1 GSW… Non-Linear Effects

One well-known characteristic of linear circuits is that they don't generate any new frequencies: the only frequencies in the output are those present at the input. If we're actually trying to generate new frequencies in the output (which we often have to do in radio transmitters and receivers), we'll clearly need something non-linear. Even when we're not trying to generate any new frequencies, real components tend not to be perfectly linear, and it's useful to be able to predict how much of a problem these non-linearities will cause.

1.1 Non-Linear Components in General

Many[1] non-linear components can have their transfer functions (output signal amplitude versus input signal amplitude) plotted on a graph, and approximated by a Taylor series of the form:

$$y(t) = A + Bx(t) + Cx^2(t) + Dx^3(t) + \ldots \qquad (1.1)$$

where $y(t)$ is the output from the component, and $x(t)$ is the input. The term A is a DC offset in the output, which for radio frequency circuits isn't usually very important, it's easy to remove with a capacitor. An example transfer function might look like this (with A set to zero):

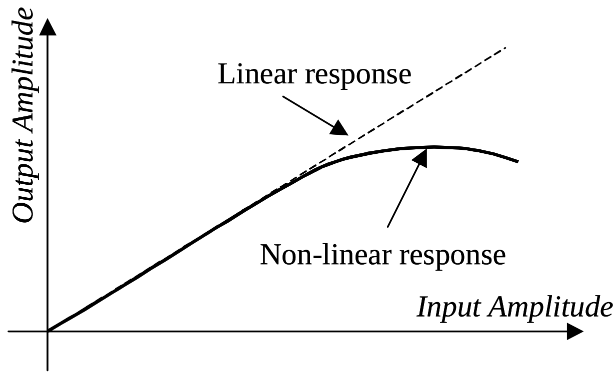

Figure 1-1 Non-Linear Characteristic

As an example, consider the following signal waveforms, typical of the input and output from a non-linear component (albeit rather exaggerated). If the component were perfectly linear, then the output would look exactly the same shape as the input (although perhaps a bit larger or smaller). A non-linear component changes the shape of the waveforms.

[1] There are some important non-linear devices that cannot be characterised in this simple way. Most importantly, this system does not work for any component with memory, so that the output at any time *t* is a function not only of the input at time *t*, but of the input at some previous times as well. Also, I'm assuming that there is no sudden discontinuity in the input-output characteristic, or the gradient of this characteristic. Despite these restrictions, it's a useful model in practice.

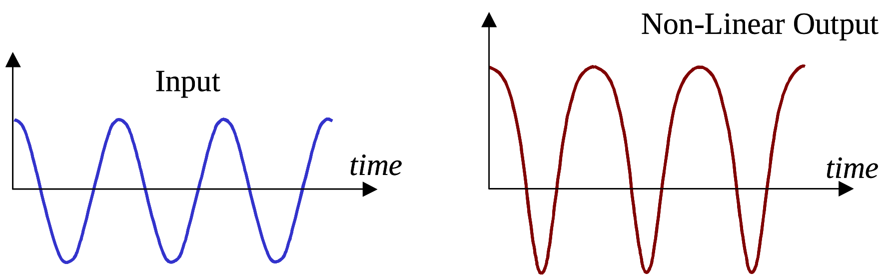

Figure 1-2 Typical Effect of a Non-Linear Stage on a Cosine Wave

Linear circuits have transfer functions of the form:

$$y(t) = B\,x(t) \tag{1.2}$$

where the amplitude gain is B (and hence the power gain is B^2). Note that for small input powers (where $x^2(t)$, $x^3(t)$ and so on can be neglected in comparison with $x(t)$) just about any component appears to be linear (apart from the DC offset term A, which is easy to get rid of).

1.1.1 Non-Linear Components with One Input Frequency

Put just one frequency ω_1 into a non-linear component characterised by the Taylor series expansion given in equation (1.1) above, and you'll get at the output:

$$y(t) = A + B\cos(\omega_1 t) + C\cos^2(\omega_1 t) + D\cos^3(\omega_1 t) + \ldots \tag{1.3}$$

and so on. Applying a few trigonometric identities to this gives:

$$y(t) = A + B\cos(\omega_1 t) + \frac{C}{2} + \frac{C}{2}\cos(2\omega_1 t) + \frac{3D}{4}\cos(\omega_1 t) + \frac{D}{4}\cos(3\omega_1 t) + \ldots \tag{1.4}$$

The output contains a whole set of different frequencies, all multiples of the input frequency ω_1. In general, a term in the Taylor series of the form $\cos^n(\omega_1 t)$ contains some power at the frequencies $n\omega_1$, $(n-2)\omega_1$, $(n-4)\omega_1$, and so on until you reach either DC or ω. For example, the term $C\cos^2(\omega_1 t)$ produces power at $2\omega_1$ and DC, and the term $D\cos^3(\omega_1 t)$ produces power at $3\omega_1$ and ω_1.

