1 Linear and Nonlinear Circuits

This chapter has a two-fold objective. First, it introduces the nomenclature that will be used throughout the book. Second, it presents the basic mathematical theory necessary to describe nonlinear systems, which will help the reader to understand their rich set of behaviors. This will clarify several important distinctions between linear and nonlinear circuits and their mathematical representations.

We shall start with a brief review of linearity and linear systems, their main properties and underlying assumptions. A reader familiarized with the linear system realm can understand the limitations of the theoretical abstraction framed in the linearity mathematical concept, realizing its validity borders and so be prepared to cross them, i.e., to enter the natural world of nonlinearity. We will then introduce nonlinear systems and the responses that we should expect from them. After this, we will study one static, or memoryless, nonlinearity and a dynamic one, i.e., one that exhibits memory. This will then establish the foundations of nonlinear static and dynamic models and their basic extraction procedures.

The chapter is presented as follows: Section 1.1 is devoted to nomenclature and Section 1.2 reviews linear system theory. Sections 1.3 and 1.4 illustrate the types of behaviors found in general nonlinear systems and, in particular, in nonlinear RF and microwave circuits. Then, Sections 1.5 and 1.6 present the theory of nonlinear static and dynamic systems that will be useful to understand the nonlinear circuit simulation algorithms treated in Chapter 2 and the device modeling techniques of Chapters 3–6. Mathematics of nonlinear systems, and in particular dynamic ones, is not easy or trivial. So, we urge you to not feel discouraged if you do not understand it after your first read. What you will find in the next chapters will certainly help provide a physical meaning and practical usefulness to most of these sometimes abstract mathematical formulations. Finally, Section 1.7 closes this chapter with a brief conclusion.

1.1 Basic Definitions

We will frequently use the notion of model and system, so it is convenient to first identify these concepts.

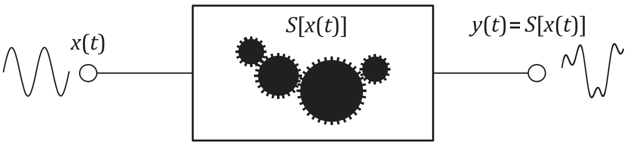

Figure 1.1 Illustration of the system concept.

1.1.1 Model

A model is a ***mathematical description***, or representation, of a set of particular features of a physical entity that combines the observable (i.e., measurable) magnitudes and our previous knowledge about that entity. Models enable the simulation of a physical entity and so allow a better understanding of its observed behavior and provide predictions of behaviors not yet observed. As models are simplifications of the physically observable, they are, by definition, an approximation and restricted to represent a subset of all possible behaviors of the physical device.

1.1.2 System

As depicted in Figure 1.1, a system is a model of a machine or mechanism that transforms an input (excitation, or stimulus, usually assumed as a function of time), $x(t)$, into an output (or response, also varying in time), $y(t)$. Mathematically, it is defined as the following operator: $y(t) = S[x(t)]$, in which $x(t)$ and $y(t)$ are, themselves, mathematical representations of the input and output measurable signals, respectively. Please note that, contrary to ordinary mathematical functions, which operate on numbers (i.e., that for a given input number, x, they respond with an output number, $y = f(x)$), mathematical operators map functions, such as $x(t)$, onto other functions, $y(t)$. So, they are also known as *mathematical function maps*. And, similar to what is required for functions, a particular input must be mapped onto a particular, unique, output.

When the operator is such that its response at a particular instant of time, $y(t_0)$, is only dependent on that particular input instant, $x(t_0)$, i.e., the system transforms each input value onto the corresponding output value, the operator is reduced to a function and the system is said to be ***static or memoryless***. When, on the other hand, the system output cannot be uniquely determined from the instantaneous input only but depends on $x(t_0)$ and its $x(t)$ past and future values, $x(t \pm \tau)$, i.e., the system is now an operator of the whole $x(t)$ onto $y(t)$, we say that the system is ***dynamic*** or that it exhibits ***memory***. (In practice, real systems cannot depend on future values because they must be causal.) For example, resistive networks are static systems, whereas networks that include energy storage elements (memory), such as capacitors, inductors or transmission lines, are dynamic.

Defined this way, this notion of a system can be used as a representation, or model, of any physical device, which can either be an individual component, a circuit or a set of circuit blocks. An interesting feature of this definition is that a system is nestable, i.e., it is such that a block (circuit) made of interconnected individual systems (circuit elements or components) can still be treated as a system. So, we will use this concept of system whenever we want to refer to the properties that we normally observe in components or circuits.

