ECE 4880: RF Systems

Fall 2016

Chapter 5: Nonlinearity

Reading Assignments:

1. W. F. Eagan, *Practical RF System Design*, Wiley, 2003, Chap. 4.

Game Plan for Chap. 5: Nonlinearity

1. Representing module nonlinearity in the Taylor series of v_{in}
2. 2^{nd} order and 3^{rd} order terms in the two-tone signal treatment and intermodulation (IM)
3. Representing nonlinearity with intercept point (IP)
4. Nonlinearity in the module cascade

5.1 Representing module nonlinearity

RF signals are composed of a carrier (modulated and demodulated by LO) and the frequency components around the carrier. The bandwidth of a channel corresponds to the bandwidth of the baseband. The goal of a successful RF transmission is to retrieve the baseband after modulation in the transmitter (shifting to higher frequency for **smaller antenna** and **more available spectral space**) and demodulation in the receiver (shifting back to be decoded). In Fig. 5.1, we can see how a "RF signal" is composed. There are desirable spectral components from the baseband, and flicker LO phase noise together with the thermal/shot noise floor. The frequency components in the original data can "intermodulate" to cause additional interference under module nonlinearity. These additional frequency components can interfere with the spectral composition of the original baseband and cause errors in data retrieval. Therefore, we need to understand and control the magnitude of the intermodulation (IM) for successful RF transmission.

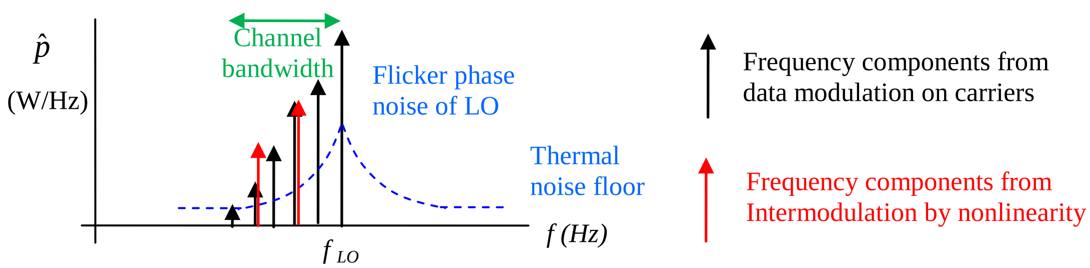

Fig. 5.1. Noise and interference in an RF signal. The original data are modulated on the "in-channel" bandwidth (here shown single-side band SSB) represented by black arrows. Nonlinearity in RF modules can cause "in-band" intermodulation represented by red arrows. Flicker phase noise from LO generation and thermal noise floor are also shown in the blue envelope.

Nonlinearity in modules are sometime "useful", as it is the base for data modulation and demodulation. However, nonlinearity can also give rise to "spurious" intermodulation (IM) that distorts the original baseband. The good news is: nonlinearity is power related. When the input signal power of a functional module is very small, all modules behave "linearly" (all except the direct cross product term in the

Gilbert-cell multiplication, which will multiply small signals accordingly). We need to know at what power level nonlinearity has to be considered for its IM effects in signal processing. This is the idea behind the **"intercept point" (IP)**, an extrapolation point when IM is as large as the fundamental components. When we know how many dBs the input signal is below IP, we can estimate how large IM is.

We will use a two-tone signal to illustrate IM by nonlinearity, although we remember that RF data signals have many spectral components and noises.

$$v_{in} = A\cos\varphi_a + B\cos\varphi_b \qquad (5.1)$$

where φ_a and φ_b correspond to the frequency components in f_a and f_b by:

$$\varphi_a = \omega_a t + \theta_a; \qquad \varphi_b = \omega_b t + \theta_b \qquad (5.2)$$

For a functional module, we will approximate its voltage transfer functions (only S_{21} or T_{11}, and we will ignore other reflection components for now) by a Taylor series:

$$v_{out} = a_0 + a_1 v_{in} + a_2 v_{in}^2 + a_3 v_{in}^3 + ... \qquad (5.3)$$

We show a typical voltage transfer curve (VTC) in Fig. 5.2. We will treat all a_i as **real**, lumping all phase/frequency shift to treatment before or after the nonlinearity.