This is one way to generate new frequencies when required: put a single frequency into a non-linear component, generate all the harmonics, and then filter out the ones you don't want. It's sometimes the only way to generate energy at very high frequencies.

1.1.2 Non-Linear Components with Two Input Frequencies

Put a signal containing two different frequencies ω_1 and ω_2 into a non-linear component, and something much more interesting happens:

$$\begin{aligned} y(t) = A &+ B\left(\cos(\omega_1 t) + \cos(\omega_2 t)\right) \\ &+ C\left(\cos(\omega_1 t) + \cos(\omega_2 t)\right)^2 \\ &+ D\left(\cos(\omega_1 t) + \cos(\omega_2 t)\right)^3 \\ &+ \ldots \end{aligned} \tag{1.5}$$

which (after applying a few more well-known trigonometric identities) gives:

$$\begin{aligned} y(t) = A + B\left(\cos(\omega_1 t) + \cos(\omega_2 t)\right) \\ + \frac{C}{2}\cos(2\omega_1 t) + \frac{C}{2}\cos(2\omega_2 t) + C\cos\left((\omega_1 + \omega_2)t\right) + C\cos\left((\omega_1 - \omega_2)t\right) \\ + \frac{9D}{4}\left(\cos(\omega_1 t) + \cos(\omega_2 t)\right) + \frac{D}{4}\left(\cos(3\omega_1 t) + \cos(3\omega_2 t)\right) \\ + \frac{3D}{4}\left(\cos\left((2\omega_1 + \omega_2)t\right) + \cos\left((\omega_1 + 2\omega_2)t\right)\right) \\ + \frac{3D}{4}\left(\cos\left((2\omega_1 - \omega_2)t\right) + \cos\left((2\omega_2 - \omega_1)t\right)\right) \end{aligned} \tag{1.6}$$

and that's just from the terms up to $x^3(t)$. The terms in the sum and difference between the two input frequencies are sometimes used to make a simple mixer[2], these are the *second-order intermodulation products*. The terms at the frequencies $2\omega_1 - \omega_2$ and $2\omega_2 - \omega_1$ are also very important; and since they result from raising the input signal to a power of three, they are known as the *third-order intermodulation products*.

(A single non-linear circuit element used to provide the function of a mixer in this way is known as a *single-ended mixer*. It's a very simple way of generating the sum and difference frequencies, but it generates a lot of energy in unwanted frequencies (known as 'spurs'). About the only reasons to use a single-ended mixer is if you're trying to minimise the cost of the design and you're not too worried about performance; or if you're working at such high frequencies that nothing else works. At most radio frequencies, and with a bit of additional complexity (and hence cost), we can do much better.)

1.2 High-Power Amplifiers

Real high-power amplifiers (HPAs) often have significant non-linearity: in fact the most efficient HPAs (in terms of power output divided by battery consumption) are non-linear. In mobile devices, it's very good idea to minimise the power taken from the battery (and hence extend the battery life), but that means having to deal with the non-linearity of the HPA.

In terms of a signal being transmitted, this non-linearity has several serious effects. The first is that a range of harmonics of the transmitted signal are generated (twice the input frequencies, three times the input frequencies, and so on). This is what the final transmit filter in any transmitter is for: getting rid of these unwanted frequencies that would otherwise radiate energy in parts of the radio spectrum very likely being used by someone else.

The second is the distortion caused to the amplitude component of any modulation scheme. As the amplitude of the input signal increases, the amplitude of the output will deviate from a linear relationship with the amplitude of the input, and this will cause problems with any sort of amplitude modulation: doubling the input will no longer produce double the output. (In many cases there is an input level that produces the maximum possible output amplitude, and inputting a greater amplitude than this actually reduces the output amplitude.)

[2] In this sense, a mixer is a circuit that produces output at the sum and difference frequencies of its input.

The third, rather less obvious problem is that the phase of the transmitted signal becomes a function of amplitude. Increase the amplitude of the input signal, and the phase difference between the input and the output changes. This can also cause problems to any modulation scheme in which different amplitude signals are transmitted.

1.2.1 Constant-Envelope Signals

There is one important point here that might not be immediately obvious: a non-linear amplifier with an input signal that consists of a cosine wave with a constant envelope[3] (for example a system that uses phase modulation or frequency modulation) does not suffer any significant problems when a non-linear amplifier is used.

For example, suppose the input to this non-linear amplifier was a phase modulated signal with a constant envelope a, and a phase modulation of $\theta(t)$:

$$x(t) = a\cos\left(\omega_1 t + \theta(t)\right) \tag{1.7}$$

Consider what happens when you put such a signal through a non-linear amplifier:

$$y(t) = A + Ba\cos\left(\omega_1 t + \theta(t)\right) + Ca^2\cos^2\left(\omega_1 t + \theta(t)\right) + Da^3\cos^3\left(\omega_1 t + \theta(t)\right) + \ldots \tag{1.8}$$

Collecting only the terms around the carrier frequency (the others can usually be filtered out quite easily), this gives:

$$y(t) = \frac{4B + 3Da^3}{4}\cos\left(\omega_1 t + \theta(t)\right) \tag{1.9}$$

and the amplifier appears to have a gain of $(4B + 3Da^3 / 4)$. Yes, this gain is a function of the envelope of the input signal a, but with this phase modulated signal that envelope is constant. The amplifier (including the filter to get rid of the harmonics) is behaving exactly like a linear amplifier with a gain of $(B + 3Da^3 / 4)$.