1.1.3 Time Invariance

Although the system response, $y(t)$, varies in time, that does not necessarily mean that the system varies in time. The change in time of the response can be only a direct consequence of the input variation with time. This time-invariance of the operator is expressed by stating that the system reacts exactly in the same way regardless at which time it is subjected to the same input. That is, if the response to $x(t)$ is $y(t) = S[x(t)]$, and another test is made after a certain amount of time, τ, then the response will be exactly the same as before, except that now it will be naturally delayed by that same amount of time $y(t - \tau) = S[x(t - \tau)]$. This defines a ***time-invariant*** system. If, on the other hand, $y(t - \tau) \neq S[x(t - \tau)]$, then the system is said to be ***time-variant***.

The vast majority of physical systems, and thus of electronic circuits, are time-invariant. Therefore, we will assume that all systems referred to in this and succeeding chapters are time-invariant unless otherwise explicitly stated.

After finalizing the study of this chapter, the reader may try Exercise 1.5 which constitutes a good example of how we can make use of this time-variance property for enabling us to treat, as a much simpler linear time-variant system, a modulator that is inherently nonlinear and time-invariant.

1.2 Linearity and the Separation of Effects

Now we will define a linear system as one that obeys superposition and recall how we use this property to determine the response of a linear system to a general excitation.

1.2.1 Superposition

A system is said to be linear if it obeys the principle of superposition, i.e., if it shares the properties of additivity and homogeneity.

The additivity property means that if $y_1(t)$ is the system response to $x_1(t)$, $y_1(t) = S[x_1(t)]$, $y_2(t)$ is the system's response to $x_2(t)$, $y_2(t) = S[x_2(t)]$, and $y_T(t)$ is the response to $x_1(t)+x_2(t)$, then

$$y_T(t) = S[x_1(t) + x_2(t)] = S[x_1(t)] + S[x_2(t)] = y_1(t) + y_2(t) \tag{1.1}$$

The additivity property is the mathematical statement that affirms that a linear system reacts to an additive composition of stimuli as an additive composition of responses, as if the system could distinguish each of the stimuli and treat them separately. In practical terms, this would mean that, if, in the lab, the result of an experiment with a cause $x_1(t)$ would produce an effect $y_1(t)$, and another, independent, experiment, on another cause $x_2(t)$, would produce $y_2(t)$, then, a third experiment, now made on a third stimulus $x_1(t)+x_2(t)$, would produce a response that is the numerical summation of the two previously obtained effects $y_1(t) + y_2(t)$.

On the other hand, the homogeneity property means that if α is a constant, then the response to $\alpha x(t)$ will be $\alpha y(t)$, i.e.,

$$S[\alpha x(t)] = \alpha S[x(t)] = \alpha y(t) \tag{1.2}$$

The homogeneity property is the mathematical description of proportionality that says that an α times larger cause produces an α times larger effect. However, it does not necessarily state that the effects are proportional to their corresponding causes. For example, although the current and the voltage in a constant (linear) capacitance obey the homogeneity principle, they are not proportional to each other. In fact, since the current in a capacitor is given by (1.3), the current to a twice as large $v_c(t)$ will be twice as large as $i_c(t)$. However, that does not mean that $i_c(t)$ is proportional to $v_c(t)$, as can be readily noticed when $v_c(t)$ is a ramp in time and $i_c(t)$ is a constant.

$$i_c(t) = C\frac{dv_c(t)}{dt} \tag{1.3}$$

In summary, ***linear systems*** obey the principle of ***superposition***,

$$\begin{aligned} S[\alpha_1 x_1(t) + \alpha_2 x_2(t)] &= S[\alpha_1 x_1(t)] + S[\alpha_2 x_2(t)] = \alpha_1 S[x_1(t)] + \alpha_2 S[x_2(t)] \\ &= \alpha_1 y_1(t) + \alpha_2 y_2(t) \end{aligned} \tag{1.4}$$

1.2.2 Response of a Linear System to a General Excitation

Superposition has very useful consequences that we now briefly review. They all revolve around that idea of the separation of effects, whereby we can expand any previously untested stimulus into a summation of previously tested excitations, making general predictions about the system responses.

