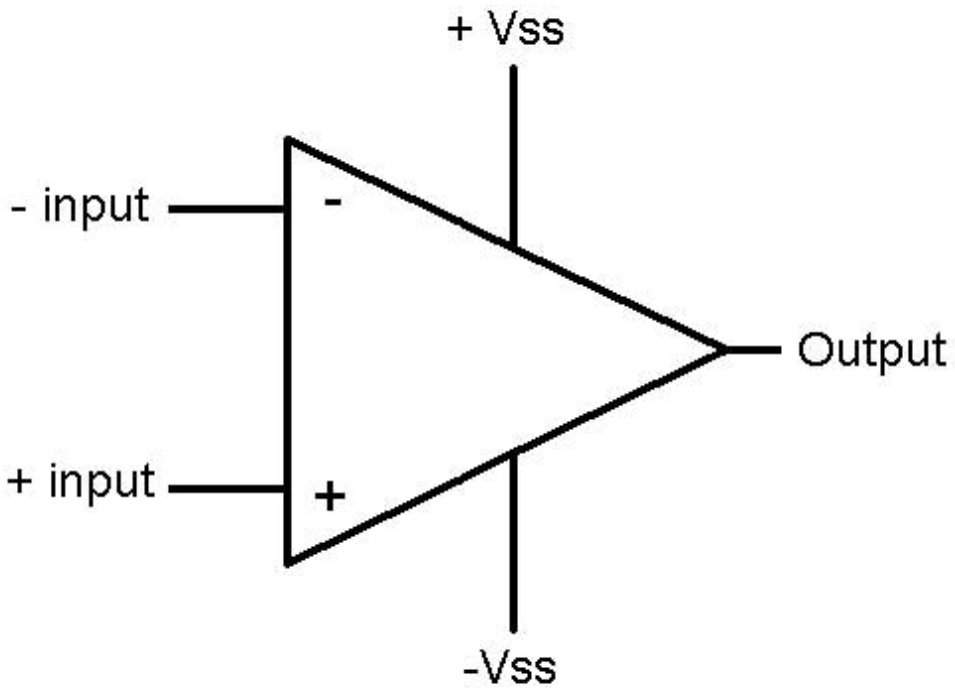
Notes on Electronics

A.Y. 2017/2018



Contents

Course introduction

These notes comes from the course "Electronics" taught by Prof. A. Spinelli at Politecnico di Milano in the second semester of the academical year 2017-2018. The goal of the course is to give the student some concepts about electronics that are useful for an engineering physicist. In fact, electronics influences physics just as physics is important for electronics. Further developing this concept, we know that there is a branch of physics that deals with electron devices, that are the basic elements of electronics. On the other hand, electronics develops the so called enabling technologies that are fundamental for the development of physics (e.g., Data Acquisition Systems). An example of this close relationship between physics and electronics can be found in the number of Nobel prize in physics that profoundly influenced also electronics (the discovery of the Giant Magnetoresistance, the invention of the CCD, experiments about graphene and the invention of blue LEDs). In particular, the most important electron device developed is the transistor (1947), that set a revolution in electronics, leading to the vast number of applications that we have nowadays. Analysing then the development of this devices, we can observe that they tend to become smaller and smaller every year, increasing the number of transistors that we can have in a certain device and the frequency at which they can work. This leads to an increase of the performances and a decrease of the scaling of these devices that are described by the Moore Law and, as a side effect, bring to question: when will we hit the scaling limit? This is actually a debated problem in science and a few promising alternatives are present, even though they seem to be still far from replacing the silicon-based technology.
However, electron devices are not the topic of this course; we will focus on the enabling technologies. In fact, sensors, instrumentation and data acquisition are important topics in every branch of science, not just in physics. The key feature of these technologies are data acquisition systems and signal processing, therefore it is important for an engineer to be able to correctly size and design every acquisition chain, dealing with these problems. The problem of signal conditioning is that, in applications, we need to deal with signals that are non-linear, high or low amplitude, noisy and analog and obtain from them linear signals, attenuated or amplified, noise free and digital. To do this, we need to design an interface circuitry, in which we have an amplification stage, a condition circuitry (in which the noise is reduced) and, finally, an analog to digital converter.
The course will therefore consist of two topics:

- Operation Amplifier circuits, in which we will deal with feedback, impedances, parameters, linear applications of OAs and their frequency response, stability and compensation;

- signal recovery from noise, in which we will deal with sensors, noise, random process and noise filtering techniques.

Dealing with Operation Amplifiers, the goal is to be able to analyse and design simple circuits, with an emphasis on their applications to the problem of data acquisition. Moreover, we will investigate the stability issue of these devices and the real parameters that can make devices different from what we theoretically designed. The sensors we will briefly introduce are thermocouples, thermistors and strain gages, while when dealing with noise we will investigate mainly the white noise and the $1/f$ noise.
At the end of this course, the student should be able to analyse and design simple circuits with operation amplifiers, understand simple problems regarding data acquisition and involving sensors, preamplification and noise filtering and, finally, he or she must be able to adopt an engineering approach to problems.
Lessons will be devoted to theory and basic concepts, while numerical example and exercises will be investigate during drills sessions. The teacher is at students' disposal for appointments, that can be asked by sending an email. Lecture slides, past exams' solutions, drills, recommended books and extra material can be found at `home.deib.polimi.it/spinelli`. A few prerequisites about linear networks, Fourier and Laplace transforms and linear systems are needed. The final examination is written and, by the end of the course, students will have to decide if it will be a regular exam or an open-book exam. The regular exam will consist in two exercises, each one made up by four questions and about the two main topics (OA and noise and filtering and signal conditioning), and a theoretical question; it will be 3 hours long. The open-book exam will consist in just two questions (as before) to be solved in 2 hours and 45 minutes. A 30/30 mark is equivalent to correctly solving (it means that also numerical results are correct, not only the method used) 75% of the exam. Exams will be generally scheduled at 13.00.
Finally, a few suggestions. The exam is about solving problems, not just describing how this can be done, and it is important to show your understanding of the subject. Therefore, the hints are to pose a lot of questions to the instructor, to first make sure to have understood theory before moving to exercises and to remember that learning is a long and slow process. It is important to be conscious of your own preparation, so that also the instructor can more easily help you, and make sure to have understood the theory first, otherwise exercises would be impossible.

Chapter 1

Operation amplifier

1.1 Equivalent circuits

Before starting, it is important to review some basic concepts that will be used in the following part of the course; in particular, we will concentrate on equivalent circuits.
The voltage equivalent circuit, also called Thévenin equivalent circuit, was first formulated by H. von Helmholtz[1] in 1853 and was then rediscovered in 1883 by L. C. Thévenin, that named it. On the other hand, the current equivalent circuit, also called the Norton equivalent circuit, was independently discovered in 1926 by H. F. Mayer and E. L. Norton. The fundamentals of electronics, therefore, dates back to the 19th century.
The subject of these two theorems are linear networks, that are defined as networks (in other words, circuits) that are made only by elements that can be described using linear differential equations. In out applications, we will deal with resistors R, inductors L and capacitors C, that in general will be assumed constant, while source terms will be represented by voltage V sources and current I sources, that will be assumed constant or linearly dependent on quantities measured over other components (e.g. a voltage source controlled in voltage will produce a voltage that will be proportional to the one measured over, for example, a certain resistor). We can then say that every linear network observed by any pair of terminals will behaves as if it were composed only by a source element and an impedance. This leads to the following two theorems:

- Thévenin equivalent circuit: the circuit equivalent to any linear network can be represented as a voltage source in series with an impedance;
- Norton equivalent circuit: the circuit equivalent to any linear network can be represented as a current source in parallel to an impedance.

It is important to remember that the equivalence is only from the viewpoint of the external load. The power dissipation in the network, having a quadratic dependence on the elements of the network, will not be equal. Only the voltage-current ($V - I$) characteristic of the network will be identical.

[1] This important physicist is also responsible for writing the first wave equation and discovering the superposition principle.

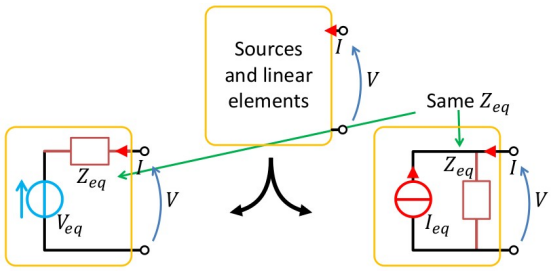


Figure 1.1: Equivalent circuits for the Thévenin and the Norton circuits.

The value of the voltage V_{eq} applied by the voltage source in the Thévenin equivalent circuit is the open-circuit voltage at the terminals of the linear network, while the current I_{eq} imposed by the current generator in the Norton equivalent circuit is the short-circuit current passing through the terminals of the linear network[2]. On the other hand, the equivalent impedance in both circuits can be found or as the ratio between the equivalent voltage and the equivalent current:

$$Z_{eq} = \frac{V_{eq}}{I_{eq}}$$

or shutting down[3] every source in the linear network and calculating the equivalent impedance from scratch. Obviously, to exploit this second definition we need to know the structure of the linear network; it is not possible to use a black-box approach.
Equivalent circuits will then be fundamental for understanding amplifiers.

1.2 Voltage and current amplifiers and related impedances

To start studying amplifiers, we can adopt a black box approach, not considering what is inside an amplifier (thus thinking it as a black box) but studying its equivalent circuit with respect to the input pins and the output pins. Depending on the type of input (voltage or current) and on the type of output (again, voltage or current), it is possible to identify four different types of amplifiers:

- voltage amplifiers: voltage as an input, voltage as an output;
- current amplifiers: current as an input, current as an output;
- transconductance amplifiers: voltage as an input, current as an output;
- transresistance amplifiers: current as an input, voltage as an output.

[2]This phrase states an important result that have to be stressed: to measure a voltage between two pins, we need to open the circuit that connects them, while to measure a current between two pins we need to short-circuit them.

[3]By shutting down every source, we mean that every voltage source is replaced by a short circuit ($\Delta V = 0$) and every current source is replaced by an open circuit ($I = 0$).

We will mainly study the first two of them, leaving the analysis of the second two to the student. Both will be one-directional amplifiers, not allowing a reverse transfer of the signal from the output to the input, and, in them, for the sake of simplicity we will consider only resistors, even though the whole reasoning can be made more general by considering complex impedances.

1.2.1 Voltage amplifier

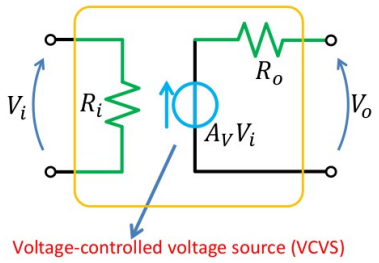


Figure 1.2: Equivalent circuit of a voltage amplifier.

A voltage amplifier, by definition, takes a voltage V_i as an input and gives a voltage V_o as an output. Due to this characteristic, we can consider what is inside the amplifier using the Thévenin equivalent circuit of the amplifier. On the input side, since we do not want any reverse transfer of signal, the voltage source of the equivalent circuit will be identically equal to zero (thus being a short-circuit), while the input impedance can be considered as an input resistance R_i. On the output side, we will have a generic output resistance R_o and, since we want the output voltage V_o to depend linearly[4] on the input voltage V_i, we will need a voltage-controlled voltage source (VCVS) that imposes a voltage difference equal to $A_V V_i$, where A_V is a certain, constant gain.

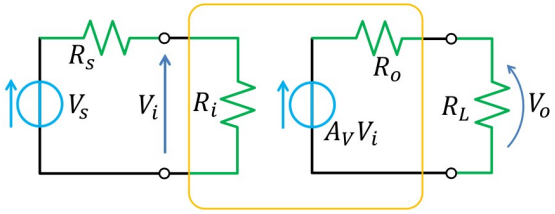


Figure 1.3: Voltage amplifier connected to a source and a load.

To study this amplifier, we can add a source (in particular, its equivalent circuit using Thévenin's theorem) and a load resistance. The source is made up by a voltage generator V_S and a series resistance R_S, while the load is represented by a resistance R_L. From the partition of the voltage, we can calculate the input voltage across the input resistance R_i:

$$V_i = V_S \frac{R_i}{R_i + R_S}$$

[4]We are dealing with equivalent circuits and the superposition principle must hold.

and again, considering the voltage-controlled voltage source to impose a voltage equal to $A_V V_i$, we can calculate the partition of the voltage on the load and on the output resistances, obtaining the following output voltage:

$$V_o = A_V V_i \frac{R_L}{R_o + R_L}.$$

We can define the total gain of the voltage amplifier as the ratio between the output voltage and the voltage imposed by the source:

$$\frac{V_o}{V_S} = A_V \frac{R_i}{R_i + R_S} \cdot \frac{R_L}{R_o + R_L}.$$

The first observation that we can make from this expression is that, since the two ratios $R_i/(R_i + R_S)$ and $R_L/(R_o + R_L)$ are always lower than one (since resistance are always positive quantities), then the total gain must always be lower than A_V, that represents the maximum, theoretical value of the gain. Moreover, we can immediately see that the gain depends both on the source series resistance R_S and on the oad resistance R_L, thus not depending only on the elements of the voltage amplifier but also on the type of source and load connected. This is an important drawback, since a change at the input source or at the output load will change the gain in an often uncontrolled way (since the values of these resistances are not always easy to predict). To avoid it, we can state the following requirement for having an ideal voltage amplifier:

- the input impedance must be very high, such that:

$$R_i \to \infty \quad \Rightarrow \quad \frac{R_i}{R_i + R_s} \simeq 1;$$

- the output impedance must be very low, such that:

$$R_o \simeq 0 \quad \Rightarrow \quad \frac{R_L}{R_o + R_L} \simeq 1.$$

These are fundamental criteria in the design of a good voltage amplifier and therefore, when dealing with an ideal voltage amplifier, we will replace the input impedance with an open circuit and the output impedance with a short-circuit.

1.2.2 Current amplifier

A current amplifier, by definition, takes a current I_i as an input and gives a current I_o as an output. Due to this characteristic, we can consider what is inside the amplifier using the Norton equivalent circuit of the amplifier. On the input side, since we do not want any reverse transfer of signal, the current source of the equivalent circuit will be identically equal to zero (thus being an open circuit), while the input impedance can be considered as an input resistance R_i. On the output side, we will have a generic output resistance R_o and, since we want the output current I_o to depend linearly on the input current I_i, we will need a current-controlled current source (CCCS) that imposes a current equal to $A_I I_i$, where A_I is a certain, constant gain.

Again, it is possible to connect this amplifier to a source and a load that, since we are dealing with currents, will be expressed using the Norton equivalent

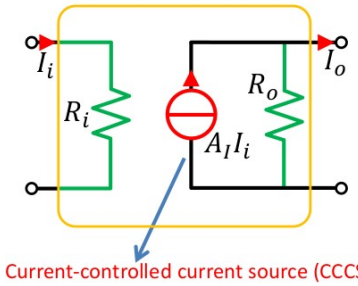


Figure 1.4: Equivalent circuit of a current amplifier.

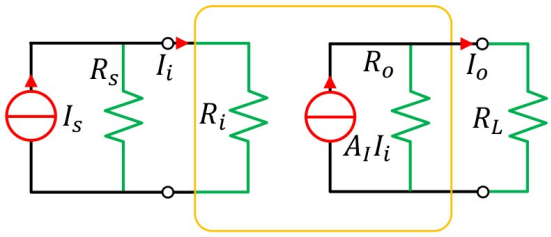


Figure 1.5: Current amplifier connected to a source and a load.

circuits. Studying the left-hand side of this circuit, we can observe that the input current, from a current partition, can be written as:

$$I_i = I_S \frac{R_S}{R_i + R_S}$$

and this current, amplified of a factor A_I, will be imposed by the current-controlled current source in the right-hand side. Therefore, we can calculate the output current I_o flowing through the load resistance by writing a current partition:

$$I_o = A_I I_i \frac{R_o}{R_o + R_L}.$$

Plugging the previous two expressions one into the other, we can obtain the expression of the total gain of the amplifier:

$$\frac{I_o}{I_S} = A_I \frac{R_o}{R_o + R_L} \frac{R_S}{R_i + R_S}.$$

Also in this case we can observe that since the ratios $R_o/(R_o+R_L)$ and $R_S/(R_i+R_S)$ are lower than one (since resistances are positive quantities), the total gain is lower than its maximum theoretical value, that is A_I, the gain of the current-controlled current source. Moreover, the gain depends on R_S and R_L in a significant way and, as discussed for the voltage amplifier, this is an important drawback, making the gain of the amplifier be different depending on the source and on the load connected to it. These comments, that are identical to one we have made for the voltage amplifier, lead in this case to different design criteria:

- the input impedance must be very low, such that:

$$R_i \simeq 0 \quad \Rightarrow \quad \frac{R_S}{R_S + R_i} \simeq 1;$$

Type	Input	Output	R_i	R_o
Voltage amp.	V	V	∞	0
Current amp.	I	I	0	∞
Transconductance amp.	V	I	∞	∞
Transresistance amp.	I	V	0	0

Table 1.1: Summary of the main types of amplifier.

- the output impedance must be very high, such that:

$$R_o \to \infty \quad \Rightarrow \quad \frac{R_o}{R_o + R_L} \simeq 1.$$

In an ideal current amplifier, therefore, we will replace the input impedance with a short-circuit (that will make the current I_S imposed by the source generator to be identical to the input current I_i) and the output impedance with an open circuit (thus making the whole current coming from the current-controlled current source to be the output current).

1.2.3 Summary

The detailed study of the transconductance amplifier and of the transresistance amplifier and of their input and output impedances in ideal cases is left to student. Doing it, it is important to remember that the source side will be different from the load's one, using a Thévenin equivalent circuit when dealing with voltages and a Norton equivalent circuit when dealing with currents. In these cases, since the gain is the ratio between the output signal and the input signal, it will have a certain dimension, while in the cases of the current and the voltage amplifiers it was non-dimensional. In the case of the transconductance amplifier, since the input is represented by voltage and the output by a current, it will have the dimension of a conductance (that is measured in Siemens, that are the reciprocal of Ohms). On the other hand, in a transresistance amplifier the input will be a current and the output a voltage, therefore the gain will have the dimension of a resistance (measured in Ohms).
The mainly characteristics of an ideal amplifier are summarized in Table 1.1.

1.3 Negative feedback and application to amplifier design

1.3.1 Historical introduction

After this brief introduction to amplifiers, we can go a little more into details about their design criteria, studying negative feedback. From an historical point of view, the problem whose solution was represented by negative feedback dates back to the first fourth of the previous century, when the American telephone lines built the first and the second transcontinental telephone lines, made up by a certain number of channels and amplifiers. After this initial success, a further increase in the number of channels was extremely challenging due to a

problem with amplifiers. From our previous courses on electromagnetism and waves, we know that an electromagnetic signal propagating through a medium (in this case, wires) will be attenuated by a number of effects and this makes necessary the regeneration of the signal along the line. This can be done using amplifiers, that at that time, before the invention of the transistor, consisted in vacuum-tubes with a certain gain. This gain, due to the design of these elements, was extremely dependent on a number of parameters (plate voltage, temperature, humidity, ambient conditions, ...), thus being unstable and not really linear devices. These non-linearities created intermodulation distortion in multi-channel systems, thus making difficult the improvement of this system.
As an example, we can consider a quadratic amplifier that receives as an input a sinusoidal signal at frequency ω:

$$\sin(\omega t).$$

This non-linear amplifier will give as an output the square of the input, therefore:

$$\sin^2(\omega t)$$

that will contain, among other terms, also the second harmonic of the signal, that will oscillate at frequency 2ω. In real amplifiers, that have much more complex non-linearities, this will lead to the presence of many more harmonics, making every channel interact with others and leading to a cross-talk between different communications.
This significant limitation was solved by H. S. Black with the idea of negative feedback. His goal, in fact, was to improve the stability and the linearity of the amplification chain. He realized that the output of a non-linear amplifier will contain some information about the non linearity present in the gain of the amplifier G_{ol} (we will soon understand why it is called like this) and that this information can be used to correct the input signal, compensating (up to a certain degree) non-linearities.

1.3.2 Theoretical explanation

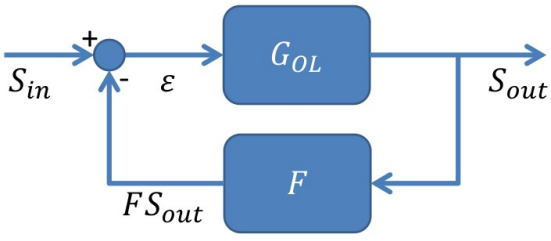


Figure 1.6: General schematics of a negative feedback system.

The general drawing of a negative feedback system is represented in Figure 1.6. In it, we can recognize a non-linear amplification stage whose gain is called G_{ol}, a feedback stage whose gain is called F and, in the left-hand side, a difference operator that subtracts, from the input signal, the signal coming through the feedback. This difference operator is fundamental for the effectiveness of the

negative feedback system.
The signal ϵ coming out from the difference operator is the error of the amplification. Calling S_{in} the input signal of the system and S_{out} the output signal, we can immediately see that it will be the difference (due to the presence of the difference operator) between the input signal and the output one multiplied by the gain feedback stage:

$$\epsilon = S_{in} - FS_{out}.$$

However, the output signal will be the product of the gain of the non-linear amplifier G_{ol} times the error signal:

$$S_{out} = G_{ol}\epsilon$$

and substituting the previous expression in this last one, we can write the overall gain G of the negative feedback system as the ratio between the output signal and the input one:

$$G = \frac{S_{out}}{S_{in}} = \frac{G_{ol}}{1 + G_{ol}F}.$$

We can now investigate this gain in two limiting cases, since the asymptotic behaviour is easier to discuss and every other case will be between these two. Considering the product at the denominator, we can first assume:

$$|G_{ol}F| \ll 1$$

and therefore the one will be the dominant term at the denominator:

$$G = \frac{G_{ol}}{1 + G_{ol}F} \simeq \frac{G_{ol}}{1} \simeq G_{ol}.$$

Therefore, in this case, we obtained the gain of the non-linear amplifier, that we can also call open-loop gain (from which the name G_{ol}: it is the gain that we have if we cut the feedback loop), just if there was not any feedback (since the loop is open).
The opposite assumption is to have a strong loop:

$$|G_{ol}F| \gg 1$$

thus obtaining:

$$G = \frac{G_{ol}}{1 + G_{ol}F} \simeq \frac{G_{ol}}{G_{ol}F} \simeq \frac{1}{F} = G_{id}.$$

This gain is called the ideal gain and, as we can see, it is independent from the open-loop gain, depending only on the feedback network. This is extremely important: in fact, the F block is not an amplifier, therefore it will be more easy to build an it will more linear and stable than the open-loop block G_{ol}.

1.3.3 Loop gain

An important parameter of the feedback system is represented by the loop gain. To calculate it, we need to impose the input signal equal to zero, to cut the loop (in an arbitrary point, since the result is independent from the cutting point) and to inject a test signal S_{test} inside the loop. In Figure 1.7 it is possible to

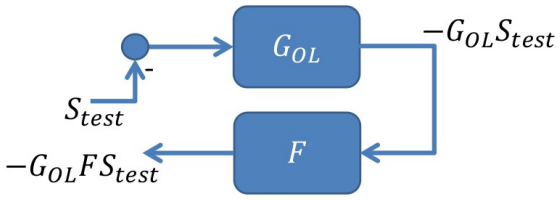


Figure 1.7: Calculation of the loop gain.

observe an example of this procedure. Doing this, we can obtain the expression of the gain of the loop:

$$G_{loop} = -G_{ol}F$$

that, therefore, is a measure of the strength of the feedback system. Since, from what we have said before, a good feedback system will have:

$$|G_{ol}F| \gg 1$$

and, since both G_{ol} and F are positive:

$$G_{loop} < 0$$

we need to have:

$$G_{loop} \ll -1 \;\Rightarrow\; |G_{loop}| \gg 1.$$

Moreover, the overall gain of the feedback system can now be written as:

$$G = \frac{G_{ol}}{1 - G_{loop}}$$

from the definition of loop gain we have just given.
It is then possible to rewrite this gain as:

$$G = \frac{G_{ol}}{1 + G_{ol}F} = \frac{\frac{1}{F}}{1 + \frac{1}{G_{ol}F}} = \frac{G_{id}}{1 - \frac{1}{G_{loop}}}$$

observing that the ideal gain G_{id} is the gain we would obtain if G_{loop} were infinite. Moreover, the quantity $1/|G_{loop}|$ is the error between the gain G and the ideal gain G_{id}. For example, we can assume the following data:

$$G_{ol} = 10^5, \quad F = 10^{-2}$$

from which we can immediately calculate the loop gain as:

$$G_{loop} = -G_{ol}F = -10^3.$$

The ideal gain will be:

$$G_{id} = \frac{1}{F} = 10^2$$

while the actual gain:

$$G = \frac{G_{ol}}{1 + G_{ol}F} = 99.9.$$

This gives a relative error between the two loops that is identical to the inverse of the absolute value of the loop gain:

$$\frac{G_{id} - G}{G_{id}} = 0.001 = \frac{1}{|G_{loop}|}.$$

This property can be demonstrated by writing the error signal:

$$\epsilon = S_{in} - FS_{out} = S_{in} - FGS_{in}$$

therefore:

$$\frac{\epsilon}{S_{in}} = 1 - FG = \frac{G_{id} - G}{G_{id}} = \frac{1}{1 - G_{loop}} \simeq \frac{1}{|G_{loop}|}$$

where the last equality holds under the assumption of strong loop and where, on the left-hand side, we can recognize the definition of relative error of the gain:

$$G_{loop} \ll 1 \;\Rightarrow\; \frac{G_{id} - G}{G_{id}} \simeq \frac{1}{|G_{loop}|}.$$

1.3.4 Sensitivity of the system

We can now investigate the sensitivity of the overall gain G to the two elements that composes it.
Consider first the open-loop gain G_{ol}. To understand its influence on G, we must calculate:

$$\frac{dG}{dG_{ol}} = \frac{d}{dG_{ol}}\left(\frac{G_{ol}}{1 + G_{ol}F}\right) = \frac{1}{(1 + G_{ol}F)^2} = \frac{1}{(1 + G_{ol}F)^2} \cdot \frac{G_{ol}}{G_{ol}} =$$
$$= \frac{G}{G_{ol}} \cdot \frac{1}{1 - G_{loop}}.$$

This allows us to write the relative variation of the gain as:

$$\frac{dG}{G} = \frac{dG_{ol}}{G_{ol}} \frac{1}{1 - G_{loop}}$$

where the second term in the right-hand side is much lower than one if the system has a strong loop:

$$G_{loop} \ll 1 \;\Rightarrow\; \frac{1}{1 - G_{loop}} \ll 1.$$

Again, we can start from the previous example:

$$G_{ol} = 10^5,\; F = 10^{-2} \;\Rightarrow\; G = 99.9$$

and try to see what happens when we double the open-loop gain:

$$G_{ol} = 2 \cdot 10^5,\; F = 10^{-2} \;\Rightarrow\; G = 99.95$$

therefore, if the open-loop gain doubles, the relative variation of the gain is much smaller:

$$\frac{\Delta G_{ol}}{G_{ol}} = 2 \;\Rightarrow\; \frac{\Delta G}{G} = \frac{0.05}{99.9} \ll 2.$$

The situation is different when we consider the sensitivity of the gain to a change in the feedback network, that gives:

$$\frac{dG}{dF} = \frac{d}{dF}\left(\frac{G_{ol}}{1+G_{ol}F}\right) = \frac{G_{ol}^2}{(1+G_{ol}F)^2} = -G^2$$

that gives a relative variation:

$$\frac{dG}{G} = -G\ dF = -\frac{dF}{F}\frac{FG_{ol}}{1+G_{ol}F} = \frac{dF}{F}\cdot\frac{G_{loop}}{1-G_{loop}} \simeq -\frac{dF}{F}$$

where under the strong loop assumption the second term is similar to one:

$$G_{loop} \ll 1 \ \Rightarrow\ \frac{G_{loop}}{1-G_{loop}} \simeq -1.$$

Again, from the previous example:

$$G_{ol} = 10^5,\ F = 10^{-2} \ \Rightarrow\ G = 99.9$$

while doubling the feedback gain:

$$G_{ol} = 10^5,\ F = 2\cdot 10^{-2} \ \Rightarrow\ G = 49.98$$

we obtain a much more important variation.
Therefore, the gain is almost independent from the open-loop gain (for which the only requirement is to be big enough to have a strong loop), while it significantly depends on the feedback network, that must be stable.

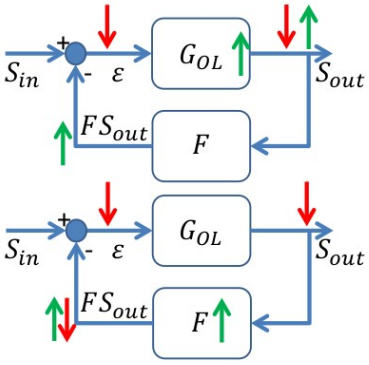


Figure 1.8: Interpretation of the different sensitivity of G from G_{ol} and F.

This result can also be obtained from a physical interpretation of the circuit. Considering the upper part of Figure 1.8, we can observe that an increase in the value of the open-loop gain G_{ol} causes an increase in the output signal S_{out} and, consequently, an increase in the feedback signal FS_{out}. However, this signal is the subtracted to the input signal, thus determining a decrease of the error signal ϵ that, passing through the increased value of G_{ol} compensates the variation of the open-loop gain.
On the other hand, if we have an increase of F in the feedback loop, we will increase also the feedback signal FS_{out} and, consequently, decrease the error

signal ϵ. This causes a decrease of the output signal that tends to compensate the variation of the feedback signal FS_{out}. However, the system now is not compensating the variation of the output, as it was doing for a variation of the open-loop gain. This is the difference between the open-loop gain sensitivity (much lower) and the feedback sensitivity (quite high).
These considerations reflect on the design requirements for negative feedback systems:

- the only requirement on the open-loop gain is that it is high enough to cause the loop to be strong:

$$|G_{ol}| \gg 1 \;\Rightarrow\; |G_{loop}| \gg 1$$

  while it can also be unstable and non-linear, being made of active elements[5] and amplifiers (since their fluctuations will be reduced by the loop gain);

- on the other hand, the feedback gain F must be very stable to not vary the gain and therefore it will be made by passive elements.

The difference in stability between active elements and passive elements is that the first ones are usually made by semiconductors, whose carrier densities depends on a lot of factors (in particular temperature), thus being much less stable, while passive elements are usually made by metals, whose characteristics are much more stable.

1.4 Elementary linear stages and impedances

1.4.1 The operation amplifier

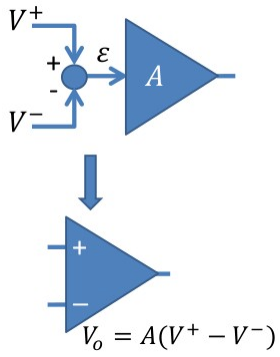


Figure 1.9: Drawing of an operation amplifier.

The main element at our disposal to build feedback systems as we have seen in the previous section is the so called operational amplifier (OA). Operational amplifiers are integrated voltage differential amplifiers, that take as an input the

[5] In general, we define active elements those elements that are able to provide a gain.

difference between two voltages (V^+ and V^-) and give as an output a voltage V_o that is equal to a certain gain A times the input difference:

$$V_o = A \cdot (V^+ - V^-).$$

In general, real operation amplifiers are quite close to ideal operation amplifiers, that are characterized by the following input and output impedances and open-loop gain:

$$R_{in} \to \infty, \quad R_o \to 0, \quad A \to \infty.$$

Even though in the following discussion we will assume to have an ideal operational amplifier, in general real values are slightly different:

$$R_{in} \simeq 10^6 - 10^9\ \Omega, \quad R_o \simeq 100\ \Omega, \quad A \simeq 10^5 - 10^6.$$

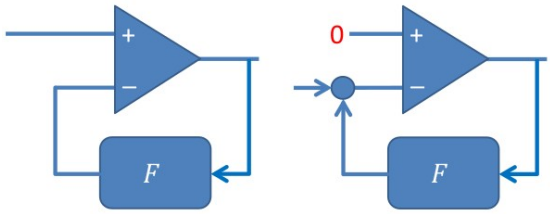


Figure 1.10: Use of an operation amplifier in a negative feedback.

We need now to design a feedback network and to use the negative input of the amplifier to subtract from a reference signal the feedback signal, implementing in this way a lot of mathematical operations.

1.4.2 Non-inverting amplifier

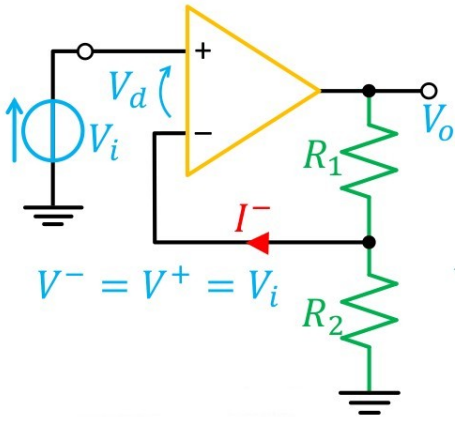


Figure 1.11: Non-inverting amplifier.

The non-inverting configuration of an operation amplifier is shown in Figure 1.11. We can study it assuming to have an ideal operation amplifier and we

will see that the main characteristic of this configuration is that it gives at the output an amplified voltage equal to the input voltage difference with the same sign. Moreover, we can immediately recognized the two resistors on the right as part of the feedback block of the negative feedback network and we can observe that this part must have a connection to the ground.
To determine the behaviour of this circuit, we need to start from the assumption of ideal operation amplifier. Defining I^- the current entering the inverting pin of the operation amplifier, since the input impedance of this ideal element is infinite we can write:

$$I^- = 0$$

and from this we can say that the whole current passing through the resistor R_1 will pass also through R_2, thus allowing us to write the voltage at the negative pin as, from the partition of the output voltage:

$$V^- = V_o \frac{R_2}{R_1 + R_2}.$$

However, since we are dealing with an ideal operation amplifier, the output voltage will be proportional to the difference between the two input pins:

$$V_o = A\left(V^+ - V^-\right) \;\Rightarrow\; V^+ - V^- = V_d = \frac{V_0}{A} \to 0$$

where we have considered that the open-loop gain of an ideal amplifier must tend to infinity. Therefore, the voltage at the negative imput must be identical to the one at the positive one and thus to the input voltage:

$$V^- = V^+ = V_i$$

therefore we are now able to relate input voltage and output voltage, obtaining the following gain:

$$G_{id} = \frac{V_o}{V_i} = \frac{R_1 + R_2}{R_2}.$$

From this derivation, we can make a few comments:

- the gain is always positive, since resistances are always positive, therefore the sign of the output voltage is always equal to the sign of the input's one, from which the name of non-inverting amplifier;
- the gain can be rewritten as:

  $$G_{id} = 1 + \frac{R_1}{R_2} > 1$$

  that is always larger than one;
- the gain is exclusively related to the ratio between the two resistances R_1 and R_2 and not on their absolute value.

We can the investigate the input and the output impedances of this circuit. To calculate the input impedance, we need to impose, for example, a voltage source V_S at the input pin and determine the input current I_S through that pin; the input impedance will be the ratio between these two quantities. Since

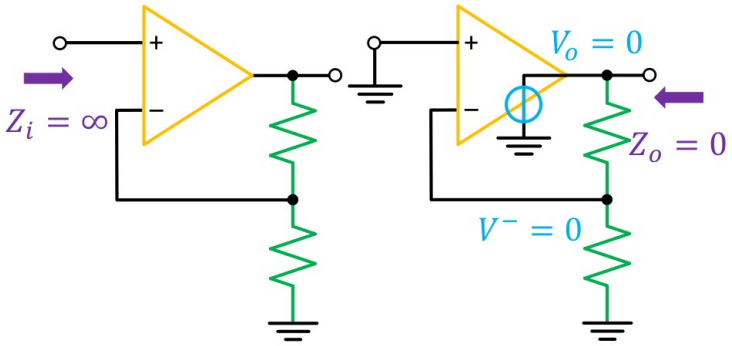


Figure 1.12: Calculation of the input and output impedances in a non-inverting amplifier.

the input impedance of an operation amplifier is infinite, we can immediately see that:

$$I_S = 0 \;\Rightarrow\; Z_i = \frac{V_S}{I_S} \to \infty.$$

To determine the output impedance, we need to shut down every input signal, connecting the input pin to the ground, and we need to impose a certain (in this case, for example) current source I_S at the output node, thus determining the voltage at that node. Since we have:

$$V^+ = 0$$

and the gain of an ideal operation amplifier is infinite, also V^- will be zero and we will not have any current flowing through R_1 and R_2. This means that the current I_S will go inside the operation amplifier and, from the Norton equivalent circuit of this side of the operation amplifier, the current will go inside this pin, whose voltage (equal to V_S) will be zero. Therefore, the output impedance:

$$Z_{out} = \frac{V_S}{I_S} = 0.$$

These two calculations clearly show us that this circuit is equivalent to having an ideal voltage amplifier; the only possible drawback is that it cannot change the sign of the input voltage and, sometimes, it may be useful.

1.4.3 Voltage follower and buffer stage

At this point, we can ask ourselves: what can we do if we would like to have an ideal gain equal to one? Since the gain of the non-inverting amplifier can be written as:

$$G_{id} = 1 + \frac{R_1}{R_2}$$

we can assume to have:

$$R_1 \simeq 0,\; R_2 \to \infty \;\Rightarrow\; G_{id} \simeq 1.$$

In principle, we could have chosen a value for R_1 different from zero, but since if R_2 is very big we cannot have any current passing neither through it nor

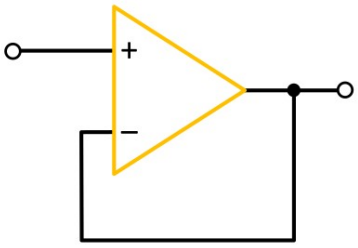


Figure 1.13: Voltage follower.

through the inverting pin of the amplifier, therefore through R_1, it is useless to have this elements and we can replace it with a short-circuit.

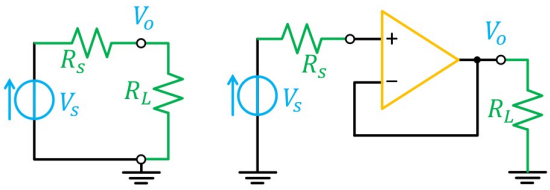


Figure 1.14: On the left, a source is measured without using a buffer stage, while on the right it is present.

This configuration is particularly useful as a buffer stage. Consider the circuit on the left-hand side of Figure 1.14. It is composed by a voltage source V_S and a series resistance R_S (thus being the Thévenin equivalent of a generic source circuit) and on the load side it is composed just by a load resistance R_L, for example belonging to a certain instrument. We want to measure, over R_L, the voltage source V_S. From the partition of the voltage:

$$V_o = V_S \frac{R_L}{R_L + R_S}$$

and we can immediately see that the voltage measured across the load resistance will be different from the voltage imposed by the source. In particular, it will be significantly affected by the presence of the load resistance and this measurement will change using different source with different series resistances R_S. This is an obvious drawback of this circuit, thus we need to find an alternative.
A possibility, represented in the right-hand side of Figure 1.14, is to use a voltage follower as a buffer stage between the source and the load. Since the input resistance of the voltage follower is very high (it tends to infinity), through the source part of the circuit there will not be any current flowing and the voltage at the positive pin of the operational amplifier V^+ will be identical to the one imposed by the voltage source V_S. However, this means that also the voltage of the negative pin is identical to the voltage of the positive one, therefore, the output voltage V_o across the load resistance will be identical to the voltage of the source V_S:

$$V_o = V_S.$$

This means that the instrument that we are using for the measurement (namely, the resistance R_L) does not affect in any way the source and therefore the result

of our measure; in fact, from the viewpoint of the output we have an output resistance that tends to zero. The effect of this buffer stage, made up by a voltage follower, is therefore to decouple the input from the output, removing load effects.

1.4.4 Inverting amplifier

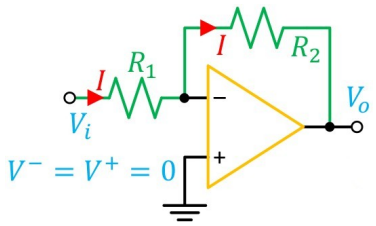


Figure 1.15: An inverting amplifier.

In many applications, however, it is useful to change the sign of the input voltage. Therefore, to reach this goal, we need to apply the signal to the inverting pin of the operation amplifier, thus obtaining an inverting amplifier.
The circuit of an inverting amplifier[6] is represented in Figure 1.15. Again, since we are dealing with an ideal operation amplifier, the voltage of the positive pin will be identical to the voltage of the negative one. Since the positive pin is directly grounded ($V^+ = 0$), the operational amplifier in this feedback system will set:

$$V^- = 0$$

and this node will be called the virtual ground, since it stays at the ground potential even though it is not physically grounded. Applying a certain input voltage V_i, therefore, we will have a certain current flowing through R_1:

$$I = \frac{V_i - V^-}{R_1} = \frac{V_i}{R_1}$$

and since the input impedance of the operation amplifier is infinite, it must be equal to the current flowing through the resistance R_2. This allows us to write the voltage drop across the resistance R_2 and, since it is connected at the output and at the virtual ground, the output must be:

$$V_o = -IR_2 = -V_i \frac{R_2}{R_1}$$

that gives the following ideal gain of the circuit:

$$G_{id} = \frac{V_o}{V_i} = -\frac{R_2}{R_1}.$$

[6]When solving these kind of circuits, the starting point is always to identify whether is V^+ that is set by V^- or vice versa. Once we have done this, we have determined the value of both nodes and we are able to solve the surrounding network.

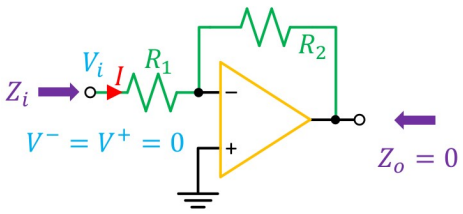


Figure 1.16: Input and output impedances of an inverting amplifier.

Again, we can now study the input and output impedances[7] of this amplifier. For the input resistance, we can set through a generator an input voltage V_i and, from what we have said before, we will obtain the following input current and thus input impedance:

$$I_i = I = \frac{V_i}{R_1} \;\Rightarrow\; Z_i = \frac{V_i}{I_i} = R_1.$$

We can immediately see that the input impedance is not infinite, as we had for the non-inverting configuration: this means that if we are applying a signal that has its own series resistance R_S we will obtain a partition between R_S and R_1, therefore the source will affect the overall gain of the circuit. However, we can immediately solve this problem by posing a buffer stage between the series resistance R_S and the resistance R_1: this will lead to increase the complexity of the circuit, but also to increase its performances.
The output resistance, on the other hand, can be found imposing a current generator I_S at the output. Since now the input voltage is grounded, the resistance R_1 will be between a virtual ground (V^-) and the real ground, therefore we will not have any current flowing neither through R_1 nor through R_2. The output voltage will then be set at ground as well as the input nodes, thus leading to:

$$V_o = 0 \;\Rightarrow\; Z_o = \frac{V_o}{I_S} = 0$$

a zero output impedance, as in the ideal case.
From this configuration, it is possible to build circuits that realize some important mathematical operations on signals.

Adder circuit

Starting from the design of the inverting amplifier, it is possible to build an adder circuit as in Figure 1.17. It consists in an inverting amplifier in which we have connected, to the inverting pin of the operation amplifier, different input branches (each one with its own resistance) one in parallel with the other. The easiest way to study this kind of circuit is exploiting the linearity property, through the superposition principle.

As in Figure 1.18, we can assume to have an input V_1 at the first input node, while every other input node is grounded. Again, since we are dealing

[7] In general, we use a voltage generator at the input and a current generator at the output when studying these impedances. This is a consequence of the fact that we already know the result of this study and this choice will allow us to better understand limiting cases. In general, however, it should be possible also to change this choice without changing the result.

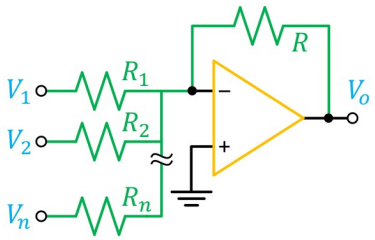


Figure 1.17: An adder circuit.

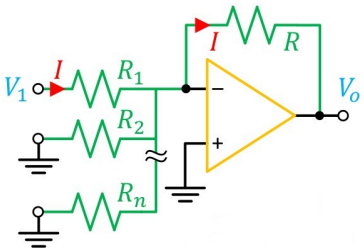


Figure 1.18: The superposition principle applied to an adder circuit.

with an ideal operation amplifier, we can set the negative input of the operation amplifier at ground potential. From this, we can say that there will not be any current flowing through the resistors $R_2, \ldots, R_n$, since they have both ends connected to ground; the only current present will flow through R_1 and it will be equal to:

$$I = \frac{V_1}{R_1}.$$

This current, then, will necessarily pass through the resistor R and it will set the potential of the output, that in case of all input equal to zero except for the first one, will give:

$$V_{o1} = -\frac{R}{R_1} V_1.$$

Repeating an identical analysis for every other condition, in which we will have $V_j \neq 0$ and every other input different from the j-th one set at zero, we will obtain the following output:

$$V_{oj} = -\frac{R}{R_j} V_j$$

and, since by the superposition principle the overall output will be the sum of the outputs in each one of these conditions:

$$V_o = -\frac{R}{R_1} V_1 - \frac{R}{R_2} V_2 - \cdots - \frac{R}{R_n} V_n$$

and assuming the resistances of every input to be each equal to the others:

$$R_1 = R_2 = \cdots = R_n \;\Rightarrow\; V_o = -\frac{R}{R_1} \left(V_1 + V_2 + \cdots + V_n\right).$$

The first comment we can immediately make is that at the output we will obtain the sum of the input potentials with its sign changed and, depending on the ratio

between the resistances R/R_1, amplified or reduced. Therefore, now the name "operation amplifier" can be understood: this element allows us to implement mathematical operations.
The input impedance of this circuit can be studied considering just one input and grounding all the others: as in the case of the inverting amplifier, it will be equal to the input resistance R_j (if $V_j \neq 0$). The output impedance, as in the case of the inverting amplifier, is identically equal to zero.

Subtractor circuit

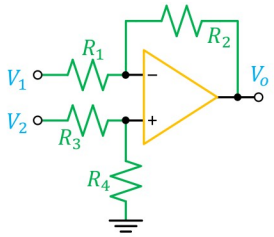


Figure 1.19: A subtractor circuit.

A different possibility is represented in Figure 1.19, where we can see a subtractor circuit. As we can easily imagine, it will make the difference of the signals applied at the two inputs. Moreover, we can observe that also in this case, as in any other case we have seen so far, the feedback is connected to the inverting pin of the operation amplifier, otherwise the system will be unstable. Also in this case, we can investigate it using the superposition principle.

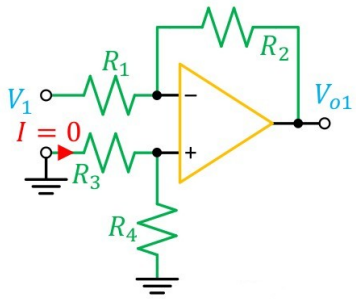


Figure 1.20: A subtractor circuit investigated using the superposition principle (part 1).

The first possibility is to impose the second input equal to the ground potential $V_2 = 0$. Since we are dealing with an ideal operation amplifier, there will not be any current coming or entering the positive pin of the amplifier, therefore the two resistances R_3 and R_4 will be in series and their ends will be both connected to ground. This means that we will not have any current flowing through them and the positive input of the amplifier will be set at ground potential. Since we are dealing with a negative feedback system, then, the negative input of the amplifier will be the virtual ground and we can then write the current flowing

through the resistance R_1 as:

$$I = \frac{V_1}{R_1}.$$

Since this current will flow also through R_2, we will obtain the following output voltage:

$$V_{o1} = -\frac{R_2}{R_1} V_1$$

as if it were an ideal inverting amplifier.

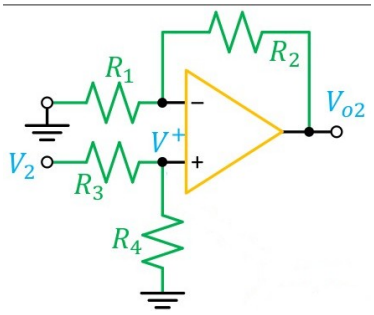


Figure 1.21: A subtractor circuit investigated using the superposition principle (part 2).

On the other hand, if we impose at ground the first input and we have a voltage V_2 at the second one, we can write the voltage of the positive pin of the operation amplifier by writing a voltage partition:

$$V^+ = V_2 \frac{R_4}{R_3 + R_4}$$

but, since we are dealing with a negative feedback system, the positive pin will be kept at the same voltage of the negative one by the circuit and this allows us to write the current flowing through the resistors R_1 and R_2 and, consequently, the voltage at the output:

$$V^- = V_{o2} \frac{R_1}{R_1 + R_2} = V_2 \frac{R_4}{R_3 + R_4} \;\Rightarrow\; V_{o2} = V_2 \frac{R_4}{R_3 + R_4} \cdot \frac{R_1 + R_2}{R_1}.$$

It is possible also to note that this second circuit, with this configuration of the inputs, can be rearranged as a non-inverting amplifier, thus explaining the result we have just obtained.
Summing the results obtained by superposition principle:

$$V_o = V_{o1} + V_{o2} = -\frac{R_2}{R_1} V_1 + \frac{R_4}{R_3 + R_4} \frac{R_1 + R_2}{R_1} V_2 = \frac{R_2}{R_1} \left(-V_1 + \frac{1 + \frac{R_1}{R_2}}{1 + \frac{R_3}{R_4}} V_2 \right)$$

and if we make the following choice of resistances:

$$\frac{R_1}{R_2} = \frac{R_3}{R_4} \;\Rightarrow\; \frac{1 + \frac{R_1}{R_2}}{1 + \frac{R_3}{R_4}} = 1$$

we obtain:

$$V_0 = \frac{R_2}{R_1}(V_2 - V_1)\,.$$

Therefore, this circuit allows us to calculate the difference between the voltages of the two inputs, increased or reduced by a factor R_2/R_1 dependent on the ratio of the two resistance considered.

A few considerations

Before ending this section, it is important to make a few comments. First of all, up to now we have only considered voltage amplifiers, but it is possible to create these kind of operations also dealing with other kind of amplifiers. Moreover, we have seen that often the closed-loop gain of the circuits depend on the ratio between the resistors, not on their absolute value, thus allowing us to make a choice. In particular, low-value resistors will give us a better frequency response, but they will draw more current. On the other hand, high-value resistors are more noisy and enhance leakage currents. In general, therefore, resistors adopted are in the range $10 - 100$ kΩ.

1.4.5 Current-voltage converter

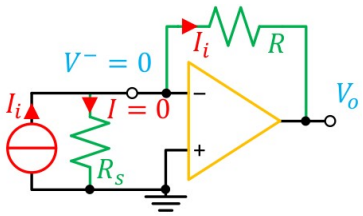


Figure 1.22: A current-voltage (I-V) converter.

Another useful linear circuit is the current-voltage (I-V) converter, represented in Figure 1.22. This kind of circuit takes as an input the current coming from a source (that we have represented through its Norton equivalent circuit) and gives as an output a voltage. Analysing this circuit, we can immediately see that since the positive pin of the operation amplifier is grounded and the system has a negative feedback, the negative pin of the operation amplifier will be set at virtual ground. This means that there will not be any current passing through the resistance of the source R_S, since it will be connected on one side to the ground and on the other side to the virtual ground. Therefore, the whole current imposed by the generator I_i will pass through the feedback resistance R, thus determining an output voltage that is equal to:

$$V_o = -RI_i$$

converting the input current in an output voltage through a conversion factor represented by the input resistance. The gain, therefore, has obviously the dimension of a resistance and we have built a transresistance amplifier. It is important to note that the vast majority of the circuits that we have seen so far relies on the inverting architecture, since it is the most flexible one.

The input resistance of this circuit can be calculated observing that the generator will set a certain non-zero current I_i but that the voltage across the generator V_S is zero due to the presence of the virtual ground; this makes the input impedance to be identically zero:

$$Z_i = 0.$$

This is good result, that makes this circuit resemble the ideal one, since the gain of the amplifier (or converter, depending on its usage) is independent from the resistance of the source. From the analysis that we have carried out many times, it is possible to observe that also the output impedance will be zero:

$$Z_o = 0.$$

1.4.6 Voltage-current converter

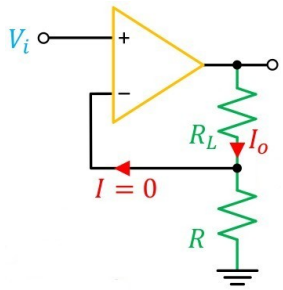


Figure 1.23: A voltage-current (V-I) converter.

In Figure 1.23 it is represented a voltage-current (V-I) converter, also called a transconductance amplifier. In this case, the input is represented by a certain input voltage V_i, while the output is a current I_o passing through a load resistance R_L. Again, since we are dealing with a negative feedback system, the negative pin of the operation amplifier will be set by the system at the same voltage of the positive pin, thus we have that the voltage drop across the resistance R is the input voltage V_i. This will allow us to calculate the current passing through R and, since we are dealing with an ideal operation amplifier, it will be equal to the output current passing through the output resistance:

$$I_o = \frac{V_i}{R}.$$

We can immediately see that in this case the gain has the dimension of a conductance, as expected.
The calculation of the input impedance is quite straightforward: the input set a voltage V_i but we cannot have any current flowing through the pin of the operation amplifier, therefore the input impedance must be equal to the impedance of the operation amplifier:

$$Z_i \to \infty.$$

To calculate the output impedance, we need to shut off any input (thus connecting the positive pin of the operation amplifier to the ground) and to substitute

the load resistance with, for example, a voltage source V_S. Due to the presence of a negative feedback, both input pins of the operation amplifier will be at ground, therefore we cannot have any current flowing neither the negative pin nor the R resistance (since both its ends are connected to ground) and this means that the output current I_S will be equal to zero, thus giving:

$$Z_o = \frac{V_S}{I_S} \to \infty.$$

In this circuit, however, the load is floating, since it is not set at a reference potential. If I want to have one end of the load to be grounded, I need to switch the position of the load R_L and of the resistance R. What will this lead to? The student is ask to think about that.

1.5 Non-linear stages

We have seen in the previous section how it is possible to implement linear operations using the negative feedback system and linear components. However, this is not the end of the story: using non-linear components (namely, capacitors) it is possible to realize circuits that allows us to implement differential operations. Before starting studying them, we need to remind a few, basic properties about capacitors. Assuming $\pm Q$ to be the electric charge on each plate of the capacitor, V the voltage drop across it and I the current (real current in wires or displacement current in the dielectric material inside the capacitor) passing through it, we can write:

$$Q = C \cdot V$$

that the charge on the plates is linearly proportional to the voltage drop between them; the proportionality constant is the capacity of the capacitor. Since we know that the current is the time derivative of the charge and, in general, we can assume the capacity to be constant:

$$I = \frac{dQ}{dt} = C\frac{dV}{dt}.$$

This relationship must be kept in mind while solving the following circuits.

1.5.1 Integrator

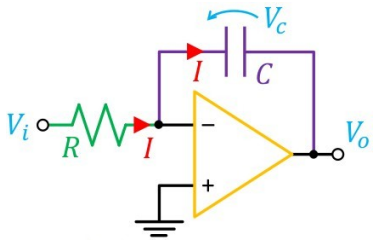


Figure 1.24: An integrator.

The circuit of an integrator is represented in Figure 1.24. As we have said many times, the presence of a feedback circuit[8] and the fact that we are dealing with an ideal amplifier set the negative pin of the operation amplifier at the same potential of the positive pin and, since this last one is grounded, we will have a virtual ground. This means that we are able to write the current flowing through the resistance R as:

$$I = \frac{V_i}{R}$$

and since the impedance of the negative pin of an ideal amplifier is infinite, it will be equal to the current passing through the capacitor:

$$I = C\frac{dV_c}{dt}.$$

Integrating this relationship, we obtain the voltage drop across the capacitor and, consequently, the output voltage (that is equal to the voltage drop across the capacitor with its sign changes):

$$V_o = -V_c = -\frac{1}{C}\int_0^T I\ dt + V_o(0) = -\frac{1}{RC}\int_0^T V_i\ dt + V_o(0)$$

where, in general, the output voltage at $t = 0$ is assumed to be zero. Therefore, the output will be the integral of the input multiplied by a certain constant that we can set designing the circuit.

1.5.2 Differentiator

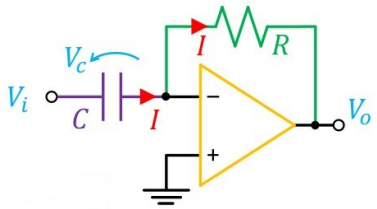


Figure 1.25: A differentiator.

A differentiator is represented in Figure 1.25. Comparing its schematic with the one of the integrator in Figure 1.24, we can see that the only difference between them is in the position of the resistance R and of the capacitor C: they are swapped. Again, since the system has a negative feedback and we are dealing with an ideal operation amplifier, the inverting pin of the operation amplifier will be at the same voltage of the positive one, thus representing a virtual ground. This means that the current I flowing through the capacitor, from the fundamental relationship of the capacitor, can be written as:

$$I = C \cdot \frac{dV_i}{dt}$$

[8] Always remember that this works since we are dealing with a negative feedback and closed-loop circuit, otherwise it will not work.

where V_i, that is the input voltage, is equal to the voltage drop across the capacitor. Then, since the impedance of the negative pin of the operation amplifier is infinite, this current will pass through the feedback resistance R and we can determine the output voltage as the voltage drop across this resistance with changed sign:

$$V_o = -RI = -RC\frac{dV_i}{dt}.$$

We can immediately see that now the output voltage is proportional to the first derivative of the input.
At this point, we are able to implement a lot of mathematical operations using circuits and, as we have already said, from this property comes the name "operation amplifier". In fact, before the invention of the transistor, this was the only way in which it was possible to implement analogically these operations, thus solving differential equations.
Moreover, we can observe that the property we have discussed of the differentiator and of the integrator that, being one the reciprocal operation of the other, they can be realized only swapping two elements of the circuit, is a general property and it will hold also for many others circuits.

1.5.3 Impedance representation

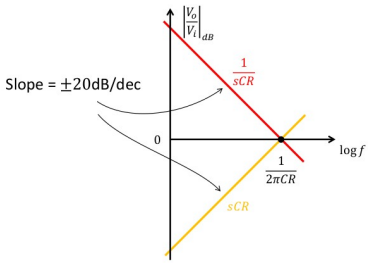


Figure 1.26: Bode plots of a differentiator and an integrator.

Usually, in electronics, it is widely the so called symbolic (or impedance) representation. In it, we get rid of derivatives and integrals by using the Laplace transform and the Laplace operator s. Considering for example the fundamental relation of a capacitor in the Laplace domain:

$$I = C\frac{dV_c}{dt} \;\Rightarrow\; I = sCV_c \;\Rightarrow\; Z_c = \frac{1}{sC}$$

we are able to write a Ohm's law for impedances. Applying it to the previous circuits and substituting the capacitor with its complex impedance Z_c, the relation between input and output in an integrator can be written as:

$$V_o = -\frac{1}{sCR}V_i = -\frac{Z_c}{R}V_i$$

while for a differentiator:

$$V_o = -sRCV_i = -\frac{R}{Z_c}V_i.$$

This notation is particularly useful since it allows us to represent the frequency response of these circuits in Bode plots, as in Figure 1.26.

1.6 Real operation amplifiers: DC and AC parameters

Up to now, we have studied only ideal operation amplifiers. However, these kind of devices cannot be obtained, therefore we need to discuss the differences between real operation amplifiers and ideal ones. Before starting, it is useful to remind the main characteristics of an ideal operation amplifier:

- infinite voltage gain at every frequency;
- infinite bandwidth (as a consequence of the previous requirement);
- infinite input impedance;
- output impedance equal to zero.

The fundamental relationship for an ideal operation amplifier can be written as:

$$V_o = A\left(V^+ - V^-\right) = A \cdot V_d$$

where V_d is the differential input voltage, since we are dealing with a differential amplifier.
Obviously, the real characteristics of an operation amplifier will be approximations of these ones: for example, instead of an infinite input impedance and voltage gain we will have large (but finite) values of the input impedance and of the voltage gain and instead of having an output impedance equal to zero we will have a very small value of the output impedance. These real values can be found in data-sheets.
Data-sheets are, in general, a list of parameters given by the manufacturer that should be useful for describing the behaviour of a certain operation amplifier depending on its model, on its characteristics and on the working conditions. Data-sheets are usually divided into the following sections:

- "Features and General description", sometimes also containing block diagrams, schematics and examples of applications of the operation amplifier; it gives an overview of the operation amplifier, possibly stressing[9] the strong points of that particular model with respect to, for example, low noise, high frequency, power, capability, ... depending on the target;
- "Absolute maximum ratings", that states the condition in which the operation amplifier is no longer able to work;
- "Operating conditions", that states typical working conditions;
- "Electric characteristics", that states the characteristics of the operation amplifier from an electrical point of view;
- "Typical performance characteristics", that states the typical performances of the operation amplifier in certain working conditions;
- other informations, such as dimensions, package, ordering informations and so on.

We can now study in further details the most relevant sections.

[9] Advertising is everywhere!

1.6.1 Absolute maximum ratings

Absolute maximum ratings are the maximum values of certain parameters that the operation amplifier can safely tolerate. It is important to note that operation beyond these values can possibly lead to a permanent damage of the device. In particular, these parameters can be:

- input voltage;
- supply voltage;
- temperature;
- power dissipation.

Again, since these maximum values must not be exceeded, otherwise we can permanently damage the device, we need to check these values during the circuit design.
An immediate example of this is represented by the power dissipation. In fact, the operation amplifier, aside from the input and output voltages that we have previously seen, will require a connection to a supply voltage, in general much higher than the input voltage, that will allow the operation amplifier to correctly set the output. Considering an ideal amplifier, it will have an infinite input impedance, therefore if we consider for example the design of the buffer stage represented in the right-hand side of Figure 1.14 at page 16, we will not have any dissipation of power at the input of the operation amplifier. However, the operation amplifier will be connected to a supply voltage[10] much higher than the input voltage, thus setting an output voltage and leading to the following power dissipation over the load resistance:

$$P = \frac{V_i^2}{R_L}.$$

From these values we can calculate the current[11] coming from the output as:

$$I = \frac{V_i}{R_L}$$

and therefore we will have a voltage difference between the supply voltage and the output voltage equal to $V_{sup} - V_i$, thus determining a dissipation of power inside the internal circuitry of the operation amplifier[12] equal to:

$$P_{oa} = (V_{sup} - V_i) \cdot I.$$

This power, as any dissipated power, will generate heat, that will increase the temperature of the operation amplifier with respect to the environment. However, we need to limit this temperature due to the characteristics of the device.

[10] For example, we can assume to have a supply voltage of 15 V, an output voltage equal to the input one and equal to 2 V and a load resistance of 500 Ω. From these values, the dissipated power over the load resistance will be equal to 8 mW.

[11] In the example considered, it will be 4 mA.

[12] Again, in our numerical example we have that the voltage difference between the supply voltage and the output is equal to $15 - 2 = 13$ V thus determining a dissipated power over the operation amplifier of $13\ \text{V} \cdot 4\ \text{mA} = 52$ mW.