However, for an amplitude modulation scheme, where the input to the non-linear amplifier has a non-constant envelope, for example:

$$x(t) = a(t)\cos\left(\omega_1 t\right) \tag{1.10}$$

then the output around the carrier frequency is now:

$$y(t) = \frac{4B + 3Da^3(t)}{4}\cos\left(\omega_1 t\right) \tag{1.11}$$

---

[3] There's some room for confusion in the term "amplitude" when used with oscillations. Sometimes the difference between the zero and maximum values of a cosine wave is referred to as the "amplitude". When this "amplitude" can vary, perhaps as a result of the modulation, I'll refer to it as an "envelope", and keep the term "amplitude" for the magnitude of the transmitted signal. So for example, with an amplitude modulated signal like:

$$x(t) = a(t)\cos\left(\omega_1 t\right)$$

I'll call $a(t)$ the *envelope*, and $a(t)\cos(\omega_c t)$ the *amplitude* of the signal.

and that's an amplifier without a constant gain: at different times it will have a different gain. That causes distortion, and that's a problem for the receiver.

1.2.2 Phase Responses

This non-linear relationship between input and output signal envelopes isn't the only problem real non-linear components cause. All the maths above assumed that the non-linear component had no memory: the output was a function of the input at the present time only, and not at any previous time. For real radio-frequency components this isn't a very good assumption, as the time it takes for a signal to get through a component can be a significant proportion of the period of the carrier frequency, and there are often several different paths that large signals can take through the component, each with different delays.

Consider a very simple (and completely unrealistic) example: suppose we had a non-linear device that we could represent using the response:

$$y(t) = Bx(t) + Dx^3(t-\tau) \tag{1.12}$$

Use a single frequency cosine wave as an input, and at low input signal levels we can neglect the third-order term, and this amplifier will have no phase difference between the input and the output signals:

$$y(t) = B\cos(\omega_1 t) \tag{1.13}$$

However at much higher signal levels, when the third-order term dominates, we can approximate the output as:

$$\begin{aligned} y(t) &= D\cos^3(\omega_1(t-\tau)) \\ &= \frac{3D}{4}\cos(\omega_1(t-\tau)) + \frac{D}{4}\cos(3\omega_1(t-\tau)) \end{aligned} \tag{1.14}$$

Filter out the signal around three times the input frequency, and we're left with:

$$y(t) = \frac{3D}{4}\cos(\omega_1(t-\tau)) \tag{1.15}$$

and that's an oscillation at the same frequency ω_1, but now there is a phase difference of $\omega_1\tau$ between the input and output signals. In-between these two extremes, the phase difference between the input and output will vary smoothly between zero and $\omega_1\tau$.

Real amplifiers have much more complex transfer functions than this, but they share the same symptoms: the phase change between the input and output of a non-linear component is a function of the input envelope. To characterise them, however, we have to use more approximate, empirical models.

1.2.3 The Saleh Model

One common empirical model of non-linearity in high-power amplifiers is due to Saleh[4]. Consider an input modulated signal of the form:

$$x(t) = a(t)\cos(\omega_1 t + \theta(t)) \tag{1.16}$$

In general, the output of a non-linear amplification stage around the carrier frequency (i.e. after filtering out any terms at DC and multiples of the input frequency ω_1) can be described as:

$$y(t) = b(t)\cos(\omega_1 t + \theta(t) + \phi(a(t))) \tag{1.17}$$

where the gain of the amplifier stage is $b(t)$ / $a(t)$ and the phase difference between the input and the output is $\phi(a(t))$, a function of the envelope of the input signal.

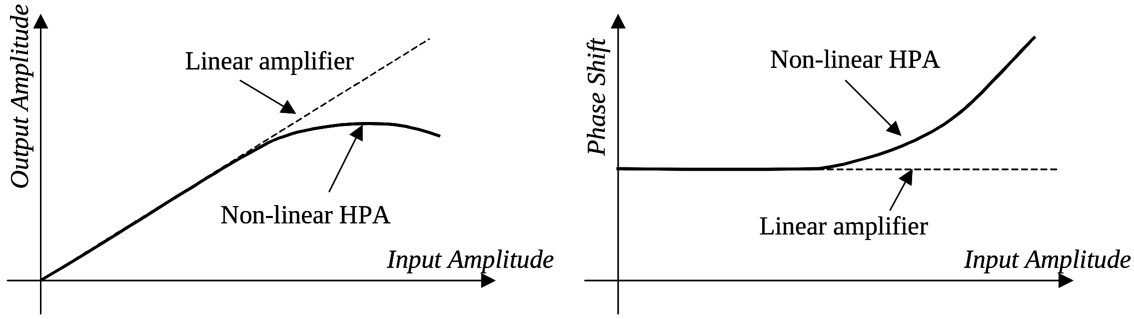

Figure 1-3 Transfer Characteristics of Non-Linear High-Power Amplifiers

In Saleh's model, the output envelope $b(t)$ is related to the input envelope $a(t)$ by an expression of the form:

