation is given by (see Example 7.5)

$$e_k = \frac{H(\omega + k\omega_m)}{\mathcal{T}\omega_0}; \qquad \mathcal{T} = 2\pi/\omega_m\,.$$

We now calculate the first order perturbation term. The first step consists in calculating the new "input signal" produced by the perturbation $\tilde{A}$

$$x_1(t) = \tilde{A}(t)\hat{h}_0(t,\omega) = -\frac{\Delta\omega}{2}\big(\mathrm{e}^{\jmath\omega_m t} + \mathrm{e}^{-\jmath\omega_m t}\big)\hat{h}_0(t,\omega)\,.$$

The first order perturbation term of the frequency response $\hat{h}_1(t,\omega)$ is then obtained by applying this signal to the unperturbed system

$$\big(D + \jmath\omega + \omega_0\big)\hat{h}_1(t,\omega) = -\frac{\Delta\omega}{2}\big(\mathrm{e}^{\jmath\omega_m t} + \mathrm{e}^{-\jmath\omega_m t}\big)\hat{h}_0(t,\omega)\,.$$

The solution of the equation is given by

$$\begin{aligned}\hat{h}_1(t,\omega) = &-\frac{\Delta\omega}{2\omega_0}H(\omega)\big[H(\omega+\omega_m)\mathrm{e}^{\jmath\omega_m t} + H(\omega-\omega_m)\mathrm{e}^{-\jmath\omega_m t}\big]\\ &-\Big(\frac{\Delta\omega}{2\omega_0}\Big)^2\Big\{H(\omega)\big[H(\omega+\omega_m) + H(\omega-\omega_m)\big]\\ &+ H(\omega+\omega_m)H(\omega+2\omega_m)\mathrm{e}^{\jmath 2\omega_m t}\\ &+ H(\omega-\omega_m)H(\omega-2\omega_m)\mathrm{e}^{-\jmath 2\omega_m t}\Big\}\,.\end{aligned}$$

Note that both $\hat{h}_0$ and $\hat{h}_1$ include terms of order $\Delta\omega$. Since $\tilde{A}$ is proportional to $\Delta\omega$ and all terms of $\hat{h}_1$ are proportional to powers of this quantity, no higher perturbation term will include a contribution of order $\Delta\omega$. The first two terms $\hat{h}_0$ and $\hat{h}_1$ are therefore enough to establish the effects of the perturbation of order $\Delta\omega$. To obtain an estimate to second order in $\Delta\omega$ we would need to calculate $\hat{h}_2$ as well.

With these results the first order response of the system when driven by a tone at ω is given by

$$y(t) = \big[\hat{h}_0(t,\omega) + \hat{h}_1(t,\omega)\big]\mathrm{e}^{\jmath\omega t}\,.$$

It is comprised by tones at $\omega + n\omega_m$, $n = -2, -1, 0, 1, 2$. It's not difficult to see that if we would calculate higher order terms we would obtain similar tones for larger values of $|n|$ and in the limit, when including all perturbation terms, for all $n \in \mathbb{Z}$.

Let's consider more closely the component at $\omega + \omega_m$

$$\begin{aligned}y_1(t) &= \frac{\Delta\omega}{2\omega_0}H(\omega+\omega_m)\big[1 - H(\omega)\big]\mathrm{e}^{\jmath(\omega_m+\omega)t}\\ &= \frac{\Delta\omega}{2\omega_0}H(\omega+\omega_m)\frac{\jmath\frac{\omega}{\omega_0}}{1+\jmath\frac{\omega}{\omega_0}}\mathrm{e}^{\jmath(\omega_m+\omega)t}\end{aligned}$$

and assume that $\omega_m \gg \omega_0$. If the filter is part of a transmitter and used to suppress noise outside the channel allocated to the user or service, then a $2\pi/\omega_m$-periodic perturbation is seen to create spurious emissions that can fall in frequency ranges reserved for other users or services and violate the maximum allowed emission levels. From the above expression we note that a wide nominal filter bandwidth ω_0 helps in reducing the emission level caused by the perturbation. This can be interpreted intuitively as follows. If the input signal frequency is much smaller than the 3 dB cut-off frequency of the filter, then it produces a very small current flowing through the filer components and, in the limit of zero current, the output signal doesn't depend on the value of the filter components.

If the input tone is well above the nominal 3 dB cut-off frequency of the filter $|\omega| \gg \omega_0$ then $|H(\omega)| \ll 1$ and the output tone at $\omega + \omega_m$ can be approximated by

$$y_1(t) \approx \frac{\Delta\omega}{2\omega_0} H(\omega + \omega_m) \mathrm{e}^{\jmath(\omega_m + \omega)t} .$$

If the frequency of the input tone is such that $|\omega + \omega_m| < \omega_0$ then the tone falls in a spurious pass band of the filter and for $|\omega + \omega_m| \ll \omega_0$ it can be approximated by

$$y_1(t) \approx \frac{\Delta\omega}{2\omega_0} \mathrm{e}^{\jmath(\omega_m + \omega)t} .$$

If the filter is part of a communication receiver responsible to suppress interfering signals (the *channel filter*) then we see that $2\pi/\omega_m$-periodic perturbations introduce spurious responses in the stop band of the filter at multiples of ω_m that down-convert interfering signals in band, possibly masking the wanted signal. The amplitude of the dominant spurious response is proportional to the perturbation magnitude $\Delta\omega$ relative to the nominal 3 dB cut-off frequency of the filter.

Example 12.9: Quadrature (De-)Modulator

Consider the frequency translating system of Example 12.4 with time-varying impulse response

$$h_{\text{mod}}(t, \xi) = \mathrm{e}^{\jmath w_0 t} \delta(\xi) .$$

It is a complex system in the sense that if we apply a real valued input signal its response is complex valued. In this example we show that the system can be implemented using two real sub-systems.

Let's decompose the input signal into its real and imaginary parts

$$x(t) = r(t) + \jmath q(t).$$

The system response is given by

$$y(t) = h_{\text{mod}}(t, \xi) *_t x(t) = [r(t) + \jmath q(t)]\mathrm{e}^{\jmath w_0 t}$$

and can be written as

$$[r(t) + \jmath q(t)] \cos w_0 t - [q(t) - \jmath r(t)] \sin w_0 t\,.$$

In this form the system response is seen to be the sum of the responses of two real systems driven by correlated signals (see Fig. 12.8). By linearity, if the two systems are driven by the real part only of the input signals, that is by r and q respectively, then the response of the system is

$$y(t) = \Re\{[r(t) + \jmath q(t)]\mathrm{e}^{\jmath w_0 t}\}\,.$$

The combination of the two real systems is called a *quadrature modulator.* Each of the two real subsystems is called *mixer* and effectively multiply the input signal with a second real valued signal l called the local oscillator (LO) signal. A mixer can therefore be considered a system having two input ports.

Consider now a system that shifts the spectrum of the input signal in the opposite direction

$$h_{\text{demod}}(t, \xi) = \mathrm{e}^{-\jmath w_0 t}\delta(\xi)\,.$$

We would like to find a real system implementation that when driven by the signal

$$[r(t) + \jmath q(t)]\mathrm{e}^{\jmath w_0 t}$$

allows us to recover the signals used at the input of the quadrature modulator used to generate it. Such a system is readily found by observing that

$$[r(t) + \jmath q(t)]\mathrm{e}^{\jmath w_0 t} \cos w_0 t = [r(t) + \jmath q(t)]\frac{1}{2}[\mathrm{e}^{\jmath 2 w_0 t} + 1]$$

and similarly

$$[r(t) + \jmath q(t)]\mathrm{e}^{\jmath w_0 t}(-1) \sin w_0 t = [r(t) + \jmath q(t)]\frac{-\jmath}{2}[\mathrm{e}^{\jmath 2 w_0 t} - 1]\,.$$

Thus, if the signals r and q are band-limited to frequencies smaller than w_0, the original signals can be recovered (up to a fixed scaling factor) by use of two mixers driven by quadrature (orthogonal) local oscillator signals followed by low-pass filters (see Fig. 12.9). Such a system is called a *quadrature demodulator.* By linearity, if the system is driven by the real signal

$$\Re\{[r(t) + \jmath q(t)]\mathrm{e}^{\jmath w_0 t}\}$$

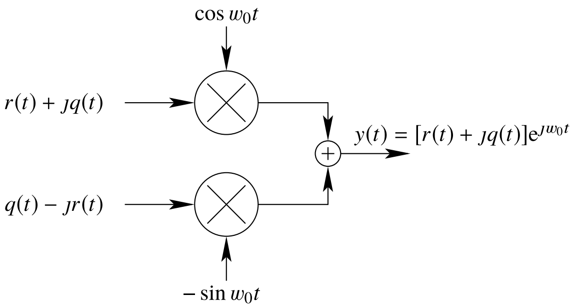


Fig. 12.8 Quadrature modulator

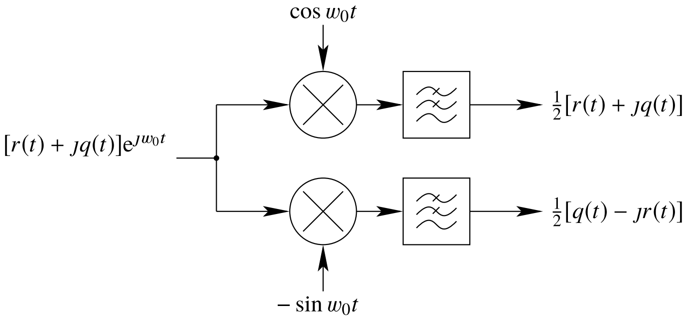

$\cos w_0 t$
$[r(t) + \jmath q(t)]e^{\jmath w_0 t}$
$\frac{1}{2}[r(t) + \jmath q(t)]$
$\frac{1}{2}[q(t) - \jmath r(t)]$
$-\sin w_0 t$

Fig. 12.9 Quadrature demodulator

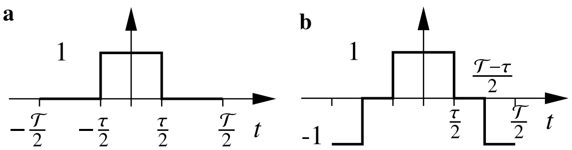


Fig. 12.10 **a** Typical local oscillator unipolar waveform **b** Typical local oscillator bipolar waveform

the two output signals are the real parts of what we found above, that is $r/2$ and $q/2$ respectively.

Example 12.10: Harmonic-Reject Mixer

We saw in Example 12.9 that a mixer is a system multiplying the input signal with a $\mathcal{T}$-periodic signal called the local oscillator signal

$$h(t, \xi) = l(t)\delta(\xi)\,.$$

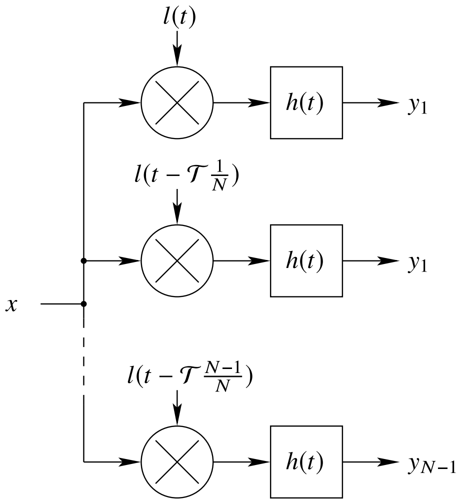

$l(t)$
$h(t)$
y_1
$l(t - \mathcal{T}\frac{1}{N})$
$h(t)$
y_1
x
$l(t - \mathcal{T}\frac{N-1}{N})$
$h(t)$
y_{N-1}

Fig. 12.11 Generic N-path receiver

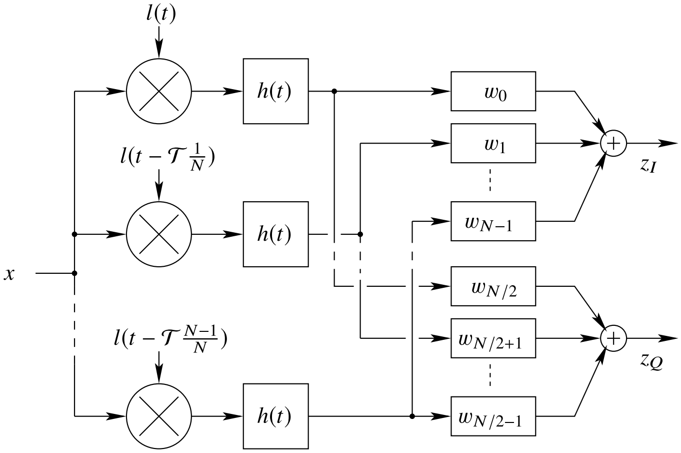


Fig. 12.12 Quadrature N-path demodulator

In practical implementations, to minimise the signal-to-noise degradation caused by the circuit, the local oscillator signal is not a pure sinusoidal. Instead, it is most often designed to approach a rectangular waveform as depicted in Fig. 12.10b. Being periodic the signal l can be represented by a Fourier series

$$l(t) = \sum_{n=-\infty}^{\infty} a_n e^{\jmath n \omega_{\mathcal{T}} t}, \qquad \omega_{\mathcal{T}} = 2\pi/\mathcal{T}$$

with

$$a_n = \begin{cases} \frac{\tau}{\mathcal{T}} & n = 0 \\ \frac{1}{\pi n}\sin(n\pi\frac{\tau}{\mathcal{T}}) & n \neq 0 \end{cases}$$

for the waveform in Fig. 12.10a and

$$a_n = \begin{cases} 0 & n \text{ even} \\ \frac{2}{\pi n}\sin(n\pi\frac{\tau}{\mathcal{T}}) & n \text{ odd} \end{cases}$$

for the one in Fig. 12.10b. Therefore, a mixer driven by an input tone

$$x(t) = \mathrm{e}^{\jmath(n\omega_{\mathcal{T}}+\omega_1)t} \tag{12.42}$$

produces an output tone at ω_1 for every value of n for which $a_n \neq 0$. When the mixer is part of a receiver designed to down-convert a signal at $\omega_{\mathcal{T}} + \omega_1$ to ω_1 for further processing and detection, the spurious responses ($n \neq -1$) are undesired as they could cause an interfering signal to overlap in frequency with the desired signal and prevent reception of the latter. The spurious responses are most often suppressed by preceding the mixer with a suitable filter. However, in some situations such a filter is undesired. In the following we present a method to suppress the dominant spurious responses of a mixer without the need for filters and still using rectangular waveforms as local oscillator signals.

Note that, while the idealised local oscillator waveforms shown in Figs. 12.10a and 12.10b are discontinuous, their Fourier series representations truncated at an arbitrarily high value of $|n|$ are indefinitely differentiable functions. Suitably truncated Fourier series are adequate representations of practical signals and do not cause any mathematical difficulty.

Consider the generic N-path receiver shown in Fig. 12.11. It is composed by N subsystems that are equal apart from the fact that the local oscillator signal of path $k, k = 0, \ldots, N-1$ is delayed by $\mathcal{T}k/N$ with respect to path 0. The blocks preceding the output signals y_k represent LTI subsystem with impulse response h. Let the input signal be as in (12.42). Then, due to the tone at $-n\omega_{\mathcal{T}}$ in the Fourier series of l, the kth output signal includes a tone at ω_1 given by

$$\begin{aligned} y_{k,-n} &= h(t) * [a_{-n}\mathrm{e}^{-\jmath n\omega_{\mathcal{T}}(t-\mathcal{T}\frac{k}{N})}\delta(\xi) *_t \mathrm{e}^{\jmath(n\omega_{\mathcal{T}}+\omega_1)t}] \\ &= [a_{-n}H(\jmath\omega_1)\mathrm{e}^{\jmath\omega_1 t}]\mathrm{e}^{\jmath nk\frac{2\pi}{N}} \\ &= y_{0,-n}(t)\mathrm{e}^{\jmath nk\frac{2\pi}{N}} \end{aligned}$$

with H the Laplace transform of h. This shows that the output components of interest (at ω_1) are the product of the signal $y_{0,-n}$ and the constants $\mathrm{e}^{\jmath n2\pi\frac{k}{N}}$, $k = 0, \ldots, N-1$. By exploiting the properties of trigonometric functions we can form weighted sums of the outputs y_k such that the resulting tone at ω_1 vanishes for some values of n

$$z_{-n}(t) = \sum_{k=0}^{N-1} w_k y_{k,-n}(t) = y_{0,-n}(t) \sum_{k=0}^{N-1} w_k \mathrm{e}^{\jmath nk\frac{2\pi}{N}} .$$

Note that the sum on the right-hand side corresponds to a discrete Fourier transform of the weighting coefficients. For example, by choosing

$$w_k = \cos\left(\frac{2\pi}{N}k\right) \tag{12.43}$$

we obtain

$$\begin{aligned} z_{-n}(t) &= y_{0,-n}(t) \sum_{k=0}^{N-1} \cos\left(\frac{2\pi}{N}k\right) \mathrm{e}^{\jmath nk\frac{2\pi}{N}} \\ &= \frac{y_{0,-n}(t)}{2} \sum_{k=0}^{N-1} \mathrm{e}^{\jmath \frac{2\pi}{N}(n+1)k} + \mathrm{e}^{\jmath \frac{2\pi}{N}(n-1)k} . \end{aligned}$$

The sums are geometric series that evaluate to

$$\sum_{k=0}^{N-1} \mathrm{e}^{\jmath \frac{2\pi}{N}(n\pm1)k} = \begin{cases} \frac{1-\mathrm{e}^{\jmath 2\pi(n\pm1)}}{1-\mathrm{e}^{\jmath \frac{2\pi}{N}(n\pm1)}} = 0 & n \pm 1 \neq Nm, m \in \mathbb{Z} \\ N & \text{otherwise} \end{cases}$$

and therefore the signal z_{-n} is

$$z_{-n}(t) = \begin{cases} 0 & n \neq Nm \pm 1 \\ \frac{N}{2} y_{0,-n}(t) & \text{otherwise} . \end{cases}$$

For example, for $N = 8$ all harmonics below the 15th except for the 7th and the 9th are suppressed. A mixer with no spurious responses at some odd harmonics is called a *harmonic-reject mixer.*

The weighting factors of (12.43) are not the only possible choice. For example, any rotation of the indexes $w_{(k+m) \bmod N}$ produces a similar result with the addition of a phase factor to the output signal. For N even we can thus construct a full quadrature demodulator by building two weighted sums, one with weighting factors as given by (12.43) and the other by factors rotated by $N/2$ ($w_{k+N/2}$; see Fig. 12.12). The case with $N = 4$ corresponds to classical situation with differential output signals. Further choices of weighting factors allow isolating responses at values of n different from 1.

While we discussed summing the signals after the LTI systems characterised by h, the same results apply if the signals are summed right after the mixers. The rejection obtained in practice is limited by mismatch between the paths. The place where the summation is implemented plays a role in this respect.

If we revert the direction of the signals in the system of Fig. 12.11 we obtain an N-path transmitter. This is a generalisation of the classic case with $N = 4$ with the

4 input signals being differential versions of the r and q modulator input signals. As with the receiver, a larger value of N allow suppressing spurious emissions at harmonics of the local oscillator signal without the use of filters.

Chapter 13
Weakly Nonlinear Time-Varying Systems

The theory of linear time-varying systems can be extended to weakly-nonlinear time-varying (WNTV) systems in a similar way as we did for linear time-invariant systems. In this chapter we first define WNTV systems mathematically and highlight some important differences from the theory of WNTI systems. We then discuss weakly-nonlinear periodically time-varying (WNPTV) systems. These type of systems generate a characteristic spectrum that is relatively easy to describe and is relevant, for example, in the study and design of communication systems.

13.1 Weakly Nonlinear Time-Varying Systems

13.1.1 Definition

A *Weakly-nonlinear time-varying system* is defined as a system whose response to the input signal x can be described by

$$y(t) = \sum_{k=1}^{\infty} w_k(t, \tau_1, \dots, \tau_k) \star x^{\otimes k}(\tau_1, \dots, \tau_k)\,. \tag{13.1}$$

The operator $\star$ is the extension of the operator introduced in Sect. 12.2.2 to higher dimensions. For causal systems described by regular distributions and driven by a right sided input x it is defined by

$$w_k(t, \tau_1, \dots, \tau_k) \star x^{\otimes k}(\tau_1, \dots, \tau_k) := \int_0^t \cdots \int_0^t w_k(t, \tau_1, \dots, \tau_k) x(\tau_1) \cdots x(\tau_k) \mathrm{d}\tau_1 \cdots \mathrm{d}\tau_k\,. \tag{13.2}$$

F. Beffa, *Weakly Nonlinear Systems*, Understanding Complex Systems,
https://doi.org/10.1007/978-3-031-40681-2_13

w_k is the *kth order fundamental kernel* of the system. As for WNTI systems, to guarantee uniqueness, we require it to be symmetric in the variables $\tau_1, \dots, \tau_k$.

Generalizations valid for a wider class of input signals can be done in the same way as was done for LTV systems. Note that $w_k \star x^{\otimes k}$ is a distribution of the single variable t and not a higher dimensional distribution as for WNTI systems. The reason for this is explained next.

Consider a WNTV system described by a differential equation of the form

$$L(t, D)y = N(t, D)x + c_2(t)y^2 + c_3(t)y^3 + \cdots$$

with

$$\begin{aligned} L(t, D) &= D^m + a_{m-1}(t)D^{m-1} + \cdots + a_0(t) \\ N(t, D) &= b_n(t)D^n + b_{n-1}(t)D^{n-1} + \cdots + b_0(t) \end{aligned}$$

and where all coefficients a_i, b_i and c_i are indefinitely differentiable functions. The equation can be solved iteratively as in the case of WNTI systems. We first solve the linear part of the equation. The solution y_1 is then used in the nonlinear terms to compute “nonlinear sources” of second order. With them we solve the part of the equation consisting of terms of second order only, a linear equation, and so on.

There is an important difference compared to the case of WNTI systems: in the case of WNTI systems, to get around the lack of a general multiplication between arbitrary distributions, we made use of a direct product of distributions and introduced a multiplication based on the tensor product. Here the same method doesn't work as the coefficients of the differential equation are functions of the single time variable t and it is unclear how to adapt them for use with higher order distributions. For this reason here the responses of all orders y_k are distributions of the single variable t. To solve the equation we must therefore assume the existence of all appearing multiplications and powers y^n, $k = 2, 3, \dots$.