In fact, an increase in the temperature of the device (that is basically a junction) will generate an increase of the carriers density in the junction, thus increasing the current flowing through the device. But this increase in the current will increase the dissipated power in the operation amplifier, that will further heat up the device, increasing the current and so on and so forth with a positive feedback that, at the end, leads to the thermal runaway of the device.
In general, we can define $\theta_{j,A}$ the thermal resistance of our device (and it is a parameter listed in the data-sheet with respect to a certain temperature of the environment), therefore we can write the power dissipated across the operation amplifier as:

$$P = \frac{T_j - T_A}{\theta_{j,A}}$$

where T_j is the temperature of the junction and T_A the temperature of the environment. Obviously, to a certain maximum value of the dissipated power P_{max} will correspond a certain maximum temperature $T_{j,max}$ of the junction that must not be exceeded. From the value of the thermal resistance and of the temperature range of the junction it is possible to work out the maximum power dissipation across the operation amplifier, thus checking if we are respecting this value during the design of a circuit. For example[13], we can have:

$$T_{j,max} = 150°\text{C},\ \theta_{j,A} = 103°\text{C/W}$$

and if the environment temperature is 25°C we have that the maximum dissipated power will be:

$$P_{max} = 1.2\ \text{W}.$$

In general, we can say that the power dissipated in an operation amplifier can reach several tens of watts.

1.6.2 Operating conditions

Operating conditions are recommended working conditions that are specified by the manufacturers and in which the gain and the input and output impedances are controlled and equal to certain expected values. In these intervals, therefore, the operation amplifier is guaranteed to work as specified in the data-sheet. It is important to note that these values will be in general lower than the absolute maximum ratings that we have introduced in the previous section; moreover, exceeding these conditions but being below the absolute maximum ratings the operation amplifier will still be working but its characteristics will be somehow unpredictable and different from the expected ones (that are specified in the data-sheet). In particular, they are ranges of supply voltage, input voltage and temperature (that are in general always specified) and of others parameters depending on the choice of the manufacturer and on the strong points of that model of operation amplifier.

With respect to the supply voltage, a dual symmetrical power supply is almost always used[14], as represented in Figure 1.27. This means that one supply pin of the operation amplifier will be connected to a certain positive voltage while the other will be connected to a certain negative voltage and that, apart

[13] These values are taken from slide 9 of lecture "L03" available on the teacher's website.
[14] In general, we will implicitly assume this kind of power supply.

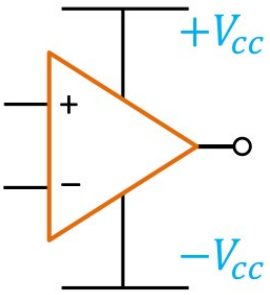


Figure 1.27: Representation of an operation amplifier in which is represented also the dual symmetrical power supply.

from the sign, these two voltages will be equal. These two pins will be called, respectively, positive and negative power supply. Another possibility is to use a single supply, in which one pin is at a certain positive voltage $+V_{cc}$ while the other is grounded. This second choice is in general easier in terms of sources, but it will lead to additional complications in the design of the circuit. On the data-sheet it is usually possible to find the supply voltage range, that states a minimum and maximum value for the supply voltage $\pm V_{cc}$.

1.6.3 Electrical characteristics

This section of the data-sheet shows the most important properties of an operation amplifier, usually reporting the minimum and maximum values of certain variables at the operating conditions.

Input and output voltage ranges

In general, the input and output voltages (that are the voltages of the input pins and of the output pin) must be between the positive supply voltage and the negative supply voltage and, moreover, they cannot be chosen too close to the power supply voltages. This means that the input and output voltage range will be a little smaller that the power supply range (except in the case of special designs) and, in general, we can assume, as a rule of thumb, this range to be equal to:

$$\pm \left(|V_{cc}| - 1 \right)$$

for the input, where $|V_{cc}|$ is the absolute value of the dual symmetrical power supply voltage. Moreover, it is also possible to define a common-mode input range, that will be related to the input voltages of the two input pins by the following relationship:

$$\frac{V^+ - V^-}{2}.$$

For the output voltage, again, we can say that it must always be between the positive power supply voltage and the negative one (in the case of dual symmetrical power supply). Therefore, the supply voltage range will limit the so

called output voltage swing (that is how it is generally called the output voltage range). If we are out of this condition, the transistors inside the operation amplifier will saturate.

Large-signal voltage gain

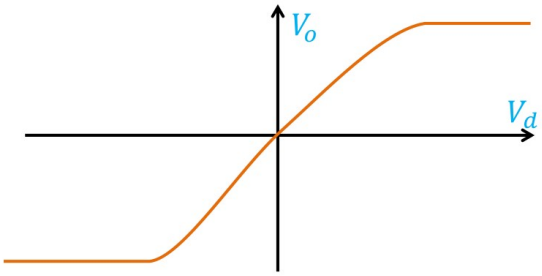


Figure 1.28: Example of a non-linear input-output characteristic of an operation amplifier.

In Figure 1.28 it is possible to observe a possible representation of the input-output characteristic of a real operation amplifier. This curve can in general be measured by setting a voltage difference V_d between the two input pins of the operation amplifier and measuring the output voltage V_o of the amplifier and it represents the fourth check[15] that we need to perform over an operation amplifier before using it. In this graph, we can observe that the unit of measurement for the vertical axis will be, in general, volts, while for the horizontal axis we will use fractions of millivolts; this means that, in reality, the slope of this curve should be quite high.
For small values of the differential input voltage V_d, it is possible to write a linear approximation of this curve, while near to the limiting values of the output voltage (that will obviously be within the supply voltage range) this linear approximation will not be accurate and we can say that the operation amplifier will be saturating. The slope of the linear approximation of this curve for small values of the differential input voltage is the gain A_V in the operating conditions. In general, it is written using decibels:

$$A_V|_{dB} = 20 \cdot \log_{10}(|A_V|)$$

and it is called large-signal voltage gain since it will approximate the behaviour of the output with respect to the input over the whole output range (except for saturation). Typically, this value is between 80 and 120 dB and it will not be constant, depending on a lot of parameters such as the supply voltage V_{cc}, the load resistance R_L, the temperature T and so on. In general, manufacturers specify, in data-sheets, an interval of typical maximum and minimum values for this gain and, since it will be almost equal to the open-loop gain of negative feedback systems, it will be useful to control if the minimum value of this gain allows us to have a strong enough loop.

[15] Before this, we need to check the absolute maximum ratings, the operating conditions and the input and output voltages.

Input offset voltage

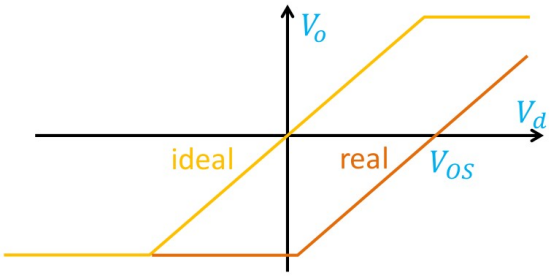


Figure 1.29: Example of an input-output characteristic of an operation amplifier with an input offset voltage.

A second non-linearity of the input-output characteristic, different from the previous one, is represented by the input offset voltage in Figure 1.29. While in an ideal amplifier the input-output characteristic can be described by the following relationship:

$$V_o = A_V \cdot V_d$$

in a real operation amplifier it is possible to have an input offset voltage V_{os}:

$$V_o = A_V \cdot (V_d - V_{os}) .$$

This offset voltage can be either positive or negative and it is usually specified, in data-sheets, as an absolute value. Typically, it will be in the order of a few millivolts (from 1 to 5) or smaller (precision operation amplifiers will have an offset in the order of a few hundreds of microvolts).
In general, this offset is strongly technology dependent, since it is a consequence of the inner design of the operation amplifier. In fact, the two input pins of an operation amplifier are in general the base or gate pin of a transistor (depending on the kind of transistor considered). Ideally, these transistors are completely identical, thus treating the same input voltage in the same way. However, as any physical device, it will be impossible to have two transistors completely identical and therefore even applying the same voltage at the two input pins we will obtain a slightly unbalanced circuit, with currents flowing more in one pin than in the other. Therefore, the mismatch in the input transistors will cause the presence of an input offset voltage and, since the gate or base voltage will determine different currents in the transistor depending on the type of transistor we are dealing with, we will have that this input offset voltage will be strongly technology dependent. This means that, depending on the type of transistors used in the internal design of the operation amplifier, we will obtain different offsets:

- in the best BJT transistors, it will be between 10 and 25 μV;
- in the best JFET transistors, it will be between 100 μV and 1 mV;
- in the best CMOS transistors, it will be between slightly less than 100 μV and 1 mV.

These differences are due to the fact that is much easier to make two BJTs that are similar, than two similar JFETs or CMOSs. Designing a circuit, therefore, we need to ask ourselves whether the input offset voltage V_{os} is an important parameter of the circuit, that can possibly have a strong influence on its behaviour, giving problems or limitations: this will guide our choice toward one technology instead of another one.

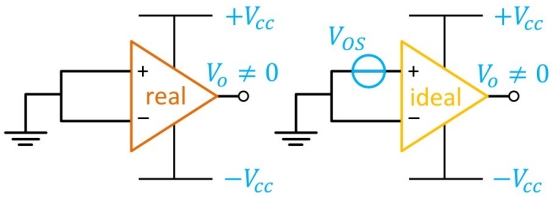


Figure 1.30: On the left-hand side, a real amplifier with a certain input offset voltage; on the right-hand side, the same circuit is represented using an ideal operation amplifier.

Since the input offset voltage is nothing but a shift of the zero of the input-output characteristic (and thus an error in the value of the output voltage), in which to a zero input will correspond a non-zero output, this characteristic can be taken into account by adding an ideal voltage source equal to the input offset voltage to one of the two input pins, as represented in Figure 1.30.
However, another important parameter that we need to take into account is the offset voltage drift[16]. This parameter is defined as the derivative of the input offset voltage with respect to the temperature:

$$\frac{dV_{os}}{dT} \quad \left[\frac{\mu\text{V}}{^\circ\text{C}}\right].$$

This represents an important check, since we have to be sure that in the temperature range of our device the offset input voltage is not only tolerable but also it cannot change to much if the temperature changes. Typically, values for general purposes operation amplifiers are between 1 and 10 μV/$^\circ$C, while for low drift operation amplifiers (in which this term is more dependent on the design of the circuit than in its technology) it will be lower than 0.3 μV/$^\circ$C. A particular family of devices called "chopper-stabilized" or "auto-zero" operation amplifiers then have an offset input voltage lower than 1 μV and have an offset voltage drift lower than 30 nV/$^\circ$C.

Input bias current

In the previous section, we have said that the input pins of an operation amplifier consist in two transistors. From this, we can add another possible difference between an ideal operation amplifier and a real device. In fact, it is possible to have currents flowing through the input pins of these transistors, thus making the input impedance different from infinite; these currents are called bias currents and they are strongly dependent on the technology adopted.

[16]In general, the word drift means the dependency from the temperature of a certain parameter.

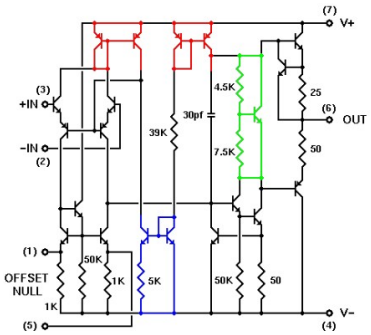


Figure 1.31: Schematic representation of the internal circuitry of an operation amplifier.

In the case of BJTs as input stages, we know that these transistors requires a continuous base current I_b that, therefore, will be between 10^{-8} and 10^{-6} A. If we use JFETs as input devices, the transistors are reverse-biased p-n junctions and therefore the base current will be the reverse bias current of the junction, thus being between 10^{-13} and 10^{-10} A.
In the case of MOSFETs, the situation need to be further explained. In fact, we know that the gate of this transistor is separated from the remaining part of the device by an insulating layer that, ideally, will have a very small leakage current. However, since these transistors are actually extremely small, they will have a very small capacitance, therefore:

$$Q = CV \;\rightarrow\; V = \frac{Q}{C}$$

an extremely small change in the charge at the interface with the oxide (for example, determined by the static electricity that is always present) will cause a significant change in the voltage across the oxide layer, eventually breaking it. Therefore, some protection devices are needed for not breaking the oxide layer. These protections are called Electro-Static Discharge (ESD) protection circuits and they are, basically, diodes, thus having a certain leakage current that represents the bias current of the device and that will be between 10^{-13} and 10^{-10} A. The previous numbers, aside from the technology, will then depend also on the design of the circuit.

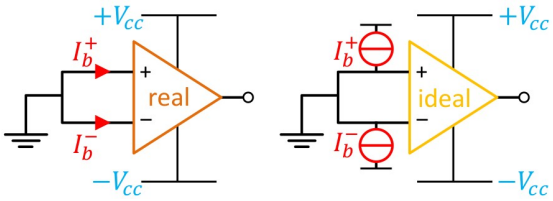


Figure 1.32: On the left-hand side, a real amplifier with a certain input bias current; on the right-hand side, the same circuit is represented using an ideal operation amplifier.

In real devices, therefore, we have certain currents entering the input pins.

This means that, for representing a circuit using an ideal operation amplifier as in Figure 1.32, we need to add two ideal current sources[17]. In general, then, the bias current entering the positive pin I_b^+ is different from the bias current entering the negative pin I_b^-, therefore in data-sheets we can usually find the average value of these currents I_b and the offset current I_{os}, that is defined as the difference between them:

$$I_b = \left| \frac{I_b^+ + I_b^-}{2} \right|, \quad I_{os} = \left| I_b^+ - I_b^- \right|.$$

As a rule of thumb, we can say that:

$$I_{os} \simeq \frac{I_b}{10}.$$

Moreover, both these values are in general dependent on the temperature, thus determining a drift of the bias current and of the offset current. Since then the bias current is larger than the offset one, the drift of the bias current will be more important and it will be strongly dependent on the technology adopted. In general, in FET and CMOS operation amplifiers both the bias current and the offset current are increased by a factor two every 10°C, since they are reverse currents in p-n junctions[18]. Therefore, when designing the circuit we need to check that the bias and offset currents are lower than the maximum value over the whole temperature range. On the other hand, BJT operation amplifiers have usually lower drifts, between 10 and 100 pA/°C, with an almost linear dependency of the current from the temperature. However, we have previously seen that in BJTs (that have better performances with respect to FETs and CMOSs in terms of drift) will have a bias and offset current higher than the values for FETs and CMOSs. Therefore, a trade off is needed in the choice of the technology for the operation amplifier, depending on the drift and the values needed.

Input offset voltage compensation

Since we have previously seen the presence of an offset voltage in our device, we can now investigate the techniques that are at our disposal for compensating this offset. In particular, there are two main ways of doing it: one is provided by the manufacturer, the other requires a do-it-yourself approach.

In the first solution, the operation amplifier has two additional pins, called the offset null pins, to which we can connect an element (for example a variable resistance) according to a detailed compensation scheme provided by the manufacturer in the data-sheet. Trimming this variables resistance and making some measurements on the operation amplifier (in particular, after having connected the two input pins with a short-circuit), it is thus possible to compensate the presence of an input offset voltage.

In this second case, we can bias one of the two input pins (for example, the positive input pin, as represented in the right-hand side of Figure 1.34) at

[17] Carefully think to the orientation of these sources and the current flows that they determine: you always want the bias currents to seem to be entering the pins of the operation amplifier.

[18] We remember that in semiconductor devices the carrier density, that determines these currents, has an exponential dependency on the temperature, thus giving a significant drift.

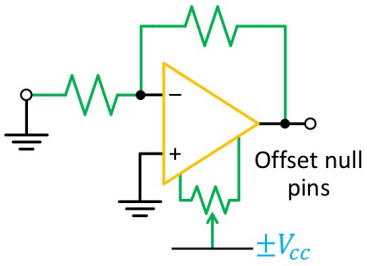


Figure 1.33: Manufacturer-provided solution for offset compensation.

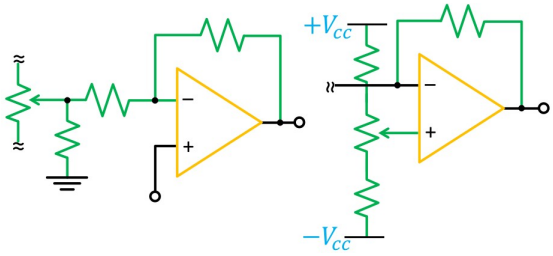


Figure 1.34: Do-it-yourself solution for offset compensation.

a voltage equal to $-V_{os}$. This can be done using a suitable voltage partition involving a variables resistor, that will allow the trim of the circuit.

Bias current compensation

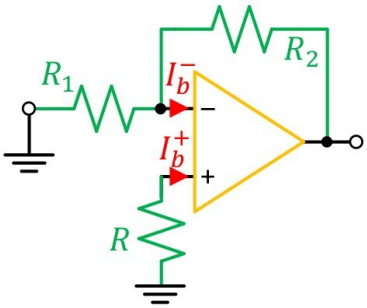


Figure 1.35: Bias current compensation.

In the same way, we can study the effect of a bias current on the output and how it is possible to compensate it. Considering Figure 1.35, we know that if we set the negative input pin to ground and we neglect the bias currents, we should obtain that also the output is at ground (since the positive input pin is already at ground). However, this is not possible and thus we need to further study the circuit.
First of all, we can start considering:

$$R = 0.$$

Under this condition, the circuit can be studied by superposition principle and, setting the negative bias current at zero:

$$I_b^- = 0 \;\Rightarrow\; V_o = 0$$

while if we set the positive bias current to zero then the non-zero negative bias current will flow through the resistance R_2 (but not through R_1), thus leading to the following output:

$$I_b^+ = 0 \;\Rightarrow\; V_o = I_b^- R_2.$$

Therefore, the overall output will be obviously different from zero:

$$V_o = I_b^- R_2.$$

This means that we need to compensate this effect that is related to the presence of the bias currents and, to do this, we can assume to have a resistance R on the circuit that is different from zero. Again, also in this case we can study the circuit applying the superposition principle and therefore observe that if the positive bias current is zero the response of the circuit is identical:

$$I_b^+ = 0 \;\Rightarrow\; V_o = I_b^- R_2$$

while if the negative bias current is zero the output voltage will be:

$$I_b^- = 0 \;\Rightarrow\; V^+ = -I_b^+ R \;\Rightarrow\; V_o = -I_b^+ R \frac{R_1 + R_2}{R_1}$$

since we have recognized that this is a non-inverting amplifier[19]. By superposition principle, therefore, we get:

$$V_o = I_b^- R_2 - I_b^+ R \left(\frac{R_1 + R_2}{R_1} \right) = I_b R_2 \left(1 - \frac{R}{R_1 \| R_2} \right) + I_{os} \frac{R_2}{2} \left(1 + \frac{R}{R_1 \| R_2} \right)$$

where the symbol $\|$ means that the two resistors are in parallel. If we set the following value of R:

$$R = R_1 \| R_2$$

that, from a physical point of view, is equivalent to setting the equivalent resistance seen from the positive input pin equal to the equivalent resistance seen from the negative input pin, we obtain that the effects of the bias current are compensated:

$$V_o = I_{os} R_2$$

and we are left only with the effects of the offset current. Obviously, this kind of compensation is useful only if the offset current is much lower than the bias one:

$$I_{os} \ll I_b.$$

Moreover, we know that the bias and the offset currents will depend significantly on the temperature, therefore a change in the temperature will make the circuit

[19] It is important to be able to study circuits recognizing inverting and non-inverting amplifiers and thus applying their fundamental relationships. In fact, solving every time from scratch the circuit can be extremely time consuming.

not compensated. This means that it has to be periodically compensated using other techniques.
Summing up, not compensating the effects of the offsets and of the bias, we can get errors in the output voltage. Moreover, we must very carefully pick compensation resistors in order to not alter, with their value, the gain of the operation amplifier. Finally, we will not be able to compensate drifts and long-term instabilities by using these solutions.

Total error for the inverting amplifier

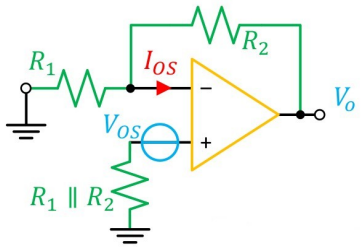


Figure 1.36: Computation of the total error for the inverting amplifier.

We can now calculate the effects of an offset voltage and an offset current (the bias current has been removed as described in the previous section) on the output of an inverting amplifier. This can be done, considering the circuit in Figure 1.36, by linear superposition[20], thus obtaining the following expression for the output:

$$V_o = -\frac{R_2}{R_1}\left(V_i - V_{os}\frac{R_1 + R_2}{R_2} - I_{os}R_1\right).$$

Observing this expression, we can see that it is composed by three terms: the first relates the input to the output, the second relates the offset voltage to the output and the third one relates the offset current to the output. The second term will then be called offset voltage error and the third one the offset current error and, together, they will be the error.
For having a certain gain, we need to impose:

$$R_2 > R_1$$

and this means that the offset error we will have:

$$\frac{R_1 + R_2}{R_2} \simeq 1.$$

Therefore, the overall error will depend on the size of the input voltage with respect to the offset voltage and on the size of R_1 (in fact, decreasing it we can make the offset current error smaller). However, since R_1 is also the input impedance of the operation amplifier, we need to have an high value of R_1 for making the amplifier similar to an ideal one. These two contrasting requirements, therefore, will give the need of a trade off in the choice of the input resistance R_1.

[20] The student is ask to solve this circuit as an exercise.

Total error for the non-inverting amplifier

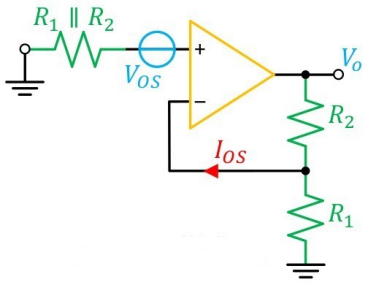


Figure 1.37: Computation of the total error for the non-inverting amplifier.

We can represent a non-inverting amplifier with a non-zero offset current and a non-zero offset voltage as in Figure 1.37. Solving again the circuit by superposition[21], we can obtain the following output voltage:

$$V_o = \frac{R_1 + R_2}{R_1} \left[V_i + V_{os} + I_{os} \cdot (R_1 \| R_2) \right]$$

where, again, the last two terms represent the error. In this case, we can observe that the offset voltage is directly competing with the input voltage, therefore again we need to have an input voltage that is significantly higher than the offset one. Moreover, in this case we do not have any specific trade off, since the parallel $R_1 \| R_2$ controls the offset current error and we can safely minimize it, while the gain is controlled by the ratio $(R_1 + R_2)/R_1$. Finally, the input impedance of this non-inverting amplifier is infinite. These properties, therefore, makes the minimization of the error easier in the non-inverting amplifier.

Input and output resistance

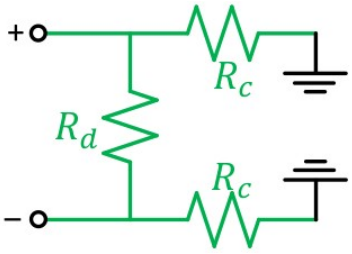


Figure 1.38: Input impedances.

Another difference between an ideal operation amplifier and a real operation amplifier is that the input impedance is not infinite. Since the input of an operation amplifier is characterized by two pins, we need to describe them with an equivalent circuit as in Figure 1.38. This can be done by adding a differential resistance R_d that connects the two pins and by two common-mode resistances

[21] This solution is left to the student.

R_c that connects each input pin to the ground and that, in general, can be assumed to be identical. Moreover, we can say that in general:

$$R_c \gg R_d$$

and, in many cases, the common-mode resistances are negligible. Sometimes, the values of these resistances are specified by the manufacturer in the data-sheet, but it can happen to have only the value of a certain, not further specified input resistance, that can in general be approximated with R_d. Moreover, it is possible to add a small input capacitance C_{in} (its value will be approximately of a few picofarads) in parallel to the differential resistance and that must be considered at high frequencies (in fact, even if small, it can give some problems in the stability of the circuit).
The values of these resistance are largely dependent on the technology: for a typical BJT, they will be between 10^5 and 10^8 Ω; for a typical JFET, they will be between 10^9 and 10^{10} Ω; for a typical CMOS, they will be in the order[22] of 10^{12} Ω. Moreover, it is important to remember that the network to which these resistances are equivalent can be extremely complicated.

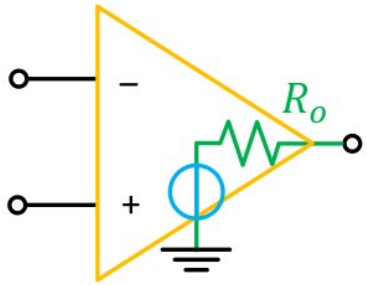


Figure 1.39: Output impedances.

The output resistance, on the other hand, will in general be a series resistance R_o connected between the voltage-controlled voltage source and the output pin, as shown in Figure 1.39. It will be typically smaller than 100 Ω or slightly higher in the case of CMOS operation amplifiers.

1.7 Instrumentation amplifiers, Common-Mode Rejection Ratio and Power Supply Rejection Ratio

Now, we are able to deal with some characteristics of the operation amplifiers that are more related to the use we make of it in a certain instrumentation.

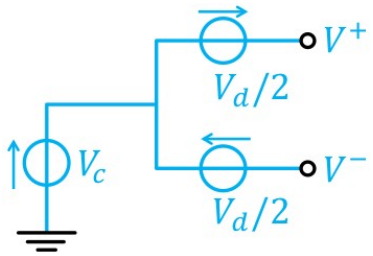


Figure 1.40: Common and differential modes.

1.7.1 Common and differential modes

In an operation amplifier, given the input voltages at the positive pin V^+ and at the negative pin V^-, we can define the common-mode voltage:

$$V_c = \frac{V^+ + V^-}{2}$$

and the differential-mode voltage:

$$V_d = V^+ - V^-$$

as a possible definition of different input variables. We can immediately see that they can be represented as in Figure 1.40 and their values are usually given by the manufacturer. From them, reversing these relationships, it is possible to obtain back the voltages at two input pins:

$$V^+ = V_c + \frac{V_d}{2}, \quad V^- = V_c - \frac{V_d}{2}.$$

The meaning of these two new variables can be understood by linear superposition. In fact, if the differential-mode voltage is equal to zero, then the common mode voltage V_c will be the voltage applied to both the positive input pin and the negative input pin, thus being the voltage in common. On the other hand, if we set the common-mode voltage to zero, we will have a perfectly symmetric and zero-average signal applied to the two pins:

$$V^- = -V^+$$

thus being a truly differential signal. Again, the output of the ideal operation amplifier will depend only on the differential-mode voltage:

$$V_o = A(V^+ - V^-) = AV_d.$$

1.7.2 Common-mode rejection ratio

In a real amplifier, however, the output voltage is not only related to the differential-mode voltage, but it is also related to the common-mode voltage through a suitable gain:

$$V_o = A_d V_d + A_c V_c = A_d \left(V_d + \frac{A_c}{A_d} V_c \right) = A_d \left(V_d + \frac{V_c}{CMRR} \right).$$

[22]This value is extremely high and, in general, it will be really difficult to measure.

In an ideal amplifier, this common-mode gain A_c will tend to zero, but in a real amplifier it will be a finite quantity. From the previous formula, therefore, we are able to define a factor $CMRR$ that stands for Common-Mode Rejection Ratio that express the residual part of the common-mode voltage that pass to the output.
Again, in ideal amplifiers the common-mode rejection ratio is infinite, thus stating the impossibility for the common-mode signal to be transformed into a differential one and thus passing to the output. For example, we can consider the following real amplifier:

$$CMRR = 120 \text{ dB} = 10^6, \quad V_c = 5 \text{ V}, \quad V_d = 0$$

that will have the following output voltage (if the differential signal is zero):

$$V_o = A_d \cdot \frac{V_c}{CMRR} = A_d \cdot 5 \ \mu\text{V}.$$

This means that a common-mode signal of 5 V is passed to the output as if it were a 5 μV differential signal.

1.7.3 Power supply rejection ratio

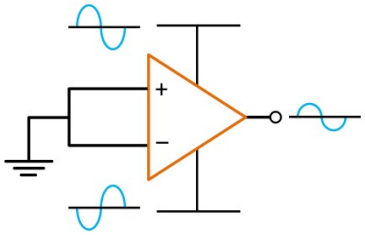


Figure 1.41: A residual oscillation in the supply voltages will pass to the output voltage reduced by the power supply rejection ratio.

Aside from the previous possibility, it is also possible that a small fluctuation in the power supply voltage is passed to the output of the operational amplifier. To study it, we need to set both the common-mode and the differential-mode voltages to zero, as in Figure 1.41. In fact, we will never have a constant power supply, since it will always show small fluctuations that will be, for example, residuals of the rectification of the supply signal (or bias, or regulators), and that will always have a non-zero transfer to the output (since it is not possible to perfectly decouple these signals). Again, we can define a certain gain for these power supply fluctuations V_{ps}, thus obtaining:

$$V_o = A_d V_d + A_{ps} V_{ps} = A_d \left(V_d + \frac{A_{ps}}{A_d} V_{ps} \right) = A_d \left(V_d + \frac{V_{ps}}{PSRR} \right).$$

Therefore, we can define a Power Supply Rejection Ratio[23] $PSRR$ that is the ratio between the power supply disturbs and the differential signal that gives the same output:

$$PSRR = \frac{A_d}{A_{ps}}.$$

[23] It can also be called Voltage Supply Rejection Ratio $VSRR$.

This term will represent an additional source of noise and it will depend on how noisy is the environment of the operation amplifier.

1.8 Frequency response of OA circuits

We can now start to study the properties of operation amplifiers with respect to the frequency on the input signal. Using the Laplace transform[24], the open-loop gain, that is the gain of the operational amplifier, can be usually described as a decreasing function of the frequency and, therefore, as a single-pole function:

$$A(s) = \frac{A_0}{1 + s\tau}.$$

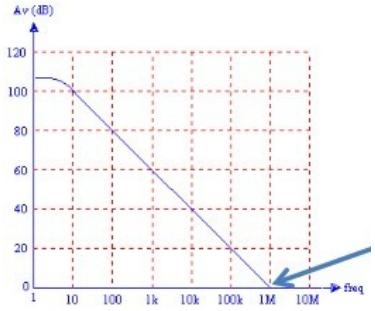


Figure 1.42: An example of a single-pole transfer function.

As it can be seen from the Bode diagram in Figure 1.42, at a certain frequency the gain of the operation amplifier will become unitary. Since the frequency can be related to the angular frequency as:

$$f = \frac{\omega}{2\pi}$$

we can define the gain-bandwidth product $GBWP$ as the frequency at which the gain is unitary; from the expression of the gain we have written before, it means that:

$$GBWP = \frac{A_0}{2\pi\tau}.$$

In general, the gain-bandwidth product goes from the hundreds of kilohertz to a few megahertz, while the pole of the transfer function is in general between 1 and 10 Hz. In data-sheets, manufacturers usually give the zero gain A_0 and the gain-bandwidth product. It is important to note that, since the gain of the operation amplifier is the open-loop gain in negative feedback systems, the decrease of the gain can leads to problems when dealing with high-frequency signals (where the word "high" means in the order of the gain-bandwidth product).

To understand the slew rate, then, we need to further investigate the internal circuit and the block scheme of an operation amplifier, that is represented in

[24]We remember that the Laplace operator can be written as:

$$s = j\omega.$$

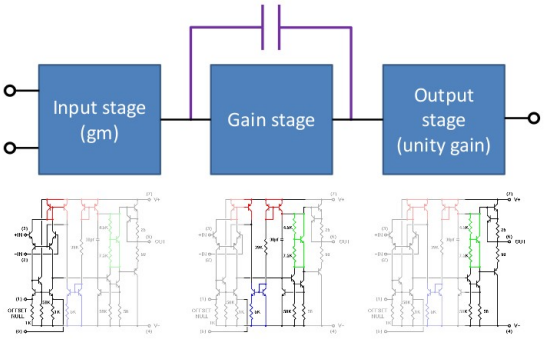


Figure 1.43: Block scheme of an operation amplifier and corresponding internal circuit highlighted.

Figure 1.43. Basically, any operation amplifier can be divided into three blocks, that can be individually optimized, thus having an higher flexibility. The first, input stage consists in a transconductance amplifier, that takes as an input a differential voltage and gives as an output a current. The second stage is the gain stage, in which the input signal is actually multiplied by a certain factor. The third stage is the output stage, that has a unity gain and that gives the output voltage. In the Figure, then, we can notice the presence of a capacitance that is in parallel with respect to the gain stage. In fact, without this capacity, the system will never be stable, due to the presence of poles in the gain stage. This means that we need to add this compensation capacitance across the gain stage, that acting as an integrator will make it stable.
From what we have discussed in previous sections, in a certain interval of input differential voltage the input-output characteristic of the amplifier is almost linear. Outside from the input range, the system saturates and therefore, from the input stage, we will obtain a maximum, saturation current that can be written as I_{sat}. This current, then, will flow through the gain stage in parallel with the capacitance, therefore across this integrator. This means that we will be charging the capacity C and, calling t the time passed from an initial moment set to be the one in which the current has saturated to I_{max}, we can write the voltage across the capacitor as:

$$V_c = \frac{I_{max}}{C} t.$$

This means that when we have a certain instantaneous modification of the input voltage from one value to another one that causes the saturation, the output voltage will not pass instantaneously to its saturation value, but it will increase (at least initially) linearly with time toward its saturation value.

Therefore, while the current coming from the input stage can change abruptly, reflecting the change in the input differential voltage, the output voltage will increase up to its maximum value in a certain finite time (in which the capacitor is charged). The derivative of the increase of the output voltage as a consequence of an abrupt increase in the input voltage is defined as the slew rate:

$$SR = \left.\frac{dV_o}{dt}\right|_{max} = \frac{I_{max}}{C}.$$

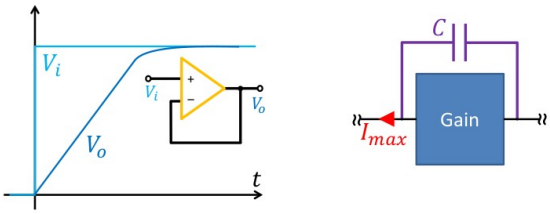


Figure 1.44: Representation of the slew rate due to the presence of a compensation capacitance across the gain stage.

Typical values for this coefficient are between 1 V/μs and 100 V/μs or more. This is the first explanation of the non-linearity of the behaviour of the device over an infinite interval of inputs.
We can now investigate what happens when we apply a sinusoidal input. For example, we can consider a buffer stage at which we are applying a sinusoidal input. If the frequency of the sinusoidal input is low enough to make the increase in the sinusoidal signal be below the slew rate, then the output will be proportional to the input, thus being sinusoidal. If we write the output signal as a consequence of a sinusoidal input as:

$$V_o = V_M \sin(\omega t)$$

we can express the previous condition as:

$$\left.\frac{dV_o}{dt}\right|_{max} = \omega V_M < SR$$

and from the definition of frequency with respect to the angular frequency:

$$f < \frac{SR}{2\pi V_M}.$$

If we assume that the operation amplifier is working at its maximum swing V_o^{max}, then for not being limited by the slew rate the frequency of the input (and thus of the output) signal must be lower than a parameter called full-power bandwidth $FPBW$:

$$f \leq \frac{SR}{2\pi V_o^{max}} = FPBW.$$

Increasing the frequency above this value, we will obtain distortions in the output, that will not follow a sinusoidal shape but will rise in a straight line limited by the slew rate. Therefore, in a more formal way we can define the full-power bandwidth as the frequency at which we can work at full power supply without obtaining any distortion.

1.9 Typical performance characteristics

We are now able to study a few examples of data-sheets, observing what are some typical values for the previously defined parameters. Again, it is worthy to recall that there a few common data that are always reported, while others

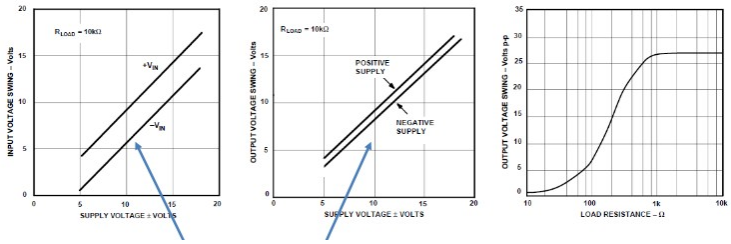


Figure 1.45: (a) Input voltage swing; (b) Output voltage swing; (c) Output voltage swing as a function of the load resistance.

depends on the choice of the manufacturer, somehow reflecting also the strong points of that kind of device.

In Figure 1.45 we can observe on the left-hand side the input voltage swing and in the centre the output voltage swing of a certain amplifier. It is important to note that both are a little smaller than the power supply voltage and, moreover, they are not symmetrical with respect to the input and to the output, thus determining the fact that positive and negative swings are not the same. On the right-hand side, we can observe the dependence of the output voltage swing on the load resistance. To understand it, we can observe that if we have a load resistance between the output voltage and the ground, to a certain output voltage it will correspond a certain current passing through the load resistance and that must be provided by the power supply. Since the power supply can provide only a limited amount of current, if the load resistance is too small the limitations to the current will limit the output voltage at a value that is lower than the saturation one. On the other hand, if the load resistance is big enough we can easily reach the saturation value of the output swing since the current needed will always be provided by the voltage supply.

In Figure 1.46, we can immediately recognize the Bode diagram of the gain in the top left corner. In this graph, we can find the value of the gain for continuous signals and determine the value of the gain-bandwidth product, that can be further analysed in the zoom in the top right corner. In the bottom left corner we can find the open-loop gain as a function of the temperature for different values of the load resistance; also here it is possible to observe that the positive and negative gains are not exactly equal, thus giving a non-symmetrical behaviour. In the bottom right graph, we can in particular observe how the supply voltage specified is almost equal to the saturation output voltage. All the graphs show various non-linear behaviours that make clearer why we need to create circuits with negative feedback loops.

In Figure 1.47 we can study the offset voltage and its drift for different operation amplifiers. Immediately, we can observe how the offset voltage can be either positive or negative and has a certain statistical distribution with an average value (that in general is zero) and a certain standard deviation. In the right-hand side we can find the drift of the offset voltage, that in the bottom graph is expressed as the drift coefficient on the horizontal axis. Moreover, we can observe that in a JFET the input offset voltage is almost an order of magnitude higher, while the BJT has worse performances in terms of drift.

In Figure 1.48 it is shown the direct dependency of the bias current from

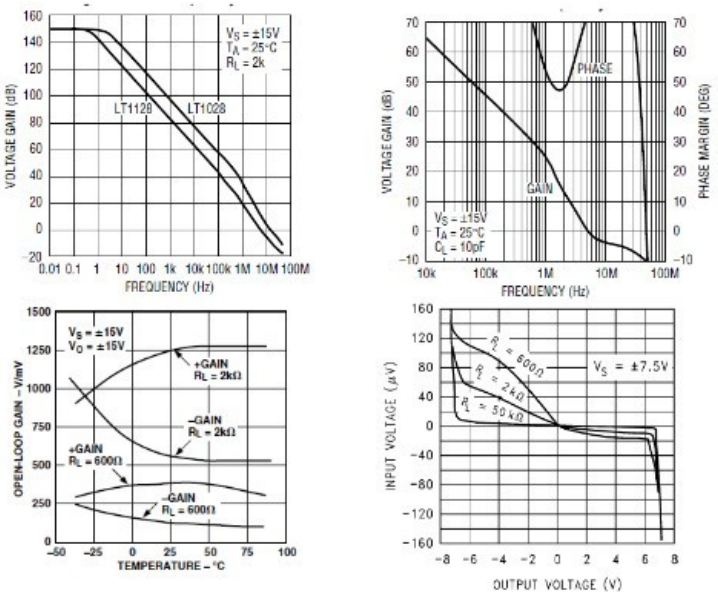


Figure 1.46: From top left to bottom right: voltage gain as a function of the frequency, zoom of the previous graph, open-loop gain as a function of the temperature, input voltage as a function of the output voltage in a low-power CMOS.

the temperature in different technologies. Immediately, we can observe that while in the BJT the scale is linear on the vertical axis, in the JFET and in the CMOS we have a logarithmic scale, thus determining an exponential dependence and a much more significant variation of these currents. However, the unit of measurement for these axis are nano-ampere for the BJT and pico-ampere for JFET and CMOS, thus being much lower in these last cases. This means that while the size of the bias current increases from right to left, the drift of the bias current increases from left to right, depending on the technology. The dotted lines in the CMOS graph means that that measurement was impossible since the value was too low. It is then possible to observe that at very high temperatures (for example, 125°C) the three technologies have almost the same value of the bias current; moreover, we have to remember that the temperature depends also on the power dissipation.

In Figure 1.49 these characteristics about the bias current come from a particular model in which the bias current is extremely low, being in the order of a few femto-amperes. In this case, offset currents are comparable to the bias current and this can be achieved using a special design with an additional current-providing circuit called bias-cancellation circuitry. The take-home message, in this case, is that it is possible to obtain extreme performances using a suitable design.

In Figure 1.50, the open-loop output impedance is represented as a function of the frequency and of the output current. In the left-hand side, where it is represented as a function of the frequency, we can observe that the output impedance is not exactly a pure resistance, since it is not independent from the

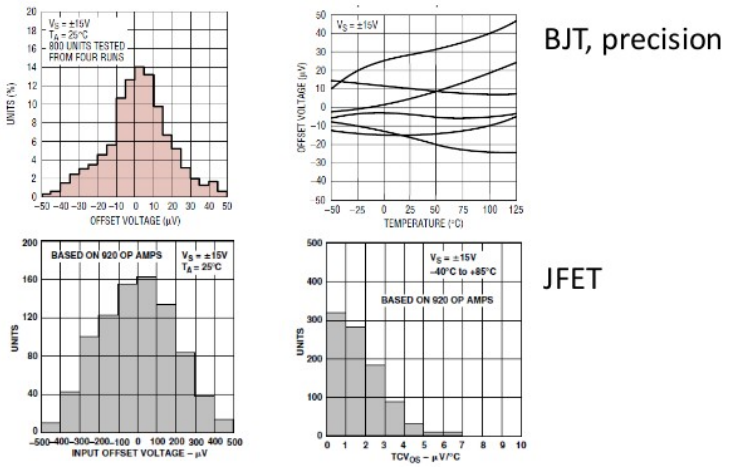


Figure 1.47: On top, the statistical distribution of the offset voltage and the drift in precision BJT operation amplifier; below, the same parameters for a JFET operation amplifier.

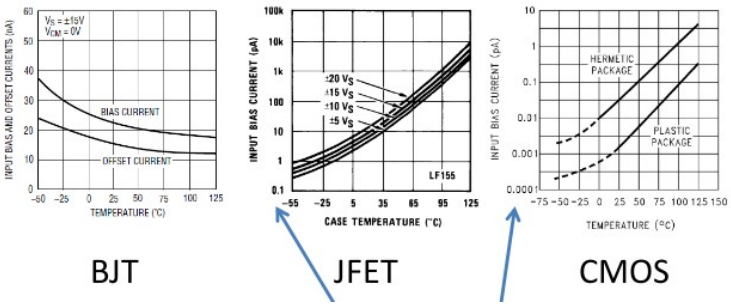


Figure 1.48: Bias current as a function of the temperature for a BJT (left), a JFET(centre) and a CMOS (right).

frequency, even though this can be a quite good approximation. In the right-hand side, we can observe that the output impedance tends to become smaller when we increase the output current of the device. These data come from a power operation amplifier, that is designed for high power applications, thus being able to tolerate quite high currents.

In Figure 1.51 we can observe how the common-mode rejection ratio and the power supply rejection ratio decrease with frequency. In general, this is not good, since it makes very challenging to be able to reject high frequency common-mode noise signals and power supply noise signals. In this case, the main check that we need to perform is that these two parameters are large enough in particular at 100 Hz (approximately), where we have the residual of the rectification of the sinusoidal power supply.

In Figure 1.52 we have the representation of the gain-bandwidth product (and in the first case, also of the phase margin) for three totally different operation amplifiers, that show a completely different behaviour as a function of the temperature. The one on the right-hand side, in particular, is an operation

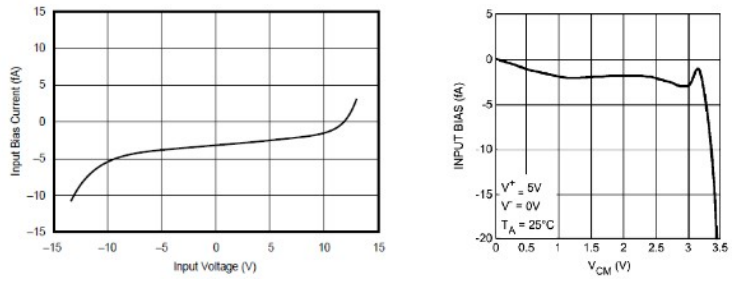


Figure 1.49: Input bias current as a function of the input voltage and of the common-mode voltage in a particular model.

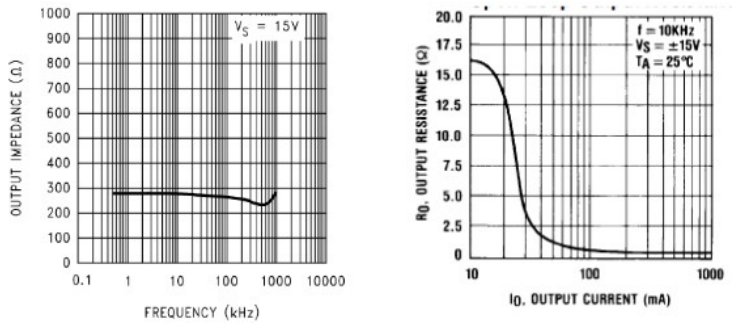


Figure 1.50: Output impedance as a function of the frequency and of the output current.

amplifier that works better when we have large loads.

In Figure 1.53 we can observe the temperature dependence and the differential input voltage dependence of the slew rate. In particular, we can observe that we have a difference between the positive slew rate, associated to a rise in the signal, and the negative slew rate, that is associated to a fall in the signal. In the bottom left picture, it is represented the effect of the finite slew rate on a square wave signal and, on the right, the same square wave gives a different output on a different operation amplifier, with a very peculiar initial infinite value of the slew rate, that after a while becomes finite.

In Figure 1.54, we can observe the peak-to-peak output voltage as a function of the frequency of the signal. For low frequencies, we are able to completely recover the signal up to the saturation value of the output voltage, thus giving the straight line behaviour on the left side of the graph. At a certain frequency, this behaviour changes: the frequency at which we have this change (represented in the Figure by the first arrow) is the so called full-power bandwidth. Further increasing the frequency, we need to limit the output swing of the signal, otherwise we will not be able to fully recover it. Since we must have:

$$2\pi f V_o = SR \;\Rightarrow\; V_o = \frac{SR}{2\pi f} \propto \frac{1}{f}$$

we can immediately explain the dependency of the maximum output voltage with respect to the frequency. In fact, further increasing the frequency we need

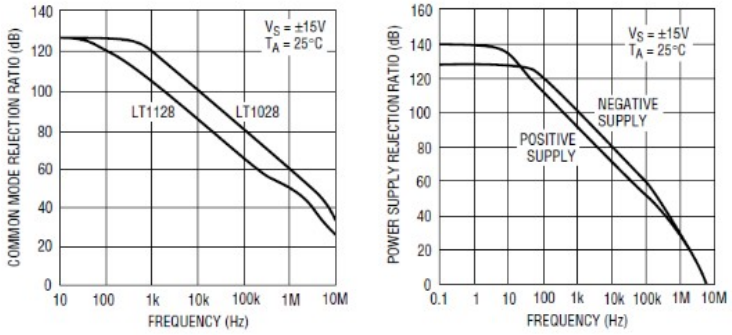


Figure 1.51: Common-mode rejection ratio and power supply rejection ratio as a function of the frequency.

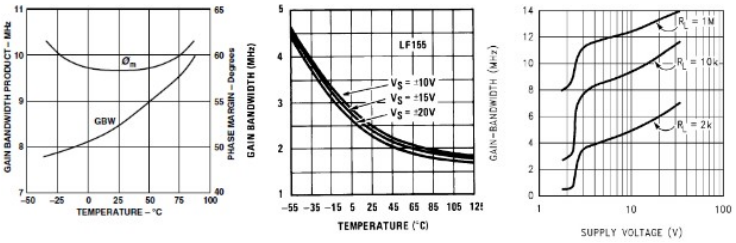


Figure 1.52: Gain-bandwidth product for three different operation amplifiers as a function of the temperature.

to decrease the output voltage in order to not have any distortion (as shown from the second arrow). It is important to note that this limitation is due to the fact that we are dealing with a single-pole network and that other operation amplifiers will have a full-power bandwidth even much smaller than this one.

In Figure 1.55 we can observe some typical outputs from step inputs and a possible graphical representation of the slew rate depending on the size of the step imposed. All these graphs are shown in the approximation of small-signal response.
As a conclusion, we can say that data-sheets provide an extensive characterization of the behaviour of an operation amplifier and that it is not necessary to consider all these information. In fact, we only need to check possible limiting factors depending on our application and using this as a guideline for choosing the best operation amplifier depending on our requisites.

1.10 Stability of the feedback loop

In the previous sections, we have seen that the gain of a real operation amplifier is always finite. We can now investigate the effects of this finite gain on the loop gain in feedback circuits.

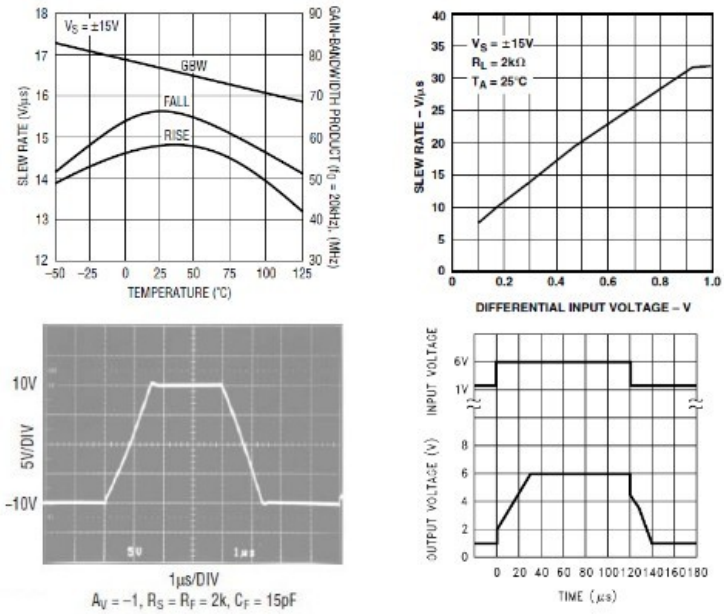


Figure 1.53: Above, slew rate as a function of the temperature and as a function of the differential input voltage; below, two effects of the finite slew rate.

1.10.1 Loop gain

Given the circuit in Figure 1.56, we can remember that, in this general circuit, the operation amplifier is generally hidden inside the open-loop gain G_{ol}. If the gain A of the operation amplifier is not infinite, also the gain G_{ol} of this element will not be infinite and therefore, as we will show, the loop gain will not be infinite. To calculate the loop gain, we need to set the input signal of the whole network to be zero. Then, we cut the feedback loop (for example as in Figure, but it is possible to use every other point of the loop) and we inject, in the sense of the feedback, a test signal S_{test}. We can then study the output signal coming from this input one and observe that it will be equal to:

$$S_{out} = -G_{ol}FS_{test} = G_{loop}$$

where we have defined the loop gain as:

$$G_{loop} = -G_{ol}F.$$

Note that the loop gain is a property of the loop and it is independent from the breaking point. Moreover, we can remember that the actual gain of the whole network with respect to the usual input will be:

$$G = \frac{G_{ol}}{1 - G_{loop}} = \frac{G_{id}}{1 - \frac{1}{G_{loop}}}$$

and it will be different from the ideal gain G_{id} if, due to the fact that we are dealing with a real operation amplifier, the open-loop gain (and consequently

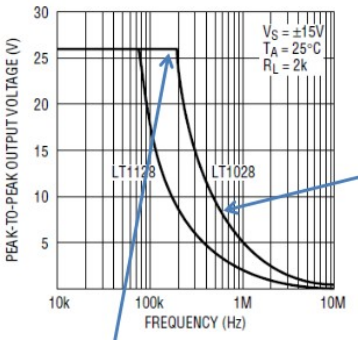


Figure 1.54: Peak-to-peak output voltage as a function of the frequency.

the loop gain) is not going to infinite. We can therefore define $1/G_{loop}$ as the error we are committing on the gain of the network with respect to the ideal one.

First of all, we can try to calculate the loop gain in a non-inverting amplifier, as shown in Figure 1.57. The starting point is to turn off the input, thus setting the positive pin of the operation amplifier to ground. Then, we can break the loop: note that in the Figure we have chosen two different points (in the left-hand side with respect to the right-hand side) for breaking the loop. Starting from the left-hand side network, that is simpler, we are directly applying a test signal V_T to the inverting pin of the operation amplifier. This gives, as an output of the operation amplifier:

$$A(V^+ - V^-) = -AV_T$$

and performing a voltage partition we are immediately able to calculate the output test signal:

$$V_o = -AV_T \frac{R_1}{R_1 + R_2}$$

and thus the loop gain of the network:

$$G_{loop} = \frac{V_o}{V_T} = -A\frac{R_1}{R_1 + R_2}.$$

Since the loop gain is a property of the network and it is independent from the breaking point, we should be able to retrieve the same result from the right-hand side network. In this second case, a voltage partition is needed to find the inverting pin of the amplifier:

$$V^- = V_T \frac{R_1}{R_1 + R_2}$$

and then, from the constitutive relationship of the operation amplifier:

$$V_o = A(V^+ - V^-) = -A\frac{R_1}{R_1 + R_2}V_T$$

we can obtain again the previous loop gain:

$$G_{loop} = -A\frac{R_1}{R_1 + R_2}.$$

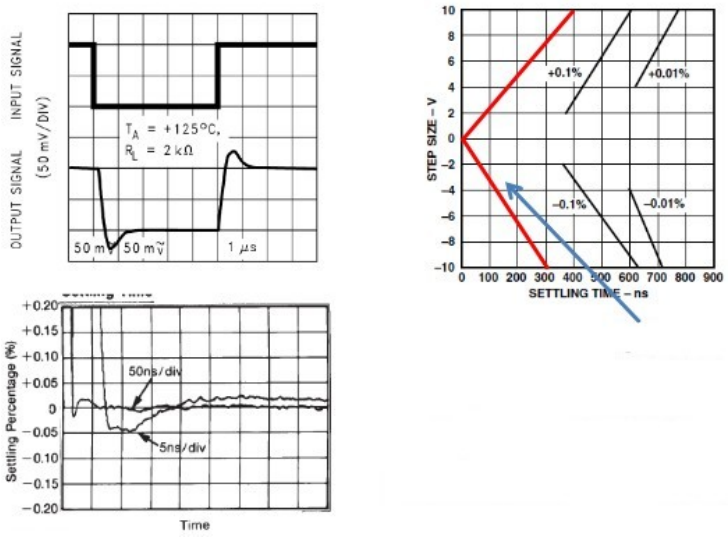


Figure 1.55: Top left corner: comparison between a square wave input signal and the output obtained; top right corner: dependency between the size of the step at the output and the time needed for covering the whole increment, or the 99%, or the 99.9%; bottom left corner: example of signal.

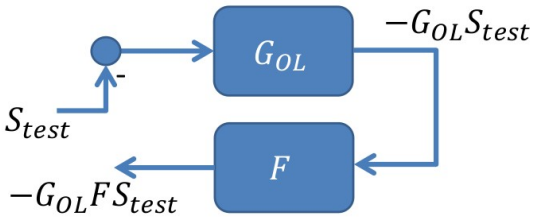


Figure 1.56: Calculation of the loop gain in a general circuit.

From the previous discussion, we can consider now a new breaking point, indicated in the right-hand side of Figure 1.58 by the word "cutline". Applying a test signal, for example a current I_T to this point, we can observe that this current will flow only through R_1 (the input impedance of the operation amplifier is infinite, we are removing one ideal characteristic of the operation amplifier at a time), thus giving the voltage of the inverting pin:

$$V^- = R_1 I_T = V_T$$

and, from the characteristic of the operation amplifier we obtain:

$$V_o = -AV_T$$

a different loop gain with respect to the previous case. This result is therefore not consistent with what we have said, so we must have committed some errors. To obtain the correct loop gain, we need to perform the so called impedance reconstruction. In fact, if the loop is closed, the same current is flowing through

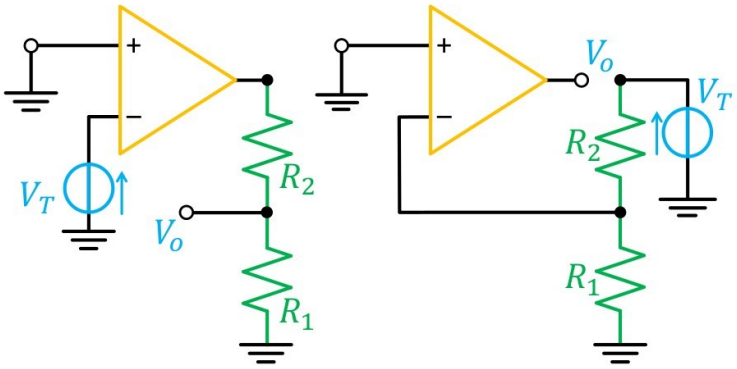


Figure 1.57: Calculation of the loop gain in a non-inverting amplifier.

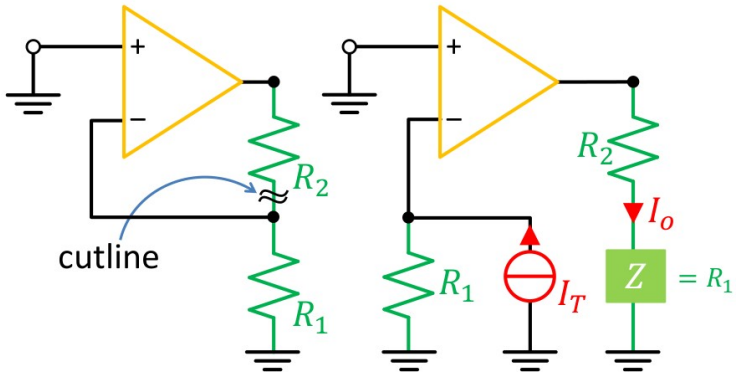


Figure 1.58: A different breaking point will lead to a different loop gain, therefore impedance reconstruction is needed.

the resistance R_1 and through the resistance R_2 due to the properties of the network. However, using this last cutting point, if we do not modify the network we will not have any current flowing through the resistance R_2 and this means that we are studying a different network: we have changed the working point of the loop. This means that when we analyse the loop gain we must ensure that we are not changing the working conditions and, in particular, that the resistive load of each node is identical before and after the cut. Therefore, every time we break a loop we must perform the so called impedance reconstruction, adding an impedance Z connected to the breaking point and assuming it to be equal to the impedance seen from the breaking point in the loop direction. It is extremely important that this reconstruction impedance is calculated along the forward direction of the loop and connected to the breaking point: every other change will lead to a wrong result. This impedance reconstruction has been performed in the right-hand side of Figure 1.58 and, from a brief analysis of the circuit, we can immediately observe that we obtain the correct result.
However, we are now able to go back to the examples in Figure 1.57 at page 54 and ask ourselves why we obtained the correct result even without the impedance reconstruction. To understand it, we can calculate the equivalent

impedance seen from the breaking point in the forward direction and observe that in the left-hand side network it will be infinite, therefore not affecting the circuit. This leads us to a quite important general rule: points with infinite impedance are always good points for breaking the loop, since there will not be any need for impedance reconstruction. In the right-hand side example, however, the reconstruction impedance can be calculated as $R_1 + R_2$. However, since the output voltage is connected to an ideal voltage source, that will fix a certain output voltage regardless of the following resistive load and therefore of the current needed, we can say that the output voltage will be independent from this reconstruction impedance and therefore neglect it. As a second rule, we can say that points connected to ideal voltage sources are good points for breaking the loop, since the effect of the reconstructed impedance to those points will be negligible. Last, we can note that we can assume as a test signal both a voltage or a current and the result, namely the gain loop, will not change.
Once we know the loop gain, we can easily calculate the closed-loop gain for that network as:

$$G = \frac{V_o}{V_i} = \frac{G_{id}}{1 - \frac{1}{G_{loop}}} = \frac{\frac{R_1+R_2}{R_1}}{1 + \frac{R_1+R_2}{AR_1}} = \frac{A}{1 + \frac{AR_1}{R_1+R_2}} = \frac{G_{ol}}{1 - G_{loop}}$$

where we have assumed the open-loop gain G_{ol}, for this circuit, to be equal to the gain of the operation amplifier A. We can then note that this result can be computed (and rearranged in this form) also by directly solving the network connecting the input V_i to the output V_o. However, this is not always easy, so it is generally better to study a network through the feedback theory; the willing student can try to retrieve the previous result from a direct inspection of the network.
Another, important observation that we can make is that if we consider a non-inverting amplifier and we set the input voltage to ground $V_i = 0$ we obtain exactly the same network of an inverting amplifier whose input is set to ground. This means that, for both networks, we will obtain the same loop gain G_{loop}. This leads to another important property: since the input is set at zero, the loop gain does not depend neither on the breaking point nor on the position of the input.

1.10.2 Open-loop gain

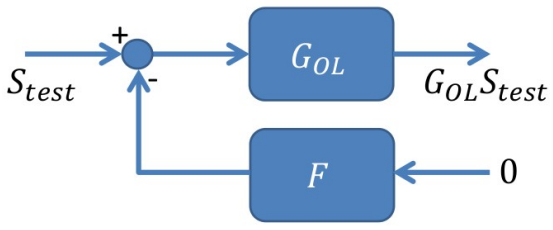


Figure 1.59: Calculation of the open-loop gain G_{ol} for a general network.

We can now study the open-loop gain G_{ol} of a certain network. The general case is represented in Figure 1.59. To calculate it, we break the feedback network

(in this case, just before the feedback element F) and we feed it with a zero-signal, while a test signal S_{test} is applied to the input of the network. Solving this network, we obtain that the output will be:

$$S_{out} = G_{ol} S_{test}.$$

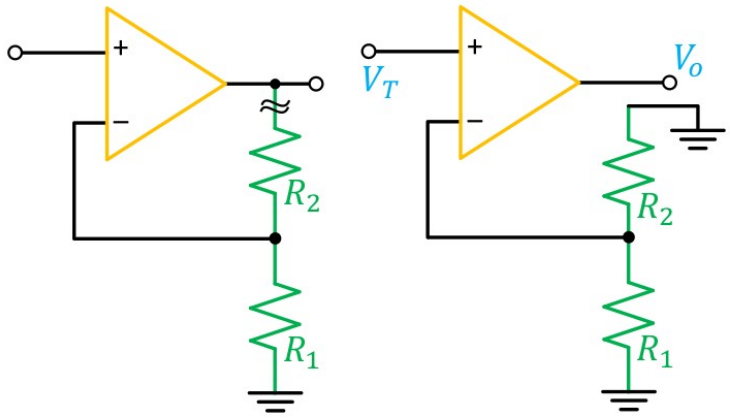


Figure 1.60: Calculation of the open-loop gain G_{ol} for a non-inverting amplifier.

To study the open-loop gain of a non-inverting amplifier, the first thing we need to do is to identify the feedback network. We can immediately observe that it is the network connecting the output to the inverting pin of the operation amplifier through the two resistances R_1 and R_2, since it is not directly related to the input (that is connected to the positive pin). We can therefore cut this network as in Figure 1.60 and ground it. This will make the negative pin of the operation amplifier to be grounded and, applying a test signal to the input, we obtain that the output will be:

$$V_o = A(V^+ - V^-) = A(V_T - 0) = AV_T$$

thus giving an open-loop gain equal to:

$$G_{ol} = \frac{V_o}{V_T} = A$$

that is exactly what we have previously assumed for this kind of circuit. It is important to note that also in this case we should have added a reconstructed impedance to the output of the operation amplifier but, since that node is connected to an ideal voltage source, this reconstructed impedance has been neglected.