1.2.2.1 Linear Response in the Time Domain

In the time domain, this means that, if we represent any input, $x(t)$, as composed of the succession of its time samples, taken at regular intervals, T_s, of a constant sampling frequency $f_s = 1/T_s$, so that they asymptotically produce the same effect of $x(t)$, $x(nT_s)T_s$,

$$x(t) \approx \sum_{n=-N}^{N} x(nT_s)T_s\delta(t - nT_s) \tag{1.5}$$

in which $\delta(t - nT_s)$ is the Dirac delta, or impulse, function centered at nT_s, where n is the number of samples, (see Figure 1.2(a)), and we know the response of the system to one of these impulse functions of unity amplitude, $h(t) = S[\delta(t)]$ (see Figure 1.2(b)), then we can readily predict the response to any arbitrary input $x(t)$ as

$$\begin{aligned} y(t) \equiv S[x(t)] &\approx S\left[\sum_{n=-N}^{N} x(nT_s)T_s\delta(t - nT_s)\right] = \sum_{n=-N}^{N} S[x(nT_s)T_s\delta(t - nT_s)] \\ &= \sum_{n=-N}^{N} x(nT_s)T_s S[\delta(t - nT_s)] = \sum_{n=-N}^{N} x(nT_s)h(t - nT_s)T_s \end{aligned} \tag{1.6}$$

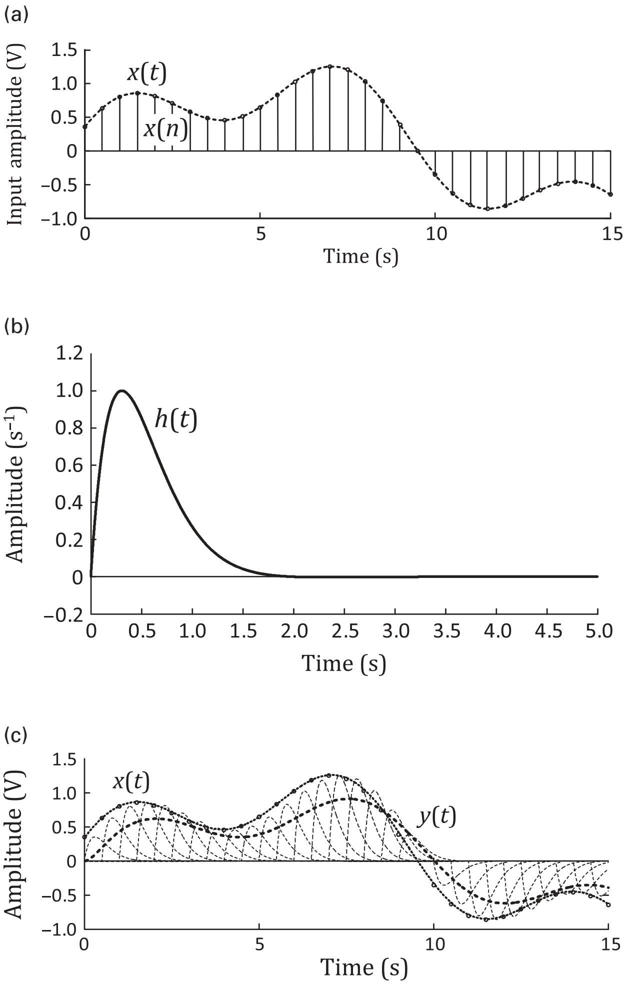

Figure 1.2. Response, $y(t)$, of a linear dynamic and time-invariant system to an arbitrary input, $x(t)$, when this stimulus is expanded in a summation of Dirac delta functions. (a) Input expansion with the base of delayed Dirac delta functions $x(n) = x(nT_s)\delta(t - nT_s)$. (b) Impulse response of the system, $h(t) = S[\delta(t)]$. (c) Response of the system to $x(t)$, $y(t) = S[x(t)]$.

by simply making use of the additivity and homogeneity properties (as shown in Figure 1.2(c)). Expression (1.6) is exact in the limit when the sampling interval, T_s, tends to zero and N tends to infinity, becoming the well-known convolution integral:

$$y(t) \equiv S[x(t)] = \int_{-\infty}^{\infty} x(\tau)h(t-\tau)d\tau = \int_{-\infty}^{\infty} h(\tau)x(t-\tau)d\tau \qquad (1.7)$$

1.2.2.2 Linear Response in the Frequency Domain

So, in time domain, we only needed to know the system response to one input basis function – the impulse response, $h(t) = S[\delta(t)]$, to be able to predict the response to any other arbitrary input. Similarly, in the frequency domain we only need to know the response to one input basis function, the cosine, although tested at all frequencies, to predict the response to any arbitrary periodic input.