- The first constant term a_0 can be dealt with by a level shifter or taking v_{out} in a differential manner.
- The second term $a_1 v_{in}$ is the transducing gain term of S_{21} or T_{11}. We assume that this is the desirable linear function of the module, which will not change the spectral shape of the mix of tones within the channel.
- The third term $a_2 v_{in}^2$ and fourth term $a_3 v_{in}^3$ and thereafter will generate the harmonics of the single tone (which can often be filtered out readily) and intermodulation (IM) of the multi-tone signal.
- If VTC is an odd function (as shown in Fig. 5.2), then all even-order terms of a_0, a_2, a_4, … will be zero.
- A diode VTC is classical example when all a's can be non-zero (except a_0 = 0 if VTC passes through the origin). The even-order term can generate a DC offset regardlessly.

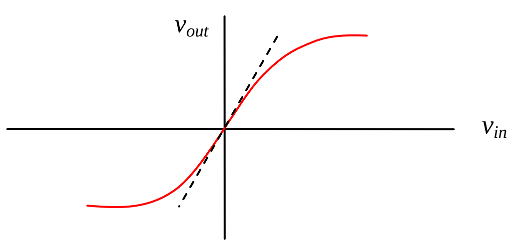

Fig. 5.2. A generalized voltage transfer curve (VTC) that can be approximated with a Taylor polynomial function (red solid curve). The linear approximation is shown in a black dashed line. Here a_0 = 0, and we can also observe that nonlinearity only affect v_{out} when v_{in} is large. The VTC here is a special case of "odd function" symmetry, where all even-order terms of VTC will be zero.

5.2 2^{nd} order terms in the two-tone signal treatment and intermodulation (IM)

We can see the quadratic term would expand to the 2^{nd} harmonics and IM terms of the two-tone signal as:

$$
\begin{aligned}
v_{out,2nd} &= a_2 v_{in}^2 = a_2\left(A\cos\varphi_a + B\cos\varphi_b\right)^2 \\
&= a_2\left(A^2\cos^2\varphi_a + 2AB\cos\varphi_a\cos\varphi_b + B^2\cos^2\varphi_b\right) \\
&= a_2\left\{\frac{A^2}{2}\left[1+\cos 2\varphi_a\right] + AB\left[\cos(\varphi_a-\varphi_b)+\cos(\varphi_a+\varphi_b)\right] + \frac{B^2}{2}\left[1+\cos 2\varphi_b\right]\right\} \\
&= a_2\left\{\underbrace{\frac{A^2+B^2}{2}}_{DC} + \underbrace{\frac{A^2}{2}\cos 2\varphi_a + \frac{B^2}{2}\cos 2\varphi_b}_{H2} + \underbrace{AB\left[\cos(\varphi_a-\varphi_b)+\cos(\varphi_a+\varphi_b)\right]}_{IM2}\right\}
\end{aligned}
\tag{5.4}
$$

We can see how the quadratic nonlinearity generates the additional DC, 2^{nd} harmonic (H2) and intermodulation (IM2) from Eq. (5.4). Notice that if φ_a and φ_b defines the data channel bandwidth, no H2 or IM2 will be in the channel, as shown in Fig. 5.3

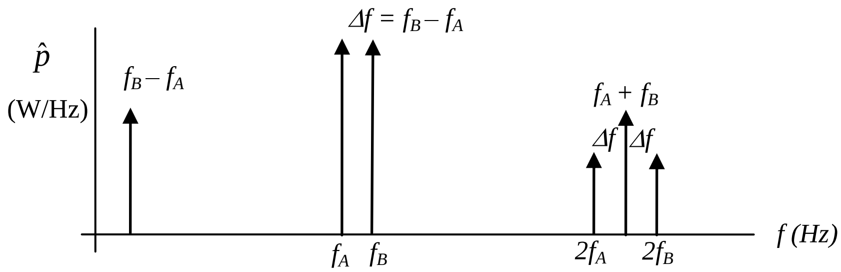

Fig. 5.3. Frequency components of 2^{nd} harmonic (H2) and intermodulation (IM2) generated by the quadratic term in the two-tone signal. Notice that if $A = B$, the H2 terms are 2 times smaller in voltage (–6dB) than the IM2 terms.