The situation however is different if we have a non-zero output impedance R_o of the operation amplifier, as shown in Figure 1.61: in this case, in fact, the impedance reconstruction is crucial. In this case, in fact, the output pin is neither a point of infinite impedance nor it is directly connected to an ideal voltage source (due to the presence of this resistance). Analysing the network, since the output impedance is the series between R_1 and R_2, from a voltage partition we can then write:

$$V^- = 0,\ V_d = V_T \ \Rightarrow\ V_o = AV_d \frac{R_1 + R_2}{R_1 + R_2 + R_o} = A \frac{R_1 + R_2}{R_1 + R_2 + R_o} V_T = G_{ol} V_T$$

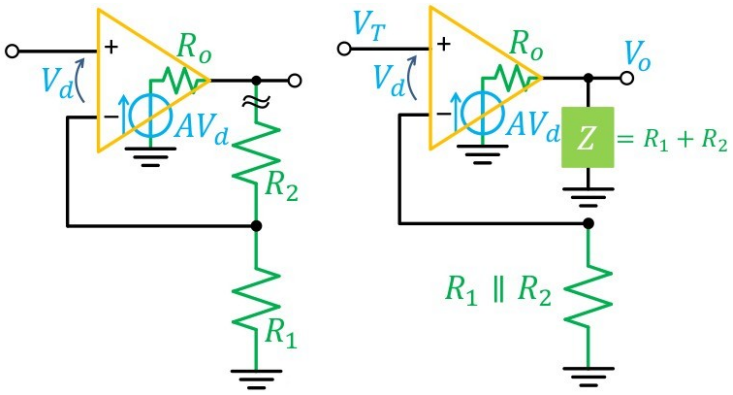


Figure 1.61: Calculation of the open-loop gain G_{ol} for a non-inverting amplifier: the presence of an output resistance makes the reconstructed impedance non-negligible.

thus obtaining the following open-loop gain[25]:

$$G_{ol} = A\frac{R_1 + R_2}{R_1 + R_2 + R_o}.$$

It is important to note that if the output resistance tends to the ideal case (thus being a short-circuit) also the open-loop gain will tend to its ideal value:

$$R_o \to 0 \quad \Rightarrow \quad G_{ol} \to A.$$

We can observe that, in this case, breaking the loop means that we have a zero-signal at the inverting pin of the operation amplifier. Therefore, from the topology of the network, an alternative breaking point is just before the negative pin of the operation amplifier this point will not need any impedance reconstruction and it will lead to the previous, correct result.

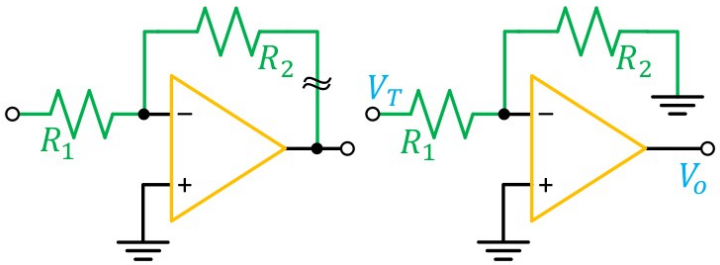


Figure 1.62: Calculation of the open-loop gain G_{ol} for an inverting amplifier.

A more complex case is the inverting amplifier, that is represented in Figure 1.62. In this network, it is more difficult to distinguish between input line and feedback line and, for example, cutting just before the inverting pin of the operation amplifier will remove both the input and the feedback signal. The

[25]The willing student is asked to find the loop gain G_{loop} for this kind of network.

best breaking point, in this case, is between the output voltage and the feedback resistor R_2 where, as we have said many times before, it is possible to neglect the reconstructed impedance since the output is connected to an ideal voltage source. Since the input impedance of the operation amplifier is infinite (again, remove one ideal characteristic at a time) we can write through a voltage partition:

$$V^- = V_T \frac{R_2}{R_1 + R_2} \;\rightarrow\; V_o = -A \frac{R_2}{R_1 + R_2} V_T$$

thus obtaining the following open-loop gain:

$$G_{ol} = \frac{V_o}{V_T} = -A \frac{R_2}{R_1 + R_2}$$

that is different from the one of the non-inverting amplifier. We can observe that this is consistent with the overall gain G that we expect from this network:

$$G = \frac{G_{id}}{1 - \frac{1}{G_{loop}}} = \frac{-\frac{R_2}{R_1}}{1 + \frac{R_1 + R_2}{A R_1}} = \frac{-\frac{A R_2}{R_1 + R_2}}{1 + \frac{A R_1}{R_1 + R_2}} = \\ = \frac{G_{ol}}{1 - G_{loop}}.$$

Remember that, in this expression, the crucial quantities are the ideal gain G_{id}, that gives a first approximation of the behaviour of the circuit, and the loop gain G_{loop}, that is a key property of the feedback system and must always be computed. Also the open-loop gain G_{ol}, then, can be computed, even though in general it is not strictly necessary, if we already have the ideal and the loop gain.
Also in the inverting case, then, the presence of a non-zero output resistance R_o inside the operation amplifier will make the calculation of the reconstructed impedance crucial. The willing student is invited to calculate how the open-loop gain is affected by this resistance.
As a final remark, we can stress the take-home messages:

- always remember to reconstruct the impedance when dealing with non-ideal cases;
- the open-loop gain can be calculated as:

  $$G_{ol} = -G_{loop} G_{id}$$

  whereas the other gain terms are related one with the other as:

  $$G_{loop} = -G_{ol} F,\; G_{id} = \frac{1}{F},\; G_{loop} = -\frac{G_{ol}}{G_{id}}$$

  and therefore we only need two of these three gain terms to determine the other one;
- the closed-loop gain may contain an additional term, that is called direct gain or feed-through gain, due to the direct transfer of the input through the feedback network, even though this term is usually small and negligible.

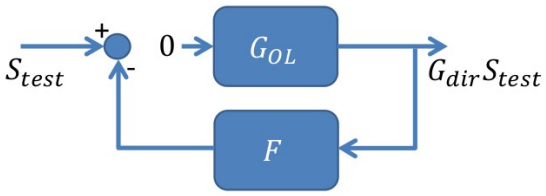


Figure 1.63: Calculation of the direct gain G_{dir} for a general network.

1.10.3 Direct gain

As we have defined it at the end of the previous section, the direct gain is the gain connected to a direct transfer of the input signal to the output one through the feedback network. For a general network, represented in Figure 1.63, we can assume to have directional elements inside the feedback network and cutting the input of the open-loop block and setting its signal to zero, due to the presence of these directional elements we will not have any transfer from the input to the output through the feedback networks. In real devices, however, the feedback network has not a preferred direction, being made of reciprocal components, therefore it is possible to have such a signal transfer. Therefore, we can compute this direct transfer from the input through the output via the feedback network by cutting the open-loop network at its input or at its output.

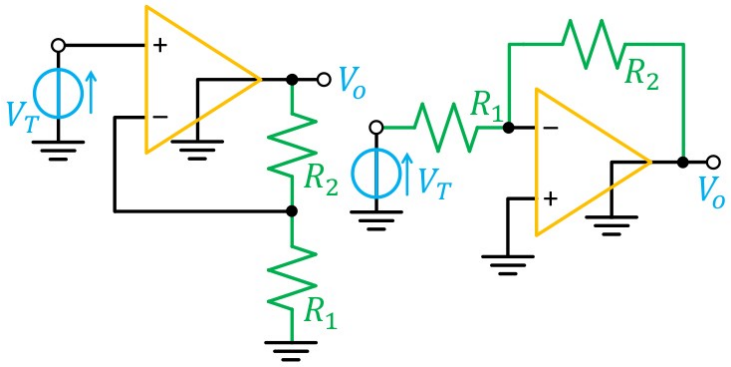


Figure 1.64: Calculation of the direct gain G_{dir} for a non-inverting (on the left) and an inverting (on the right) amplifier.

We can now investigate the direct gain, both for an inverting and a non-inverting amplifier, in the case of an ideal operation amplifier as shown in Figure 1.64. For the non-inverting amplifier, the resistors R_1 and R_2 are clearly part of the feedback loop and not of the open-loop gain, therefore this circuit is easier. For the inverting amplifier, things are more complicated since R_1 and R_2 belong both to the open-loop gain and to the feedback. In both cases, to cut the open-loop network, since we know that the only element that for sure belongs only to the open-loop network is the operation amplifier, we can set the ideal output voltage of the operation amplifier at ground. This will make the output voltage to be zero and, therefore, we obtain the predicted value for the direct gain:

$$G_{dir} = 0$$

in this ideal case. In real cases, however, it is not zero, but often it is so small that it can neglected.

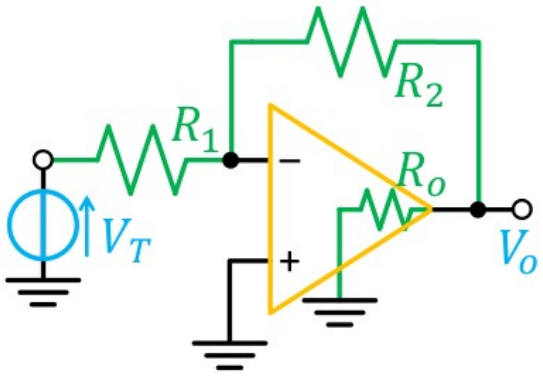


Figure 1.65: Calculation of the direct gain G_{dir} for an inverting amplifier with non-zero output impedance.

We can now calculate the direct gain for an inverting amplifier if we assume to be dealing with an operation amplifier with finite output resistance R_o. In this case, to disconnect the open-loop gain means, as in the previous one, that the ideal voltage source at the output is replaced by a short-circuit connected to the output resistance R_o:

$$A = 0.$$

In this case, since the output resistance is different from zero, we expect to have a direct gain that is different from zero and, by writing a voltage partition we can write:

$$V_o = V_T \frac{R_o}{R_o + R_1 + R_2} \;\Rightarrow\; G_{dir} = \frac{V_o}{V_T} = \frac{R_o}{R_o + R_1 + R_2}.$$

From this example, we can infer[26] that the overall gain of the network can be written as:

$$G = \frac{G_{ol} + G_{dir}}{1 - G_{loop}}.$$

In general, since the open-loop gain G_{ol} contains the gain of the operation amplifier A, that is a very large quantity, while the direct gain G_{dir}, from the expression obtained in this example, is a small quantity (in particular, lower than one), we can neglect the direct gain:

$$G_{dir} \ll G_{ol} \;\Rightarrow\; G \simeq \frac{G_{ol}}{1 - G_{loop}}$$

unless we are working at very high frequencies, where the open-loop gain drops. Moreover, it is important to note that solving by direct inspection the network we obtain an overall gain that contains, for free, also the contribution of the direct gain.

[26] This is not a rigorous demonstration, it is just an observation based on this example.

1.10.4 Input and output impedances

We can now discuss the effects, on the feedback system, of the presence of input and output impedances. As before, we will start making some examples and then we will somehow infer from them the general formula.

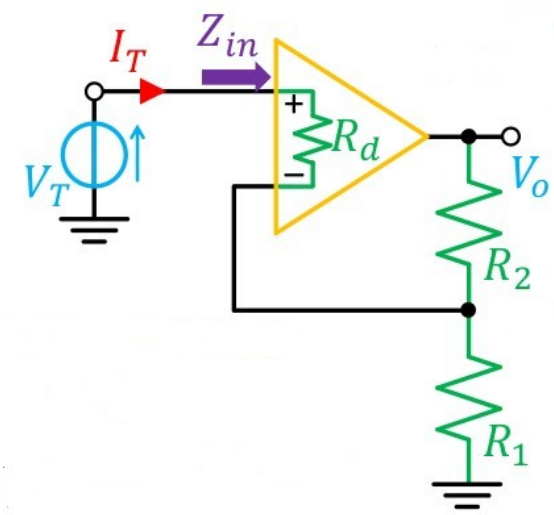


Figure 1.66: Non-inverting amplifier with a differential input impedance.

First, we can consider a non-inverting amplifier with a finite differential input impedance[27], represented in Figure 1.66. In an ideal case, the gain A of the operation amplifier is infinite, therefore it will make the two input pins to be equal and thus the voltage drop V_d across the differential resistance will be zero:

$$A \to \infty \quad \Rightarrow \quad V^+ = V^- \quad \Rightarrow \quad V_d = 0.$$

This means that, in ideal cases, the feedback not only stabilizes the gain but also improves the input impedance of the operation amplifier, leading it to be equal to the one we would have had in an ideal case, reducing the effect connected to the presence of the finite input impedance.
However, if the gain is finite, under the assumption of having a large differential impedance:

$$R_d \gg R_1,\ R_2$$

we can write the output voltage (neglecting the direct gain) as:

$$V_o = GV_T = \frac{G_{id}}{1 - \frac{1}{G_{loop}}} = \frac{R_1 + R_2}{R_1} \cdot \frac{V_T}{1 - \frac{1}{G_{loop}}}$$

and this allows us to write:

$$V^- \simeq V_o \cdot \frac{R_1}{R_1 + R_2} \simeq \frac{V_T}{1 - \frac{1}{G_{loop}}} = \frac{G_{loop}}{G_{loop} - 1} V_T$$

[27] In general, we neglect common mode impedances since:

$$R_c \gg R_d$$

and we neglect also differential capacities C_d.

where we have considered that the voltage of the inverting pin V^- is determined exclusively by the current flowing through R_1 and R_2 (from which the voltage partition) under the assumption of very small differential resistance. This allows us to calculate the test current I_T that will flow through the pins of the operation amplifier:

$$I_T = \frac{V^+ - V^-}{R_d} = \frac{V_T - V^-}{R_d} \simeq \frac{V_T}{R_d}\left[1 - \frac{G_{loop}}{G_{loop} - 1}\right] \simeq \frac{V_T}{R_d} \cdot \frac{-1}{G_{loop} - 1} = \\ = \frac{V_T}{R_d(1 - G_{loop})}$$

and from this we get the input impedance of this feedback network:

$$Z_{in} = \frac{V_T}{I_T} = R_d\,(1 - G_{loop})\,.$$

We can observe that, if the loop gain G_{loop} is equal to zero, therefore if the loop is open, the input impedance will be equal to the differential resistance R_d, that therefore is called the open-loop input impedance Z_{ol}. The input impedance, therefore, is varied with respect to its open-loop value of a quantity that depends on G_{loop}, that is a key parameter of the feedback system:

$$Z_{in} = Z_{ol}(1 - G_{loop}).$$

In an ideal case, as the one we have discussed before, if the loop gain tends to infinity also the input impedance tends to infinity, as in an ideal case, regardless of the value of the open-loop impedance; this is consistent with what we have discussed before.

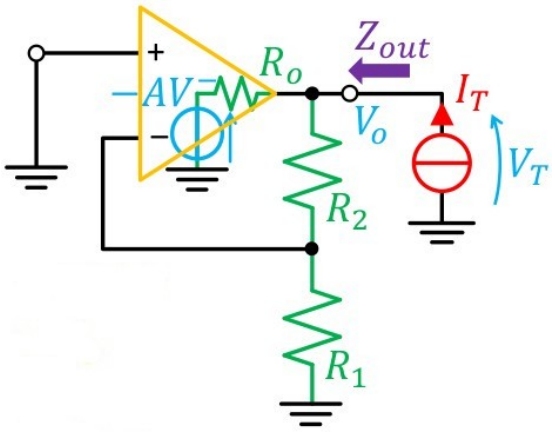


Figure 1.67: Non-inverting amplifier with an output impedance.

For the same non-inverting amplifier, is then possible to study the effect of a non-zero output impedance as in Figure 1.67. Also in this case, we can assume to not have any signal at the input and to apply a test signal at the output. This is a general procedure: assuming a load at the output, we can disconnect it from the input and observe the circuit from the perspective of the output

voltage V_o, disconnecting the load and calculating the output impedance across the load (that can be connected to ground or to another, different voltage). In this case, we will assume the output resistance R_o to be small with respect both to R_1 and R_2:

$$R_o \ll R_1,\ R_2.$$

Under this assumption, the test current I_T at the output will flow mainly through R_o, while the current flowing through R_1 and R_2 will be much smaller. From this, we can write:

$$V_o = V_T,\ R_o I_T \simeq V_T - A(V^+ - V^-) = V_T + AV^-$$

where we have that:

$$V^- = V_T \cdot \frac{R_1}{R_1 + R_2}.$$

Therefore, the current:

$$I_T \simeq \frac{V_T + AV^-}{R_o} = \frac{V_T}{R_o}\left(1 + A\frac{R_1}{R_1 + R_2}\right) \simeq \frac{V_T}{R_o}\left(1 - G_{loop}\right)$$

where we have assumed:

$$-G_{loop} = A\frac{R_1}{R_1 + R_2}.$$

This allows us to write the output impedance as:

$$Z_{out} = \frac{V_T}{I_T} \simeq \frac{R_o}{1 - G_{loop}} = \frac{Z_{ol}}{1 - G_{loop}}$$

and again we can observe that the presence of a feedback modifies the output impedance. Moreover, since R_o is the output impedance when the loop gain is equal to zero, thus for an open-loop condition, we can define an open-loop output impedance Z_{ol}. In an ideal case, the loop gain tends to be infinite and, therefore, the output impedance tends to zero:

$$G_{loop} \to \infty \quad \Rightarrow \quad Z_{out} \to 0$$

as in the case of an ideal operation amplifier.
We are now able to generalize this kind of reasoning for any given network. We only need to follow a procedure:

1. Compute the input Z_{in} or output Z_{out} impedance in the ideal case. In general, this is an easy task and, regardless of the fact that we are considering an input or an output impedance, we will obtain zero or an infinite value. Depending on what we have found as an ideal value, we can then add the effect of the loop gain.

2. Choose between the two forms:

   - if the ideal value is infinite:

$$Z_{ideal} \to \infty \quad \Rightarrow \quad Z = Z_{ol}\left(1 - G_{loop}\right);$$

- if the ideal value is zero:

$$Z_{ideal} = 0 \quad \Rightarrow \quad Z = \frac{Z_{ol}}{1 - G_{loop}}.$$

This allows us to be consistent with the fact that, in the ideal case, when the loop gain tends to infinite, we need to find the value calculated at the previous point.

3. Compute the missing terms, in particular the open-loop impedance Z_{ol} and the loop gain G_{loop}.

Considering again the example represented in Figure 1.66 at page 1.66, we can observe that if we set the loop gain equal to zero by imposing $A = 0$, we are actually grounding the output and, therefore, the open-loop input impedance is:

$$Z_{ol} = R_d + \frac{R_1 R_2}{R_1 + R_2}.$$

In the ideal case, since we are dealing with an input impedance, it will be infinite, therefore at the end we can write:

$$Z = Z_{ol}\,(1 - G_{loop})$$

where Z_{ol} is the value we have just calculated and where G_{loop} is the loop gain calculated considering also the presence of the differential resistance R_d.
In the example in Figure 1.67 at page 62, we can again impose the open-loop condition as $A = 0$, thus obtaining:

$$Z_{ol} = \frac{R_o(R_1 + R_2)}{R_o + R_1 + R_2}.$$

Since the ideal resistance in this case is zero:

$$Z_{id} = 0$$

the overall output resistance will then be:

$$Z = \frac{Z_{ol}}{1 - G_{loop}}$$

where, again, the loop gain G_{loop} is calculated taking into account the presence of the output resistance R_o.
It is important to remember that the feedback theory we are developing is just a way of studying this kind of networks. The same results (maybe a little messed up) can be obtained directly solving the network, even though this first method is usually simpler and clearer.
An alternative method is represented by Blackman's impedance formula:

$$Z = Z_{ol} \cdot \frac{1 - G_{loop}|_{sc}}{1 - G_{loop}|_{oc}}$$

where *oc* stands for open circuit and *sc* for short-circuit. The open-loop impedance can be calculated as in the previous case, while for *sc* we need to calculate

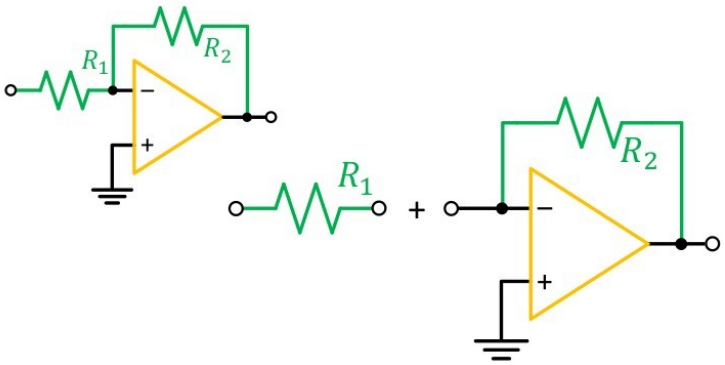


Figure 1.68: Calculation of the input impedance of an inverting amplifier.

the loop gain when the pin considered for the calculation of the impedance is short-circuited to ground and for *oc* it is left as an open circuit.

As a last example, we can study the case of an inverting amplifier, represented in Figure 1.68. In this case, we can calculate the ideal input impedance as:

$$V^+ = V^- \;\Rightarrow\; Z_{id} = R_1$$

and we can observe that it is nor zero neither an infinite value. This is due to the fact that attaching an element in series or in parallel to the real impedance of the network, we will observe as an ideal impedance the series or the parallel of this element and of the actual ideal impedance. In this case, we can immediately observe that R_1 is in series with the actual input impedance, so we can disconnect it and apply the previous procedure to the remaining part, assuming an ideal input impedance equal to zero:

$$Z'_{id} = 0.$$

This gives:

$$Z' = \frac{Z_{ol}}{1 - G_{loop}}$$

where the open-loop impedance Z_{ol} is calculated as usual, thus obtaining an overall input impedance of:

$$Z = R_1 + \frac{Z_{ol}}{1 - G_{loop}}.$$

1.11 Frequency behaviour, system stability and pole compensation

1.11.1 Frequency response of feedback amplifiers

In the frequency domain, all the properties that we have seen to be valid in the previous section will hold as functions of the incoming frequency. In the case of

the loop gain, for example, in the Laplace domain, defining the Laplace operator s, we can obtain:

$$G_{loop}(s) = -\frac{G_{ol}(s)}{G_{id}(s)}$$

that in a magnitude diagram, where the scale is in decibels[28], gives:

$$|G_{loop}(s)|_{\text{dB}} = |G_{ol}(s)|_{\text{dB}} - |G_{id}(s)|_{\text{dB}} \,.$$

In the same way, we can define the gain of a given feedback network as:

$$G(s) = \frac{G_{ol}(s)}{1 - G_{loop}(s)} = \frac{G_{id}(s)}{1 - \frac{1}{G_{loop}(s)}}$$

and also in this case we can distinguish between a strong loop condition and a weak loop condition:

$$\begin{cases} G_{loop} \gg 1 & \Rightarrow \quad G \simeq G_{id} \\ G_{loop} \ll 1 & \Rightarrow \quad G \simeq G_{ol} \end{cases} .$$

The main quantity, therefore, is again the loop gain $G_{loop}(s)$, that now will be dependent on the frequency. These quantities can then be represented in Bode plots as in Figure 1.69.

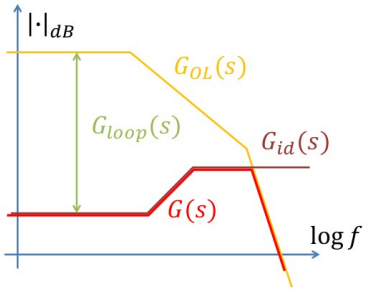


Figure 1.69: Example of Bode plot of the magnitude for the open-loop gain and for the ideal one. From them, it is then determined the loop gain and the gain of the network.

In this Figure, we have plot a possible frequency behaviour of the open-loop gain $G_{ol}(s)$ and of the ideal gain $G_{id}(s)$. From the previous relationship, then, we can observe that, since the unit of measurement of the vertical axis is the decibel, we can determine from the graph the loop gain $G_{loop}(s)$ at each frequency just by calculating the difference between the open-loop gain and the ideal gain at that frequency. It is important to note that in the point where the open-loop gain and the ideal gain are identical, the value in decibels of the loop gain is zero, thus being one in usual units:

$$|G_{ol}|_{\text{dB}} = |G_{id}|_{\text{dB}} \;\Rightarrow\; |G_{loop}|_{\text{dB}} = 0 \;\Rightarrow\; G_{loop} = 1.$$

[28] We remind that:

$$G_{loop}(s)\big|_{\text{dB}} = 10 \log_{10}(G_{loop}(s)).$$

Beyond that point, then, the loop gain will be lower than one. For the closed loop gain, therefore, we can define observe the following asymptotic behaviour:

$$G \simeq \begin{cases} G_{id}, & \text{for } G_{loop} \gg 1 \\ G_{ol}, & \text{for } G_{loop} \ll 1 \end{cases}$$

as in Figure 1.69 and, as a rule of thumb, we can extend it in order to have a piecewise continuous behaviour:

$$G = \begin{cases} G_{id}, & \text{for } G_{loop} > 1 \\ G_{ol}, & \text{for } G_{loop} < 1 \end{cases}.$$

This means that the gain of the network will always be equal to the minimum value between the ideal and the loop gain:

$$G \simeq \min(G_{id},\ G_{loop}).$$

Obviously, these are called asymptotic Bode plots and, drawing them, we are committing some errors due to some approximations.

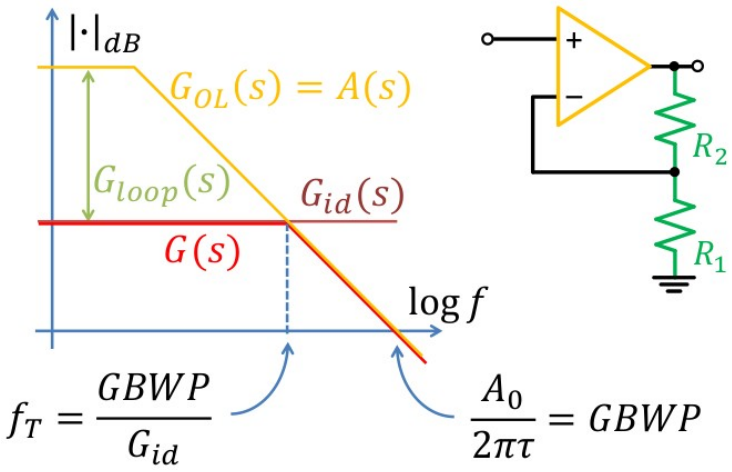


Figure 1.70: On the left, Bode plot of the magnitude of the involved gains; on the right, network that we are considering.

The first application of these plots is to a single-pole amplifier, represented in the right-hand side of Figure 1.70. In fact, it is important to remember that the vast majority of the operation amplifiers will have the following transfer function between input and output:

$$V_o = A(V^+ - V^-), \quad A(s) = \frac{A_0}{1 + s\tau}.$$

Therefore, calculating the open-loop gain, we can cut the loop at the output (as we have done in the previous section), obtaining:

$$G_{ol} = A(s)$$

that is represented in Figure 1.70. The ideal gain G_{id}, on the other hand, has been studied many times and it is constant and independent from the frequency. This allows us to study the loop gain $G_{loop}(s)$ and to draw the gain $G(s)$. The crossing point at which the gain $G(s)$ pass from being equal to the ideal gain to being equal to the open-loop gain is the frequency represented in Figure as f_T. Since we know that the frequency at which the open-loop gain crosses the horizontal axis is called gain-bandwidth product $GBWP$ and it is a characteristic of our device:

$$\frac{A_0}{2\pi\tau} = GBWP$$

we can write the frequency f_T in the Figure as[29]:

$$f_T = \frac{GBWP}{G_{id}}$$

either from a direct calculation involving the expression of the open-loop gain and the value of the ideal gain or by studying the graph. From the expression of the loop gain of the network considered:

$$G_{loop} = \frac{A(s)R_1}{R_1 + R_2} = \frac{A_0}{1 + s\tau} \cdot \frac{R_1}{R_1 + R_2}$$

we can write the gain of the closed-loop network:

$$G = \frac{G_{id}}{1 - \frac{1}{G_{loop}(s)}} = \frac{G_{id}}{1 + \frac{R_1 + R_2}{A(s)R_1}}$$

and in the denominator we will have a pole determined by the frequency behaviour of the operation amplifier. This pole can be found by solving the polynomial at the denominator:

$$\frac{(R_1 + R_2)(1 + s\tau)}{A_0 R_1} = -1$$

thus giving:

$$s = -\frac{1}{\tau}\left(1 + \frac{A_0 R_1}{R_1 + R_2}\right) = -\frac{1}{\tau}\left(1 - G_{loop}(0)\right)$$

and the frequency of the pole is:

$$f_p = \frac{1}{2\pi\tau}\left(1 + \frac{A_0 R_1}{R_1 + R_2}\right) = \frac{1}{2\pi\tau}\left(1 - G_{loop}(0)\right).$$

It is possible to note that $1/(2\pi\tau)$ is the frequency of the pole of the open-loop network, when:

$$G_{loop}(0) \to 0$$

and therefore it is also called open-loop pole.
We can observe that, without any feedback, the gain is just the open-loop gain

[29] Always remember that when we are dealing with a line with slope $-n \cdot 20$ dB/dec the gain A_1 at a certain frequency f_1 will become, at another frequency f_2 along the same line, such that the following relationship is satisfied:

$$A_1 \cdot f_1^n = A_2 \cdot f_2^n.$$

and the only pole present is the open-loop one. The presence of the feedback network, then, will change the position of this pole, as seen in Figure 1.70, and therefore also the bandwidth of the network that we are considering; in general, this is a good property and we will use it. Since we can observe that:

$$\frac{A_0 R_1}{R_1 + R_2} \gg 1$$

we can in general write that, for a non-inverting amplifier as the considered one:

$$f_p \simeq \frac{A_0}{2\pi\tau}\frac{R_1}{R_1 + R_2} = -\frac{1}{2\pi\tau}\frac{A_0}{G_{id}}$$

and inverting this relationship we obtain:

$$|f_p|\frac{R_1 + R_2}{R_1} = |f_p|G_{id} = \frac{A_0}{2\pi\tau} = GBWP.$$

This makes clearer the meaning of the gain-bandwidth product: it is a constant factor that relates the gain and the position of the pole. If the absolute value of the frequency of the pole increases, then the ideal gain of the network will diminish, while if the absolute value of the frequency of the pole decreases, the ideal gain of the bandwidth increases. Since the minimum value of the ideal gain is one (corresponding to zero decibels), then the maximum bandwidth is given by $GBWP$, as we have seen in the Bode diagram in Figure 1.70.
In an inverting configuration, the same relationship will lead to:

$$|f_p|(1 + |G_{id}|) = GBWP$$

where the factor multiplying the frequency of the pole is different from the ideal gain of the network. At the end, therefore, the feedback loop reduces the open-loop gain by a factor of $1 - G_{loop}(0)$ and widens the bandwidth of an identical factor.
However, the fact that the position of a pole in a network is changed depending on the ideal gain of the network can, in principle, lead to instabilities even if the open-loop system is stable. We need, therefore, to discuss the stability of a feedback system.

1.11.2 Stability of feedback amplifiers

Given a generic feedback system as the one represented in Figure 1.71, it is possible to demonstrate a fundamental properties of feedback systems: the stability of the closed-loop system only depends on the loop gain G_{loop}. An alternative possibility, obviously, is to study directly the gain G of the network but, since we have this property, it is also possible to demonstrate that the critical condition for having instability is:

$$G_{loop} = 1$$

or, better:

$$-G_{loop} = -1.$$

This can be done by studying the Nyquist criterion, but this kind of reasoning is quite complicated and not very useful from our perspective.
The main tool that we will use to study the stability of a feedback system is the Bode stability criterion (1945). It states that if:

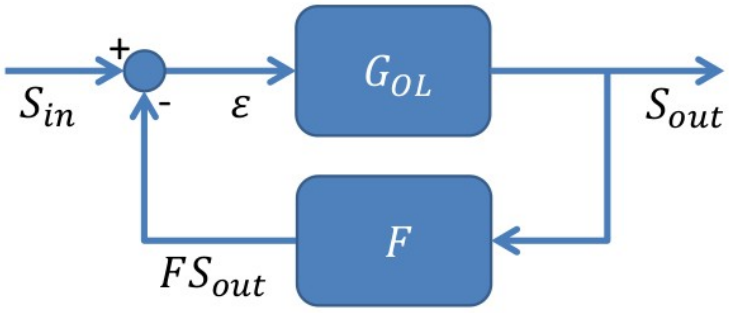


Figure 1.71: A generic feedback system.

- $G_{loop}(s)$ has only one pole in the left-hand side plane (therefore, with real part negative) or in the origin ($s = 0$);
- there is only one critical frequency $f_{0\ \mathrm{dB}}$ where the magnitude of $-G_{loop}$ is equal to 0 dB, thus crossing the horizontal axis;
- the phase of the loop gain at that frequency is such that:

$$\angle\left(-G_{loop}(f_{0\ \mathrm{dB}})\right) < 180^\circ$$

  therefore,we have a positive phase margin;

then the system is stable.
It is important to observe that the previous condition are sufficient for determining the stability of the system, but they are not necessary: if the hypotheses of this theorem are not satisfied, we cannot say that the system is unstable. Moreover, this first formulation involves the requirement of having just one crossing of the horizontal axis.
A second formulation, on the other, sets the requirement of having just one cross of the 180° phase in the phase diagram. This alternative formulation states that if:

- $G_{loop}(s)$ has only one pole in the left-hand side plane (therefore, with real part negative) or in the origin ($s = 0$);
- there is only one frequency f_{180° where the phase of $-G_{loop}$ is 180° (plus or minus integer multiples of 360°);
- at that frequency:

$$|G_{loop}(f_{180^\circ})| < 1;$$

then the system is stable.
Considering what we have discussed in the previous section regarding the gain-bandwidth product, we can immediately observe that a change in the zero-frequency gain A_0 of the operation amplifier can possibly rise or lower the magnitude Bode diagram, thus changing the critical frequency and possibly affecting the stability property of such a system. The gain margin, defined as:

$$G_m = \frac{1}{|G_{loop}(f_{180^\circ})|}, \quad G_m|_{\mathrm{dB}} = -\,G_{loop}(f_{180^\circ})|_{\mathrm{dB}}$$

therefore represent how much the gain can change without affecting the stability property of the system. On the other hand, a fluctuation in the position of a pole can affect the stability of the system as well. We can therefore define the phase margin as the maximum variation of the phase that it is possible to face without loosing the stability property of the system:

$$\phi_m = 180^\circ + \angle\left[-G_{loop}(f_{0\ \mathrm{dB}})\right].$$

These parameters are extremely important in real systems, where the transfer functions are subjected to some tolerances due to the fact that the systems are realized with real components, significantly different from ideal ones.

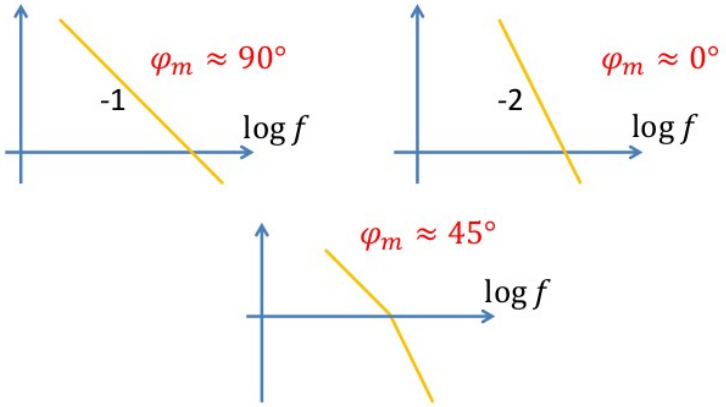


Figure 1.72: Simplified version of the Bode criterion.

A simplified version of the Bode criterion can then be adopted when the real part of every zero and pole is negative. In this case, in fact, every pole will determine a phase shift of $-90°$ and every zero of $90°$, thus making possible to reconstruct the phase Bode diagram just from the magnitude one. Systems satisfying this hypothesis are called minimum phase systems. This simplified Bode criterion, therefore, states that if all the poles and zeros of the transfer function are in the left-hand side plane, it is possible to infer the stability of a system from its Bode plot and, in particular:

- if we are cutting the horizontal axis, in the magnitude Bode diagram, with a slope of -20 dB/dec, the system is stable and the phase margin is $\phi_m \simeq 90°$;
- if we are cutting the horizontal axis, in the magnitude Bode diagram, with a slope of -40 dB/dec or more inclined, the system is unstable and the phase margin will be zero degrees or even negative;
- if the point at which we are cutting the horizontal axis in the magnitude Bode diagram is exactly a pole where the slope passes from -20 dB/dec to -40 dB/dec, then the system is stable and the phase margin is approximately $45°$.

These properties can be demonstrated by studying the Bode diagrams associated to minimum phase systems in these conditions.

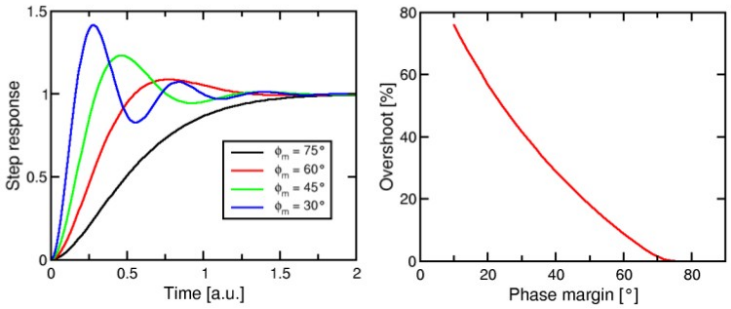


Figure 1.73: On the left, temporal step response of a system for different phase margins; on the right, overshoot on the step response as a function of the phase margin.

We can then study the step response of a system and observe that the phase margin influences the overshoot in this step response, even though the system is stable. Therefore, what is an acceptable value for the phase margin? It depends on the overshoot that our system can face. Another possible problem is the settling time, that can be influenced by the phase margin. In general, in our cases, we will be satisfied with a phase margin of:

$$\phi_m \simeq 45°.$$

However, in practical application this may not be enough.
At this point, a problem may arise: what can we done if the phase margin of a certain system is not enough? This problem introduces us to the next topic, the frequency compensation, through a suitable tailoring of the loop gain G_{loop}.

1.11.3 Compensation

As we have said at the end of the previous section, the frequency compensation of an operation amplifier is the tailoring of the loop gain $G_{loop}(s)$ in order to improve the circuit stability. Most of the operation amplifiers at our disposal are in general "internally compensated" in order to be easier to use with a resistive feedback. This means that the manufacturer modifies them in order to have just a single pole when the magnitude of the gain is above 0 dB. However, when we have frequency dependent feedback systems, we need to check the stability of the system and, eventually, compensate it to obtain a stable system.

A first example of compensation technique is the so called dominant pole compensation. Consider, for example, the original magnitude Bode diagram of the loop gain represented as a dashed line in Figure 1.74. This gain is cutting the 0 dB axis with a very high slope, thus the phase margin of this system will be very low or even negative. To solve this problem, it is possible to introduce an additional pole, whose frequency is indicated with a cross on the frequency axis in Figure, at a very low frequency. This additional pole, being at very low frequency, will be the first pole of the loop gain function and, therefore, will make the magnitude diagram of the loop gain cross the 0 dB axis with a better slope, thus increasing the phase margin that, in the example in Figure, will be about 45°. This kind of compensation technique is generally adopted in

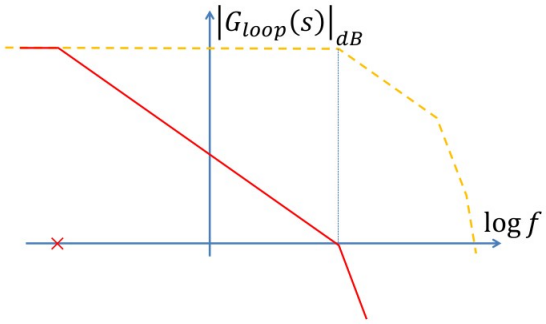


Figure 1.74: An example of dominant pole compensation.

every kind of "internally compensated" operation amplifier. In fact, every real operation amplifier will have a lot of poles[30] and to compensate them, we can introduce a capacitor, represented in Figure 1.43 at page 44, that will control the slew rate and that will be at such a low frequency that we can assume, for this kind of device, a single-pole transfer function[31]. However, this is not the only possibility: there exist also uncompensated operation amplifiers, where poles are not compensated, that will be described, for example, by a transfer function with two poles.

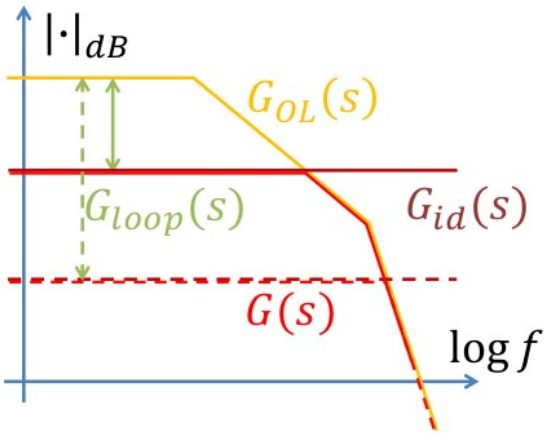


Figure 1.75: Example of Bode diagram of the magnitude of the gains for an uncompensated operation amplifier.

In Figure 1.75 it is represented an example of Bode diagram of the gains for an uncompensated operation amplifier. We can immediately note that, when the ideal gain is represented by the solid line in Figure, the gain G will probably be such that the system is unstable, while for a lower ideal gain, described by the dashed line, the loop gain is higher and the gain G will probably describe

[30] In general, every capacity of a certain system will add a pole and, in an operation amplifier, we will have a lot of transistors and capacitors, each one of them adding one or more poles.

[31] In fact, every other pole will play a role only at high enough frequency, where the gain is well lower than the 0 dB axis.

a stable system. In general, uncompensated operation amplifiers can achieve better performances, but they must be carefully tailored on the specific circuit. We will almost always deal with compensated operation amplifiers, due to the higher difficulties involved in the design of circuits for uncompensated operation amplifiers.

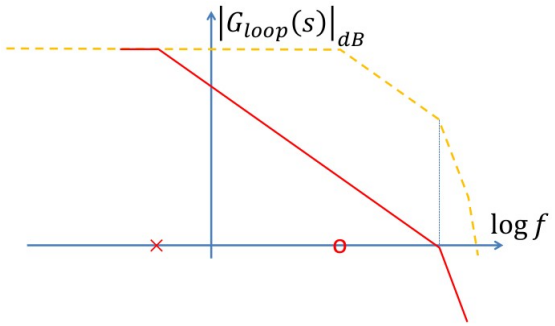


Figure 1.76: Example of Bode diagram of the magnitude for a pole-zero compensation.

A different kind of compensation technique is the so called pole-zero compensation, represented in Figure 1.76. In this kind of technique, the additional, low-frequency pole (represented with a cross in Figure) is not enough to obtain a stable system. Therefore, we need to include an additional zero (represented with a circle in Figure) at exactly the same frequency of one of the poles of the original loop gain function. In the given example, the additional zero cancels out a pole and therefore allows the loop gain function to cut the horizontal axis with a slope that ensure an high enough phase margin (while this was not possible without the additional zero). To add a zero in the loop gain function, the only possibility is to act on the feedback network.
From the previous descriptions, it should be intuitive that there is not a definitive path or technique to obtain a stable system: every solution is different and it has its own strong and weak points.

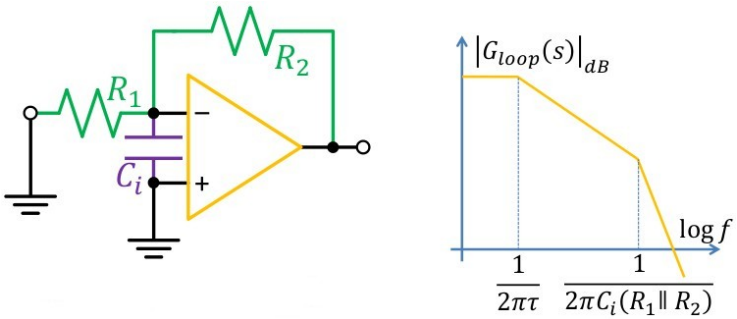


Figure 1.77: Inverting amplifier with an input capacitance and the associated Bode diagram for the magnitude of the loop gain.

A first example of system in need of pole compensation is an inverting amplifier with an input capacitance. Since we want to calculate the loop gain of

this network, we can ground the input, cut the loop between the resistance R_2 and the output of the operation amplifier and add, in that point, a test voltage source. In general, we have that at the input of the operation amplifier there is only a differential resistance R_d that is much higher than R_1 and R_2 and, therefore, it can be neglected. At high frequencies, however, a small input capacitance inside the operation amplifier, indicated with C_i in Figure, becomes important. Observing that, in this network, the resistance R_1 and the capacity C_i are in parallel (both being between V^- and $V^+ = 0$, we can replace them with a complex impedance Z connected to the two input pins of the operation amplifier. This complex impedance will be:

$$Z = \frac{R_1}{R_1 + sC_iR_1}$$

therefore the output voltage will be:

$$V_o = -A(s)V^- = -A(s)\frac{Z}{Z + R_2}V_T$$

thus giving the following loop gain:

$$G_{loop} = -A(s)\frac{Z}{Z + R_2} = -A(s)\frac{R_1}{R_1 + R_2} \cdot \frac{1}{1 + sC_i(R_1 \| R_2)}.$$

We can therefore observe that every time we add a reactive element (in general capacitors and inductors, even though we will deal only with capacitors) to a network, we are also introducing a new pole, where the time constant will be:

$$\tau = CR_{eq}$$

where C is the capacity introduced and R_{eq} is the equivalent resistance seen across the capacitor (in this case, the parallel between R_1 and R_2).

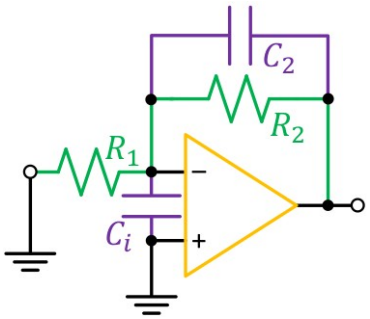


Figure 1.78: Inverting amplifier with an input capacitance and a compensation capacitance.

The previous circuit, since the slope of the loop gain at the crossover frequency is -2, has some issues regarding stability. Therefore, we need to add a compensation capacitance, as in Figure 1.78. In fact, since every capacitor in general generates a pole in the transfer function, without the additional capacitance C_2 we have two poles (one given from the input capacitance C_i, the other

coming from the operation amplifier) and the system is probably unstable. The two poles are placed in:

$$s = -\frac{1}{\tau}, \quad s = -\frac{1}{C_i(R_1 \| R_2)}.$$

Therefore, neglecting the presence of C_2, we obtain the dashed behaviour represented in Figure 1.79. The phase margin, in this case, is clearly lower than 45°, therefore we have a stability issue.

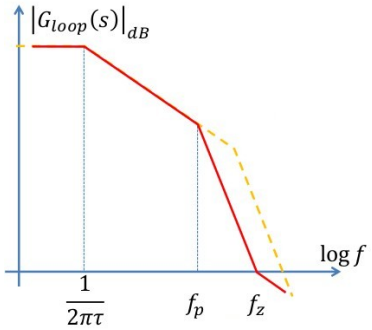


Figure 1.79: Bode diagram of the magnitude of the loop transfer function with or without compensation capacitance.

To compensate the stability problem of the previous circuit, we can add a capacitor C_2 as represented in Figure 1.78 and we will find out that this new capacitor will a zero to the transfer function. Defining the following two complex impedances:

$$Z_1 = C_i \| R_1, \quad Z_2 = C_2 \| R_2$$

we therefore obtain the following loop gain:

$$G_{loop} = -A(s)\frac{Z_1}{Z_1 + Z_2} = -A(s)\frac{\frac{R_1}{1+sC_iR_1}}{\frac{R_1}{1+sC_iR_1} + \frac{R_2}{1+sC_2R_2}} =$$
$$= -A(s)\frac{R_1}{R_1 + R_2}\frac{1 + sC_2R_2}{1 + s(C_i + C_2)\cdot(R_1 \| R_2)}.$$

Therefore, we have clearly added a zero at the following frequency:

$$f_z = \frac{1}{2\pi C_2 R_2}$$

while the poles will be at the following frequencies:

$$f_{p1} = \frac{1}{2\pi\tau}, \quad f_{p2} = \frac{1}{2\pi(C_i + C_2)(R_1 \| R_2)}$$

thus it is possible to note that the zero will be at a frequency higher than the one of the two poles:

$$f_z > f_{p2} > f_{p1}.$$

Adding this zero, properly selecting the compensation capacitance C_2 it is possible to place it exactly at the crossover frequency, changing the slope from -2

to -1 and increasing the phase margin. Note that now we have two different capacitors C_i and C_2 but we do not have two different poles (one from each of them, apart from the one of the operation amplifier): this is due to the fact that the two capacitors are not independent. In electronics, two elements are said to be independent when they are neither in parallel nor in series. This can be clearly observed by calculating the equivalent resistance of C_i or C_2 when calculating the loop gain: both, in fact, will have one end connected to ground and the other connected to the inverting pin of the operation amplifier.
The additional capacitance, that gives the stability of the circuit, however, has also a drawback. In fact, the capacitance C_2 modifies the closed-loop gain, reducing the bandwidth of the circuit, that now is limited by the frequency of the zero f_z. In general, in fact, we have that the resistance R_2 is higher than R_1 in order to have an amplifier. In particular, we can observe that:

$$R_1 \ll R_2 \;\Rightarrow\; R_1 \| R_2 \simeq R_1$$

while, in general, compensation capacitance are extremely low:

$$C_i \gg C_2.$$

This gives the fact that:

$$f_{p2} < f_z$$

because the frequency of the pole is almost unchanged with respect to the uncompensated case, while the zero will clearly be at an higher frequency. The idea, in particular, is to choose a compensation capacitance C_2 small enough to have the frequency of the zero f_z in a place suitable for giving an effective compensation (otherwise, it will be beyond the crossover frequency). In the design of a circuit, therefore, we are tailoring the position of the zero f_z depending on the desired phase margin. We are, somehow, trading off the stability of the circuit against its bandwidth and this can be studied by calculating the ideal gain of the compensated circuit represented in Figure 1.78. Considering for example a signal V_i applied to the first resistance R_1, due to the effect of the negative feedback, considering an ideal operation amplifier, we will have that:

$$V^+ = V^- = 0$$

therefore the presence of the input capacitance C_i can be neglected in the calculation of the ideal gain. Defining the following complex impedance:

$$Z_2 = C_2 \| R_2$$

we obtain the following ideal gain:

$$G_{id} = -\frac{Z_2}{R_1} = -\frac{R_2}{R_1} \cdot \frac{1}{1 + sC_2R_2}$$

and this clearly show that the ideal gain, now, has a certain bandwidth that is limited by the frequency of the zero of the loop transfer function. We are thus trading off between the stability (in particular, the phase margin) of the loop gain (and thus of the whole closed-loop system) and the bandwidth of the ideal gain.
A possible, alternative choice is to have:

$$C_2R_2 = C_iR_1 \;\Rightarrow\; f_{p2} = f_z$$

obtaining what is generally called a pole-zero cancellation. However, this is usually difficult to obtain, since in real devices the input capacitance C_i is never constant[32].
Last, in differential amplifiers another possibility is to use a symmetric compensation. From an ideal point of view, in fact, anything that is placed between the positive and the negative input pins of the operation amplifier should not affect the ideal gain of the device, since the voltage drop across these elements is identically equal to zero (when the loop is closed and the operation amplifier is considered as an ideal element). This observation leads us to put compensating elements between these two nodes, thus adding a zero[33] at a frequency higher than the one of the pole given by the capacity. This kind of circuit is represented in Figure 1.80 and it is called lag network.

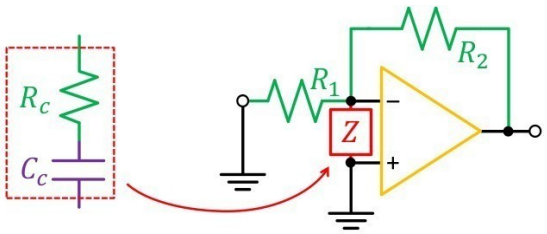


Figure 1.80: A lag network.

Studying the loop gain of this network, it is possible to calculate that:

$$G_{loop} = -A(s)\frac{R_1}{R_1 + R_2} \cdot \frac{1 + sC_C R_C}{1 + sC_C(R_C + R_1 \| R_2)}$$

after the definition of the complex impedance Z as the series between the resistance R_C and the capacitance C_C. We can observe that, from the previous reasoning, this network will not affect the ideal gain G_{id}, but it will possibly degrade the input impedance Z_{in} of the operation amplifier when we are dealing with non-inverting amplifiers. We can define a network in which:

$$f_p < f_z$$

a lag network, since we will first loose a certain amount of phase that will be the gained again later. On the contrary, if we have a network in which:

$$f_z < f_p$$

it will be called a lead network, since we will first gain some phase and then we will loose it.
Another circuit that whose compensation can be studied is the differentiator, first analysed at page 25, to which we can add an input capacitance C_i as in Figure 1.81.

[32] We have seen in a previous section that the parameters of a real operation amplifier changes significantly depending on several variables.

[33] In general, there is not any clever rule for finding out the position of a zero by studying the topology of a network. The only way of finding it is to explicitly calculate the loop gain of the whole network.

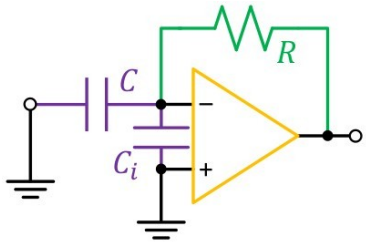


Figure 1.81: A differentiator with an input capacitance.

As we have seen in the related exercise, this circuit has stability problems also without the addition of the capacity C_i. Adding this input capacity, we can see that it is in parallel with the capacitance C, therefore they will give just one pole of equivalent capacity $C + C_i$. Calculating the loop gain of this network, we obtain:

$$G_{loop} = -A(s)\frac{1}{1 + sR(C + C_i)}$$

and, in general, the input capacitance is much smaller than the other one:

$$C_i \ll C$$

and therefore we can neglect it.
A first possibility for compensating the circuit is to add a resistor R_C in parallel to C_i, thus connected to the positive and negative input pins of the operation amplifier. Neglecting, then, for the sake of simplicity, the small input capacitance it is possible to calculate the loop gain of this compensated circuit[34], obtaining:

$$G_{loop} = -\frac{A_0}{1 + s\tau} \cdot \frac{1}{1 + sC(R_C \| R)} \cdot \frac{R_C}{R + R_C}.$$

Adding a very small compensating capacitance:

$$R_C \ll R$$

we are thus decreasing the gain at low frequency, thus lowering the whole Bode diagram of the magnitude (that however will maintain the same shape) with respect to the zero decibels axis. At a certain point, the gain will be so low that the crossover frequency will be placed between the pole coming from the operation amplifier and the pole coming from the capacitors, thus crossing the zero decibels axis with a slope of -1 and thus giving a stable system. This kind of compensation, obviously, works if we do not have any alternative option, due to its major drawback. In this way, in fact, we are loosing loop gain and, therefore, accuracy on the output, thus increasing significantly the error.

An alternative possibility for compensating the circuit is to add a resistor R_C is series with the capacitor C, as in Figure 1.83. Again, we could study the loop gain of this circuit and then observe the position of the zeros and poles to obtain the phase margin: this kind of analysis is left to the student. A possible, alternative analysis of the circuit is based on the approximated

[34]Every time that a circuit is not drawn, it is left as an exercise to the willing student.

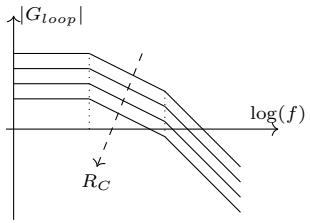


Figure 1.82: Variation of the magnitude of the loop gain depending on the value of the compensation capacitance.

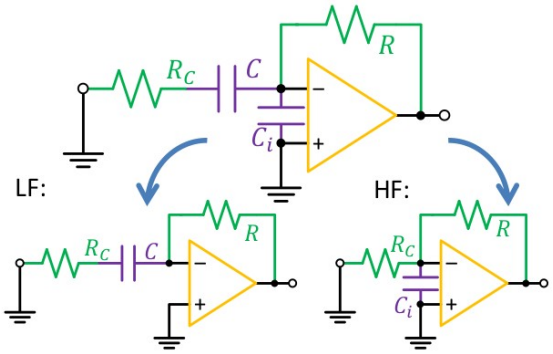


Figure 1.83: A compensated differentiator and its analysis in different frequency regimes.

evaluation of the poles. First of all, we can observe that the capacitors C and C_i are neither in parallel nor in series, thus being independent, and therefore we expect to have two poles (plus the one coming from the operation amplifier). To compute the position of the pole associated to C_i, we can calculate the equivalent resistance of the network as seen from this capacitor. However, we can immediately observe that, in principle, this will not be a resistance but it will be a complex impedance, varying with the frequency and containing the effect of the presence of the other capacitor C. The same can be observed by trying to calculate the equivalent resistance from the point of view of the C capacitor, that will contain the effects given from the capacity C_i. The effects of the two capacitors, therefore, are mixed up: by directly evaluating the loop gain of this network, we would have obtained a second order equation at the denominator of the transfer function, thus complicating the solution of this circuit. However, we can observe that in general the input capacitance is much smaller than the other one:

$$C_i \ll C$$

and therefore we can assume the two associated poles to be well separated:

$$f_{p1} = \frac{1}{2\pi C R_{eq,C}} \ll f_{p2} = \frac{1}{2\pi C_i R_{eq,C_i}}$$

if the two equivalent resistors $R_{eq,C}$ and R_{eq,C_i} are not too different[35]. If this

[35]This happens seldom, but it may happen, giving two poles one near to the other and

holds, we can evaluate the low-frequency pole f_{p1}, related to the capacitor C, observing that the capacitor C_i will be well below the frequency of its pole and, therefore, it can be approximated with an open circuit (as in the bottom left-hand side of Figure 1.83). This allows us to evaluate the equivalent resistance of the capacitor C as usual. In the same way, to find the high-frequency pole f_{p2}, that will be related to the input capacitance C_i, we can observe that the capacitor C will be operating well above its associated frequency and, therefore, it can be approximated by a short-circuit (as represented in the bottom right-hand side of Figure 1.83). This allows us to evaluate also the equivalent resistance of the capacitor C_i. Performing this approximate evaluation, we obtain that:

$$f_{p1} \simeq \frac{1}{2\pi C(R_C + R)}, \quad f_{p2} \simeq \frac{1}{2\pi C_i(R_C \| R)}$$

while from a complete, algebraic solution of the network we can obtain the frequency of the zero:

$$f_z = \frac{1}{2\pi C R_C}.$$

Moreover, we must consider the presence of the pole given by the operation amplifier. In general, the second pole f_{p2} is at high frequency and it can be neglected. In fact, the capacitance C_i is small, while the compensating resistance is much smaller than the feedback one:

$$R_C \ll R \;\Rightarrow\; R_C \| R \simeq R_C$$

thus begin in general placed in a position where the magnitude of the loop transfer function is well below the zero decibels. For the same reason, we can say that we are using a lag network for compensating this circuit, since:

$$R_C \ll R \;\Rightarrow\; f_z = \frac{1}{2\pi C R_C} > f_{p1} \simeq \frac{1}{2\pi C R}.$$

The closed-loop gain bandwidth of the circuit will be then limited by the frequency of the zero:

$$f_z = \frac{1}{2\pi R_C C}.$$

Considering the ideal gain and the fact that, since the loop is closed and the operation amplifier is an ideal element, the two input pins are at the same voltage and therefore every element placed between these two nodes can be neglected[36], we can write the ideal gain as:

$$G_{id} = -\frac{R}{R_C + \frac{1}{sC}} = \frac{-sCR}{1 + sCR_C}$$

obtaining the relationship represented in Figure 5.21 at page 297.

1.11.4 Capacitive load

We can now study what happens when we have an amplifier connected to a capacitive load C_L, as represented in Figure 1.84. This load will determine the

making this argument not valid.

[36]Since there will not be any voltage drop across these elements.

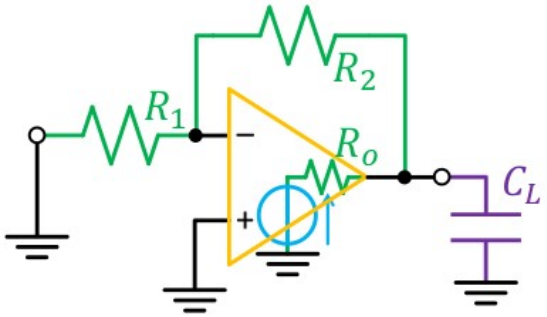


Figure 1.84: An amplifier connected to a capacitive load.

presence of an additional pole in the loop gain G_{loop} that we would like to be at a frequency higher than the gain-bandwidth product $GBWP$ of a certain factor (for example $10 \cdot GBWP$). We can evaluate the frequency of this additional pole as:

$$f_{pL} \simeq \frac{1}{2\pi C_L [R_o \| (R_1 + R_2)]}$$

but since we have that the output resistance is generally much smaller than any other resistance in the circuit:

$$R_o \ll R_1,\ R_2$$

we can write it as:

$$f_{pL} \simeq \frac{1}{2\pi C_L R_o}$$

and thus it could possibly lead to a significant decrease in the phase margin. Since the frequency of this pole is inversely proportional to the load capacitance:

$$f_{pL} \propto \frac{1}{C_L}$$

we could possibly have problems for large values of the load capacitance. We can thus calculate a maximum value of the load capacitance associated to our particular amplifier. Assuming to have a single-pole operation amplifier, we can observe that the crossover frequency of the loop gain will be, in principle the gain-bandwidth product $GBWP$ regardless of the initial gain considered.

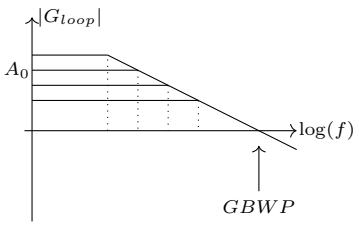


Figure 1.85: Representation of the Bode diagram of the magnitude of the operation amplifier.

Therefore, if we want that this pole does not affect the stability of the system regardless of the value of the gain A_0 considered, we want the pole to be placed at a frequency significantly higher than the gain-bandwidth product:

$$f_{pL} \gg GBWP \;\Rightarrow\; C_L \ll \frac{1}{2\pi R_o GBWP}.$$

Considering for example an acceptable margin a factor of ten between the gain-bandwidth product and the frequency of the pole, we can write the maximum value of the load capacity as:

$$C_L \simeq \frac{1}{2\pi R_o GBWP}.$$

This is just a quite rough estimate, giving a very stringent condition, and for high values of the load capacitance we need to compensate their presence.

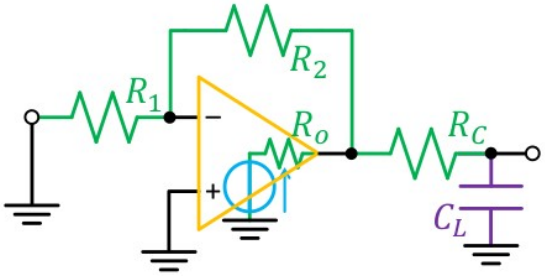


Figure 1.86: Compensation of a capacitive load.

A first possibility of compensating a capacitive load is to add a resistance R_C between the output of the operation amplifier and the capacitive load, as shown in Figure 1.86. This will lead us to add a zero and a pole to loop gain of this network. Evaluating the equivalent resistance of the load capacitance:

$$R_C + R_o \| (R_1 + R_2) \simeq R_C + R_o$$

we can write the frequency of the pole as:

$$f_p = \frac{1}{2\pi C_L [R_C + R_o \| (R_1 + R_2)]} \simeq \frac{1}{2\pi C_L (R_C + R_o)}$$

while the zero can be evaluated directly from the loop gain, obtaining:

$$f_z = \frac{1}{2\pi C_L R_C}.$$

In general, since the compensation resistance is small:

$$f_z > f_p$$

and by suitably choosing it we can reach the required phase margin and thus the stability of the network.
A more refined compensation scheme is represented in Figure 1.87. The analysis of this circuit, finding the positions of the two poles and of the two zeros and the conditions for having two pole-zero cancellations are left to the willing student.