If we consider t as a fix parameter then the multiplication between components of y act as a tensor product like operation. Consider the product between y_k and y_l

$$\begin{aligned} y_k(t)y_l(t) &= \int_0^t \cdots \int_0^t w_k(t, \tau_1, \dots, \tau_k)x^{\otimes k}(\tau_1, \cdots, \tau_k)\mathrm{d}\tau_1 \cdots \mathrm{d}\tau_k \\ &\quad \cdot \int_0^t \cdots \int_0^t w_l(t, \tau_1, \dots, \tau_l)x^{\otimes l}(\tau_1, \cdots, \tau_l)\mathrm{d}\tau_1 \cdots \mathrm{d}\tau_l \\ &= \int_0^t \cdots \int_0^t w_k(t, \tau_1, \dots, \tau_k)w_l(t, \tau_{k+1}, \dots, \tau_{k+l}) \\ &\qquad \cdot x^{\otimes k+l}(\tau_1, \cdots, \tau_{k+l})\mathrm{d}\tau_1 \cdots \mathrm{d}\tau_{k+l}\,. \end{aligned} \tag{13.3}$$

The result has the form of a response of order $k + l$ which can be interpreted as a "nonlinear source" generated by nonlinearities and lower order responses as desired.

To solve the equation we must be able to solve the equation for each order independently and verify that it has the desired form. Solving the equations is (in principle) simple as all equations are linear. The solution of the equation consisting of terms of order k is given by

$$y_k(t) = \int_0^t \int_0^\tau \cdots \int_0^\tau v(t,\tau) z_k(\tau,\tau_1,\ldots,\tau_k) x^{\otimes k}(\tau_1,\cdots,\tau_k) \mathrm{d}\tau_1 \cdots \mathrm{d}\tau_k \mathrm{d}\tau$$

with $z_k \star x^{\otimes k}$ the nonlinear source and v the fundamental kernel of the equation.

To show that this expression can be transformed in the desire form, consider the integral

$$\int_0^t \int_0^\tau \int_0^\tau f(\tau,\tau_1,\tau_2) \mathrm{d}\tau_2 \mathrm{d}\tau_1 \mathrm{d}\tau \,.$$

As a first step we exchange the order of integration between τ and τ_1 and obtain

$$\int_0^t \int_{\tau_1}^t \int_0^\tau f(\tau,\tau_1,\tau_2) \mathrm{d}\tau_2 \mathrm{d}\tau \mathrm{d}\tau_1 \,.$$

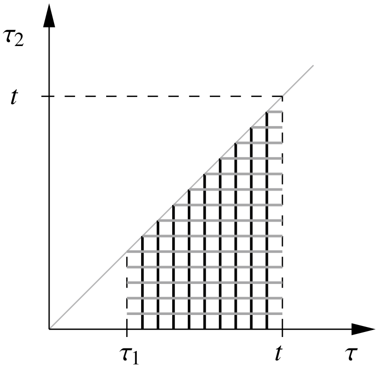
We then perform a second exchange between τ_2 and τ (refer to Fig. 13.1) which results in

$$\int_0^t \int_0^t \int_{\max(\tau_1,\tau_2)}^t f(\tau,\tau_1,\tau_2) \mathrm{d}\tau \mathrm{d}\tau_2 \mathrm{d}\tau_1 \,.$$

If the integral would involve more integrations between 0 and τ then we could repeat the last step more times giving

$$\int_0^t \cdots \int_0^t \int_{\max(\tau_1,\ldots,\tau_k)}^t f(\tau,\tau_1,\ldots,\tau_k) \mathrm{d}\tau \mathrm{d}\tau_k \cdots \mathrm{d}\tau_1 \,. \tag{13.4}$$

Fig. 13.1 Domain of integration (refer to text)



Using this result we can transform the above expression for $y_k(t)$ into

$$y_k(t) = \int_0^t \cdots \int_0^t \int_{\max(\tau_1,\ldots,\tau_k)}^t v(t,\tau) z_k(\tau,\tau_1,\ldots,\tau_k) \mathrm{d}\tau$$
$$\cdot x^{\otimes k}(\tau_1,\ldots,\tau_k) \mathrm{d}\tau_1 \cdots \mathrm{d}\tau_k$$

which has the desired form $w_k \star x^{\otimes k}$.

13.1.2 Time-Varying Nonlinear Impulse Responses

As for LTV systems, the response of WNTV systems can also be expressed in terms of the *time-varying nonlinear impulse responses*

$$h_k(t,\xi_1,\ldots,\xi_k) := w_k(t, t-\xi_1,\ldots,t-\xi_k) \tag{13.5}$$

and the convolution operator $*_t$ for time varying systems

$$h_k(t,\xi_1,\ldots,\xi_k) *_t x^{\otimes k}(\xi_1,\ldots,\xi_k) :=$$
$$\int_0^t \cdots \int_0^t h_k(t, t-\xi_1,\ldots,t-\xi_k) x(\xi_1) \cdots x(\xi_k) \mathrm{d}\xi_1 \cdots \mathrm{d}\xi_k \,. \tag{13.6}$$

Example 13.1

Consider a WNTV system described by the following differential equation

$$Dy + a(t)y = x + y^2\,.$$

We are interested in the second order fundamental kernel of the system.

The fundamental kernel of the linearized equation is given by (12.26) which, taking into account the commutativity of the product of scalar functions simplifies to

$$w_1(t,\tau_1) = \mathrm{e}^{-\int_{\tau_1}^{t} a(\lambda)\mathrm{d}\lambda}\,.$$

With it the linear response of the system is

$$y_1(t) = w_1(t,\tau_1) \star x(t)\,.$$

Given y_1 we can compute the "nonlinear source" of second order

$$y_1^2(t) = \int_0^t \int_0^t w_1(t,\tau_1) w_1(t,\tau_2) x(\tau_1) x(\tau_2) \mathrm{d}\tau_1 \mathrm{d}\tau_2\,.$$

With it we can then solve the equation consisting of terms of second order only

$$\big(D + a(t)\big) y_2 = y_1^2\,.$$

The fundamental kernel of this equation is the same as the one of the first order equation. The second order response of the system is therefore

$$\begin{aligned} y_2(t) &= \int_0^t w_1(t,\tau) y_1^2(\tau)\mathrm{d}\tau \\ &= \int_0^t \int_0^t \int_{\max(\tau_1,\tau_2)}^{t} w_1(t,\tau) w_1(\tau,\tau_1) w_1(\tau,\tau_2) \mathrm{d}\tau\, x(\tau_1) x(\tau_2) \mathrm{d}\tau_1 \mathrm{d}\tau_2\,. \end{aligned}$$

The second order fundamental kernel of the system can be found by comparing this expression with $y_2 = w_2(t,\tau_1,\tau_2) \star x^{\otimes 2}(\tau_1,\tau_2)$ giving

$$w_2(t,\tau_1,\tau_2) = \int_{\max(\tau_1,\tau_2)}^{t} \mathrm{e}^{-\int_{\tau}^{t} a(\lambda)\mathrm{d}\lambda - \int_{\tau_1}^{\tau} a(\lambda)\mathrm{d}\lambda - \int_{\tau_2}^{\tau} a(\lambda)\mathrm{d}\lambda} \mathrm{d}\tau\,.$$

As a check we verify that in the special case in which $a(t)$ is constant we obtain the same result as in Example 9.5. Evaluating the integrals gives

$$w_2(\tau_1, \tau_2) = \frac{1}{a}\Big(\mathrm{e}^{-a[t-\min(\tau_2,\tau_1)]} - \mathrm{e}^{-a(2t-\tau_1-\tau_2)}\Big)$$

and, after the variable substitutions $\xi_i = t - \tau_i$, $i = 1, 2$ we indeed obtain an expression equivalent to h_2 in Example 9.5.

13.1.3 Time-Varying Nonlinear Frequency Responses

Weakly-nonlinear time-varying systems can equivalently be characterised by *time-varying nonlinear frequency responses*. The kth order one is defined as the Fourier transform with respect to $\xi_1, \ldots, \xi_k$ of the impulse response $h_k(t, \xi_1, \ldots, \xi_k)$. For regular distributions

$$\hat{h}_k(t, \omega_1, \ldots, \omega_k) := \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h_k(t, \xi_1, \ldots, \xi_k)\mathrm{e}^{-\jmath(\omega,\xi)}\mathrm{d}^k\xi \tag{13.7}$$

with $\omega, \xi \in \mathbb{R}^k$.

The response of order k of a system can be calculated by

$$y_k(t) = \frac{1}{(2\pi)^k}\hat{h}_k(t, \omega_1, \ldots, \omega_k)\,\mathrm{e}^{\jmath(\omega_1+\cdots+\omega_k)t} \star \hat{x}^{\otimes k}(\omega_1, \ldots, \omega_k)\,. \tag{13.8}$$

The derivation is entirely analogous to the one dimensional case carried out in Sect. 12.4.1.

13.2 Weakly Nonlinear Periodically Time-Varying Systems

Weakly nonlinear periodically time-varying (WNPTV) systems are weakly nonlinear systems whose characteristics vary periodically in time. In other words, their fundamental kernels, impulse responses and frequency responses are periodic functions of time and can therefore be expanded in Fourier series. For example, the kth order frequency response of a $\mathcal{T}$-periodic system can be represented by the series

$$\hat{h}_k(t, \omega_1, \ldots, \omega_k) = \sum_{n=-\infty}^{\infty} \hat{h}_{k,n}(\omega_1, \ldots, \omega_k)\mathrm{e}^{\jmath n\omega_{\mathcal{T}} t}$$

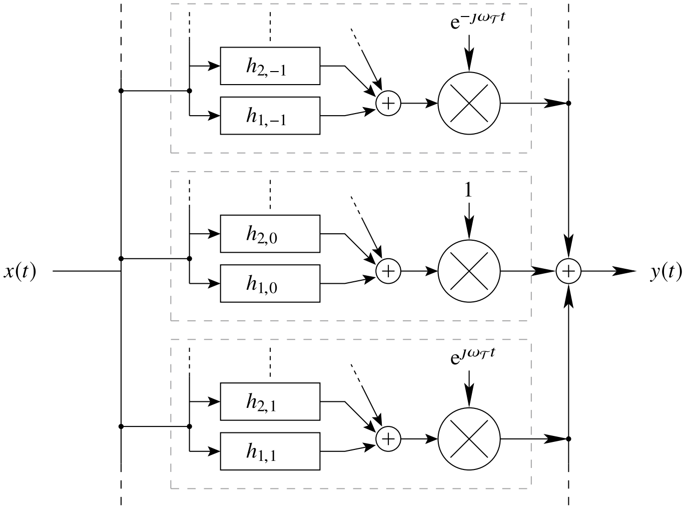


Fig. 13.2 Generic representation of a WNPTV system

with $\omega_{\mathcal{T}} = 2\pi/\mathcal{T}$. This representation highlights the fact that such systems can be represented by a parallel connection of a countable set of weakly nonlinear *time-invariant* networks whose outputs are shifted in frequency by a multiple of $\omega_{\mathcal{T}}$ (see Fig. 13.2). Practical applications where this representation is particularly useful include the analysis and design of communication systems.

In the rest of this section we focus on the special case in which weakly nonlinear periodically time-varying systems are driven by a set of tones. This will reveal a spectrum characteristic of this type of systems.

13.2.1 Discrete Convolution

Before turning to actually calculating the response of WNPTV systems driven by a set of tones, it's convenient to introduce some notation that will simplify many expressions.

A series $\sum\limits_{n=-\infty}^{\infty} a_n$ is *absolutely convergent* if the sum of the absolute values of the terms converges

$$\sum_{n=-\infty}^{\infty} |a_n| < \infty\,.$$

In this case the value of the series doesn't depend on the order of the elements. The product of two absolutely convergent series $\sum_{n=-\infty}^{\infty} a_n$ and $\sum_{n=-\infty}^{\infty} b_n$ is also absolutely convergent

$$|\sum_{n=-\infty}^{\infty} a_n \sum_{n=-\infty}^{\infty} b_n| \leq \sum_{n=-\infty}^{\infty} |a_n| \sum_{n=-\infty}^{\infty} |b_n| < \infty$$

and can be expressed as

$$\sum_{n=-\infty}^{\infty} a_n \sum_{n=-\infty}^{\infty} b_n = \sum_{n=-\infty}^{\infty} \Big(\sum_{q=-\infty}^{\infty} a_q b_{n-q} \Big) .$$

The inner sum in the last expression is called *discrete convolution* (or *Cauchy product*). For convenience, we are going to denote it by

$$(a_. *_d b_.)_n := \sum_{q=-\infty}^{\infty} a_q b_{n-q} \tag{13.9}$$

The discrete convolution is associative and commutative

$$\big((a_. *_d b_.) *_d c_.\big)_n = \big(a_. *_d (b_. *_d c_.)\big)_n$$
$$(a_. *_d b_.)_n = (b_. *_d a_.)_n$$

and has a unit element, the *Kronecker delta*

$$\delta_n = \begin{cases} 1 & n = 0 \\ 0 & n \neq 0 . \end{cases} \tag{13.10}$$

13.2.2 Product of Fourier Series

In the following we use the convention introduced in Sect. 4.5 of denoting the kth Fourier coefficient of a distribution f by $c_k(f)$.

It is well known that if $t \mapsto f(t)$ is a continuous $\mathcal{T}$-periodic function, its Fourier series is absolutely convergent for all values of t [23]. If f and g are two such functions then their product is well-defined and continuous. In addition, the Fourier coefficients of the product can be expressed in terms of the coefficients of the individual series

$$\sum_{n=-\infty}^{\infty} c_n(f) e^{\jmath n \omega_{\mathcal{T}} t} \sum_{n=-\infty}^{\infty} c_n(g) e^{\jmath n \omega_{\mathcal{T}} t} = \sum_{n=-\infty}^{\infty} \Bigg(\sum_{q=-\infty}^{\infty} c_q(f)\, c_{n-q}(g) \Bigg) e^{\jmath n \omega_{\mathcal{T}} t} .$$

The coefficients of the product are evidently the convolution product of the coefficients of the two series

$$\big(c_{.}(f) *_d c_{.}(g)\big)_n = \sum_{q=-\infty}^{\infty} c_q(f)\, c_{n-q}(g)\,. \tag{13.11}$$

Let now f and g be two $\mathcal{T}$-periodic *distributions*. Let further introduce the sequences (f_k) and (g_k) defined by

$$f_k = f * \beta_k \qquad \text{and} \qquad g_k = g * \beta_k$$

with (β_k) a sequence of functions in $\mathcal{D}$ converging to δ (for example the sequence of Example 2.5). Then (f_k) and (g_k) are sequences of indefinitely differentiable functions converging as distributions to f and g respectively. The Fourier series of each member of each sequence is thus absolutely convergent.

If the product $f\,g$ exists, then it defines a $\mathcal{T}$-periodic distribution which must coincide with the limit of the sequence

$$f\,g = \lim_{k\to\infty} f_k\, g_k\,.$$

The Fourier series of each member of the sequence can be written as

$$f_k\, g_k = \sum_{n=-\infty}^{\infty} \big(c_{.}(f_k) *_d c_{.}(g_k)\big)_n \mathrm{e}^{\jmath n \omega_{\mathcal{T}} t}\,.$$

Therefore, from the assumption of convergence and the uniqueness of the Fourier series representation of periodic distributions we conclude that the Fourier coefficients of $f\,g$ must be

$$c_n(f\,g) = \big(c_{.}(f) *_d c_{.}(g)\big)_n := \lim_{k\to\infty} \big(c_{.}(f_k) *_d c_{.}(g_k)\big)_n\,.$$

Example 13.2

Consider the regular $\mathcal{T}$-periodic distribution shown in Fig. 13.3 that we denote by f and whose Fourier coefficients are

$$c_n(f) = \begin{cases} 0 & n \text{ even} \\ \frac{2}{\pi n}(-1)^{\frac{n-1}{2}} & n \text{ odd}\,. \end{cases}$$

From the graph it's apparent that the product of f with itself is well-defined and produces the regular distribution with constant value 1. The Fourier coefficients are evidently all zero apart from the zeroth one whose value is one $c_0(f\,f) = 1$. We show that, despite the fact that the Fourier series of f is not absolutely convergent, $c_{.}(f) *_d c_{.}(f)$ produces the right answer.

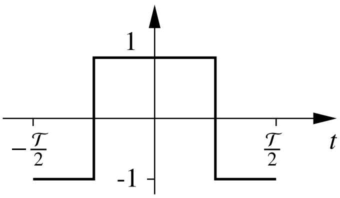


Fig. 13.3 Square regular $\mathcal{T}$-periodic distribution

First note that for n odd either $c_q(f)$ or $c_{n-q}(f)$ is zero for every value of q. Hence,

$$\big(c_.(f) *_d c_.(f)\big)_n = 0 \qquad n \text{ odd}\,.$$

For n even the convolution product is

$$\begin{aligned}\big(c_.(f) *_d c_.(f)\big)_n &= \sum_{k=-\infty}^{\infty} \frac{2}{\pi(2k+1)}(-1)^k \frac{2}{\pi(n-(2k+1))}(-1)^{\frac{n-2(k+1)}{2}}\\ &= (-1)^{\frac{n}{2}-1}\Big(\frac{2}{\pi}\Big)^2 \sum_{k=-\infty}^{\infty} \frac{1}{(2k+1)(n-(2k+1))}\,.\end{aligned}$$

For the particular case $n=0$ the summation in the last expression can be written as

$$\sum_{k=-\infty}^{\infty} \frac{-1}{(2k+1)^2} = -2\sum_{k=0}^{\infty} \frac{1}{(2k+1)^2} = -\frac{\pi^2}{4}\,.$$

The zeroth coefficient is therefore

$$\big(c_.(f) *_d c_.(f)\big)_0 = \Big(\frac{2}{\pi}\Big)^2 \frac{\pi^2}{4} = 1\,.$$

To evaluate the Fourier coefficient for $n \neq 0$ it's convenient to rewrite the summation as

$$\sum_{k=-\infty}^{\infty} \frac{1}{(2k+1)(n-(2k+1))} = \sum_{k=-\infty}^{\infty} \frac{1/n}{2k+1} + \frac{1/n}{(n-(2k+1))}\,.$$

In this form it's apparent that for each value of n all terms cancel in pair (the kth with the $(n/2+k)$th), thus giving

$$\big(c_.(f) *_d c_.(f)\big)_n = 0 \qquad n \neq 0 \text{ even}\,.$$

13.2.3 Response to Multi-tones

Consider a weakly nonlinear periodically time-varying system described by the differential equation

$$L(t, D)y = N(t, D)x + c_2(t)y^2 + c_3(t)y^3 + \cdots$$

with

$$\begin{aligned} L(t, D) &= D^m + a_{m-1}(t)D^{m-1} + \cdots + a_0(t) \\ N(t, D) &= b_n(t)D^n + b_{n-1}(t)D^{n-1} + \cdots + b_0(t) \end{aligned}$$

and where all coefficients a_i, b_i and c_i are smooth $\mathcal{T}$-periodic functions. We assume that the system is driven by N complex tones

$$x(t) = A_1 e^{\jmath\omega_1 t} + \cdots + A_N e^{\jmath\omega_N t}$$

with $A_1, \ldots, A_N$ the phasors of the tones.

In Sect. 12.4.2 we saw that the solution of the linear part of the equation is given by

$$y_1(t) = \sum_{n=1}^{N} A_n \hat{h}_1(t, \omega_n) e^{\jmath\omega_n t}$$

with $\hat{h}_1$ the (first order) time-varying frequency response of the system. We also saw (Sect. 12.5.2) that $t \mapsto \hat{h}(t, \omega_1)$ is a $\mathcal{T}$-periodic function. Expanding it in a Fourier series, y_1 can be written as

$$y_1(t) = \sum_{n=1}^{N} A_n e^{\jmath\omega_n t} \sum_{q=-\infty}^{\infty} \hat{h}_{1,q}(\omega_n) e^{\jmath q\omega_\mathcal{T} t} .$$

y_1 is therefore a sum of tones at $q\omega_\mathcal{T} + \omega_n$.

We now solve the nonlinear equation by adding terms to y_1 in a similar way as we did for weakly nonlinear time invariant systems in Sect. 9.5. As explained in Sect. 13.1 here we must assume the existence of the powers y_1^k, $k = 2, 3, \ldots$ and the others that will appear below.

For the sake of solving the equation let's assume that the frequencies $\omega_\mathcal{T}, \omega_1, \ldots, \omega_N$ are all incommensurate. Under this assumption, the only power resulting in terms proportional to $A_j A_l e^{\jmath(\omega_j+\omega_l)t}$; $j, l = 1, \ldots, N$ is the second order one

$$\begin{aligned} c_2(t)y_1^2(t) = \sum_{|m|=2} & \frac{2!}{m!} A_1^{m_1} \cdots A_N^{m_N} e^{\jmath\omega_m t} \\ & \sum_{q=-\infty}^{\infty} \left(c_.(c_2) *_d \hat{h}_{1,.}(\omega_1)^{*_d m_1} *_d \cdots *_d \hat{h}_{1,.}(\omega_N)^{*_d m_N}\right)_q e^{\jmath q\omega_\mathcal{T} t} \end{aligned}$$

with m the multi-index $m = (m_1, \dots, m_N)$ whose elements range from 0 to k (=2) and ω_m as defined in (9.27) and repeated here for convenience

$$\omega_m = \sum_{n=1}^{N} m_n \, \omega_n = m_1\omega_1 + \cdots + m_N\omega_N \, .$$

Similarly to the time invariant case we can assume that the solution of the nonlinear differential equation includes a term of second order y_2 proportional to $A_j A_l \mathrm{e}^{\jmath(\omega_j+\omega_l)t}$; $j, l = 1, \dots, N$. y_2 can be found by retaining only those terms in the equation that are proportional to $A_j A_l \mathrm{e}^{\jmath(\omega_j+\omega_l)t}$. The resulting equation is linear with $c_2(t)y_1^2(t)$ playing the role of a source composed by tones. Exploiting linearity we can solve the equation for a single tone at $q\omega_{\mathcal{T}} + \omega_1 + \omega_2$ and combine the results at the end

$$L(t, D)\hat{g}_{2,q}(t, \omega_1, \omega_2)\mathrm{e}^{\jmath(q\omega_{\mathcal{T}}+\omega_1+\omega_2)t} = \mathrm{e}^{\jmath(q\omega_{\mathcal{T}}+\omega_1+\omega_2)t} \, .$$

$t \mapsto \hat{g}_{2,q}(t, \omega_1, \omega_2)$ is also $\mathcal{T}$-periodic and can be expanded in a Fourier series

$$\hat{g}_{2,q}(t, \omega_1, \omega_2)\mathrm{e}^{\jmath(q\omega_{\mathcal{T}}+\omega_1+\omega_2)t} = \sum_{q_2=-\infty}^{\infty} c_{q_2}(\hat{g}_{2,q})\mathrm{e}^{\jmath\left((q+q_2)\omega_{\mathcal{T}}+\omega_1+\omega_2\right)t} \, .$$

With it the second order term y_2 is given by

$$\begin{aligned} y_2(t) = \sum_{|m|=2} \frac{2!}{m!} A_1^{m_1} \cdots A_N^{m_N} \mathrm{e}^{\jmath\omega_m t} \\ \cdot \sum_{q_1=-\infty}^{\infty} \left(c_{.}(c_2) *_d \hat{h}_{1,.}(\omega_1)^{*_d m_1} *_d \cdots *_d \hat{h}_{1,.}(\omega_N)^{*_d m_N}\right)_{q_1} \mathrm{e}^{\jmath q_1\omega_{\mathcal{T}} t} \\ \cdot \sum_{q_2=-\infty}^{\infty} c_{q_2}(\hat{g}_{2,q_1})\mathrm{e}^{\jmath q_2\omega_{\mathcal{T}} t} \end{aligned}$$

which, with the change of variable $l = q_1 + q_2$, can be rewritten as

$$y_2(t) = \sum_{|m|=2} \frac{2!}{m!} A_1^{m_1} \cdots A_N^{m_N} \sum_{l=-\infty}^{\infty} \hat{h}_{2,m,l}\mathrm{e}^{\jmath(l\omega_{\mathcal{T}}+\omega_m)t}$$

with

$$\hat{h}_{2,m,l} := \sum_{q_1=-\infty}^{\infty} \left(c_{.}(c_2) *_d \hat{h}_{1,.}(\omega_1)^{*_d m_1} *_d \cdots *_d \hat{h}_{1,.}(\omega_N)^{*_d m_N}\right)_{q_1} c_{l-q_1}(\hat{g}_{2,q_1}) \, .$$

The second order response of the system thus consists of tones at all possible sums of two of the input tone frequencies at a time, around each of the harmonics of the fundamental frequency of the system.