$$b(t) = \frac{\alpha\, a(t)}{1 + \beta\, a^2(t)} \tag{1.18}$$

At low input amplitudes the gain is α, and at higher amplitudes the gain decreases. The phase difference between the input and output signals is given by:

$$\phi(a(t)) = \frac{\gamma\, a^2(t)}{1 + \delta\, a^2(t)} \tag{1.19}$$

so that at low input amplitudes, there is no phase change between the input and output, and as the amplitude increases, the phase difference increases. α, β, γ and δ are characteristic parameters of each HPA.

It's important to note here that $a(t)$ and $b(t)$ are envelopes of the input and output signals, not the fast-oscillating carrier frequency signals themselves. The envelope is the amplitude of the cosine wave, for example, see the diagram below:

[4] Saleh, A. "Frequency-independent and frequency-dependent non-linear models of TWT amplifiers", IEEE Transactions on Communications, vol. COM-29, no.11, November 1981, pp.1715-20.

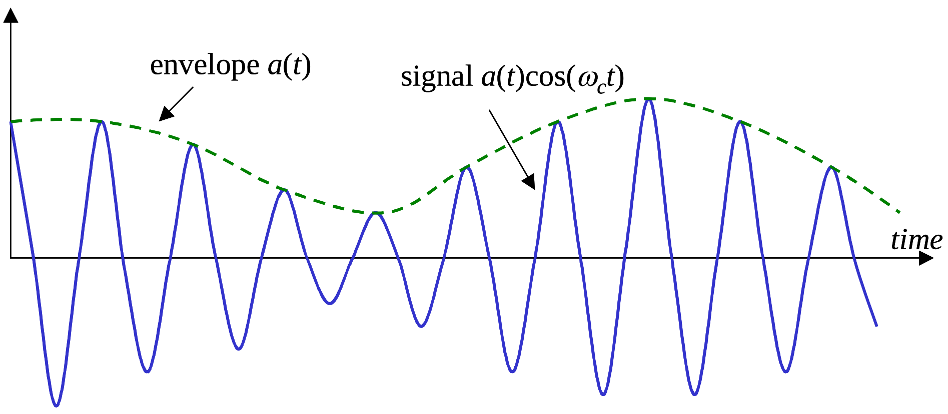

Figure 1-4 Envelope and Signal of Amplitude-Modulated Cosine Wave

Although this model was first produced for travelling-wave tube (TWT) amplifiers, it seems to work quite well for a wide range of amplifiers, with appropriate values for the parameters. Note that for a perfectly linear amplifier, β and γ must be zero.

1.2.4 The 1-dB Compression Point

Instead of the two parameters (α and β) used to specify the non-linearity in the Saleh model, the non-linearity of amplifiers (and other component) is more often quoted in terms of their small-signal gain and 1-dB compression point. The 1-dB compression point is the input power level that results in an output 1 dB less than the output power of a perfectly linear amplifier with the same small-signal gain.

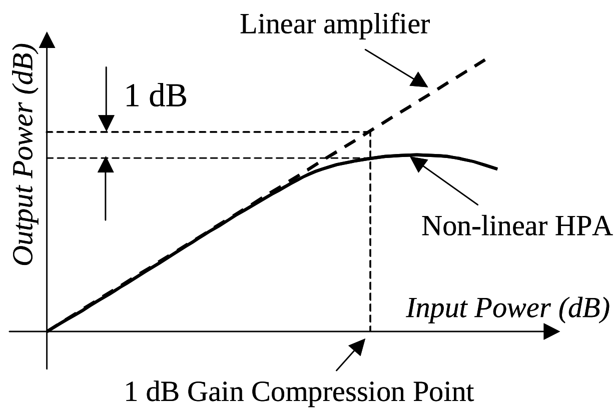

Figure 1-5 The 1-dB Gain Compression Point

In the case of the Saleh model, the 1-dB gain compression point occurs when:

$$a^2(t) = \frac{10^{0.05} - 1}{\beta} \tag{1.20}$$

where $a(t)$ is the envelope of the signal at the 1-dB gain compression point.