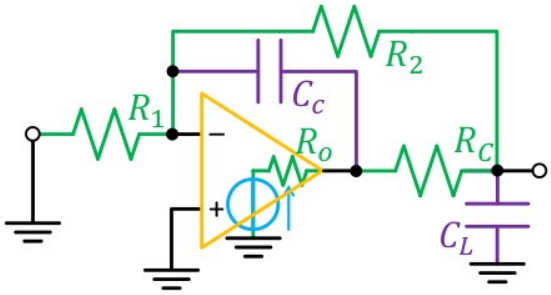


Figure 1.87: A more refined compensation scheme of a capacitive load.

1.12 Amplifiers and signals

1.12.1 Single-ended and differential signals

Up to now, we have studied how it is possible to build an amplifier depending on the characteristics of the operation amplifier at our disposal and on all the possible problems deriving from the fact that we are dealing with non-ideal devices. However, how is it possible to apply these amplifiers to the signals coming from a sensor? In this section and in the next chapter, therefore, we will mainly deal with what we called the input of the operation amplifier.

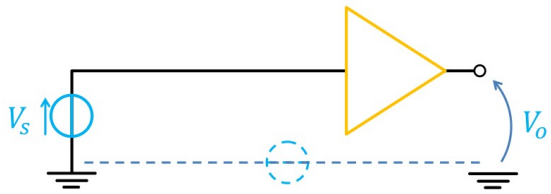


Figure 1.88: A single-ended signal.

In general, an input signal is defined as a voltage difference between two nodes. A possible choice is to have one node connected to the input of the amplifier (represented with a triangle in Figure 1.88 even though it is not just an operation amplifier) and the ground. In this case, therefore, at the output of the amplification stage we will obtain a voltage difference between the output pin of the amplifier and the ground and this kind of signal is called a single-ended signal. The peculiarity of these signals, therefore, is that one pin of the voltage difference is always considered to be at ground. A question, however, may arise: what is this ground? In fact, the ground potential is defined as a reference potential, that is arbitrarily chosen to be at:

$$V = 0$$

since the only physically meaningful quantities are the voltage differences, not the voltages themselves. However, in reality, it is impossible to define a unique ground, since nothing will be a perfect metal, in which every point is set exactly at the same potential by definition. The ground, therefore, will have a certain, intrinsic resistivity, thus making any local ground different from any other one.

It will be, therefore, just a reference voltage of the circuit. This kind of signals, therefore, will be suitable if and only if the input and the output of the amplifier are in the same place and they are referring to the same ground. If we consider remote signals coming to an amplifier that is placed in a different place, the two grounds (the one of the signal and the one of the amplifier) may be different[37] and they may be varying in a very complicated way. These changes in the ground are represented by the dashed voltage generator in Figure 1.88 and they are then amplified through the amplifier, possibly representing a source of troubles in our device. Therefore, this kind of signals have two important drawbacks: they are sensitive to noise and interferences coming from external electromagnetic fields and the are sensitive to differences in ground potentials. However, being simple and cheap, they are still used in non-demanding applications, in which we have a clear, strong and local signal.

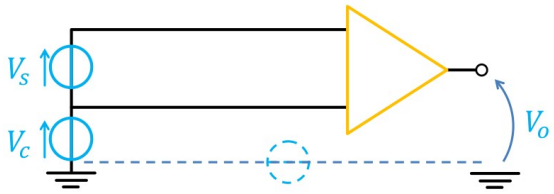


Figure 1.89: A differential signal.

On the other hand, if we are dealing with a differential amplifier, we will be able to amplify the difference between two input pins. Therefore, we will not care of the common-mode signal V_C, since it will be applied to both pins, and of the ground fluctuations (that will be common-mode signals too). This network, in fact, is sensitive only to differential signals. Moreover, most of the time the interferences will act in the same way on both input pins, since they are near one to the other, thus cancelling out almost completely their effects. These signals are called differential signals. In this case, however, we need to have an high common-mode rejection ratio ($CMRR$) in order to reject the noise and the fluctuations of the ground potential. Since this ratio is defined as:

$$CMRR = \frac{A_{dm}}{A_{cm}}$$

where A_{dm} is the differential-mode amplification of the amplifier and A_{cm} is the common-mode amplification. This is a strong additional requirement, on the amplifier, with respect to what we had in single-ended signals, but it is extremely important since in general the differential signal V_S is much smaller than the superimposed common-mode signal V_C. The devices needed for dealing with these signals are in general more expensive (requiring more components and more complex networks) but they can give rise to higher performances.

1.12.2 Subtractor circuit

A possible amplifier for a differential signal is the subtractor circuit, represented in Figure 1.90. In general, this circuit has two main drawbacks: the input

[37] Even of a few volts, thus surely not a negligible quantity.

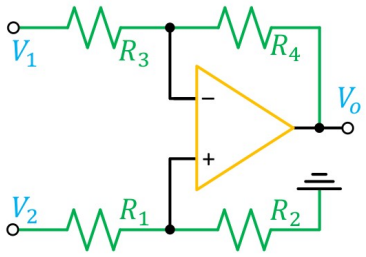


Figure 1.90: A subtractor circuit.

impedance of the network is low and asymmetric and, moreover, the common-mode rejection ratio is limited by the resistive matching.
To compute the common-mode gain of this stage, it is necessary to apply to both input pins the same, common-mode signal:

$$V_1 = V_2 = V_{cm}.$$

We can then study the network, that is linear, by exploiting the superposition principle. First, we can switch off the second input, obtaining:

$$V_2 = 0 \;\rightarrow\; V^+ = V^- = 0$$

and therefore the output is:

$$V_{o1} = -\frac{R_4}{R_3}V_1 = -\frac{R_4}{R_3}V_{cm}.$$

On the other hand, switching off the signal at the first input:

$$V_1 = 0 \;\rightarrow\; V^+ = V^- = \frac{R_2}{R_1 + R_2}V_{cm}$$

and thus the output is:

$$V_{o2} = \frac{R_3 + R_4}{R_3} \cdot \frac{R_2}{R_1 + R_2}V_{cm}.$$

Superimposing these two outputs, we obtain the overall output:

$$V_o = \left[\frac{R_3 + R_4}{R_3} \cdot \frac{R_2}{R_1 + R_2} - \frac{R_4}{R_3}\right] V_{cm}.$$

We can thus write the common-mode amplification factor as:

$$\begin{aligned} A_{cm} &= \frac{R_3 + R_4}{R_3}\frac{R_2}{R_1 + R_2} - \frac{R_4}{R_3} = \frac{R_2}{R_1 + R_2}\left(1 + \frac{R_4}{R_3} - \frac{R_4}{R_3}\frac{R_1 + R_2}{R_2}\right) = \\ &= \frac{R_2}{R_1 + R_2}\left(1 + \cancel{\frac{R_4}{R_3}} - \cancel{\frac{R_4}{R_3}} - \frac{R_4 R_1}{R_2 R_3}\right) = \frac{R_2}{R_1 + R_2}\left(1 - \frac{R_1 R_4}{R_2 R_3}\right) = \\ &= \frac{1}{1 + \frac{R_1}{R_2}}\left(1 - \frac{R_1 R_4}{R_2 R_3}\right). \end{aligned}$$

In order to have a fully differential amplifier, we would like this common-mode amplification factor to be equal to zero:

$$A_{cm} = 0 \;\rightarrow\; \frac{R_1}{R_2} = \frac{R_3}{R_4}.$$

However, a problem may arise when trying to impose this condition: resistors, in fact, are never exactly equal to a certain value, but they are different within a certain tolerance and certain differences related to external parameters. Therefore, instead of having a defined value R_i of resistance we will have an average value $\bar{R}_i$ and a certain tolerance ΔR_i such that:

$$\bar{R}_i \pm \Delta R_i = \bar{R}_i \, (1 \pm x)$$

where:

$$x = \frac{\Delta R_i}{\bar{R}_i}$$

depends on the manufacturing process, the environmental conditions and many other parameters. Therefore, at the numerator of the common-mode amplification factor we will have:

$$\frac{\bar{R}_1(1 \pm x)}{\bar{R}_2(1 \pm x)} \cdot \frac{\bar{R}_4(1 \pm x)}{\bar{R}_3(1 \pm x)}$$

and in the worst case it will give:

$$\frac{\bar{R}_1(1 + x)}{\bar{R}_2(1 - x)} \cdot \frac{\bar{R}_4(1 + x)}{\bar{R}_3(1 - x)} = (1 + x)^2 \left(\frac{1}{1 - x}\right)^2.$$

However, in general the relative tolerance x is small[38], therefore we can apply the following series expansion:

$$\frac{1}{1 - x} \simeq 1 + x$$

and neglecting second order terms:

$$\begin{aligned}(1 + x)^2 \left(\frac{1}{1 - x}\right)^2 &\simeq (1 + x)^2(1 + x)^2 = (1 + x^2 + 2x)^2 \simeq \\ &\simeq (1 + 2x)^2 \simeq (1 + 4x + 4x^2) \simeq 1 + 4x.\end{aligned}$$

Note that the same will happen also in others "worst case" conditions, depending on the choice of the sign of the tolerance, giving a fact of $\pm 4x$ at the numerator of the amplification factor of the common-mode signals that is not surprisingly equal to the number of independently fluctuating elements (the four resistors) multiplied by the associated tolerance, since tolerances sum up. After this reasoning, the common-mode amplification factor can be written as:

$$A_{cm} = \frac{4x}{1 + \frac{R_1}{R_2}}$$

[38]This value is indicated on the resistance by a colour scale: a silver ring is a tolerance of 10%, a gold ring corresponds to 5% and a violet one 0.1%.

and observing that, assuming valid, at least in a first order approximation, the resistor matching condition, the quantity at the denominator is almost unitary, as we can see from page 20 and sequent ones:

$$A_{cm} \simeq 4x$$

we can write the common-mode rejection ratio as:

$$CMRR = \frac{A_{dm}}{A_{cm}} \simeq \frac{A_{dm}}{4x}.$$

Therefore, the limitations to the common-mode rejection ratio comes from the tolerances of the resistors.

1.12.3 Instrumentation amplifiers

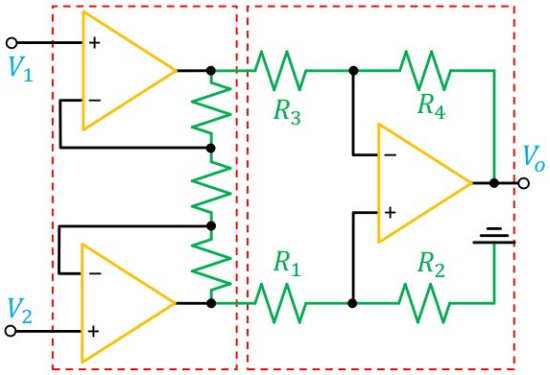


Figure 1.91: An instrumentation amplifier.

To overcome the difficulty that we have just described, a new kind of amplifier, called instrumentation amplifier and represented in Figure 1.91, was developed. Even though it can be found as a single component, it basically consists of two different stages: a sort of differential amplifier, with two inputs and two outputs, and a subtractor identical to the one we have just discussed. We can now study in further details the first stage.

To study the behaviour of the first stage when we have a common-mode input signal, we can assume the two input pins to be identically equal to V_C. This means that both the negative pins of the two operation amplifiers, since we are dealing with a negative feedback system with ideal operation amplifiers, will be set at:

$$V^- = V_C.$$

This means that there will not be any current flowing through the resistor R_G, since it will be placed between two nodes at the same voltage and, since the operation amplifier will have an infinite input impedance, there will not be any current flowing through R too. This means that we will not have any voltage drop across the resistors R and thus both the outputs will be at the common-mode voltage:

$$V_{o1} = V_{o2} = V_C.$$

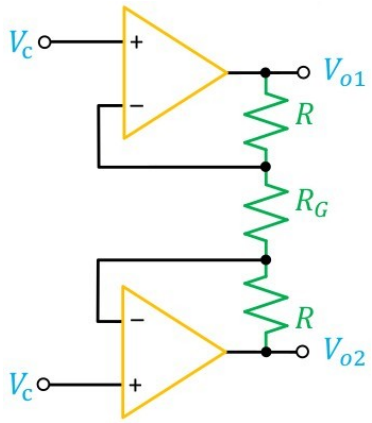


Figure 1.92: Study of the first stage of an instrumentation amplifier under common-mode signal.

This stage, therefore, does not reject the common-mode signal, but it has a unitary common-mode amplification factor:

$$A_{cm1} = 1.$$

It is possible to note that nothing will change if one of the two resistors indicated with R changes, therefore the common-mode amplification factor is almost independent from the resistor matching and from any problem involving the tolerances of the resistors.

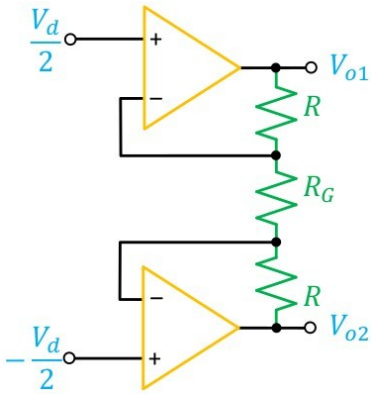


Figure 1.93: Study of the first stage of an instrumentation amplifier under differential-mode signal.

Assuming now a differential-mode signal as in Figure 1.93:

$$V_1^+ = \frac{V_d}{2}, \quad V_2^+ = -\frac{V_d}{2}$$

where we used the subscript 1 for the upper operation amplifier and the 2 for the lower one, since we are dealing with negative feedback systems and ideal

operation amplifiers:

$$V_1^- = \frac{V_d}{2}, \quad V_2^- = -\frac{V_d}{2}.$$

Therefore, there will be a voltage drop across the resistor R_G and thus a current flowing through it:

$$I = \frac{V_1^- - V_2^-}{R_G} = \frac{\frac{V_d}{2} + \frac{V_d}{2}}{R_G} = \frac{V_d}{R_G}.$$

However, this current flows through both resistors R (since the operation amplifier has an infinite input impedance, thus giving the following output voltages:

$$V_{o1} = \frac{V_d}{2} + IR = \frac{V_d}{2} + V_d\frac{R}{R_G} = \left(\frac{1}{2} + \frac{R}{R_G}\right)V_d$$

$$V_{o2} = -\frac{V_d}{2} - IR = -\left(\frac{V_d}{2} + V_d\frac{R}{R_G}\right) = -\left(\frac{1}{2} + \frac{R}{R_G}\right)V_d$$

thus giving the following differential-mode amplification factor[39]:

$$A_{dm1} = \frac{V_{o1} - V_{o2}}{V_d} = 1 + 2\frac{R}{R_G}$$

that is mainly set by the value of the external resistor R_G. In real instrumentation amplifiers, in fact, all these components will be part of an integrated device except for the R_G resistor, that will be external and that can be changed depending on the application considered.
It is possible to observe that this result is extremely similar to the one of a non-inverting amplifier. In fact, rewriting the resistor R_G as the series of two resistors equal to R_G, we can observe that a node between them will be set a ground potential. The circuit, therefore, will be symmetric with respect to this point and the same symmetry is reflected in the voltages applied. Therefore, this point must necessarily be equal to the ground potential for symmetry reasons. The overall gain, therefore, can be calculated considering a non-inverting amplifier as the one represented in Figure 1.11 at page 13 where:

$$R_1 = R, \quad R_2 = \frac{R_G}{2}.$$

Considering now also the second stage of the instrumentation amplifier, it is possible to calculate the overall common-mode rejection ratio of the network. Since a common-mode signal entering the first stage will be again a common-mode signal at the end of the first stage and, in the same way, a differential-mode signal is again differential at the end of the first stage[40], we can factorize the differential-mode or common-mode amplification factors as the factors related to the two stages:

$$CMRR = \frac{A_{dm}}{A_{cm}} = \frac{A_{dm1}A_{dm2}}{A_{cm1}A_{cm2}}$$

but since we have just demonstrated that:

$$CMRR_1 = \frac{A_{dm1}}{A_{cm1}} = \frac{A_{dm1}}{1}$$

[39] It is defined as the differential output over the differential input.
[40] And this is not as obvious as it seems to be.

we obtain that:

$$CMRR = A_{dm1} CMRR_2.$$

We are therefore improving the overall common-mode rejection ratio of a factor A_{dm1} given by the first stage. This is due to the fact that the first stage is not directly rejecting the common-mode signal, it is just amplifying in a significant way the differential-mode signal. We are therefore separating the two tasks, the amplification of the differential-mode and the rejection of the common-mode, between the two stages of the instrumentation amplifier.
The common-mode rejection ratio is therefore actually limited by the operation amplifiers and, to a first approximation, the common-mode errors will be cancelled by the second stage, since it is the difference between the common-mode rejection ratios that is important. This leads us to achieve common-mode rejection ratios up to 90 or 140 dB.
We can now give some specifications that are useful for understanding the behaviour of an instrumentation amplifier. An usual gain range, for an instrumentation amplifier, is generally between 1 (that is the value obtained without any resistance R_G) and 1000 (under a suitable choice of the resistance R_G). In this range, the device is guaranteed to work as specified regardless of the choice of R_G. Exceeding this value, the circuit will probably continue to work but its performances will be no longer guaranteed. The gain (or equation) error represents the maximum deviation of experimental data about the gain from the equation that is assumed to be describing the gain. Usual values are, for example, 0.5% and it is important to note that the gain equation[41] may be not linear. Another important parameter is the non-linearity, that is defined as the maximum deviation from the interpolating line; a typical value is 100 ppm. Note that, if the gain equation is linear, this parameter will be conceptually similar to the gain error we have previously defined; however, this is not always true. The last parameter, the offset voltage, deserves a further study.

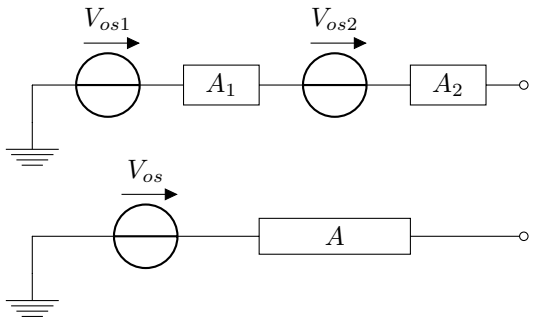


Figure 1.94: A cascade of two operation amplifier with their associated offset voltage and the representation of the equivalent amplifier and the associated offset voltage.

As an example, we can consider a network with multiple amplification stages. Each one of these stages will have its own offset voltage and, in the example, we

[41] The gain equation is a characteristic equation that states the relationship between the gain of the instrumentation amplifier and $1/R_G$. This error, therefore, will give the precision of this equation in describing the real value of the gain.

are considering two amplification stages, thus obtaining two offset voltages V_{os1} and V_{os2}. Grounding as in Figure 1.94 the input, we can find the output as:

$$V_o = A_1 A_2 V_{os1} + A_2 V_{os2} = A_1 A_2 V_{os}$$

since the overall amplification factor is:

$$A = A_1 A_2$$

and where V_{os} is the overall offset voltage. This gives, therefore:

$$V_{os} = V_{os1} + \frac{V_{os2}}{A_1}$$

the overall offset voltage as expressed in terms of the offset voltages of each stage. In the particular case we are describing, since the gain of the first stage is:

$$A_1 = G = 1 + \frac{2R}{R_G}$$

we can write the overall offset voltage of an instrumentation amplifier as:

$$V_{os} = V_{os1} + \frac{V_{os2}}{G}.$$

A typical value for the offset voltage is 500 μV and we can observe that the more we increase the gain of the first stage, the lower will be the offset voltage, even though it is limited by the value of the offset voltage of the first stage.
A typical gain-bandwidth product for an operation amplifier is about 100 kHz, that we can observe to not be very large, in particular being smaller than in circuits in which we only have operation amplifiers, due to the fact that this circuit is generally optimized only for having an high common-mode rejection ratio. We can then obtain a drift of the offset voltage that typically is lower than 0.5 μV/°C and a bias current that is typically lower than 2 nA. In general, these devices are made using a bipolar technology.

1.13 Single power supply operation amplifiers

In general, we will always assume to have a simple, symmetric power supply network for the operation amplifiers that we are using. However, there are cases in which this is not possible for many reasons: in these cases, we need to have a single power supply.

Consider, for example, the inverting amplifier represented in Figure 1.95, that is assumed to have a unitary gain. Since the output voltage range is limited by the power voltage supply, the output range will be slightly smaller than the interval $[0, V_{cc}]$, therefore we cannot have a negative output regardless of the input, as shown in Figure 1.96. Moreover, also the input swing is in general limited by the power supply, therefore also the first part of the output voltage could be wrong.

Back to the dual power supply network, we can observe that it will work fine since:

$$V_i = 0 \;\Rightarrow\; V_o = 0$$

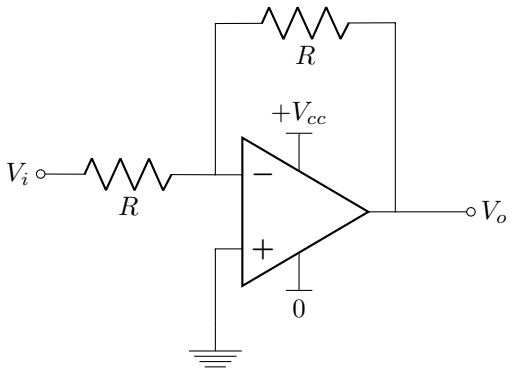


Figure 1.95: A single-power supply operation amplifier in an inverting configuration.

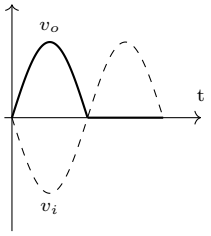


Figure 1.96: An example of output for a sinusoidal input.

is the equilibrium point of the circuit. We can note that this working point is exactly in the middle between the two power supply voltages $V_{cc}/2$ and $-V_{cc}/2$. In analogous way, we can try to find the equilibrium or reference point of the single power supply network. According to the idea that we have just expressed, it will be $V_{cc}/2$. Therefore, to pass from the symmetric power supply to a single power supply we need to add to every node (except for the power supply nodes, obviously) $V_{cc}/2$, thus biasing our network. Assuming thus to have a zero-signal at the input, the output will be identically equal to $V_{cc}/2$. To bias the positive input pin to $V_{cc}/2$, we need to add a suitable bias network as in Figure 1.98.

Assuming to have this bias network, we can decouple the input signal V_i from the bias voltage $V_{cc}/2 = V^-$ by adding a capacitor C. The drawback of this addition, however, is that the transfer function of this inverting configuration has been modified:

$$T(s) = -\frac{R}{R + \frac{1}{sC}} = -\frac{sCR}{1 + sCR}$$

thus integrating the signal up to a certain frequency at which we have a pole and where the gain reach its desired value of -1 (given this choice of the resistors). Sometimes we can also add a capacity in parallel to the R_B resistor that is placed between V^- and ground, in order to get rid of the ripples that are always present in the power supply. Therefore, we are able to amplify signals only above a certain cut-off frequency:

$$f_c = \frac{1}{2\pi CR}.$$

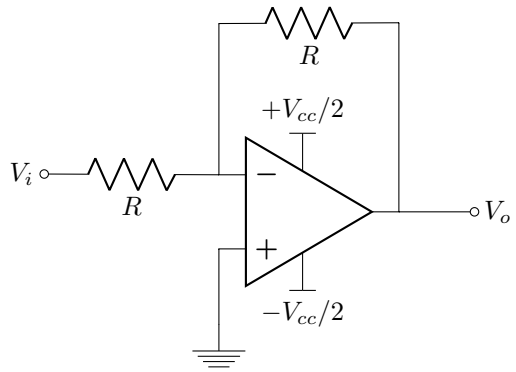


Figure 1.97: A double-power supply operation amplifier in an inverting configuration.

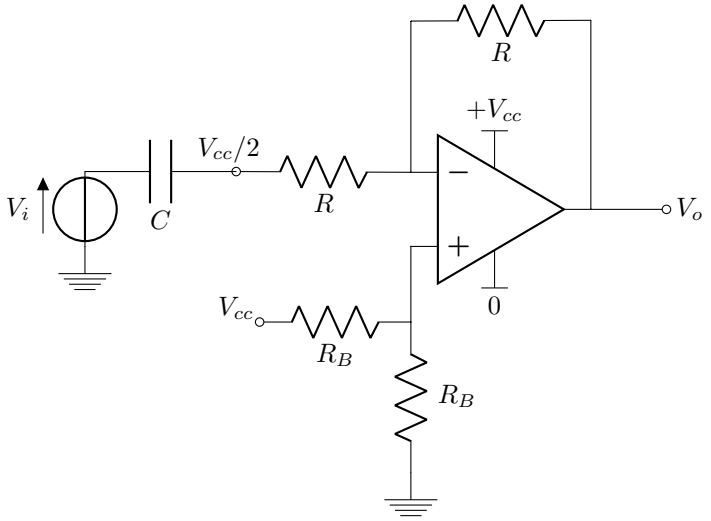


Figure 1.98: A single-power supply operation amplifier in an inverting configuration with the associated bias network.

Also in the non-inverting case, if we have a single-power supply, we are rising everything of $V_{cc}/2$. The operation amplifier in non-inverting configuration with the associated bias network is represented in Figure 1.99.

In this network, the capacitor C_2 does not change the gain given by the partition resistances of the feedback network R_1 and R_2. This capacitor, however, will affect the bandwidth of the amplifier, therefore an amplification is not possible in the low-frequency limit. It is then possible also to have a capacitor C_3 at the output of the network to obtain an output signal referred to zero, thus being decoupled from the bias.

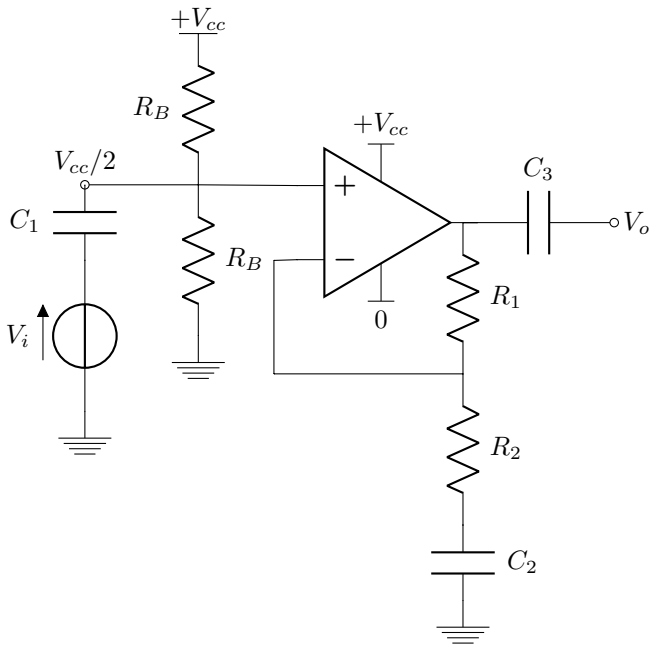


Figure 1.99: A single-power supply operation amplifier in a non-inverting configuration with the associated bias network.

Chapter 2

Sensors

2.1 Signal readout from resistive sensors

2.1.1 Resistive sensors

After this description of the operation amplifiers and how they can be used to amplify a signal, we can focus on the sensors, that are, in general, the source of the input signal V_i considered in the previous chapter. A large and important class of sensors is represented by the so called resistive sensors. These sensors track a certain physical quantity (for example, temperature, strain, magnetic field, ...) as a change in the resistance of a conductive element.
Considering S to be the considered physical quantity, assuming to have a small change of this quantity we can in general associated to it a small variation of the resistance, thus expanding with a linear approximation this dependence:

$$R = R_0 + \Delta R = R_0 \left(1 + \frac{\Delta R}{R}\right) = R_0(1 + \alpha S) = R_0(1 + x)$$

where we have defined the following coefficient:

$$\alpha = \frac{1}{R_0} \frac{dR}{dS}\bigg|_{R=R_0}$$

that depends on the physical phenomenon we are considering in this sensor.

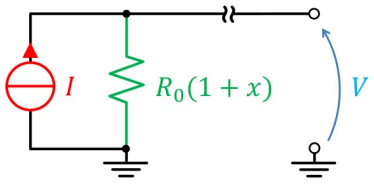


Figure 2.1: Example of single-ended resistive measurement.

A problem, however, at this point may arise: how is it possible to measure only variations in the resistance (that are, in general, extremely small) and not the average value R_0 of this resistance? In other words, we can consider

single-ended measurements, represented in Figure 2.1. In this case, we bias a resistor using a certain current generator I and we measure the variations of the associated voltage drop:

$$V = IR = IR_0(1 + x) = IR_0 + I\Delta R.$$

The problem, however, is that in general we have a large bias signal IR_0 on top of a very small variation $I\Delta R$ that we would like to measure. Therefore, an extremely high precision is required for measuring the value of x with the needed accuracy. Moreover, we can have noise, interferences, ground potentials fluctuations and any other problem generally affecting single-ended measurements that come to degrade the performance of our system. This kind of measurement, therefore, can be used only for high-level signal, that are characterized by a low noise and that are propagated for short distances through the environment.
How is it possible, however, to have a differential measurement for resistive sensors?

2.1.2 Wheatstone bridge

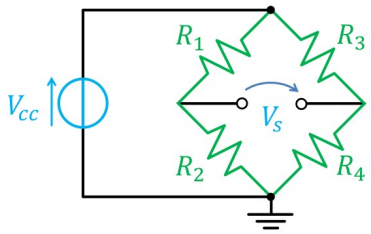


Figure 2.2: The Wheatstone bridge.

The answer to the concluding question of the previous section is called Wheatstone bridge. This kind of network, represented in Figure 2.2, allows differential measurements and that can be used associated to suitable amplifiers. In it, we are measuring the so called unbalance of the network, that is the difference in the two middle voltages of the two arms of the bridge; V_{cc} is the bias voltage of the network. The unbalance V_S can be found by writing a partition for the two arms of the bridge:

$$V_S = V_{cc}\left(\frac{R_4}{R_3 + R_4} - \frac{R_2}{R_1 + R_2}\right) = V_{cc}\left(\frac{1}{1 + \frac{R_3}{R_4}} - \frac{1}{1 + \frac{R_1}{R_2}}\right).$$

At our reference value, ideally, the unbalance must be identically equal to zero, thus giving the following conditions:

$$V_S = 0 \;\Rightarrow\; \frac{R_1}{R_2} = \frac{R_3}{R_4} = k.$$

To choose the value of this parameter k, we can require that the maximum sensitivity of the measured voltage V_s to a variation of the resistances:

$$\frac{dV_s}{dR_1} = \frac{V_{cc}}{\left(1 + \frac{R_1}{R_2}\right)^2} \cdot \frac{1}{R_2} = \frac{V_{cc}}{R_1} \cdot \frac{k}{(1 + k)^2}.$$

From this expression, we can observe that it depends exclusively on the absolute value of the supply voltage V_{cc} and on the resistance R_1, obviously, apart from being a function of the ratio k. The sensitivity of the bridge with respect to this parameter can be observed in Figure 2.3.

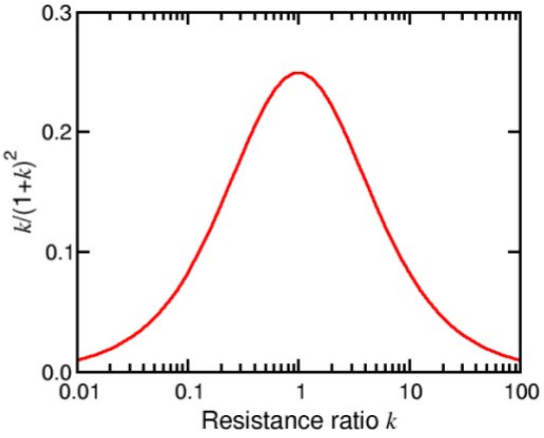


Figure 2.3: The sensitivity of a Wheatstone bridge as a function of the ratio k.

It is important that, even though it seems, this is not a gaussian behaviour and it attains its maximum in:

$$k = 1.$$

This means that, to maximize the sensitivity, we have to require:

$$R_1 = R_2 = R_3 = \langle R_4 \rangle$$

where three resistors are equal and the fourth one is equal to them on average (apart from its variations). Assuming therefore:

$$R_1 = R_2 = R_3 = R, \quad R_4 = R(1+x)$$

we have that the output voltage can be written as:

$$V_s = V_{cc}\left(\frac{R(1+x)}{R(2+x)} - \frac{1}{2}\right) = V_{cc}\frac{x}{2(2+x)}.$$

Note that the exact relationship that we have obtained, in this case, is not a linear one. In general, this is not good property of this device, since we would like to have linear sensors, otherwise we would need to invert a non-linear relationship to determine the desired physical quantity. However, since the normalized variation x of the resistance is in general a small quantity (a few percent at the very most), we can make the following approximation:

$$V_{cc}\frac{x}{4\left(1+\frac{x}{2}\right)} \simeq \frac{V_{cc}}{4}\cdot x \cdot \left(1-\frac{x}{2}\right) \simeq \frac{V_{cc}}{4}x - \frac{V_{cc}}{8}x^2 \simeq \frac{V_{cc}}{4}x = V_s$$

where we have used a first order approximation of the exact denominator and we have neglected second order terms. In general, the measured voltage V_s is really small, as in the following example:

$$V_{cc} = 10 \text{ V}, \quad x = 0.002 = 0.2\% \;\Rightarrow\; V_s = 5 \text{ mV}.$$

In this case, we can calculate the linearity error as:

$$\epsilon = \left| \frac{\frac{x}{2(2+x)} - \frac{x}{4}}{\frac{x}{2(2+x)}} \right| = \left| 1 - \frac{x}{4} \cdot \frac{2(2+x)}{x} \right| = \frac{x}{2}$$

and assuming the values that we have defined in this example:

$$\epsilon = 10^{-3} = 0.1\%.$$

The maximum tolerable value, for this error, depends on the specifications of the problem we are dealing with.

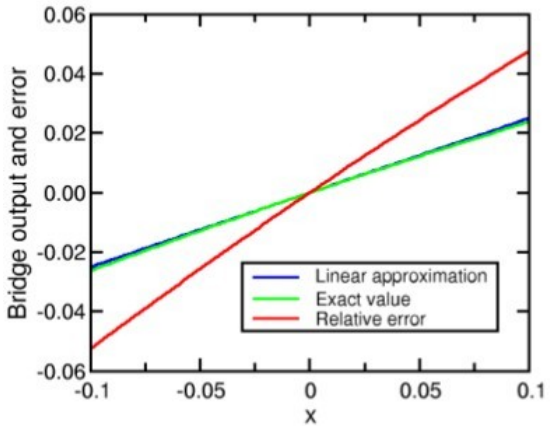


Figure 2.4: Linear approximation and exact behaviour of the output voltage of the bridge and relative error in this approximation.

A possible solution to increase the output signal with respect to the previous case is to have two different active elements that can vary their resistances. In particular, assuming the bottom right resistor to be variable, the only other possible variable resistor will be top left one, as represented in Figure 2.5.

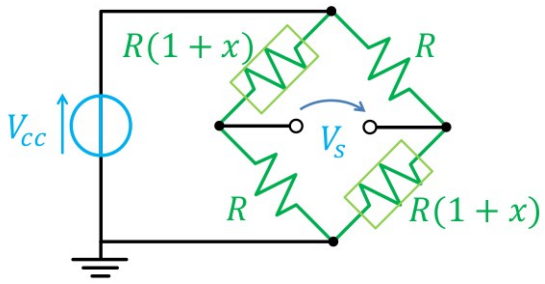


Figure 2.5: Wheatstone bridge with doubled sensitivity.

In fact, if on the other hand we were using the top right resistor as a variable one, we would obtain a zero output signal, since we would have the same ratio between the resistors in the two branches. On the other hand, using the bottom left resistor as the second variable resistor, the output signal would have been

again zero due to a cancellation between the two terms that compose the output signal V_s. The only possibility left is therefore the one that is represented in Figure 2.5, since a normalized variation x will increase the term corresponding to one branch of the bridge and it will decrease the term related to the other one. From calculations, in fact:

$$V_s = V_{cc}\left(\frac{R(1+x)}{R(2+x)} - \frac{R}{R(2+x)}\right) = V_{cc}\frac{x}{2+x}$$

and again, since:

$$x \ll 2 \;\rightarrow\; 2 + x \simeq 2$$

we obtain:

$$V_s \simeq V_{cc}\frac{x}{2}.$$

Comparing this formula with the sensitivity that we have previously obtained, we can immediately observed that this approximated value is equal to twice the approximated value that we had using just one variable resistor: we have doubled the sensitivity. Moreover, also in this case it is possible to calculate the non-linearity error (that arises from the fact that we are approximating a non-linear relationship with a linear one), obtaining a value that is identical to the one we get in the previous case:

$$\epsilon = \frac{x}{2}.$$

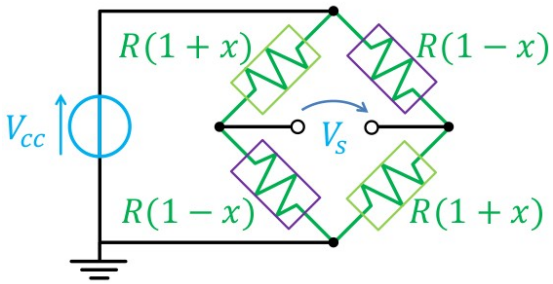


Figure 2.6: Wheatstone bridge with the maximum possible sensitivity.

In certain cases, the sensitivity of a Wheatstone bridge can be further increased if we can use sensors with a different dependence (with respect to the sign, but same modulus) on the normalized variation x, as represented in Figure 2.6. An example of these kind of devices are strain sensors. The sensitivity, in this case, can be calculated as:

$$V_s = V_{cc}\left(\frac{R(1+x)}{2R} - \frac{R(1-x)}{2R}\right) = V_{cc}x$$

and we can immediately observe that it is four times higher than the value we get using just one single active element. Moreover, this relationship is exact, therefore the linearity error in this case can be assumed to be ideally equal to zero. The problem, however, is to find sensors where this kind of circuit is

possible, since it is not always available and it can be quite costly due to the fact that we are using a lot of sensors for a single measurement.
We are now able to define a few important parameters for the Wheatstone bridge. The first one is the sensitivity of the bridge, that is defined as the output voltage for a unitary bias voltage:

$$V_{cc} = 1 \text{ V}$$

that is associated to the maximum variation of the normalized resistance:

$$x = x_{max}.$$

This parameter is usually expressed in millivolts over volts. In the previous case:

$$V_s = 5 \text{ mV}, \; V_{cc} = 10 \text{ V} \; \Rightarrow \; S = 0.5 \text{ mV/V}.$$

A second parameter is the accuracy, that is defined as the difference between the real characteristic equation and the linear one expressed as a percentage. In reality, this parameter will be higher than the non-linearity error that we have calculated in the previous examples, since passive and active elements of the circuit will not behave like ideal elements.
Last is the resistance, that is the resistance of the bridge when it is measured between the output terminals. It can computed as its nominal value, therefore when:

$$x = 0$$

and from the circuit represented in Figure 2.2, we can observe that it is the parallel between the two branches of the bridge, each equal to the series between two identical resistances R:

$$R_{eq} = \frac{(R+R)(R+R)}{(R+R)+(R+R)} = R$$

and therefore, if the bridge is balanced, the equivalent resistance of the bridge in nominal conditions will be equal to the nominal value of the resistances.

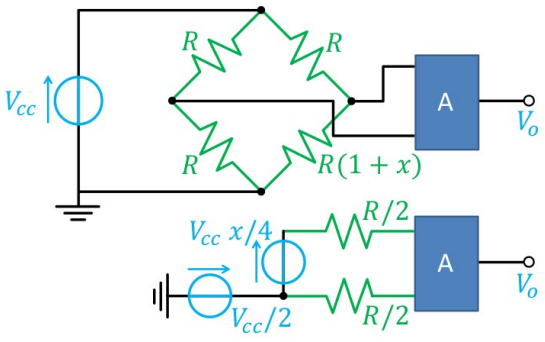


Figure 2.7: A Wheatstone bridge connected to an amplifier and the equivalent circuit of the bridge when it is seen from the amplifier.

The output signal of a Wheatstone bridge, then, will be the measured voltage that will be passed to a certain amplifier. We can therefore discuss the

requirements that are needed for having a good and reliable amplification of the signal. Considering the circuit in the upper part of Figure 2.7, therefore, we can study the Thévenin equivalent circuit of the bridge with respect to the input pins of the amplifier. Considering the lower pin in the drawing, therefore, we can observe that to pass from it to the ground we have voltage equal to $V_{cc}/2$ and an equivalent resistance (that can be calculated switching off the voltage source V_{cc}, thus replacing it with a short-circuit) that is equal to $R/2$, since it is the parallel between two identical resistances R. For the upper pin, on the other hand, we can say that in nominal behaviour ($x = 0$) the voltage applied will be $V_{cc}/2$ and then, when we have variations of the normalized resistance x, it will have an additional voltage that can be approximately written as $V_{cc}x/4$. The equivalent resistance of this pin, on the other hand, can be calculated again switching off any voltage source (and replacing it with a short-circuit), thus resulting to be the parallel between R and $R(1 + x)$. However, since we are dealing with small variations of the normalized resistance x, also in this case the equivalent resistance will be the parallel between two identical resistors R, thus being $R/2$. From this reasoning, therefore, we obtain the equivalent circuit represented in Figure 2.7.
We can now, referring to the previous example:

$$V_{cc} = 10 \text{ V}, \quad V_s = V_{cc}\frac{x}{4} \simeq 5 \text{ mV}, \quad x_{max} = 0.2\%$$

investigate the requirements for having a good amplifier. First of all, we can observe that the amplifier is a differential amplifier, therefore the common-mode signal $V_{cc}/2$ should not be amplified. Assuming to be exploiting the full dynamic of the signal, thus requiring that for the maximum signal we obtain as an output a voltage that is equal to V_{cc}:

$$\begin{cases} V_s = 0 & \rightarrow \ V_0 = 0 \\ V_s = 5 \text{ mV} & \rightarrow \ V_0 = V_{cc} = 10 \text{ V} \end{cases}$$

we obtain the following gain:

$$G = 2000.$$

Therefore, we would like to have a large value of the gain.
Then, we can consider for example to have a finite input impedance R_i between the two input pins of the amplifier. In this case, if the differential signal is $V_{cc}x/4$, we will not measure this value but we will measure the following partition V_i:

$$V_i = V_{cc}\frac{x}{4} \cdot \frac{R_i}{R_i + R}$$

where R is the overall resistance of the loop. This value can then be approximated as:

$$V_i \simeq V_{cc}\frac{x}{4}\left(1 - \frac{R}{R_i}\right)$$

and therefore we can write the error of this value as:

$$\epsilon = \frac{R}{R_i}.$$

This means that we would like to have an high value of the input impedance for this amplifier. In particular, assuming:

$$R = 100\ \Omega, \quad \epsilon < 0.1\%$$

we are making the following requirement on the input resistance:

$$R_i \geq 1000R = 100\ \text{k}\Omega.$$

Last, the maximum common-mode signal that we are applying to our circuit is:

$$V_{cm} = 5\ \text{V}$$

and we want to reject it as much as possible, while the maximum differential input voltage is equal to 5 mV. Assuming that this circuit will work with an 8-bit system (we will study later on how the performance of sensors and amplifiers influences converters and other part of the acquisition system), we can say that the minimum signal that we would like to distinguish[1] is equal to the maximum differential input signal divided by the number of channels that we have:

$$V_{LSB} = V_{cc}\frac{x}{4}\Big|_{max} \cdot \frac{1}{2^8} \simeq 20\ \mu\text{V}.$$

Since at the output of the amplifier we will have a superposition of the amplified differential signal and of the rejected common-mode signal:

$$V_0 = V_{cm}A_{cm} + V_{dm}A_{dm} = A_{dm}\left(V_{dm} + V_{cm}\frac{A_{cm}}{A_{dm}}\right)$$

since we want the differential mode signal to be dominant over the common-mode signal:

$$V_{cm}\frac{A_{cm}}{A_{dm}} \ll V_{dm}$$

and we can rewrite this requirement according to the definition of common-mode rejection ratio $CMRR$:

$$CMRR = \frac{A_{dm}}{A_{cm}} \gg \frac{V_{cm}}{V_{dm}}.$$

This allows us to make the following estimate of the minimum common-mode rejection ratio that is desired:

$$CMRR \geq \frac{V_{cm}}{V_{LSB}} = \frac{5\ \text{V}}{20\ \mu\text{V}} \simeq 108\ \text{dB}.$$

This is a very conservative estimate, in reality this minimum value is slightly smaller, even remaining quite high. These requirements are quite stringent and difficult to match: this can be generally done only using an instrumentation amplifier.

[1] The subscript LSB, in this quantity, means "least significant bit" and it is the smallest signal that we need to distinguish from the zero-signal.

2.1.3 2-, 3- and 4-wire connections

In general, we assume wirings in circuits to be ideal links, and this is a good approximation when we are in a controlled environment, such as a lab. In reality, however, this approximation does not hold, since dealing with remotely located bridges (useful for example when measuring particularly high temperatures) cable resistances and noise pickup is a big source of errors. In general, the resistances in cables are much smaller than the resistances in the bridge (thus being negligible) or they give a constant offset error, that can be suitably compensated. However, changes in cable resistances during the operation of the circuit (for example, with temperature) lead to an error signal (since they gives a variation of the gain) at the bridge output. We can thus investigate the effect of the connections on the system.

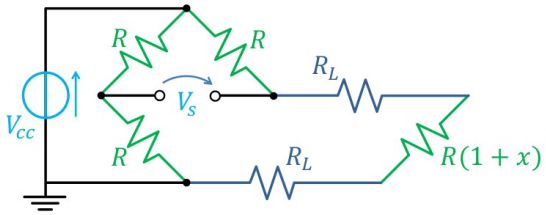


Figure 2.8: The 2-wire connection of a Wheatstone bridge.

The first way of connecting a Wheatstone bridge with a remotely located variable resistance is the 2-wire connection, represented in Figure 2.8. Assuming R_L to be the additional resistances due to the link, solving this network we obtain:

$$\begin{aligned} V_s &= V_{cc}\left(-\frac{1}{2}+\frac{R(1+x)+2R_L}{R(2+x)+2R_L}\right)= \\ &= V_{cc}\left(\frac{-R(2+x)-2R_L+R(2+2x)+4R_L}{2R(2+x)+4R_L}\right)= \\ &= V_{cc}\frac{Rx+2R_L}{2R(2+x)+4R_L}. \end{aligned}$$

In a first order approximation, since:

$$x \ll 2$$

we can write:

$$V_s \simeq V_{cc}\frac{Rx+2R_L}{4R} = V_{cc}\left(\frac{x}{4}+\frac{R_L}{2R}\right) = \frac{V_{cc}}{4}\left(x+\frac{2R_L}{R}\right).$$

Therefore, we can observe that $2R_L/R$ is the error related to the presence of the links. Assuming the following typical values for the resistances, the error can be calculated as:

$$2R_L = 0.5\ \Omega,\ R = 100\ \Omega\ \rightarrow\ \epsilon = 0.5\%.$$

We need thus to compare this value to the desired precision on the specific application we are dealing with. If the link resistances R_L were constant, they

would give a constant offset in value of the measured voltage and therefore we could compensate for their presence. The link resistances, however, varies due to the fact that they are sensitive to the environmental conditions, giving a certain noise term in the output.

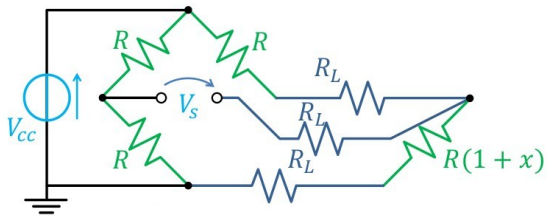


Figure 2.9: The 3-wire connection of a Wheatstone bridge.

If the previous error is too large for the application we are dealing with, we can try to reduce it. The problem, in the 2-wire connection, is that the link resistances R_L are in series to the variable resistor $R(1+x)$ and therefore we are not measuring the voltage drop across this variable resistor but also the voltage drops across the two links. Adding a third connection, as represented in Figure 2.9, we create the 3-wire connection and try to measure, at least on one side, the voltage at the remote node that is between one link resistance and the variable resistance, where we have added the third connection. From this circuit, we can write:

$$\begin{aligned} V_s &= V_{cc}\left(-\frac{1}{2}+\frac{R(1+x)+R_L}{R(2+x)+2R_L}\right)= \\ &= V_{cc}\left(\frac{-R(2+x)-\cancel{2R_L}+R(2+2x)+\cancel{2R_L}}{2R(2+x)+4R_L}\right)= \\ &= V_{cc}\left(\frac{Rx}{2R(2+x)+4R_L}\right)= \\ &= V_{cc}\frac{Rx}{4R\left(1+\frac{x}{2}+\frac{R_L}{R}\right)} \end{aligned}$$

and again, in a first order approximation, since:

$$\frac{x}{2}+\frac{R_L}{R} \ll 1$$

we obtain:

$$V_s \simeq V_{cc}\frac{x}{4}\left(1-\frac{x}{2}-\frac{R_L}{R}\right)=V_{cc}\frac{x}{4}\left(1-\frac{Rx+2R_L}{2R}\right).$$

Since we have assumed, in the previous section, to be reading the output signal V_s with an high-impedance amplifier, from the middle link resistance R_L there will not be any current flowing and, therefore, the voltage at both ends of the resistance will be identical. This is the difference with the previous case, where we were measuring also the load effects of the two link resistances due to the fact that there was a current flowing through the link used for measuring the voltage.

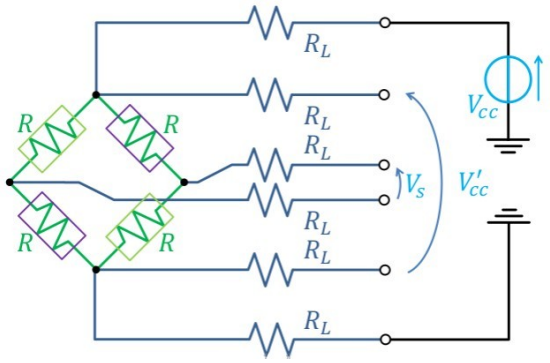


Figure 2.10: The Kelvin (or 4-wire) connection of a Wheatstone bridge.

An further extension of this way of reasoning is represented by the so called Kelvin (or 4-wire) connection. In this case, the whole bridge is remotely located and we have four different active elements. In this kind of circuit, also the bias voltage V_{cc} and the ground are influenced by the presence of the resistances of the link R_L, therefore we need to refer everything to the same local ground and to take into account the voltage drop across the wirings, that have a certain resistance R_L. It is important to note that, even though it is called 4-wire connection, in reality we have six wires connecting the local circuit to the Wheatstone bridge and bringing back the signal. Moreover, the ground of the bias voltage and the one of the bridge are equal and, referring every voltage to the same local ground, we are getting rid of the ground fluctuations. It is also possible to have a different configuration (in which the bridge is excited with a constant current) with respect to the one represented in Figure 2.10, that will reduce the number of wirings and thus of resistances to four. Note that, apart from the wirings that are connecting V_{cc} and the ground to the bridge, on the other wirings there will not be any voltage drop since, at least ideally, there will not be any current flowing across these resistances. Apart from the intrinsic problems of a Wheatstone bridge with four active elements, additional problems may arise from the cost of this circuit and from the fact that, in some applications, there is not enough space for running all these wires. As a rule of thumb, the 3-wire connection is suitable up to a few tens of metres of distance between the varying the resistor and the rest of the bridge; for longer connections, different schemes must be used. Note that, in all these schemes, the stability of the bias voltage V_{cc} is a common concern.

2.1.4 Temperature compensation

In general, it is possible to study the dependence of the output of a Wheatstone bridge (considering, for simplicity, the full bridge with four active elements) from a certain physical quantity S as:

$$V_s = V_{cc}x = V_{cc}\alpha S$$

where we have defined the following proportionality constant:

$$\alpha = \frac{1}{R_0}\frac{dR}{dS}\bigg|_{R_0}.$$

In reality, this proportionality constant, regardless of the physical quantity we are dealing with, is a temperature-dependent quantity:

$$\alpha = \alpha(T)$$

thus introducing inaccuracies in the output (unless we are measuring the temperature). If we are not measuring the temperature, in fact, a change in it and not in the physical quantity S can be interpreted as a variation of the physical quantity that we would like to measure and, therefore, we want to compensate it. On the other hand, if we are measuring the temperature, our sensor will be non-linear with it (since α has, in general, a non-linear dependence from the temperature), but we can account for this effect.
In general, when we are not measuring a temperature, we need to compensate for its variations, and there are several ways of doing this; we will study only the simplest one. The starting point, in this case, is the derivative of the output voltage with respect to the temperature that, linking variations of the temperature with variations of the output voltage, we would like to have identically equal to zero:

$$\frac{dV_s}{dT} = S\left(\alpha\frac{dV_{cc}}{dT} + V_{cc}\frac{d\alpha}{dT}\right) = 0.$$

From this expression, we can say that the only way of doing it is to ensure that the changes corresponding to the two terms between parenthesis compensate each other. This means that the relative variation of the bias voltage due to the temperature must be equal to the relative variation of the proportionality constant α:

$$\alpha\frac{dV_{cc}}{dT} + V_{cc}\frac{d\alpha}{dT} = 0 \;\rightarrow\; \frac{1}{V_{cc}}\frac{dV_{cc}}{dT} = -\frac{1}{\alpha}\frac{d\alpha}{dT} = -\beta.$$

Therefore, the bridge excitation voltage must be temperature dependent and it must have an opposite rate of variation with respect to α.

A simple temperature compensation scheme that we can implement is to added a fixed resistor in series to the Wheatstone bridge, as it is represented in Figure 2.11. This additional resistance R_T is assumed to be temperature independent (or, at least, with a much smaller dependence than the variable resistors in the bridge) and we can define the equivalent resistance of the bridge R_B when seen from the node at voltage V_{cc} as:

$$R_B \simeq R.$$

In this case, the bias voltage of the bridge can be written as:

$$V_{cc} = V'_{cc}\frac{R_B}{R_B + R_T}$$

and therefore the variation of the bias voltage with respect to the temperature:

$$\frac{dV_{cc}}{dT} = V'_{cc}\frac{R_T}{(R_B + R_T)^2}\frac{dR_B}{dT}$$

where we have considered, as we have said, only R_B to be a temperature dependent quantity. Considering again the expression for V_{cc}, this relationship can be rewritten as:

$$\frac{1}{V_{cc}}\frac{dV_{cc}}{dT} = \frac{R_T}{R_B + R_T}\frac{1}{R_B}\frac{dR_B}{dT}$$

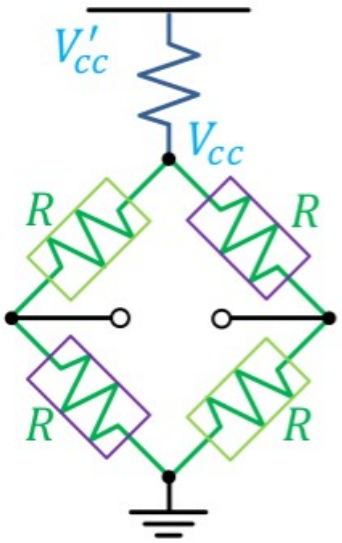


Figure 2.11: A simple temperature compensation scheme for a Wheatstone bridge.

and since we can define the relative variation of the bias voltage as:

$$\frac{1}{V_{cc}}\frac{dV_{cc}}{dT} = -\beta$$

and the relative variation of the resistance of the bridge:

$$\frac{1}{R_B}\frac{dR_B}{dT} = \frac{1}{R}\frac{dR}{dT} = \gamma$$

where γ is the temperature-resistance dependence coefficient, we obtain that:

$$-\beta = \frac{1}{V_{cc}}\frac{dV_{cc}}{dT} = \frac{\gamma R_T}{R_B + R_T} \simeq \frac{\gamma R_T}{R + R_T}.$$

Rewriting this result in terms of the temperature independent resistance:

$$R_T = -\frac{\beta}{\beta + \gamma}R$$

we obtain a very simple and popular solution of compensating for temperature variations. However, this compensation scheme has a few disadvantages. In fact, it is only possible if and only if:

$$\beta < 0$$

otherwise the required resistance will be negative, and if:

$$\beta + \gamma > 0 \;\rightarrow\; \gamma > -\beta = |\beta|.$$

Moreover, the implement it we need to accurately know the temperature dependence of all the resistances, thus precisely knowing both β and γ. Last, it leads

to a reduced output signal and in some applications this may be a problem. This is the reason why this compensation scheme is usually adopted only in the range:

$$25 \pm 15 \ ^\circ\mathrm{C}$$

while for larger temperature ranges we need to implement more complicated compensation schemes.

2.2 Sensor generalities and parameters

We can now deal with sensors from a more general perspective. A sensor is defined as a device that convert an input physical property, also called the *stimulus*, to a different (in our case, electrical) signal. They are, actually, energy converters, passing energy from one form to another one. A certain relation, called the characteristic relation of the sensor, is present between the input range and the output range of the sensor. Most of the times (but not always) both ranges start from zero and they reach a maximum value that is called full-scale. An often used synonymous of range is the word span. It is important to remember that a lot of disquisitions are possible on the formal differences between a sensor and a transducer.
Sensors, in general, can be distinguished depending on various properties:

- the measured quantity, that can be a temperature, a pressure, a velocity, a current, ...;
- the detection mean, that can be biological, chemical, electrical, mechanical, ...;
- the sensor material, that can be a semiconductor, an organic material, a liquid, ...;
- the field of application, that can be a scientific research, an industrial project, a medical application and so on.

Also the characteristics of sensor can be divided into various classes:

- the static parameters, such as the transfer function, the accuracy, the resolution, ..., that are related to the steady-state performances of the system;
- the dynamic parameters, such as the frequency response, the settling time, ..., that are related to the dynamical performances;
- other parameters, for example the operating and storage conditions, the reliability and so on.

To fully characterize a device, a fundamental quantity is the input-output characteristic of the sensor considered. It is a relationship relating the input of the sensor (that is the measured quantity) to its output (in our case, in general a voltage, a current, a charge, a capacitance or any other electrical quantity). When these devices are used as detectors, the inverse function of this relationship is needed, since we want to be able to relate a certain output electrical signal to the input physical quantity that we are measuring.

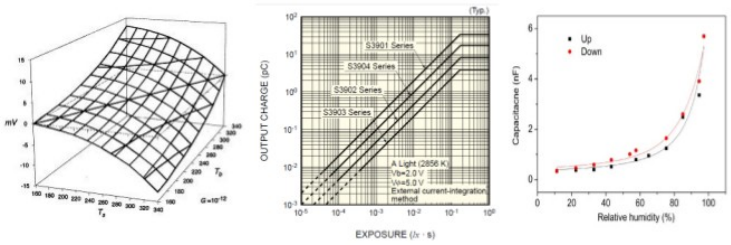


Figure 2.12: Examples of input-output characteristics.

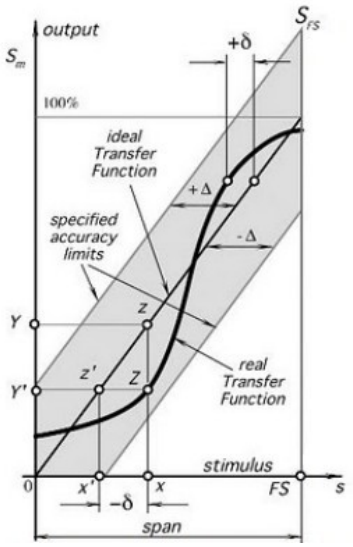


Figure 2.13: Example of input-output characteristics in which we have defined the input and output ranges the full-scale inputs and outputs.

Once we have defined a characteristic relation, we can identify the input and output ranges as the range in which the input and the output of the sensor can vary. Moreover, we call the full-scale input and the full-scale output, respectively, the maximum values of the input and of the output. Both quantities can be identified in the example of characteristic equation that is represented in Figure 2.13.
We can now move to the definition of a few quantities that are important for describing the performances of our system.

2.2.1 Sensitivity

The sensitivity of a sensor is defined as the ratio between the output and the input variations (respectively, dS_o and dS_i):

$$S = \frac{dS_o}{dS_i}.$$

In linear sensors, the sensitivity is constant. However, in general this is not the case, and we can only approximate this quantity as linear through a process

called linearisation, that can be performed (without excessive errors) only over a limited input range. On the other hand, if we do not want (or we cannot) linearise this relationship, we need data processing for inverting the characteristic equation of the sensor and thus trace back the value of the input physical quantity from the output electrical one. It is important to note, moreover, that the characteristics of sensors can also be non-monotonic, and this is an even bigger problem.

2.2.2 Linearity

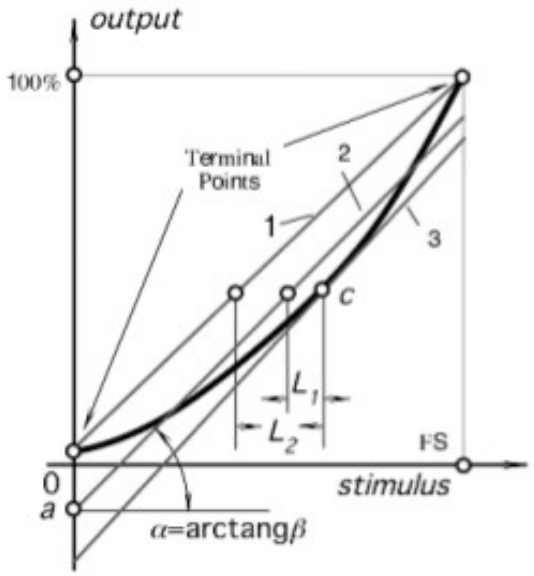


Figure 2.14: Example of linearisation of a characteristic equation.

As we have said before, in general the characteristic equation is a non-linear relationship and, for the sake of simplicity, we can linearise it in a certain interval. Therefore, we define the linearity (or non-linearity) error as the maximum difference between the real transfer function of the sensor and its linear approximation calculated at a fixed output and divided by the full-scale input:

$$\epsilon = \frac{\max(\Delta)}{FS_i}.$$

Note that, in the Figure 2.14, the quantity that we called Δ is represented as L_2 or L_1 depending on the linear approximation considered. In fact, there are several different ways of linearising a transfer function and, therefore, of expressing this error; we need thus to understand what is the one a certain data-sheet is referring to.

The first way of linearising a function, that is represented in the left hand-side of Figure 2.15, is to draw a line that starts from the origin and best approximates the curve (through a least squares approximation). Alternatively, we can define a linear relationship that tracks the edges of the transfer function; this kind of

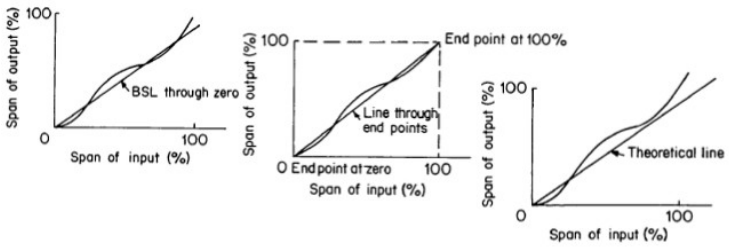


Figure 2.15: Example of different kind of linearisation.

curve is represented in the central graph in Figure 2.15. Last, in the right hand-side of the same Figure the linear characteristic is the theoretical behaviour of the device and, therefore, it can be derived from the underlying theory. It is important to note that all these different linear relations will give different non-linearity errors.

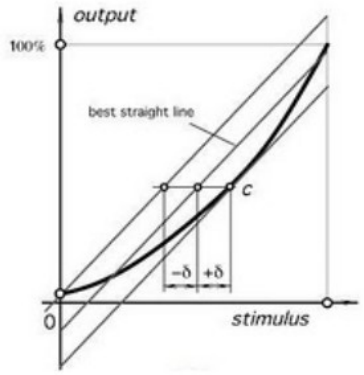


Figure 2.16: Example of best linear relationship approximating the real transfer function of the device.