Actually, since the cosine can be given as the additive combination of two complex exponentials

$$A\cos(\omega t) = A\frac{e^{j\omega t} + e^{-j\omega t}}{2} \tag{1.8}$$

from a mathematical viewpoint, we only need to know the response to that basic complex exponential. This response can be obtained from (1.7) as

$$\int_{-\infty}^{\infty} h(\tau)e^{j\omega(t-\tau)}d\tau = e^{j\omega t}\int_{-\infty}^{\infty} h(\tau)e^{-j\omega\tau}\,d\tau = H(\omega)e^{j\omega t} \tag{1.9}$$

in which $H(\omega)$ is the Fourier transform of $h(\tau)$. This is an interesting result that tells us that the response to an arbitrary $x(t)$ can be easily computed by summing up the Fourier components of that input scaled by the system's response to each particular frequency. Indeed, if $R(\omega)$ is the frequency-domain Fourier representation of a time-domain signal $r(t)$, so that

$$R(\omega) = \int_{-\infty}^{\infty} r(t)e^{-j\omega t}dt \tag{1.10a}$$

and

$$r(t) = \frac{1}{2}pi\int_{-\infty}^{\infty} R(\omega)e^{j\omega t}d\omega \tag{1.10b}$$

then, the substitution of (1.10) into (1.7) would lead to

$$Y(\omega) = H(\omega)X(\omega) \tag{1.11}$$

where $Y(\omega)$ can be related to $y(t)$ – as $X(\omega)$ is related to $x(t)$ – by the Fourier transform of (1.10). This expression tells us the following two important things.

First, the time-domain convolution of (1.7) between the input, $x(t)$, and the impulse response, $h(\tau)$, becomes the product of the frequency-domain representation of these two entities, $X(\omega)$ and $H(\omega)$, respectively.

Second, the response of a linear time-invariant system to a continuous-wave (CW) signal (an unmodulated carrier of frequency ω, specifically $\cos(\omega t)$) is another CW signal of the same frequency with, possibly, different amplitude and phase. Consequently, the response to a signal of complex spectrum will only have frequency-domain components at the frequencies already present at the input. A time-invariant linear system is incapable of generating new frequency components or of performing any qualitative transformation of the input spectrum.

Finally, equation (1.11) tells us that, in the same way we only needed to know the system's impulse response to be able to predict the response to any arbitrary stimulus in the time domain, we just need to know $H(\omega)$ to predict the response to any arbitrary periodic input described in the frequency domain. As an illustration, Figure 1.3 depicts the measured transfer function $S_{21}(\omega)$, in amplitude and phase, of a microwave filter.