The higher order responses can be calculated in a similar manner. The kth order response has the form

$$y_k(t) = \sum_{|m|=k} \frac{k!}{m!} A_1^{m_1} \cdots A_N^{m_N} \sum_{l=-\infty}^{\infty} \hat{h}_{k,m,l} \mathrm{e}^{\jmath(l\omega_{\mathcal{T}}+\omega_m)t} \tag{13.12}$$

and is composed by tones at all possible sums of k input tone frequencies at a time, around each of the harmonics of the system fundamental frequency. A comparison of the typical two tones response of LTI-. WNTI-, LPTV- and WNPTV-systems is shown in Fig. 13.4.

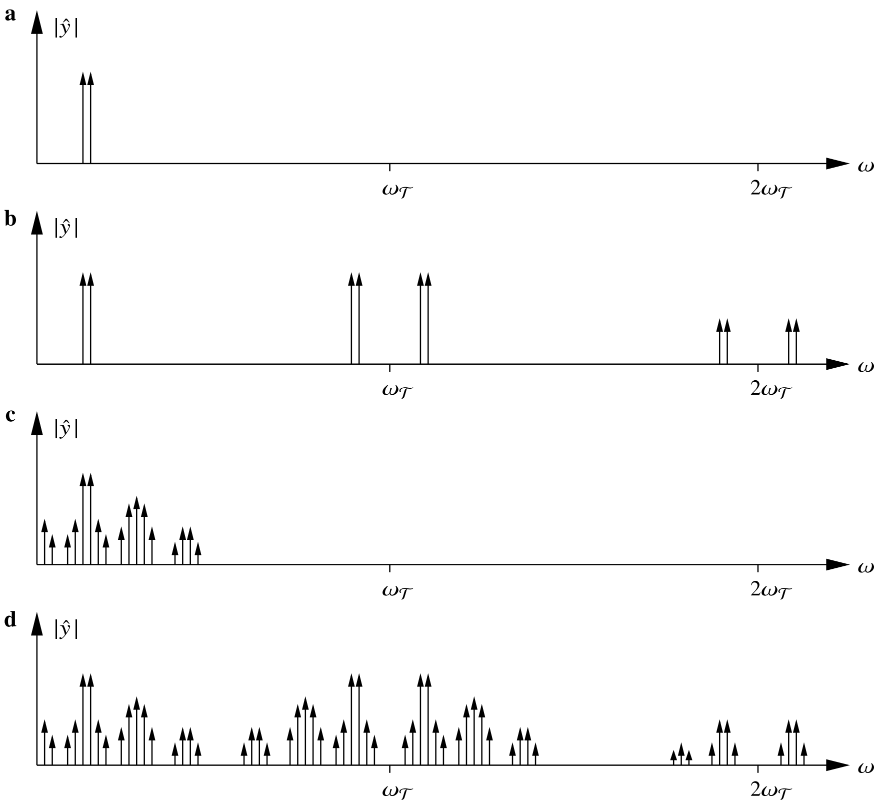


Fig. 13.4 Comparison of typical two (real) tones spectral response of LTI-, LPTV-, WNTI- and WNPTV-systems

Note that the factor

$$\sum_{l=-\infty}^{\infty} \hat{h}_{k,m,l}\mathrm{e}^{\jmath l\omega_{\mathcal{T}} t}$$

appearing in the kth order response y_k is the Fourier series of the time-varying kth order nonlinear frequency response of the system $\hat{h}_k(t, \omega_1, \ldots, \omega_k)$. It is related to $\hat{h}_{k,m,l}$ by

$$\hat{h}_{k,m,l} = c_l\big(\hat{h}_k(t, \underbrace{\omega_1, \ldots, \omega_1}_{m_1}, \ldots, \underbrace{\omega_N, \ldots, \omega_N}_{m_N})\big)\,, \qquad |m| = k\,.$$

Chapter 14
Periodically Switched Circuits

This chapter is devoted to illustrating applications of the theory of weakly nonlinear time-varying systems with practical examples. After introducing the class of electrical networks called switched circuits, we analyse in details some instantiations. We analyse a practical implementation of a quadrature modulator, including its distortion and other aspects of fundamental importance in applications. As another example of the usefulness of time-varying circuits, we illustrate how they allow implementing highly selective filters that are otherwise unfeasible in small integrated form.

14.1 Switched Circuits

An important class of circuits that finds many applications is the *Switched circuits* one. These are circuits whose only time varying components are switches. Despite the fact that switches are time-varying resistors, this class of systems can be analysed as a sequence of time invariant circuits, each one valid over an interval over which all switches remain in the same state.

Let O denote an open interval of $\mathbb{R}^n$. The space $\mathcal{D}_O$ of test functions ϕ with support contained in O is a vector subspace of $\mathcal{D}$. A distribution on O is a distribution defined on $\mathcal{D}_O$. The vector space of distributions defined on O is denoted by $\mathcal{D}'{}_O$.

Consider a linear switched circuit including at least a capacitor or an inductor. Without loss of generality we can assume the network to be driven by a single independent source x (superposition principle). Let $t_i, i \in \mathbb{N}$ denote the times at which any of the ideal switches changes state and O_i the open intervals (t_i, t_{i+1}). In any of these intervals the network is described by a system of first order linear differential equations

$$Du = A_i u + B_i x$$

F. Beffa, *Weakly Nonlinear Systems*, Understanding Complex Systems,
https://doi.org/10.1007/978-3-031-40681-2_14

with u the state of the network that we can represent by the voltage across the capacitors and the currents through the inductors. Suppose that the network is in the zero state and that we apply a Dirac impulse at time τ with $t_i < \tau < t_{i+1}$. Then, for $\tau < t < t_{i+1}$ the state u evolves as a continuous function. At time t_{i+1} some of the switches change state. If we exclude circuits including closed loops composed exclusively by ideal inductors, ideal voltage sources and (closed) ideal switches, then at this time one of the following happens

- If an ideal switch across a capacitor is closed, then the voltage across that capacitor immediately after t_{i+1} becomes zero: $v_C(t_{i+1}+) = 0$. Since charge must be conserved, this change in state must be accompanied by a current impulse $v_C(t_{i+1}-)C\delta(t - t_{i+1})$ discharging the capacitor through the switch.
- If the closing of a switch forms a closed loop formed exclusively of (ideal) capacitors then at time t_{i+1} the charge in the capacitors will instantly redistribute under the constraint of charge and energy conservation. This will be accompanied by current impulses through the capacitors. For example, if at t_{i+1} two capacitors with capacitance C_1 and C_2 respectively are connected in parallel by an ideal switch then the voltage across the capacitors immediately after the closing of the switch will be

$$v_1(t_{i+1}+)[C_1 + C_2] = v_2(t_{i+1}+)[C_1 + C_2] = v_1(t_{i+1}-)C_1 + v_2(t_{i+1}-)C_2 .$$

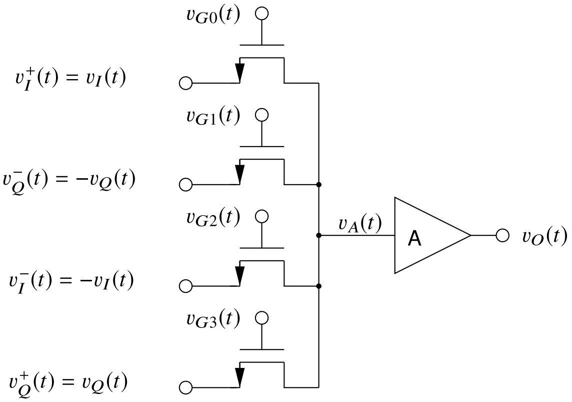
- If an ideal switch in series with an inductor is opened then the current through the inductor immediately after t_{i+1} becomes zero: $i_L(t_{i+1}+) = 0$. Faraday's law implies that this change in state is accompanied by a voltage impulse $-i_L(t_{i+1}-)L\delta(t - t_{i+1})$ across the inductor.
- In all other cases the voltages across the capacitors and the currents through the inductors remain unchanged. In other words the state component u_m representing any of those quantities immediately after t_{i+1} must equal the state component before that time instant: $u_m(t_{i+1}+) = u_m(t_{i+1}-)$.

These conditions specify initial conditions for the interval O_{i+1} that, together with the differential equation, allow to calculate the evolution of the state u of the network in that interval. The same arguments apply to all subsequent switching times. The state u is therefore fully determined and can be extended to a distribution on the hole of $\mathbb{R}$. The fundamental kernel W is therefore well defined for $\tau \in O_i, i \in \mathbb{N}$.

The only problematic cases are when τ coincides with one of the switching times. To work around this problem we limit the set of allowed input signals to the set of regular bounded distributions. Then, assuming further x to be right-sided, the state of the circuit can be represented by the integral

$$u(t) = \int_0^t W(t, \tau)B(\tau)x(\tau)\mathrm{d}\tau .$$

Fig. 14.1 Voltage-mode quadrature modulator



Since the values of τ at which W is not defined is a set of zero measure, the output is well-defined at all times.

The above discussion shows that for switched circuits the computation of the time-varying impulse response is a straightforward process. In addition, since the impulse response in each interval O_i corresponds to the one of an LTI system, the computation of the time varying frequency response by Fourier transformation of $h(t, \xi)$ does not pose any problem.

Linear periodically switched circuits are linear switched circuits in which the operation of the switches is periodic. For these circuits the time-varying impulse response $h(t, \xi)$ and the time-varying frequency response $\hat{h}(t, \omega)$ are periodic in time.

In the following we illustrate the use of this technique to analyse idealised versions of some practical periodically switched circuits used in communication receivers and transmitters.

14.2 Voltage-Mode Quadrature Modulator

In this section we analyse an implementation of the quadrature modulator of Example 12.9 suitable for realisation in a CMOS technology and shown in Fig. 14.1. The input signals v_I and v_Q (called r and q in Example 12.9) are applied differentially. We assume the LO signals to be non-overlapping and to have *very fast edges*. In fact we model the gate signals as having rectangular waveform high 25% of the time and with a relative delay among them of $\mathcal{T}/4$. We assume further that when the corresponding LO signal is high each MOSFET can be modeled as a resistor of value r_{ON}, while when the LO signal is low it can be modeled as an open circuit (infinite resistance). We are interested in the signal v_A at the input of the amplifier following the switching transistors and assume that the input impedance of the latter can be adequately modeled as a capacitor.

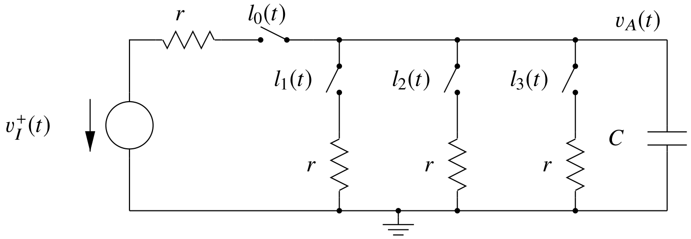


Fig. 14.2 Voltage-mode quadrature modulator model

Under these assumptions the circuit is *linear*. Hence, we can analyse the contribution to the output of each input signal independently. Figure 14.2 shows the model used to analyse the contribution of signal v_I^+ where we have combined r_{ON} with the source resistance (assumed equal at all inputs)

$$r = r_{ON} + R_S$$

and where we assume the switches to be closed when the corresponding LO control signal $l_i, i = 0, \ldots, 3$ is high and open when low. The control signals are defined by

$$l_i(t) := l(t - i\mathcal{T}/4) \qquad i = 0, \ldots, 3$$

with l the signal introduced in Example 12.10 and shown in Fig. 12.10a with $\tau = \mathcal{T}/4$.

14.2.1 Single Input Response

In this subsection we analyse the response to v_I^+. To compute its contribution to the output, we apply an input impulse at time τ. If the impulse is applied when the input switch is open, then its contribution is zero. If it's applied when the switch is closed, $\tau \in (-\mathcal{T}/8, \mathcal{T}/8) \mod \mathcal{T}$, its contribution is described by the differential equation

$$(1 + rCD)v_A = \delta(t - \tau) .$$

Note that the capacitor C is connected in parallel with a resistor of value r at all times! The fundamental kernel is therefore

$$W(t, \tau) = \omega_{3dB}\, \mathrm{e}^{-\omega_{3dB}(t-\tau)}\, 1_+(t - \tau)\, l_0(\tau)\, , \qquad \omega_{3dB} = \frac{1}{rC}$$

and can be interpreted as the impulse response of an LTI system whose input signal is $v_I^+(t)l_0(t)$.

The time-varying impulse response of the system can be obtained from the above fundamental kernel by applying the variable transformation $\xi = t - \tau$

$$h(t,\xi) = \omega_{3dB}\,\mathrm{e}^{-\omega_{3dB}\xi}\,1_+(\xi)\,l_0(t-\xi)\,.$$

We compute the response of the system to a complex tone through the time-varying frequency response. The latter is obtained by Fourier transforming $h(t,\xi)$ with respect to ξ

$$\hat{h}(t,\omega) = \int_{-\infty}^{\infty} h(t,\xi)\,\mathrm{e}^{-\jmath\omega\xi}\,\mathrm{d}\xi\,.$$

Using the Fourier series of l_0

$$l_0(t) = \sum_{n=-\infty}^{\infty} a_n \mathrm{e}^{\jmath n\omega_{\mathcal{T}} t}\,, \qquad a_n = a_{-n} = \begin{cases} \frac{1}{4} & n = 0 \\ \frac{1}{\pi n}\sin(n\frac{\pi}{4}) & n > 0 \end{cases}$$

where $\omega_{\mathcal{T}} = 2\pi/\mathcal{T}$, we have

$$\begin{aligned} \hat{h}(t,\omega) &= \omega_{3dB}\int_0^{\infty} \mathrm{e}^{-(\omega_{3dB}+\jmath\omega)\xi} \sum_{n=-\infty}^{\infty} a_n \mathrm{e}^{\jmath n\omega_{\mathcal{T}}(t-\xi)}\,\mathrm{d}\xi \\ &= \omega_{3dB}\sum_{n=-\infty}^{\infty} a_n \int_0^{\infty} \mathrm{e}^{-[\omega_{3dB}+\jmath(\omega+n\omega_{\mathcal{T}})]\xi}\,\mathrm{d}\xi\;\mathrm{e}^{\jmath n\omega_{\mathcal{T}} t} \\ &= \sum_{n=-\infty}^{\infty} \hat{h}_n(\omega)\,\mathrm{e}^{\jmath n\omega_{\mathcal{T}} t} \end{aligned} \tag{14.1}$$

with

$$\begin{aligned} \hat{h}_n(\omega) &= \omega_{3dB} a_n \int_0^{\infty} \mathrm{e}^{-[\omega_{3dB}+\jmath(\omega+n\omega_{\mathcal{T}})]\xi}\,\mathrm{d}\xi \\ &= \frac{a_n}{1+\jmath\frac{\omega+n\omega_{\mathcal{T}}}{\omega_{3dB}}}\,. \end{aligned} \tag{14.2}$$

The response of the system to a complex tone of angular frequency ω is thus

$$v_A(t) = \sum_{n=-\infty}^{\infty} \hat{h}_n(\omega)\,\mathrm{e}^{\jmath(\omega+n\omega_{\mathcal{T}})t} \tag{14.3}$$

and, as remarked above, is seen to be equal the response of an LTI system with transfer function

$$H(s) = \frac{1}{1 + \frac{s}{\omega_{3dB}}} \tag{14.4}$$

to the input

$$v_I^+(t) l_0(t) = \sum_{n=-\infty}^{\infty} a_n \mathrm{e}^{\jmath(\omega + n\omega_{\mathcal{T}})t} \,.$$

14.2.2 Input Current and Switched Resistor

Before combining the outputs from the four input signals v_I^+, v_I^-, v_Q^+ and v_Q^- we compute the input current drawn from the input v_I^+ when the other sources are disabled. When the input switch is closed, the input current is equal to the current flowing into the capacitor, while when the switch is open, the current is zero

$$i_S(t) = l_0(t) i_C(t), \qquad i_C(t) = C D v_I^+(t) \,.$$

The current is therefore given by

$$\begin{aligned} i_S(t) &= \left| \sum_{n=-\infty}^{\infty} a_n \mathrm{e}^{\jmath n \omega_{\mathcal{T}} t} \right| \left| \sum_{n=-\infty}^{\infty} \frac{a_n}{r} \frac{\jmath \frac{(\omega + n\omega_{\mathcal{T}})}{\omega_{3dB}}}{1 + \jmath \frac{\omega + n\omega_{\mathcal{T}}}{\omega_{3dB}}} \mathrm{e}^{\jmath(\omega + n\omega_{\mathcal{T}})t} \right| \\ &= \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} \frac{a_m a_n}{r} \frac{\jmath \frac{(\omega + n\omega_{\mathcal{T}})}{\omega_{3dB}}}{1 + \jmath \frac{\omega + n\omega_{\mathcal{T}}}{\omega_{3dB}}} \mathrm{e}^{\jmath[\omega + (n+m)\omega_{\mathcal{T}}]t} \\ &= \sum_{k=-\infty}^{\infty} y_k(\omega)\, \mathrm{e}^{\jmath(\omega + k\omega_{\mathcal{T}})t} \end{aligned} \tag{14.5}$$

with

$$y_k(\omega) := \sum_{n=-\infty}^{\infty} a_{k-n} a_n \frac{\jmath(\omega + n\omega_{\mathcal{T}})C}{1 + \jmath \frac{\omega + n\omega_{\mathcal{T}}}{\omega_{3dB}}} \tag{14.6}$$

where in the last step we made the substitution $k = m + n$.

Let's consider more closely the term for $k = 0$. Substituting the expression for a_n we obtain

$$y_0(\omega) = \jmath\omega \frac{C}{16} + \sum_{n \neq 0} \frac{\sin^2(n\frac{\pi}{4})}{(\pi n)^2} \frac{\jmath(\omega + n\omega_{\mathcal{T}})C}{1 + \jmath \frac{\omega + n\omega_{\mathcal{T}}}{\omega_{3dB}}} \,.$$

To find an approximate value for this series it's useful to separate real- and imaginary-parts

$$y_0(\omega) = g_0(\omega) + \jmath b_0(\omega) \,.$$

We start by simplifying the imaginary part

$$b_0(\omega) = \omega\frac{C}{16} + \sum_{n\neq 0} \frac{\sin^2(n\frac{\pi}{4})}{(\pi n)^2} \frac{(\omega + n\omega_{\mathcal{T}})C}{1+\left(\frac{\omega+n\omega_{\mathcal{T}}}{\omega_{3dB}}\right)^2}$$

The first thing to note is that the terms proportional to $n\omega_{\mathcal{T}}$ with $n > 0$ cancel with the ones for $n < 0$ so that we obtain

$$b_0(\omega) = \omega C\left[\frac{1}{16} + \sum_{n\neq 0} \frac{\sin^2(n\frac{\pi}{4})}{(\pi n)^2} \frac{1}{1+\left(\frac{\omega+n\omega_{\mathcal{T}}}{\omega_{3dB}}\right)^2}\right]. \tag{14.7}$$

All terms in the square bracket are positive. If we assume $|\omega| \ll \omega_{\mathcal{T}} < \omega_{3dB}$ the terms decrease as $1/n^2$ for $n < \omega_{\mathcal{T}}/\omega_{3dB}$ and as $1/n^4$ for larger values of n. We can therefore bound the series by

$$\sum_{n\neq 0} \frac{\sin^2(n\frac{\pi}{4})}{(\pi n)^2} \frac{1}{1+\left(\frac{\omega+n\omega_{\mathcal{T}}}{\omega_{3dB}}\right)^2} < \sum_{n\neq 0} \frac{1}{(\pi n)^2}\,.$$

Using the known result

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$$

we thus obtain the upper bound

$$\frac{b_0(\omega)}{\omega C} < \frac{1}{16} + 2\sum_{n=1}^{\infty} \frac{1}{(\pi n)^2} = \frac{19}{48} \approx 0.40\,.$$

This bound is tighter for large values of $\omega_{3dB}/\omega_{\mathcal{T}}$.

To obtain a value closer to the actual value of the series we note that for $N \gg 1$

$$\sum_{n=1}^{N} \sin^2(n\frac{\pi}{4}) \approx \frac{N}{2}. \tag{14.8}$$

Instead of bounding $\sin^2(n\pi/4)$ by 1, we approximate its value by $1/2$ independently of n to obtain

$$\frac{b_0(\omega)}{\omega C} < \frac{1}{16} + \frac{1}{6} = \frac{11}{48} \approx 0.23\,.$$

Figure 14.3 shows the normalized value of b_0 as a function of $\omega_{3dB}/\omega_{\mathcal{T}}$ computed from (14.7). We see that, while the argument to obtain the above approximate value

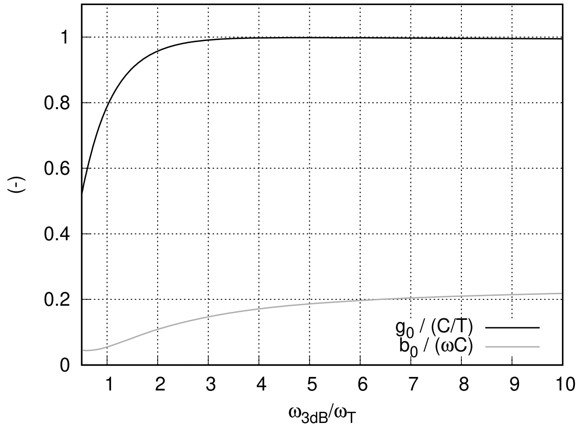


Fig. 14.3 Normalized values of the real- and imaginary parts of $y_0(0.1\omega_{\mathcal{T}})$ as a function of the ratio $\omega_{3dB}/\omega_{\mathcal{T}}$

is questionable, for large values of $\omega_{3dB}/\omega_{\mathcal{T}}$ the approximation is remarkably close to the real value.