An way of defining the non-linearity error that is independent from the choice of the linearisation method considered is the independent non-linearity. In this case, we adopt as linear characteristic the straight line that minimizes the maximum absolute non-linearity error. This line can be found by applying a least square approximation starting from a generic straight line on the graph:

$$y = mx + q$$

and determining the two unknown parameters q and m through the minimization of the least square error. In this case, the data-sheet will specify the two parameters m and q to which all other specifications are referred.

2.2.3 Resolution, precision and accuracy

The resolution of a sensor:

$$\frac{\Delta}{FS_i}$$

is defined as the smallest increment in the *stimulus* that can be sensed; it can be specified as an absolute quantity or as a percentage of the full-scale of the input. In principle, in fact, even the smallest increase in the input will lead to a variation of the output. However, in real devices, due to the presence of the noise, this is not true; therefore, we must be able to distinguish this variation on top of the noise. This is the reason why, in real devices, the resolution is ultimately determined by the noise of the sensor itself. Moreover, in ideal devices this parameter will tend to zero: the lower is the resolution, the better is the device. Other factors (such as, for example, the noise in electronics front-end, the digitization and so on) can further degrade it.
A different parameter (even though in common language they are often used as synonymous) is the precision, that in this field is related to the reproducibility of the results. It is defined, therefore, as the ability of the sensor to reproduce the same result after repetitive experiments in the same conditions. It is extremely important to note that precision is *not* resolution: a bad digital clock with a lot of digits may have an high resolution, for example measuring even tiny time intervals, but worse precision, since repetitively measuring the same time interval it will give every time a different measurement. Take care of not using these terms as synonymous.

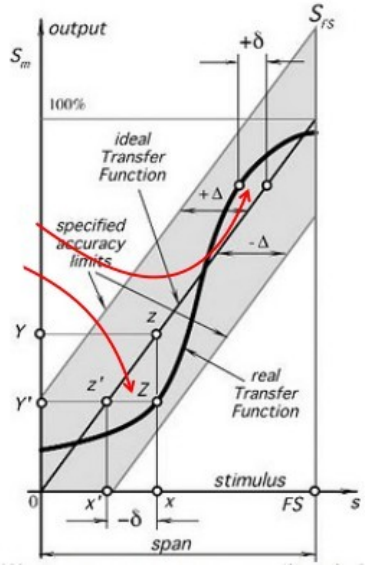


Figure 2.17: An intuitive representation of the concept of accuracy.

The last fundamental parameter is the accuracy of the sensor, that is defined as the maximum deviation from the ideal measured value for many nominally identically sensors. If, in these measurements, we have a strong random component, the average value should be considered.
Therefore, while the precision measures the spread and, therefore, the repeatability of the measurements, the accuracy measures the error, thus the difference between the expected values that come from an ideal characteristic. An intuitive representation of the difference between accuracy and precision is represented in Figure 2.18. Note that the simplest way of increasing the accuracy is to change

the equation describing the ideal behaviour of the device; this process is called calibration.

Figure 2.18: Difference between precision and accuracy.

2.2.4 Dynamic parameters

Under the name of the dynamic parameters we can collect many other parameters that we have already, at least partially, discussed. For example, to this category will belong the frequency response of the system, its response time and its bandwidth.

2.3 Deformation sensors

A first example of resistive sensors that we can study are the deformation sensors, that are used to measure strain. Before studying how it is possible to measure these quantities, we need to briefly study the theory underlying them. Assuming a load F applied on a material in a direction that is perpendicular to one of its surfaces of area S, we define a stress as:

$$\sigma = \frac{F}{S}.$$

This stress will determine a deformation of the material considered; we define strain this deformation per unit length:

$$\epsilon = \frac{\Delta L}{L}.$$

Typical values for a strain are in the order of 10^{-3} and may typically expressed in micro-strains or μstrain, that means in units of 10^{-6}:

$$10^{-3} = 10^{6} \ \mu\text{strain}.$$

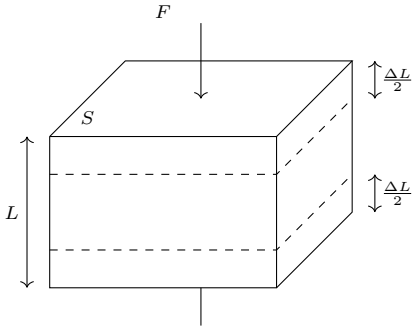


Figure 2.19: A material when a certain stress is applied.

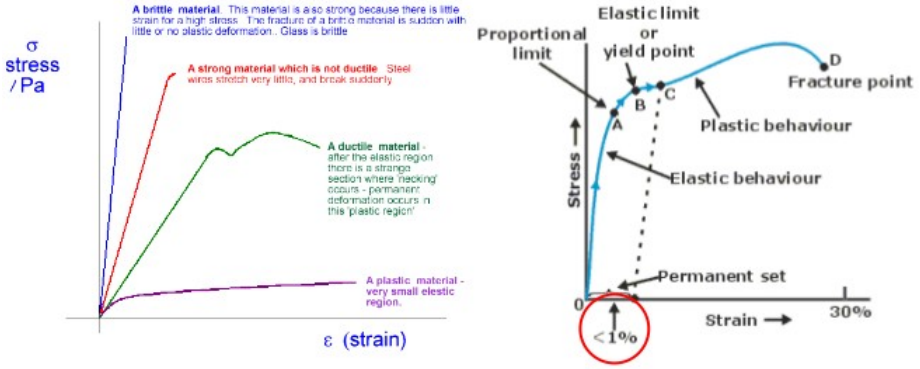


Figure 2.20: Stress as a function of the strain for different classes of materials.

Material, under the effect of a stress, behaves differently. In particular, we can define four different classes of materials:

- brittle materials;
- strong materials;
- ductile materials;
- plastic materials;

and distinguish between them according to their characteristics. In particular, in general we can study the strain of a material as a function of the stress applied, as represented in Figure 2.20, defining the following ranges:

- elastic range: the system is represented by a linear relationship between the stress σ and the strain ϵ according to Hooke's law:

$$\sigma = E\epsilon$$

where E is the so called Young's modulus; in this range every deformation is reversible;

- plastic range: in this range, that is limited on one side from the elastic limit and on the other from the fracture point, the deformation is permanent.

Depending on the presence or absence and size of these regions, it is possible to distinguish between the different classes of materials.
In the elastic region, where as we have said the Hooke's law is valid, an axial strain ϵ_{ax} is always accompanied by a lateral strain ϵ_{lat} of opposite sign in the two perpendicular directions. The ratio between these two quantities is called the Poisson ratio:

$$\nu = -\frac{\epsilon_{lat}}{\epsilon_{ax}}.$$

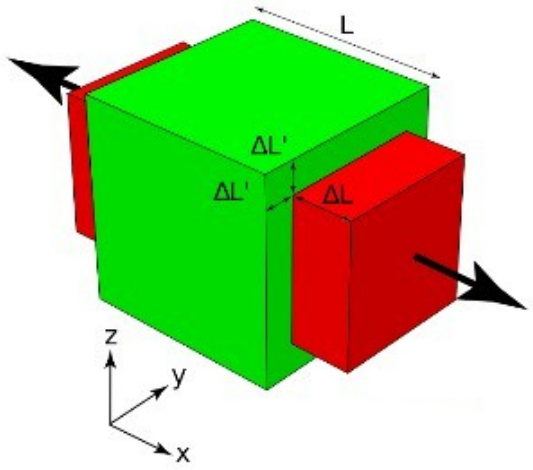


Figure 2.21: Deformation of a material in the various directions.

Referring to Figure 2.21, we can write the elongation in the axial direction as:

$$2\Delta L = \epsilon$$

and therefore the elongation in one of the two non-axial directions will be:

$$-2\Delta L' = -\nu\epsilon.$$

Assuming the original volume to be a cube with length L of the edges, we can write the new volume when a stress is applied as:

$$V' = L^3(1+\epsilon)(1-\nu\epsilon)^2.$$

Expanding this product under the assumption of having a small strain ϵ:

$$\begin{aligned} V' &= L^3(1+\epsilon)(1+\nu^2\epsilon^2-2\nu\epsilon) \simeq L^3(1+\epsilon)(1-2\nu\epsilon) \simeq \\ &\simeq L^3(1+\epsilon-2\nu\epsilon) \simeq L^3\left[1+\epsilon(1-2\nu)\right] \end{aligned}$$

and therefore the relative variation of the volume can be written as:

$$\frac{\Delta V}{V} = \epsilon(1-2\nu).$$

We can thus immediately observe that, if the volume does not change when a stress is applied, the Poisson ratio can be written as:

$$\nu = 0.5$$

and this condition will be satisfied for the majority of the incompressible materials. Moreover, the following theoretical limits are established:

$$-1 < \nu \leq 0.5.$$

Therefore, for compact and weakly compressible materials (liquids and rubbers, in particular) the stress will primarily result in a shape change, therefore:

$$\nu \simeq 0.5.$$

For the majority of the well-known solids (metals, polymers, ceramics, ...) this ratio is slightly lower:

$$0.25 < \nu < 0.35.$$

For glasses and minerals, that are more compressible, this ratio tends to zero:

$$\nu \to 0.$$

For gases, that can be assumed to be perfectly compressible:

$$\nu = 0.$$

A particularly strange class of materials is the one of the materials with negative Poisson's ratio: they are called auxetic materials and an example is represented by Gore-Tex.
In our case, since we are dealing with sensors, we will use metals and we now have to study how it is possible to make use of this phenomenon for creating strain gauges (or gages). The most immediate way of doing it is starting from the second Ohm's law:

$$R = \rho \frac{L}{S}$$

and taking the logarithm of this relationship:

$$\log(R) = \log(\rho) + \log(L) - \log(S)$$

and differentiating it:

$$\frac{\Delta R}{R} = \frac{\Delta \rho}{\rho} + \frac{\Delta L}{L} - \frac{\Delta S}{S}$$

we can obtain the relative variation of the resistance depending on the other factors. In particular, recognizing the variations of the length and of the surface as the strain along the axial and perpendicular directions, we can write:

$$\frac{\Delta L}{L} = \epsilon, \quad \frac{\Delta S}{S} = (1 - \nu\epsilon)^2 - 1 = 1 + \nu^2\epsilon^2 - 2\nu\epsilon - 1 \simeq -2\nu\epsilon$$

and thus we obtain:

$$\frac{\Delta R}{R} = \frac{\Delta \rho}{\rho} + \epsilon(1 + 2\nu).$$

The term $\Delta\rho/\rho$ is called piezoresistivity and is related to the fact that a change in the pressure on a solid can change the band structure of the material and, therefore, also its resistivity. Therefore, this piezoresistivity coefficient is proportional to the strain ϵ and it is generally tabulated. We can thus define the gauge factor as:

$$GF = \frac{\frac{\Delta R}{R}}{\epsilon} = 1 + 2\nu + \frac{\frac{\Delta\rho}{\rho}}{\epsilon}.$$

The gauge factor gives the slope of the relation between the relative variation of the resistance and the strain. From the value of the Poisson's ratio in metals, we expect this gauge factor to be approximately equal to be between 1.5 and 1.8. If the gauge factor is approximately equal to 2, it means that the material we are considering has a large piezoresistivity and, therefore, it is particularly good for a sensor, since this will increase the sensitivity of the device. Typical value of the gauge factor as reported in Table 2.1.

Material	Gauge factor GF
Constantans (Ni-Cu alloys)	$1.8 - 2.2$
Ni-Cr alloys	~ 1.9
Ni	-12
Pt-Ir	~ 5
Doped Si (with impurities)	$\pm 100 - \pm 200$
Poly-Si	± 30

Table 2.1: Gauge factors for certain materials.

A part from the piezoresistivity effect, then, also the temperature dependence of the materials will matter. In fact, since Ni-Cu alloys has a small dependence from the temperature, that is defined through the following coefficient:

$$TCR = \frac{\Delta R}{\Delta T}$$

they are widely used for this kind of sensors. On the contrary, silicon has an high TCR coefficient.

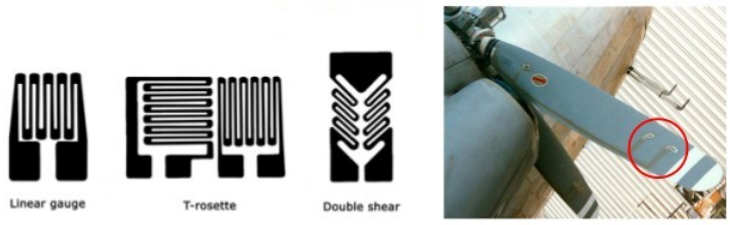


Figure 2.22: Metal-foil strain gauges.

The first example of these sensors is represented by metal-foil strain gauges. These devices, that can have a size that is up to a few centimetres, consists of a grid of fine metallic wire or of a foil. The wire or the file must be bonded (glued) to the strained surface or to a carrier matrix by a thin layer of epoxy, which must transmit the mechanical strain while being an electrical insulator.

The advantages of this device are that it is extremely simple, has a large area, is not expensive and not demanding: all these advantages make this kind of devices extremely common. The disadvantages are that they can work in a limited temperature range and that we must ensure that the deformation of the bulk structure must be equal to the deformation of the metallic wires; in other words, the epoxy must not absorb the deformation.

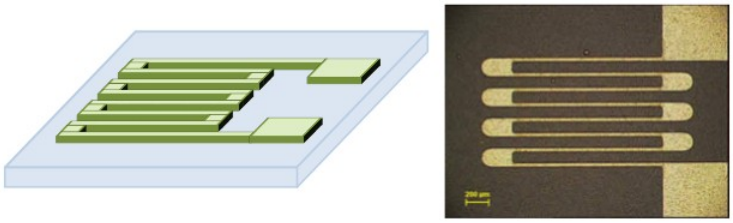


Figure 2.23: Thin film strain gauges.

An alternative solution is represented by thin film strain gauges, that can be built using the thin film technology, for example sputtering individual atoms or molecules, as it is often done in microelectronics. In this case, an insulation layer, typically a ceramic, is deposited on the stressed metal surface and then the strain gauge is deposited onto this layer. Vacuum deposition and sputtering techniques are therefore used for bonding these materials from a molecular point of view. These kind of sensors are much smaller (just a few millimetres) and are directly attached to the substrate, thus having an high range of possible working temperatures.

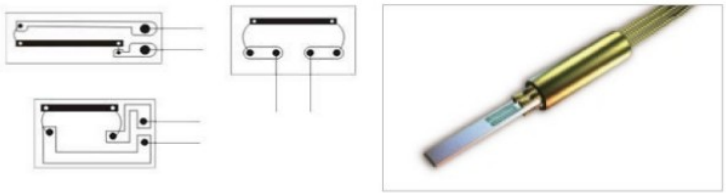


Figure 2.24: Semiconductor strain gauges.

Similar to these devices are semiconductor strain gauges, but in this case on the insulator we are depositing a semiconductor instead of a metal. These devices, since they are built using semiconductors, will have a larger gauge factor GF due to the presence of a significant piezoresistance; however, they are not linear and they are significantly temperature dependent. In the case of larger devices, they can be bonded using the same epoxy that were used for foil gauges, while for smaller ones we can use the same technology that is usually adopted for integrated circuits.
A first, important problem of strain gauges is that they are also sensitive to strains that are perpendicular to a certain longitudinal axis. This is an undesired effect: in fact, we would like to know the direction in which a certain stress is applied and, therefore, we would like to be measuring a stress, with a certain device, exclusively along a given direction. In plane wire strain gauges, this transverse sensitivity is related to the presence of portions in the end loops that

lay in the transverse direction and, therefore, that will be strained even if we are applying a stress in a different direction from the desired one. In foil strain gauges, it is not possible to identify just a single contribution to the transverse sensitivity: it will be the sum of many factors, such as the thickness and the elastic moduli of the backing and of the foil, the with-to-thickness ratio of the foil grid-lines and so on. In general, defining the transverse gauge factor as GF_T and the axial gauge factor as GF_L, we can define the transverse sensitivity factor as:

$$K = \frac{GF_T}{GF_L}$$

and it will be usually between zero and 100%.

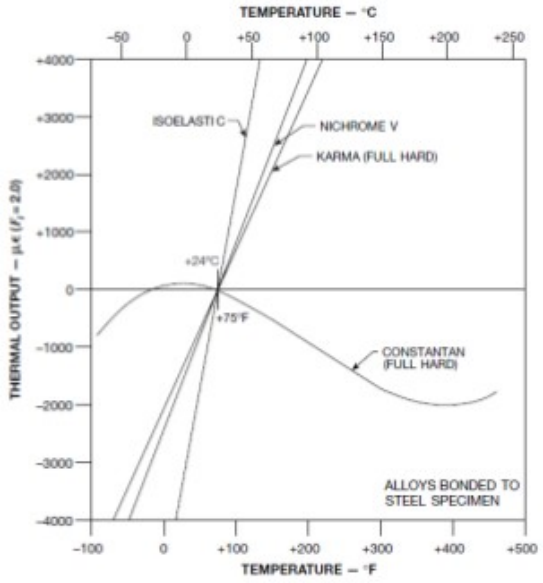


Figure 2.25: Temperature dependent behaviour of a strain gauge.

Another problem is represented by temperature effects. This is a particularly complicated problem since, by varying the temperature, we can determine two main effects:

- direct variations of the resistivity of the material and consequently variations of the the resistance R of the device;
- deformations of the sensor and of the material related to the variations of the temperature and, consequently, variations of the measured resistance.

In general, therefore, the temperature dependence gives the worst error that we can commit and it sets a limit to the strain error; however, we are able to compensate it. To take into account these effects without directly compensating them, we can use suitable correction curves as the one represented in Figure 2.25.

An example of compensation scheme is the dummy gauge compensation. In this scheme, we have a Wheatstone bridge with four sensitive elements where

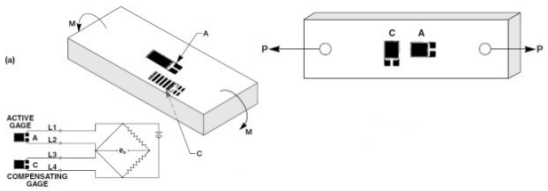


Figure 2.26: Dummy gauge compensation.

one arm is attached to the material and thus varies with both the temperature and the strain, while the other is not attached to the material, thus varying only with the temperature (it does not undergo to any mechanical deformation).
Note that strain sensors are suitable for realizing Kelvin bridges. Considering for example a bar that is bending toward a certain direction. If on one face of the bar, for example the one that is elongating, we have two strain sensors belonging to different arms of the bridge (in the positions that we have previously studied), while on the other side, that is contracting, we have two other strain sensors, then we can determine a variation with same magnitude but different sign of the resistances on different sides, thus allowing us to obtain this 4-wire connection.

2.4 Temperature sensors

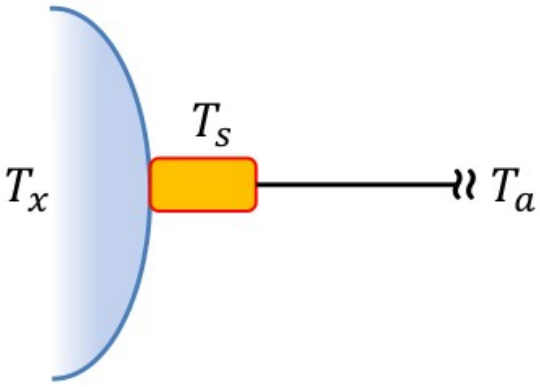


Figure 2.27: A contact sensor at temperature T_s is used for measuring a system at temperature T_x when the environment is at temperature T_a.

A first way of measuring a temperature is to put a certain sensor in contact with the system that we want to measure. This situation is represented in Figure 2.27, where T_x is the temperature of the system, T_s is the temperature of the sensor and T_a the temperature of the environment. The goal, to obtain a correct measurement, is therefore to make the temperature of the sensor T_s to be as close as possible to the temperature of the environment.

From Table 2.2, we can draw an equivalent electrical circuit to the thermal circuit that we are considering as in Figure 2.28. This is due to the fact that, formally, the same equations are relating the quantities considered in the Table

Thermal	Electrical
Temperature [K]	Voltage
Heat flow [W]	Current
Thermal resistance [K/W]	Resistance
Heat capacity [J/K]	Capacitance
$Q = \frac{\Delta T}{R_T}$	$I = \frac{\Delta V}{R}$
$Q = C_T \frac{dT}{dt}$	$I = C \frac{dV}{dt}$

Table 2.2: Equivalence between thermal quantities and electrical quantities.

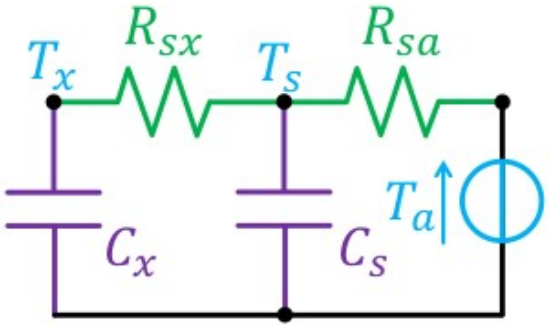


Figure 2.28: Equivalent circuit to the thermal problem.

and therefore every time we have a thermal problem we can draw its equivalent electrical problem, that in principle should be easier to solve. Temperatures, in fact are considered as voltages of certain nodes, thermal resistance R_{sx} and R_{sa} are related to the thermal contacts, the capacitors C_x and C_s are related to the finite heat capacities of the system and of the sensor and we have assumed the environment to have an infinite heat capacity (thus being represented with an ideal voltage source).

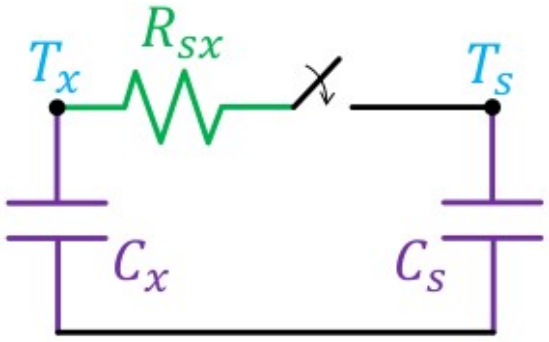


Figure 2.29: Equivalent circuit to the situation of isolated system and sensor.

The first approximated problem that we can consider is when we have an isolated system and sensor, as in Figure 2.29. In this case, at a certain time the switch will close and the sensor will be posed in contact with the system. Since

the two temperatures can be assumed to be initially different:

$$T_x \neq T_s$$

after a transient (that takes a certain amount of time) the system will reach an equilibrium at a certain final temperature:

$$T'_x = T'_s = T_f.$$

To solve this circuit, from an electrical point of view, we have to impose the conservation of the charge before and after the transient:

$$Q = CV = CT$$

and therefore:

$$Q_{in} = Q_f \ \rightarrow \ C_s T_s(0) + C_x T_x(0) = (C_s + C_x) T_f$$

we obtain the following final temperature:

$$T_f = \frac{C_s T_s(0) + C_x T_x(0)}{C_s + C_x}.$$

Again, from the equivalence between electrical quantities and physical quantities, the transient will have the following time constant:

$$\tau = \frac{R_{sx}}{\frac{1}{C_s} + \frac{1}{C_x}}.$$

It is important to note that the final temperature is not necessarily equal to the initial temperature of the system: it can be even significantly different. This is an unwanted effect, since we would like these two values to be at least similar. Since the only parameter that we can modify is the capacity of the sensor C_s, we would like to have this parameter as small as possible, thus to have a small heat capacity of the sensor. It is important to note that the value of the thermal resistance R_{sx} between the sensor and the system will not affect the static value of the temperature; it will only modify the transient. To obtain a fast transient, therefore, we would like to have a small time constant and this implies a small value of this thermal resistance.

We can then consider the case in which also the environment is present but the system is assumed to have an infinite thermal capacitance. In this case, we can write the voltage generator T_a as the series between a voltage generator $T_a - T_x$ and a voltage generator T_x. Applying then the superposition principle, we can write that:

$$T_a - T_x = 0 \ \rightarrow \ T_{f1} = T_x$$

while, on the other hand:

$$T_x = 0 \ \rightarrow \ T_{f2} = (T_a - T_x) \cdot \frac{R_{sx}}{R_{sx} + R_{sa}}.$$

This means that the final temperature will be the sum of these two partial final temperatures:

$$T_f = T_x + (T_a - T_x) \cdot \frac{R_{sx}}{R_{sx} + R_{sa}}$$

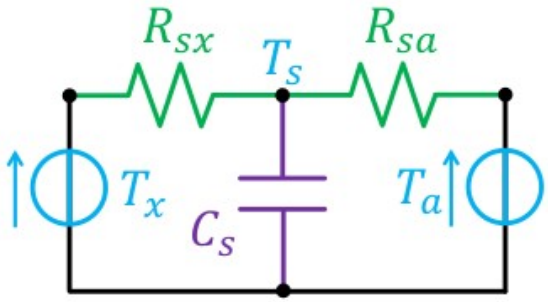


Figure 2.30: Equivalent circuit when the environment is present and the system has an high thermal capacitance.

and, therefore, the second term is actually the error with respect to the quantity we would like to measure. The minimization of this error intuitively leads to have a large ratio R_{sa}/R_{sx} and this means that we would like to have a bad thermal contact between the sensor and the environment and a good thermal contact between the sensor and the system to improve the accuracy of the sensor. In this case, the time constant of the transient will be:

$$\tau = C_s \cdot (R_{sx} \| R_{sa}) \,.$$

Therefore, we can translate these ideal requirements in real devices as it follows:

- the small heat capacity of the sensor C_s can be obtained using small sensors;
- the small thermal resistance between the sensor and the system can be obtained through a good thermal contact, thus maximizing the contact[2] are between them and, for solids, using good thermal grease;
- the large thermal resistance between the sensor and the environment can be obtained using the correct sensor connections, thus designing sensors with long and narrow connections, with a low thermal conductivity (for example, using stainless steel) and good electrical conductivity.

2.4.1 RTD

We can now start our overview on different temperature sensors. The first sensors that we will study are called Resistance Temperature Detectors (RTDs) and they exploit the variation in the resistance of certain metals with temperature. They can provide highly accurate results, with a minimum detectable temperature difference that goes from 0.1 °C to 0.0001 °C, and they are usually adopted in the temperature range between 14 and 1200 K. In general, in a metal the resistance increases with the temperature due to the fact that the probability of a scattering event between an electron and a phonon increases. We can thus expand the dependence of the resistance in power series:

$$R = R_0 \left(1 + \alpha_1 T + \alpha_2 T^2 + \cdots + \alpha_n T^n\right)$$

[2]Note that this requirement seems to be in contrast with the previous one.

observing from Figure 2.31 that in general this relationship can be approximated as linear over a quite large range of temperatures. In this interval where we can assume to have a linear dependence, the sensitivity of the device will be related to the slope of this curve.

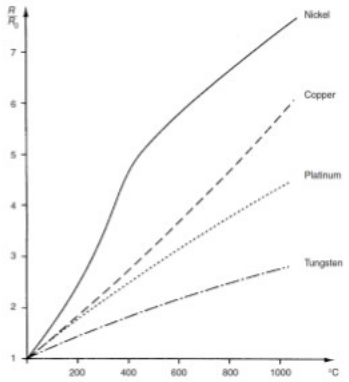


Figure 2.31: Temperature dependence of the resistance of a few common metals.

We can thus define the so called temperature coefficient of the resistance TCR as:

$$TCR = \frac{1}{R_0} \left. \frac{dR}{dT} \right|_{R_0}$$

and it will be related to the sensitivity of the sensor. In the interval in which the variation of the resistance with the temperature is linear, we can write this value as:

$$\alpha = TCR = \frac{1}{R_0} \frac{\Delta R}{\Delta T}$$

and in general this holds between 100 °C and 0 °C. In Table 2.3 are reported a few temperature coefficients of resistance with the corresponding linearity range. From this Table, it is possible to observe that for pure platinum this linearity range is pretty large, thus making it a quite common choice for Resistance Temperature Detectors. Moreover, we can note that the temperature coefficients are in general quite small.

Metal	Temp. range (°C)	TCR ($°C^{-1}$)
Pt	$[-200, +850]$	$3.85 \cdot 10^{-3}$
Ni	$[-100, +200]$	$6.72 \cdot 10^{-3}$
Cu	$[-100, +250]$	$4.27 \cdot 10^{-3}$
W	$[-100, +400]$	$4.8 \cdot 10^{-3}$

Table 2.3: Temperature coefficient of resistance and linearity interval for a few metals.

In general, platinum is the mostly used metal for Resistance Temperature Detectors. Its first advantage, in fact, is that it is inert from a chemical point of view and this is particularly important when we are working at extremely high temperatures, where a lot of chemical reactions are thermally activated,

possibly leading to the chemical contamination of the sensors. Moreover, it has a large enough temperature coefficient of resistance, even though it is smaller than the one of other materials, and its resistance-temperature relationship is almost linear over a large range of temperatures. Last, the fabrication process can be in general free from strains and the resistance R resulting from it is only weakly dependent on the strain. The main disadvantage is that it is extremely expensive. In general, these kind of sensors are built in order to have a fixed resistance at the reference temperature of 0 °C, for example equal to 100 Ω (thus giving the PT100 Resistance Temperature Detectors) or to 1000 Ω (in this case, we have the PT1000 Resistance Temperature Detectors); this value of resistance can in general be controlled during the manufacturing process.

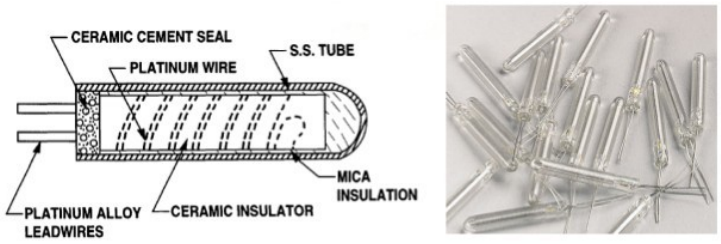


Figure 2.32: Wire-wound Resistance Temperature Detectors.

From a practical point of view, a first way in which these devices can be built is the so called Wire-wound Resistance Temperature Detectors, that are represented in Figure 2.32. In these devices, a small sensing platinum wire (in general with a diameter between 7 and 50 μm) is wound around a cylindrical ceramic mandrel. This wounding must be non-inductive: this means that we do not want to create loops around the mandrel, otherwise the system will have a certain value of inductance and thus it will be sensitive also to magnetic fields. Therefore, we want to reduce as much as possible the area of the loop and this can be done coupling the wire with itself and then wounding the two different ends of the same wire together around the insulator. The mandrel and the wire are then usually covered with a thin layer of material that will provide an electrical insulation and a mechanical protection of the device. The whole device (that can be approximately of the size of the mandrel) has a length between 2 and 3 cm, with a diameter of the mandrel between 1 and 5 mm. The disadvantage of these devices, however, is that they are in general extremely sensitive to vibrations and mechanical shocks, therefore a better variation is represented by the coil suspension Resistance Temperature Detectors.

An example of coil suspension RTD is represented in Figure 2.33. In these devices, a small coil of fine platinum wire is assembled into small holes inside a cylindrical ceramic mandrel. In each device, we have a pair of these coils. These coils are supported by a ceramic powder that is needed to stabilize the structure from a mechanical point of view; both ends of the mandrel are sealed. The ceramic powder, therefore, will allow the expansion and contraction of the platinum wire, thus reducing the effects not only of vibrations and mechanical shocks but also the effects of strains. Typically, these device have a length between 10 and 30 cm and a diameter between 5 and 6 mm. These kind of devices are suitable for many applications, but surely not for all: due to their quite big

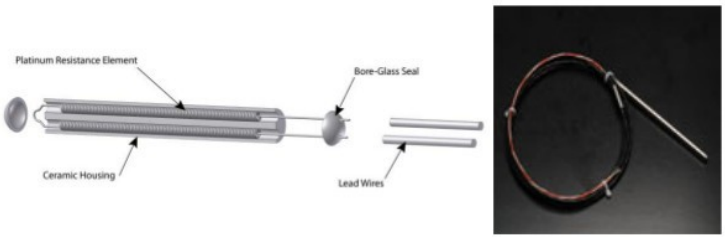


Figure 2.33: Coil suspension Resistance Temperature Detectors.

dimensions, they will not be suitable for measuring local temperatures, where smaller sensors will be needed. However, they can provide a large contact area between the sensor and the system and, depending on the application, this might be a significant advantage.
Note that, in this case, since the platinum wire is not directly attached to the mandrel, the thermal expansions and contractions of the mandrel and the consequent strains will be much less relevant in coil suspension RTDs with respect to wire-wound RTDs.

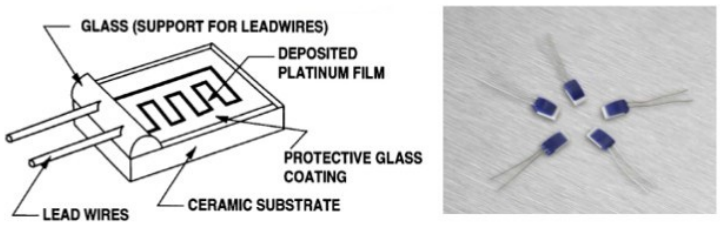


Figure 2.34: Thin-film RTDs.

A much smaller alternative to the previous devices are the thin-film RTDs, that are represented in Figure 2.34. In these devices, that are built using the same techniques usually adopted for thin-film strain gauges, a thin film of platinum is deposited onto a ceramic substrate and then it is etched, leaving the element pattern, that is finally covered with a glass material to protect the device from humidity and contaminants. In general, the length of these devices is between 1 and 10 mm, with an height between 1 and 2 mm: they are thus much smaller than the previous sensors. This gives some advantages: first of all, we are now able to sense very localized temperatures (while in previous detectors we were averaging them). Then, these devices have a very small thermal capacitance, thus having a small time constant, being particularly fast sensors. Last, due to their small size they are not expensive.

2.4.2 Thermistors

The idea underlying these devices is similar to the one beneath RTDs, but in this case we are using different elements. In particular, now we are dealing with transition metals oxides, such as chromium (Cr), cobalt (Co), copper (Cu), manganese (Mn) and nickel (Ni), that will show a semiconductor-like behaviour.

They are characterized by a strongly non-linear resistance-temperature characteristic, but with a quite high temperature coefficient of resistance, that can be either positive (thus being indicated as PTC) or negative (NTC). In general, only negative temperature coefficients (NTCs) are useful for sensing applications, while PTCs will have other uses.
Either from the Maxwell-Boltzmann or the Fermi-Dirac statistics, we know that the carrier density in a semiconductor (or in an oxide behaving similarly) can be related to the exponential of the temperature through suitable coefficients. This dependence is reflected in the resistance-temperature characteristic of these devices, that can be written as:

$$R(T) = R(T_0)e^{B\left(\frac{1}{T}-\frac{1}{T_0}\right)}.$$

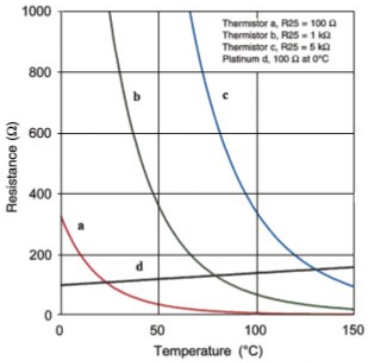


Figure 2.35: Resistance-temperature characteristic of certain thermistors.

It is important to note that, due to this strongly non-linear characteristic, the same minimum detectable change in resistance, when centred across two different regions of the curve, will give a different minimum measurable variation in temperature. This means that the resolution of the sensor is not constant: given a minimum measurable value ΔR, the associated temperature variation ΔT will depend on the temperature T. Depending on the applications, this might or might not be a useful effect. In general, the reference value for the resistance of such a device is referred to a temperature of 25 °C and the associated resistance can vary, depending on the temperature, between 100 Ω and 100 kΩ.
From the definition of temperature coefficient of resistance TCR that we have previously given:

$$TCR = \frac{1}{R_0}\left.\frac{dR}{dT}\right|_{R_0}$$

substituting the expression for the exponential temperature dependence of the resistance we can obtain:

$$TCR = \alpha = -\frac{B}{T^2}.$$

Typical values for this coefficient are between -3 and $-5 \cdot 10^{-2}$ °C^{-1}, thus being approximately one order of magnitude larger than what we had in RTDs. Therefore, for the same temperature variation a thermistor will give rise to

a larger signal, thus allowing us to perform the detection using a standard Wheatstone bridge with just one active element. NTC thermistors are generally used between -50 °C and 150 °C, with the upper limit that can be extended up to 300 °C for some glass-encapsulated units.

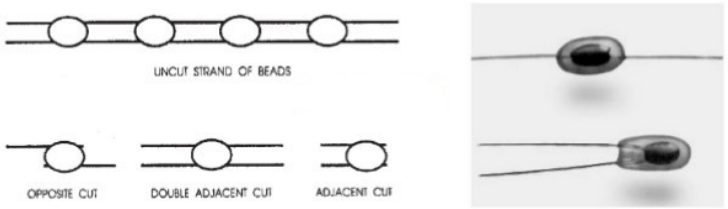


Figure 2.36: Bead thermistors.

From a practical point of view, a first way of creating these devices is represented by bead thermistors, that can be observed in Figure 2.36. In these devices, around two parallel wires of platinum is placed a certain mixture of metal oxide and binder. When this material condenses around the wires, it forms some beads. The obtained strand is then sintered, allowing the contacts to form intimate bonds with the thermistor. These beads are then cut depending on the desired geometry and coated. Each bead will have a size approximately equal to 1 mm.

Figure 2.37: Surface contact thermistors.

An alternative that can be adopted using thin-films technology are surface contact thermistors, that are represented in Figure 2.37. They are in general fabricated by layer deposition (tape-casting on a chip or compressed metal powders in disks), with the following creation of metal contacts (being applied either by spraying, painting or sputtering and then being fired onto the ceramic body). These devices are much smaller than 1 mm and there exist also leadless versions for hybrid- or surface-mount types.

PTC thermistors, on the other hand, have a resistance R that increases with temperature, on the contrary with what we had with the thermistors that we have previously studied. This means that their temperature coefficient of resistance is positive:

$$TCR > 0.$$

However, they are almost never used as sensors, since they are largely unstable, have a weird behaviour above the Curie point and have an extremely small temperature range of linearity. They are obtained from polycrystalline ceramic materials that are made semiconductive by the addition of dopants. They are

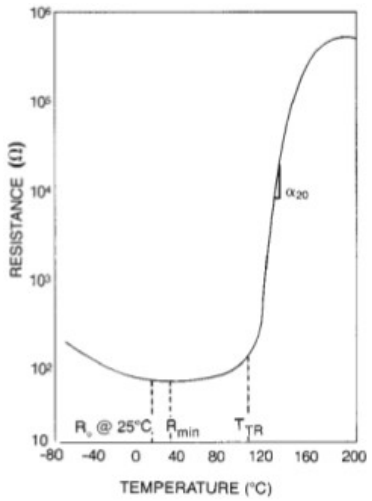


Figure 2.38: Characteristic equation for a PTC thermistor.

generally used as overcurrents protections or heaters. In fact, in the case of an overcurrent, the Joule effect increases the power dissipation in the circuit and, therefore, also the temperature increases. At a certain temperature, however, the resistance of these devices significantly increases, actually shutting down the whole circuit before it is burned.

2.4.3 Sensor self-heating

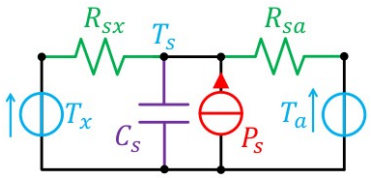


Figure 2.39: Equivalent thermal circuit for self-heating.

In sensors, as in any other device in an electrical circuit, a current flowing will lead to a certain power dissipation that is related to the Joule effect. This generally leads to an increase of temperature of the circuit and, as we can immediately understand, this is a particularly bad effect in temperature sensors, representing a source of errors: this effect is called self-heating. To study it, we can consider the equivalent circuit that is represented in Figure 2.39. Neglecting the presence of the environment (and thus the right hand-side of the circuit):

$$R_{sa} \to \infty$$

by linear superposition, first switching off the infinite thermal capacitance of the system T_x and then the self-heating power P_s, after a transient we can obtain the following value for the temperature of the sensor:

$$T_s = T_x + P_s R_{sx}.$$

Therefore, the term $P_s R_{sx}$ will represent a source of error in our temperature measurement: the so called self-heating error. The power dissipation in the sensor, therefore, can be represented as a power flow from the sensor to the system. Assuming for example to have a PT100 sensor through which flows the following current:

$$I = 1 \text{ mA}$$

with the following thermal resistance between the system and the sensor:

$$R_{sx} = 1 \text{ °C/mW}$$

we can obtain that the sensor has a resistance R_s and thus it dissipates a power P_s:

$$R_s = 100 \ \Omega \ \Rightarrow \ P_s = 100 \ \mu\text{W} = 0.1 \text{ mW}$$

thus giving the following self-heating error:

$$\Delta T = P_s \cdot R_{sx} = 0.1 \text{ °C}.$$

This value seems to be pretty small, and for many applications it can be considered negligible, but it actually depends on the application we are considering. Moreover, the considered values represent a pretty fair case: things can be much worse in reality.
To reduce the self-heating error, one possibility is to improve the thermal contact between the sensor and the system, thus lowering the value of the thermal resistance R_{sx}. Alternatively, if this is not possible, we can try to reduce the current I flowing through the sensor. This necessarily implies a reduction of the bias voltage V_{cc} of the Wheatstone bridge used for this kind of sensor, thus lowering the output signal of the bridge and possibly degrading its resolution.
Alternatively, it is possible to adopt a different perspective. In fact, since the self-heating mechanism has a certain time constant that depends on the thermal capacitance of the sensor C_s, instead of driving the Wheatstone bridge with a DC voltage we can use as a bias voltage V_{cc} a rectangular wave signal (therefore, an AC signal). If the time intervals in which the signal V_{cc} is at its high level are smaller than the time constant for self-heating, the sensor will not have the time to significantly modify its temperature due to self-heating:

$$\Delta t \ll R_{sx} C_s = \tau.$$

The equivalent circuit, in this case, will be similar to the one represented in Figure 2.39, with the addition of a switch in series to P_s. Every time the switch is closed, therefore when V_{cc} is high, the system will be measuring the temperature and thus it will be (slightly) self-heating, while when the circuit is open and the bias voltage V_{cc} is equal to zero we will not have any current flowing through the sensor.

2.5 Thermoelectric effect and thermocouples

Under the name of thermoelectric effects are grouped a large class of quite interesting and difficult effects. From our perspective, we will focus only on the Seebeck effect, that is exploited in a class of temperature sensors: the thermocouples.

The Seebeck effect, as many physical effects, was accidentally discovered by T. Seebeck in 1826. According to this effect, a temperature difference between two conductors (or semiconductors) generates a voltage difference or a current flow. Therefore, a thermal gradient translates in an electric field or in an electromotive force, therefore in a voltage gradient. The vice versa occurs in the dual phenomenon: the Peltier effect.

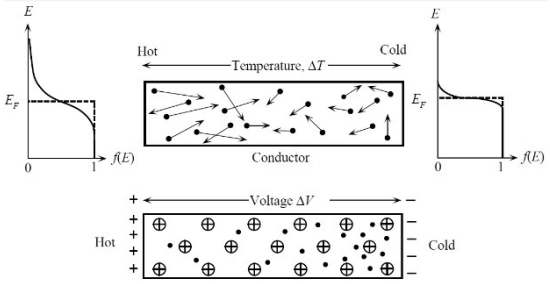


Figure 2.40: Physical picture of the Seebeck effect.

To understand the physics of this phenomenon, consider a certain slab of a conductor (either a metal or a semiconductor), as it is represented in Figure 2.40, to which a certain temperature gradient ΔT is applied. From previous courses on solid state physics, we know that the energy distribution of the electrons is controlled by the Fermi-Dirac distribution, that depends on the temperature. This means that at the hot end of the slab we will have an higher probability of having electrons with high kinetic energy and a lower probability of having electrons with a low kinetic energy, and vice versa at the cold end. This means that, at the hot end of the slab, the electrons will have, on average, an higher kinetic energy and they will tend to diffuse towards other regions of the slab, included the cold side. This means that, at some point, the density of the electrons in the material will be higher at the cold side than at the hot side and, therefore, a certain electric field directed toward the cold end will be established. In steady-state conditions, this electric field will be such that the drift of the electrons toward the hot side of the slab due to the presence of the electric field will be perfectly balancing the diffusion process due to the difference in temperature between the two ends of the slab. Since the presence of an electric field is always related to a voltage gradient inside the material, we will have a certain voltage drop ΔV inside the material and thus, from the direction of the electric field, the hot end of the conductor will be at an higher voltage with respect to the cold one. From a microscopic point of view, we can say that a gradient in the concentration of the electrons (that is responsible for the electric field and thus the voltage difference) is able to compensate the effects of a gradient in temperature. Considering Fermi-Dirac statistics[3], in solid state physics it is possible to calculate the average energy of the electrons, depending on the temperature, as:

$$\mathcal{E} = \frac{3}{5}\mathcal{E}_F \left[1 + \frac{5\pi^2}{12}\left(\frac{k_B T}{\mathcal{E}_F}\right)^2\right]$$

[3]This physical model and the following one that is used for explaining the Seebeck effect in semiconductors have not been discussed in details during the lecture.

and therefore, differentiating this relationship, it is possible to calculate the variation of energy related to a variation ΔT of the temperature:

$$\Delta\mathcal{E} = \frac{\partial\mathcal{E}}{\partial T}\Delta T = \frac{\pi^2 k_B^2 T}{2\mathcal{E}_F} \cdot \Delta T$$

that will be balanced by the electrostatic energy $-q\Delta V$. Equating therefore these two relationships, we can define the Seebeck coefficient[4] as:

$$S = \frac{\Delta V}{\Delta T} = -\frac{\pi^2 k_B^2 T}{2q\mathcal{E}_F}$$

thus determining a relationship between the temperature gradient ΔT and the voltage gradient ΔV. It is important to note that ΔV will be the voltage of the cold side of the conductor, since in general the voltage of the hot end is taken as a reference value.
From a more general point of view, the Seebeck coefficient can be defined as:

$$S = \frac{dV}{dT}$$

and from the previous convention of the voltages its sign will be equal to the sign of the voltage of the cold side with respect to the hot side. A few Seebeck coefficients are reported in Table 2.4.

Metal	S (μV/K)	Metal	S (μV/K)
Sb	42	Bi	-68
Li	14	K	-13
Mo	4.7	Pd	-9
Cd	2.6	Na	-6.5
W	2.5	Pt	-4.5
Cu	1.6	C	-2
Ag	1.5	Al	-1.6
Ta	0.05	Pb	-1.1

Table 2.4: Seebeck coefficients for a few reference metals at 0 °C.

The important point that we need to observe from this Table is that these coefficients are in general really small and that it is possible to have a sign difference between them.

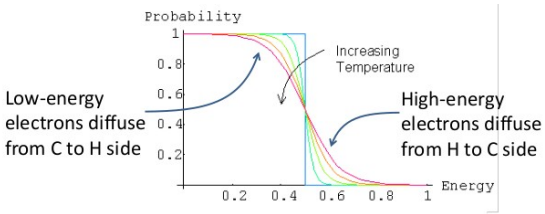


Figure 2.41: Fermi-Dirac distribution plotted for different temperatures.

[4] Also called thermoelectric power, even though it is definitively not a power.

To understand why this sign difference is possible, we need to consider the Fermi-Dirac distribution that is represented in Figure 2.41. In particular, we can observe that depending on the temperature we will have not only more or less high-energy electrons that will diffuse from the hot end to the cold one, but also that it is possible to have more or less low-energy electrons that can diffuse from the cold end to the hot one. In the previous description, we have implicitly assumed that an higher energy of the electrons were related to an higher kinetic energy, therefore to an higher velocity and an higher mobility of the electrons coming from the hot side, but cannot always be taken from granted. The sign of the Seebeck coefficient S, therefore, is related to the energy dependence of the diffusion coefficient and thus to the efficiency of the scattering mechanisms at different energies. In metals with a negative Seebeck coefficient, the probability of a scattering event is much higher for high-energy electrons than for low-energy ones, and therefore the diffusion of the electrons will take place from the cold side toward the hot side, determining the apparently strange sign of the Seebeck coefficient. This presence of negative coefficients was completely surprising at Seebeck's time and, depending on the scattering mechanisms, it has been satisfactorily described only after the development of the solid state theory.

Semic.	S (μV/K)
Se	900
Te	500
Si	435
Ge	300
PbTe	-180
PbGeSe	-2000 or +1700
BiTe	-230

Table 2.5: Seebeck coefficient for some reference semiconductors.

In Table 2.5 are reported a few Seebeck coefficients for some semiconductors. It is immediately possible to note that these coefficients are much larger than the one we have obtained in metals; this is due to the fact that in metals the density of carriers is fixed, while in semiconductors it will be exponentially dependent from the temperature[5].

We define thermocouples the sensors that exploit the Seebeck effect. A first tentative of creating a thermocouple is represented in Figure 2.42. In this case, we are connecting two points at different temperature:

$$T_2 > T_1$$

with a wire of a certain metal (A). Then, since we want to measure the voltage drop across the two ends of this wire, we connect them to a voltmeter (a device that is used for measuring voltage differences) with two wires of the same metal (A). Also the voltmeter, being a physical device, will be at a certain temperature T_0 that, for example, we can assume to be between T_2 and T_1. Across the first wire, connecting the hot and the cold regions, there will be a certain voltage

[5] A simple model for the calculation of the Seebeck coefficient in this case can be found on the slides, even though it has not been discussed during the lecture.

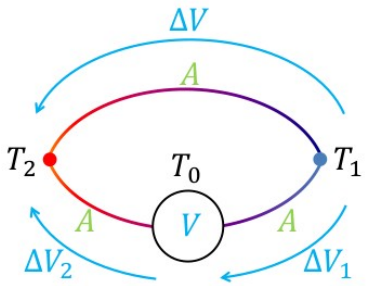


Figure 2.42: A first and not successful tentative of creating a thermocouple.

drop that is related to the Seebeck coefficient S_A of the metal considered and to the temperature difference between the two ends:

$$S_A(T_2 - T_1) = S_A \Delta T.$$

However, also the connections from T_2 to T_0 and from T_0 to T_1 will give, for the same effect, a certain voltage drop, that we can write as:

$$\Delta V_2 = S_A(T_2 - T_0), \quad \Delta V_1 = S_A(T_0 - T_1).$$

Therefore, the voltage drop that we can measure with our voltmeter will be equal to the voltage drop across the metallic wire minus the two drops related to wirings of the voltmeter:

$$\begin{aligned} \Delta V &= S_A(T_2 - T_1) - S_A(T_2 - T_0) - S_A(T_0 - T_1) = \\ &= S_A(T_2 - T_1) - S_A(T_2 - \cancel{T_0} + \cancel{T_0} - T_1) = 0. \end{aligned}$$

Therefore, we have obtained that this device, in this configuration, will not measure any voltage regardless of the temperature T_0 at which we have placed the voltmeter. Therefore, we need to change the setup in order to be able to measure a certain voltage difference.

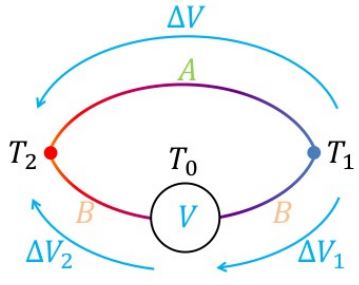


Figure 2.43: Drawing of a thermocouple.

To change the previous device, we can assume, as in Figure 2.43, to use a different metal (B) for wiring the hot and cold ends of the wire (A) to the voltmeter. This second metal, therefore, will have a different Seebeck coefficient S_B.

From the previous reasoning, therefore, the voltage difference that is measured across the voltmeter can be written as:

$$\begin{aligned}\Delta V &= S_A(T_2 - T_1) - \Delta V_2 - \Delta V_1 = \\ &= S_A(T_2 - T_1) - S_B(T_2 - T_0) - S_B(T_2 - T_0) = \\ &= S_A(T_2 - T_1) - S_B(T_2 - \cancel{T_0} + \cancel{T_0} - T_1) = \\ &= (S_A - S_B) \cdot (T_2 - T_1) = S_{AB}\Delta T.\end{aligned}$$

In this way, therefore, we are actually observing a voltage drop that is proportional to the difference between the hot and cold temperature through the difference of the Seebeck coefficients. This means that for any given thermocouple A-B we need to know the difference between the two Seebeck coefficients.

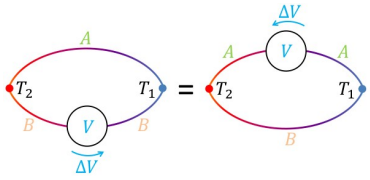


Figure 2.44: Regardless of the position of the voltmeter, a thermocouple will give the same voltage drop.

Summing up, a circuit made with just one conductor will not generate any voltage drop regardless of the temperature gradient, as well as a circuit made with two different conductors will not generate any voltage drop if there is not any temperature gradient: the only way of obtaining a voltage drop is to use different materials whose ends are at different temperatures. In this case, the thermocouples can be considered similar to sources of electromotive force, since regardless of the position of the voltmeter we will always obtain (if present) the same voltage drop. If we close the loop without placing any voltmeter, a current I will flow through this loop, generating a magnetic field (that can be sensed) and dissipating some energy over the small (but finite) resistances of the metals. Considering the last equation that we have written, from it we can derive the law of intermediate temperatures:

$$\begin{aligned}V &= S_{AB}(T_2 - T_1) = S_{AB}(T_2 - T_0 + T_0 - T_1) = \\ &= S_{AB}(T_2 - T_0) - S_{AB}(T_1 - T_0)\end{aligned}$$

that states that if the voltage drops can be related to a known reference temperature T_0, then we are able to compute the electromotive force for any temperature difference ΔT. Analogously, we only need to know the temperature T_1 to determine the dependence of the Seebeck coefficient on the second temperature T_2. Moreover, this relationship is valid also if the Seebeck coefficient is a function of the temperature:

$$\int_{T_1}^{T_2} S_{AB}(T)\ dT = \int_{T_1}^{T_0} S_{AB}(T)\ dT + \int_{T_0}^{T_2} S_{AB}(T)\ dT.$$

Alternatively, we can derive the law of intermediate metals:

$$\begin{aligned}V &= (S_A - S_B)\Delta T = (S_A - S_M + S_M - S_B)\Delta T = \\ &= (S_A - S_M)\Delta T - (S_B - S_M)\Delta T\end{aligned}$$

that states that if the electromotive forces with respect to a reference electrode are known, they can be used to compute the electromotive force for any couple of metals. Again, also this relationship is valid if the Seebeck coefficient is dependent on temperature. However, due to the possibility of chemical modifications at the junctions between the metals, this last law will hold only as a first order approximation, while the law of intermediate temperatures will have a broader validity.

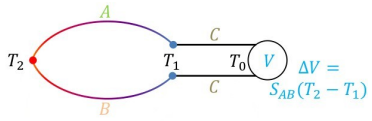


Figure 2.45: A measurement with a thermocouple.

We can now study how it is possible to perform a measurement from a practical point of view. The first thing that we need to make sure is that the soldering metals, that are represented as dots in Figure 2.45, are placed at the same temperature T_1 or T_2 depending on their position: this will make us sure of the fact that they will not generate any additional voltage drop. Moreover, the voltmeter, that will then send a signal to an amplifier, must be connected to the thermocouple by two wires of the same material C: their voltage drop will cancel out, while if we were using two different metals these contributions will lead to an additional voltage drop. The voltage measured by the voltmeter can thus be written as:

$$\begin{aligned}\Delta V &= \cancel{S_C(T_1 - T_0)} + S_A(T_2 - T_1) + S_B(T_1 - T_2) + \cancel{S_C(T_0 - T_1)} = \\ &= S_A(T_2 - T_1) - S_B(T_2 - T_1) = S_{AB}(T_2 - T_1).\end{aligned}$$

The problem, now, is how it is possible to measure an absolute temperature T_2. In fact, from a thermocouple we are only able to measure the temperature difference $T_2 - T_1$, therefore we need to know the value of T_1 with a certain accuracy (and stability) if we want to be able to measure the absolute value of T_2.

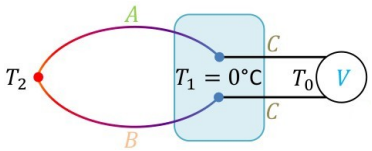


Figure 2.46: A thermocouple with an ice bath.

One of the first arrangements for measuring absolute temperatures is the ice bath represented in Figure 2.46. In this case, the temperature T_1 is kept at the reference, stable value of 0 °C using an ice bath: this gives a constant offset voltage that can be subtracted. However, this is impractical in modern systems, since we need always a bath at a reference temperature.

A more recent scheme, that is represented in Figure 2.47, is the cold junction compensation technique. In this circuit, the temperature T_1, that is measured through a thermistor or an RTD and its contribution is subtracted from the

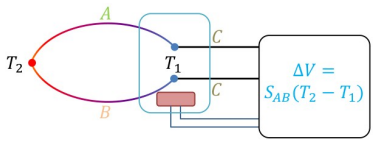


Figure 2.47: A thermocouple with a thermistor or an RTD.

output voltage of the thermocouple. Since we are actually measuring the absolute value of temperature T_1, we do not it to be stable, at least within a range in which the adopted thermistor or RTD works properly. We are thus inferring the absolute value of the temperature T_2 from its relative measurement and from an absolute measurement of the first temperature T_1. A question, at this point, may arise: why do not we directly measure the temperature T_2 with a thermistor or an RTD, instead of performing two measurements? In general, this is because thermocouples are used for extremely low or high temperatures, outside the usage ranges for thermistors and RTDs. Therefore, we can use these device to measure a temperature that in general is the temperature of the environment, thus being stable, in the linearity range of the device and in a region in which we have a good enough resolution (that will be actually better in the case of thermistors) and use this measurement as a reference for the one of the extreme temperature T_2. Moreover, the same temperature T_1 can be used as the reference temperature for many different thermocouples, thus allowing us to have multi-point measurements, that are especially useful in industrial environments.
A few of considerations, when performing this kind of measurements, must be taken into account. First of all, we want to measure an open-circuit voltage, therefore the input impedance of the voltmeter or of the amplifier that is connected to the thermocouple must be large. If this is not the case, we are actually measuring the partition of the voltage related to the thermocouple over the series of the input impedance Z_{in} and the small resistances of the metals used, due to the fact that a current will flow through the whole thermocouple, determining a voltage drop across these resistances. Moreover, since we have previously demonstrated that a good sensor must have a small thermal capacitance, we can use thin thermocouple wires. However, these wires will give, from the second Ohm's law, and high value of the resistance and, consequently, an high contribution to the noise: to counteract these effect we must keep the thermocouple as short as possible and we must use very thick connection wires. Then, since the temperature T_2 that is measured through the thermocouple is generally quite high, a lot of chemical reactions will be activated and accelerated by this temperature, contributing to the chemical contamination of the metals of the thermocouple. This will change the chemical composition of the junction between the two metals, de-calibrating the thermocouple over time due to a change in the Seebeck coefficients. This means that, in general, thermocouples have a short lifetime due to their exposure to an extreme environment and therefore a recalibration process (whether it is possible) may be periodically needed. Last, several different types of thermocouples junctions are available: they might be grounded, ungrounded or exposed.

2.6 Summary and comparison

Before ending this section on thermal detectors, we can make a brief summary of what we have seen, comparing the various types of thermal detectors.
In the case of thermocouples, we have seen that they are simple and rugged sensors that are suitable for operating at high temperature. These are low cost devices and have a very fast response (actually, the fastest among the other devices) to a temperature change. On the other hand, the associated disadvantages are that they are the lest stable and repeatable class of devices and that they have a low sensitivity to small temperature changes, thus providing only small signals. Moreover, they always need the measurement of a reference temperature.
A possible alternative is represented by RTDs, that are the most stable over time and the most accurate devices. However, they have an high cost, the slowest response among all the devices and a low sensitivity to small temperature changes, thus providing only small signals.
The last class of devices is represented by thermistors, whose advantage is to have an high sensitivity to small temperature changes, thus being able to provide large signals. However, they can work only in a limited temperature range and they are particularly fragile.
A fourth possibility, that were not discussed during lectures, is represented by infrared optical devices.

Chapter 3

Noise

3.1 Signal and noise in time and frequency domains

In this second part of the course, we can now deal with what comes from the two elements that we have previously studied, the sensor and the amplifier: a signal. In general, in fact, we will obtain a small signal that is superimposed over a large noise, therefore our goal is the reduction of this noise in order to be able to retrieve the desired signal.
A signal is defined as a physical quantity (for example, in our case, it will generally be a voltage or a current) that varies with time and that contains a certain amount of information. It is intrinsically defined as a deterministic quantity and we can assume it to be described by its time or frequency behaviour. All random or stochastic components of the signal will be considered as noise. Given a certain signal $x(t)$ in the time domain, it may be useful to study it also in the frequency domain and, to pass from one domain to the other, we can define the Fourier transform. In particular, we define the Fourier transform of the signal in the angular frequencies domain as:

$$X(\omega) = \int_{-\infty}^{+\infty} x(t) e^{-j\omega t}\, dt$$

while in the frequency domain it will be:

$$X(f) = \int_{-\infty}^{+\infty} x(t) e^{-j2\pi f t}\, dt.$$

Once we have the signal in the frequency (or angular frequency) domain it is possible to retrieve the signal in the time domain by means of the inverse transform:

$$x(t) = \int_{-\infty}^{+\infty} X(\omega) e^{j\omega t}\, \frac{d\omega}{2\pi} = \int_{-\infty}^{+\infty} X(f) e^{j2\pi f t}\, df.$$

It is important to note that when we consider the frequency domain we have a perfect duality between the Fourier transform and the inverse transform operation, while in the angular frequency domain there is a different constant factor.