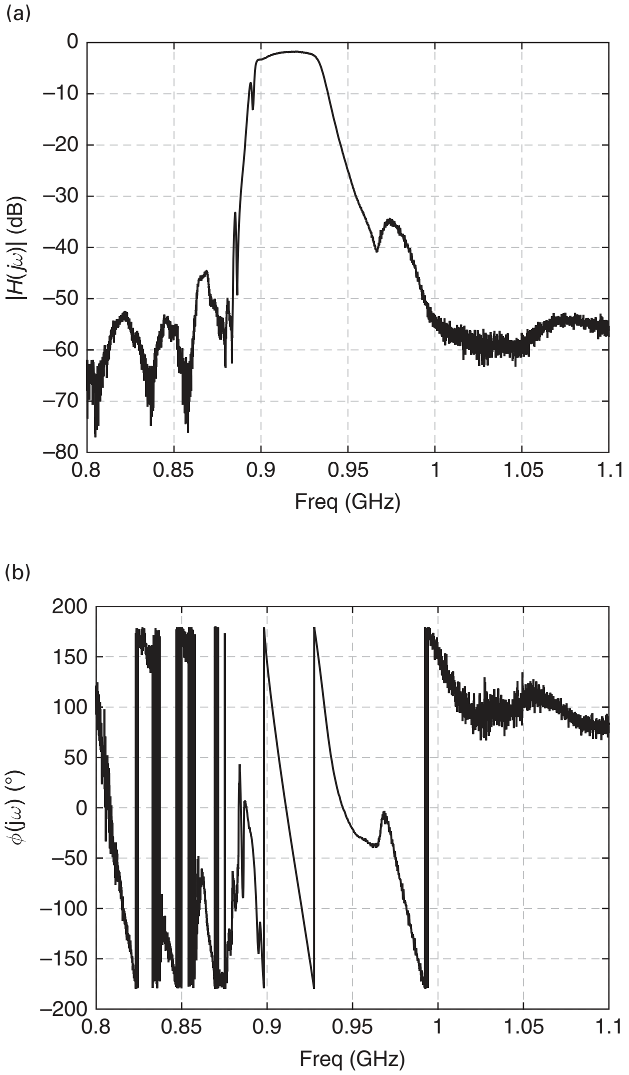

Figure 1.3 Example of the frequency-domain transfer function of a linear RF circuit, $H(\omega)$: measured forward gain, $S_{21}(\omega)$, in amplitude – (a) and phase – (b), of a microwave filter.

1.3 Nonlinearity: The Lack of Superposition

As all of us have been extensively taught and trained in working with linear systems, and with the additivity and homogeneity properties being so intuitive, we may easily fall into the trap of believing that these should be properties naturally inherent to all physical systems. But this is not the case. In fact, most of macroscopic physical systems behave very differently from linear systems, i.e., they are not linear. Actually, we use the term nonlinear systems to identify them.

Since we have been making the effort to define all important concepts used so far, we should start by defining a nonlinear system. But that is not a straightforward task as there is no general definition for these systems. There is only the unsatisfying definition of defining something by what it is not: a nonlinear system is one that is not linear, i.e., a nonlinear system is one that does not obey the principle of superposition. This is an intriguing, but also revealing, situation, which tells us that if linear systems are the ones that obey a precise mathematical principle, nonlinear systems are all the other ones. Hence, from an engineering standpoint the relevant question to be answered is: Are nonlinear systems often seen, or used, in practice? To demonstrate their importance, let us try a couple of very common, RF electronic examples. But, before these, the reader may want to try the two simpler examples discussed in Exercises 1.1–1.4.

Example 1.1 Active Devices and Amplifiers In this example we will show that any active device must be nonlinear.

As a first step, we will show that all active devices depend on two different excitations. One is the input signal and the other is the dc power supply. This means, as illustrated in Figure 1.4, that amplifiers are transducers that convert the power supplied by a dc power source into output signal power, i.e. they convert dc into RF power.

Now, as the second step in our attempt to prove that any active device must be nonlinear, let us assume, instead, that it could be linear. Then, it would have to obey the additivity property, which means that the response to each of the inputs, the signal and the power supply, should be determined separately. That is, the response to the auxiliary supply and to the signal should be obtained as if the other stimulus would not exist. And we would come back to an amplifier that could amplify the signal power without requiring any auxiliary power, thus violating the energy conservation principle.

Although this argument seems quite convincing, it raises a puzzling question, because, if it is impossible to produce amplifiers without requiring nonlinearity, we should be magicians as we all have already seen and designed linear amplifiers. So, how can we overcome this paradox?

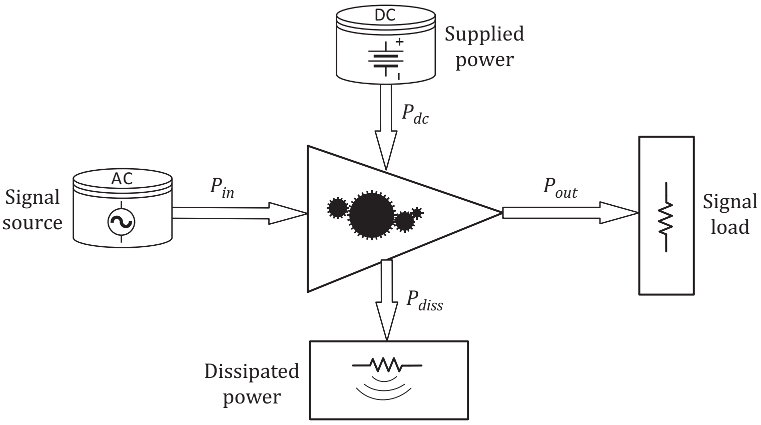

Figure 1.4 Illustration of the power flow in a transducer or amplifier.

According to the power flow shown in Figure 1.4, where P_{in}, P_{out}, P_{dc} and P_{diss} are, respectively, the signal input and output powers, the supplied dc power and the dissipated power (herein assumed as all forms of energy that are not correlated with the information signal, such as heat, harmonic generation, intermodulation distortion, etc.), the amplifier gain, G, can be defined by

$$G \equiv \frac{P_{out}}{P_{in}} \tag{1.12}$$

And this G must be constant and independent of P_{in} for preserving linearity.