We next turn to the real part of $y_0(\omega)$

$$g_0(\omega) = \frac{1}{r} \sum_{n \neq 0} \frac{\sin^2(n\frac{\pi}{4})}{(\pi n)^2} \frac{\left(\frac{\omega + n\omega_{\mathcal{T}}}{\omega_{3dB}}\right)^2}{1 + \left(\frac{\omega + n\omega_{\mathcal{T}}}{\omega_{3dB}}\right)^2} \tag{14.9}$$

If we assume $|\omega| \ll \omega_{\mathcal{T}} < \omega_{3dB}$ then to a good approximation we have

$$g_0(\omega) \approx \frac{2}{r} \sum_{n=1}^{\infty} \frac{\sin^2(n\frac{\pi}{4})}{\pi^2} \frac{\left(\frac{\omega_{\mathcal{T}}}{\omega_{3dB}}\right)^2}{1 + \left(\frac{n\omega_{\mathcal{T}}}{\omega_{3dB}}\right)^2}$$

If $\omega_{\mathcal{T}}/\omega_{3dB} \ll 1$ then the quadratic term in the denominator can be neglected in a large number of terms up to approximately $N = \lceil \frac{\omega_{3dB}}{\omega_{\mathcal{T}}} \rceil$. The first N terms of the series contribute the largest part of its total value. Therefore, referring again to the approximation (14.8), we approximate again $\sin^2(n\pi/4)$ by $1/2$. Instead of neglecting the terms for $n > N$ we approximate the series by the integral

$$\frac{1}{\pi^2 r} \frac{\omega_{\mathcal{T}}}{\omega_{3dB}} \int_0^{\infty} \frac{1}{1 + x^2} \, \mathrm{d}x$$

with

$$n \frac{\omega_{\mathcal{T}}}{\omega_{3dB}} \to x, \qquad \frac{\omega_{\mathcal{T}}}{\omega_{3dB}} \to \mathrm{d}x.$$

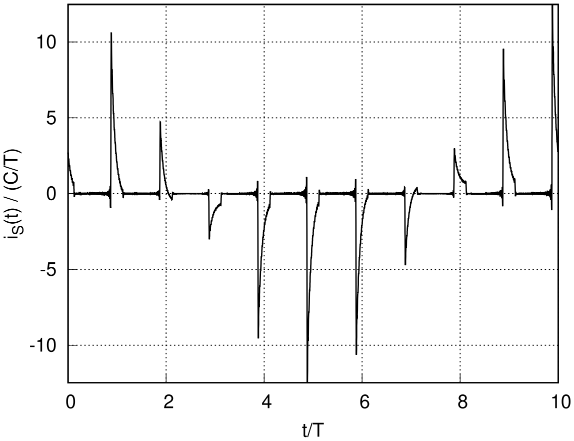


Fig. 14.4 Normalized $i_S(t)$ for a cosinusoidal input with $\omega = 0.1\omega_\mathcal{T}$, $\omega_{3dB} = 2\omega_\mathcal{T}$ computed using (14.5) truncated at $|k| = 60$ and $|n| = 200$

This integral is easily solved and we finally obtain

$$g_0(\omega) \approx \frac{1}{r\pi^2}\frac{\omega_\mathcal{T}}{\omega_{3dB}}\frac{\pi}{2} = \frac{C}{\mathcal{T}}\,.$$

Figure 14.3 shows the normalized value of g_0 as a function of $\omega_{3dB}/\omega_\mathcal{T}$ computed from Eq. (14.9). For $\omega_{3dB}/\omega_\mathcal{T} > 3$ it's in very good agreement with the given approximation.

Figure 14.4 shows the normalized current $i_S(t)$ for a cosinusoidal input with $\omega = 0.1\omega_\mathcal{T}$ and $\omega_{3dB} = 2\omega_\mathcal{T}$. The curve consists of peaks in concomitance with the closing instants of switch 0, followed by an exponential decay with a time constant of approximately $1/\omega_{3dB}$ and a sudden jump to zero at the instants where switch 0 is opened. (The oscillations around the instants where switch 0 changes state are due to the Gibbs phenomenon of the Fourier series.) As the time constant is shortened by reducing the value of r, the curve converges to a series of Dirac pulses at the closing instants of switch 0. If we shift the closing instants of switch 0 at multiples of $\mathcal{T}$ we can express this behavior by

$$\lim_{\substack{r\to 0\\ \omega\to 0}} i_S\left(t - \frac{\mathcal{T}}{8}\right) = \sum_{k=-\infty}^{\infty} \frac{C}{\mathcal{T}} \mathrm{e}^{\jmath k\omega_\mathcal{T} t}\,.$$

The discrete spectrum of $i_S(t - \mathcal{T}/8)$ for $\omega_{3dB} = 3\omega_\mathcal{T}, 20\omega_\mathcal{T}$ and $\omega = 0.1\omega_\mathcal{T}$ is shown in Fig. 14.5. The figure shows that as the value of $\omega_{3dB}/\omega_\mathcal{T}$ is increased, an increasing number of coefficients $y_k(\omega)$ tend to approach the value of $C/\mathcal{T}$ as expected. At a value of $k \approx \omega_{3dB}/\omega_\mathcal{T}$ the real and imaginary parts have roughly the same value and for larger values of k the magnitude of $y_k(\omega)$ decreases.

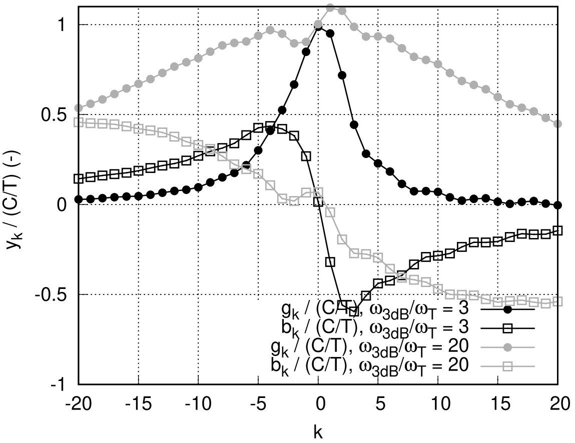


Fig. 14.5 Normalized real- and imaginary-part of $y_k(\omega)\mathrm{e}^{-\jmath(\omega+k\omega_{\mathcal{T}})\mathcal{T}/8}$ for $\omega = 0.1\omega_{\mathcal{T}}$ as a function of k computed with (14.6) truncated to $|n| = 2000$. The points between the discrete values of k were joined to better highlight the trend

14.2.3 Full Response

We now go back to the voltage across the capacitor v_A and calculate the combined response to all four signals v_I^+, v_I^-, v_Q^+ and v_Q^-. To distinguish the four responses we will add a subscript equal to the one of the corresponding LO signal. Thus in the following we will denote the response to the signal v_I^+ given in (14.3) by $v_{A,0}$. The contribution of v_I^- differs from the one of v_I^+ by (i) a shift by $\mathcal{T}/2$ in the LO waveform and (ii) a reversal of sign of the input signal. Its contribution to v_A is therefore

$$\begin{aligned} v_{A,2}(t) &= -\mathrm{e}^{\jmath\omega t}\sum_{n=-\infty}^{\infty}\hat{h}_n(\omega)\,\mathrm{e}^{\jmath n\omega_{\mathcal{T}}(t-\mathcal{T}/2)} \\ &= \sum_{n=-\infty}^{\infty}(-1)^{n+1}\hat{h}_n(\omega)\,\mathrm{e}^{\jmath(\omega+n\omega_{\mathcal{T}})t}\,. \end{aligned}$$

Note that the even harmonics have opposite sign compared to the ones of $v_{A,0}$, while the odd ones have the same sign. Therefore the combined response of v_I^1 and v_I^- consists of odd harmonics only

$$v_{A,0}(t) + v_{A,2}(t) = 2\sum_{n\ \text{odd}}\hat{h}_n(\omega)\,\mathrm{e}^{\jmath(\omega+n\omega_{\mathcal{T}})t}\,.$$

The response to the signal v_Q^+ differs from the one to v_I^+ by (i) a shift by $-\mathcal{T}/4$ of the LO signal and (ii) a shift of $\mathcal{T}/4$ in the input signal

$$v_Q^+ = -\jmath\mathrm{e}^{\jmath\omega t}\,.$$

Its contribution to v_A is therefore

$$\begin{aligned} v_{A,3}(t) &= -\jmath \mathrm{e}^{\jmath\omega t} \sum_{n=-\infty}^{\infty} \hat{h}_n(\omega)\, \mathrm{e}^{\jmath n\omega_{\mathcal{T}}(t+\mathcal{T}/4)} \\ &= -\sum_{n=-\infty}^{\infty} \jmath^{n+1} \hat{h}_n(\omega)\, \mathrm{e}^{\jmath(\omega+n\omega_{\mathcal{T}})t}\,. \end{aligned}$$

Similarly, the response to the signal v_Q^- differs from the one to v_I^+ by (i) a shift by $\mathcal{T}/4$ of the LO signal and (ii) a shift of $-\mathcal{T}/4$ in the input signal

$$\begin{aligned} v_{A,1}(t) &= \jmath \mathrm{e}^{\jmath\omega t} \sum_{n=-\infty}^{\infty} \hat{h}_n(\omega)\, \mathrm{e}^{\jmath n\omega_{\mathcal{T}}(t-\mathcal{T}/4)} \\ &= -\sum_{n=-\infty}^{\infty} (-\jmath)^{n+1} \hat{h}_n(\omega)\, \mathrm{e}^{\jmath(\omega+n\omega_{\mathcal{T}})t}\,. \end{aligned}$$

Again we note that odd harmonics of $v_{A,3}$ and $v_{A,1}$ have the same sign, while even ones have opposite sign. The combined response of these two signals is therefore also composed of odd harmonics only

$$\begin{aligned} v_{A,3}(t) + v_{A,1}(t) &= -2 \sum_{n \text{ odd}} \jmath^{n+1} \hat{h}_n(\omega)\, \mathrm{e}^{\jmath(\omega+n\omega_{\mathcal{T}})t} \\ &= -2 \sum_{n \text{ odd}} (-1)^{(n+1)/2} \hat{h}_n(\omega)\, \mathrm{e}^{\jmath(\omega+n\omega_{\mathcal{T}})t}\,. \end{aligned}$$

We now combine the two partial sums $v_{A,0} + v_{A,2}$ and $v_{A,3} + v_{A,1}$. The terms for $n = 1 + 4m$, $m \in \mathbb{Z}$ have the same sign, while the terms at $n = -1 + 4m$ have opposite sign. The total sum is therefore

$$\begin{aligned} v_A(t) &= v_{A,0}(t) + v_{A,2}(t) + v_{A,3}(t) + v_{A,1}(t) \\ &= 4 \sum_{m=-\infty}^{\infty} \hat{h}_{1+4m}(\omega)\, \mathrm{e}^{\jmath[\omega+(1+4m)\omega_{\mathcal{T}}]t}\,. \end{aligned}$$

Note again that the response of the system is equal to the one of an LTI system with the transfer function given by (14.4) and driven by the input signal

$$x(t) = v_I^+(t)\, l_0(t) + v_I^-(t)\, l_2(t) + v_Q^+(t)\, l_3(t) + v_Q^-(t)\, l_1(t)\,.$$

Using four signal paths (two differential) this quadrature modulator cancels three spurious emission tones every four. It's a transmitter implementation of the harmonic-reject mixer presented in Example 12.10 where we showed that to suppress more harmonics requires a larger number of signal paths.

The above derivation of the output signal highlights the fact that suppression of harmonics relies on exact cancelling of strong tones. We will investigate some imperfections limiting the amount of cancelling achievable in practical implementations in later sections. Before turning to that question we investigate the effect called carrier leakage.

14.2.4 Carrier Leakage

Carrier leakage refers to the presence of a tone at $\pm\omega_{\mathcal{T}}$ in the output spectrum of the modulator. In transmitters, it is one of the undesired tones close to the signal of interest (or, depending on the architecture, indeed overlapping with the modulated wanted signal) that can't be easily filtered. It is caused by the presence of small, DC offset voltages at the inputs of the modulator. These offset voltages are the result of mismatch in the driving circuits and are therefore Gaussian random variables.

We denote the DC offset random variables by $X_k, k = 0, \ldots, 3$ where the index matches the one of the corresponding switch. They form the following signal at the input of the equivalent LTI system

$$\begin{aligned} &X_0\, l_0(t) + X_1\, l_1(t) + X_2\, l_2(t) + X_3\, l_3(t) \\ &\quad = \sum_{n=-\infty}^{\infty} \big(X_0 + X_1 \mathrm{e}^{-\jmath\frac{\pi}{2}n} + X_2 \mathrm{e}^{-\jmath\pi n} + X_3 \mathrm{e}^{-\jmath\frac{3\pi}{2}n}\big) a_n \mathrm{e}^{\jmath n\omega_{\mathcal{T}} t} \\ &\quad = \sum_{n=-\infty}^{\infty} \big(X_0 + X_1(-\jmath)^n + X_2(-1)^n + X_3(\jmath)^n\big) a_n \mathrm{e}^{\jmath n\omega_{\mathcal{T}} t}\,. \end{aligned} \tag{14.10}$$

Since usually the most problematic tone is the one at $\pm\omega_{\mathcal{T}}$ we only consider the terms for $n = -1, 1$. The term for $n = 1$ is

$$\big[X_0 - X_2 + \jmath(X_3 - X_1)\big] a_1 \mathrm{e}^{\jmath\omega_{\mathcal{T}} t}$$

and the one at $n = -1$ is its conjugate complex. The sum of the two terms gives

$$2a_1\big[X_c \cos(\omega_{\mathcal{T}} t) - X_s \sin(\omega_{\mathcal{T}} t)\big], \qquad X_c = X_0 - X_2\,, \quad X_s = X_3 - X_1\,.$$

Linear combinations of independent Gaussian random variables are Gaussian. Therefore, if we assume $X_k, k = 0, \ldots, 3$ to be independent of each other, X_c and X_s are independent Gaussian random variables as well. We denote the standard deviation of X_c and X_s by σ_X. Their joint probability density function (PDF) is

$$p_{X_c,X_s}(x_c, x_s) = p_{X_c}(x_c)\, p_{X_s}(x_s) = \frac{1}{2\pi\sigma_X^2} \mathrm{e}^{-\frac{x_c^2 + x_s^2}{2\sigma_X}}\,.$$

It is now convenient to pass to polar random variables. Specifically, using the relation

$$\cos(\omega t + \phi) = \cos(\phi)\cos(\omega t) - \sin(\phi)\sin(\omega t)$$

the sum of the input terms for $n = 1$ and -1 can be rewritten as

$$2a_1 X_r \cos(\omega_{\mathcal{T}} t + X_\phi)$$

with the new polar random variables

$$X_r = \sqrt{X_c^2 + X_s^2}$$
$$X_\phi = \arctan\frac{X_s}{X_c}\,.$$

Given that the probability density in terms of X_c, X_s must agree with the one in terms of X_r, X_ϕ, we must have

$$p_{X_c,X_s}(x_c, x_s)\mathrm{d}x_c\mathrm{d}x_s = p_{X_r,X_\phi}(x_r, x_\phi)\mathrm{d}x_r\mathrm{d}x_\phi\,.$$

From this equation and $\mathrm{d}x_c\mathrm{d}x_s = x_r\mathrm{d}x_r\mathrm{d}x_\phi$ we therefore deduce

$$p_{X_r,X_\phi}(x_r, x_\phi) = \frac{x_r}{2\pi\sigma_X^2}\mathrm{e}^{-\frac{x_r^2}{2\sigma_X^2}}\,.$$

This joint probability density function is easily factored

$$p_{X_r,X_\phi}(x_r, x_\phi) = p_{X_r}(x_r)p_{X_\phi}(x_\phi)$$

which implies that X_r and X_ϕ are independent random variables with the following probability density functions

$$p_{X_r}(x_r) = \frac{1}{\sigma_X^2}x_r\mathrm{e}^{-\frac{x_r^2}{2\sigma_X^2}}\,, \qquad x_r \geq 0 \tag{14.11}$$

$$p_{X_\phi}(x_\phi) = \begin{cases} \frac{1}{2\pi} & 0 \leq x_\phi < 2\pi \\ 0 & \text{otherwise}\,. \end{cases} \tag{14.12}$$

The phase random variable X_ϕ is uniformly distributed over the full circle. The distribution of the variable X_r is called *Rayleigh distribution*. Its PDF and complementary cumulative density function $1 - F_{X_r}(x_r)$ are plotted in Fig. 14.6. The PDF assumes its maximum at $x_r = \sigma_X$. The expected value and variance of X_r are

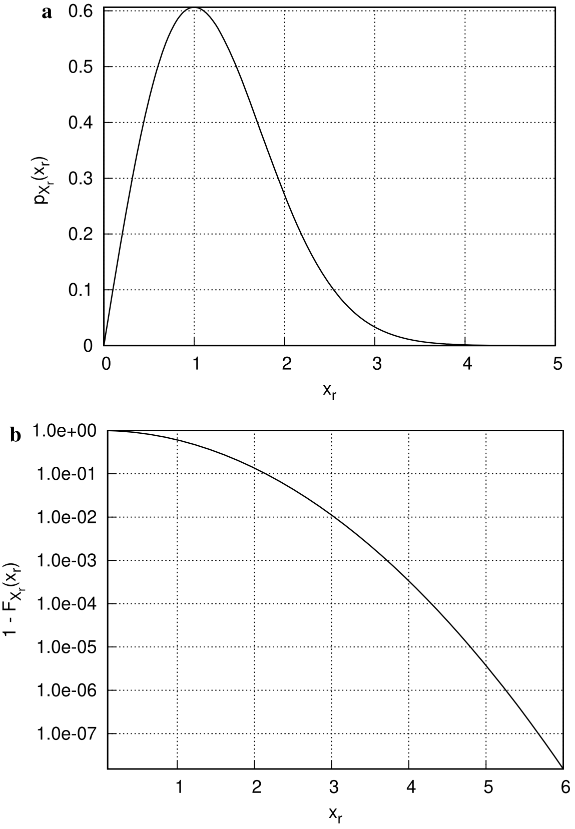


Fig. 14.6 Rayleigh distribution for $\sigma_X = 1$. **a** Probability density function. **b** Complementary cumulative density function $1 - F_{X_r}(x_r)$

$$\mathrm{E}[X_r] = \int_0^\infty x_r p_{X_r}(x_r) dx_r = \sqrt{\frac{\pi}{2}} \sigma_X$$

and

$$\mathrm{Var}(X_r) = \mathrm{E}[(x_r - \mathrm{E}[X_r])^2] = \frac{4-\pi}{2} \sigma_X^2$$

respectively.

The carrier leakage of the modulator is therefore given by

$$X_r \frac{\sqrt{2}}{\pi} \Re\{H(\jmath\omega_{\mathcal{T}}) \mathrm{e}^{\jmath(\omega_{\mathcal{T}} t + X_\phi)}\}$$

with the phase uniformly distributed over the full circle, The magnitude of the tone is Rayleigh distributed and if $\omega_{\mathcal{T}} \ll \omega_{3dB}$ has an expected value of

$$\frac{\sigma_X}{\sqrt{\pi}}$$

From Fig. 14.6 we read that, under the same assumption, 0.1% of the modulators have a carrier leakage magnitude exceeding

$$3.7\frac{\sqrt{2}}{\pi}\sigma_X \approx 1.67\sigma_X \,.$$

14.2.5 Image-Rejection

In this subsection we come back to the finite cancelling of harmonics in practical implementations. We assume again the common case of a transmitter up-converting the input signal to $\omega + \omega_{\mathcal{T}}$ with $|\omega| \ll \omega_{\mathcal{T}} \lesssim \omega_{3dB}$.

In previous calculations we assumed perfectly balanced signals $v_I^- = -v_I^+, v_Q^- = -v_Q^+$, equal amplitudes for all signals, a phase difference between v_I^+ and v_Q^+ of exactly $\pi/2$ and delays between the LO signals of exactly $\mathcal{T}/4$. If we now introduce small differences in the amplitudes

$$v_I^+(t) = (A_I + \Delta A_I/2)\mathrm{e}^{\jmath\omega t}\,, \quad v_I^-(t) = -(A_I - \Delta A_I/2)\mathrm{e}^{\jmath\omega t}$$

the even harmonics of $v_I^+(t)l_0(t) + v_I^-(t)l_2(t)$ do not cancel perfectly anymore

$$v_I^+(t)l_0(t) + v_I^-(t)l_2(t) = \sum_{n \text{ odd}} 2A_I a_n \mathrm{e}^{\jmath(n\omega_{\mathcal{T}}+\omega)t} + \sum_{n \text{ even}} \Delta A_I a_n \mathrm{e}^{\jmath(n\omega_{\mathcal{T}}+\omega)t}$$

and similarly for the signal $v_Q^+(t)l_3(t) + v_Q^-(t)l_1(t)$

$$\begin{aligned} v_Q^+(t)l_3(t) + v_Q^-(t)l_1(t) = &-\sum_{n \text{ odd}} \jmath^{n+1} 2A_Q a_n \mathrm{e}^{\jmath[(n\omega_{\mathcal{T}}+\omega)t - n\Delta\phi]} \\ &- \sum_{n \text{ even}} \jmath^{n+1} \Delta A_Q a_n \mathrm{e}^{\jmath[(n\omega_{\mathcal{T}}+\omega)t - n\Delta\phi]} \end{aligned}$$

where in addition we have added a small delay error in l_3 and l_1 of $\Delta\tau$ and set $\Delta\phi = 2\pi\Delta\tau/\mathcal{T}$. If we now sum these partial sums, the complete cancelling that was happening for three harmonics out of four becomes a partial cancelling. In particular this partial cancelling causes the appearance of a tone at $|\omega_{\mathcal{T}} - \omega|$ (for $n = -1$) which is difficult to filter as, in a single-sided representation, it appears very close to the wanted signal at $\omega_{\mathcal{T}} + \omega$. The tone at $|\omega_{\mathcal{T}} - \omega|$ is called the *image* of the wanted signal. The ratio of the magnitude of the image to the one of the signal is called the *image-reject ratio* (IRR) and is given by

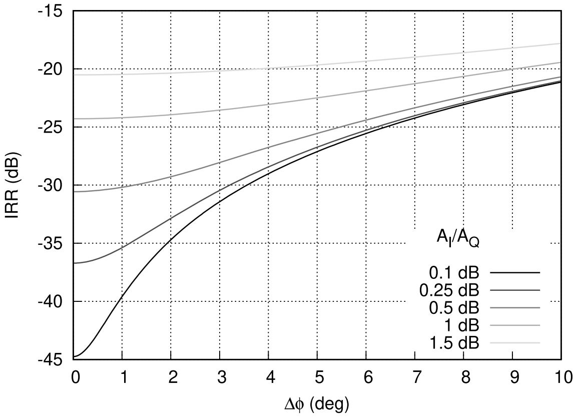


Fig. 14.7 Mixer image-reject ratio

$$
\begin{aligned}
IRR &= \left|\frac{A_I - A_Q e^{\jmath\Delta\phi}}{A_I + A_Q e^{-\jmath\Delta\phi}}\right| = \left|\frac{A_I e^{-\jmath\Delta\phi/2} - A_Q e^{\jmath\Delta\phi/2}}{A_I e^{\jmath\Delta\phi/2} + A_Q e^{-\jmath\Delta\phi/2}}\right| \\
&= \sqrt{\frac{(A_I - A_Q)^2\cos^2(\Delta\phi/2) + (A_I + A_Q)^2\sin^2(\Delta\phi/2)}{(A_I + A_Q)^2\cos^2(\Delta\phi/2) + (A_I - A_Q)^2\sin^2(\Delta\phi/2)}} \\
&\approx \sqrt{\left(\frac{A_I - A_Q}{A_I + A_Q}\right)^2 + \tan^2(\Delta\phi/2)}
\end{aligned}
\tag{14.13}
$$

where in the last step we neglected the term $(A_I - A_Q)^2\sin^2(\Delta\phi/2)$ in the denominator which is of second order in the errors. Note that part of the phase error could well come from the input signal. This is the IRR of the effective signal at the input of the LTI system $H(s)$. It is plotted in Fig. 14.7.