The Fourier transform of a signal leads us to a few properties that may be useful when dealing with signals. The first one is the so called initial-value theorem, that states that:

$$X(0) = \int_{-\infty}^{+\infty} x(t)\ dt, \quad x(0) = \int_{-\infty}^{+\infty} X(\omega)\ \frac{d\omega}{2\pi} = \int_{-\infty}^{+\infty} X(f)\ df.$$

Therefore, we are able to relate the DC component of a signal to its time integral over the whole domain and, analogously, the initial value of the signal to the integral of the whole Fourier transform of the signal. It is important to note that these theorems are a direct and immediate consequence of the previous definitions of Fourier transform and of inverse transform.
A second important property is represented by the time and frequency shift. Assuming the following Fourier transform to be valid:

$$\mathcal{F}[x(t)] = X(f)$$

then we can express the Fourier transform of a signal shifted in time as:

$$\begin{aligned}\mathcal{F}[x(t+\tau)] &= \int_{-\infty}^{+\infty} x(t+\tau)e^{-j2\pi ft}\ dt = \quad \text{but } z = t+\tau \\ &= \int_{-\infty}^{+\infty} x(z)e^{-j2\pi f(z-\tau)}\ dz = e^{j2\pi f\tau}X(f)\end{aligned}$$

where we have observed the last exponential term to be independent from the integration variable z, thus being extracted from the integral. Analogously, anti-transforming a signal shifted in frequency:

$$\mathcal{F}^{-1}[X(f+f_0)] = e^{-j2\pi f_0 t}x(t)$$

and we can obtain that we have also in this case multiplied the signal by an exponential factor with a different sign with respect to the previous case.
A third set of properties are the so called scaling properties, both in the time or in the frequency domain. Therefore, given again the usual signal:

$$\mathcal{F}[x(t)] = X(f)$$

it is possible to demonstrate that:

$$\mathcal{F}[x(at)] = \frac{1}{|a|}X\left(\frac{f}{a}\right)$$

and therefore we can immediately observe that a narrowing in the temporal domain is equivalent to a broadening in the frequency domain and vice versa. In the particular case of the time reversal:

$$a = -1 \ \Rightarrow\ \mathcal{F}[x(-t)] = X(-f)$$

and if we assume the signal in the time domain $x(t)$ to be real, as it will usually be, we obtain the complex conjugate of the Fourier transform:

$$x(t) \in \mathbb{R} \ \Rightarrow\ \mathcal{F}[x(-t)] = X^*(f).$$

From this property, it is possible to demonstrate that if $x(t)$ is even, also its Fourier transform will be even:

$$x(t) = x(-t) \;\Rightarrow\; X(-f) = \mathcal{F}[x(-t)] = \mathcal{F}[x(t)] = X(f)$$

and also that if $x(t)$ is real and even, also its Fourier transform is real and even. Another important property is related to the convolution of two signals. Defining therefore the following two Fourier transform of the signal $x(t)$ and $y(t)$:

$$\mathcal{F}[x(t)] = X(f), \quad \mathcal{F}[y(t)] = Y(f)$$

defining the convolution product as it follows:

$$x(t) * y(t) = \int_{-\infty}^{+\infty} x(\tau) y(t-\tau)\ d\tau$$

it is possible to demonstrate that a convolution in the time corresponds to a product in the frequency domain and vice versa:

$$\mathcal{F}[x(t) * y(t)] = X(f)Y(f), \quad \mathcal{F}[x(t)y(t)] = X(f) * Y(f).$$

According the same definition of the Fourier transform of the signals, then, it is possible to demonstrate also the so called Parseval's theorem:

$$\int_{-\infty}^{+\infty} x(t) y^*(t)\ dt = \int_{-\infty}^{+\infty} X(f) Y^*(f)\ df$$

and in the special case in which:

$$x(t) = y(t)$$

it gives:

$$\int_{-\infty}^{+\infty} |x(t)|^2\ dt = \int_{-\infty}^{+\infty} x^2(t)\ dt = \int_{-\infty}^{+\infty} |X(f)|^2\ df.$$

Another important property states that the Fourier transform of a gaussian signal is again a gaussian signal in the frequency domain:

$$\mathcal{F}\left[e^{-\pi\sigma_t^2 t^2}\right] = e^{-\pi\sigma_f^2 f^2} = e^{-\pi\frac{f^2}{\sigma_t^2}}$$

where the widths of the functions are related by the following uncertainty relationship:

$$\sigma_t \cdot \sigma_f = 1.$$

This means that a broad signal in time will give a narrow signal in the frequency domain and vice versa, with a certain relationship between the two widths. Since it is possible to demonstrate that a similar property will hold for any other signal and the associated Fourier transform and that, however, gaussian pulses are the one for which the product of the two widths is smaller, we are now able to estimate the minimum value of the bandwidth that is required for representing a certain signal in the frequency domain.

To generalize the previous property, we can consider a generic signal represented in Figure 3.1 and the associated Fourier transform in the frequency

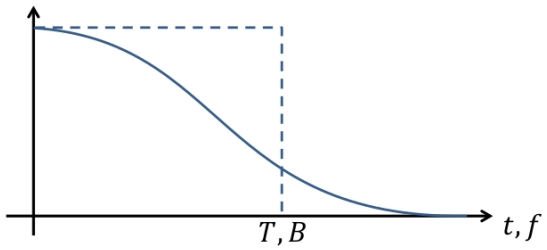


Figure 3.1: A signal in the time domain or in the frequency domain and the associated rectangular version.

domain, that we can assume to be represented in the same Figure. For the sake of simplicity, it is possible to define two square signals, one in the time domain and the other in the frequency domain, with an amplitude equal to the initial amplitude of the signal and of its Fourier transform and with a duration, respectively T and B, such that the area under the curve is preserved:

$$\int_{-\infty}^{+\infty} x(t)\ dt = x(0) \cdot T, \quad \int_{-\infty}^{+\infty} X(f)\ df = X(0) \cdot B.$$

In this case, from the initial value theorem, we can write:

$$X(0) = \int_{-\infty}^{+\infty} x(t)\ dt = x(0)T$$

where the last equality holds from our definition of T, and analogously:

$$x(0) = \int_{-\infty}^{+\infty} X(f)\ df = X(0)B_f = x(0)TB_f$$

where in the last equivalence we have substituted the previous equation. From this last expression, simplifying common terms, it is possible to obtain that:

$$TB_f = 1$$

or, in the angular frequency domain:

$$TB_\omega = 2\pi.$$

In general, this will hold only as a first order approximation but, as we have said before, it can be useful for estimating the bandwidth requirements.

3.2 Cross-correlation and autocorrelation

A particular class of signals is represented by the signals $x(t)$ such that:

$$x(t) \in \mathrm{L}^2(\mathbb{R})$$

where L^2 is the Lebesgue space of the functions whose integral of the square modulus, according to the Lebesgue definition of integral, converges. These signals are called energy signals. Given two of such signals $x(t)$ and $y(t)$, it is

possible to define the cross-correlation between these two signals as:

$$k_{xy}(\tau) = \int_{-\infty}^{+\infty} x(t)y(t+\tau)\ dt.$$

This is a time-dependent quantity and it measures the "similarity" between the two signals considered as a function of their reciprocal time difference τ. This function is useful since in many cases we approximately know the shape of the signal coming from a certain sensor, while the associated unknown quantities may be its amplitude, its arrival time and the noise superimposed to it. Therefore, to evaluate how much this signal is similar to the reference, expected signal, we can evaluate this function depending on the reciprocal time-delay. From the definition of the cross-correlation, we can immediately demonstrate the following property:

$$\begin{aligned} k_{xy}(\tau) &= \int_{-\infty}^{+\infty} x(t)y(t+\tau)\ dt = \quad \text{but } z = t+\tau \\ &= \int_{-\infty}^{+\infty} x(z-\tau)y(z)\ dz = k_{yx}(-\tau). \end{aligned}$$

Moreover, the following two inequalities will hold:

$$|k_{xy}(\tau)| \leq \sqrt{k_{xx}(0)k_{yy}(0)}, \quad |k_{xy}(\tau)| \leq \frac{1}{2}\left[k_{xx}(0) + k_{yy}(0)\right].$$

In them, we can observe that we are calculating the cross-correlation of a signal with itself: this quantity is defined autocorrelation. The autocorrelation of a signal, therefore, will measure the similarity of a signal with a shifted replica of itself, thus measuring the "predictability" of a signal over time. It is a real and even function that can be defined as:

$$k_{xx}(\tau) = \int_{-\infty}^{+\infty} x(\tau)x(t+\tau)\ dt$$

and it is possible to show that:

$$|k_{xx}(\tau)| \leq k_{xx}(0) = \int_{-\infty}^{+\infty} x^2(t)\ dt = E$$

where E is the so called energy of the signal. This inequality seems to be quite obvious: when the two signals are not delayed one with respect to the other ($\tau = 0$) they will be actually identical and, therefore, their cross-correlation will be maximum. Moreover, it is important to note that the signal energy is only proportional (with a proportionality constant that in general depends on the signal considered) to the physical energy of the signal, but it is not strictly speaking an energy (since its dimension is not Joules).

Comparing, as in Figure 3.2, the autocorrelation[1] of a signal:

$$k_{xx}(t) = \int_{-\infty}^{+\infty} x(\tau)x(\tau+t)\ d\tau$$

[1]Note that in the previous definition the roles of τ and t where exchanged, but this does not lead to any difference.

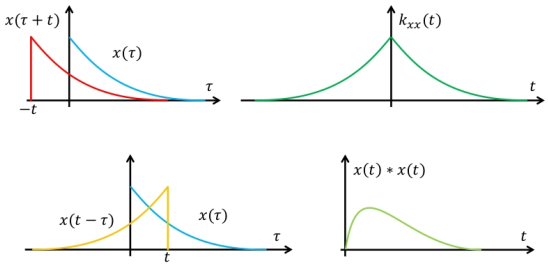


Figure 3.2: Autocorrelation (above) and convolution (below) of a signal with itself.

with the convolution of a signal with itself:

$$x(t) * x(t) = \int_{-\infty}^{+\infty} x(\tau)x(t-\tau)\ d\tau$$

it is possible to observe that in one case we are comparing the signal with a delayed version of itself for different delays, while in the second case we are comparing the signal with a delayed an time-reversed version of itself: we are thus performing two different operations, obtaining two different results. Therefore, considering the expression for the autocorrelation:

$$k_{xx} = k_{xx}(-t) = \int_{-\infty}^{+\infty} x(\tau)x(\tau-t)\ d\tau = x(t) * x(-t)$$

and assuming to have a real and even signal, Fourier transforming the autocorrelation we will obtain:

$$\mathcal{F}[k_{xx}(t)] = X(f)X^*(f) = |X(f)|^2.$$

From this expression of the Fourier transform of the autocorrelation, if we then apply the initial value theorem[2], we can write the energy of the signal as:

$$E = k_{xx}(0) = \int_{-\infty}^{+\infty} |X(f)|^2\ df$$

and from this expression we can define $|X(f)|^2$ as the energy spectral density of the signal. In fact, integrating it over the whole range of the frequencies we will obtain the energy of the signal and, moreover, at a given frequency it will give how much energy is carried by that frequency component of the overall signal.
A different class of signals is represented by power signals. These signals do not belong to space L^2:

$$x(t) \notin \mathrm{L}^2(\mathbb{R})$$

since their energy diverges. A classical example, in this case, are sinusoidal signals, but also any other periodic signal. A possibility, in this case, is to restrict our analysis only over a period (or a certain portion) of the signal, defining a truncated energy signal as:

$$x_T(t) = \begin{cases} x(t), & \forall |t| \leq T \\ 0, & \forall |t| > T \end{cases}.$$

[2]Alternatively, this can be obtained by applying the Parseval's theorem.

For these truncated signals, it is possible to define the associated Fourier transform $X_T(f)$ and the following autocorrelation function:

$$k_{xx}^T(\tau) = \int_{-\infty}^{+\infty} x_T(t)x_T(t+\tau)\ dt$$

where, again, we can define the energy power density of the truncated signal:

$$\mathcal{F}[k_{xx}^T(\tau)] = |X_T(f)|^2.$$

Then, to retrieve the fact that the original signal is not truncated, in calculating the autocorrelation we need to calculate the following limit:

$$K_{xx}(\tau) = \lim_{T\to+\infty} \frac{1}{2T} k_{xx}^T(\tau) = \lim_{T\to+\infty} \frac{1}{2T} \int_{-T}^{+T} x(t)x(t+\tau)\ dt.$$

Calculating the Fourier transform of this autocorrelation, then, it is possible to define the power spectral density of the signal $S(f)$:

$$\mathcal{F}[K_{xx}(\tau)] = \lim_{T\to+\infty} \frac{1}{2T} |X_T(f)|^2 = S(f)$$

and its meaning can be understood by applying the definition of autocorrelation and the Parseval's theorem:

$$P = K_{xx}(0) = \lim_{T\to+\infty} \frac{1}{2T} \int_{-T}^{+T} x^2(t)\ dt = \int_{-\infty}^{+\infty} S(f)\ df.$$

In fact, the quantity in the first integral is the energy of the signal (that we know to be infinite) that is divided by the time interval considered (that is tending to infinite in the limit), thus giving the power P associated to the signal. This power, then, will be equal to the integral over the whole frequency range of a quantity that can be recognized to be the power spectral density. Again, also in this case this power is not a physical power, since it will only be proportional to physically defined quantities. The associated power spectral density, analogously to the energy spectral density, will give the contribution of each harmonic component to the power of the signal.
Summing up, in energy signals the energy is a finite quantity and thus the power associated is identically equal to zero, while in power signals the power of the signal is a finite quantity and therefore the energy associated to the signal tends to infinity.

3.3 Random processes

Stochastic or random processes represent the time dependence of a random variable and they are generally superimposed to a deterministic signal. An immediate example of a random process is represented by the current flowing through a certain device, that will continuously undergo to random fluctuations. Given therefore x a certain random variable and t the time variable, we can define a random process as the probability density function $p(x,t)$ that will depend both on the random variable considered and the time or, alternatively, by the joint probabilities:

$$p(x_1,\dots,x_n;t_1,\dots,t_n).$$

At any given time instant, therefore, the process can be studied as a random variable, while for any given random variable the process will be a deterministic function of time, that will be called the realization. The stochastic nature of such a process is evident only when we are considering different replicas of the same system to describe a certain phenomenon: the measurement of that phenomenon will undergo to various, unpredictable fluctuations.
A specific class of processes are the so called stationary processes, that are independent from the time shift considered. In this case, indicating with x_1, x_2, ..., x_n the samples that we are investigating, we can observe that the joint probability distributions for these samples are independent from the time shift:

$$p(x_1, \ldots, x_n; t_1, \ldots, t_n) = p(x_1, \ldots, x_n; t_1 + T, \ldots, t_n + T) \quad \forall n,\ t,\ T.$$

A certain time window, therefore, can be arbitrarily shifted in time without changing the associated probability for a given event. As a consequence, we can say that since the joint probability density functions are the same for any time shift, this means that these probabilities are independent from time:

$$p(x, t) = p(x, t + T)\ \forall t, T \quad \Rightarrow \quad p(x, t) = p(x).$$

Considering therefore two different samples and expressing the time, in one, as the time of the other sample shifted of a certain quantity τ:

$$p(x_1, x_2; t_1, t_2) = p(x_1, x_2; t_1, t_1 + \tau), \quad t_2 = t_1 + \tau\ \forall t_1$$

we can observe that the joint probability density function for these two samples will depend, exclusively on the reciprocal time difference:

$$p(x_1, x_2; t_1, t_2) = p(x_1, x_2; \tau)$$

that has been defined as:

$$\tau = t_2 - t_1.$$

As we have done in many other courses, we can then define a mean value for a certain random variable:

$$\bar{x} = \int x p(x)\ dx$$

and this is completely analogous to the definition of average value in stationary processes, and a variance:

$$\sigma_x^2 = \int (x - \bar{x})^2 p(x)\ dx = \overline{x^2} - \bar{x}^2.$$

It is important to note that, in principle, the probabilities $p(x)$, involved in both the definitions, are time-dependent quantities. However, in the case of stationary processes, they will be independent from time, thus making also the mean value and the variance independent from time. In the vast majority of the cases, we will assume that the mean value of a certain random variable is identically equal to zero:

$$\bar{x} = 0.$$

From signal theory, then, we can define a few tools that will be needed. In particular, we define the autocorrelation as:

$$R_{xx}(\tau) = \overline{x_1 x_2} = \iint x_1 x_2 p(x_1, x_2; \tau)\ dx_1\ dx_2$$

where, since by definition we have that:

$$\tau = t_2 - t_1$$

then the autocorrelation will be implicitly dependent from these times t_1 and t_2. Only in the case of stationary processes, this dependence will be only on the difference τ between them, as we have written in the previous definition.
We define then the covariance as:

$$C_{xx}(\tau) = \overline{(x_1 - \bar{x}_1)(x_2 - \bar{x}_2)} = \iint (x_1 - \bar{x}_1)(x_2 - \bar{x}_2) p(x_1, x_2; \tau)\ dx_1\ dx_2$$

and also in this case the previous observation holds: this quantity will depend only on the time difference τ if and only if we are dealing with stationary processes.
In the general case, after a few calculations it is possible to show that:

$$C_{xx} = R_{xx} - \bar{x}_1 \bar{x}_2.$$

In the same way, we are able to calculate the values when the two signals have not a reciprocal delay:

$$\tau = 0: \quad R_{xx}(0) = \int x_1^2 p(x_1)\ dx_1 = \overline{x^2}$$
$$C_{xx}(0) = \int (x_1 - \bar{x}_1)^2 p(x_1)\ dx_1 = \sigma_x^2.$$

Since we will generally consider processes with zero mean value, this means that:

$$\bar{x} = 0 \ \Rightarrow \ R_{xx} = C_{xx}.$$

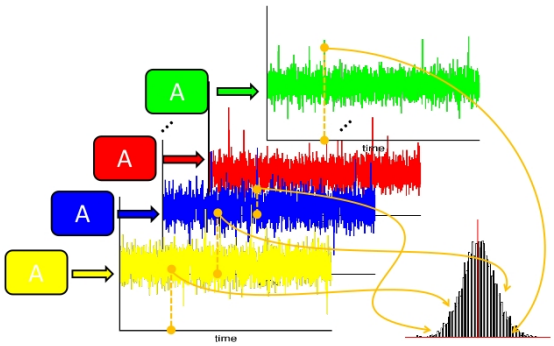


Figure 3.3: Possible outcomes of four different amplifiers sampled at the same time instant.

Even when the average value of a certain random variable is different from zero, a measurement of this variable will show a certain disturbance that will make the outcome of these measurements to be similar but not exactly equal to the outcomes from any different measurement. We can then calculate an average of these measurements and this can be done in two different ways: averaging over an ensemble or averaging over time.

The meaning of an ensemble average is shown in Figure 3.3. As an example, we can consider the noise that is coming from four different amplifiers whose input is grounded. Sampling the obtained waveforms for this noisy signal at a certain, fixed time instant t_0, the probability density function for a certain outcome of the measurement will be the usual probability density function evaluated at that time instant $p(x, t_0)$. Therefore, for any different amplifier the random variables $x_1, x_2, \ldots, x_n$ will have a different value and this can be plotted in the histogram represented in Figure, thus determining the probability density function at a fixed time instant.

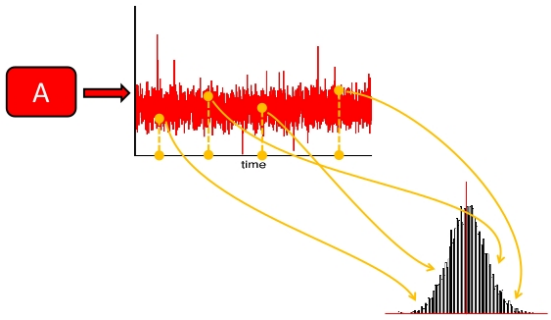


Figure 3.4: Possible outcomes of an amplifier sampled at different time instants.

Alternatively, as it is represented in Figure 3.4, we can focus on a single amplifier x_0 and observe it at certain, different time instants: this is what is called a temporal average. In this case, therefore, we are obtaining the probability density function for a certain random variable $p(x_0, t)$. Collecting this sampled value as a function of time in an histogram, we have thus obtained another (and in principle different) probability distribution function.
It is important to stress the difference between these two averages: in the first case, we are sampling many, different amplifiers at a fixed time instant, while in the second case we have fixed the amplifier that we are considering and we have sampled it at different time instants. Are the two probability distribution functions the same? In general, no: the ensemble average is usually different from the temporal average, apart when we are dealing with ergodic processes.
Ergodicity, therefore, is a property of a particular class of systems (that are called ergodic) in which the ensemble average $\bar{x}$ is equal to the temporal average $\langle x \rangle$. From a mathematical point of view, the equivalence between these two averages can be written as:

$$\langle x \rangle = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} x(t)\ dt = \int xp(x)\ dx = \bar{x}$$

and in analogous way for these processes the ensemble autocorrelation:

$$R_{xx}(\tau) = \iint x_1 x_2 p(x_1, x_2; \tau)\ dx_1\ dx_2$$

is equal to the temporal autocorrelation:

$$K_{xx}(\tau) = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} x(t)x(t+\tau)\ dt$$

thus giving:

$$R_{xx}(\tau) = K_{xx}(\tau).$$

This means that, for all the moments, temporal statistics will converge to the ensemble ones. Moreover, it is possible to demonstrate that all ergodic processes are also stationary processes (thus giving sense to all the temporal averages that we have written), while the opposite is not true.

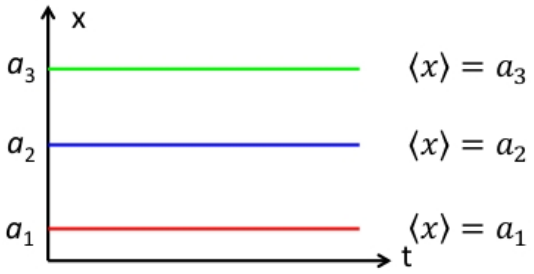


Figure 3.5: A counterexample of a system that is stationary but it is not ergodic.

A simple counterexample of a system that is stationary but it is not ergodic can be found in Figure 3.5. Consider the process:

$$x(t) = A, \quad A \in [0, 1]$$

where A is a random variable in the interval defined. We can immediately observe that, from this definition, the ensemble average of x will be coincident with the average of the random variable A:

$$\bar{x} = \bar{A}.$$

Considering now a fixed realization of this system, for example a_1 as it is represented in Figure, since it is independent from time it will be equal to the temporal average for that realization. However, considering a different realization, this temporal average will be different and, in general, any temporal average will be different from the ensemble average:

$$\langle x \rangle = a_1 \neq \langle x \rangle = a_2 \neq \langle x \rangle = a_3 \neq \cdots \neq \bar{A}.$$

We have thus proven the existence of at least one stationary system $x(t)$ that is not ergodic.
A different, but useful perspective also when dealing with stochastic processes is the frequency domain. In particular, since a single realization of the process is, actually, a signal (that is defined just as a time dependent variable), its Fourier transform will be well defined. Considering therefore $x_i(t)$ as a single realization, thus being a power signal, we can deal with its truncated version and then perform a limit, thus obtaining the following power spectral density $S_i(f)$ for the single realization:

$$S_i(f) = \lim_{T \to \infty} \frac{1}{2T} |X_{T,i}(f)|^2.$$

We can then observe that this power spectral density $S_i(f)$ is a random variable that will depend on the realization considered, therefore calculating its ensemble average:

$$S(f) = \overline{S_i(f)} = \lim_{T\to\infty} \frac{1}{2T}\overline{|X_{T,i}(f)|^2}$$

we define the power spectral density $S(f)$ of the random process. From our previous introduction to signals, we have defined the power spectral density of a signal as the Fourier transform of the autocorrelation. Is this true also for random processes? The answer is yes and it is stated in the so called Wiener-Kintchine theorem. To prove it, we can try to calculate the inverse Fourier transform of the power spectral density:

$$\begin{aligned}\mathcal{F}^{-1}[S(f)] &= \int \lim_{T\to\infty} \frac{1}{2T}\overline{|X_T(f)|^2} e^{j2\pi f\tau}\ df = \\ &= \lim_{T\to\infty} \frac{1}{2T}\int \overline{X_T(f)X_T^*(f)} e^{j2\pi f\tau}\ df.\end{aligned}$$

However, from the definition of Fourier transform:

$$X_T(f) = \int_{-T}^{T} x(t_1)e^{-j2\pi f t_1}\ dt_1, \quad X_T^*(f) = \int_{-T}^{T} x(t_2)e^{j2\pi f t_2}\ dt_2$$

where, in general:

$$t_1 \neq t_2$$

we can write:

$$\mathcal{F}^{-1}[S(f)] = \lim_{T\to\infty} \frac{1}{2T}\int \overline{\int_{-T}^{T} x(t_1)e^{-j2\pi f t_1}\ dt_1 \int_{-T}^{T} x(t_2)e^{j2\pi f t_2}\ dt_2}\ e^{j2\pi f\tau}\ df.$$

Switching time integrals with ensemble averages and observing that the only random terms over which we are performing the ensemble average are $x(t_1)$ and $x(t_2)$, while all the exponential are deterministic quantities for which we can apply the usual properties of the exponentials, we can write:

$$\mathcal{F}^{-1}[S(f)] = \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T}\int_{-T}^{T} \overline{x(t_1)x(t_2)} \int e^{j2\pi f(t_2-t_1+\tau)}\ df\ dt_1\ dt_2.$$

From the temporal shift property of the Fourier transform, observing that in this case in the last integral in the frequency domain we are calculating the inverse Fourier transform of an exponential, obtaining a time shift Dirac's delta, we obtain:

$$\mathcal{F}^{-1}[S(f)] = \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T}\int_{-T}^{T} R_{xx}(t_1,t_2)\delta(\tau+t_2-t_1)\ dt_1\ dt_2$$

where we have recognized the definition of autocorrelation, and since the Dirac's delta is just sampling the previous function in a point where the argument of the delta is zero:

$$\mathcal{F}^{-1}[S(f)] = \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} R_{xx}(\tau)\ dt_1 = R_{xx}(\tau)$$

since the autocorrelation function that we are integrating is independent from the variable of integration t_1, thus making the limit equal to the function itself. We have therefore just proven that:

$$S(f) = \mathcal{F}[R_{xx}(\tau)]$$

therefore, also in random processes, the power spectral density can be obtained as the Fourier transform of the autocorrelation. It is important to note that the crucial step, in this demonstration, is the calculation of the following ensemble average:

$$\overline{X_T(f)X_T^*(f)}$$

using the definition of the Fourier transform for power signals and observing which terms are actually random variables and which ones are deterministic quantities.

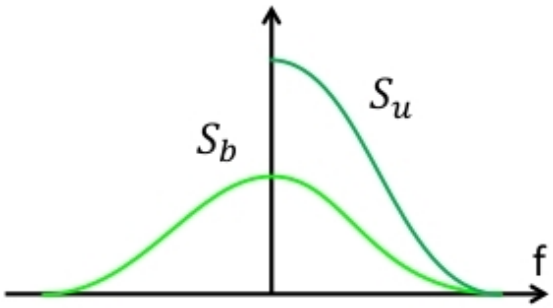


Figure 3.6: A bilateral and an unilateral power spectrum.

In general, the power spectral density $S(f)$ is real and even and it extends from $-\infty$ to $+\infty$, therefore also the autocorrelation $R_{xx}(\tau)$ is real and even and we can define a bilateral power spectrum $S_b(f)$. However, in circuit calculations, we are usually interested in positive frequencies, therefore a unilateral power spectral density $S_u(f)$ is defined, extending from 0 to $+\infty$ in the frequency domain. Between the two power spectral densities, the following relationship holds:

$$S_u(f) = \begin{cases} 2S_b(f) & \forall f \geq 0 \\ 0 & \forall f < 0 \end{cases}$$

as a consequence of the fact that the bilateral power spectral density is even. It is fundamental to note that the previous demonstration of the Wiener-Kintchine theorem holds only or the bilateral power spectral density:

$$\mathcal{F}[R_{xx}(\tau)] = S_b(f) \neq S_u(f).$$

3.4 White noise

A particular kind of noise is the so called white noise. From a statistical point of view, it is defined as a noise signal whose autocorrelation is equal to a Dirac's delta:

$$R_{nn}(\tau) = \lambda \cdot \delta(\tau).$$

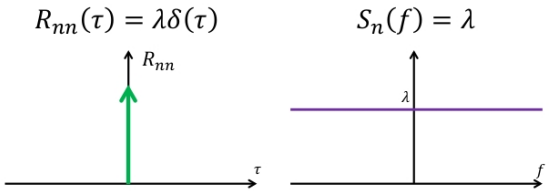


Figure 3.7: Autocorrelation and power spectral density for a white noise.

This means that the autocorrelation of this signal is identically equal to zero for any delay τ different from zero:

$$R_{nn}(\tau) = 0 \quad \forall \tau \neq 0$$

and therefore this process is completely uncorrelated with itself, thus not allowing any prediction on it. From the Wiener-Kintchine theorem we can write the power spectral density of this noise as:

$$S_n(f) = \mathcal{F}[R_{nn}(\tau)] = \lambda$$

and therefore the power spectral density is constant and different from zero all over the spectrum of the frequencies. This is a direct consequence of what we have said before: if the process is completely uncorrelated with itself, it will be a completely unpredictable signal and this means that the power spectral density of this signal will have an equal contribution from all the frequencies in the process. Assuming the average of this signal to be equal to zero:

$$\bar{n} = 0$$

we can calculate its variance as:

$$\sigma_n^2 = \overline{n^2} - \bar{n}^2 = \overline{n^2} = R_{nn}(0) = \int_{-\infty}^{+\infty} S_n(f)\ df = \infty.$$

This means that it is impossible, for a physical process, to be described through a white noise: in fact, any physical process will always contain a finite amount of energy.
We can thus try to approximate real noise terms as white noise on a limited spectral (or temporal) range. It is possible to state the following uncertainty relation between the time and the bandwidth of this real noise term: in fact, if the autocorrelation is equal to zero only for big enough delays:

$$R_{nn}(\tau) = \lambda \cdot g(\tau),\ g(\tau) \simeq 0 \quad \forall |\tau| > \tau_0$$

then the power spectral density will be constant only in a limited range of frequencies:

$$S_n(f) \simeq \text{const.} \quad \forall |f| < \frac{1}{\tau_0}.$$

A first way of approximating a real noise term with a white noise is the so called triangular approximation. In this approximation, the autocorrelation of the noise is represented as a triangle extending from $-\tau_0$ to $+\tau_0$ and, making this

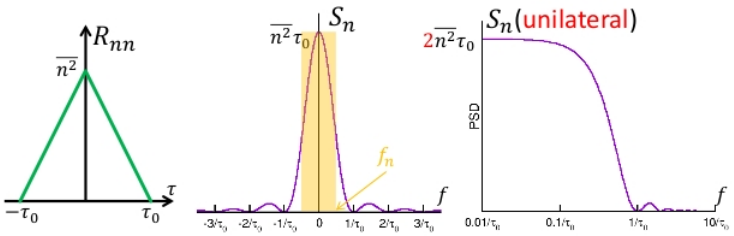


Figure 3.8: Triangular approximation.

width tend to zero, we can immediately retrieve the Dirac's delta that is typical of a white noise. Calculating the Fourier transform of this autocorrelation term, we obtain the spectral noise density for this signal:

$$\mathcal{F}[R_{nn}(\tau)] = S_n(f) = \overline{n^2} \cdot \left(\frac{\sin(\pi f \tau_0)}{\pi f \tau_0} \right)^2$$

that is represented, both in its bilateral and unilateral version, in Figure 3.8. From the calculation of the average value of the square of the noise term in the so called equivalent rectangle approximation:

$$\overline{n^2} = \int S_n(f)\ df = \overline{n^2}\tau_0 \cdot \int \left(\frac{\sin(\pi f \tau_0)}{\pi f \tau_0} \right)^2\ df = S_n(0) \cdot 2 f_n$$

we can calculate the width of this equivalent rectangle as:

$$f_n = \frac{1}{2\tau_0}.$$

This result is not surprising at all, since the equivalent rectangle to the triangle representing the autocorrelation of the signal, if we want to preserve the area, will be extended from $-\tau_0/2$ to $\tau_0/2$.

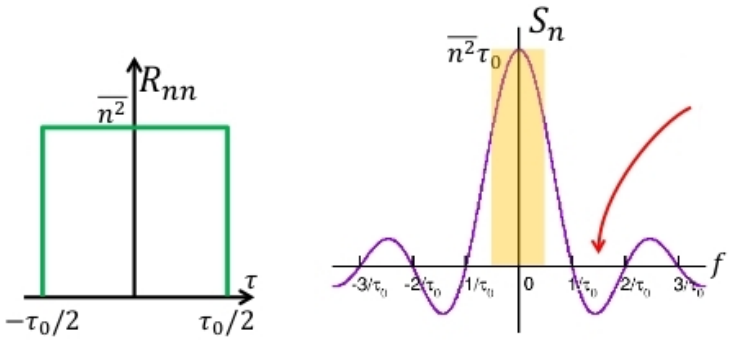


Figure 3.9: Rectangular approximation.

A different approximation is the rectangular one, that is represented in Figure 3.9. In this case, the autocorrelation of the signal is approximated with a rectangle, thus obtaining, from the Fourier transform of this autocorrelation,

the following spectral power density:

$$S_n(f) = \overline{n^2}\tau_0 \frac{\sin(\pi f \tau_0)}{\pi f \tau_0}$$

and also in this case we can use an equivalent rectangle approximation also for this power spectral density:

$$\overline{n^2} = \int S_n(f)\ df = S_n(0) \cdot 2f_n$$

thus obtaining the following width for the equivalent rectangle:

$$f_n = \frac{1}{2\tau_0}.$$

If we want, in both cases, to approximate this behaviour with a white noise, we can assume that the typical temporal behaviour of the system involves times much larger than τ_0 or, equivalently, in the frequency domain we must be dealing with frequencies much lower than $1/\tau_0$. It is then possible to remember that the power spectral density was also defined as:

$$S_n(f) = \lim_{T\to\infty} \frac{1}{2T} \overline{|X_T(f)|^2}$$

where we are averaging over a square modulus, that should be positive. However, considering for example the part indicated, in Figure 3.9, with a red arrow, we can clearly observe that there are portions of the spectral power density that are negative. What does it mean? This question is left to the student.

3.5 Thermal noise in resistors

We can now focus more specifically on the types of noise that we have to face in electronics. The first kind of noise we will be dealing with is the thermal noise, that was originally observed in the '20s in vacuum tubes technology.

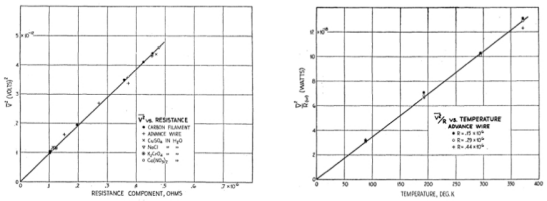


Figure 3.10: Root mean square value of the voltage fluctuations depending on the resistance and on the temperature of the conductor.

In Figure 3.10 are represented two dependencies originally obtained in the first observation of the thermal noise. In particular, in a conductor we can have certain voltage fluctuations whose root mean square value can be observed to be linearly proportional to the resistance and to the temperature of the conductor considered. Moreover, Johnson discovered that this noise term is almost white, thus being constant at any frequency he could measure. Therefore, considering for example a resistor without any voltage applied to its ends, this noise will have a zero average and a root mean square value of these fluctuations with the above dependencies.

3.5.1 Nyquist derivation

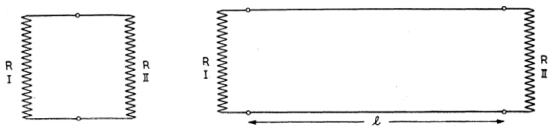


Figure 3.11: Transmission line considered in Nyquist derivation.

The first explanation of thermal noise is the so called Nyquist derivation. Consider, as represented in Figure 3.11, a circuit with two resistors, R_I and R_{II}. If the two resistors are connected, the voltage fluctuations at thermal equilibrium in one resistor will be completely identical to the voltage fluctuations in the other one. At thermal equilibrium, therefore, the same power will be transmitted from the first resistor to the second one and vice versa, giving a flow of current in the two conductors. We can now suppose to add a lossless, adaptive and ideal transmission line of length l and with a characteristic impedance R between these two resistors. In this case, a travelling wave, related to these fluctuations, will travel along the line and, if at a certain point the line is short-circuited over itself, we are somehow "trapping" the energy into the line and we must have, in it, some standing waves. These standing waves will be also called modes of the circuit and, indicating with n the maximum order of these modes[3], λ their wavelength and l the length of this transmission line, the following relationship will hold:

$$n\lambda = l, \quad n \in \mathbb{N}$$

where the following relationship between frequency f and wavelength λ, given v the speed of the signal across the line, holds:

$$\lambda f = v.$$

Analogously to what we do in Bloch's theory in solid state physics, we can now calculate the number of these modes from the previous relations:

$$n = \frac{l}{\lambda} = \frac{l}{v} f$$

and differentiating this relationship we can obtain the number of modes contained in an infinitesimal interval of frequencies:

$$dn = \frac{l}{v}\, df.$$

From the equipartition theorem, since any mode will have two degrees of freedom (an electric degree of freedom and a magnetic one), the energy $\mathcal{E}(f)$ associated to a certain mode will be equal to $k_B T$ and, therefore, we can write the energy contained in an infinitesimal frequency interval as:

$$\mathcal{E}(f)\, df = k_B T\, dn = k_B T \frac{l}{v}\, df$$

[3]Thus being equal to the number of modes that are allowed in the line.

thus being the energy transferred at a given frequency in the time interval $\tau = l/v$ equal to a transit time of the line. Since the power density is defined as the energy transferred per unit time at a given frequency f, it can be written as:

$$P(f)\ df = \frac{\mathcal{E}(f)\ df}{\frac{l}{v}} = k_B T\ df$$

and therefore we have determined that the power density, from the equipartition theorem, will be equal to $k_B T$. Now, we need to relate this physically meaningful quantity to some electrical parameters. A generic voltage fluctuation V will generate a current:

$$I = \frac{V}{2R}\ \rightarrow\ \overline{I^2} = \overline{\left(\frac{V}{2R}\right)^2} = \frac{\overline{V^2}}{4R^2}$$

in the given circuit, thus dissipating a certain power over each resistor that, on average, is equal to:

$$\bar{P} = R\overline{I^2} = R \cdot \left(\frac{\overline{V^2}}{4R}\right) = \int_0^{+\infty} \frac{S_V}{4R}\ df$$

where we have assumed that the negative frequencies are not meaningful, thus using an unilateral power spectral density. However, this last equation gives:

$$\bar{P} = \int_0^{+\infty} \frac{S_V}{4R}\ df = \int P(f)\ df = \int k_B T\ df$$

where $P(f)$ is the power spectral density that we have previously defined. This gives therefore:

$$\frac{S_V}{4R} = k_B T\ \rightarrow\ S_V = 4k_B T R$$

the power spectral density S_V in terms of voltage of the noise that we have defined, since:

$$\overline{V^2} = \int_0^{+\infty} S_V\ df.$$

As an example, we can consider the following resistor:

$$R = 1\ \text{k}\Omega$$

at room temperature, obtaining:

$$S_V \simeq 1.66 \times 10^{-20}\ \text{V}^2/\text{Hz} = \left(4.07\ \text{nV}/\sqrt{\text{Hz}}\right)^2$$

where we have approximated this noise term as if it were a white noise, even though this is not true. To understand whether this white noise approximation is sufficiently good, we need to study the high-frequency behaviour of this power spectral density. The starting point, in this case, is the Planck equation, that describes the energy associated to an oscillator at frequency f at a certain temperature T:

$$\mathcal{E} = \frac{hf}{e^{\frac{hf}{k_B T}} - 1}$$

and assuming:

$$\frac{hf}{k_BT} \ll 1 \ \rightarrow \ f \ll \frac{k_BT}{h}$$

it gives the following approximation:

$$\mathcal{E} \sim \frac{hf}{1+\frac{hf}{k_BT}-1} \sim k_BT$$

that is exactly the one that we adopted for evaluating, through the equipartition theorem, the energy of an oscillator in the previous case. The previous calculation, therefore, shows us that at room temperature the power spectral density can be assumed to be white up to a frequency that is approximately equal to 100 GHz. In all the applications that we will consider, we will be far below this frequency, therefore we can approximate this noise contribution to be white. From Quantum Mechanics, we know that to a quantum harmonic oscillator should be associated also a zero-point energy; however, this contribution can be generally neglected and therefore it is not included.

3.5.2 Brownian motion

A more intuitive derivation of the power spectral density of this noise term can be obtained studying the random motion of the electrons in a conductor due to the thermal energy. In fact, we can immediately observe that on average the sum of the effects of the random motions of the electrons in a conductor will be equal to zero, but in a certain time interval it is possible to obtain a value, for example of a voltage related to these motions, that is slightly different from zero, since we are dealing with a stochastic process. The starting point in this case is the diffusion equation, also called Brownian motion equation:

$$\frac{\partial n}{\partial t} = D\frac{\partial^2 n}{\partial x^2}$$

where D is the so called diffusion coefficient and n is the electron density. From this equation, especially when dealing with large-scale diffusion processes (since for electrons the limit we will consider is not really meaningful), the concentration term at the initial time $t = 0$ can be associated to a Dirac's delta, completely concentrated in one point of the space, that will then spread all over the domain while time goes by. In one of the important papers published by Einstein in 1905, it is possible to demonstrate that the following relation holds:

$$\overline{x^2(t)} = 2Dt$$

and, moreover, that the spectral density of the difference in the speed of the electrons $S_{\Delta v_x}$ can be written as:

$$S_{\Delta v_x} = 4D.$$

The noise, therefore, will be a consequence of the random motion of the electrons and, therefore, of the fluctuations of their velocity due to the thermal energy.
To understand how a single electron and the fluctuations in its velocity can contribute to the presence of a current and thus of a voltage, we can consider a single electron moving in a plane capacitor whose electrodes are both grounded.

Assuming the electron to be moving from left to right with an axial velocity v_x, assuming 0 to be the coordinate of the left hand-side armature, x to be the coordinate of the electron and L to be the coordinate of the right hand-side armature, since the electron has a charge $-q$ it will induce two spatial charge densities on the armatures. In particular, the charge induced on the left hand-side armature will be:

$$q\frac{L-x}{L}$$

while the charge induced on the right hand-side armature will be:

$$q\frac{x}{L}.$$

When the electron is moving through this conductor, it will change its position x and thus also the charges induced on the electrodes will change, resulting in a current i flowing from one electrode to the other in the opposite direction with respect to the motion of the electron:

$$i = \frac{q}{L}v_x.$$

Differentiating this relationship, we obtain:

$$\Delta i = \frac{q}{L}\Delta v_x$$

but since the speed of the electron is a random variable, we can write the autocorrelation of the random variation of the current as a function of the random variation of the velocity:

$$R_{\Delta i} = \overline{\Delta i(t)\Delta i(t+\tau)} = \frac{q^2}{L^2}\overline{\Delta v_x(t)\Delta v_x(t+\tau)} = \frac{q^2}{L^2}R_{\Delta v}.$$

Calculating the Fourier transform of both sides of this equation, we can write a new equation involving the power spectral densities and, since we know that power spectral density associated to a variation of the speed of an electron:

$$S_{\Delta i} = \frac{q^2}{L^2}S_{\Delta v_x} = \frac{q^2}{L^2}4D$$

we have obtained a relation linking the power spectral density associated to variations of the current to the dispersion coefficient. Assuming now to have:

$$N = nSL$$

independent electrons, where n is the electronic density of the conductor, S the surface of the previous electrodes (that will be the faces of a conductor) and L the length along the direction of motion considered of the conductor (or alternatively the distance between the electrodes), we can write the power spectral density associated to the current of every electron as:

$$S_I = \sum_{m=0}^{N} S_{\Delta i} = N \cdot S_{\Delta i} = \frac{q^2}{L^2}4DnSL = \frac{4nSq^2}{L}\frac{k_B T}{q}\mu$$

where we have summed all the single-electron contributions since we assumed them to be independent and we have used the following definition that comes from Einstein's relation:

$$\frac{D}{\mu} = \frac{k_B T}{q}$$

where μ is the mobility of the electrons. However, remembering the definition of the conductivity of a metal:

$$\sigma = qn\mu$$

we obtain:

$$S_I = 4k_B T \frac{\sigma S}{L}$$

and since, from the second Ohm's law:

$$R = \rho \frac{L}{S} = \frac{1}{\sigma} \frac{L}{S}$$

we obtain:

$$S_I = \frac{4k_B T}{R}.$$

This gives the following spectral density in terms of voltage:

$$S_V = 4k_B T R$$

consistently with what we have obtained in the previous derivation.

3.6 Shot noise and Poisson random process model

As we have seen in the previous section, the current in an electronic device can be written, from a physical point of view, as:

$$I = qnvA$$

where q is the charge of the electron, n is the density of carriers, v its their velocity and A is the area of the conductor observed. In this expression, only two quantities can show significant fluctuations: the carrier density n and the speed v. Fluctuations in the velocity of the electrons lead, as we have seen in the previous section, to the so called thermal noise. Fluctuations in the carrier density n, on the other hand, lead to a new source of noise that we can study: the shot noise.
This kind of noise was first observed by W. Schottky in 1918, while he was studying the current fluctuations in vacuum tubes. In particular, as we have just seen, it is related to the fluctuations in the number of charge carriers, rather than in their velocity.

As represented in Figure 3.12, we can consider one electron travelling through the depletion region W of a semiconductor. Assuming v to be the speed of this electron, then the associated current is constant and can be written as:

$$I = \frac{qv}{W} = \frac{q}{\tau}$$

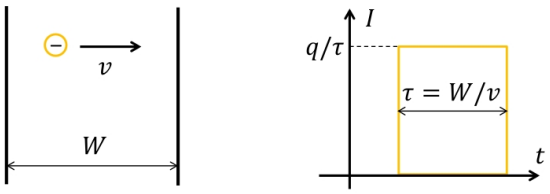


Figure 3.12: An electron travelling the depletion region of a semiconductor and the associated current pulse.

where we have defined the transit time of the electron in the depleted region:

$$\tau = \frac{W}{v}.$$

We will thus obtain a current pulse as the one represented in Figure, that will be constant for the whole time τ in which the electron is travelling through the depleted region, while it will be zero before the start and after the arrival of the electron. We can now ask ourselves: what happens when we have more than one electron travelling through this depleted region?

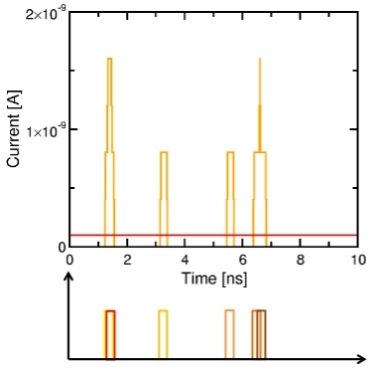


Figure 3.13: Current fluctuations related to six different electrons.

In Figure 3.13, we have a very small biased device. The associated transit time of this device is:

$$\tau = 100 \text{ ps}$$

the average current of this bias is:

$$\bar{I} = 10^{-10} \text{ A}$$

and the average number of charge carriers per unit time interval will be:

$$\bar{n} \simeq 6.2 \cdot 10^{8} \text{ s}^{-1}.$$

This means that, in the time interval considered (approximately 10 ns), we are observing the transit of about six electrons. As we can seen from the Figure, while the average value of the current will be constant, in reality the current term will show a series of peaks much higher than the average value due to

the presence of different, independent current pulses related to the different electrons flowing through the device. Averaging on these current fluctuations, we will obtain again the average value of the current considered. This current pulses, in general, are randomly located along the temporal axis, maybe even overlapping one with the other, thus giving rise to some random fluctuations and, therefore, to a noise term.

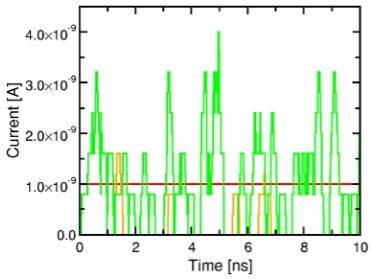


Figure 3.14: Current fluctuations related to sixty different electrons.

Increasing the average value of the current and therefore the number of different electrons that are crossing the device in the time interval considered:

$$\tau = 100 \text{ ps}, \quad \bar{I} = 10^{-9} \text{ A}, \quad \bar{n} \simeq 6.2 \cdot 10^{9} \text{ s}^{-1}$$

as represented in Figure 3.14, an increased number of electrons will cross the device, thus giving an increased number of pulses that can overlap one with the other. The current will still have a certain average value (that will obviously be higher than in the previous case), but also the current fluctuations will be increased. It is important to note that the physical origin of this noise term is completely different from the one of the thermal noise, that was related to the fluctuations in the value of the speed of the electrons.

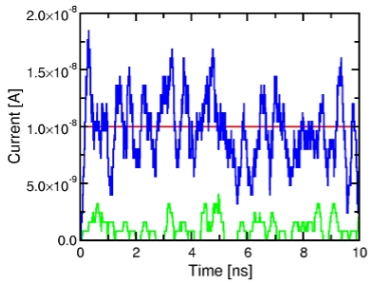


Figure 3.15: Current fluctuations related to six hundred different electrons.

Further increasing the average current and the number of electrons:

$$\tau = 100 \text{ ps}, \quad \bar{I} = 10^{-8} \text{ A}, \quad \bar{n} \simeq 6.2 \cdot 10^{10} \text{ s}^{-1}$$

we can observe in Figure 3.15 that these fluctuations are even more significant. To describe the whole process $x(t)$, that is the current flowing at a certain time instant t through the device, we can write it as the sum of all the different

normalized[4] current pulses $h(t-t_k)$ multiplied by the charge of each electron q:

$$x(t) = \sum_k qh(t-t_k).$$

Note that, in this description, t_k is the arrival time of the Poisson process and therefore it is a random variable (thus being related to the randomness of the process) and that we can sum over all the different pulses only since we are considering individual and independent[5] (thus totally uncorrelated) single-electron pulses. It is important to note that this signal will be a power signal and, therefore, we can define the power spectral density for such a signal as:

$$S(f) = \lim_{T\to\infty} \frac{1}{2T}\overline{|X_T(f)|^2}.$$

In this definition, we have defined $X_T(f)$ the truncated Fourier transform of $x(t)$ for:

$$t_k \in [-T, T]$$

thus obtaining:

$$X_T(f) = qH(f)\sum_k e^{-j2\pi f t_k}.$$

In this expression, we have clearly defined the Fourier transform of the shape of the single pulse as:

$$H(f) = \mathcal{F}[h(t)]$$

and the sum of the exponential terms is related to the fact that each one of these pulses will have a different arrival time in the interval $[-T, T]$, thus being shifted differently in time and giving rise to sum of various exponential terms. Calculating therefore this average to obtain the power spectral density, we can observe that the only randomly varying quantity will be this sum of exponentials (every other quantity, in fact, is deterministic), therefore the ensemble average will only involve the exponential terms, while other terms are constant and can be moved out of the limit:

$$\begin{aligned} S(f) &= q^2|H(f)|^2 \lim_{T\to\infty} \frac{1}{2T}\overline{\left|\sum_k e^{-j2\pi f t_k}\right|^2} = \\ &= q^2|H(f)|^2 \lim_{T\to\infty} \overline{\left(\sum_k e^{-j2\pi f t_k}\right)\cdot\left(\sum_m e^{j2\pi f t_m}\right)} = \\ &= q^2|H(f)|^2 \lim_{T\to\infty} \frac{1}{2T}\sum_{k,m}\overline{e^{-j2\pi f(t_k-t_m)}}. \end{aligned}$$

At this point, we can split the sum in two different sums:

$$\begin{aligned} S(f) &= q^2|H(f)|^2 \lim_{T\to\infty} \frac{1}{2T}\sum_{k=m}\overline{e^{-j2\pi f(t_k-t_m)}} + \\ &\quad + q^2|H(f)|^2 \lim_{T\to\infty} \frac{1}{2T}\sum_{k\neq m}\overline{e^{-j2\pi f(t_k-t_m)}} = \\ &= S(f)|_{k=m} + S(f)|_{k\neq m}. \end{aligned}$$

[4] Therefore, with unitary area.

[5] This crucial assumption allows us to use Poisson statistics.

and study them differently. It is important to note that this can be done due to the linearity of the ensemble average: the ensemble average of a sum is the sum of the ensemble averages. Studying the first term, we can observe that in the sum for $k = m$ all the exponentials will be equal to one, whose ensemble average is equal to one:

$$\sum_{k=m} \overline{e^{-j2\pi f(t_k - t_m)}} = \sum_{k=m} \overline{1} = \sum_{k=m} 1 = N_{2T}$$

and therefore we are just calculating the number of different values of $k = m$ we have; this will be equal to the number of independent electrons we have in the time interval $[-T, T]$ of amplitude $2T$, therefore it can be written as N_{2T}. This first term, therefore, can be rewritten as:

$$S(f)|_{k=m} = q^2 |H(f)|^2 \lim_{T\to\infty} \frac{N_{2T}}{2T} = q^2 \bar{n} |H(f)|^2 = q\bar{I}|H(f)|^2$$

where we have considered that:

$$\bar{n} = \lim_{T\to\infty} \frac{N_{2T}}{2T}$$

is the average number of electrons in a unitary time interval and where:

$$\bar{I} = q\bar{n}$$

is the average current, since it is the average number of carriers per unitary time interval multiplied by their charge. The second term, on the other hand, can be demonstrated[6] to be equal to:

$$S(f)|_{k\neq m} = q^2 \bar{n}^2 \delta(f) = \bar{I}^2 \delta(f)$$

thus giving the following total, bilateral power spectral density:

$$S(f) = \bar{I}^2 \delta(f) + q\bar{I}|H(f)|^2.$$

Studying this expression, we can observe that in a random process, if its average value is different from zero (for example, due to the addition of a constant) we expect to obtain a power spectral density in which one term is related to the fluctuations (and in this case, it is the second term), while the other is related to the DC average value (that will be represented by a Dirac's delta function). In this case, since the average current is obviously different from zero, we need to have a DC term associated to the average value.
Since then it is possible to demonstrate that:

$$H(0) = \int h(t)\ dt = 1$$

then the Fourier transform of the pulse $H(f)$ will be approximately white up to a certain high frequency. From the equivalent rectangle approximation, since the transit time is τ, this white noise approximation will hold in the following limit:

$$f \ll \frac{1}{\tau}.$$

[6]This has not been explicitly demonstrated during the lectures, but it can be found in an appendix to the slides.

The unilateral power spectral density, therefore, will be white up to a frequency equal to the reciprocal of the transit time, giving:

$$S(f) = 2q\bar{I}$$

where the extra factor 2 is needed to pass from the bilateral to the unilateral power spectral density.

3.7 Flicker noise

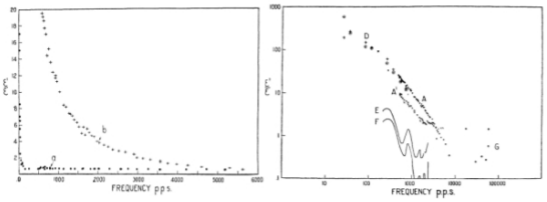


Figure 3.16: Spectral behaviour of the Flicker noise.

The last noise term that we will study is the flicker (or $1/f$) noise. This kind of noise was discovered by J. Johnson in 1925 while he was investigating the shot noise in vacuum tubes and, then, it has been found in many different electron devices. The name of this noise comes from the fact that it is similar to a white noise at high frequencies, while at low frequencies it clearly shows an $1/f$ dependency.

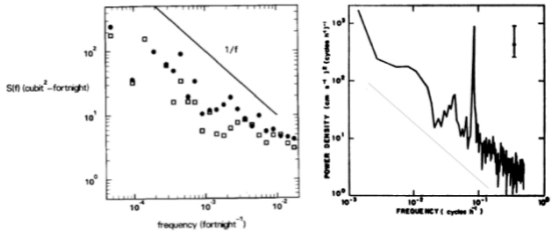


Figure 3.17: Some examples of Flicker noise in geophysics.

This noise term, then, has been found in several different processes, even completely uncorrelated with electronics. This means that this is a very general process, that can be found almost everywhere; it is therefore very difficult to develop a common theory or explanation for it, being intrinsically related to the physical behaviour of all the phenomena. In Figure 3.17 have been reported two examples from geophysics: on the left we can observe the levels of the Nile river flood, while on the right the velocity of some oceanic currents, both as a function of the frequency. In Figure 3.18, on the other hand, flicker noise has been studied in music, depending on the different authors and types of music considered.

A first, important property of the flicker noise that we can study is its scale

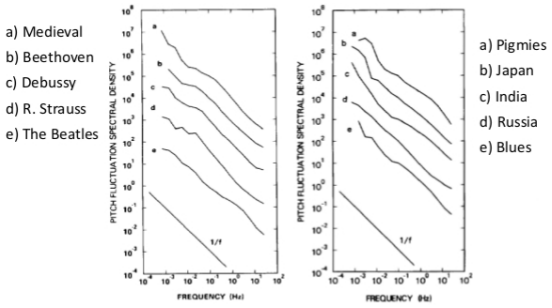


Figure 3.18: Some examples of flicker noise in music.

invariance. Considering in fact its power spectral density to be equal to:

$$S(f) = \frac{K}{f}$$

where K is a suitable constant, we can write the power P associated to this noise in the bandwidth $[f_L, f_H]$ as:

$$P = \int_{f_L}^{f^H} S(f)\ df = \int_{f_L}^{f_H} \frac{K}{f}\ df = K \ln\left(\frac{f_H}{f_L}\right)$$

and we can immediately observe that this power will exclusively depend on the ratio between the two limiting frequencies of the bandwidth. As an example, therefore, we can observe that the power of this noise between 0.1 and 1 Hz will be exactly equal to the one between 100 kHz and 1 MHz, while the second bandwidth is significantly large than the first one. This is completely different from what we had in white noise, where the power in the second bandwidth would have been one million times larger than the one in the first one. This properties is due to the fact that the ratios between the two limiting frequencies of the bandwidth are identical:

$$\frac{0.1 \text{ Hz}}{1 \text{ Hz}} = \frac{100 \text{ kHz}}{1 \text{ MHz}} = 0.1$$

and it is called scale invariance. It will obviously not hold for every kind of noise whose power depends linearly on the frequency. Another characteristic of this kind of noise is that it diverges both in the high-frequency and in the low-frequency limit. This can be immediately demonstrated from the expression of the power associated to this noise:

$$P = K \ln\left(\frac{f_H}{f_L}\right) = \begin{cases} \to +\infty & \text{if } f_H \to +\infty \\ \to +\infty & \text{if } f_L \to 0 \end{cases}$$

and this is a particularly problematic issue of this kind of noise. This was obviously not true for white noise, that is diverging only in the high-frequency limit. Therefore, how it is possible to deal with these problems?
First of all, it is important to observe that a band-limited flicker noise is stationary. This would not be true if we were dealing with an unlimited bandwidth, but

it can be arbitrarily limited or through filtering (that will define a bandwidth and a power for this noise signal) or, as we will briefly see, as a consequence of the physical properties of any system. From a practical point of view, in fact, cut-off frequencies always occur. For the high-frequency limit, for example, we can consider the response time of the system (that is in the order of 1 ps) as the limit, obtaining a cut-off frequency approximately equal to 160 GHz, or the time λ/c taken by light to cross an atom, that gives an approximate frequency of 10^{21} Hz. For the low-frequency limit, the cut-off frequency is set by the observation time we are considering: for a one day observation, it will be 10^{-5} Hz; for a one year observation, it will be $3 \cdot 10^{-8}$ Hz; for an observation as long as the age of the Universe, the cut-off frequency will be 10^{-17} Hz. These are clearly exaggerated limits (in reality, the bandwidth will be much narrower), but are useful for understanding that the previous limits can never be observed, since the low-frequency limit will involve an infinite observation time, while the high-frequency limit will require a zero time constant in the system response, that are clearly impossible. In the worst case, therefore, using the previous exaggerated estimates, the bandwidth will be extended for approximately 38 decades and, therefore, the maximum power associated to the flicker noise will only be 38 times larger than the one in the bandwidth going from 1 Hz to 10 Hz.

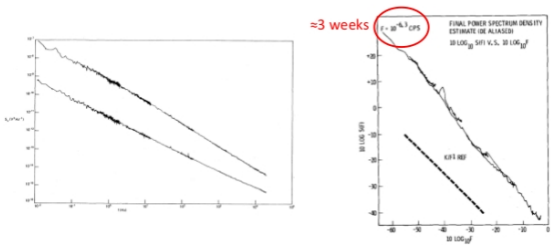


Figure 3.19: Experimental results for finding a minimum cut-off frequency for flicker noise.

Since every model for flicker noise predicts the existence of a minimum cut-off frequency for flicker noise, several different experiments have been performed for measuring it. However, this value is in general extremely small and, therefore, it has never been observed. Two examples of these results have been reported in Figure 3.19.

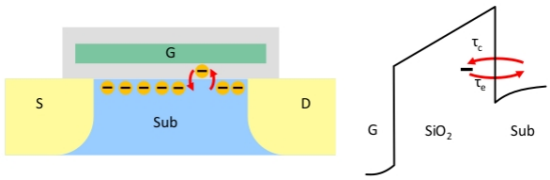


Figure 3.20: An MOS structure needed for studying the random telegraph noise.

In particular, we can study a model for the flicker noise in MOS structures, as represented in Figure 3.20: the random telegraph noise (RTN). If this MOS structure is suitably biased:

$$V_S = 0, \quad V_G > 0, \quad V_D > 0$$

it is possible to form a conductive channel below the oxide layer, thus going from the source to the drain of the device. In this channel, then, we will have a flow of electrons, thus giving a current that will flow from the drain to the source. The problem, in this case, is that just above this conductive channel we have an interface between an oxide (in particular, silicon oxide SiO_2) and the bulk semiconductor (in general, doped silicon Si). This kind of interfaces presents many defects, in particular bonds that can accept or emit electrons, thus being called traps. This means that it may happen that one of the electrons in the conductive channel is captured (respectively, released) from these traps. From an electrostatic point of view, nothing changes in the structure; however, this will determine a random decrease (respectively, increase) in the current that is flowing through the device. The capture and the emission of single electrons from the oxide traps will have two different time constants τ_c and τ_e and the fluctuations in the drain current will determine fluctuations in the threshold voltage.

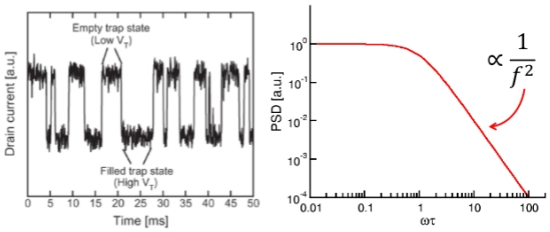


Figure 3.21: Random telegraph noise connected to the presence of a single trap.

Considering the presence of a single trap, as in Figure 3.21, this will determine some current fluctuations between a low threshold voltage level (in which the trap state is empty and the current is higher) and an high threshold voltage level (in which the trap state is filled and the current is lower). Therefore, these current fluctuations are related to the catch and release of electrons from this single trap state. Studying the power spectral density of this single-trap noise, we can observe that it shows an $1/f^2$ dependency from the frequency, thus not being a flicker noise.

If we now consider the presence of many different trap states in an MOS device, each one with its own different pole in the power spectral density due to the fact that each trap state has a different emission or capture time constant, we can obtain the power spectral density in Figure 3.22. Since these traps are independent one from the other, in fact, we can add the various different power spectral densities, obtaining the $1/f$ dependency that is typical of the flicker noise.

In semiconductors, therefore, the flicker noise is mainly related to the presence of distributed capture and emission processes. This is the reason why MOS devices have an high flicker noise: the presence of a silicon-oxide interface, that presents a large density of defects, determines an high trap density. On the other hand, JFET and BJT technologies will have better performances, since the capture and emission processes will take place in the space-charge regions, not near to a silicon-oxide interface, thus presenting a much lower density of trap states. The flicker noise, therefore, will be an issue mainly in MOS amplifiers.

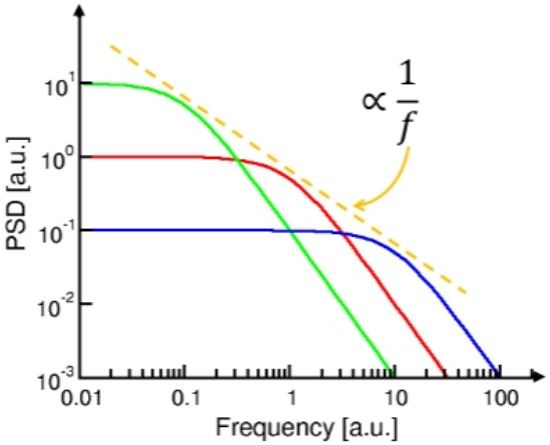


Figure 3.22: Power spectral density of the random telegraph noise connected to the presence of various traps.

3.8 Noise in linear circuits and OAs

After this introduction to noise, we are now able to study the effect of the noise on linear circuits and operation amplifiers. In particular, therefore, we will study a special class of systems: the linear and time-invariant (LTI) systems. This class of systems is, in general, quite broad and it contains all the systems and circuits that we have previously described.
The first properties of these systems is their linearity. It can be defined as the fact that, if an input $x_i(t)$ produces an output $y_i(t)$, then the following discrete sum of inputs over some coefficients c_k:

$$\sum_k c_k x_k(t)$$

will produce as an output the sum of the single output weighted on the same coefficients:

$$\sum_k c_k y_k(t).$$

This means that the superposition principle holds. Generalizing even more, we can say that a continuous sum of infinitesimal inputs, that is an integral:

$$\sum c(\tau)x(t,\tau)\ d\tau$$

will produce an analogous sum, with the same coefficients, of the outputs:

$$\int c(\tau)y(t,\tau)\ d\tau.$$

The second peculiar property of these systems is their time invariance, that states that a temporal shifted input will produce an output that is shifted in time of the same time interval. Therefore, if $x(t)$ is the input that produces an output $y(t)$, then feeding an LTI system with $x(t+T)$ as an input we will obtain $y(t+T)$ as an output.