1.2.5 The Rapp Model

Another common model for non-linear transmit amplifiers is the Rapp model[5]. This does not model the phase change at all, but defines the amplitude gain as:

$$g\left(A\right)=\frac{gA_{in}}{\left(1+\left(\frac{A_{in}}{A_{sat}}\right)^{2p}\right)^{\frac{1}{2p}}} \qquad (1.21)$$

where g is the small-signal gain, p is typically between two and three and A_{sat} is the saturation amplitude, and determines at what input level the non-linear effects begin to be significant.

1.3 Intermodulation Products

Measuring the non-linearity of components with a single frequency input in terms of the 1-dB compression point is typical of how the HPA used at the transmitter could be characterised.

At the receiver, signal levels are typically much lower, and there isn't just one frequency to consider: the antenna picks up lots of frequencies from all over the radio spectrum. In this case, it's more useful to specify non-linearity in terms of *intermodulation products*: frequencies that depend on both input frequencies, rather than being harmonics of each individual input frequency. Usually only the second-order intermodulation products (resulting from the term in $Cx^2(t)$), and the third-order intermodulation products (resulting from the term in $Dx^3(t)$) are considered, since higher-order intermodulation products don't often produce any significant additional problems.

Note from equation (1.6) that putting two frequencies into a non-linear component results in intermodulation products at the sum and difference frequencies (the second-order terms), and at $2\omega_1 - \omega_2$ and $2\omega_2 - \omega_1$, (the third-order terms).

These intermodulation products can cause real problems to a receiver even when the signal levels are much lower than those found in transmitters. Consider an FM radio receiver trying to receive a low-power signal at 104 MHz, right next door to two powerful TV transmitters operating at 474 MHz and 578 MHz (UHF band 21 and 34 in the UK). If a non-linear component in the receiver produces a significant amount of power at the difference frequency between these two higher power TV signals, it could easy disrupt the reception of the FM radio signal (578 – 474 = 104).

(This is how second-order intermodulation terms can be measured[6]: two input frequencies that differ by the carrier frequency are input to the receiver. Mixing of these frequencies in the non-linear component produces a term at the carrier frequency. The amplitude of these two interfering signals is then increased until the quality of the received signal at the carrier frequency is degraded in some defined way.)

[5] Rapp, C., "Effects of HPA-Nonlinearity on a 4-DPSK/OFDM-Signal for a Digital Sound Broadcasting System", Proceedings of the Second European Conference on Satellite Communications, Oct. 22-24, 1991, pp. 179-184.

[6] From T/R 24-03 "Radio Characteristics of Cordless Telephones", by working group T/WG3 of the European Radiocommunications Office.

1.3.1 Intermodulation Intercept Points

The amount of intermodulation interference produced by a non-linear component given several input frequencies is usually specified in terms of an *intercept point*. This is a theoretical point on a graph of the output power of a non-linear component (or the whole receiver) in terms of the input powers.

For example, modelling the non-linearity with the transfer function:

$$y(t) = A + Bx(t) + Cx^2(t) + Dx^3(t) + \ldots \qquad (1.22)$$

and using a signal input at the carrier frequency ω_c with an envelope a, so that:

$$x(t) = a\cos(\omega_c t) \qquad (1.23)$$

would result in an output with envelope Ba at low signal powers (where the DC, second and higher-order terms can be neglected.)

1.3.1.1 Second-Order Intermodulation Intercept Points

Apply two frequencies spaced apart by the carrier frequency ω_c, both with an envelope a to the input:

$$x(t) = a\cos(\omega_1 t) + a\cos((\omega_1 + \omega_c)t) \qquad (1.24)$$

and amongst the output frequencies produced:

$$\begin{aligned} y(t) = A &+ B\left(a\cos(\omega_1 t) + a\cos((\omega_1 + \omega_c)t)\right) \\ &+ \frac{Ca^2}{2}\cos(2\omega_1 t) + \frac{Ca^2}{2}\cos(2(\omega_1 + \omega_c)t) \\ &+ Ca^2\cos((2\omega_1 + \omega_c)t) + Ca^2\cos(\omega_c t) \\ &+ \ldots \end{aligned} \qquad (1.25)$$

is one term at the carrier frequency the receiver is tuned to receive. That's the term:

$$Ca^2\cos(\omega_c t) \qquad (1.26)$$

If you assume that all three input signals (the two interfering signals that produce the intermodulation product at ω_c, and the original signal also at ω_c) arrive at the receiver with the same envelope a, then you can plot the output power resulting from the desired signal the receiver is trying to listen to:

$$Ba\cos(\omega_c t) \qquad (1.27)$$

and the output power from the intermodulation product of the two interfering signals ω_c apart in frequency:

$$Ca^2\cos(\omega_c t) \qquad (1.28)$$

against the input power (the same power for all three inputs) on the same graph. Noting that the power in a cosine wave is equal to one-half of the square of the envelope[7], this gives as the output signal power:

$$\begin{aligned}\text{Output Signal Power}(\text{W}) &= \frac{(Ba)^2}{2} \\ \text{Output Signal Power}(\text{dBW}) &= 20\log_{10}(B) + 20\log_{10}(a) - 10\log_{10}(2)\end{aligned} \quad (1.29)$$

and the input power is similarly:

$$\begin{aligned}\text{Input Power}(\text{W}) &= \frac{a^2}{2} \\ \text{Input Power}(\text{dBW}) &= 20\log_{10}(a) - 10\log_{10}(2)\end{aligned} \quad (1.30)$$