RF engineers would like to denote the nonlinearity NOT by a_2, but by intercept points (IP), as this notation offers a shorthand calculation in dB for an upper bound, as illustrated by Fig. 5.4. Here we plot v_{in} and v_{out} in dB, i.e., log-log plots. The linear a_1 term from Eq. (5.3) will be:

$$
\begin{aligned}
&20\log_{10}(a_1 v_{in}) = 20\log_{10}(a_1) + 20\log_{10}(v_{in}) = 20\log_{10}(v_{out,linear}) \\
&10\log_{10}\left(\frac{a_1^2 v_{in}^2}{2R}\right) = 20\log_{10}(a_1) + 10\log_{10}(p_{in}) = 10\log_{10}(p_{out,linear})
\end{aligned}
\tag{5.5}
$$

That is, the linear response will be a straight line of slope 1. The transducing gain is a_1 in dB.

The quadratic term from Eq. (5.3) will be:

$$\begin{aligned} 20\log_{10}\left(a_2 v_{in}^2\right) &= 20\log_{10}(a_2) + 40\log_{10}(v_{in}) = 20\log_{10}\left(v_{out,quadratic}\right) \\ 10\log_{10}\left(\frac{a_2^2 v_{in}^4}{(2R)^2}\right) &= 20\log_{10}(a_2) + 20\log_{10}(p_{in}) = 10\log_{10}\left(p_{out,quadratic}\right) \end{aligned} \tag{5.6}$$

That is, the quadratic response will be a straight line of slope 2 in voltage or in power. Notice that the total response of v_{out} will remain curving in the log-log plot when ALL terms in the Taylor series are considered. It is the "individual Taylor terms" that are the straight lines with various slopes. Now we define IP_{H2} as the intercept point where the second harmonic has the same voltage (or power) as the fundamental, as shown in Fig. 5.4.

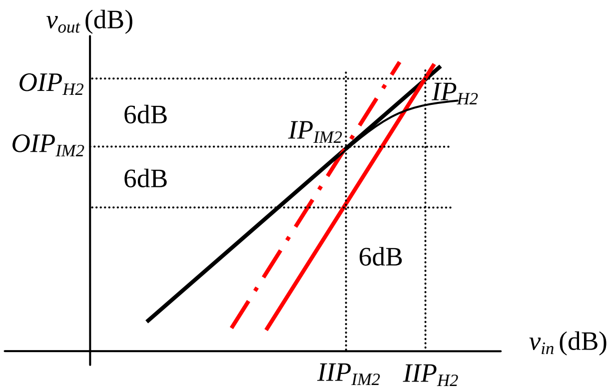

Fig. 5.4. Use of intercept points to represent nonlinearity from the quadratic term.

Notice that IP_{H2} denotes the point on tht VTC where $p_{out1} = p_{outH2}$, and then the fundamental frequency decreases with slope 1 with v_{in} (i.e., p_{out1} decreases 10dB with every 10dB of p_{in}) and the 2nd harmonic term decreases with slope 2 with v_{in} (i.e., p_{outH2} and p_{outIM2} decreases 20dB with every 10dB of p_{in}). From Eq. (5.4), we also notice that IM2 voltage is 2 times larger than H2 voltage when $A = B$, i.e., $p_{outIM2} = 4 \times p_{outH2}$ or p_{outIM2} is 6dB above p_{outH2}. IIP is defined by the intercept of the fundamental frequency, and the H2 and IM2 terms. Accordingly, IIP_{IM2} will be 6dB lower than IIP_{H2}, as shown in Fig. 5.4.

Let's use a realistic amplifier to illustrate the shorthand calculation of 2nd harmonic by IP. Assume the amplifier has a gain of 20dB and IIP_{H2} at –5dBm (by definition OIP_{H2} = 15dBm as the intercept is on the fundamental line as well). We can estimate the 2nd harmonic terms readily. For an input at –30dBm, Output of the fundamental will be at –10dBm. As input is 25dB lower than the IIP_{H2}, p_{outH2} will be at a further 25 dB lower than p_{out1} at –35dBm, i.e., p_{in} = –30dBm; p_{out1} = –10dBm; p_{outH2} = –35dBm.

For the same gain of 20dB and IIP_{H2} = –5dBm, if p_{in} = –50dBm, then p_{out1} = –30dBm, and p_{outH2} = –75dBm.