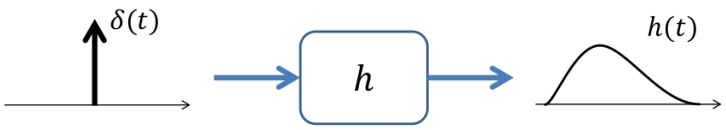


Figure 3.23: Impulse response of a linear and time-invariant system.

These two properties have an important consequence: a linear and time-invariant system can be completely characterized by its impulse response $h(t)$. To understand its meaning, if we are feeding a certain system with an impulse $\delta(t)$ at the input, we will obtain the impulse response $h(t)$ of the system as an output. Moreover, this quantity will control the response of the signal to any system, not only to impulses, since any signal can be considered as a sum of several different time-shifted delta signals.

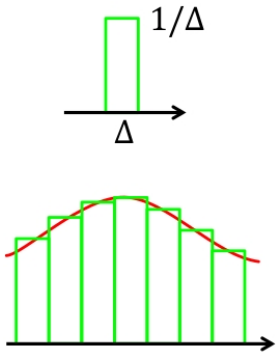


Figure 3.24: The "sifting" principle.

To better understand what we meant in the previous sentence, we can try to study the so called "sifting" principle. In particular, we can define an approximated delta function with unitary area δ_Δ that is represented in Figure 3.24. It will be a rectangle of amplitude Δ and height (from the unitary are requirement) $1/\Delta$, and this clearly leads to the following property:

$$\delta_\Delta(t) \xrightarrow[\Delta \to 0]{} \delta(t).$$

Given therefore a certain signal $x(t)$, we can obtain a piecewise constant approximation of this signal by evaluating it at different time instants a multiplying it for the previously define approximated delta function:

$$x_\Delta(t) = \sum_k x(k \cdot \Delta)\delta_\Delta(t - k \cdot \Delta) \cdot \Delta.$$

Then, in the limit in which the approximated delta function tends to the delta function, we can write:

$$x(t) = \lim_{\Delta \to 0} x_\Delta(t) = \int x(\tau)\delta(t - \tau)\, d\tau$$

where the last integral is the convolution of the signal with a delta. This is nothing but a property of the convolution operation: the convolution of a generic

signal with a delta function is equal to the evaluation of the signal in the point in which the argument of the delta function is identically equal to zero.
If this property holds, we can consider the output of a linear and time-invariant system when we are feeding it with an $x_\Delta(t)$ piecewise constant signal. This output, from the linearity property[7], will be the superposition of the different response to the various time-shifted approximated delta signals $\delta_\Delta(t-k\cdot\Delta)$, thus being several time-shifted[8] impulse responses, multiplied by suitable coefficients:

$$y_\Delta(t) = \sum_k x(k\cdot\Delta)\cdot h_\Delta(t-k\Delta)\cdot\Delta.$$

In the limit in which the approximated delta function tends to a delta function, we can write:

$$y(t) = \lim_{\Delta\to 0} y_\Delta(t) = \int x(\tau)h(t-\tau)\,d\tau = x(t) * h(t).$$

We have thus demonstrated that the output of a system for any given input signal will be the convolution of this input signal with the impulse response of the filter[9], that is therefore also defined as the weighting function of the filter.

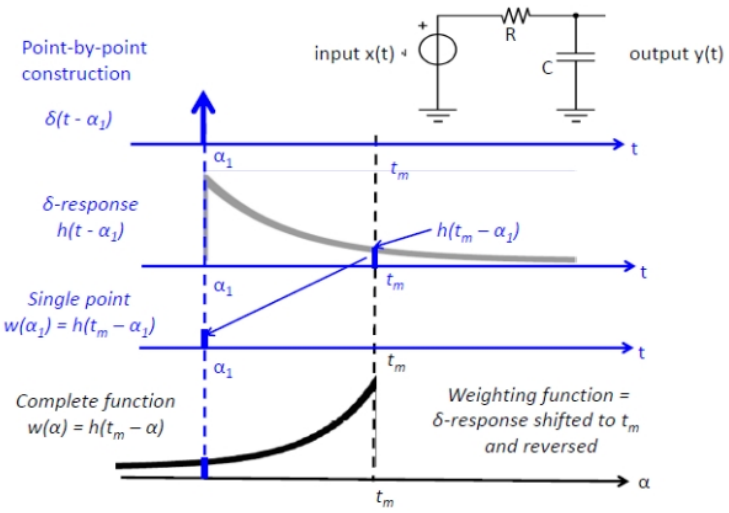


Figure 3.25: Interpretation of the impulse response of a filter.

To better understand the meaning[10] and the usefulness of the impulse response of a system, we can study the problem reported in Figure 3.25. In this case, we are studying the circuit represented in the upper part of the drawing, that is clearly a linear and time invariant system. Assuming therefore an impulse

[7] This is the important of the linearity assumption: it will allow us to write the output as a linear combination of the responses to the different inputs.

[8] Note that this is a consequence of the time invariance of the system: since the output of the system to a time-shifted input is a time-shifted output, we can properly define a time-shifted impulse response.

[9] From now on, the word filter will be often used as a synonymous of linear and time-invariant system.

[10] This kind of interpretation is strongly encouraged by the teacher, since it can be useful for understanding more complicated mathematical properties that we will study later on.

input $\delta(t-\alpha_1)$, we know that the output of the system will be a decaying exponential[11] that will be equal, from what we have previously said, to the impulse response of the system:

$$h(t-\alpha_1) = \frac{1}{\tau} e^{-\frac{t-\alpha_1}{\tau}}.$$

Assuming now to evaluate the contribution of this impulse at time α_1 to the output at time t_m, we can write it as:

$$w(\alpha_1) = h(t_m - \alpha_1).$$

If the input is a more complicated signal, we can then consider it as a series of delta functions, each one with a different amplitude and a different temporal position α and study the effect of this signal on the output at time t_m by considering the weighting factor, for each one of these components, depending on the position of the impulse α:

$$w(\alpha) = h(t_m - \alpha).$$

We can therefore immediately observe that this weighting function will be equal to the impulse response of the system shifted in time of t_m and reversed and it will be useful for calculating the contribution of a certain component of the input related to time α on the output evaluated at time t_m.
We can now try to study how a linear and time-invariant system reacts to noise. In this case, since the input of the system is a random process, we can say that also the output of the system will be a random process, that therefore can be studied through its autocorrelation. In the more general case, in which the process is not stationary, this autocorrelation will depend explicitly on both time t_1 and t_2, therefore we can write:

$$R_{yy}(t_1, t_2) = \overline{y(t_1)y(t_2)}.$$

According to the definition of the output of a system as the convolution of the input with the impulse response of the system, then, we can write it as:

$$R_{yy}(t_1, t_2) = \overline{\int x(\alpha)h(t_1-\alpha)\ d\alpha \cdot \int x(\beta)h(t_2-\beta)\ d\beta}$$

and changing the position of the integrals, it can be rewritten as:

$$R_{yy}(t_1, t_2) = \overline{\iint x(\alpha)h(t_1-\alpha)x(\beta)h(t_2-\beta)\ d\alpha\ d\beta}.$$

Note that, in the previous calculations, we have considered that the ensemble average is performed over all the different realizations but, since they are output signals, we can write them using the impulse response of the system. However, this impulse response is a deterministic quantity, therefore the only quantities that will be affected by the ensemble average will be the input signal $x(\alpha)$ and $x(\beta)$. Remembering the definition of autocorrelation of a signal, this gives:

$$\begin{aligned} R_{yy}(t_1, t_2) &= \iint \overline{x(\alpha)x(\beta)}h(t_1-\alpha)h(t_2-\beta)\ d\alpha\ d\beta = \\ &= \iint R_{xx}(\alpha, \beta)h(t_1-\alpha)h(t_2-\beta)\ d\alpha\ d\beta. \end{aligned}$$

[11] In fact, the step response of this system is an increasing exponential, and the two are obviously related one to the other.

Since the two impulse responses will depend only on α or on β, we can first integrate on one variable, considering the other as a constant, and then integrate on the other; recognizing the definition of a convolution product:

$$\begin{aligned} R_{yy}(t_1,t_2) &= \int h(t_2-\beta) \int R_{xx}(\alpha,\beta) h(t_1-\alpha)\ d\alpha\ d\beta = \\ &= \int h(t_2-\beta)\left[R_{xx}(t_1,\beta) * h(t_1)\right]\ d\beta = \\ &= h(t_2) * \left[R_{xx}(t_1,t_2) * h(t_1)\right] = R_{xx}(t_1,t_2) * h(t_2) * h(t_1) \end{aligned}$$

where we have considered that the convolution product satisfies the distributive, associative and commutative properties.
If we now assume to have a stationary input noise (flicker noise would be the only exception if it were considered on an unlimited bandwidth), defining the reciprocal delay between the two signals:

$$\tau = t_2 - t_1$$

since the output of a stationary input noise is stationary as well, we need to calculate the following autocorrelation:

$$R_{yy}(\tau) = \overline{y(t)y(t+\tau)}$$

where:

$$t_1 = t, \quad t_2 = t+\tau.$$

From the previous calculation, that was more general, we can directly write:

$$R_{yy}(\tau) = \iint R_{xx}(\beta-\alpha) h(t-\alpha) h(t+\tau-\beta)\ d\alpha\ d\beta$$

where we have considered that since the noise is stationary, the autocorrelation depends exclusively on the reciprocal delay:

$$\gamma = \beta - \alpha.$$

Changing variables, therefore:

$$R_{yy}(\tau) = \iint R_{xx}(\gamma) h(t-\alpha) h(t+\tau-\gamma-\alpha)\ d\alpha\ d\gamma$$

and integrating first on α and then on γ we obtain:

$$R_{yy}(\tau) = \int R_{xx}(\gamma) \int h(t-\alpha) h(t-\alpha+\tau-\gamma)\ d\alpha\ d\gamma.$$

Applying the following change of variable:

$$z = t - \alpha$$

we can obtain[12]:

$$R_{yy}(\tau) = \int R_{xx}(\gamma) \int h(z) h(z+\tau-\gamma)\ dz\ d\gamma$$

[12]Note that we have considered that, applying rigorously the change of variable, the internal integral would have been:

$$-\int h(z) h(z+\tau-\gamma)\ dz = -k_{hh}(\tau-\gamma) = k_{hh}(\tau-\gamma)$$

due to the symmetry of the autocorrelation function $k_{hh}(\tau-\gamma)$, that is even and real. Note that this is due to the fact that we are dealing with the autocorrelation of a deterministic signal and not of a stochastic signal, where things are more complicated.

that gives, recognizing the autocorrelation of a deterministic signal and a convolution product:

$$R_{yy}(\tau) = \int R_{xx}(\gamma) k_{hh}(\tau - \gamma)\ d\gamma = R_{xx}(\tau) * k_{hh}(\tau).$$

We have thus proven that the ensemble autocorrelation of the output is the convolution of the ensemble autocorrelation of the input with the autocorrelation of the impulse response function. Assuming a stationary input noise, therefore, the mean square value of this noise can be written as:

$$\overline{n_y^2} = R_{yy}(0) = \int R_{xx}(\gamma) k_{hh}(\gamma)\ d\gamma$$

while if the input noise is non-stationary also the output noise will be non-stationary and we must use the more complicated complete expression:

$$\overline{n_y^2(t)} = \iint R_{xx}(\alpha, \beta) h(t - \alpha) h(t - \beta)\ d\alpha\ d\beta.$$

In the frequency domain, as a consequence of the property of the convolution and of the Fourier transform, considering a deterministic signal as an input and, therefore, a deterministic signal as an output, we will obtain:

$$Y(f) = X(f) H(f).$$

On the other hand, if the input and consequently also the output are stationary noises, applying the Fourier transform to the following relationship:

$$R_{yy}(\tau) = R_{xx}(\tau) * k_{hh}(\tau)$$

we obtain:

$$S_y(f) = S_x(f) \cdot \mathcal{F}[k_{hh}(\tau)] = S_x(f) \left|H(f)\right|^2.$$

The mean square value of this noise, then, can be written as:

$$\overline{n_y^2} = \int S_y(f)\ df = \int S_x(f) |H(f)|^2\ df$$

and this expression could have been obtained also starting from the one that we have previously proven:

$$\overline{n_y^2} = R_{yy}(0) = \int R_{xx}(\tau) k_{hh}(\tau)\ d\tau$$

by means of the Parseval's theorem.
A particular case is the one of white stationary noise, whose ensemble autocorrelation can be written as:

$$R_{xx}(\tau) = \lambda \cdot \delta(\tau)$$

and therefore the noise at the output of the filter will have the following autocorrelation:

$$R_{yy}(\tau) = R_{xx}(\tau) * k_{hh}(\tau) = \lambda \cdot k_{hh}(\tau).$$

To obtain the previous expression, we have considered that the convolution of a certain function with a delta function gives the evaluation of the function in the point in which the argument of the delta is equal to zero. This clearly shows that the output noise is no longer a white noise. From an intuitive point of view, this is a consequence of the fact that the output of the system is a weighted average of all the previous inputs, and since this will be true for any output at any time, we will have a certain autocorrelation in the noise at the output of the filter, thus not making it white. The mean square value of the noise, then, can be calculated as:

$$\overline{n_y^2} = R_{yy}(0) = \lambda k_{hh}(0) = \lambda \int h^2(t)\ dt.$$

Alternatively, the same analysis can be performed in the frequency domain, starting from the power spectral density of the input white noise:

$$S_x(f) = \lambda$$

and calculating, using the previous formula, the power spectral density of the output noise:

$$S_y(f) = S_x(f)|H(f)|^2 = \lambda|H(f)|^2.$$

Also in this case, we can immediately observe that the output noise is not a white noise and we are filtering the frequencies with the function $|H(f)|^2$, thus enhancing some components and rejecting the others. Integrating this power spectral density, we can obtain the mean square value of this noise:

$$\overline{n_y^2} = \int S_y(f)\ df = \lambda \int |H(f)|^2\ df$$

and observe that, as we have previously stated, the two relations for the mean square value of the noise could have been directly related using Parseval's theorem.

3.9 Noise factor, noise figure and signal-to-noise ratio

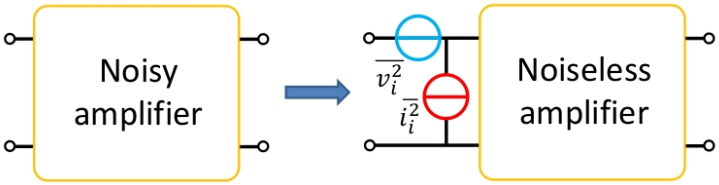


Figure 3.26: A noisy amplifier and the equivalent noise representation.

At this point, we can study how the presence of noise affects the performances of an amplifier. We can start from a general consideration: every component of an amplifier has its own sources of noise. Therefore, considering an amplifier whose input pins are grounded, we will obtain from this amplifier an output

that is different from zero and that varies depending on all these noise terms. It is clearly impossible to take explicitly into account all these sources of noise; we need thus to find an alternative way of representing the noise that is affecting the overall amplifier. This can be done, for example, as it is represented in Figure 3.26. In this case, we are considering a noiseless amplifier and we have connected at its input pins an equivalent noise voltage generator $\overline{v_i^2}$ and an equivalent noise current generator $\overline{i_i^2}$. Therefore, these two generators will give, at the output of the noiseless amplifier, the same noisy signal that we would have obtained from a real amplifier. Moreover, they can be used also for representing noise terms for every input condition of our device.
A question may arise: why do we need two different generators, one for a voltage and the other for a current? To understand it, we can consider only the presence of the current generator $\overline{i_i^2}$, neglecting the voltage one. In this case, if we assume, as a special input condition, that we have short-circuited the two input pins of the amplifier, we are easily getting rid of this noise term and obtaining zero at the output of the amplifier, without any noise signal. On the other hand, if we assume to have only the voltage generator $\overline{v_i^2}$ we can observe that if the two pins are two open circuits we will not have any current flowing through the circuit and the two pins (that are connected, inside the amplifier, from the input impedance of the amplifier) will be at the same voltage imposed by the equivalent noise voltage generator; also in this case, we are obtaining zero at the output of the amplifier, without any noise. Therefore, both the equivalent noise generators are fundamental.
It is important to note that, in principle, these two parameters come form several different noise sources in the various elements of the device and, from this reason, they will surely be somehow (at least partially) correlated. However, in order to be able to solve any practical problem, since we are not able to express explicitly this very difficult correlation, we will neglect this correlation between the different noise sources and we will assume them to be totally uncorrelated.

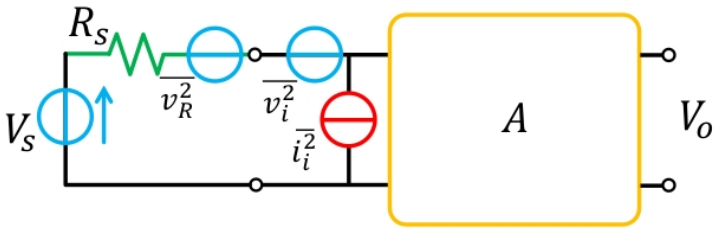


Figure 3.27: The equivalent noise representation of an amplifier to which we have connected a certain signal.

We can now assume to have a certain signal at the input of our noisy amplifier, that can be represented as in Figure 3.27 through the Thévenin equivalent of the input network. We can thus add at the input of the amplifier a voltage generator V_S that will give the signal at the input of the amplifier, a source resistance R_S at the input of the amplifier and a noise equivalent voltage generator $\overline{v_R^2}$ that will be related to the thermal noise in the source resistor R_S:

$$\overline{v_R^2} = 4k_B T R_S.$$

Considering a sinusoidal signal at the input of the amplifier:

$$v_S = V_S \cos(\omega t)$$

and a certain frequency interval Δf, we can immediately write the signal at the output of the amplifier, that will be equal to:

$$V_o = AV_S \cos(\omega t).$$

However, superimposed to this signal we will find, at the output, also a noise signal that can be written, in the frequency domain, considering three different noise sources:

- the thermal noise in the source resistance, whose power spectral density (in terms of voltage) is:
$$S_{VR} = 4k_B T R_S;$$
- the equivalent voltage noise of the amplifier, whose power spectral density is S_V;
- the equivalent current noise of the amplifier, whose power spectral density is S_I.

Therefore, the output power spectral density of the noise will be a linear superposition of the power spectral densities at the input (assuming an infinite input impedance) multiplied by the square of the amplification factor[13]:

$$S_{Vo} = S_{VR}A^2 + S_V A^2 + S_I R_S^2 A^2.$$

From the definition of the mean square value of the noise output signal, then, we can write:

$$\overline{V_o^2} = \int S_{Vo}\, df = \int \left(S_{VR}A^2 + S_V A^2 + S_I R_S^2 A^2\right)\, df.$$

In this case, however, we have assumed the signal at the input to be at a very precise frequency ω, therefore we can filter any other signal coming to the amplifier with a filter with a small bandwidth equal to Δf centred around ω and, assuming the power spectral density to be constant over this domain[14] we can calculate this integral as:

$$\overline{V_o^2} = \left(S_{VR} + S_V + S_I R_S^2\right) A^2 \Delta f.$$

It is important to note that this term that we have just calculated is a variance and it is related to the noise; this means that the larger will be this value, the larger will be the noise contribution superimposed to the signal.

A good parameter for evaluating the contribution of noise on a signal and therefore the cleanness of a certain signal is the so called signal-to-noise ratio S/N. It is defined as the ratio between the maximum amplitude of the signal

[13] In fact, we know that the power spectral density is proportional to the square of the noise signal and, since the amplification factor will amplify this signal, it will be squared in the power spectral density.

[14] This is not a trivial assumption: the white noise will be for sure constant, but the flicker noise will show an $1/f$ dependency that makes this assumption relevant.

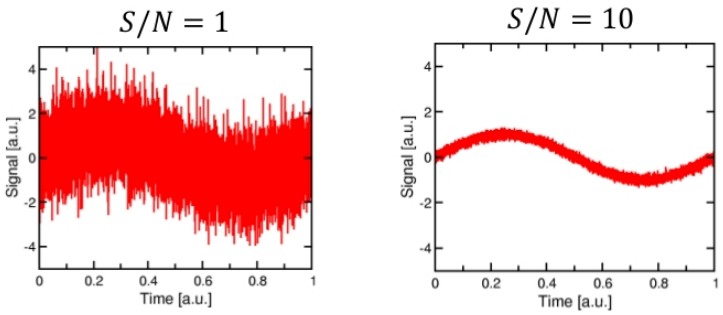


Figure 3.28: Two examples of signal-to-noise ratio.

that we are dealing with and the root mean square value of the amplitude of the noise[15]:

$$\left(\frac{S}{N}\right)_{out} = \frac{\text{maximum signal amplitude}}{\text{noise rms amplitude}} = \frac{V_o}{\sqrt{\overline{V_o^2}}}.$$

From the examples represented in Figure 3.28 we can immediately see that the higher is the signal-to-noise ratio, the smaller is the noise component as compared to the signal and therefore the better is our device. Applying this formula to the previous example, we can observe that:

$$\left(\frac{S}{N}\right)_{out} = \frac{V_o}{\sqrt{\overline{V_o^2}}} = \frac{AV_S}{\sqrt{\left(S_{VR} + S_V + S_I R_S^2\right) A^2 \Delta f}} = \\ = \frac{V_S}{\sqrt{\left(S_{VR} + S_V + S_I R_S^2\right) A^2 \Delta f}}.$$

To evaluate the effect of the amplifier, we can now compare the signal-to-noise ratio before the amplifier, where the only noise term is the thermal noise in the resistance of the source:

$$\left(\frac{S}{N}\right)^2 = \frac{V_S^2}{S_{VR}\Delta f}$$

with the signal-to-noise ratio after the amplifier, where also the equivalent noise generators of the amplifier have come into play:

$$\left(\frac{S}{N}\right)^2 = \frac{V_S^2}{(S_{VR} + S_V + S_I R_S^2)\Delta f}.$$

We can therefore immediately observe that the amplifier has reduced the signal-to-noise ratio, because it has introduced its own noise that is related to the power spectral density of the noise equivalent voltage generator and of the noise equivalent current generator (this last one evaluated on the resistance of the source). Amplifying a signal, therefore, we reduce its signal-to-noise ratio and to evaluate this reduction we can define two new parameters: the noise factor and the noise figure. The noise factor F is defined as the ratio between the square of

[15] Both evaluated at the output or at the input of a certain device.

the signal-to-noise ratio at the input and the square of the signal-to-noise ratio at the output of the amplifier:

$$F = \frac{\left(\frac{S}{N}\right)^2_{in}}{\left(\frac{S}{N}\right)^2_{out}}.$$

In this particular case, it can be evaluated, from the previous expressions:

$$F = 1 + \frac{S_V(f) + S_I(f)R_S^2}{S_{VR}}$$

where S_{VR} is the power spectral density of the thermal noise in the source resistance, that in principle, from the model that we have studied, is white:

$$S_{VR} = 4k_BTR_S$$

while the power spectral densities of the equivalent current and voltage noise generators will depend more or less significantly on the frequency. In an ideal amplifier, the amplification process will not add any noise to the signal and therefore the noise factor will be:

$$S_V = 0,\ S_I = 0 \ \Rightarrow\ F = 1.$$

In a real device, however, we will always have these noise sources, thus giving a noise factor that is greater than one:

$$F > 1.$$

Moreover, since the power spectral density of the noise equivalent generators depend on the frequency (since they include terms related to the flicker noise, for example), also the noise factor of a certain amplifier will depend on the frequency. It is then possible to observe that:

$$S_{VR} = 4k_BTR_S \propto R_S$$

while at the numerator:

$$S_IR_S^2 \propto R_S^2$$

and therefore it will be possible to find an optimum value of the resistance of the source for minimizing the noise factor. Making this dependence explicit:

$$F = 1 + \frac{1}{4k_BT} \cdot \left(\frac{S_V}{R_S} + S_IR_S\right)$$

we can find the optimum value by imposing:

$$\frac{\partial F}{\partial R_S} = 0 \ \rightarrow\ -\frac{S_V}{R_S^2} + S_I = 0$$

thus obtaining an optimum resistance that is equal to:

$$R_{S,opt} = \sqrt{\frac{S_V(f)}{S_I(f)}}$$

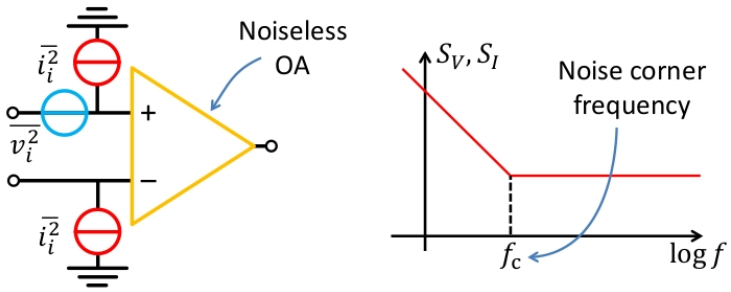


Figure 3.29: Equivalent representation of a noisy operation amplifier and the associated power spectral density.

and it will obviously depend on the frequency. An equally useful parameter that can be used for characterizing the behaviour of an amplifier with respect to the noise is the noise figure, that is simply defined as[16]:

$$NF = 10 \cdot \log_{10}(F) \quad [\text{dB}].$$

We can now specialize these general considerations for the case of an operation amplifier. In this case, we will have again one noise equivalent voltage generator that is applied to one of the input pins, while we will have two noise equivalent current generators, one for each input pin. These two current generators will have, at least in principle, the same power spectral density, but they will be totally independent random processes, thus having a small (and considered negligible) correlation. The spectral densities of all these noise equivalent generators, that are represented in the right hand-side of Figure 3.29, will contain both a white component and a flicker one. At low frequencies, therefore, the flicker component will be predominant over the white one, while at high enough frequencies the white noise component will prevail. The frequency at which these two components have the same magnitude and thus we are changing from a $1/f$ behaviour (related to the flicker component) to a constant one (that comes from the white noise) is called the noise corner frequency f_c and it is a parameter of the device considered. This frequency will be useful for understanding the behaviour of the prevailing noise sources for a particular range of frequencies. From an indicative point of view, we can say that the square of the voltage noise power spectral density for the white noise $\sqrt{S_V}$ will be between 1 and 10 or 20 nV/$\sqrt{\text{Hz}}$ for the BJT technology, while in the JFET and MOS it is usually higher, between 20 and 30 nV/$\sqrt{\text{Hz}}$. On the other hand, the square of the current noise power spectral density for the white noise $\sqrt{S_I}$ will be about a few pA/$\sqrt{\text{Hz}}$ for the BJT technology, while it will be lower, being only a few fA/$\sqrt{\text{Hz}}$ for the JFET and even lower[17] for CMOS one. This difference can be immediately explained if we consider that BJT devices have larger bias currents with respect to the other ones. Considering the noise corner frequency, it is possible to observe that it is between one and 100 Hz for the BJT and the JFET

[16]Note that this definition is nothing but a conversion of the noise factor in decibels. In fact, since the noise factor is the ratio between the squares of the signal to noise ratio, it will be proportional to the ratio between the squares of two signals, thus being proportional to the ratio between two powers, from which comes the factor 10 that is multiplying the logarithm.

[17]Sometimes, this value is so small that it is impossible to be measured.

technology, while it is up to a few kilohertz or even more for MOS devices (even though a few devices of this kind with a much smaller noise corner frequency exist). This difference in the noise corner frequency can be immediately related to the value of the flicker noise in the device we are considering. In the MOS technology, in fact, the generation and recombination processes at the interfaces between oxide and semiconductor are significantly increasing the importance of the flicker noise in these devices, thus increasing also the frequency at which the contribution of the flicker noise is equal to the contribution of the white noise (that is constant and independent from the frequency). This makes the noise corner frequency larger for MOS device than for BJT or JFET ones. A few examples of the roots of the power spectral density with respect to voltage and to current can be found, respectively, in Figure 3.30 and 3.31.

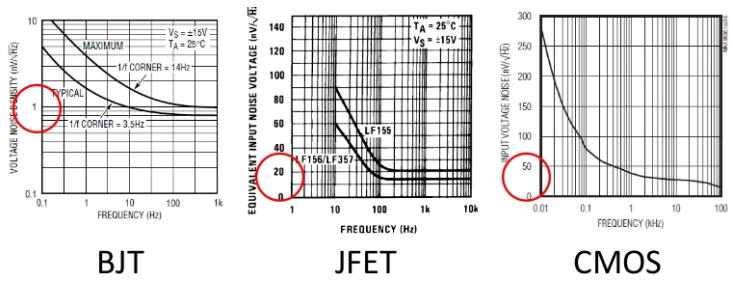


Figure 3.30: Root of the noise equivalent voltage power spectral density in different classes of devices.

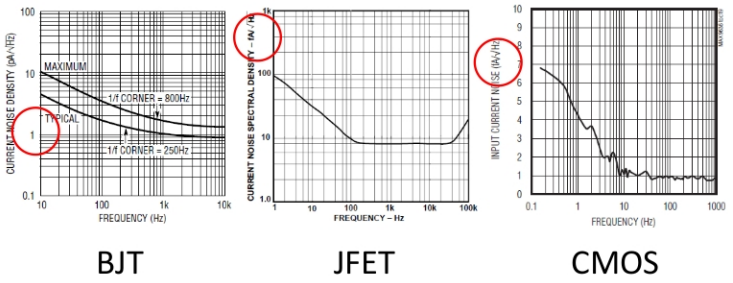


Figure 3.31: Root of the noise equivalent current power spectral density in different classes of devices.

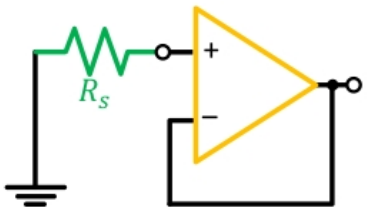


Figure 3.32: Buffer circuit considered.

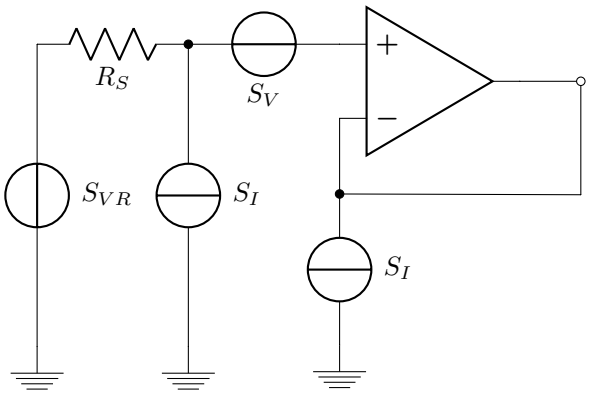


Figure 3.33: Buffer circuit where we added the noise equivalent sources.

We can now try to extend this comparison of the different operation amplifier to the buffer circuit that is represented in Figure 3.32. Assuming that the input is grounded and therefore that we do not have any signal at the output, adding the noise equivalent sources as in Figure 3.33 and noting that the noise equivalent current generator connected to the inverting pin will not have any contribution to the noise on the output we can write the power spectral density of the output as:

$$S_{Vo} = S_{VR} + S_V + S_I R_S^2 = 4k_B T R_S + S_V + S_I R_S^2.$$

We can plot these three different contributions as in Figure 3.34 and observe that in different region we will have a different prevailing noise source.

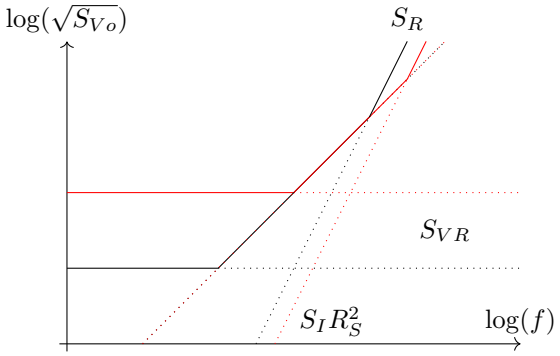


Figure 3.34: Different noise contributions in a BJT technology (in black) and in a JFET/CMOS technology (in red). The prevailing contribution is represented with solid lines.

Summing all the different contributions that are represented separately in Figure 3.34, it is possible to obtain the smooth behaviour that is represented in Figure 3.35. We can immediately observe, therefore, that there is a first region in which the power spectral density of the noise on the output of a BJT operation amplifier is lower with respect to the one of JFET and CMOS ones; then we have an intermediate region in which the two noise contributions are similar;

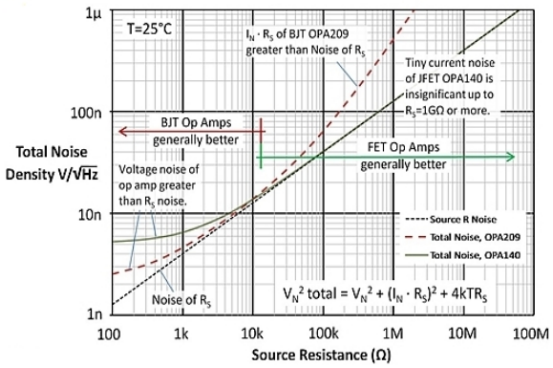


Figure 3.35: Comparison between two different technologies of operation amplifiers.

last, we have a region in which the power spectral density of the noise is lower for the JFET and CMOS technology.

3.10 Feedback and noise

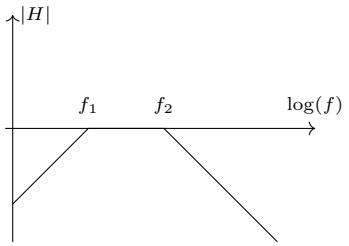


Figure 3.36: The pass-band filter considered.

We can now consider explicitly the noise in feedback networks. In particular, what will happen when we filter the noise? Assuming to have a pass-band filter as the one represented in Figure 3.36:

$$H(s) = \frac{s\tau_1}{(1 + s\tau_1)(1 + s\tau_2)}$$

we can investigate the effect of this filtering operation on the noise.
The first case we can consider is the one in which the noise at the input is assumed to be white. In this case, the power spectral density at the input of the system will be:

$$S_{in} = \lambda$$

and at the output of the filter it will be:

$$S_o = \lambda |H(\omega)|^2.$$

It is important to note that in a filter, since this is a physical system, we will be dealing only with positive frequencies, therefore the power spectral densities we

are considering are unilateral power spectral densities. Integrating this power spectral density, we can write the mean square value of the output signal as:

$$\overline{V_o^2} = \int_0^\infty S_o(\omega)\ \frac{d\omega}{2\pi} = \lambda \int_0^\infty |H(\omega)|^2\ \frac{d\omega}{2\pi} = \\ = \lambda \int_0^\infty \frac{(\omega\tau_1)^2}{(1+\omega^2\tau_1^2)(1+\omega^2\tau_2^2)}\ \frac{d\omega}{2\pi}$$

and after a few calculation we can rewrite the argument of this integral as:

$$\overline{V_o^2} = \frac{\lambda}{2\pi}\frac{\tau_1^2}{\tau_2^2-\tau_1^2}\int_0^\infty \left(\frac{1}{1+\omega^2\tau_1^2} - \frac{1}{1+\omega^2\tau_2^2}\right)\ d\omega.$$

Since from basic calculus we know that:

$$\frac{d}{dx}\left(\arctan(x)\right) = \frac{1}{1+x^2}$$

we can calculate this integral to be equal to:

$$\overline{V_o^2} = \frac{\lambda}{2\pi}\cdot\frac{\tau_1^2}{\tau_2^2-\tau_1^2}\cdot\left(\frac{1}{\tau_1}\arctan(\omega\tau_1)\Big|_0^\infty - \frac{1}{\tau_2}\arctan(\omega\tau_2)\Big|_0^\infty\right) = \\ = \frac{\lambda}{2\pi}\frac{\tau_1^2}{\tau_2^2-\tau_1^2}\left(\frac{\pi}{2\tau_1} - \frac{\pi}{2\tau_2}\right) = \frac{\lambda}{4}\frac{\tau_1}{\tau_2(\tau_1+\tau_2)}.$$

Under the assumption of having a filter with a large band:

$$f_2 \gg f_1 \ \rightarrow\ \tau_1 \gg \tau_2$$

we can write that:

$$\frac{\tau_1}{\tau_2(\tau_1+\tau_2)} = \frac{1}{\tau_2\left(1+\frac{\tau_2}{\tau_1}\right)} \simeq \frac{1}{\tau_2}\left(1-\frac{\tau_2}{\tau_1}\right)$$

and this gives:

$$\overline{V_o^2} \simeq \frac{\lambda}{4}\left(\frac{1}{\tau_2} - \frac{1}{\tau_1}\right) = \lambda\frac{\pi}{2}(f_2-f_1)$$

where we have considered that:

$$f_1 = \frac{1}{2\pi\tau_1}, \quad f_2 = \frac{1}{2\pi\tau_2}.$$

Note that since the power spectral density of the noise is constant, if we had an ideal filtering in the band defined we would expect to have a mean square value of the noise that is equal to $\lambda(f_2-f_1)$. However, this filter is not a perfect window, with an infinitely abrupt cut-off, therefore we need to take into account the additional contribution of the ending portions of our band-pass filter. These parts are responsible for the $\pi/2$ factor.
On the other hand, in the case of flicker noise[18], the mean square value of the

[18] In the angular frequency domain, the associated power spectral density will be:

$$S_n = \frac{K}{f} = \frac{2\pi K}{\omega}.$$

output voltage can be written as:

$$\overline{V_o^2} = \int_0^\infty S_o(\omega)\,\frac{d\omega}{2\pi} = \int_0^\infty \frac{2\pi K}{f}|H(\omega)|^2\,\frac{d\omega}{2\pi} = \\ = \int_0^\infty \frac{2\pi K}{\omega}\frac{(\omega\tau_1)^2}{(1+\omega^2\tau_1^2)(1+\omega^2\tau_2^2)}\,\frac{d\omega}{2\pi}$$

and after a few calculation we can rewrite the argument of this integral as:

$$\overline{V_o^2} = \frac{K}{2}\frac{\tau_1^2}{\tau_1^2-\tau_2^2}\int_0^\infty \left(\frac{2\tau_1^2\omega}{1+\omega^2\tau_1^2} - \frac{2\tau_2^2\omega}{1+\omega^2\tau_2^2}\right)\,d\omega.$$

Noting that the following equivalence holds regardless of the value of τ that we are considering:

$$\frac{d}{d\omega}\left[\ln\left(1+\omega^2\tau^2\right)\right] = \frac{2\omega\tau}{1+\omega^2\tau^2}$$

exploiting the properties of the logarithm we can write this integral as:

$$\overline{V_o^2} = K\frac{\tau_1^2}{\tau_1^2-\tau_2^2}\left[\ln\left(\sqrt{\frac{1+\omega^2\tau_1^2}{1+\omega^2\tau_2^2}}\right)\Bigg|_0^\infty\right] = K\frac{\tau_1^2}{\tau_1^2-\tau_2^2}\ln\left(\frac{\tau_1}{\tau_2}\right)$$

where the square root in the logarithm comes from the $1/2$ factor that we had in front of the integral. If the bandwidth of the filter is quite wide:

$$\tau_2 \ll \tau_1 \;\Rightarrow\; \frac{\tau_1^2}{\tau_1^2-\tau_2^2} \simeq \frac{\tau_1^2}{\tau_1^2} \simeq 1$$

and this gives the expected proportionality[19]:

$$\overline{V_o^2} \simeq K\ln\left(\frac{f_2}{f_1}\right).$$

Now, we want to assess the effect of a negative feedback on the performances with respect to the noise, that can be measured either using the noise factor F or the signal-to-noise ratio S/N. To do this, we can consider the following simple example: the inverting amplifier that is represented in Figure 3.37.

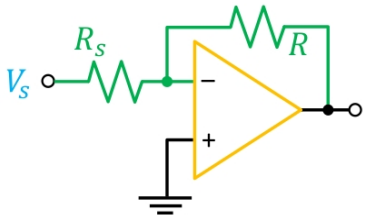


Figure 3.37: Inverting amplifier considered.

First of all, we have to calculate the noise factor F in an open-loop case. The noise sources of this network, therefore, will be the thermal noise in the

[19] In fact, since the power spectral density of the flicker noise is proportional to $1/f$, we expected to have for the mean square value of this noise a logarithmic behaviour with respect to the frequency, if the filtering where perfectly abrupt.

resistances R and R_S (that we will represent with the equivalent current sources) and the current and the voltage noise that are present in the amplifier. Cutting the feedback at its output and connecting it to the ground, we can obtain the network that is represented in Figure 3.38. In particular, noting that all the current sources will be in parallel one with the other, we can sum them, obtaining a single noise equivalent current generator whose power spectral density will be:

$$S_{I_n} = \frac{4k_BT}{R_S} + \frac{4k_BT}{R} + S_I.$$

This last equivalence comes from the fact that the variances of the noise terms, that are related to the power spectral densities, sum up when we are dealing with uncorrelated processes.

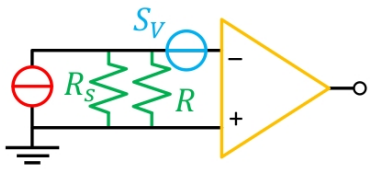


Figure 3.38: Inverting amplifier in an open-loop configuration with the equivalent noise sources.

Directly analysing this circuit, then, we can write the output power spectral density in an open-loop configuration as:

$$S_{V_o,OL} = [S_{I_n}(R_S \| R)^2 + S_V] \cdot |A(s)|^2$$

while if we consider only the output noise that is related to the noise in the source of the signal, thus being only the thermal noise in the source resistance R_S that generates a current (through the noise equivalent current generator) that will flow through both R_S and R:

$$S_{V_o,OL,source} = \frac{4k_BT}{R_S}(R_S \| R)^2 \cdot |A(s)|^2.$$

From the definition of noise factor, then, we can write it as:

$$F_{OL} = 1 + \frac{S_{V_o,OL}}{S_{V_o,OL,source}} = 1 + \left(\frac{4k_BT}{R} + S_I + \frac{S_V}{(R_S \| R)^2}\right) \cdot \frac{R_S}{4k_BT}.$$

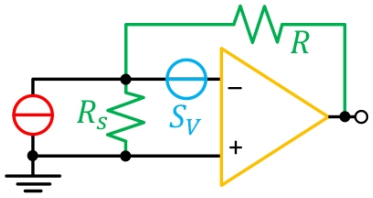


Figure 3.39: Inverting amplifier in a closed-loop configuration with the equivalent noise sources.

The closed-loop configuration of the same circuit is represented in Figure 3.39. From the analysis of the behaviour of this circuit, since we know that the effect of the loop will be to divide the obtained quantities by $|1 - G_{loop}|^2$, we can write the power spectral density on the output as:

$$S_{V_o,CL} = \frac{S_{V_o,OL}}{|1 - G_{loop}|^2}$$

while considering only the noise terms associated to the source of the signal it gives:

$$S_{V_o,CL,source} = \frac{S_{V_o,OL,source}}{|1 - G_{loop}|^2}.$$

Again, from the definition of the noise factor, we can write:

$$F_{CL} = 1 + \frac{S_{V_o,CL}}{S_{V_o,CL,source}}$$

but since the two $|1 - G_{loop}|^2$ terms will cancel out, we can write:

$$F_{CL} = 1 + \frac{S_{V_o,OL}}{S_{V_o,OL,source}} = F_{OL}$$

and therefore the feedback will not affect the performances in terms of noise with respect to the noise factor.
Assuming the flicker noise to be negligible[20], we can try to compute the signal-to-noise ratio in the open-loop case and in the closed-loop one. In this case, assuming the input signal V_S to be a step signal, we know that the behaviour of the operation amplifier will be described by its gain $A(s)$. This will give rise, due to the presence of a pole in the operation amplifier, an exponential behaviour toward the steady-state value of the output. To calculate the signal-to-noise ratio, then, since we know that we have to take the maximum value of the output, we will write it as the steady-state output voltage, that from the final value theorem is equivalent to the DC gain of the operation amplifier:

$$V_{o,OL} = A_0 \frac{R}{R + R_S} V_S$$

Starting from the open-loop power spectral density that we have described before $S_{V_o,OL}$, then, we can write the mean square value of the output voltage as:

$$\overline{V_o^2} = \int_0^\infty S_{V_o,OL}\, df = \left(S_{I_n}(R_S \| R)^2 + S_V\right) A_0^2 \int_0^\infty \frac{1}{1 + \omega^2 \tau_{OL}^2} \frac{d\omega}{2\pi} =$$
$$= \left(S_{I_n}(R_S \| R)^2 + S_V\right) \frac{A_0^2}{4\tau_{OL}}$$

thus giving the following root mean square value of the output voltage:

$$\sqrt{\overline{V_o^2}} = \sqrt{(S_{I_n}(R_S \| R)^2 + S_V) \cdot \frac{A_0^2}{4\tau_{OL}}}$$

[20] This is required in order to make the following computation easier.

thus obtaining the following open-loop signal-to-noise ratio:

$$\left(\frac{S}{N}\right)_{OL} = \frac{V_S \frac{R}{R+R_S} 2\sqrt{\tau_{OL}}}{\sqrt{S_{I_n}(R_S \| R)^2 + S_V}} = K \cdot \sqrt{\tau_{OL}}.$$

Note that, in all the previous derivation, we have defined as τ_{OL} the time constant of the open-loop pole of the operation amplifier.
However, as we have seen in Chapter 1.11.1 at page 68, closing the loop we are modifying the position of this pole and, therefore, we can write the new time constant of the closed-loop pole as τ_{CL} (since closing the loop we are lowering the gain and enlarging the bandwidth). From the same theory, the output voltage in closed-loop can be written as the output voltage in open-loop conditions divided by the steady-state variation associated to the loop gain:

$$V_{o,CL} = \frac{V_{o,OL}}{1 - G_{loop}(0)}$$

while the root mean square value of the output voltage can be written as:

$$\sqrt{\overline{V_o^2}} = \sqrt{\frac{S_{V_o,OL}(0)}{|1 - G_{loop}(0)|^2} \cdot \frac{1}{4\tau_{CL}}}$$

from which, after a few calculations[21], we can derive the following closed-loop signal-to-noise ratio:

$$\left(\frac{S}{N}\right)_{CL} = K\sqrt{\tau_{CL}}.$$

We can thus observe that the closed-loop signal-to-noise ratio can be seen as the following product between the open-loop signal-to-noise ratio and another term:

$$\left(\frac{S}{N}\right)_{CL} = \left(\frac{S}{N}\right)_{OL} \cdot \sqrt{\frac{\tau_{CL}}{\tau_{OL}}}$$

that again from the theory at page 68:

$$\tau_{CL} = \frac{\tau_{OL}}{1 - G_{loop}(0)}$$

can be written as:

$$\left(\frac{S}{N}\right)_{CL} = \left(\frac{S}{N}\right)_{OL} \cdot \frac{1}{\sqrt{1 - G_{loop}(0)}}.$$

Since the loop gain G_{loop} is in general very large, we can observe that:

$$\left(\frac{S}{N}\right)_{CL} < \left(\frac{S}{N}\right)_{OL}$$

and therefore the presence of a negative feedback loop will worsen the signal-to-noise ratio.
Summing up, we have demonstrated that the presence of a negative feedback

[21] These calculations are long and they have not been discussed during lectures; they can be found as an appendix to the slides given by the teacher (L12).

will modify all the open-loop transfers of the same quantity, thus not affecting the noise factor F. However, it will also widen the bandwidth of the system, thus making it collect more noise and degrading the signal-to-noise ratio S/N. In reality, however, the bandwidth of a certain circuit is generally set independently from the feedback, since we have many other requirements that we have to satisfy. This means that this change in the signal-to-noise ratio is misleading, since the limiting factor will be the bandwidth of the signal, thus being the only parameter that matters for the calculation of the signal-to-noise ratio. At the end, this means that the presence of a negative feedback will not modify significantly the performances in terms of noise.

Chapter 4

Signal recovery

4.1 Introduction

After having studied sensors, amplifiers and the problems connected to the noise, we can try to deal with the problem of signal recovering and signal conditioning. In general, the building blocks that are responsible for these operations are placed after the amplifier, thus being the last part of the data acquisition system that we have to analyse.
The problem, in this case, is that the signal coming from an amplifier is, in general, not acceptable, with a very low signal-to-noise ratio. This means that we need to be able to clean the signal from this noise increasing the signal-to-noise ratio and this operation is generally called signal recovery (or conditioning). Dealing with this stage of the data acquisition chain, we will discuss different techniques, but in the time and in the frequency domain, that will be more or less useful depending on the type of signal and on the type of noise considered. We will first study how to deal with high-frequency noise (basically, white noise) both in low-frequency signals and, then, in high-frequency signals. Then, we will move to the study of the low-frequency noise (in this case, mainly the flicker noise) and again we will first see the case of low-frequencies signals (that we will approximate as constants or slowly variable) and then to the high-frequency ones (that will be considered as pulses).
The first element that we can analyse, therefore, is the so called low-pass filter (LPF), that will allow us to reduce the bandwidth of the noise just after the bandwidth of the signal.

4.2 White noise

4.2.1 Low-pass filter

The first example of filter that we can study is the so called low-pass filter (LPF), that is represented in Figure 4.1. From a temporal perspective, this is a linear and time-invariant filter, therefore we can define its delta-function response. From our previous knowledge of these networks, we know that the delta-function response will be a decaying exponential with a time constant:

$$T = RC$$

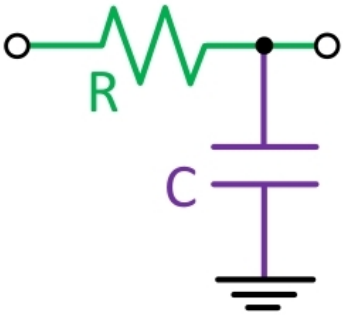


Figure 4.1: A low-pass filter.

that depends on the elements used in the circuit, thus giving:

$$y(t) = x(t) * h(t) = \int h(\tau)h(t-\tau)\ d\tau$$

where[1]:

$$h(t) = \frac{1}{T}e^{-\frac{t}{T}}u(t).$$

Alternatively, we could have considered that since the delta-function is the time derivative of the step function, then the delta-function response will be the derivative of the step response. In the frequency domain, the associated transfer function will be:

$$H(s) = \mathcal{L}\{h(t)\} = \frac{1}{1+sT}.$$

We can now further study the behaviour of this filter in response to a signal in the time domain. The simplest low-frequency signal that we can study is the step:

$$x(t) = Au(t)$$

where A is the amplitude of this step. Since in the time domain the output of the network will be the convolution of the input signal with the delta-function response of the filter, we can write:

$$y(t) = x(t) * h(t) = \int_{-\infty}^{t} x(\tau)h(t-\tau)\ d\tau = \int_{0}^{t} \frac{A}{t}e^{-\frac{t-\tau}{T}}\ d\tau = A\left(1-e^{-\frac{t}{T}}\right)$$

where the integration is extended only on the interval in which both signals are different from zero. We can thus identify the function $h(t-\tau)$ as the weighting function of the filter, since it will weight differently, on the output at a certain time t, all the contributions coming from the delta-functions at previous time instants τ. This means that our low-pass filter has a sort of exponential memory, weighting more the portion of the signal that is quite close to the time instant considered on the output.

[1]Remember that $u(t)$ is the step function, that is needed since this step response is well defined only for $t > 0$.

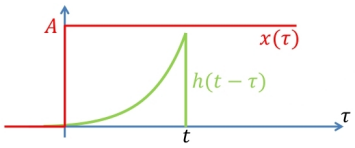


Figure 4.2: Weighting function and step input for a low-pass filter.

Calculating then the autocorrelation of the delta-function response of the filter, from its definition:

$$k_{hh}(\tau) = \int_0^\infty h(t)h(t+\tau)\ dt = \frac{1}{T^2}\int_0^\infty e^{-\frac{t}{T}}u(t)e^{-\frac{t+\tau}{T}}u(t+\tau)\ dt = \\ = \frac{e^{-\frac{\tau}{T}}}{T^2}\int_0^\infty e^{-\frac{2t}{T}}\ dt = \frac{e^{-\frac{\tau}{T}}}{2T}$$

and since this function must be an even function as a consequence of the fact that it does not matter which one of the two delta-function responses is coming first, we can write:

$$k_{hh}(\tau) = \frac{e^{-\frac{|\tau|}{T}}}{2T}.$$

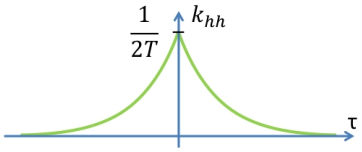


Figure 4.3: Autocorrelation of the delta-function response of a low-pass filter.

We can now consider to have, at the input of our filter, a white noise signal; then, from its definition:

$$R_{xx}(\tau) = \lambda\delta(\tau).$$

The autocorrelation of the output noise, from the previous calculation of the autocorrelation of the delta-function response of the filter, can be written as:

$$R_{yy}(\tau) = R_{xx}(\tau) * k_{hh}(\tau) = \lambda k_{hh}(\tau) = \frac{\lambda}{2T}e^{-\frac{|\tau|}{T}}.$$

From this definition, we can write the mean square value of the noise at the output of the filter as:

$$\overline{n_y^2} = R_{yy}(0) = \lambda k_{hh}(0) = \frac{\lambda}{2T} \propto \frac{1}{T}.$$

This means that the noise power is reduced when the time constant T of the network is increased. In fact, in this case we are averaging the noise over a longer time interval, thus getting closer to the true average value of the noise, that is expected to be zero.
In the frequency domain, in the case of an input signal, we can consider a step signal at the input:

$$X(s) = \frac{A}{s}$$

thus obtaining, at the output:

$$Y(s) = H(s)X(s) = \frac{1}{1+sT}\frac{A}{s} = A\left(\frac{1}{s} - \frac{T}{1+sT}\right).$$

Transforming back from the Laplace domain, this gives:

$$y(t) = A\left(1 - e^{-\frac{t}{T}}\right)u(t)$$

and it is immediately possible to notice that this is exactly the same result that we obtained from a time domain perspective.
Considering now the noise in the frequency domain, we know that the output power spectral density of the noise can be related to the input power spectral density of the noise through the square modulus of the transfer function of the filter. Since we assume to have a white input noise:

$$S_x(\omega) = \lambda$$

then we get:

$$S_y(\omega) = S_x(\omega)\left|H(\omega)\right|^2 = \lambda|H(\omega)|^2$$

and since the mean square value of the output noise is the integral over the whole frequency axis of the power spectral density of the noise, we can write:

$$\overline{n_y^2} = \int_{-\infty}^{+\infty} S_y(\omega)\,\frac{d\omega}{2\pi} = \frac{\lambda}{2\pi}\int_{-\infty}^{+\infty}\frac{d\omega}{1+(\omega T)^2} = \frac{\lambda}{2\pi T}\arctan(\omega T)\Big|_{-\infty}^{+\infty} = \frac{\lambda}{2T}$$

and we obtain, once again, exactly the same result that we have derived in a time domain perspective. It is particularly important to notice that, from the expression of the transfer function of the filter, it will be decreasing with the frequency (above the frequency of the pole), therefore the output noise will not be white but it will have a certain, finite bandwidth, since the power spectral density goes to zero in the high-frequency limit. In the case of a signal, therefore, the maximum allowed bandwidth is defined by the frequency of the pole, thus giving:

$$BW_s = \frac{1}{2\pi T}$$

but what is the bandwidth of the noise? Using an equivalent rectangle approximation, we can assume the power spectral density of the noise to be constant and equal to its maximum value $S_y(0) = \lambda$ over the whole bandwidth. Since this equivalent rectangle must have the same integral of the original power spectral density (in fact, the mean square value of the noise, that is the integral of these quantities, must be the same when calculated with both methods), that is a known quantity, we can impose:

$$\lambda 2 \cdot BW_n = \frac{\lambda}{2T}$$

thus obtaining the following bandwidth for the noise:

$$BW_n = \frac{1}{4T} = \frac{\pi}{2}BW_s.$$

Note that, therefore, the bandwidth seen from the noise is larger than the bandwidth of the signal and this is a consequence of the fact that the filter is not abruptly decreasing after the pole. In fact, it shows an $1/f$ behaviour, thus giving a region in which the signal is attenuated (thus not being useful for it) but in which we are still collecting some noise. From the previous calculations, this region has been demonstrated to have an area equal to $\pi/2$. Increasing the time T and moving to the left the position of the pole we are narrowing the bandwidth over which the noise is collected.

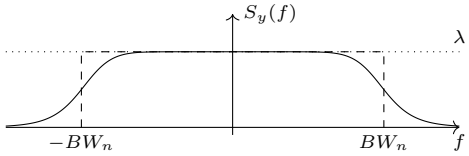


Figure 4.4: Power spectral density of the output noise and equivalent rectangle approximation.

We can thus try to evaluate the improvement in the signal-to-noise ratio that comes from this filter. A problem, however, arises. If we consider a truly white noise at the input of our device, then the mean square value of this input noise will tend to infinite and the signal-to-noise ratio at the input of the filter will be identically equal to zero:

$$\overline{n_x^2} \to \infty \quad \Rightarrow \quad \left(\frac{S}{N}\right)_x \to 0.$$

We need thus to define a quasi-white noise[2], this means a noise that is white only on an equivalent bandwidth that is defined as[3]:

$$f_n = \frac{1}{2T_n}, \quad T_n \ll T, \; f_n \gg f_p = \frac{1}{2\pi T}.$$

In this case, assuming V_i to be the maximum value of the signal at the input (and consequently also at the output) of the device, since the mean square value of the input noise will be:

$$\overline{n_x^2} = \int_{-\infty}^{+\infty} S_x(f)\, df = \int_{-f_n}^{+f_n} \lambda \, df = 2\lambda f_n$$

we can write the signal-to-noise ratio at the input of the device as:

$$\left(\frac{S}{N}\right)_x = \frac{V_i}{\sqrt{\overline{n_x^2}}} = \frac{V_i}{\sqrt{2\lambda f_n}} = V_i\sqrt{\frac{T_n}{\lambda}}$$

where in the last equality we used the expression that links the bandwidth of the input noise with the associated time constant. At the output, the signal

[2]To understand these requirements, consider the triangular approximation of the white noise that has been previously discussed.

[3]Note that in this way the input noise power spectral density is flat over the frequency range of interest.

will have the same maximum amplitude V_i (since the gain of the filter is one) while the mean square value of the noise has been shown, from the previous calculations, to be equal to $\lambda/2T$, thus giving the following signal-to-noise ratio:

$$\left(\frac{S}{N}\right)_y = \frac{V_i}{\sqrt{n_y^2}} = \frac{V_i}{\sqrt{\frac{\lambda}{2T}}} = V_i\sqrt{\frac{2T}{\lambda}} \propto \sqrt{T}.$$

From this last expression we can immediately see that the higher is the time constant T of the network the more we are reducing the noise and thus we are increasing the signal-to-noise ratio at the output of the filter.
The improvement in the signal-to-noise ratio can be directly evaluated by relating these two quantities:

$$\left(\frac{S}{N}\right)_y = \left(\frac{S}{N}\right)_x \sqrt{\frac{2T}{T_n}} = \left(\frac{S}{N}\right)_x \sqrt{\frac{f_n}{BW_n}} > \left(\frac{S}{N}\right)_x.$$

This factor, therefore, will quantify the improvement in the signal-to-noise ratio that is connected to the use of this low-pass filter. Moreover, considering the $\sqrt{2T/T_n}$ factor it makes clear that we need to apply a filter whose time constant T is larger than the correlation time of the noise. In fact, with this filter we are averaging the noise and, if this signal is still correlated during our average, we will obtain a non-zero result, while if we are integrating over a long enough time it will be an uncorrelated signal, thus averaging at zero. In a frequency perspective, the term $\sqrt{f_n/BW_n}$ gives the fact that the bandwidth of the noise after the filter must be smaller (actually, much smaller) than the bandwidth of the quasi-white noise in order to have an improvement in the signal-to-noise ratio.

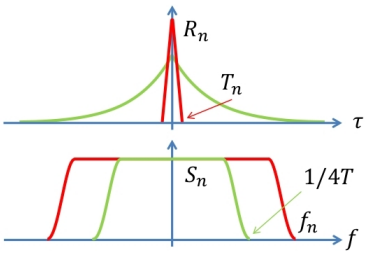


Figure 4.5: Autocorrelation of the filter and of the quasi-white noise (above) and power spectral density of the noise before and after the filter (below).

4.2.2 Time-variant filter

A different and more complicated possibility is represented by time-variant filters, in which the delta-function response is dependent from the time. In this case, given an input signal $x(t)$, the output signal will be the integral of the input with a system response function $w(t, \tau)$:

$$y(t) = \int_{-\infty}^{+\infty} x(\tau)w(t,\tau)\ d\tau$$

where every contribution for $t < \tau$ will be identically equal to zero. In the special case of a delta-function input:

$$x(\tau) = \delta(\tau - \tau_0) \;\rightarrow\; y(t) = \int_{-\infty}^{+\infty} \delta(\tau - \tau_0) w(t, \tau)\, d\tau = w(t, \tau_0)$$

and we can see that the function $w(t, \tau)$ is still the system response at time t to a delta function that is applied at the time τ, but it is not the delta-function response shifted and reversed as in the case of time-invariant filters:

$$w(t, \tau) \neq h(t - \tau)$$

since this quantity is separately a function of t and τ.

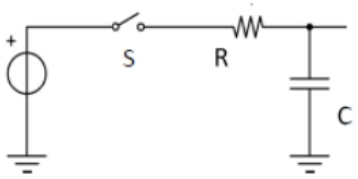


Figure 4.6: An example of time-variant filter.

As an example we can consider the time-variant filter that is represented in Figure 4.6. In this case, the dependence on the time comes from the presence of a switch S: when the switch is closed, the circuit is completely analogous to a low-pass filter, thus having a usual exponential delta-function response, while when the switch is open the value of the output will be set by the voltage drop across the capacitor. We can thus study the delta-function response of this filter depending on when the delta-function is imposed at the input of the filter. If the delta-function is imposed when the switch is closed and if the switch remains closed for the whole duration of the transient, then the network will exhibit the usual exponential decay delta-function response, as in the case of a low-pass filter. If the delta-function is imposed at the input when the switch is open, since we do not have any voltage across the capacitor the output will stay a zero for every time instant, regardless of the fact that in future time instants the switch is open or closed.

A more complicated situation is the one represented in Figure 4.7, in which a delta-function arrives at the input when the switch is closed, starting on the output an exponential decay. Suddenly, the switch opens; this freezes the output at its current voltage value, since the capacitor is charged, and the output voltage will remain constant (at least ideally) for the whole time interval in which the switch is open. When the switch closes again, the exponential decay behaviour will start again from where it stopped. This makes clearer that the delta-function response depends separately on the time instant t considered and on the arrival time τ (or α in the Figure) of the delta-function. Fixing therefore a certain time instant t_m and considering the contribution, at that time instant, of the various exponential decays for delta-functions imposed at different initial times τ, we can obtain the behaviour of the weighting function as in Figure 4.8.

Note that, as expected, any delta-function arriving at the input when the switch is open will give a zero contribution at time t_m, from which the region in which the weighting function is identically equal to zero. Moreover, for

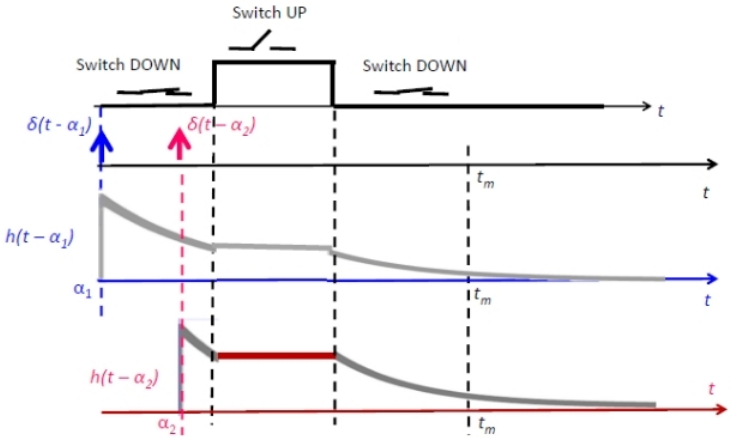


Figure 4.7: Delta-function response of a time-variant low-pass filter.