The interference power from the second-order intermodulation product of the two incoming signals at $\omega_1+\omega_c$ and ω_1 is then:

$$\begin{aligned}\text{Output Interference Power}(\text{W}) &= \frac{(Ca^2)^2}{2} \\ \text{Output Interference Power}(\text{dBW}) &= 20\log_{10}(C) + 40\log_{10}(a) - 10\log_{10}(2)\end{aligned} \quad (1.31)$$

Using these equations to express the output power (both signal and interference) in terms of the input power of the three input signals (all assumed to be equal) then gives:

$$\text{Output Signal Power}(\text{dBW}) = \text{Input Power}(\text{dBW}) + 20\log_{10}(B) \quad (1.32)$$

and

$$\begin{aligned}\text{Output Interference Power}(\text{dBW}) = {} & 2 \times \text{Input Power}(\text{dBW}) \\ & -10\log_{10}(2) + 20\log_{10}(C)\end{aligned} \quad (1.33)$$

If the output power (in dBW) is plotted against the input power (also in dBW) on the same graph, then the linear term (from equation (1.32)) is a straight line with a gradient of one, and the second-order term (from equation (1.33)) is a straight line with a gradient of two.

It looks like this:

[7] At least if we are assuming, as we usually do, that the impedance is one. If we wanted to know the real power in Watts, we'd have to divide by the impedance of the circuit at this point, since the mean-square value of a cosine wave is one-half:

$$Power = \frac{\overline{V^2}}{Z} = \frac{\overline{(a\cos(\omega_c t))^2}}{Z} = \frac{a^2}{Z}\overline{\cos^2(\omega_c t)} = \frac{a^2}{2Z}$$

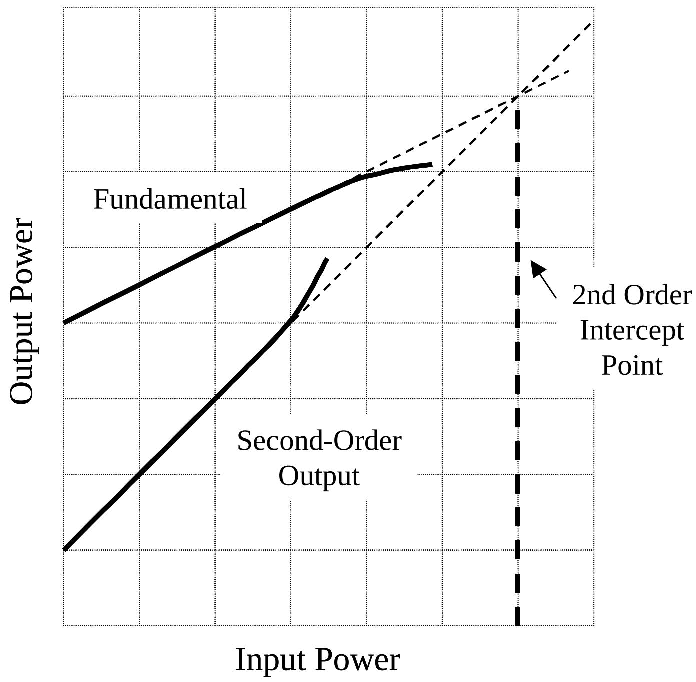

Figure 1-6 - Second Order Intercept Points

Since they have different gradients, at some point, the two curves must cross. The input power when this happens is known as the *second-order intercept point*. It's usually not achievable in practice, it would take so much input power to get there that the circuit would stop working. It's just a theoretical figure used to characterise the quality of the receiver at much lower input power levels.

This intercept point happens at an input amplitude (for all three signals) of:

$$\text{Output Interference Power} = \text{Output Signal Power}$$

$$\frac{\left(Ca^2\right)^2}{2} = \frac{\left(Ba\right)^2}{2} \tag{1.34}$$

$$a = \frac{B}{C}$$

Second-order non-linearities are not usually a performance-limiting problem with receivers, since the two input frequencies must be spaced apart by the carrier frequency of the signal being received, which means it is usually easy to attenuate at least one of them by a filter before any non-linear stages in the receiver. Much more troublesome are the third-order non linearities.