14.2.6 Effect of Mismatch

In this subsection we investigate the effect of mismatch which, as we will see, is another phenomenon limiting the amount of harmonic cancelling achievable in practical implementations.

Due to mismatch, each of the four transistors with which the modulator is implemented (see Figs. 14.1 and 14.2) presents a slightly different r_{ON} resistance and similarly for the four source resistors. Therefore, the value of the resistance connected to the capacitor is not independent of time, but is time-varying

$$
r(t) = \sum_{k=0}^{3} r_k\, l_k(t)\,, \qquad r_k \in \mathbb{R}, k = 0, \ldots, 3\,.
$$

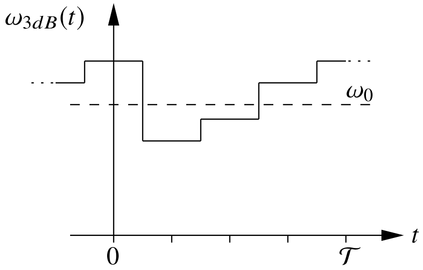


Fig. 14.8 Possible sample waveform of the time varying cut-off frequency $\omega_{3dB}(t)$. Variation greatly exaggerated for illustration

The differential equation describing the system therefore becomes

$$[D + \omega_{3dB}(t)]v_A = \omega_{3dB}(t)x(t)\,, \qquad \omega_{3dB}(t) := \frac{1}{r(t)C}$$
$$x(t) := v_I^+(t)\,l_0(t) + v_I^-(t)\,l_2(t) + v_Q^+(t)\,l_3(t) + v_Q^-(t)\,l_1(t)$$

where we denoted the sum of the input signals by x. Since the variation of the resistance from the nominal value is small, we can solve the equation using the perturbation method and proceed as in Example 12.8 (Fig. 14.8).

As a first step we develop $\omega_{3dB}(t)$ in a Fourier series and decompose it in two parts

$$\omega_{3dB}(t) = \omega_c(t) + \omega_s(t) \tag{14.14}$$

$$\omega_c(t) = r_0\,l_0(t) + r_2\,l_2(t) = \omega_{c,0} + X_c\sum_{n=1}^{\infty} w_n\cos(n\omega_{\mathcal{T}}t) \tag{14.15}$$

$$\omega_s(t) = r_1\,l_1(t) + r_3\,l_3(t) = \omega_{s,0} - X_s\sum_{n=1}^{\infty} w_n\sin(n\omega_{\mathcal{T}}t) \tag{14.16}$$

with

$$w_n = \frac{4}{\pi n}\sin(n\frac{\pi}{4})\,, \qquad n > 0\,.$$

$\omega_c(t)$ corresponds to the curve of Fig. 12.10b with $\tau = \mathcal{T}/4$ scaled by X_c plus a constant term. $\omega_s(t)$ is constructed similarly, but with the curve of Fig. 12.10b shifted by $-\mathcal{T}/4$. We use the symbols X_c and X_s to denote independent Gaussian random variables as in Sect. 14.2.4, but they are not related to the quantities of that section. We also denote again their standard deviation by σ_X.

The sum of the constant terms (which are also random variables) is the average frequency

$$\omega_0 = \omega_{c,0} + \omega_{s.0}\,.$$

The variable part of $\omega_{3dB}(t)$ can be written as

$$\sum_{n=1}^{\infty} w_n [X_c \cos(n\omega_{\mathcal{T}} t) - X_s \sin(n\omega_{\mathcal{T}} t)]$$

Proceeding as in the analysis of carrier leakage we can express it in terms of the polar random variables X_r and X_ϕ

$$\sum_{n=1}^{\infty} w_n X_r \cos(n\omega_{\mathcal{T}} t + X_\phi) .$$

Using this form for $\omega_{3dB}(t)$ the differential equation can be written as

$$(D + \omega_0)v_A = \omega_{3dB}(t)x(t) - \sum_{n=1}^{\infty} w_n X_r \cos(n\omega_{\mathcal{T}} t + X_\phi)v_A(t) .$$

We solved this equation to first order in X_r for one input tone and one cos term in Example 12.8. Referring to that example for details, we conclude that mismatch produces tones at all frequencies $n\omega_{\mathcal{T}} + \omega, n \in \mathbb{Z}$. The amplitude of these tones is proportional to the random variable X_r which is Rayleigh distributed. The phase is uniformly distributed over the full circle.

While in this subsection we focused on mismatch, the same method can be used to analyse other effects causing variations in the resistance such as overlapping LO signals.

14.2.7 Second-Order Distortion

In this subsection we analyse the distortion of second order introduced by the nonlinear characteristic of the MOSFETs. As discussed, if we neglect mismatch the circuit can be modeled as a time-invariant system drived by the input signal

$$x(t) = v_I^+(t)\, l_0(t) + v_I^-(t)\, l_2(t) + v_Q^+(t)\, l_3(t) + v_Q^-(t)\, l_1(t) . \tag{14.17}$$

To simplify the calculations we discard the source resistors R_S and consider the situation shown in Fig. 14.9a. We assume the transistor to remain in the so-called linear region of its characteristic which is described by

$$i_D = \beta(v_G - V_T)(v_D - v_S) - \frac{\beta}{2}(v_D^2 - v_S^2) .$$

In our model the gate voltage is assumed to be constant at a sufficiently high level V_G, in which case the characteristic can be modeled by the linear resistor r that we

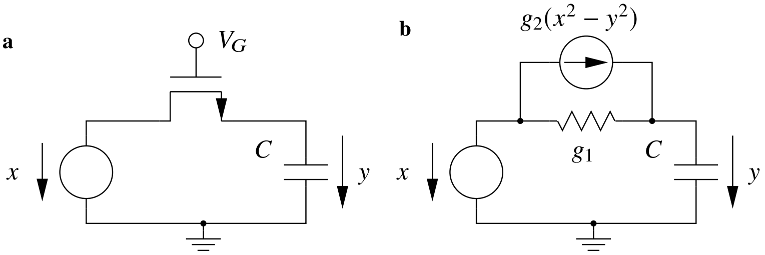


Fig. 14.9 **a** Equivalent WNTI schematic of the quadrature modulator of Fig. 14.1 **b** Equivalent WNTI circuit of the quadrature modulator of Fig. 14.1

used before and two nonlinear VCCS that we combine in a single one controlled by the two voltages v_D and v_S

$$i_D = g_1(v_D - v_S) + g_2(v_D^2 - v_S^2)$$

with

$$g_1 = \frac{1}{r} = \beta(V_G - V_T) \quad \text{and} \quad g_2 = -\frac{\beta}{2}$$

and represented in Fig. 14.9b. Using this transistor model the differential equation describing the system is

$$(1 + rCD)v_A = x + \frac{g_2}{g_1}(x^2 - v_A^2)\,.$$

The first order response of the system is described by the transfer function H that we calculated before, that we repeat here for convenience and to which we add an index representing the order as usual

$$H_1(s_1) = \frac{1}{1 + \frac{s_1}{\omega_{3dB}}}\,.$$

We compute the higher order responses by Laplace transforming the differential equation and retaining only terms of the relevant order. To obtain the transfer functions directly we use a Dirac pulse as input. The Laplace transformed of the second-order part of the differential equation is

$$[1 + rC(s_1 + s_2)]H_2(s_1, s_2) = \frac{g_2}{g_1}[1 - H_1(s_1)H_1(s_2)]\,.$$

The second-order transfer function therefore is

$$H_2(s_1, s_2) = \frac{g_2}{g_1} H_1(s_1 + s_2)[1 - H_1(s_1)H_1(s_2)] \,. \tag{14.18}$$

Consider the case in which the modulator is driven by two baseband ($v_I^+, v_I^-, v_Q^+, v_Q^-$) real tones at ω_1 and ω_2. The tones of interest in the effective input signal x of the WNTI model are at $\pm(\omega_{\mathcal{T}} \pm \omega_i)$, $i = 1, 2$. Under the assumption that $|\omega_i| \ll \omega_{\mathcal{T}} \ll \omega_{3dB}$ we can approximate H_1 at these frequencies by

$$H_1(\jmath\omega) \approx 1 - \jmath \frac{\omega}{\omega_{3dB}} \,.$$

Using this approximation in H_2 we find

$$\begin{aligned} H_2(\jmath\omega_1, \jmath\omega_2) &\approx \frac{g_2}{g_1} \frac{1 - (1 - \jmath\omega_1/\omega_{3dB})(1 - \jmath\omega_2/\omega_{3dB})}{1 + \jmath\frac{\omega_1+\omega_2}{\omega_{3dB}}} \\ &\approx \frac{-1}{2(V_G - V_T)} \frac{\jmath\frac{\omega_1+\omega_2}{\omega_{3dB}}}{1 + \jmath\frac{\omega_1+\omega_2}{\omega_{3dB}}} \,. \end{aligned} \tag{14.19}$$

where in the last step we have neglected the small quantity $\omega_1\omega_2/\omega_{3dB}^2$. This expression shows that, to reduce the principal second order distortion components under the given assumptions, it is more convenient to choose a voltage $V_G - V_T$ as large as possible than merely reduce r by using a wider transistor.

14.2.8 Third-Order Distortion

We next compute the third-order transfer function. The third-order part of the Laplace transformed differential equation is

$$[1 + rC(s_1 + s_2 + s_3)]H_3(s_1, s_2, s_3) = -2\frac{g_2}{g_1} \left[H_1(s_1)H_2(s_2, s_3)\right]_{\text{sym}} \,.$$

From it we immediately obtain

$$H_3(s_1, s_2, s_3) = -2\frac{g_2}{g_1} H_1(s_1 + s_2 + s_3) \left[H_1(s_1)H_2(s_2, s_3)\right]_{\text{sym}} \,. \tag{14.20}$$

To gain some insight from this expression we assume again two real input tones with $|\omega_i| \ll \omega_{\mathcal{T}} \ll \omega_{3dB}$. Under these assumptions we can expand H_3 in a first order Taylor polynomial and obtain

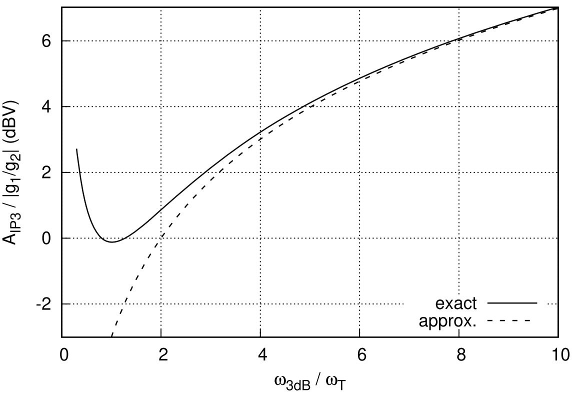


Fig. 14.10 Quadrature modulator IP3 as a function of $\omega_{3dB}/\omega_{\mathcal{T}}$ normalized to $|g_1/g_2|$. Solid line computed with the full H_3 in (14.20), dashed line IP3 computed with (14.22). $\omega_1 = \omega_2 = (1+0.1)\omega_{\mathcal{T}}$, $\omega_3 = -(1+0.2)\omega_{\mathcal{T}}$

$$\begin{aligned} H_3(\jmath\omega_1, \jmath\omega_2, \jmath\omega_3) &\approx -\jmath\frac{4}{3}\left(\frac{g_2}{g_1}\right)^2 \frac{\omega_1+\omega_2+\omega_3}{\omega_{3dB}} \\ &= -\frac{\jmath(\omega_1+\omega_2+\omega_3)C}{3\beta(V_G-V_T)^3}\,. \end{aligned} \tag{14.21}$$

As for second-order distortion we find that it's more convenient to choose a large $V_G - V_T$ than to increase the width of the transistor. Using this expression we can estimate the IP3 of the modulator as

$$A_{\mathrm{IP3}} \approx \frac{g_1}{g_2}\left|\sqrt{\frac{\omega_{3dB}}{\omega_{\mathcal{T}}}} = 2(V_G - V_T)\sqrt{\frac{\omega_{3dB}}{\omega_{\mathcal{T}}}}\right|\,. \tag{14.22}$$

The value of this approximation is compared with the value calculated from the full H_3 (14.20) as a function of $\omega_{3dB}/\omega_{\mathcal{T}}$ in Fig. 14.10 for $\omega_1 = \omega_2 = (1+0.1)\omega_{\mathcal{T}}$ and $\omega_3 = -(1+0.2)\omega_{\mathcal{T}}$. The approximation gives a reasonable value from $\omega_{3dB}/\omega_{\mathcal{T}} \gtrsim 2$. For values of $\omega_{3dB}/\omega_{\mathcal{T}} < 1$ the IP3 is seen to raise. This is however related to the fact that the wanted signals do also experience substantial attenuation compared with the case of a large ratio $\omega_{3dB}/\omega_{\mathcal{T}}$.

It is important to realize that the effective input signal x of the model includes many tones that produce many intermodulation products. In particular, the tones at $3\omega_{\mathcal{T}} - \omega_i$, $i = 1, 2$ together with the main tones at $\omega_{\mathcal{T}} + \omega_i$ produce third-order intermodulation products that falls close to the wanted signal and are difficult to suppress

$$3\omega_{\mathcal{T}} - \omega_i - 2(\omega_{\mathcal{T}} + \omega_i) = \omega_{\mathcal{T}} - 3\omega_i\,.$$

These tones are called third order *counter intermodulation products (CIM3)*. We saw in previous paragraphs that many practical imperfections introduce tones at the second harmonic of the input signals $2(\omega_{\mathcal{T}} \pm \omega_i)$. In this case second-order distortion does also produce tones around the signal of interest. In particular the combination

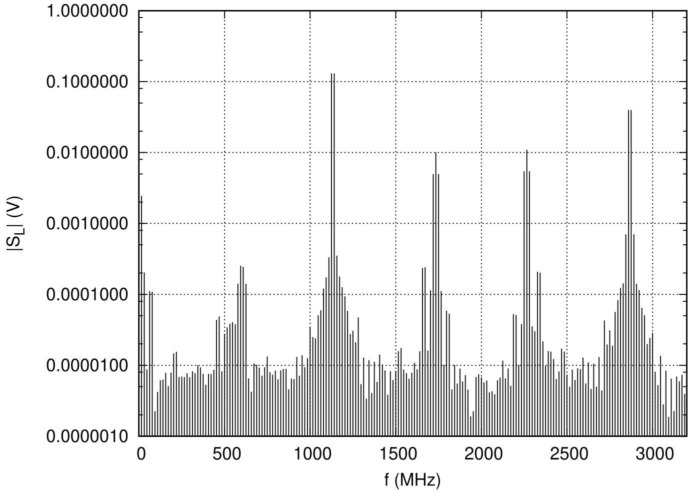


Fig. 14.11 Single-sided output spectrum of the modulator simulated with accurate transistor models

$$2(\omega_{\mathcal{T}} - \omega_i) - (\omega_{\mathcal{T}} + \omega_i) = \omega_{\mathcal{T}} - 3\omega_i$$

results in a tone at the CIM3 frequency as does the second-order distortion between $3\omega_{\mathcal{T}} - \omega_i$ and $2(\omega_{\mathcal{T}} - \omega_i)$

$$(3\omega_{\mathcal{T}} - \omega_i) - 2(\omega_{\mathcal{T}} + \omega_i) = \omega_{\mathcal{T}} - 3\omega_i\,.$$

Depending on the details of the design, these second order distortion components may contribute significantly to the overall CIM3 level of the modulator.

Figure 14.11 shows part of the output spectrum magnitude obtained by numerical simulation of the modulator with accurate transistor models, a load capacitance of 1 pF, an LO frequency of 1 GHz, and two input tones given by

$$\begin{aligned} v_I^+(t) &= -v_I^-(t) = A\cos(\omega_1 t) + A\cos(\omega_2 t) \\ v_Q^+(t) &= -v_Q^-(t) = A\sin(\omega_1 t) + A\sin(\omega_2 t) \end{aligned}$$

with

$$A = 0.15\,\text{V}, \quad \omega_1 = \frac{\omega_{\mathcal{T}}}{8}, \quad \omega_2 = \omega_1 + \frac{\omega_{\mathcal{T}}}{64}\,.$$

We used 22 nm FinFETs modelled with BSIM-CMG compact models [29] with technology parameters from [30]. The transistors were sized to have an r_{ON} of 20 Ω at $V_G - V_T = 0.5$ V. Since the threshold voltage of the transistors is 0.311 V, the LO voltage high level was chosen to be 0.811 V, while the low level was set to –0.189 V. To avoid overlapping the duration of the high pulses was reduced slightly from the nominal value of $\mathcal{T}/4$ to produce a cross point between successive LO signals ca.

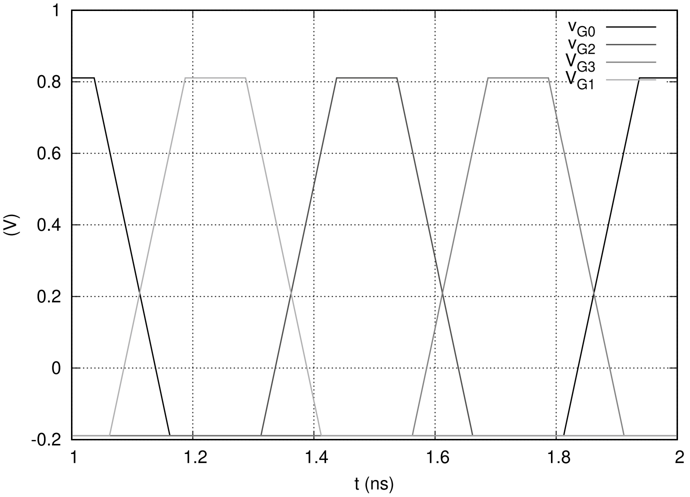


Fig. 14.12 LO signal waveforms used in the simulation of the modulator

0.1 V below V_T. The raise- and fall-times were set to 0.125 ns giving an LO transient slope K of 8 GV/s. The LO signals used in the simulation are shown in Fig. 14.12.

The spectral component levels obtained by simulation compare favorably with our analysis. The expected level of the main tones is

$$A\frac{4}{\sqrt{2}\pi} \approx 0.135\,\text{V}$$

and is very close to simulated one of 0.131 V. Our choice of parameters is such that $\omega_{3dB}/\omega_{\mathcal{T}} \approx 7.9$. We can therefore estimate the expected second- and third order intermodulation products from the approximate transfer functions given by (14.19) and (14.22) respectively. The expected level of the second order tone at $2(\omega_{\mathcal{T}} - \omega_1)$ is estimated to be

$$\left|\left(A\frac{4}{\sqrt{2}\pi}\right)^2 \frac{1}{2} H_2(\jmath(\omega_{\mathcal{T}} - \omega_1), \jmath(\omega_{\mathcal{T}} - \omega_1))\right| \approx 2.29\,\text{mV}$$

The expected IM3 level at $\omega_{\mathcal{T}} + 2\omega_1 - \omega_2$ is

$$\left|\left(A\frac{4}{\sqrt{2}\pi}\right)^3 \frac{3}{4} H_3(\jmath(\omega_{\mathcal{T}} + \omega_1), \jmath(\omega_{\mathcal{T}} + \omega_1), -\jmath(\omega_{\mathcal{T}} + \omega_2))\right| \approx 0.23\,\text{mV}\,.$$

The simulated values are 5.44 and 0.35 mV respectively, reasonably close to the predicted values.

In our simplified analysis we assumed zero LO signal raise- and fall-times. It is interesting to investigate how fast the LO transients have to be before the intermodula-

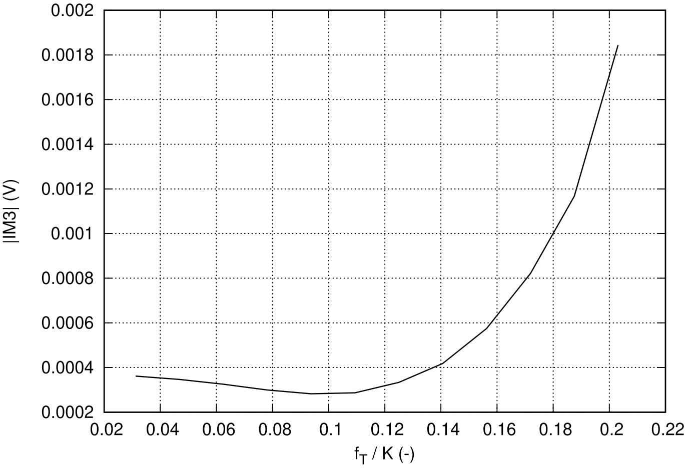


Fig. 14.13 Simulated IM3 versus LO transient slope K

tion products start to deviate significantly from the predicted values. We investigated this question by simulation. The IM3 level is plotted as a function of the LO transients slope K in Fig. 14.13. For all values of K the crossing point was kept constant. Note that the LO waveforms corresponding to the lowest K values are essentially triangular with no flat high level region. This simulation suggest that the analysis gives reasonable IM3 estimates for values of $f_{\mathcal{T}}/K \lessapprox 0.14$.