In the example above, we will also know IIP_{IM2} = –11dBm. For p_{in} = –50dBm, p_{outIM2} = –69dBm.

We can also write the algebraic equations as:

$$p_{outH2} = \frac{p_{out1}^2}{OIP_{H2}}; \qquad p_{outIM2} = \frac{p_{out1}^2}{OIP_{IM2}} \tag{5.7}$$

Use the above example, at p_{in} = –30dBm; p_{outH2} = –35dBm = 2×(–10dBm) – 15dBm. P_{outIM2} = –29dBm = 2×(–10dBm) – 9dBm. Check!

Or we can write the relation at the intercept point IP_{IM2} (in real number, not dB) as:

$$A_{OIPIM\,2} = |a_1| A_{IIPIM\,2} = |a_2| A_{IIPIM\,2}^2 \qquad (5.8)$$

$$A_{IIPIM\,2} = \left|\frac{a_1}{a_2}\right| \qquad (5.9)$$

We can see the definition of IP is nothing but a shorthand to denote the nonlinearity. We can also see when the module has larger nonlinearity (larger $|a_2|$), then IIP_{IM2} and IIP_{H2} will be lower, i.e., the output will see nonlinearity at a lower input level.

A last word on 2nd order nonlinearity. When the DC offset is important for direct conversion (i.e., the baseband is DC), the DC term in Eq. (5.4) will be critical. The DC term magnitude is the same as H2, and should be denoted accordingly.

5.3 3rd order terms in the two-tone signal treatment and intermodulation (IM)

Now we will look at the third-order nonlinearity. This is often more important than the second-order nonlinearity in RF transmitters for two reasons:

1. Many transmitters build up their signals in a differential way, and all even-order terms are rejected by the common-mode rejection ratio (CMRR).

2. The third-order intermodulation has terms "in band", i.e, very close to the fundamental frequency, and cannot be filtered without distorting the original spectrum within the channel. The second-order intermodulation has no term that is closed to the fundamentals.

We can repeat similar trigonometry function manipulation as in Eq. (5.4), and arrive at:

$$v_{out,3rd} = a_3 v_{in}^3 = \frac{a_3}{4}\left\{\begin{array}{c} \underbrace{\left(3A^3 + 6AB^2\right)\cos\varphi_a + \left(3B^3 + 6A^2B\right)\cos\varphi_b}_{Addition\ to\ fundamental} \\ + \underbrace{3\left[A^2B\cos\left(2\varphi_a - \varphi_b\right) + AB^2\cos\left(2\varphi_b - \varphi_a\right)\right]}_{In-band\ IM\ to\ fundamental} \\ + \underbrace{3\left[A^2B\cos\left(2\varphi_a + \varphi_b\right) + AB^2\cos\left(2\varphi_b + \varphi_a\right)\right]}_{IM\ around\ 3rd\ harmonic} \\ + \underbrace{A^3\cos 3\varphi_a + B^3\cos 3\varphi_b}_{3rd\ harmonic} \end{array}\right\} \qquad (5.10)$$

We can notice that Eq. (5.10) gives 8 frequency components, one group around the original band, and the other around the third harmonics, as shown in Fig. 5.5.

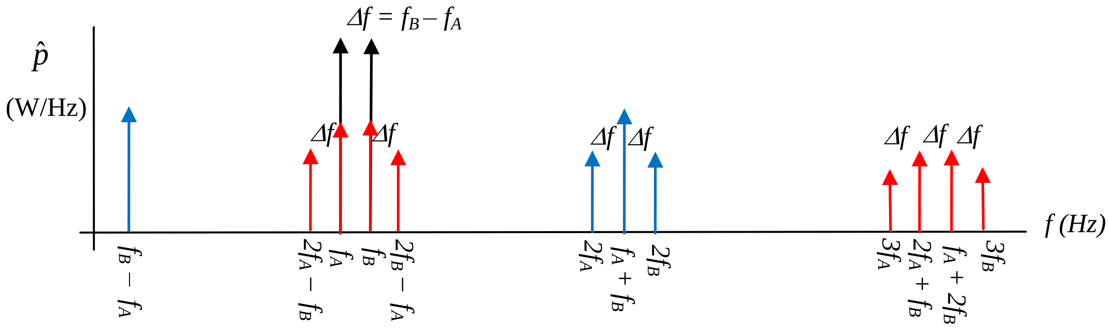

Fig. 5.5. Frequency components of 2nd (blue) and 3rd (red) order nonlinearity in the two-tone signal. Notice that we have qualitatively expressed the relative magnitude for the 8 terms in IM3 and H3.