Consider: if there are two interfering signals of amplitude a at $\omega_c - \Delta\omega$ and $\omega_c - 2\Delta\omega$, then any third-order non-linearity term in the transfer function:

$$Dx^3\left(t\right) \tag{1.35}$$

will produce an output of:

$$
\begin{aligned}
&D\left(a\cos\left(\omega_c-\Delta\omega\right)+a\cos\left(\omega_c-2\Delta\omega\right)\right)^3 \\
&\quad=\frac{9Da^3}{4}\cos\left(\omega_c-\Delta\omega\right)+\frac{9Da^3}{4}\cos\left(\omega_c-2\Delta\omega\right) \\
&\qquad+\frac{Da^3}{4}\cos\left(3\omega_c-3\Delta\omega\right)+\frac{Da^3}{4}\cos\left(3\omega_c-6\Delta\omega\right) \\
&\qquad+\frac{3Da^3}{4}\cos\left(3\omega_c-5\Delta\omega\right)+\frac{3Da^3}{4}\cos\left(\omega_c-3\Delta\omega\right) \\
&\qquad+\frac{3Da^3}{4}\cos\left(3\omega_c-4\Delta\omega\right)+\frac{3Da^3}{4}\cos\left(\omega_c\right)
\end{aligned}
\tag{1.36}
$$

Note the last term. This is proportional to the third power of the input amplitude, and lies directly at the carrier frequency ω_c: also, the two interfering signals do not have to be very far from the carrier frequency so they are much more difficult to filter out before the non-linear element. Any radio system that divides its frequency allocation into adjacent channels (e.g. GSM, Bluetooth, WiMAX, Wi-Fi etc), can be affected by these intermodulation products.

Expressing this in terms of the power in the input signals (all of envelope *a*), gives:

$$
\begin{aligned}
\text{Output Interference Power}\left(\text{W}\right)&=\frac{1}{2}\left(\frac{3Da^3}{4}\right)^2 \\
\text{Output Interference Power}\left(\text{dBW}\right)&=20\log_{10}\left(D\right)+60\log_{10}\left(a\right)+10\log_{10}\left(9/32\right)
\end{aligned}
\tag{1.37}
$$

and in terms of the input power

$$
\text{Input Power}\left(\text{dBW}\right)=20\log_{10}\left(a\right)-10\log_{10}\left(2\right)
\tag{1.38}
$$

this gives:

$$
\begin{aligned}
\text{Output Interference Power}\left(\text{dBW}\right)&=20\log_{10}\left(D\right)+10\log_{10}\left(9/4\right) \\
&\quad+3\times\text{Input Power}\left(\text{dBW}\right)
\end{aligned}
\tag{1.39}
$$

This time, if the output power (in dBW) is plotted against the input power (also in dBW) on the same graph, then as before the linear term (from equation (1.38)) is a straight line with a gradient of one, but the third-order term (from equation (1.39)) is a straight line with a gradient of three.

It looks like this:

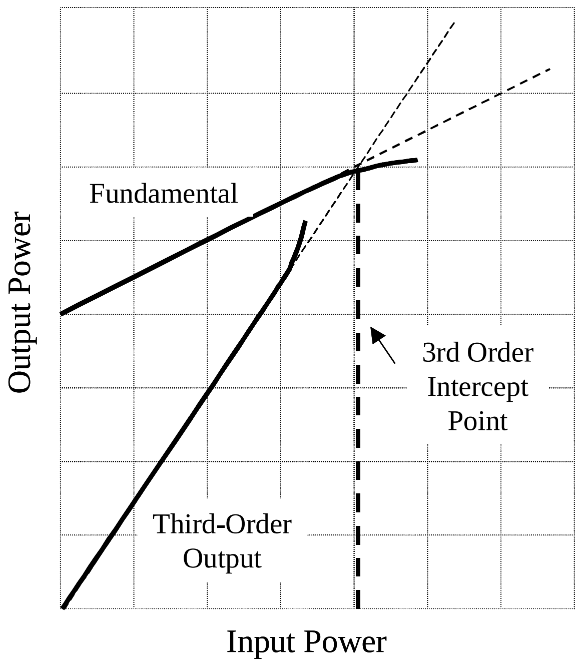

Figure 1-7 - Third-Order Intercept Point

Once again, the two lines meet when the signal output power and the intermodulation product output power are equal, and this time the input power when this occurs is known as the third order intercept point (IP_3). The zero-to-peak amplitude of the three input signals at this point can be calculated by setting the output signal power equal to the output intermodulation power:

$$\frac{1}{2}\left(\frac{3Da^3}{4}\right)^2 = \frac{1}{2}\left(Ba\right)^2$$
$$a = \sqrt{\frac{4B}{3D}} \tag{1.40}$$

so the IP_3 expressed in Watts and dBW in terms of B and D is:

$$IP_3\left(\mathrm{W}\right) = \frac{1}{2}\left(\sqrt{\frac{4B}{3D}}\right)^2 = \frac{1}{2}\left(\frac{4B}{3D}\right)$$
$$IP_3\left(\mathrm{dBW}\right) = 10\log_{10}\left(\frac{4B}{3D}\right) - 10\log_{10}\left(2\right) \tag{1.41}$$

Just like the second-order intercept point this is a theoretical input power only, no-one would ever put this much power into a non-linear component, and they'd probably permanently damage the component if they tried.