Imposing the energy conservation principle to this transducer results in

$$P_{out} + P_{diss} = P_{in} + P_{dc} \tag{1.13}$$

from which the following constraint can be found for the gain:

$$G(P_{in}) = 1 + \frac{P_{dc} - P_{diss}}{P_{in}} \tag{1.14}$$

Since P_{diss} cannot decrease below zero (100% dc-to-RF conversion efficiency) and P_{dc} must be limited (as is proper from real power sources), $G(P_{in})$ cannot be kept constant but must decrease beyond a certain maximum P_{in}.

In RF amplifiers, this gain decrease with input signal power is called gain compression. In practice, amplifiers not only exhibit a gain variation when their input amplitude changes, but also an input-dependent phase shift. This is particularly important in RF amplifiers intended to process amplitude modulated signals as this input modulation is capable of inducing nonlinear output amplitude and phase modulations. These are the well-known AM/AM and AM/PM nonlinear distortions, often plotted as shown in Figure 1.5(a) and (b), respectively.

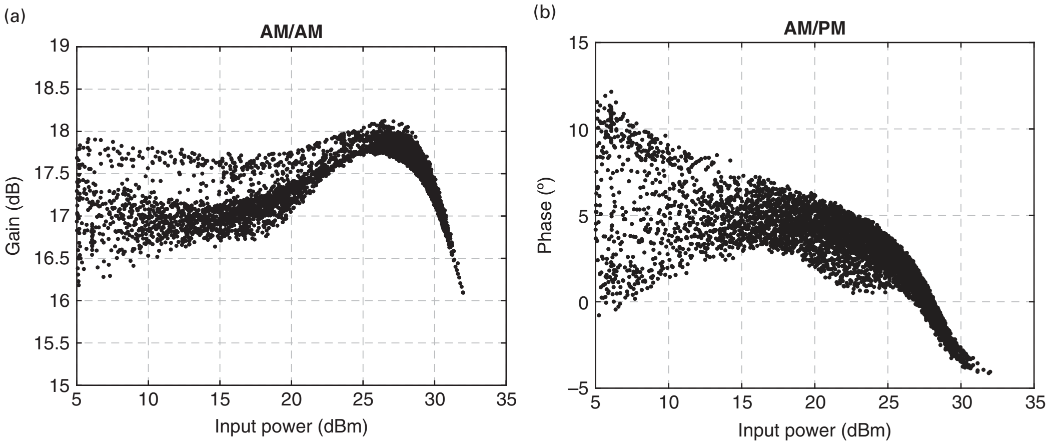

Figure 1.5 Illustration of measured (a) amplitude – AM/AM – and (b) phase-shift – AM/PM – gain variations as a function of input signal amplitude. Please note how these plots are not any idealized lines, but a cloud of dots that reveal hysteretic trajectories.

This analysis shows that linearity can only be obeyed at sufficiently small signal levels, and that it is only a matter of excitation amplitude to make an apparently linear amplifier expose its hidden nonlinearity.

Actually, this study provided us a much deeper insight of linearity and linear systems. Linearity is what we obtain when looking only at the system's input to output signal mapping (leaving aside the dc-to-RF energy conversion process) and when the signal is a very small perturbation of the dc quiescent point. So, linear systems are the conceptual mathematical model for the behaviors obtained from analytic operators (i.e., that are continuous and infinitely differentiable mappings), when these are excited with signals whose amplitudes are infinitesimally small as compared with the magnitude of the quiescent points. And it is under this small-signal operation regime that the linear approximation is valid. We will come back to this important concept later.

Example 1.2 A Sinusoidal Oscillator A sinusoidal oscillator is another system that depends on nonlinearity to operate. Although in basic linear system analysis we learned how to predict the stable and unstable regimes of amplifiers, and so to predict oscillations, we were not told the complete story. To understand why, we can just use the above results on the analysis of the amplifier and recognize that, by definition, an oscillator is a system that provides an output even without an input. That is, contrary to an amplifier that is a nonautonomous, or forced, system, an oscillator is an autonomous one. So, if it would not rely on any external source of power, it would violate the energy conservation principle. Like an amplifier, it is, instead, a transducer that converts energy from a dc power supply into signal power at some frequency ω. Hence, like the amplifier, it must rely on some form of nonlinearity. But, unlike the amplifier, in which we have shown that, seen from the input signal to the output signal, it could behave in an