14.3 Sampling Mixer

A communication receiver should ideally be able to detect a single signal on a channel of a frequency band allocated to the service of interest and completely suppress all other signals. Due to limitations in the selectivity of filters and the difficulty of implementing tuneable filters, this can only be achieved approximately. Virtually all receivers are composed by a fixed highly selective filter (the preselection filter) typically implemented with surface- (SAW) or bulk-acoustic wave (BAW) technologies suppressing all signals outside the band of interest. This filter is then followed by some signal amplification and by a shift of the signal of interest to a lower fixed frequency with the help of a mixer. At this lower frequency another fixed filter (the channel filter) with a bandwidth corresponding to the bandwidth of a channel separates the wanted signal from other signals on adjacent channels. The channel of interest is selected by shifting in frequency of the input spectrum in such a way that the desired signal falls in the passband of the channel filter. This is done by appropriately choosing the frequency of the so-called local oscillator (LO) driving the LO port of the mixer.

The sampling mixer analysed hereafter is an attempt to remove the preselection filter and implement a tuneable filter capable of selecting a single channel only using components available in a standard CMOS technology. This is driven by the desire for miniaturisation and cost reduction. While this type of circuits have their own drawbacks they are a nice example showing some capabilities of time-varying systems that can't be matched by LTI ones.

14.3.1 Time-Varying Impulse Response

Consider the highly idealised sampling mixer shown in Fig. 14.14. The input signal is represented by the voltage source V_s and the nodes labelled $V_0, \dots, V_{N-1}$ represent output signals. The ideal switches S_n, $n = 0, \dots, N-1$ are driven by the $\mathcal{T}$-periodic clock signals ϕ_n. A switch is closed when the corresponding clock signal is high and open when low. We assume that the clock signals are non-overlapping. Since no reactive component is present on the source side of the switches the output signals can be analysed independently of each other. In the following we assume that each clock phase has the same duration so that

$$t_n = n\frac{\mathcal{T}}{N}; \qquad n = 0, \dots, N-1.$$

In this case it's enough to compute the time-varying impulse response of one output signal only. The other ones are then obtained by a simple translations in time. We will therefore compute the time-varying impulse response corresponding to the output V_0 that in the following we will denote by y. Similarly, we will denote the source signal by x.

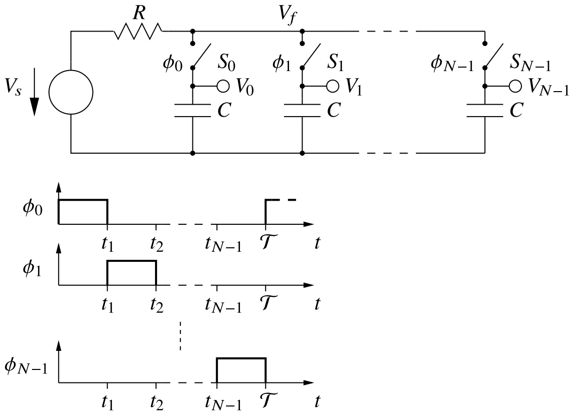


Fig. 14.14 Linear periodic switched capacitor circuit that can be used as a sampling mixer or as an N-path filter

The circuit, having two phases, is described by two differential equations. Between $0 < t < t_1$, when the switch is closed, it is described by

$$Dy + \omega_0 y = \omega_0 x, \qquad \omega_0 := \frac{1}{RC}$$

while for $t_1 < t < \mathcal{T}$, when the switch is open, by

$$Dy = 0\,.$$

We start by computing the fundamental kernel of the system $W(t, \tau)$ which is the solution of the differential equation when driven by a Dirac impulse occurring at time τ. Since the circuit varies periodically in time, it's enough to compute it for $0 < \tau < \mathcal{T}$.

For $0 < \tau < t_1$ the output is zero up to time τ at which point it will jump to $1/RC$ and start to decay exponentially as in an LTI system. At time t_1 the switch is opened, leaving the output capacitor floating. The output voltage will therefore remain constant up to time $\mathcal{T}$. At time $\mathcal{T}$, since there is a resistor between the capacitor and the source, the output will simply start to decrease exponentially again with the same time constant $1/RC$ as during the first part of the response. Continuing this process we obtain (see Fig. 14.15)

$$\underset{0<\tau<t_1}{W(t,\tau)} = \begin{cases} 0 & t < \tau \\ \omega_0 \mathrm{e}^{-\omega_0(t-\tau)} & \tau < t < t_1 \\ \omega_0 B A^{k-1} \mathrm{e}^{-\omega_0(t-k\mathcal{T})} & k\mathcal{T} < t < k\mathcal{T} + t_1\,, k \geq 1 \\ B A^k & t_1 + k\mathcal{T} < t < (k+1)\mathcal{T}\,, k \geq 0 \end{cases}$$

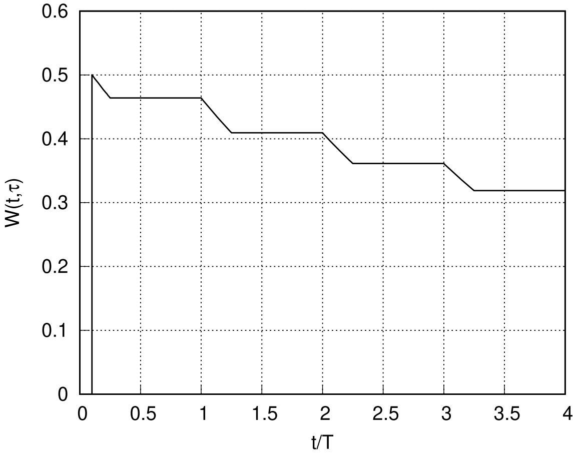


Fig. 14.15 Fundamental kernel of the sampling mixer for $\tau = 0.1\mathcal{T}$, $N = 4$, $\omega_0\mathcal{T} = 0.5$

with

$$A := \mathrm{e}^{-\omega_0 t_1} \qquad B := \mathrm{e}^{-\omega_0 (t_1-\tau)}\,.$$

For $t_1 < \tau < \mathcal{T}$, given that the output is disconnected from the input, the output remains zero

$$\underset{t_1<\tau<\mathcal{T}}{W(t,\tau)} = 0\,.$$

The time varying impulse response can be derived from the fundamental kernel with the help of the variable substitution $\xi = t - \tau$ and by keeping in mind that the value of the impulse response $h(t,\xi)$ is the value of the output at time t assuming that a Dirac impulse was applied ξ seconds in the past. As the impulse response is periodic, it's enough to compute its value over the first period. For $0 < t < t_1$ it is given by (see Fig. 14.16a)

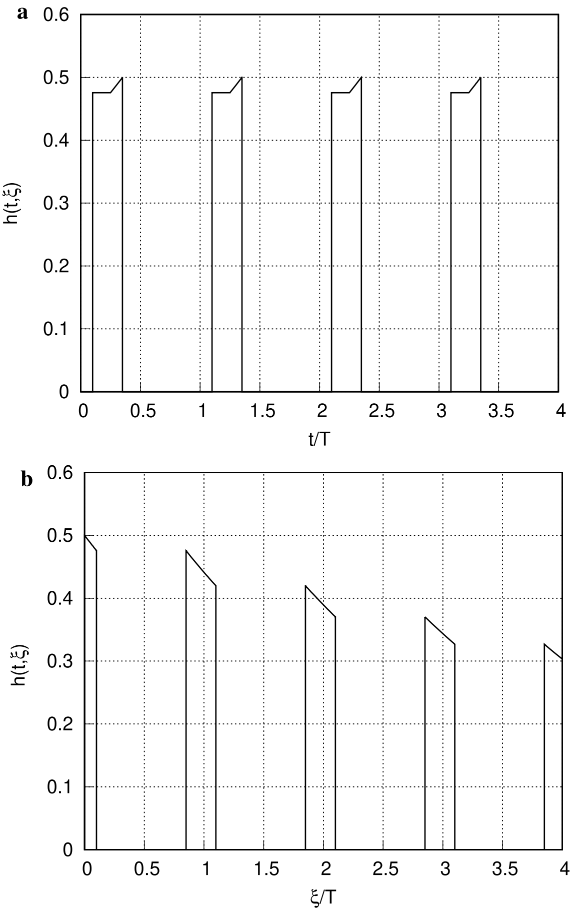


Fig. 14.16 **a** Time-varying impulse response of the sampling mixer as a function of time for $\xi = 0.1T$, $N = 4$, $\omega_0 = 0.5\mathcal{T}$ **b** Time-varying impulse response of the sampling mixer as a function of ξ for $t = 0.1T$, $N = 4$, $\omega_0 = 0.5\mathcal{T}$

$$\underset{0<t<t_1}{h(t,\xi)} = \begin{cases} \omega_0 \mathrm{e}^{-\omega_0 \xi} & 0 < \xi < t \\ \omega_0 A^k \mathrm{e}^{-\omega_0(\xi - k\mathcal{T})} & t - t_1 + k\mathcal{T} < \xi < t + k\mathcal{T}, k \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

and for $t_1 < t < \mathcal{T}$ by

$$\underset{t_1<t<\mathcal{T}}{h(t,\xi)} = \begin{cases} \omega_0 A^k \mathrm{e}^{-\omega_0[\xi - (k\mathcal{T}+t-t_1)]} & t - t_1 + k\mathcal{T} < \xi < t + k\mathcal{T}, k \geq 1 \\ 0 & \text{otherwise}\,. \end{cases}$$

14.3.2 Time-Varying Transfer Function

While the circuit is fully characterised by the above time-varying impulse response, its filtering characteristics are best understood by analysing its time-varying frequency response. This will allow us to easily obtain the output signal when the circuit is driven by a tone.

We compute the time-varying transfer function $\hat{h}(t,\omega)$ by Fourier transforming $h(t,\xi)$. For $0 < t < t_1$ we have

$$\hat{h}(t,\omega) = \int_0^t \omega_0 \mathrm{e}^{-\omega_0 \xi} \mathrm{e}^{-\jmath\omega\xi} \mathrm{d}\xi + \sum_{k=1}^{\infty} \int_{t-t_1+k\mathcal{T}}^{t+k\mathcal{T}} \omega_0 A^k \mathrm{e}^{-\omega_0(\xi-k\mathcal{T})} \mathrm{e}^{-\jmath\omega\xi} \mathrm{d}\xi\,.$$

The terms in the right summation are powers of the summation variable k multiplied by a constant and reminds of a geometric series with a missing first term. As a first step we therefore add the missing term by adjusting the limits of the first integral

$$\hat{h}(t,\omega) = -\int_{t-t_1}^{0} \omega_0 \mathrm{e}^{-\omega_0\xi - \jmath\omega\xi} \mathrm{d}\xi + \sum_{k=0}^{\infty} \omega_0 A^k \mathrm{e}^{\omega_0 k\mathcal{T}} \int_{t-t_1+k\mathcal{T}}^{t+k\mathcal{T}} \mathrm{e}^{-\omega_0\xi - \jmath\omega\xi} \mathrm{d}\xi\,.$$

Evaluating the integrals and simplifying we find

$$\frac{1 - \mathrm{e}^{-(\omega_0+\jmath\omega)(t-t_1)}}{1 + \jmath\frac{\omega}{\omega_0}} + \mathrm{e}^{-(\omega_0+\jmath\omega)t} \frac{\mathrm{e}^{(\omega_0+\jmath\omega)t_1} - 1}{1 + \jmath\frac{\omega}{\omega_0}} \sum_{k=0}^{\infty} A^k \mathrm{e}^{-\jmath\omega k\mathcal{T}}\,.$$

Performing the summation of the geometric series we finally obtain

$$\underset{0<t<t_1}{\hat{h}(t,\omega)} = \frac{1 - \mathrm{e}^{-(\omega_0+\jmath\omega)(t-t_1)}}{1 + \jmath\frac{\omega}{\omega_0}} + \frac{\left(\mathrm{e}^{(\omega_0+\jmath\omega)t_1} - 1\right)\mathrm{e}^{-(\omega_0+\jmath\omega)t}}{\left(1 + \jmath\frac{\omega}{\omega_0}\right)\left(1 - \mathrm{e}^{-\jmath\omega\mathcal{T}}\mathrm{e}^{-\omega_0 t_1}\right)}\,. \tag{14.23}$$

For $t_1 < t < \mathcal{T}$ the time-varying frequency response is given by

$$\hat{h}(t,\omega) = \sum_{k=0}^{\infty} \int_{t-t_1+k\mathcal{T}}^{t+k\mathcal{T}} \omega_0 A^k \mathrm{e}^{-\omega_0[\xi-(k\mathcal{T}+t-t_1)]} \mathrm{e}^{-\jmath\omega\xi} \mathrm{d}\xi \,.$$

This is again a geometric series and proceeding as above we obtain

$$\underset{t_1<t<\mathcal{T}}{\hat{h}(t,\omega)} = \frac{\left(\mathrm{e}^{\jmath\omega t_1} - \mathrm{e}^{-\omega_0 t_1}\right)\mathrm{e}^{-\jmath\omega t}}{\left(1+\jmath\frac{\omega}{\omega_0}\right)\left(1-\mathrm{e}^{-\jmath\omega\mathcal{T}}\mathrm{e}^{-\omega_0 t_1}\right)} \,. \tag{14.24}$$

14.3.3 Selectivity

With $\hat{h}(t,\omega)$ the output of the circuit when driven by $x(t) = \cos(\omega t)$ is immediately obtained

$$y(t) = \Re\{\hat{h}(t,\omega)\mathrm{e}^{\jmath\omega t}\} \,.$$

The output is shown in Fig. 14.17 for two values of the input frequency and $N = 4$. During the time intervals $t_1 + k\mathcal{T} < t < (k+1)\mathcal{T}$, $k \in \mathbb{Z}$ the output is constant and assumes the value

$$y(t) = \Re\{\hat{h}(t,\omega)\mathrm{e}^{\jmath\omega t}\} = \Re\left\{\frac{\left(\mathrm{e}^{\jmath\omega t_1} - \mathrm{e}^{-\omega_0 t_1}\right)\mathrm{e}^{\jmath\omega k\mathcal{T}}}{\left(1+\jmath\frac{\omega}{\omega_0}\right)\left(1-\mathrm{e}^{-\jmath\omega\mathcal{T}}\mathrm{e}^{-\omega_0 t_1}\right)}\right\} .$$

where we used the periodicity in time of $\hat{h}(t,\omega)$ and the previously computed expression valid for $t_1 < t < \mathcal{T}$

$$\hat{h}(t,\omega) = \underset{t_1<t-k\mathcal{T}<\mathcal{T}}{\hat{h}(t-k\mathcal{T},\omega)} \,.$$

These values are the output sample values of the sampling mixer. Let's denote them by $y[k]$ and set $\omega = n\omega_s + \Delta\omega$, $n \in \mathbb{Z}$ with $\omega_s = 2\pi/\mathcal{T}$ and $\Delta\omega < \omega_s/2$. Then the above expression becomes

$$\begin{aligned} y[k] &= \Re\left\{h_{\text{eff}}(n\omega_s + \Delta\omega)\mathrm{e}^{\jmath\Delta\omega k\mathcal{T}}\right\} \\ &:= \Re\left\{\frac{\left(\mathrm{e}^{\jmath(n\omega_s+\Delta\omega)t_1} - \mathrm{e}^{-\omega_0 t_1}\right)\mathrm{e}^{\jmath\Delta\omega k\mathcal{T}}}{\left(1+\jmath\frac{n\omega_s+\Delta\omega}{\omega_0}\right)\left(1-\mathrm{e}^{-\jmath\Delta\omega\mathcal{T}}\mathrm{e}^{-\omega_0 t_1}\right)}\right\} . \end{aligned} \tag{14.25}$$

These are the samples of a sinusoidal with angular frequency $\Delta\omega$ and amplitude

$$\left|\frac{\left(\mathrm{e}^{\jmath(n\omega_s+\Delta\omega)t_1} - \mathrm{e}^{-\omega_0 t_1}\right)}{\left(1+\jmath\frac{n\omega_s+\Delta\omega}{\omega_0}\right)\left(1-\mathrm{e}^{-\jmath\Delta\omega\mathcal{T}}\mathrm{e}^{-\omega_0 t_1}\right)}\right| .$$

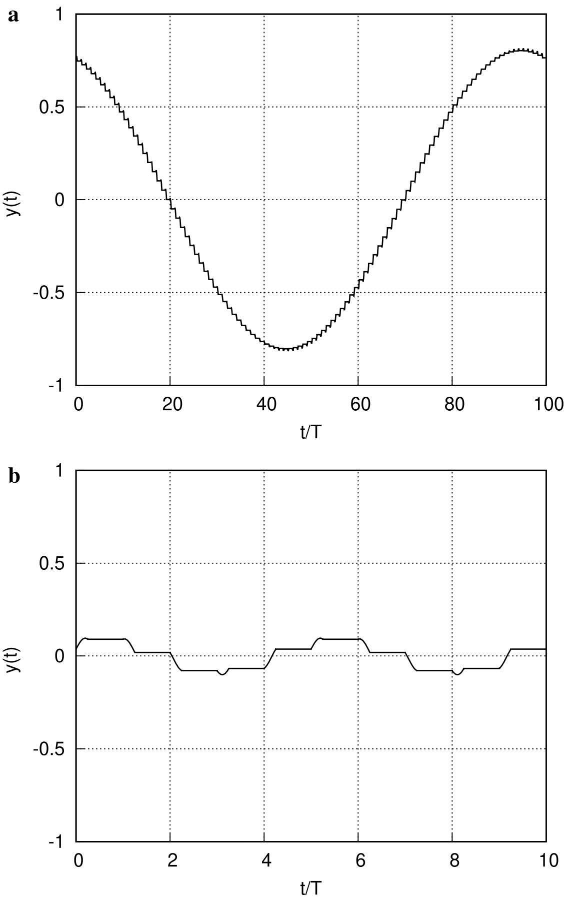


Fig. 14.17 Sampling mixer output V_0 when driven by $\cos(\omega t)$ with $N = 4$, $\omega_0 = 0.5\mathcal{T}$. **a** $\omega = 1.01\omega_s$. **b** $\omega = 1.2\omega_s$

The fact that the output samples correspond to samples of a sinusoidal with a frequency independent of n is a manifestation of *aliasing* inherent in every sampling process. The interesting aspect of the sampling mixer is the fact that only samples of tones with a frequency very close to $n\omega_s$ have a significant amplitude while the ones of signals with frequencies at a distance larger than approximately ω_0/N from $n\omega_s$ are attenuated. This effect is due to the factor

$$1 - \mathrm{e}^{-\jmath\Delta\omega\mathcal{T}}\mathrm{e}^{-\omega_0 t_1} = 1 - \mathrm{e}^{-\jmath 2\pi\frac{\Delta\omega}{\omega_s}}\mathrm{e}^{-\frac{2\pi}{N}\frac{\omega_0}{\omega_s}}$$

in the denominator of the above expression. For $\omega_0 \ll N\omega_s$ the last exponential on the right is only slightly smaller than 1. Therefore, for $\Delta\omega < \omega_0/N$ this factor becomes small thereby boosting the value of the samples around those frequencies (see Fig. 14.18). From this we conclude that the sampling mixer not only behaves as a sample and hold, but it also acts as a highly selective (large quality factor) filter

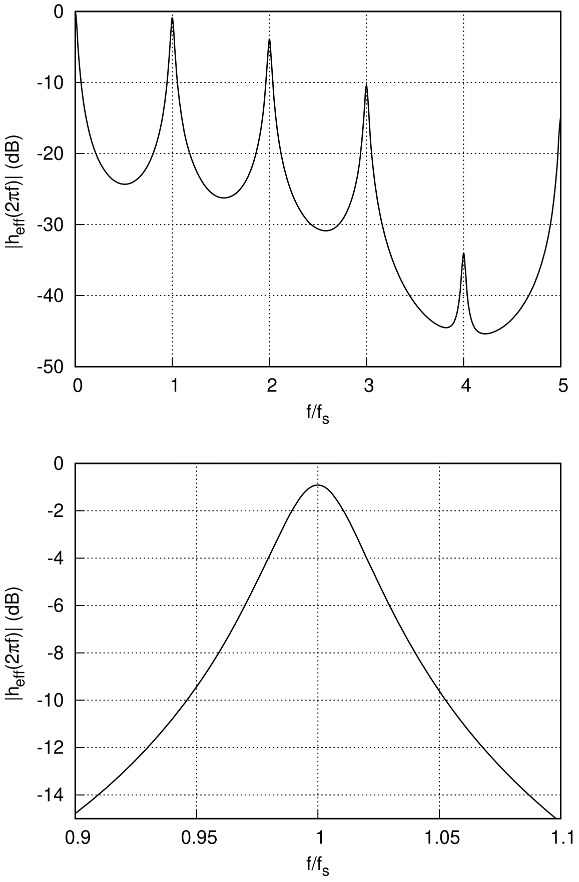


Fig. 14.18 Equivalent magnitudes of the output samples of a sampling mixer with $N = 4$, $\omega_0 = 0.5\mathcal{T}$

around the sampling frequency and its harmonics. The achievable selectivity is much higher than the one of LTI RLC filters integrable in a standard CMOS technology.

14.3.4 Even Harmonic Response Suppression

The response around the harmonics is generally undesired and can be suppressed by using weighted sums of the N outputs as discussed in Example 12.10. In the following we assume N even and investigate the possibility to suppress the responses at even harmonics by making use of the sample values on the capacitors as opposed to the full waveforms. Consider the sample value on capacitor $i = N/2$

$$V_{N/2}(t) = \Re\{\hat{h}(t - \mathcal{T}/2, \omega)\mathrm{e}^{\jmath\omega t}\}$$

Denoting the sample value held by the capacitor during the interval $t_1 + t_1 N/2 + k\mathcal{T} < t < \mathcal{T} + t_1 N/2 + k\mathcal{T}$ by $y_{N/2}[k]$ and using the periodicity in time of $\hat{h}(t,\omega)$ as before we obtain

$$\begin{aligned} y_{N/2}[k] &= \Re\left\{ \frac{\left(e^{\jmath\omega t_1} - e^{-\omega_0 t_1}\right) e^{-\jmath\omega(t-k\mathcal{T}-\mathcal{T}/2)}}{\left(1+\jmath\frac{\omega}{\omega_0}\right)\left(1-e^{-\jmath\omega\mathcal{T}}e^{-\omega_0 t_1}\right)} e^{\jmath\omega t} \right\} \\ &= \Re\left\{ \frac{\left(e^{\jmath\omega t_1} - e^{-\omega_0 t_1}\right)}{\left(1+\jmath\frac{\omega}{\omega_0}\right)\left(1-e^{-\jmath\omega\mathcal{T}}e^{-\omega_0 t_1}\right)} e^{\jmath\omega(k\mathcal{T}+\mathcal{T}/2)} \right\} \\ &= \Re\left\{ h_{\mathrm{eff}}(\omega) e^{\jmath\omega k\mathcal{T}} e^{\jmath\omega\mathcal{T}/2} \right\} \end{aligned}$$

If the frequency of the signal is close to the nth harmonic of the sampling frequency $\omega = n\omega_s + \Delta\omega$, $\Delta\omega \ll \omega_s$ the last expression becomes

$$y_{N/2}[k] = \Re\left\{ h_{\mathrm{eff}}(n\omega_s + \Delta\omega) e^{\jmath\Delta\omega k\mathcal{T}} (-1)^n e^{\jmath\pi\frac{\Delta\omega}{\omega_s}} \right\}.$$

The difference between sample values on capacitor C_0 and $C_{N/2}$ is thus given by

$$y[k] - y_{N/2}[k] = \Re\left\{ h_{\mathrm{eff}}(n\omega_s + \Delta\omega) e^{\jmath\Delta\omega k\mathcal{T}} \left[1 - (-1)^n e^{\jmath\pi\frac{\Delta\omega}{\omega_s}}\right] \right\}.$$

Under the assumption $\Delta\omega \ll \omega_s$ the right most exponential is close to 1 so that

$$y[k] - y_{N/2}[k] \approx \begin{cases} 0 & n \text{ even} \\ 2y[k] & n \text{ odd}. \end{cases}$$

For $N = 4$ the four output samples can be combined in pairs forming signals corresponding to the in-phase ($V_I[k] = V_0[k] - V_2[k]$) and quadrature ($V_Q[k] = V_1[k] - V_3[k]$) output of a quadrature mixer.