We can construct the similar input and output intercept points (IIP and OIP) for 3rd harmonics as shown in Fig. 5.6. Now the 3rd harmonics have slope of 3, i.e., p_{outH3} and p_{outIM3} decreases 30dB with every 10dB of p_{in}. We also notice that if $A = B$, the IM terms in $v_{out,3rd}$ is 3 times larger than the H3 terms, or p_{outIM3} is $20\times\log_{10}3 = 9.54$dB higher than p_{outH3}. From the different slope in the log-log (or dB-dB) plot, if p_{out1} is lower than OIP_{H3} by X dB, p_{outH3} will be lower than p_{out1} by $2X$ dB (i.e., p_{out1} with slope 1 and p_{outH3} with slope 3). Therefore, IIP_{IM3} will be just 9.54dB/2 = 4.77dB lower than IIP_{H3}. This is graphically represented in Fig. 5.6.

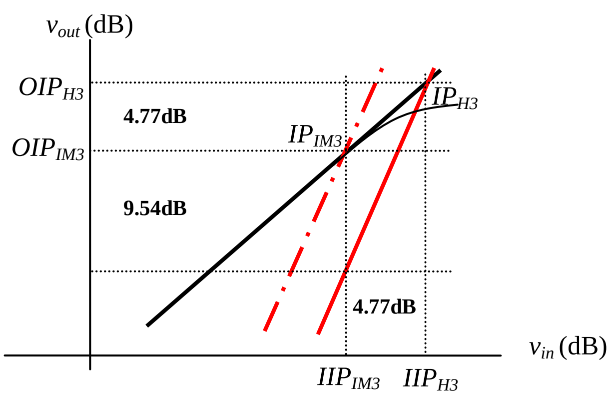

Fig. 5.6. Use of intercept points of IIP_{IM3} and IIP_{H3} to represent nonlinearity from the cubic term.

We can also work out the algebraic equations for the relation between the third harmonic terms and the fundamental term, similar to Eqs. (5.7 – 5.9).

$$p_{outH3} = \frac{p_{out1}^3}{OIP_{H3}^2} \quad \text{or} \quad \frac{p_{outH3}}{p_{out1}} = \frac{p_{out1}^2}{OIP_{H3}^2} \qquad (5.11)$$

$$p_{outIM3}(2f_a - f_b) = \frac{p_{out1a}^2 p_{out1b}}{OIP_{IM3}^2} \quad \text{or} \quad \frac{p_{outIM3}(2f_a - f_b)}{p_{out1a}} = \frac{p_{out1a} p_{out1b}}{OIP_{IM3}^2} \qquad (5.12)$$

We can also give the voltage amplitude at the IM3 intercept point (IP_{IM3}):

$$A_{OIPIM\,3} = \underbrace{|a_1| A_{IIPIM\,3}}_{\text{point on fundamental}} = \underbrace{\frac{3}{4}|a_3| A_{IIPIM\,3}^3}_{\text{point on } IM3} \qquad (5.13)$$

$$A_{IIPIM\,3}^2 = \frac{4}{3}\left|\frac{a_1}{a_3}\right| \qquad (5.14)$$

Of no surprise, IIP_{IM3} and IIP_{H3} are nothing more than short hands to denote the cubic nonlinearity with easier dB estimation!

Among all of the harmonic and intermodulation terms, terms at $2f_a - f_b$ and $2f_b - f_a$ as 3rd-order intermodulation are by far the MOST important, as they are very close to f_a and f_b. In a given channel, the spread spectrum caused by data modulation and phase noise as shown in Fig. 5.1, IM3 can distort that spread spectrum and cause data errors and non-retrievable waveforms! As mentioned before, even-order terms can be smaller by differential signaling (i.e., a_0 and a_2 are much smaller than a_1 and a_3), and IM3 is also larger than H3 by 9.54dB. Due to its importance and dominance, if you are given a module with the specification of "*IIP3*", it is implicitly assumed *IIP3* means IIP_{IM3}. Remember that for all terms in H2, IM2, H3 and IM3, only part of IM3 will make an "in-band" contribution to v_{out}.