1.3.2 Calculating the Intermodulation Product

From equation (1.37) we've already got:

$$\text{Output Interference Power}\left(\mathrm{W}\right) = \frac{1}{2}\left(\frac{3Da^3}{4}\right)^2 \tag{1.42}$$

and the IP_3 for the component in terms of the third-order gain D and linear gain B from equation (1.41) is:

$$IP_3(\mathrm{W}) = \frac{1}{2}\left(\sqrt{\frac{4B}{3D}}\right)^2 = \frac{1}{2}\left(\frac{4B}{3D}\right) \tag{1.43}$$

and the input power when the input envelope is a is:

$$S_{in}(\mathrm{W}) = \frac{1}{2}a^2 \tag{1.44}$$

Combining these equations gives an expression for the power in the intermodulation product as a function of the input power and the IP_3:

$$\begin{aligned}\text{Output Interference Power}(\mathrm{W}) &= \frac{1}{2}\left(\frac{3Da^3}{4}\right)^2 \\ &= 4B^2\left(\frac{3D}{4B}\right)^2\left(\frac{1}{2}a^2\right)^3 \\ &= B^2\frac{1}{IP_3^{\,2}}S_{in}^{\,3}\end{aligned} \tag{1.45}$$

B is the linear amplitude gain of the stage, so the power gain of the stage G will be B^2 (since power is proportional to the square of the amplitude). Therefore we can write this as:

$$\text{Output Interference Power}(\mathrm{W}) = \frac{G}{IP_3^{\,2}}S_{in}^{\,3} \tag{1.46}$$

1.3.3 Calculating the Signal to Distortion Ratio

Given a receiver with a third-order intermodulation intercept point IP_3 in watts, it's quite simple to calculate the ratio of the signal power to the power in the intermodulation product at the carrier frequency being used in the wanted channel (known as the *signal to distortion ratio*) for any given input power S_{in} for the three frequencies.

The signal to distortion ratio is given by:

$$\begin{aligned}\mathrm{SDR} &= \frac{\text{Output Signal Power}}{\text{Output Interference Power}} \\ &= \frac{(Ba)^2/2}{\left(\frac{3Da^3}{4}\right)^2/2} = \frac{B^2a^2}{\frac{9D^2a^6}{16}} = \frac{16B^2}{9D^2a^4}\end{aligned} \tag{1.47}$$

Of course, it's not usually B and D that are specified for a component, it's the IP_3 and the input signal power. We can determine the SDR in terms of these parameters by noting that the input signal power is:

$$S_{in} = \frac{a^2}{2} \tag{1.48}$$

and the IP_3 is:

$$IP_3(\mathrm{W}) = \frac{1}{2}\left(\frac{4B}{3D}\right) = \left(\frac{4B}{6D}\right) \tag{1.49}$$

so that:

$$\mathrm{SDR} = \frac{16B^2}{9D^2a^4} = \frac{16B^2/4}{9D^2a^4/4} = \frac{(4B/6D)^2}{\left(a^2/2\right)^2} = \frac{IP_3^{\,2}}{S_{in}^{\,2}} \tag{1.50}$$

or in decibels:

$$\begin{aligned} SDR\,(\mathrm{dB}) &= 2IP_3(\mathrm{dBW}) - 2S_{in}(\mathrm{dBW}) \\ &= 2IP_3(\mathrm{dBm}) - 2S_{in}(\mathrm{dBm}) \end{aligned} \tag{1.51}$$

and the same thing is true in dBμW, dBnW, and anything else (provided the IP_3 and S_{in} are both in the same units).

It’s a very easy formula to use. Take the IP_3 (in dBm), subtract the power input S_{in} (in dBm) and multiply by two. The result is the signal to distortion ratio. (Although remember that this formula only works if all three inputs (the wanted signal and the two signals that create the intermodulation product) are all at the same power level.)

1.4 Tutorial Questions

1) Explain the effect of using a non-linear HPAs on:

i) An amplitude modulated signal
ii) A frequency modulated signal

*2) A non-linear HPA can be modelled using the Saleh model with parameters α = 40, and β = 3. What is the 1 dB compression point of this amplifier? (In other words, what value of input power would produce an output 1 dB less than a perfectly linear amplifier with the same gain at low input levels?)

If your transmitter needed 99% linear amplification (in other words the power gain at the maximum input signal is 99% of the small-signal power gain), what’s the maximum input power you could use? How much lower is this than the 1 dB compression point?

3) Explain why third-order intermodulation products are much more of a problem for receivers than second-order intermodulation products.

4) A receiver with an overall IP_3 of 18 dBm is receiving a signal at 433 MHz with a received power of 3 dBm. Also present at the input to the receiver are two other signals, one at 433.01 MHz with a power of 0 dBm and one at 433.02 MHz with a power of 2 dBm. What signal-to-distortion ratio would be expected due to the third-order intermodulation product?