Before concluding this section we note that a loss of charge from the capacitor during the hold phase results in a lower boost of the samples around the frequencies $n\omega_s$, $n \in \mathbb{Z}$. A loss of charge could be caused for example by a finite load resistance or by a switched-capacitor circuit following the sampling mixer. The reduction in the magnitude of the samples comes from the fact that the value of A appearing in the definition of $h(t,\xi)$ will become smaller and as a consequence the boosting factor

$$1 - e^{-\jmath\Delta\omega\mathcal{T}} A$$

in the denominator of $\hat{h}(t,\omega)$ will not become as small as calculated above.

14.4 N-Path Filters

The block diagram of a general N-*path filter* is shown in Fig. 14.19 and can be thought of as the cascade of an N-path receiver and an N-path transmitter (compare with Example 12.10). Among other things they permit to implement transfer functions that, under suitable assumptions, mimic the ones of LTI networks that are difficult or not manufacturable with RLC elements due to the limited range of actually implementable values or due to limitations in their quality (quality factor). The time-varying transfer function of a general N-path filter can be analysed using the same methods used to analyse the N-path receiver of Example 12.10. Here instead of the general case we analyse a concrete implementation that shows some useful applications.

In the following we analyse the simple case in which the N LTI subsystems are simple shunt capacitors and where the periodic input and output functions are equal switching functions. Under these conditions, when the switches of path k are closed, the upper plate of the kth capacitor is simultaneously connected to the input as to the output and we obtain the circuit shown in Fig. 14.14 where now the output is constituted by the node labeled V_f.

14.4.1 Time-Varying Frequency Response

During clock phase $k, 0 \leq k \leq N-1$, during which the switch S_k is closed, the voltage V_f is equal to the voltage V_i. For this reason we can express V_f in terms of the time-varying frequency response that we obtained in the previous section. In particular, when the input is a complex tone $\mathrm{e}^{\jmath\omega t}$ the output is given by

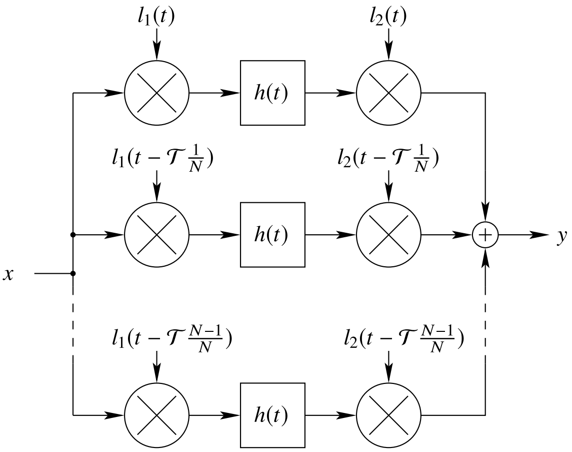

$l_1(t)$
$l_2(t)$
$h(t)$
$l_1(t-\mathcal{T}\frac{1}{N})$
$l_2(t-\mathcal{T}\frac{1}{N})$
$h(t)$
y
x
$l_1(t-\mathcal{T}\frac{N-1}{N})$
$l_2(t-\mathcal{T}\frac{N-1}{N})$
$h(t)$

Fig. 14.19 Block diagram of a generic N-path filter

$$V_f(t) = \mathrm{e}^{\jmath\omega t}\sum_{k=0}^{N-1}\hat{h}_{sm}(t-kt_1,\omega)1_{t_1}(t-kt_1 \mod \mathcal{T})$$

with

$$1_{t_1}(t) = \begin{cases}1 & 0 < t < t_1\\ 0 & \text{otherwise}\end{cases}$$

and where we denoted the time-varying frequency response of the sampling mixer by $\hat{h}_{sm}$ to avoid confusion with the one of the whole N-path filter that we'll denote by $\hat{h}$. For $0 < t < \mathcal{T}$, using (14.23) the frequency response of the filter is thus given by

$$\begin{aligned}\hat{h}(t,\omega) =& \frac{1}{1+\jmath\frac{\omega}{\omega_0}} + \Bigg[\frac{-\mathrm{e}^{(\omega_0+\jmath\omega)t_1}}{1+\jmath\frac{\omega}{\omega_0}}\\ &+ \frac{\left(\mathrm{e}^{(\omega_0+\jmath\omega)t_1}-1\right)}{\left(1+\jmath\frac{\omega}{\omega_0}\right)\left(1-\mathrm{e}^{-\jmath\omega\mathcal{T}}\mathrm{e}^{-\omega_0 t_1}\right)}\Bigg]\sum_{k=0}^{N-1}\mathrm{e}^{-(\omega_0+\jmath\omega)(t-kt_1)}1_{t_1}(t-kt_1)\,.\end{aligned}$$

As the time-varying frequency response is $\mathcal{T}$-periodic, we can expand it in a Fourier series. The nth Fourier coefficient of the summation on the right is

$$\begin{aligned}a_n &= \frac{1}{\mathcal{T}}\int_0^{\mathcal{T}}\sum_{k=0}^{N-1}\mathrm{e}^{-(\omega_0+\jmath\omega)(t-kt_1)}1_{t_1}(t-kt_1)\mathrm{e}^{-\jmath n\omega_s t}\mathrm{d}t\\ &= \frac{1}{\mathcal{T}}\sum_{k=0}^{N-1}\mathrm{e}^{(\omega_0+\jmath\omega)kt_1}\int_{k\mathcal{T}/N}^{(k+1)\mathcal{T}/N}\mathrm{e}^{-[\omega_0+\jmath(\omega+n\omega_s)]t}\mathrm{d}t\\ &= \frac{1}{\mathcal{T}}\frac{1-\mathrm{e}^{-[\omega_0+\jmath(\omega+n\omega_s)]t_1}}{\omega_0+\jmath(\omega+n\omega_s)}\sum_{k=0}^{N-1}\mathrm{e}^{-\jmath n\omega_s kt_1}\,.\end{aligned}$$

The last summation is zero unless n is a multiple of N in which case it evaluates to N

$$a_n = \begin{cases}\frac{N}{\mathcal{T}}\frac{1-\mathrm{e}^{-[\omega_0+\jmath(\omega+n\omega_s)]t_1}}{\omega_0+\jmath(\omega+n\omega_s)} & n = Nm, m\in\mathbb{Z}\\ 0 & \text{otherwise.}\end{cases}$$

The time-varying frequency response of the filter is therefore given by

$$\hat{h}(t,\omega) = \sum_{n=\ldots,-4,0,4,\ldots} h_n(\omega)\mathrm{e}^{\jmath n\omega_s t} \tag{14.26}$$

with

$$h_n(\omega) = \frac{1}{1+\jmath\frac{\omega}{\omega_0}} + \left[\frac{-\mathrm{e}^{(\omega_0+\jmath\omega)t_1}}{1+\jmath\frac{\omega}{\omega_0}} + \frac{\left(\mathrm{e}^{(\omega_0+\jmath\omega)t_1}-1\right)}{\left(1+\jmath\frac{\omega}{\omega_0}\right)\left(1-\mathrm{e}^{-\jmath\omega\mathcal{T}}\mathrm{e}^{-\omega_0 t_1}\right)}\right]a_n$$

or, after some simplification

$$h_n(\omega) = \frac{1}{1+\jmath\frac{\omega}{\omega_0}}\left[1+\frac{1}{t_1\omega_0}\frac{\left(\mathrm{e}^{-\jmath\omega(\mathcal{T}-t_1)}-1\right)\left(1-\mathrm{e}^{-[\omega_0+\jmath(\omega+n\omega_s)]t_1}\right)}{\left(1-\mathrm{e}^{-\jmath\omega\mathcal{T}}\mathrm{e}^{-\omega_0 t_1}\right)\left(1+\jmath\frac{\omega+n\omega_s}{\omega_0}\right)}\right]. \qquad (14.27)$$

The last term of $h_n(\omega)$ includes the same factor that we discussed in the analysis of the sampling mixer and responsible, under the condition $\omega_0 t_1 \ll 1$, for boosting the response of the circuit at frequencies $\omega = k\omega_s + \Delta\omega$, $k \in \mathbb{Z}$, $\Delta\omega < \omega_0/N$. Therefore, for $\omega_0 t_1 \ll 1$ the transfer function $h_0(\omega)$ represents a highly selective band-pass filter with pass bands centered at $k\omega_s$ with $k \neq Nm$, $m \in \mathbb{Z}$ (see Fig. 14.20).

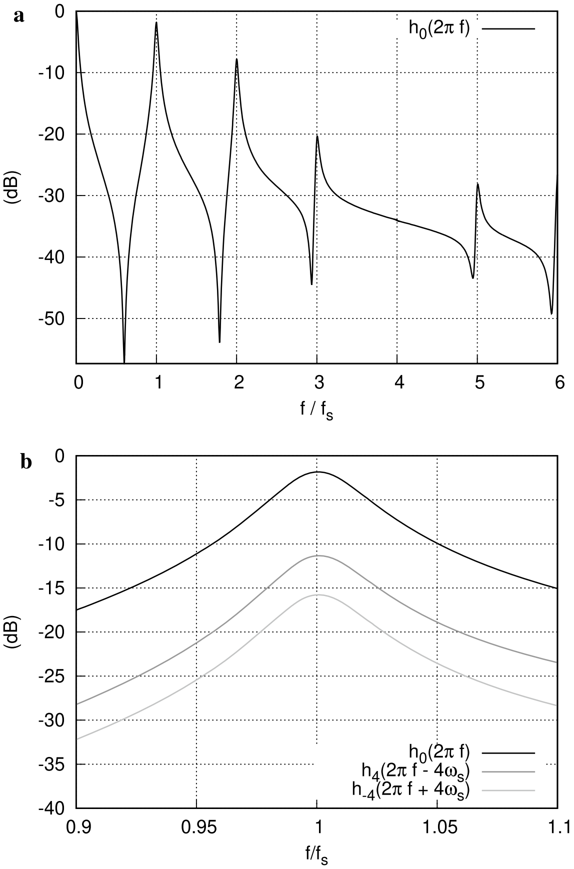


Fig. 14.20 N-path filter transfer functions magnitude for $N = 4$, $\omega_0 = 0.5\mathcal{T}$. **a** $n = 0$. **b** Detail around $f/f_s \approx 1$ for $n = -4, 0, 4$

The transfer functions $h_n(\omega)$, $n \neq 0$ are also highly selective, but they also introduce a shift in frequency. In particular an input tone at $\omega - n\omega_s$ passing through $h_n(\omega)$ results in an output tone at ω which will overlap with the response of $h_0(\omega)$ to an input tone at ω. Therefore, if this N-path filter is used in a receiver to suppress interfering signals, while it possess undesired pass-bands, the closest frequency of an interfering signal that at the output of the circuit will overlap in frequency with the wanted signal is $N\omega_s$. The magnitude of a few transfer functions producing an output tone at ω are shown in Fig. 14.20.

14.4.2 Selectivity

By applying some approximations to h_0 valid in the vicinity of ω_s we can obtain a transfer function that can be implemented with fixed RLC components. This will allow us to quantify the selectivity of the filter in terms of standard metrics.

As a first step we set again $\omega = \omega_s + \Delta\omega$ and use $\Delta\omega \ll \omega_s$ to make the following approximation

$$\begin{aligned}\frac{1}{1+\jmath\frac{\omega_s+\Delta\omega}{\omega_0}}\frac{1}{t_1\omega_0}\left(\left(\mathrm{e}^{-\jmath\omega(\mathcal{T}-t_1)}-1\right)\right) &\approx \frac{1}{\jmath\omega_s t_1}\mathrm{e}^{\jmath\omega_s t_1/2}\left(\mathrm{e}^{\jmath\omega_s t_1/2}-\mathrm{e}^{-\jmath\omega_s t_1/2}\right)\\ &= \mathrm{e}^{\jmath\pi/N}\frac{\sin(\pi/N)}{\pi/N}\,.\end{aligned}$$

Similarly, using in addition $\omega_0 \ll \omega_s$

$$\begin{aligned}\frac{\left(1-\mathrm{e}^{-[\omega_0+\jmath(\omega_s+\Delta\omega)]t_1}\right)}{\left(1-\mathrm{e}^{-\jmath(\omega_s+\Delta\omega)\mathcal{T}}\mathrm{e}^{-\omega_0 t_1}\right)\left(1+\jmath\frac{\omega_s+\Delta\omega}{\omega_0}\right)} &\approx \mathrm{e}^{-\jmath\pi/N}\frac{2\jmath\sin(\pi/N)}{(\jmath\Delta\omega\mathcal{T}+\omega_0 t_1)\jmath\frac{\omega_s}{\omega_0}}\\ &\approx \mathrm{e}^{-\jmath\pi/N}\frac{\sin(\pi/N)}{(1+\jmath N\frac{\Delta\omega}{\omega_0})\pi/N}\,.\end{aligned}$$

Finally, using these approximations and noting that the first summand in $h_0(\omega)$ is small compared to the second we obtain

$$h_0(\omega) \approx \left(\frac{\sin(\pi/N)}{\pi/N}\right)^2\frac{1}{1+\jmath N\frac{\Delta\omega}{\omega_0}}. \tag{14.28}$$

The impedance of a parallel LTI RLC resonator with a resonance frequency of ω_s is given by

$$Z_r(\omega) = \frac{\frac{\jmath\omega}{\omega_s}\frac{R_r}{q}}{\left(\frac{\jmath\omega}{\omega_s}\right)^2+\frac{\jmath\omega}{\omega_s q}+1}$$

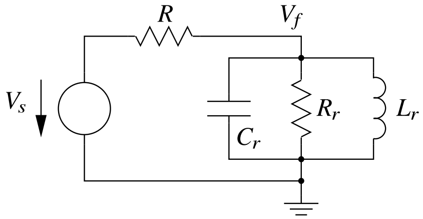


Fig. 14.21 Model for the N-path filter $h_0(\omega)$ valid around ω_s

with q the quality factor of the resonator and R_r the impedance at resonance. Around the resonance frequency it can be approximated by

$$Z_r(\omega_s + \Delta\omega) \approx \frac{R_r}{1 + \jmath 2q\frac{\Delta\omega}{\omega_s}}\,.$$

The transfer function to V_f around the resonance frequency of the circuit shown in Fig. 14.21 is therefore given by

$$\frac{R_r}{R + R_r} \cdot \frac{1}{1 + \jmath 2q\frac{R}{R+R_r}\frac{\Delta\omega}{\omega_s}}\,, \qquad \omega_s^2 = \frac{1}{\sqrt{L_r C_r}}\,, \qquad q = \frac{R_r}{\omega_s L_r}$$

and has the same form as the approximation of $h_0(\omega)$ given in (14.28). The two are equal if

$$\begin{cases} \dfrac{R_r}{R + R_r} = K \\ 2q\dfrac{R}{R + R_r}\dfrac{\Delta\omega}{\omega_s} = \dfrac{N}{\omega_0} \end{cases}$$

with

$$K := \left(\frac{\sin(\pi/N)}{\pi/N}\right)^2.$$

This shows that around ω_s $h_0(\omega)$ can be modelled as a parallel resonator with resonance frequency ω_s and characterised by

$$R_r = R\frac{K}{1 - K}\,, \qquad q = \frac{N\pi}{(1 - K)\omega_0\mathcal{T}}\,.$$

The transfer function of this model is compared with the exact $h_0(\omega)$ in Fig. 14.22. For $N = 4$, $\omega_0\mathcal{T} = 0.5$ the quality factor has a value close to 133. For comparison, the highest resonance quality factor implementable at RF frequencies with inductors and capacitors available in standard CMOS technologies is in the range of 20, with typical values substantially lower than this.

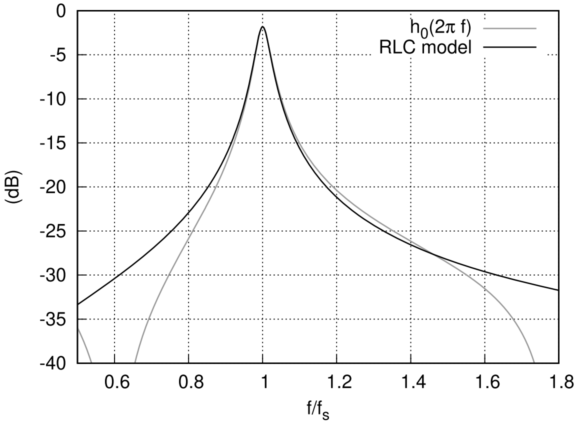


Fig. 14.22 Comparison between the N-path filter transfer function using the RLC model valid around the sampling frequency and the transfer function h_0 for $N = 4$ and $\omega_0 = 0.5\mathcal{T}$

Appendix A
Signal-Flow Graphs

Signal-flow graphs (SFGs) are a graphical way to represent systems. While this is true of block diagrams, the conventions used with the former make them simpler and, for many purposes, more powerful than the latter. SFGs are sometimes used with nonlinear systems, but are most useful to manipulate and transform linear systems. In this appendix we first review some of the most useful transformations applicable to SFGs used to describe *linear* systems, so called linear signal-flow graphs. We then review their more limited use with general nonlinear systems. Finally, we show how all the rules valid for linear SFGs can be extended to weakly-nonlinear time-invariant systems.

A.1 Linear Signal-Flow Graphs

A.1.1 Construction Rules

Signal-flow graphs are directed graphs. Each *node* represents a variable, in our case a signal. Signals flow along the *branches* (or *edges*) of the graph in the indicated direction. Each branch is labelled by the *transmission factor* of the branch. A signal flowing along a branch is composed with the branch transmission factor: in the time domain composition is effected by using the convolution product, while in the Laplace or frequency domain by using standard multiplication. The transmission factors in the time domain representation of a system are impulse responses while in the Laplace representation are transfer functions.

All signals entering a node through branches are *summed* at that node. In other words, the value of the variable represented by a node is the sum of the entering sig-

F. Beffa, *Weakly Nonlinear Systems*, Understanding Complex Systems,
https://doi.org/10.1007/978-3-031-40681-2

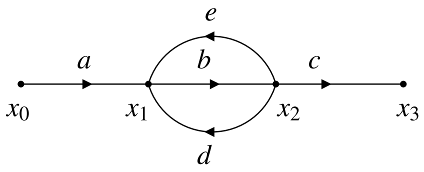


Fig. A.1 Signal-flow graph corresponding to Eq. (A.1)

nals. The value of the variable represented by any node is transmitted to all branches leaving the node. Nodes without incoming branches are *source* nodes. Nodes with only incoming branches are *sink* nodes.

Using the above rules SFGs can be used to represent systems of linear algebraic and convolution equations. The usefulness of SFGs comes from the fact that their graphical nature helps clarify the relation between variables and the existence of feedback loops. For example, the system of equations

$$\begin{aligned} x_1 &= ax_0 + dx_2 + ex_2 \\ x_2 &= bx_1 \\ x_3 &= cx_2 \end{aligned} \tag{A.1}$$

can be represented by the SFG shown in Fig. A.1.

Here and in the following, for simplicity of notation, we will use single letters to represent transmission factors and suppress all product symbols. Before proceeding discussing more aspects of SFGs it's convenient to introduce some specific terminology [37].

Intermediate node: A node with incoming and outgoing branches.

Path: A connection from a starting node to an end node by a continuous, unidirectional succession of branches all of which are traversed along the branch direction.

Open path: Any path not touching a node more than once.

Loop: A path starting and ending at the same node. All other nodes are touched at most once.

Self-loop: A loop touching a single node.

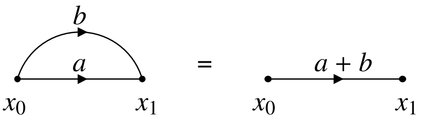
Non-touching loops: Loops without common nodes.

Path-/Loop-gain: The product of the transmission factors of the branches forming the path resp. the loop.

A.1.2 Reduction Rules

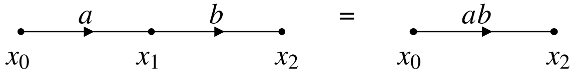
Many of the algebraic manipulations performed in solving linear equations can be translated in *reduction rules* for SFGs [37]. These are graphical rules to transform a given SFG in an equivalent, but simpler form. Hereafter we list the most useful rules together with the algebraic equations proving the equivalence.

1. *Parallel transformation*:



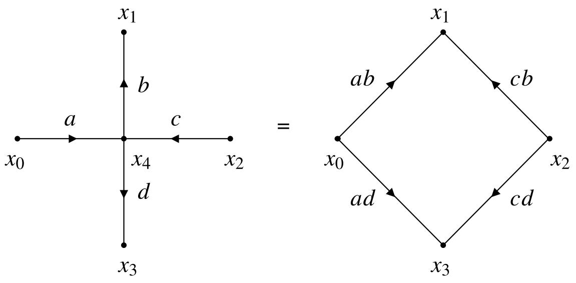
$$x_1 = ax_0 + bx_0 \qquad x_1 = (a+b)x_0$$

2. *Cascade transformation*:



$$x_2 = bx_1, x_1 = ax_0 \qquad x_2 = abx_0$$

3. *Star to mesh transformation*:



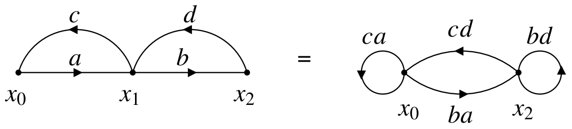
$$x_1 = bx_4 \qquad x_1 = abx_0 + cbx_2$$

$$x_3 = dx_4 \qquad x_3 = adx_0 + cdx_2$$

$$x_4 = ax_0 + cx_2$$

This rule can also be applied in the case in which a transmission factor is zero and in the case in which some nodes coincide.