Before we leave the topic, we have an important term to discuss.

$$v_{out,3rd}(f_a, f_b) = \frac{a_3}{4}\left[\left(3A^3 + 6AB^2\right)\cos\varphi_a + \left(3B^3 + 6A^2B\right)\cos\varphi_b\right] \qquad (5.15)$$

We can see that the cubic nonlinearity changes the gain of the module!!! For example, in the case where $A << B$ (i.e., in the two tone signal, one is much weaker than the other), the voltage amplitude at f_a is:

$$Amplitude(f_a) = a_1 A + \frac{a_3}{4}\left(3A^3 + 6AB^2\right) \cong A\left(a_1 + \frac{3a_3}{2}B^2\right) \qquad (5.16)$$

The gain factor at f_a after the module does not ONLY depend on the module parameters a_1 and a_3, but also the stronger signal strength B at f_b!!!! When a_1 and a_3 are of opposite signs (most often for nonlinearity due to gain saturation), that means the gain at f_a will decrease from its linear gain of a_1 with a strong signal at f_b present! This is called "**desensitization**", and is very important for the consideration of anti-jamming when the RF channels are saturated at a particular frequency (not band, as it will be more difficult) on purpose by hackers or for privacy. Also, if amplitude modulation is used, we can see the amplitude modulation at f_a signals will be affected by signals at f_b. This is called "**cross modulation**", and is an important design factor for systems using AM or the in-phase *I* component in any q-naries modulation.

One of the most straightforward ways to mitigate desensitization and cross modulation is indeed to improve the linearity of the RF modules, or make $|a_3| << |a_1|$. For BJT or MOSFET in subthreshold, when the transfer gain is nearly exponential, $|a_3| \cong |a_1|$ and is quite large. Special care will be needed to treat desensitization and cross modulation. For MOSFET in above-threshold saturation, the cubic nonlinearity in a_3 is on the other hand very small, and hence a better choice when desensitization or cross modulation is significant.

5.4 1-dB compression point when only IM3 is important

As mentioned before, IM3 is most important due to its in-band nature. However, because it is "in-band", it is difficult to measure directly as IM3 will "hide behind" the phase noise of the LO of your receiver or your spectrum analyzer. H2 and H3 are much easier to measure, and we can "theoretically derive" IM2 and IM3 from H2 and H3 by

$$IIP_{IM3} = IIP_{H3} - 4.77\text{dB} \qquad \text{and} \qquad IIP_{IM2} = IIP_{H2} - 6.0\text{dB}. \qquad (5.17)$$

This is unfortunately not a direct measurement, and hence not preferred by RF engineers, whose motto is "*Measure twice; cut once*" or "*Measure the component at hand before use even if the component specification is given*". Hence, another popular way to denote the nonlinearity is by the 1dB compression point, as shown in Fig. 5.7. We can see that as long as the output can be measured as a function of input, the 1-dB compression point ($I_{1dBcomp}$) can be directly extracted from the measurement of the module under test.

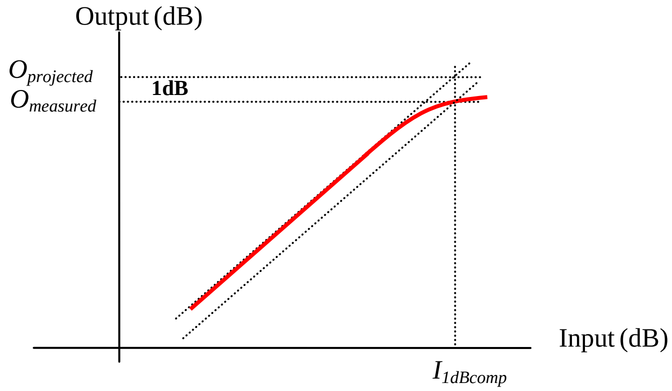

Fig. 5.7. The input level $I_{1dBcomp}$ that gives a measured output that is 1dB lower (in voltage or in power) than the linear projection from low input power.