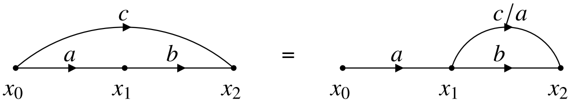
4. *Node elimination*:



$$x_0 = cx_1 \qquad\qquad x_0 = cax_0 + cdx_2$$
$$x_2 = bx_1 \qquad\qquad x_2 = bax_0 + bdx_2$$
$$x_1 = ax_0 + dx_2$$

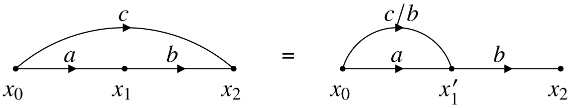
This is a special case of the star to mesh rule.

5. *Shifting start point*:



$$x_1 = ax_0 \qquad\qquad x_1 = ax_0$$
$$x_2 = bx_1 + cx_0 \qquad\qquad x_2 = (b + c/a)x_1$$

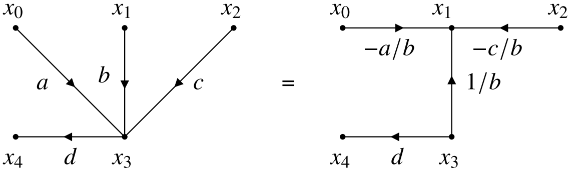
6. *Shifting end point*:



$$x_1 = ax_0 \qquad\qquad x_1' = (a + c/b)x_0$$
$$x_2 = bx_1 + cx_0 \qquad\qquad x_2 = bx_1'$$

Note that this transformation changes the variable x_1 into a new one $x_1' \neq x_1$. In spite of this the total transmission factor from x_0 to x_2 remains unaffected.

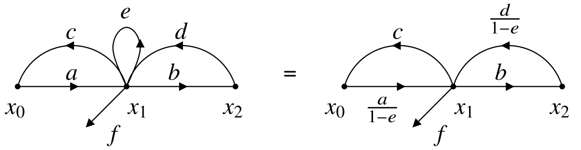
7. *Path inversion:* A branch can only be inverted if it starts at a *source* node (no incoming branches) as the branch with transmission factor b in the illustration.



$$x_3 = ax_0 + bx_1 + cx_2 \qquad\qquad x_1 = -\frac{a}{b}x_0 - \frac{c}{b}x_2 + \frac{1}{b}x_3$$

Note that application of this rule moves a source of the graph. Consecutive application of the rule allows inverting a path from a source to an arbitrary node.

8. *Self-loop elimination*:



$$\begin{aligned} x_0 &= cx_1 & \qquad x_0 &= cx_1 \\ x_2 &= bx_1 & x_2 &= bx_1 \\ x_1 &= ax_0 + dx_2 + ex_1 & x_1 &= \frac{a}{1-e}x_1 + \frac{d}{1-e}x_2 \end{aligned}$$

In the first few transformation rules only summations and multiplication of transmission factors occur. Therefore, if the initial transmission factors correspond to transfer functions (or impulse responses) of stable systems, so do the ones of the transformed graph. This is not necessarily the case with later rules entailing divisions.

A.1.3 Mason's Rule

One of the strength of SFGs is the fact that *Mason's rule* allow writing the transfer function (or impulse response) from a source node to another node by inspection [38]. Before stating the general rule we first focus on graphs with a single open path from a source node x_s to another node x_j and where the path touches every loop in the graph (that means that it has at lease a node in common with every loop). For this case the transmission is given by

$$T_{sj} = \frac{P_{sj}}{\Delta}$$

with P_{sj} the gain of the open path and Δ the graph *determinant* (or system determinant). If we denote the loop gain of loop i by L_i, the graph determinant is defined by

$$\Delta := 1 - \sum L_i + \sum L_i L_j - \sum L_i L_j L_k + \cdots ,$$

the first summation being over all loops, the second over all pairs of non-touching loops, the third over all triplets of non-touching loops, etc.

As an example consider the SFG depicted in Fig. A.2. From x_0 to x_4 there is a single path touching all three loops. The transmission T_{04} is

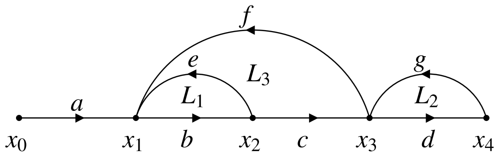


Fig. A.2 Single path multi-loop SFG

$$
\begin{aligned}
T_{04} &= \frac{P_{04}}{1-(L_1+L_2+L_3)+L_1L_2} \\
&= \frac{abcd}{1-(be+dg+bcf)+bedg}\,.
\end{aligned}
$$

Consider now the case of two open paths from a source node to another node. Since the graph represents a linear system, the output is the sum of the two contributions

$$T = \frac{P_1}{\Delta_1} + \frac{P_2}{\Delta_2}$$

where the individual contributions are calculated as for the single path case, with P_j the gain of path j and Δ_j the determinant associated with path j, that is, calculated discarding the loops not touched by path j. This expression can be rewritten as

$$T = \frac{P_1\Delta_2 + P_2\Delta_1}{\Delta}\,; \qquad \Delta = \Delta_1\Delta_2\,.$$

The denominator of this expression Δ is the determinant of the full SFG and each path transmission P_j is multiplied by the determinant of that part of the graph with no nodes in common with the path. Generalising this expression to more paths we obtain *Mason's rule*

$$T = \frac{\sum_j P_j\Delta_j}{\Delta} \tag{A.2}$$

where here, differently from above, Δ_j denotes the *co-factor* of path k defined as the determinant of that part of the graph which doesn't have any node in common with path j. A formal proof can be found in [39].

A.2 Nonlinear Systems

The use of signal-flow graphs with nonlinear systems is more limited. One (and the first) application is in studying the smallest number of implicit equations in a system of nonlinear equations [38]. In this context a branch simply represents a dependence of the variable represented by the node pointed to by the branch in question by the

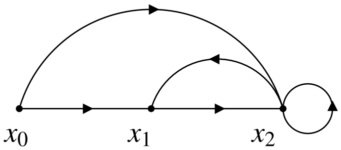


Fig. A.3 Example signal-flow graph representing Eq. (A.3)

variable represented by the node where the branch originates. For example the set of nonlinear equations

$$\begin{aligned} x_1 &= f_1(x_0, x_2) \\ x_2 &= f_2(x_0, x_1, x_2) \end{aligned} \tag{A.3}$$

is represented as in Fig. A.3. In this setting there is no concept of branch transmission. Also, parallel branches between nodes make no sense. In spite of this one can adapt those reduction rules based only on variable substitutions. For example, in the above example we can remove x_1 by a simple substitution obtaining a single implicit equation

$$x_2 = f_2(x_0, f_1(x_0, x_2), x_2) = \tilde{f}(x_0, x_2) .$$

A second way in which signal-flow graphs are used in conjunction with nonlinear systems consists in retaining most construction rules of linear SFGs, but allow the use of nonlinear functions instead of transmission factors. This approach is popular for example in the study of neural networks [40]. As an example Fig. A.4 shows the signal-flow graph of Rosenblatt's perceptron. In this model the input branches posses transmission factors labelled w_j, but the branch connecting to the output y represents the application of the nonlinear function $\varphi(.)$ to the signal v. The function $\varphi(.)$ is called activation function and in this context has a monotonic, limiting character.

This method is useful to analyse graphs without feedback loops. There is no equivalent of Mason's rule applicable in the presence of feedback loops.

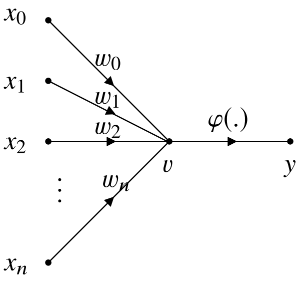


Fig. A.4 Signal-flow graph of a Perceptron

A.3 Weakly-Nonlinear Systems

The equations describing weakly-nonlinear time-invariant systems can be solved in an iterative way as described in Chap. 9. The component of order k of the Volterra series representation of a signal is calculated by solving a *linear* system of equations with products and powers of terms of order lower than k acting as sources. The signal-flow graph of a weakly-nonlinear system can thus be constructed as a linear SFG representing the linear part of the equations and where products and powers of signals are added as *source nodes*. In this way, all features of linear SFGs can be used, including Mason's rule.

One starts by calculating the first order response due to the external source (which is of order one) as with linear systems. With it one can then compute the value of the sources of second order. More generally, with the responses of order up to $k-1$ one can compute the sources of order k. The latter drive a linear system. The transfer function of order k can often be written by inspection using Mason's rule as follows

1. Compute the linear transfer function from the source to the node of interest.
2. Replace the single Laplace variable s of the linear transfer function by the sum $s_1 + \cdots + s_k$ (see Sect. 10.1).

The procedure works in the time domain as well. One has simply to use impulse responses and the convolution product, and adapt the second step.

The following example illustrates the use of SFGs to analyse WNTI systems. In Sect. 10.2 we use SFGs to illuminate the effect of distortion within a feedback loop.

Example A.1: Driven Pendulum

In this example we analyse the pendulum shown in Fig. A.5. A weight of mass m at the end of a massless rod is suspended from a pivot in Earth's gravitational field. We model the weight as a point mass whose position is specified by the angle ϕ. We assume the presence of some viscous fluid causing a drag proportional to the velocity of the weight. A motor drives the pendulum exerting a periodic torque M. The equation governing the dynamics of the system can be obtained from Newton's second law

$$D^2\phi + \frac{b}{lm}D\phi + \frac{g}{l}\sin\phi = \frac{1}{l^2m}K\cos(\omega t)$$

with $M(t) = K\cos(\omega t)$.

For convenience, we re-express the coefficients of the equation in terms of the standard parameters for systems of second order and normalise the amplitude of the torque

$$\omega_0 = \sqrt{\frac{g}{l}}\,; \qquad q = \sqrt{gl}\frac{m}{b}\,; \qquad A = \frac{K}{mlg}\,.$$

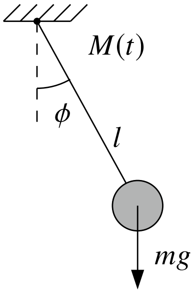


Fig. A.5 Driven pendumum in Earth's gravitational field

Further, using the generalised velocity $\alpha = D\phi$, expanding the sin function in its Taylor series

$$\sin\phi = \sum_{n=1,3,5,\dots} \frac{(-1)^{(n-1)/2}}{n!}\phi^n$$

and defining the input signal x as

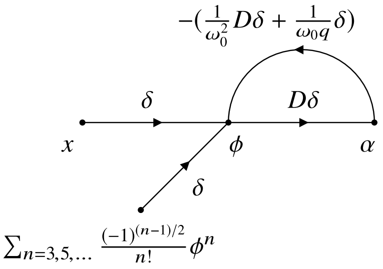
$$x(t) = A\cos(\omega t)\,,$$

the equation can be written as a system of two convolution equations relating α and ϕ

$$\begin{aligned} \alpha &= D\delta * \phi \\ \phi &= -(\frac{1}{\omega_0^2}D\delta + \frac{1}{\omega_0 q}\delta) * \alpha + x + \sum_{n=3,5,\dots} \frac{(-1)^{(n-1)/2}}{n!}\phi^n\,. \end{aligned} \tag{A.4}$$

A signal-flow graph representing these equations is shown in Fig. A.6.

We are interested in the steady-state oscillation when the pendulum is driven with a frequency equal to ω_0. For this reason it's convenient to work with transfer functions. The various transfer functions can be determined by inspection from the SFG using Mason's rule, using the transform of each transmission factor and applying the transform of an impulse as input. The first order transfer function is



$-(\frac{1}{\omega_0^2}D\delta + \frac{1}{\omega_0 q}\delta)$
δ
$D\delta$
x
ϕ
α
δ
$\sum_{n=3,5,\dots}\frac{(-1)^{(n-1)/2}}{n!}\phi^n$

Fig. A.6 Signal-flow graph corresponding to Eq. (A.4)

$$H_1(s_1) = \frac{1}{\frac{s^2}{\omega_0^2} + \frac{1}{q}\frac{s}{\omega_0} + 1} .$$

The linear part of the response is thus

$$\phi_1(t) = \Re\left\{H_1(\jmath\omega_0)A\mathrm{e}^{\jmath\omega_0 t}\right\} = Aq\sin(\omega_0 t) .$$

From the signal-flow graph it's immediately apparent that the transmission from the source representing the nonlinear terms to the node ϕ is the same as the one from x to ϕ. Using the fact that the response of a linear system to a source of order k is obtained by multiplying the Laplace transform of the source by the linear transfer function of the system and by replacing in the latter s_1 with the sum $s_1 + \cdots + s_k$ (see Sect. 10.1), the transfer function of order k from the "nonlinear" source of order k is therefore

$$H_1(s_1 + \cdots + s_k) .$$

There is no second order power in the source terms. Hence, the second order transfer function H_2 is identically zero. The third order transfer function is obtained by substituting $H_1 + H_2$ in the "nonlinear source" terms ϕ^n and retaining only terms of third order. This gives

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

hence

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

Note that H_3 is already symmetric.

The contribution of third order to the steady-stage oscillation at ω_0 is given by the frequency mix $m = (1, 2)$

$$\begin{aligned}\phi_{3,m}(t) &= \frac{A^3}{24}\frac{3!}{2!}\Re\left\{H_1(-\jmath\omega_1)H_1(\jmath\omega_1)H_1(\jmath\omega_1)H_1(\jmath\omega_1)\mathrm{e}^{\jmath\omega_0 t}\right\}\\ &= -A^3\frac{q^4}{8}\cos(\omega_0 t) .\end{aligned}$$

The fourth order transfer function is identically zero. The fifth order is obtained by inserting $H_1 + \cdots + H_4$ into the "nonlinear source" terms ϕ^n and retaining only terms of fifth order, giving

$$\frac{1}{6}\frac{3!}{2!}\left[H_1^{\otimes 2} \otimes H_3\right]_{\text{sym}} - \frac{1}{5!}H_1^{\otimes 5} .$$

The fifth order transfer function is then found by multiplying it by $H_1(s_1 + \cdots + s_5)$ giving

$$H_5(s_1, s_2, s_3, s_4, s_5) = \left\{ \frac{1}{2}\left[H_1(s_1)H_1(s_2)H_3(s_3, s_4, s_5)\right]_{\mathrm{sym}} \right.$$
$$\left. - \frac{1}{5!} H_1(s_1)H_1(s_2)H_1(s_3)H_1(s_4)H_1(s_5) \right\} H_1(s_1 + s_2 + s_3 + s_4 + s_5)\,.$$

The fifth order component at ω_0 produced by the frequency mix $m = (2, 3)$ is

$$y_{5,m}(t) = A^5 \frac{1}{2^4} \frac{5!}{2!3!} \Re\left\{ H_5(-\jmath\omega_0, -\jmath\omega_0, \jmath\omega_0, \jmath\omega_0, \jmath\omega_0)\mathrm{e}^{\jmath\omega_0 t} \right\}.$$

There are in total 10 ways to fill the slots of $\left[H_1^{\otimes 2} \otimes H_3\right]_{\mathrm{sym}}$ with permutations of the tuple $(-\jmath\omega_0, -\jmath\omega_0, \jmath\omega_0, \jmath\omega_0, \jmath\omega_0)$. One possibility is to put the negative frequencies in the first two slots giving

$$H_1(-\jmath\omega_0)H_1(-\jmath\omega_0)H_3(\jmath\omega_0, \jmath\omega_0, \jmath\omega_0) = \frac{\jmath q^5}{6(8 - \frac{3\jmath}{q})}\,.$$

There are 3 cases in which both H_1s are applied to positive frequencies

$$H_1(\jmath\omega_0)H_1(\jmath\omega_0)H_3(\jmath\omega_0, -\jmath\omega_0, -\jmath\omega_0) = \frac{q^6}{6}\,.$$

and 6 in which the two H_1s appear one with a positive frequency and the other with a negative frequency

$$H_1(-\jmath\omega_0)H_1(\jmath\omega_0)H_3(\jmath\omega_0, \jmath\omega_0, -\jmath\omega_0) = -\frac{q^6}{6}\,.$$

Summing all terms we obtain

$$H_5(-\jmath\omega_0, -\jmath\omega_0, \jmath\omega_0, \jmath\omega_0, \jmath\omega_0) = \left[\frac{1}{10}\left(\frac{\jmath q^5}{12(8 - \frac{3\jmath}{q})} - \frac{q^6}{4}\right) + \jmath\frac{q^5}{120}\right](-\jmath q)\,.$$

The first, third and fifth order approximations that we have found are compared to a numerical solution of the differential equation for $q = 7$ and $A = 0.1$ in Fig. A.7. Even though we didn't took into account the harmonics produced by nonlinear terms, the fifth order approximation at ω_0 gives a fairly accurate approximation up to a swing of ca. 40°. A linear model predicts that the pendulum dynamics settles in such a way that the peak of the applied torque occurs when the pendulum is in the vertical position ($\phi = 0°$). The more accurate model shows that at moderate swing levels this is not the case. The peak torque happens before the pendulum passes through the vertical position. It's also interesting to note that the main phase correcting term is $\phi_{3,m}$ whose phasor is perpendicular to the one of the linear term. However, being

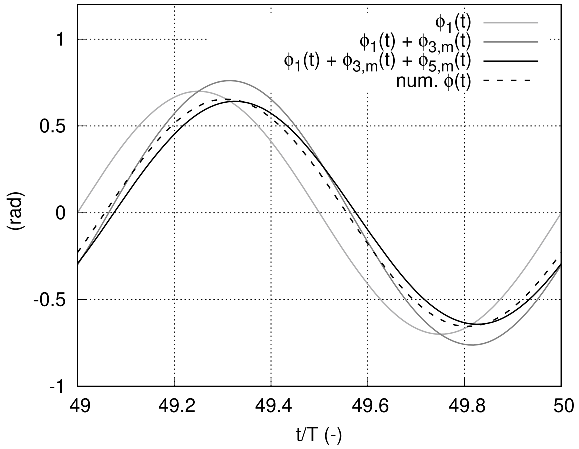


Fig. A.7 Comparison of the steady-stage solution of the driven pendumum obtained by numerical integration with the Volterra series of order one, three and five. $q = 7$, $A = 0.1$

at 90°, $\phi_{3,m}$ is unable to produce a good amplitude correction. For that we need to consider at least one more term.

References

1. Maxima—A computer algebra system. https://maxima.sourceforge.io
2. S.N. Laboratories. Xyce. https://xyce.sandia.gov
3. G.J. Sussman, J. Wisdom, *Structure and Interpretatino of Classical Mechanics*, 2nd edn. (MIT Press, 2014)
4. V. Volterra, Rend. Licei **3**(4), 97 (1887)
5. V. Volterra, *Theory of Functionals and of Integral and Integro-Differential Equations* (Dover Publications Inc., 1959)
6. M. Fréchet, Annales scientifiques de l'É.N.S., 3e séries, tome **27** (1910)
7. N. Wiener, Response of non-linear device to noise (Technical report, Massachussets Institute of Technology, 1942)
8. M.B. Brillant, Theory of the analysis of nonlinear systems (Technical report, Massachussets Institute of Technology, 1958)
9. D.A. George, Continuous nonlinear systems (Technical report, Massachussets Institute of Technology, 1959)
10. J.W. Graham, Nonlinear system modeling and analysis with applications to communication systems (Technical report, Signatron Inc., 1973)
11. N. Wiener, *Non-linear Problems in Random Theory* (Martino Publishing, 2013)
12. M. Schetzen, *The Volterra and Wiener Theories of Nonlinear Systems* (Wiley, New York, 1980)
13. D.D. Weiner, J.F. Spina, *Sinusoidal Analysis and Modeling of Weakly Nonlinear Circuits* (Van Nostrand Reinhold Ltd., 1980)
14. P.M. Dirac, *The Principles of Quantum Mechanics*, 4th edn. (Oxford University Press, 2006)
15. J. Mikusinski, *Operational Calculus* (Pergamon Press, 1959)
16. L. Schwartz, *Théorie des Distributions* (Hermann, 1966)
17. F. cois Trèves, *Topological Vector Spaces, Distributions and Kernels* (Dover Publications Inc., 1967)
18. A.H. Zemanian, *Distribution Theory and Transform Analysis* (McGraw-Hill, 1965)
19. G. van Dijk, *Distribution Theory* (de Gruyter, 2013)
20. L. Schwartz, *Mathematics for the Physical Sciences* (Dover Publications Inc., 1966)
21. H. Amann, J. Escher, *Analysis I* (Birkhäuser, 2006)
22. T. Kailath, *Linear Systems* (Prentice Hall, 1980)
23. H. Amann, J. Escher, *Analysis II* (Birkhäuser, 2006)
24. J.J.E. Slotine, W. Li, *Applied Nonlinear Control* (Prentice Hall, 1991)
25. P. Henrici, *Applied and Computational Complex Analysis* (Wiley, 1974)
26. A.S. Porret, T. Melly, C.C. Enz, E.A. Vittoz, IEEE J. Solid-State Circuits **35**(3), 337 (2000)

F. Beffa, *Weakly Nonlinear Systems*, Understanding Complex Systems,
https://doi.org/10.1007/978-3-031-40681-2

27. P.R. Gray, P.J. Jurst, S.H. Lewis, R.G. Meyer, *Analysis and Design of Analog Ingegrated Circuits*, 4th edn. (Wiley, 2001)
28. P.E. Allen, D.R. Holberg, *CMOS Analog Circuit Design* (Oxford University Press, 1987)
29. S. Khandelwal, et al., BSIM-CMG 110.0.0 multi-gate MOSFET compact model. http://bsim.berkeley.edu/models/bsimcmg/
30. S. Sinha, G. Yeric, V. Chandra, B. Cline, Y. Cao, in *Proceedings—Design Automation Conference* (2012), pp. 283–288. http://ptm.asu.edu/
31. A. Hastings, *The Art of Analog Layout* (Prentice-Hall Inc., 2001)
32. L.O. Chua, C.A. Desoer, E.S. Kuh, *Linear and Nonlinear Circuits* (McGraw-Hill Inc., 1987)
33. L. Berg, *Introduction to the Operational Calculus* (North-Holland Publishing company, Amsterdam, 1967)
34. H. Amann, *Linear and Quasilinear Parabolic Problems*, vol. I (Birkhäuser, 1995)
35. L. Schwartz, *Proceedings of the ICM* vol. 1 (1950)
36. L. Zadeh, Proc. IRE **38**(3), 291 (1950)
37. G.S. Moschytz, *Linear Integrated Networks—Fundamentals*. Bell Laboratories (Van Nostrand Reinhold Company, 1974)
38. S.J. Mason, Proc. IRE **41**(9), 1144 (1953)
39. N.J.A. Sloane, A.D. Wyner (eds.), *Claude E. Shannon: Collected Papers* (Wiley IEEE Press, 1993). Article titled “The theory and design of linear differential equation machines”
40. S. Haykin, *Neural Network and Learning Machines* (Pearson, 2009)

Index

F. Beffa, *Weakly Nonlinear Systems*, Understanding Complex Systems,
https://doi.org/10.1007/978-3-031-40681-2