If we *assume* that the nonlinearity is ONLY caused by IM3 (this can be true if the VTC is a narrow-band plot with only in-band signals), we can then relate IIP_{IM3} and $I_{1dBcomp}$ by the following derivation on the voltage amplitude $A_{1dBcomp}$ (we have assumed $a_3 < 0$ here):

$$\frac{A_{measured}}{A_{projected}} = \frac{|a_1|A_{1dBcomp} - \frac{3}{4}|a_3|A_{1dBcomp}^3}{|a_1|A_{1dBcomp}} = 1 - \frac{3}{4}\left|\frac{a_3}{a_1}\right|A_{1dBcomp}^2 = -1dB = 10^{-1/20} = 0.89125 \quad (5.18)$$

From Eq. (5.14), we can then write

$$1 - 0.89125 = \frac{A_{1dBcomp}^2}{A_{IIPIM3}^2} = \frac{I_{1dBcomp}}{IIP_{IM3}} = 0.10875 = -9.64dB \qquad (5.19)$$

Notice that Eq. (5.19) is correct in both power and voltage. Finally we can write **in dB**:

$$I_{1dBcomp} = IIP_{IM3} - 9.64\text{dB} \qquad (5.20)$$

$$O_{1dBcomp} = IIP_{IM3} + gain - 1\text{dB} - 9.64\text{dB} = OIP_{IM3} - 10.64\text{dB} \qquad (5.21)$$

where $O_{1dBcomp}$ is the measured output power that is 1dB lower than the projected linear-gain output power.

For a module with nonlinearity, most often it is given by *IIP3* (implicitly assumed as IIP_{IM3}) or $I_{1dBcomp}$ (implicitly assumed IM3 is dominant). Read the fine prints in the specification sheet carefully, and compare with your module under test.

5.5 Nonlinearity and intercept point in a two-module cascade

We will treat the nonlinearity of a two-module cascade and derive some intuition under the simplest case of IM3 adding randomly or coherently, as shown in Fig. 5.8. For multi-modules or specific consideration of the phase in IM3 terms, we will leave that to the detailed treatment in SIMULINK.

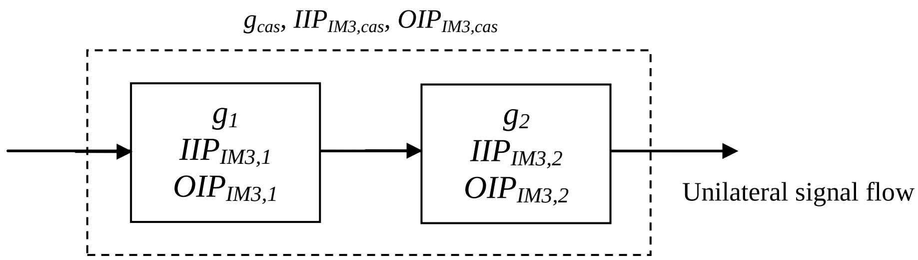

Fig. 5.8. Cascade of two nonlinear modules with respective gain, IIP_{IM3} and OIP_{IM3}.

If we assume that all IM3 terms of Module 1 and Module 2 can add up "randomly" (all phases of IM3 terms are random and uncorrelated), i.e., the combined IM3 terms can be treated with the ideal unilateral signal gain model and the IM3 power is additive (not voltage amplitude), we can write the power of the IM3 term at $2f_a - f_b$ as:

$$\begin{aligned} p_{outIM3,cas} &\cong g_2 \cdot p_{outIM3,1} + p_{outIM3,2} = g_2 \frac{p_{out1a,1}^2 \cdot p_{out1b,1}}{OIP_{IM3,1}^2} + \frac{p_{out1a,2}^2 \cdot p_{out1b,2}}{OIP_{IM3,2}^2} \\ &= \frac{\left(g_2 p_{out1a,1}\right)^2 \cdot \left(g_2 p_{out1b,1}\right)}{\left(g_2 OIP_{IM3,1}\right)^2} + \frac{p_{out1a,2}^2 \cdot p_{out1b,2}}{OIP_{IM3,2}^2} \\ &= p_{out1a,cas}^2 \cdot p_{out1b,cas}\left[\frac{1}{\left(g_2 OIP_{IM3,1}\right)^2} + \frac{1}{OIP_{IM3,2}^2}\right] \end{aligned} \qquad (5.22)$$

where we know $g_2 p_{out1a,1} = p_{out1a,2} = p_{out1a,cas}$. We conclude that when the IM3 power adds up (random phase), the cascade OIP_{IM3} (in real number, not dB) can be expressed as:

$$\frac{1}{OIP_{IM3,cas}^2} = \frac{1}{\left(g_2 \cdot OIP_{IM3,1}\right)^2} + \frac{1}{OIP_{IM3,2}^2} \qquad (5.23) \text{ (adding randomly)}$$

This result is intuitive, as the module with higher nonlinearity (smaller OIP_{IM3}) should dominate the $OIP_{IM3,cas}$. If all IM3 terms add up *coherently*, we have to use the voltage amplitude instead of the power, which will give the cascade IM3 output power at $2f_a - f_b$ as (the derivation is similar):

$$\sqrt{p_{outIM3,cas}} \cong \sqrt{g_2} \cdot \sqrt{p_{outIM3,1}} + \sqrt{p_{outIM3,2}} = \sqrt{g_2}\, \frac{p_{out1a,1} \cdot \sqrt{p_{out1b,1}}}{OIP_{IM3,1}} + \frac{p_{out1a,2} \cdot \sqrt{p_{out1b,2}}}{OIP_{IM3,2}}$$
$$= p_{out1a,cas} \cdot \sqrt{p_{out1b,cas}} \left[\frac{1}{g_2 OIP_{IM3,1}} + \frac{1}{OIP_{IM3,2}} \right] \tag{5.24}$$

or equivalently:

$$\frac{1}{OIP_{IM3,cas}} = \frac{1}{g_2 \cdot OIP_{IM3,1}} + \frac{1}{OIP_{IM3,2}}$$ (5.25) (adding coherently)

Example: For a power amplifier with g_1 = 20dB and $OIP_{IM3,1}$=40dBm cascading with a filter with g_2 = 0dB and $OIP_{IM3,2}$ = 40dBm, if we will use Eq. (5.23) for IM3 terms with random phase,

PA-filter cascade will have:

$$\frac{1}{OIP^2_{IM3,cas}} = \frac{1}{(g_2 \cdot OIP_{IM3,1})^2} + \frac{1}{OIP^2_{IM3,2}} = \frac{1}{(1 \cdot 10^{40/10})^2} + \frac{1}{(10^{40/10})^2} = \frac{1}{\left(\frac{1}{\sqrt{2}} 10^4\right)^2}$$, i.e., $OIP_{IM3,cas}$ will be 38.5dBm.

Filter-PA cascade will have:

$$\frac{1}{OIP^2_{IM3,cas}} = \frac{1}{(g_1 \cdot OIP_{IM3,2})^2} + \frac{1}{OIP^2_{IM3,1}} = \frac{1}{(100 \cdot 10^{40/10})^2} + \frac{1}{(10^{40/10})^2} \cong \frac{1}{(10^4)^2}$$, i.e., $OIP_{IM3,cas}$ will be 40 dBm.

The result makes sense. The PA-filter cascade will have higher nonlinearity than filter-PA cascade.

If we use the coherent phase in Eq. (5.25) as a worst-case nonlinear cascade, the PA-filter cascade will have $OIP_{IM3,cas}$ = 37dBm, 3dB smaller than 40dBm.

Before we close the topic on cascade, we will survey what RF modules have nonlinearity concerns, and IIP3 needs to be defined. As nonlinearity is proportional to the Taylor series by v_{in}, harmonic and intermodulation generation is only serious when v_{in} is large (most often at the end of the RF transmitter). However, once harmonic and intermodulation are generated, the effect can be felt all over the receiver system as well, as the main function of the receiver is to retrieve the information in the transmitter before going through the nonlinear distortion and harmonic generation. The first element that needs serious nonlinear treatment for IM and harmonic generation is the power amplifier, almost by its name suggests. The second possibility is filters and circulators, which are often passive in construction. Air-coil inductors and air-gap capacitors have little nonlinearity, but both have small values. To fit the desired frequency of $f = 1/2\pi\sqrt{LC}$, ferromagnetic and high-k materials are needed in high-value inductors and capacitors, respectively, which introduce nonlinearity in the permeability and permittivity.