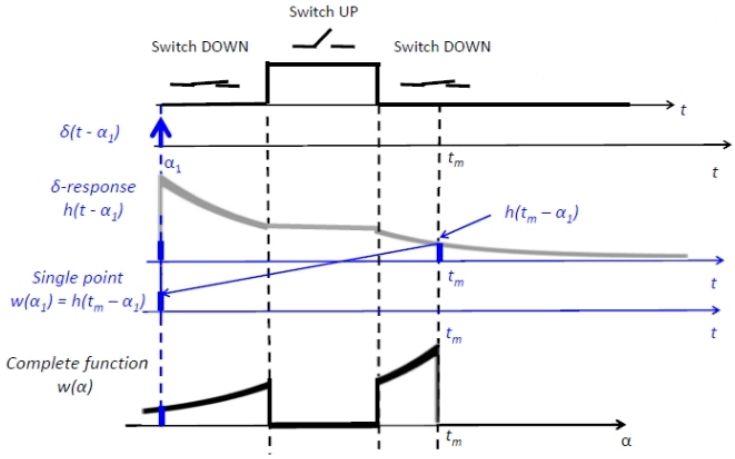


Figure 4.8: Weighting function of a time-variant low-pass filter.

delta-functions arriving before the close-to-open switching event, since when the switch is opened the output voltage is constant, the exponential behaviour of the weighting function will restart exactly from the same value after the region in which it is equal to zero. From another perspective, we have only added a region in which the weighting function is equal to zero in between the exponential weighting function that we expect for a low-pass filter. Another possible representation of the weighting function for a different switching event is represented in Figure 4.9, thus making clearer that this weighting function depends both on the arrival time of the delta-function considered and on the time considered for studying the output.

From a more numerical point of view, we can now evaluate the response to the noise of this filter. Since the filter is time-variant, the output noise will not be stationary and we will have an explicit dependence of the autocorrelation of the output noise on the time instants t_1 and t_2 used for evaluating the output.

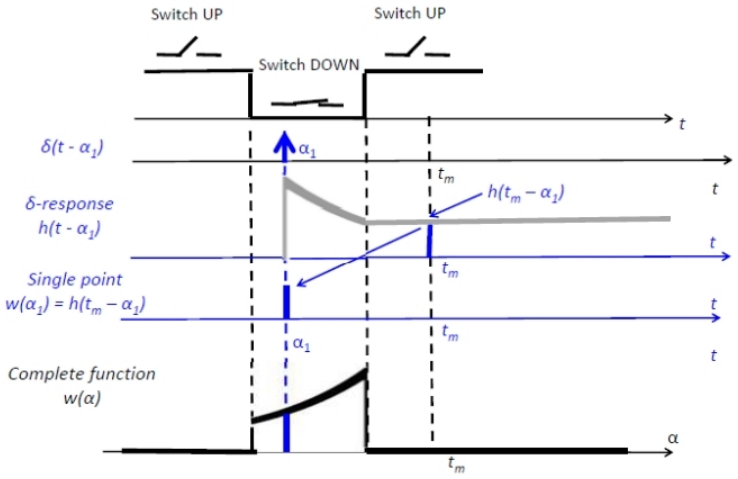


Figure 4.9: Weighting function of a time-variant low-pass filter for a different switching event.

From the definition of the autocorrelation of the output noise, we can write:

$$R_{yy}(t_1,t_2) = \overline{y(t_1)y(t_2)} = \overline{\iint x(\alpha)w(t_1,\alpha)x(\beta)w(t_2,\beta)\ d\alpha\ d\beta} =$$
$$= \iint \overline{x(\alpha)x(\beta)}w(t_1,\alpha)w(t_2,\beta)\ d\alpha\ d\beta =$$
$$= \iint R_{xx}(\alpha,\beta)w(t_1,\alpha)w(t_2,\beta)\ d\alpha\ d\beta$$

where we considered that the ensemble average of a time integral is the time integral of the ensemble averaged quantities, that the weighting function is, actually, a deterministic quantity and that the autocorrelation of the input noise is defined as:

$$R_{xx}(\alpha,\beta) = \overline{x(\alpha)x(\beta)}.$$

In this case, the mean square value of the output noise will be equal to the autocorrelation of the noise when the two signals are evaluated in the same time instant

$$\overline{n_y^2}(t_1) = R_{yy}(t_1,t_1).$$

Note that, since the noise is non-stationary, the mean square value of the output noise depends on time.
Assuming now a stationary input noise:

$$\overline{x(\alpha)x(\beta)} = R_{xx}(\alpha,\beta) = R_{xx}(\alpha-\beta)$$

and thus we can rewrite the autocorrelation of the output noise as:

$$R_{yy}(t_1,t_2) = \overline{y(t_1)y(t_2)} = \iint R_{xx}(\beta-\alpha)w(t_1,\alpha)w(t_2,\beta)\ d\alpha\ d\beta$$

where it is important to note that even if the input noise is stationary, the filter is time-variant and therefore the output noise will be non-stationary. Defining

the following variable:

$$\gamma = \beta - \alpha, \quad d\gamma = d\beta$$

we can write:

$$\begin{aligned} R_{yy}(t_1, t_2) &= \iint R_{xx}(\gamma) w(t_1, \alpha) w(t_2, \alpha + \gamma)\ d\alpha\ d\gamma = \\ &= \int R_{xx}(\gamma)\ d\gamma \int w(t_1, \alpha) w(t_2, \alpha + \gamma)\ d\alpha = \\ &= \int R_{xx}(\gamma) k_{w_{12}}(\gamma)\ d\gamma \end{aligned}$$

where we have defined:

$$k_{w_{12}}(\gamma) = \int w(t_1, \alpha) w(t_2, \alpha + \gamma)\ d\alpha.$$

In fact, if we could forget the fact that we are evaluating the weighting function in two different time instants t_1 and t_2, this would be the time correlation of the weighting function, since we are integrating the product of this function with its temporally shifted replica. However, this correlation does not depend only on the reciprocal shift γ but also on the time instants t_1 and t_2 and this is the reason of the subscript 12.
Once we have calculate the autocorrelation of the output noise, it is easy to evaluate the mean square value of the output noise in the case of a stationary input noise:

$$t_1 = t_2 = t \ \rightarrow\ \overline{n_y^2}(t) = R_{yy}(t, t) = \int R_{xx}(\gamma) k_{w_{tt}}(\gamma)\ d\gamma$$

where we have, from the previous definition, the following time correlation of the weighting function:

$$k_{w_{tt}}(\gamma) = \int w(t, \alpha) w(t, \alpha + \gamma)\ d\alpha.$$

Again, this immediately shows that the output noise is non-stationary, even though the input noise is stationary, due to the time-variant nature of the filter.
To study now the behaviour of a signal in the frequency domain, we can apply the Parseval's theorem, thus obtaining:

$$y(t) = \int x(\tau) w(t, \tau)\ d\tau = \int X(f) W^*(t, f)\ df$$

where we have defined:

$$W(t, f) = \int w(t, \tau) e^{-j2\pi f \tau}\ d\tau$$

the Fourier transform of the weighting function. However, the output signal in the frequency domain:

$$Y(f) = \mathcal{F}\{y(t)\}$$

cannot be defined in a simple way in the general case: it can be done only in a few very specific cases.

Assuming to have a stationary input noise, again from the Parseval's theorem we can write the mean square value of the output noise as:

$$\overline{n_y^2}(t) = \int R_{xx}(\gamma) k_{w_{tt}}(\gamma)\, d\gamma = \int S_x(f)|W(t,f)|^2\, df$$

where the power spectral density of the input noise is the Fourier transform of its autocorrelation and where we have defined $W(t,f)$ as:

$$S_x(f) = \mathcal{F}\{R_{xx}(\gamma)\}, \quad |W(t,f)|^2 = \mathcal{F}\{k_{w_{tt}}(\gamma)\}.$$

Note that this result is quite similar to the one that we have obtained in the case of time-invariant filters, except for the fact that $W(t,f)$ is now a time-dependent quantity.
In the case of a white stationary noise at the input, by definition its autocorrelation can be written as:

$$R_{xx}(\gamma) = \lambda\delta(\gamma)$$

therefore the mean square value of the output noise in a time-domain perspective will be:

$$\overline{n_y^2}(t) = \int R_{xx}(\gamma) k_{w_{tt}}(\gamma)\, d\gamma = \lambda \int \delta(\gamma) k_{w_{tt}}(\gamma)\, d\gamma = \lambda k_{w_{tt}}(0) =$$
$$= \lambda \int w^2(t,\alpha)\, d\alpha.$$

In a frequency-domain perspective, on the other hand, since the input white noise is defined as:

$$S_x(f) = \lambda$$

we obtain:

$$\overline{n_y^2}(t) = \int S_x(f)|W(t,f)|^2\, df = \lambda \int |W(t,f)|^2\, df$$

where we can immediately observe that these two integrals are equal one to the other as a consequence of the Parseval's theorem.

4.3 Gated integrators and improvement of S/N

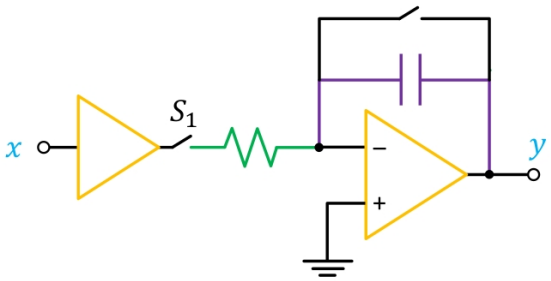


Figure 4.10: A gated integrator.

A particular type of time-varying filter is the gated integrator, that integrates the input signal for a finite time interval. This time interval, also called the integration window, is determined by the commuting of two switches, one that is indicated with S_1 in Figure 4.10, the other that is in parallel to the capacitor and that we will call S_2. When S_1 is open and S_2 is closed, the output voltage is identically equal to zero and we are not integrating the input. When we commute both the switches, thus making S_1 closed and S_2 open, we start the integration of the input signal. Another commutation of the two switches will determine the end of the time window in which are integrating. Since the two switches are always commuting together (when one is open, the other is closed and vice versa), we can study the behaviour of the circuit only depending on the position of the switch S_1.

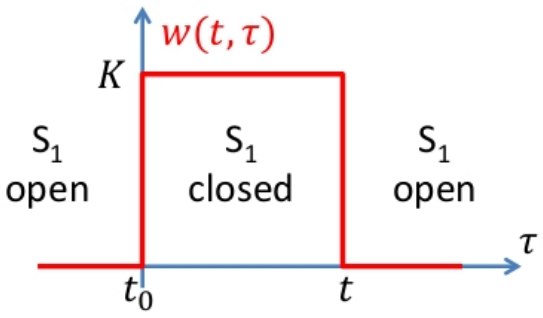


Figure 4.11: Weighting function of a gated integrator.

Since we know that the weighting function $w(t, \tau)$ is the system response at time t to a delta-function applied at time τ, we can consider that the weighting function will be identically equal to zero outside the integration window, when S_1 is open. When S_1, therefore when we are in the gate, the output will be the integral of the input and, since the integral of a delta function is a step function, we will obtain the same output regardless of the arrival time of the delta function and therefore the weighting function in this interval will be constant[4] and equal to the gain of the network K. Considering for example that the initial time of the gate is:

$$t_0 = 0$$

for the sake of simplicity, this weighting function can be written as the difference between two step functions centred in different time instants:

$$w(t, \tau) = K \cdot [u(\tau) - u(\tau - t)] .$$

Changing the duration of the gate t, we are thus stretching the rectangle that is represented in Figure 4.11.
To evaluate the signal response of this filter, we can write:

$$y(t) = \int x(\tau) w(t, \tau)\, d\tau = K \int_0^t x(\tau)\, d\tau = Kt\langle x \rangle$$

[4]It will be the sum of various steps with different initial part.

where we explicitly took into account the expression of the weighting function and the definition of time average of a signal:

$$\langle x \rangle = \frac{1}{t} \int_0^t x(\tau)\ d\tau.$$

Defining the Fourier transform of the weighting function we have previously defined, we can see that:

$$W(t,f) = \mathcal{F}\{w(t,\tau)\} = Kt\,\mathrm{sinc}(\pi f t) e^{-j\pi f t}$$

where we know that the Fourier transform of a rectangle is a sinc function whose argument is πf and where the exponential term takes into account that the rectangle we are considering is centred in $t/2$, having an additional phase term:

$$e^{-j2\pi f \frac{t}{2}} = e^{-j\pi f t}.$$

In particular, it is important to consider that we always have to be sure, when calculating the Fourier transforms, that the following theorem holds:

$$W(t,0) = Kt = \int w(t,\tau)\ d\tau$$

thus checking the zero frequency value. Therefore, from a frequency point of view the output signal could have been written as:

$$y(t) = \int X^*(f) W(t,f)\ df$$

but since:

$$X^*(f) = X(-f)$$

we obtain:

$$y(t) = \int X(-f) W(t,f)\ df = Kt \int X(-f)\,\mathrm{sinc}(\pi f t) e^{-j\pi f t}\ df.$$

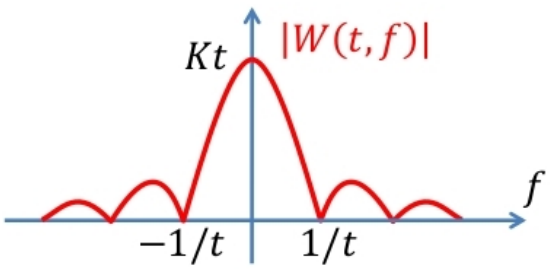


Figure 4.12: Behaviour of the filter in the frequency domain.

Considering the behaviour of the filter in the frequency domain that is represented in Figure 4.12, we can immediately note that we cannot use the following expression:

$$Y(f) \neq X(f)|W(t,f)|$$

that will hold only for linear and time-invariant filters. We can thus calculate the autocorrelation of the weighting function as:

$$k_{w_{tt}}(\tau) = \int w(t,\alpha)w(t,\alpha+\tau)\ d\alpha$$

and the square modulus of the Fourier transform of the weighting function:

$$|W(t,f)|^2 = |\mathcal{F}\{w(t,\alpha)\}|^2 = \left|Kt\,\mathrm{sinc}(\pi f t)e^{-j\pi f t}\right|^2 = K^2t^2\,\mathrm{sinc}^2(\pi f t)$$

and represent them as in Figure 4.13.

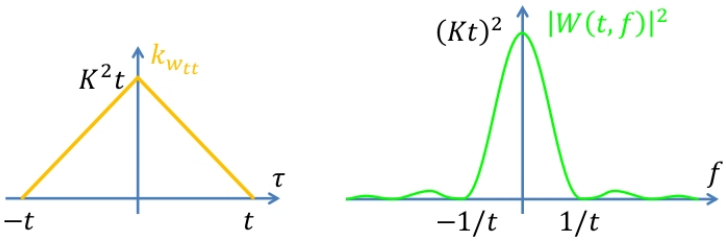


Figure 4.13: Autocorrelation of the weighting function and square modulus of its Fourier transform.

To study now the output noise of the device, from a time domain perspective we can write its means square value, under the assumption of having a white input noise:

$$R_{xx} = \lambda\delta(\gamma)$$

as:

$$\overline{n_y^2} = \int R_{xx}(\gamma)k_{w_{tt}}(\gamma)\ d\gamma = \lambda k_{w_{tt}}(0) = \lambda t K^2.$$

In the frequency domain, since from the definition of white noise:

$$S_x(f) = \lambda$$

we can write[5]:

$$\overline{n_y^2} = \int S_x(f)|W(t,f)|^2\ df = \lambda K^2t^2\int \mathrm{sinc}^2(\pi f t)\ df = \lambda t K^2$$

consistently with the Parseval's theorem.
Also in this case, considering an equivalent rectangle approximation of the power spectral density of the output noise, since it must preserve that area of the power spectral density, that we have just calculated as the mean square value of the output noise, we can write:

$$\lambda K^2t^2 \cdot 2BW_n = \lambda t K^2 \ \rightarrow\ BW_n = \frac{1}{2t}$$

[5]It is important to remember the following notable integral:

$$\int \mathrm{sinc}^2(\pi f t)\ df = \frac{1}{t}.$$

and this is the noise equivalent bandwidth of the gated integrator. Also in this case, therefore, as the integration time t increases, we are narrowing this filter. To compute the input and output signal-to-noise ratio, as in the previous case, we can approximate the input noise as a quasi-input noise with a triangular approximation of its autocorrelation on a time $T_n \ll t$, while the input noise power spectral density will be constant and equal to λ on a bandwidth $[-f_n, f_n]$ that is defined from the following relation:

$$2f_nT_n = 1 \;\rightarrow\; 2f_n = \frac{1}{T_n} \;\rightarrow\; f_n = \frac{1}{2T_n}.$$

In this case, from the result we have obtained in the previous sections, we can write the input signal-to-noise ratio, assuming V_i to be the maximum value of the input voltage, as:

$$\left(\frac{S}{N}\right)_x = \frac{V_i}{\sqrt{\overline{n_x^2}}} = \frac{V_i}{\sqrt{2\lambda f_n}} = V_i\sqrt{\frac{T_n}{\lambda}}.$$

At the output of the filter, the output signal will be the integral of the input one:

$$V_y(t) = \int x(\tau)k(t,\tau)\; d\tau = KtV_i$$

and thus it can be written as the gain multiplied by the maximum value of the input voltage and by the width of the gate. Considering also the expression of the mean square value of the output noise that we have just derived, the signal-to-noise ratio at the output can be written as:

$$\left(\frac{S}{N}\right)_y = \frac{V_y}{\sqrt{\overline{n_y^2}}} = \frac{V_iKt}{\sqrt{\lambda tK^2}} = V_i\sqrt{\frac{t}{\lambda}}.$$

Rewriting the output signal to noise ratio as a function of the input one, we can obtain:

$$\left(\frac{S}{N}\right)_y = \left(\frac{S}{N}\right)_x \sqrt{\frac{t}{T_n}} = \left(\frac{S}{N}\right)_x \sqrt{\frac{2t}{2T_n}} = \left(\frac{S}{N}\right)_x \sqrt{\frac{f_n}{BW_n}}.$$

This relationship makes clear that we have to integrate over gates whose duration t is much larger than the correlation time T_n of the input noise:

$$t \gg T_n$$

in order to have an average of the noise that is zero, since we are integrating a totally uncorrelated signal. From the second expression, on the other hand, we are obtaining that the bandwidth BW_n of the output noise signal must be much smaller than the bandwidth of the input noise.

We can now compare the bandwidth of a gated integrator with the one of a low-pass filter. From the expression of a low-pass filter, its transfer function can be written as:

$$\frac{1}{1 + (2\pi fT)^2}$$

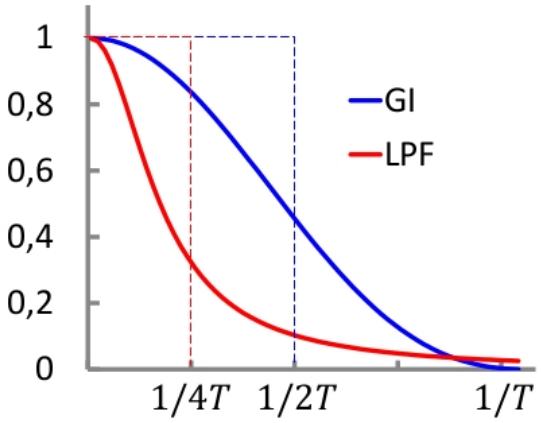


Figure 4.14: Comparison of the bandwidth of a gated integrator with a low-pass filter.

while for a gated integrator we have just shown it to be:

$$(Kt)^2 \operatorname{sinc}^2(\pi f t).$$

To be able to compare the two signal-to-noise ratio, we want the two filters to have the same low-frequency gain with respect to a signal, thus making them different only with respect to the noise, and this can be done by imposing:

$$Kt = 1.$$

Assuming then, for example:

$$t = T$$

we can study the equivalent rectangle approximation of these filters that is represented in Figure 4.14. Assuming a white input noise, then the mean square of the output noise will be proportional (through a suitable coefficient that is related to the power spectral density of the input noise) to the integral of these curves. We can thus observe that the gated integral is more noisy, as it can be seen from the equivalent rectangle approximation. The equivalent output noise bandwidth for these two filters therefore will be:

$$BW_n(GI) = \frac{1}{2T}, \quad BW_n(LPF) = \frac{1}{4T}$$

and therefore, to achieve the same signal-to-noise ratio, we have to choose the two integration times as it follows:

$$t = 2T.$$

In general, a low-pass filter is always present in an acquisition system to cut all the unwanted high-frequency noise components, that are in regions in which we will not have any signal. Gated integrators, on the contrary, are more specific for certain applications and they can be used in particular when we are dealing with fast signals such as pulses. In fact, short pulses will have, from the Fourier theorem, a large bandwidth and therefore a low-pass filter will suppress also

high-frequency components of the signal. A suitably triggered gated integrator, in this case, is able to select only the time window in which we have the pulse, thus significantly reducing the amount of noise that is collected after the filter. Moreover, the gated integrators will present several zeros in the Fourier transform of the weighting function, thus being able of completely suppressing certain well-defined frequency components on the output:

$$f_n = \frac{n}{t}.$$

This kind of filters, therefore, is used also to reject power supply disturbs or interferences[6].

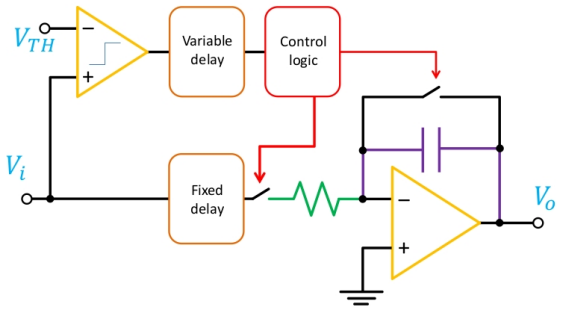


Figure 4.15: A simplified scheme of a gated integrator.

A simplified version of a gated integrator is represented in Figure 4.15. In particular, we have to add another circuit in order to trigger the window in which we are integrating. The signal V_i, therefore, is coming to a discriminator with a certain threshold V_{th} and, when the input signal is above the threshold, a certain signal arrives through a variable delay to a control logic, that starts the integration. However, the discriminator and the control logic introduce a certain delay on the trigger signal, therefore, we need to add the same delay on the input line. The problem, now, is that the delay introduced from the control network is very difficult to calculate and, moreover, it may change depending on various parameters. We thus add a big (surely bigger than the delay introduced by the control network) and fixed delay on the input line and a variable delay on the control network. By manually adjusting the variable delay on the control network, is the possible to match the fixed delay on the input line and to obtain a perfect synchronization of the input signal with the gate.

An example of input and output waveforms from a gated integrator is represented in Figure 4.16. In this case, the time interval of the gate T_G can be chosen in order to optimize the signal-to-noise ratio on the measurement we are performing. For pulses, therefore, gated integrators are better, since they can suppress the noise at any other time instant while collecting the pulse without modifying it.

Several models of gated integrators are generally at our disposal. Typical gate

[6]Interferences, in fact, are due to other systems in which we might have switchings, couplings and other effects. These effects are generally well localized in frequency, therefore if they are strong enough they can be eliminated by properly choosing the integration time, as it is often done in portable tools.

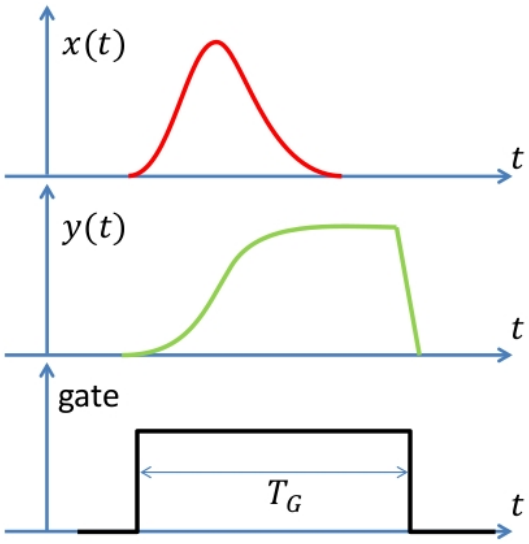


Figure 4.16: Input and output waveforms from a gated integrator.

widths (that corresponds to the integration time) can go from one or two nanoseconds to several microseconds (thus allowing the measurement of pretty short pulses), while the gain, that gives an amplification of the signal, typically ranges between one and 1000. Between two integration gates, then, a certain time interval is required for resetting all the offsets, discharging the capacitors and making the system ready for another measurement. This time interval is generally called dead time and it is typically about a few microseconds. Many other parameters, such as the linearity and the offset, might then be discussed in the data-sheets of these devices.

4.4 Boxcar averagers and ratemeters

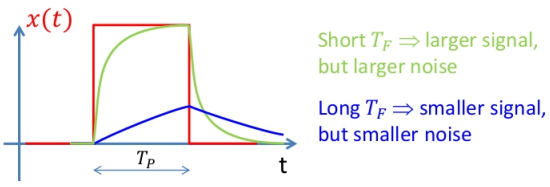


Figure 4.17: The output of a gated integrator in presence of a fast pulse and a white noise.

At this point, we can study the example of a system that have to filter a fast pulse in presence of white noise. Assuming to have a gated integrator, depending on the bandwidth of the filter we have to choose between the complete recovery of the signal or the complete reduction of the noise. In fact, if the time constant of the filter T_F is much smaller than the duration T_P of the pulse, the signal

will be larger, but also the noise measured will be larger:

$$T_F \ll T_P \;\rightarrow\; BW_n = \frac{1}{T_F} \;\rightarrow\; \overline{n_y^2} = \frac{1}{4T_F}.$$

On the other hand, if we choose a filter whose time constant T_F is much bigger than the duration of the pulse T_P, then the signal will be smaller but the noise contribution will be smaller as well. In general, we can write the signal-to-noise ratio in this case as:

$$\left(\frac{S}{N}\right)_{out} = A\frac{1-e^{-\frac{T_P}{T_F}}}{\sqrt{\frac{\lambda}{4T_F}}}.$$

The question therefore is: can we retain the advantage of a long time constant of the filter T_F without sacrificing the signal?

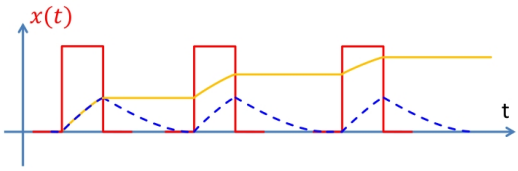


Figure 4.18: Repetitive pulse.

A particularly interesting situation, in this case, is the one in which we have a series of repetitive pulses. This situation is typical of many physical signals and it does not necessarily imply that the signal is periodic. Considering the use of a low-pass filter, there are two main situations that are possible. In the first case, that is the dashed one in Figure 4.18, the capacitor of the filter is completely discharged between one measurement and the other. We are thus performing a repetitive series of measurements as the one we have seen in the previous case. On the other hand, if the capacitor is not discharged after each pulse, as in the solid line in Figure, we can gain signal. In fact, at each different gate, we are integrating and thus summing the contribution coming from various signals, while always having the same, constant noise contribution. This allows us to obtain an higher value of the signal with respect to the noise and it can be done, in a gated integrator, by assuming that through the whole measurement the switch S_2, that is in parallel to the capacitor, is always open, thus never discharging it.

The weighting function of a gated integrator can thus be derived as in Figure 4.19. In particular, a Dirac delta coming during the period T_C in which the first switch is closed (and thus we are integrating) will give a constant contribution[7], while every delta function coming when the first switch is open (a period of time that we call T_O) will not give any effect, thus making that region of the weighting function equal to zero. This gives us a weighting function $w(t,\tau)$ that is similar to a square wave. Performing a suitable change of variables, from τ to τ', it is possible to define a new weighting function that will consist in the sequence of all the pieces of the previous one that are different from zero, resulting in a new weighting function that is constant but whose duration, instead of being a series of N rectangles of width T_C separated by intervals of width T_O, is just a

[7]The integral of a Dirac delta is a step.

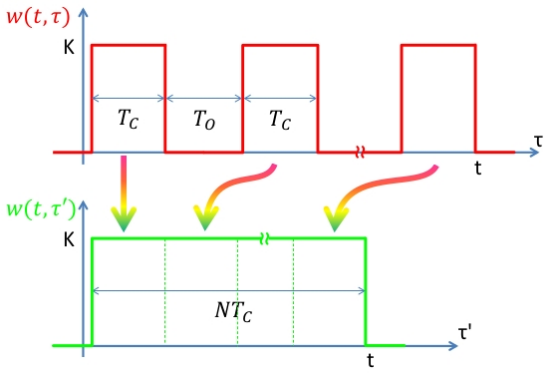


Figure 4.19: Weighting function of a gated integrator.

single rectangle of width NT_C. It is important to note that these two weighting functions are completely equivalent from the viewpoint of a white noise[8], since:

$$\overline{n_y^2} = \lambda k_{w_{tt}}(0) = \lambda \int w^2(t,\tau)\, d\tau = \lambda \int w'^2(t,\tau')\, d\tau' = \lambda K N T_C.$$

This system, therefore, behaves, from the viewpoint of the noise, as a gated integrator with an integration time that is equal to:

$$T_G = NT_C$$

therefore the signal-to-noise ratio can be written as:

$$\left(\frac{S}{N}\right)_{out} = \left(\frac{S}{N}\right)_{in} \sqrt{\frac{T_G}{T_n}} = \left(\frac{S}{N}\right)_{sp} \sqrt{N}$$

where T_n is the correlation time of the quasi-white noise at the input and we can define the signal-to-noise ratio coming from the measurement of a single pulse as:

$$\left(\frac{S}{N}\right)_{sp} = \left(\frac{S}{N}\right)_{in} \sqrt{\frac{T_C}{T_n}}.$$

Averaging over N samples, therefore, the signal-to-noise ratio improves of a factor $\sqrt{N}$ when the white noise is dominant over every other noise source. This new idea involves, however, the possibility of storing the information about the previous samples, thus extending the average over multiple pulses, and it leads to the development of the boxcar averager.

A simple schematic of a boxcar averager is represented in Figure 4.20. It is slightly different from a gated integrator, since it is a low-pass filter that is repetitively switched using a suitably triggered and controlled switch S_1. This circuit, as we will see, gives an exponential average rather than a uniform one.

Again, when the switch is closed this circuit perfectly mimics a low-pass filter, therefore the delta function response is a piece of a decaying exponential. When the switch is open, on the other hand, the contribution of a delta function to the output is identically equal to zero. This behaviour is represented in Figure

[8] And also from the viewpoint of the signal.

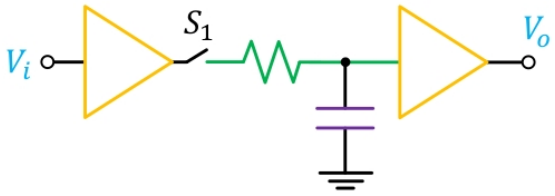


Figure 4.20: Schematic of a boxcar averager.

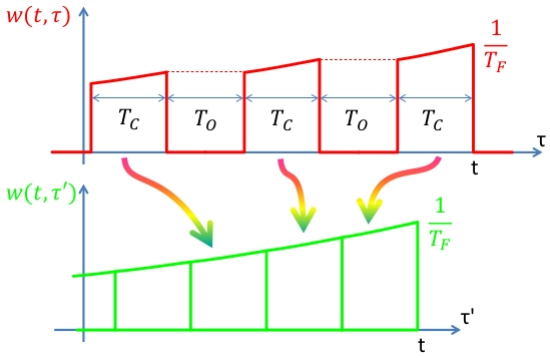


Figure 4.21: Weighting function for a boxcar averager.

4.21. Again, from the viewpoint of a white noise and of the signal, this weighting function is completely equivalent to the one we can obtain, through a suitable change of variables, by adding all the non-zero parts of the original weighting function one after the other. We have thus obtained a new weighting function $w(t, \tau')$ that is actually a true decaying exponential behaviour. This means that the expression of the signal-to-noise ratio is similar to the one of a low-pass filter.
In this new "equivalent time" τ' representation, the weighting function can be written as:

$$w(t, \tau') = \frac{1}{T_F} e^{-\frac{t-\tau'}{T_F}} u(t - \tau')$$

and thus, for a white noise, we can write the output mean square value of the noise as:

$$\overline{n_y^2} = \lambda \int w^2(t, \tau)\, d\tau = \lambda \int w^2(t, \tau')\, d\tau'$$

therefore the signal-to-noise ratio is exactly equal to the one of a low-pass filter. Note that, in this case, we have the benefits of a long integration time T_F with respect to the noise without loosing any signal, since we are accumulating over a large number of examples. The signal-to-noise ratio can thus be written as:

$$\left(\frac{S}{N}\right)_{out} = \left(\frac{S}{N}\right)_{in} \sqrt{\frac{2T_F}{T_n}}$$

and therefore the signal-to-noise ratio of the boxcar averager is identical to the one of the low-pass filter. Moreover, the output will depend on the input only when the switch is closed, while the time that is needed to reach the steady-state

condition depends on the temporal duration of the interval in which the switch is closed T_C and on the duration of the interval in which the switch is open T_O (therefore, on the sampling rate), but the signal-to-noise ratio is independent from it.
For a gated integrator, then, we have seen that the improvement in the signal-to-noise ratio is proportional to $\sqrt{N}$, where N is the number of samples we are integrating. In the case of a boxcar averager, what is the meaning of N? Since the weighting function is not constant, this is not a trivial consideration.
First of all, we can consider a single-pulse boxcar averager. Working on a single pulse, we can assume the associated weighting function to be a portion of exponential that starts at a certain value at time $t - T_C$ and reaches the value $1/T_F$ at time t, while being identically equal to zero outside from this interval. Assuming that:

$$T_C \ll T_F$$

we can neglect this exponential behaviour that is related to the discharge of the capacitor over the time interval T_C; this is a consequence of the fact that the time interval in which the switch is closed is much shorter than the time constant of the capacitor. We can thus approximate the single-pulse weighting function with a rectangle of height $1/T_F$. In this way, we are approximating the single-pulse behaviour of the boxcar averager in this condition with the one of a gated integrator (for which the weighting function will always be a rectangle) with the following gain:

$$K = \frac{1}{T_F}.$$

In this way, if we have the following mean square value of the noise and the following output noise that are calculated in this rectangular approximation:

$$\overline{n_y^2} = \lambda \frac{T_C}{T_F^2}, \quad y = A\frac{T_C}{T_F}$$

for a constant input signal with amplitude A and for a white (or quasi-white) noise with input power spectral density λ, we can write the signal-to-noise ratio for the single-pulse as:

$$\left(\frac{S}{N}\right)_{sp} = A\frac{T_C}{T_F}\sqrt{\frac{T_F^2}{\lambda T_C}} = A\sqrt{\frac{T_C}{\lambda}} = \left(\frac{S}{N}\right)_{in}\sqrt{\frac{T_C}{T_n}}$$

where we have remembered that the input signal-to-noise ratio is:

$$\left(\frac{S}{N}\right)_{in} = A\sqrt{\frac{T_n}{\lambda}}.$$

From the expression of the signal-to-noise ratio for the boxcar averager that we have previously determined (and the we called with the subscript *out*), substituting the expression of the signal-to-noise ratio for the single-pulse we can determine:

$$\left(\frac{S}{N}\right)_{BA} = \left(\frac{S}{N}\right)_{sp}\sqrt{\frac{2T_F}{T_C}} = \left(\frac{S}{N}\right)_{sp}\sqrt{N_{eq}}$$

where we have defined the following equivalent number of samples, in analogy to the case of the gated integrator, as:

$$N_{eq} = \frac{2T_F}{T_C}.$$

This means that N_{eq} represents the improvement of the signal-to-noise ratio due to the exponential average. As we have seen in the previous case, the single-pulse boxcar is usually approximated with a gated integrator and the boxcar averager will give an improvement if and only if we have a repetitive signal, thus dealing with many different pulses. In this way, we are thus considering our exponentially weighted average as a uniform average over an equivalent number of pulses under the approximation:

$$T_F \gg T_C$$

otherwise we need to explicitly consider the exponential discharge of the capacitor.
Removing this hypothesis, in the general case, in fact, the weighting function of a single-pulse boxcar can be written as:

$$w(t,\tau) = \frac{1}{T_F} e^{-\frac{t-\tau}{T_F}}, \quad t - T_C \leq \tau \leq t.$$

Assuming to have a constant input signal of amplitude A, then, the associated output can be written as:

$$y = \int x(\tau) w(t,\tau)\, d\tau = A \int_{t-T_C}^{t} \frac{1}{T_F} e^{-\frac{t-\tau}{T_F}}\, d\tau = A\left(1 - e^{-\frac{T_C}{T_F}}\right).$$

For the noise term, assuming a white stationary input noise with a bilateral power spectral density equal to λ, we can write the mean square value of the output noise as:

$$\begin{aligned}\overline{n_y^2} = \lambda k_{w_{tt}}(0) = \lambda \int w^2(t,\tau)\, d\tau = \frac{\lambda}{T_F^2} \int_{t-T_C}^{t} e^{-\frac{2(t-\tau)}{T_F}}\, d\tau = \\ = \frac{\lambda}{T_F^2} \int_0^{T_C} e^{-\frac{2\gamma}{T_F}}\, d\gamma = \frac{\lambda}{2T_F}\left(1 - e^{-\frac{2T_C}{T_F}}\right)\end{aligned}$$

where we have exploited the following change of variables:

$$\gamma = t - \tau, \quad d\tau = -d\gamma$$

inverting the extremes of the integration to absorb the minus sign. From these considerations, we can write the signal-to-noise ratio for a single pulse as:

$$\begin{aligned}\left(\frac{S}{N}\right)_{sp} = \frac{A\left(1 - e^{-\frac{T_C}{T_F}}\right)}{\sqrt{\frac{\lambda}{2T_F}\left(1 - e^{-\frac{2T_C}{T_F}}\right)}} = A\sqrt{\frac{2T_F}{\lambda}} \cdot \frac{1 - e^{-\frac{T_C}{T_F}}}{\sqrt{1 - e^{-\frac{2T_C}{T_F}}}} = \\ = \left(\frac{S}{N}\right)_{BA} \frac{1 - e^{-\frac{T_C}{T_F}}}{\sqrt{1 - e^{-\frac{2T_C}{T_F}}}}\end{aligned}$$

where we have recognized that:

$$\left(\frac{S}{N}\right)_{BA} = A\sqrt{\frac{2T_F}{\lambda}}$$

is the signal-to-noise ratio for a boxcar averager in the condition:

$$t \to \infty$$

thus being equal to the one of a low-pass filter. Since by definition the following relation holds:

$$\left(\frac{S}{N}\right)_{BA} = \left(\frac{S}{N}\right)_{sp} \cdot \sqrt{N_{eq}}$$

we can write the equivalent number of samples, in the general case, as:

$$N_{eq} = \frac{1 - e^{-\frac{2T_C}{T_F}}}{\left(1 - e^{-\frac{T_C}{T_F}}\right)^2} = \frac{\left(1 + e^{-\frac{T_C}{T_F}}\right)\left(1 - e^{-\frac{T_C}{T_F}}\right)}{\left(1 - e^{-\frac{T_C}{T_F}}\right)^2} = \frac{1 + e^{-\frac{T_C}{T_F}}}{1 - e^{-\frac{T_C}{T_F}}}.$$

From this result, under the previous hypothesis:

$$T_C \ll T_F: \quad 1 + e^{-\frac{T_C}{T_F}} \simeq 2, \quad 1 - e^{-\frac{T_C}{T_F}} \simeq 1 - 1 + \frac{T_C}{T_F} + \dots$$

we can retrieve the previous result:

$$N_{eq} \simeq \frac{2}{1 - 1 + \frac{T_C}{T_F}} = \frac{2T_F}{T_C}.$$

We can now ask to ourselves: what does it happen when the input noise is not white? In this condition, we have to discuss the noise correlation. From the expression of the mean square value of the output noise of a signal:

$$\overline{n_y^2} = \int R_{xx}(\gamma) k_{w_{tt}}(\gamma)\, d\gamma$$

in the case of a white noise we know that its correlation can be written as:

$$R_{xx}(\gamma) = \delta(\gamma)$$

thus being sampling the temporal autocorrelation of the weighting function, while in the case of a non-white noise it will explicitly depend on the autocorrelation of the input noise. Moreover, in this case we have to consider the true expression of the weighting function of the filter and not, as in the previous case, its equivalent continuous representation. Assuming to have a periodic behaviour of the weighting function of the filter, from the definition of the temporal autocorrelation of the weighting function:

$$k_{w_{tt}}(\gamma) = \int w(t,\tau) w(t,\tau+\gamma)\, d\tau$$

we can observe that for $\gamma = 0$ we are integrating a term that is equal to $w^2(t,\tau)$, thus giving the same factor that we would have obtained from a low-pass filter.

For different shifts γ, then, we will obtain a smaller and smaller superposition of the two shifted replicas of the weighting function until, for $\gamma = T_C$, this superposition becomes identically equal to zero. If we then shift the two replicas of $\gamma = T_C + T_O$ then the first pulse of a replica will be overlapped to the second one of the other, the second of one replica to the third of the other and so on, giving another peak in the temporal autocorrelation, and so on and so forth. Note that these contributions will be smaller and smaller when we are overlapping distant pulses, at the end resulting in a train of spikes at:

$$m \cdot (T_C + T_O), \quad m \in \mathbb{N}$$

enveloped in an exponential with the following time constant:

$$T_F \frac{T_C + T_O}{T_C}.$$

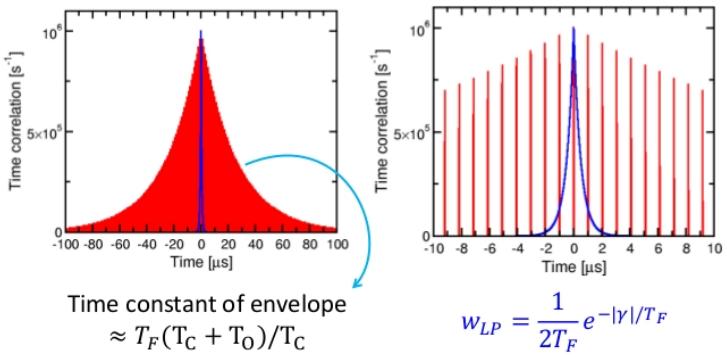


Figure 4.22: Correlation function for a boxcar averager if $T_C = 20$ ns, $T_O = 1$ μs and $T_F = 0.5$ μs. In blue it is represented the weighting function of a low pass filter for a comparison.

In Figure 4.22 we have represented the correlation function in two different scales, one for highlighting the presence of the envelope and the other for showing the fact that is given by a series of discrete spikes, and we have compared it with the weighting function of a low-pass filter. In the case of a white noise, since the autocorrelation of the noise is a delta function centred in the origin, we will be considering only the central spike of the boxcar averager and, since it is equal to the value in zero of the correlation function for the low-pass filter, we can immediately observe that if the noise is totally uncorrelated with itself the two filters have the same performances. If the noise is not white and it is self-correlated over a time that is larger than T_C and smaller than T_O:

$$T_C < T_n < T_O$$

then the the boxcar averager is clearly better than the low-pass filter, since we are considering only the central spike of the autocorrelation of the boxcar averager, that will be much narrower than the region of the autocorrelation of the low-pass filter considered. To reduce the noise we have thus to average over a certain number of uncorrelated samples and, summing a certain number of completely uncorrelated pulses, the boxcar averager will give better performances.

If the noise is strongly uncorrelated from one pulse to the other, therefore, the boxcar averager is really effective also when we are dealing with non-white noise.

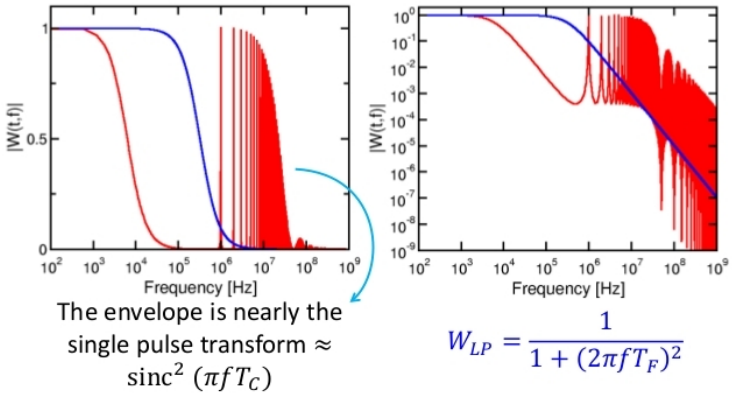


Figure 4.23: Frequency domain representation of the modulus Fourier transform of the weighting function for the same choice of the quantities considered. A comparison with the behaviour of the low-pass filter is possible.

The same comparison but in the frequency domain is represented in Figure 4.23. In this case, the modulus of the Fourier transform of the weighting function for the boxcar averager, that is represented in red, will give a much lower frequency of the pole with respect to the low-pass filter. At high frequencies, however, as it is possible to see both in a linear scale and in a logarithmic one, the boxcar averager gives rise to spikes due to the presence of harmonics, that are frequencies that are sampled in phase for many different pulses. Since the value of the autocorrelation in the origin must be equal to the integral over the whole spectrum of the square of this quantity, then the area of the two curves must be identical. If the power spectral density of the noise is relevant only at high enough frequencies (this means that the correlation time is quite small), then it may be relevant for the low-pass filter but not for the boxcar averager thus giving, in the frequency domain, the same advantages that we have seen in the time domain.
The main typical parameters of a boxcar averager are:

- the gain width T_C, that typically ranges from 1 or 2 ns to 20 or 30 μs and that can be even shorter in fast samplers;
- the equivalent number of samples, that can range from one to several thousands;
- the delay, that is intrinsically between 10 and 15 ns and that typically ranges between 3 and 300 ns, unless custom modifications, but that is not really important;
- the trigger rate, that is the maximum number of pulses that we can process per each second and that is typically lower than 100 kHz.

We can now consider a new type of behaviour for a boxcar averager: the waveform recovery mode. This way of using this device is useful when we have a

series of repetitive pulses for which we want to recover not only their amplitude but also the shape of the pulse, that is its waveform. From an ideal perspective, we could think to sample a pulse with a train of delta functions; however, these delta functions in reality do not exist and they can only be approximated using small rectangles in the time domain. However, since these rectangles are small in the time domain, they will be very large in the frequency domain, thus collecting a lot of noise and, if the noise is white, this will give rise to a terrible signal-to-noise ratio. The solution, in this case, is to average over many different pulse using a boxcar averager. In this situation, the trigger delay is not fixed: it is increment such that it is possible to sweep the different pulses between an initial and a final value. The measurement of each point (that will correspond to a certain value of the trigger delay) will then be repeated for a certain number of pulses (for example, N) in order to improve the signal-to-noise ratio before sending the data to the output. In this way, it is possible to reconstruct the waveform of the signal, in a process that is called equivalent-time sampling.

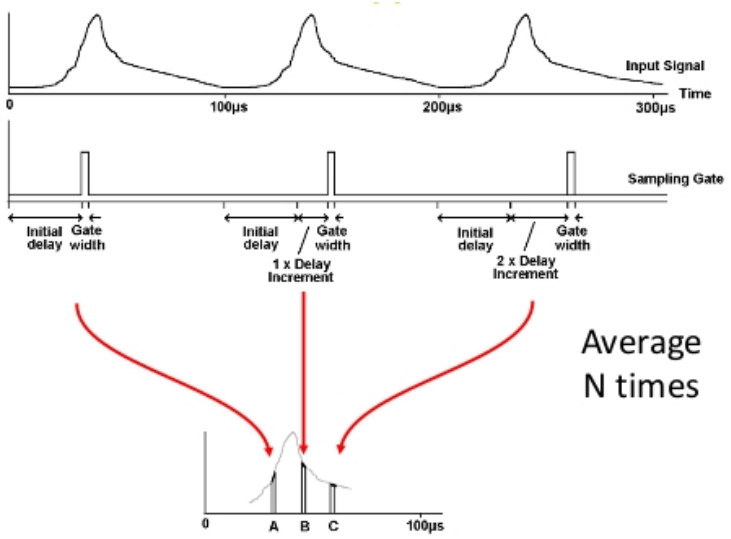


Figure 4.24: Waveform recovery mode.

In Figure 4.24 it is represented this procedure. It is important to remember that we are sampling many different pulses, not only one, for many reasons, one of which the problem of collecting a lot of noise. These different pulses, moreover, will be sampled in different position (in general, we take only one or two samples for each pulse) and we will take the average of the samples taken in the same time instant with respect to the position of the pulse. In this case, a very large number of pulses is required but it will allow a very clear signal reconstruction with a significant reduction of the noise. This kind of sampling is also called equivalent time sampling.

A slightly different device from a boxcar averager is a ratemeter, that is represented in Figure 4.25. In this device, in fact, the function of the switch is different, since it moved in front of the first buffer. This means that the switch S_1 is able to prevent the input signal to reach the capacitor but, on the other hand, it is not able of preventing the discharge of the capacitor itself, since the

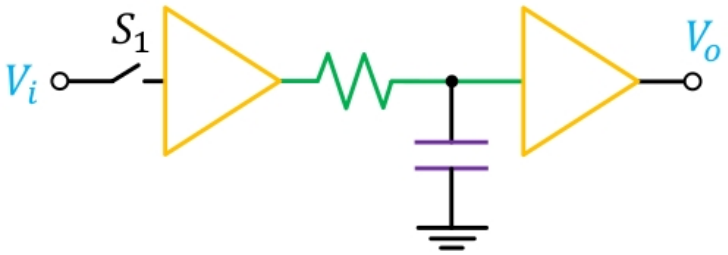


Figure 4.25: A ratemeter.

buffer will connect the capacitor to the ground through the resistor.

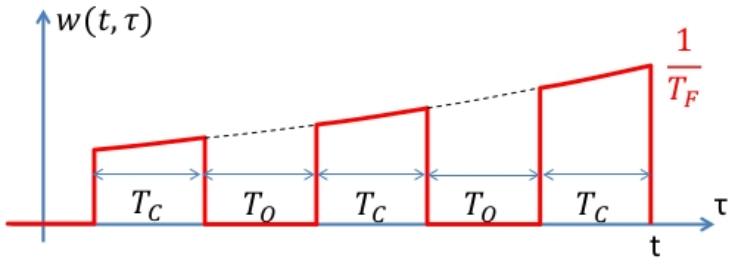


Figure 4.26: Weighting function of a ratemeter.

The weighting function of this device is represented in Figure 4.26 (even though the real behaviour may not be periodic). In this case, as we have previously said, the exponential discharge of the capacitor continues also when the switch is open, thus during the time interval T_O. This is different from the behaviour of the boxcar averager, in which we remained at the same level while the switch was open. We can thus consider the delta function response to be a continuous decreasing exponential sampled by many different rectangles.
Since the weighting function can be seen as the superposition of several different pulses, the contribution of the n-th pulse to the weighting function can be written as:

$$w_n(t,\tau) = \frac{e^{-\frac{t-\tau}{T_F}}}{T_F}, \quad \tau \in [-n(T_C+T_O) - T_C; -n(T_C+T_O)]\,.$$

Assuming then to have a constant signal with amplitude A at the input of the device, the contribution to the output of the n-th pulse, assuming for the sake of simplicity that, for this pulse, $t = 0$, can be written as:

$$y^n = A\int_{-n(T_C+T_O)-T_C}^{-n(T_C+T_O)} w_n(t,\tau)\,d\tau = A\int_{-n(T_C+T_O)-T_C}^{-n(T_C+T_O)} \frac{e^{\frac{\tau}{T_F}}}{T_F}\,d\tau =$$
$$= A\left(1 - e^{-\frac{T_C}{T_F}}\right) e^{-\frac{n(T_C+T_O)}{T_F}}.$$

Summing all these terms, then, the output considering all the different pulses

can be written as:

$$y = \sum_{n=0}^{\infty} y^n = A\left(1 - e^{-\frac{T_C}{T_F}}\right) \sum_{n=0}^{\infty} \left(e^{-\frac{T_C+T_O}{T_F}}\right)^n = A\frac{1 - e^{-\frac{T_C}{T_F}}}{1 - e^{-\frac{T_C+T_O}{T_F}}}$$

where we have recognized the presence, in the sum, of a geometric series.
From the viewpoint of noise, on the other hand, the mean square value of the output noise will be:

$$\overline{n_y^2} = \lambda k_{w_{tt}}(0) = \lambda \int w^2(t,\tau)\, d\tau = \lambda \sum_n \int w_n^2(t,\tau)\, d\tau = \lambda \sum_n k_{w_{tt}}^n(0)$$

where we have defined the contribution of the n-th pulse to the autocorrelation as:

$$k_{w_{tt}}^n(0) = \int_{-n(T_C+T_O)-T_C}^{-n(T_C+T_O)} \frac{e^{2\frac{\tau}{T_F}}}{T_F^2}\, d\tau = \frac{1 - e^{-\frac{2T_C}{T_F}}}{2T_F} e^{-\frac{2n(T_C+T_O)}{T_F}}.$$

Summing all these contributions, we obtain that:

$$k_{w_{tt}}(0) = \sum_{n=0}^{\infty} k_{w_{tt}}^n(0) = \frac{1}{2T_F}\left(1 - e^{-\frac{2T_C}{T_F}}\right) \sum_{n=0}^{\infty} \left(e^{-2\frac{T_C+T_O}{T_F}}\right)^n =$$

$$= \frac{1}{2T_F} \frac{1 - e^{-\frac{2T_C}{T_F}}}{1 - e^{-\frac{2(T_C+T_O)}{T_F}}}$$

where we have recognized the presence of a geometric series, that gives:

$$\overline{n_y^2} = \frac{\lambda}{2T_F} \frac{1 - e^{-\frac{2T_C}{T_F}}}{1 - e^{-\frac{2(T_C+T_O)}{T_F}}}.$$

From these considerations, we can write the signal-to-noise ratio of the ratemeter as:

$$\left(\frac{S}{N}\right)_{out} = A\sqrt{\frac{2T_F}{\lambda}} \frac{1 - e^{-\frac{T_C}{T_F}}}{\sqrt{1 - e^{-\frac{2T_C}{T_F}}}} \cdot \frac{\sqrt{1 - e^{-\frac{2(T_C+T_O)}{T_F}}}}{1 - e^{-\frac{T_C+T_O}{T_F}}}.$$

This expression can be divided into several different terms that allows us to better understand the behaviour of this device. The first term:

$$A\sqrt{\frac{2T_F}{\lambda}}$$

is the signal-to-noise ratio of a boxcar averager or, equivalently, of a low-pass filter, since they are the same from the viewpoint of a white noise. Adding the term:

$$\frac{1 - e^{-\frac{T_C}{T_F}}}{\sqrt{1 - e^{-\frac{2T_C}{T_F}}}}$$

we can obtain the signal-to-noise ratio of a single-pulse boxcar averager and this comes from the fact that we are considering only the last pulse that has come to the device. Last the term

$$\frac{\sqrt{1 - e^{-\frac{2(T_C+T_O)}{T_F}}}}{1 - e^{-\frac{T_C+T_O}{T_F}}}$$

gives the effect of the exponential average we are performing and it will be the square root of the equivalent number of samples. For this device, thus we can define this equivalent number of pulses as:

$$N_{eq} = \frac{1 - e^{-\frac{2(T_C+T_O)}{T_F}}}{\left(1 - e^{-\frac{T_C+T_O}{T_F}}\right)^2} = \frac{\cancel{\left(1 - e^{-\frac{T_C+T_O}{T_F}}\right)}\left(1 + e^{-\frac{T_C+T_O}{T_F}}\right)}{\left(1 - e^{-\frac{T_C+T_O}{T_F}}\right)^{\cancel{2}}} = \frac{1 + e^{-\frac{T_C+T_O}{T_F}}}{1 - e^{-\frac{T_C+T_O}{T_F}}}$$

and if we assume to have the following condition:

$$T_C + T_O \ll T_F$$

then it can be approximated as:

$$N_{eq} \simeq \frac{2}{1 - \left(1 - \frac{T_C+T_O}{T_F}\right)} = \frac{2T_F}{T_C + T_O}.$$

It is important to notice that, for this device, this condition includes also the time interval T_O in which the switch is open, while in the case of a boxcar averager the only relevant time interval was the one in which the switch was closed T_C. This means that the equivalent number of samples for a ratemer is always smaller than the equivalent number of samples of a boxcar averager:

$$\frac{2T_F}{T_C + T_O} < \frac{2T_F}{T_C}$$

and thus the signal-to-noise ratio of a boxcar averager is always smaller than the one of a ratemeter. In the case of the ratemeter, for correlated noise, the temporal autocorrelation of the weighting function will then be similar to the one that we have previously discussed for the boxcar averager, but with a smaller amplitude of the spikes. This means that our system will be collecting a lower fraction of noise but also a lower signal. It is not easy to intuitively determine whether this is an advantage or a disadvantage but, from the calculation of the signal-to-noise ratio that we have just performed, it is clear that this makes the performances of the ratemeter to be worse than the ones of the boxcar averager. This means that we have two possible behaviours for the two networks, depending on the position of a switch. In the boxcar averager, it is only the number of pulses that we are collecting that matters, while for the ratemeter also the time in which they are coming is relevant: it will be compared with the discharge time of the filter. If the number of pulses per second, and thus the rate, is high, we have that the signal-to-noise ratio in the ratemeter will be comparable to the one of the boxcar averager.

In Figure 4.27 the two networks, of a ratemeter and of a boxcar averager, are represented. In both cases, the switch is acting as a gate on the input source, but while in the ratemeter case it is decoupled from the RC circuit through the use of a voltage buffer, in the boxcar one is directly acts on the RC passive filter. In the ratemeter, therefore, the RC passive filter has constant parameters, it is unaffected by the switch and it does not have an hold state, while in the boxcar case the time constant T_F of the integrator filter is switched from a finite value set by the RC circuit when the switch is closed to an infinite value when

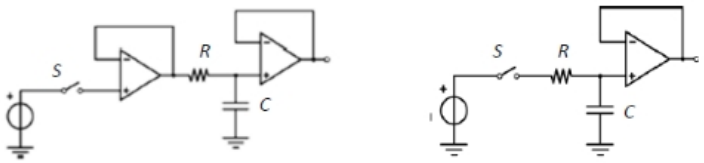


Figure 4.27: Comparison between a ratemeter and a boxcar averager.

the switch is open. In the ratemeter, therefore, the sample average is done on a given time that is defined by the RC value, while in the boxcar case it is done on a certain number of samples that is determined by the ratio between the time constant of the filter and the temporal duration of the gate.

4.5 Discrete-time filters and their representation in the frequency domain

In a discrete time filter, a certain number N of samples of the input signal and of the noise are acquired with a certain sampling time t_S. A suitable weighted average is then performed on the data, obtaining the following output:

$$y = \sum_{k=1}^{N} w_k \cdot (x_k + n_k)$$

where k is an index that is referred to the sample considered, w_k is the weight of this average, x_k is the signal at time k and n_k is the noise at the same time instant.
The simplest kind of discrete-time filter that we can consider is the uniform average. In this case, each one of the samples has the same weight, that will be equal to the inverse of the number of samples:

$$w_k = \frac{1}{N}$$

thus leading to the following output:

$$y = \frac{1}{N} \sum_{k=1}^{N} (x_k + n_k)$$

that is an average of the samples with uniform weights. In this way, we are actually building a discrete-time version of a gated-integrator, as it is represented in Figure 4.28.

Assuming for example a constant signal of amplitude A and a white noise (or a non-correlated[9] stationary noise) at the input, we can write the output as:

$$\bar{y} = \overline{\frac{1}{N} \sum_{k=1}^{N} (A + n_k)} = \frac{1}{N} \sum_{k=1}^{N} (A + \overline{n_k}) = A$$

[9]This means that the correlation time of the noise considered T_n is shorter than the time interval between two consecutive samples:

$$T_n < t_s.$$

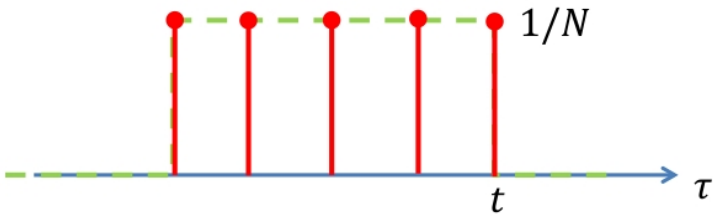


Figure 4.28: The equivalent of a gated-integrator in discrete time the uniform average.

since we have assumed the noise to have a null average, while the mean square value of the output noise can be written as:

$$\overline{n_{out}^2} = \overline{(y-A)^2} = \overline{\frac{1}{N^2}\left(\sum_{k=1}^{N} n_k\right)^2} = \frac{1}{N^2}\sum_{k=1}^{N}\overline{n_k^2} = \frac{1}{N}\overline{n_{in}^2}.$$

Calculating the ratio between these two quantities, we can determine the signal-to-noise ratio of this discrete-time filter as:

$$\left(\frac{S}{N}\right)_{out} = \frac{\bar{y}}{\sqrt{\overline{n_{out}^2}}} = \frac{A}{\sqrt{\overline{n_{in}^2}}}\sqrt{N} = \left(\frac{S}{N}\right)_{in}\sqrt{N}.$$

As expected from the result of the corresponding continuous time filter, the signal-to-noise ratio improves of a factor $\sqrt{N}$, where N is the number of uncorrelated samples that we are considering.
Setting the total time of the measurement, that is equal to N times the interval between two samples, equal to the temporal duration of the gate of a gated integrator:

$$T_M = T_G = N \cdot t_S$$

we can compare the signal-to-noise ratio for a gated integrator, that we have determined in the previous section:

$$\left(\frac{S}{N}\right)_{GI} = \left(\frac{S}{N}\right)_{in}\sqrt{\frac{T_M}{T_n}}$$

to the one of the uniform average:

$$\left(\frac{S}{N}\right)_{AV} = \left(\frac{S}{N}\right)_{in}\sqrt{\frac{T_M}{t_S}}$$

but since we have assumed the noise samples to be uncorrelated, thus meaning that the correlation time of the noise is lower between the temporal distance between two samples:

$$T_n < t_S$$

this implies that the signal-to-noise ratio of a gated integrator is higher than the one of a uniform averager:

$$\left(\frac{S}{N}\right)_{AV} < \left(\frac{S}{N}\right)_{GI}.$$

A different possibility is to consider an ideal sampling procedure. In this case, the output signal will be equal to the input one at a certain set of discrete time instants:

$$y(t) = x(t_S)$$

and thus this sampling operation, from a continuous time perspective, can be seen as the convolution of the input signal with a particular weighting function:

$$y(t) = \int x(\tau)\delta(t_S - \tau)d\tau.$$

This means that the weighting function in the case of an ideal sample is a delta function:

$$w(t,\tau) = \delta(t_S - \tau)$$

centred in the time instant that we want to sample. Considering to be sampling at many different time instants that are integer multiples of the sampling time, we can write this weighting function as:

$$w(t,\tau) = \frac{1}{N}\sum_{k=0}^{N-1}\delta(\tau - (t - kt_S)) = \frac{1}{N}\text{rect}(T_M)\sum_{k=-\infty}^{+\infty}\delta(\tau - kt_S)$$

and it is represented in Figure 4.29. Note that every delta function will be sampling the input signal and then it will be multiplied by $1/N$, thus leading to the correct output. Moreover, observe that we have written this sampling as the product between a rectangular function of width T_M centred in a certain position and an infinite comb of delta functions (also called Dirac comb) in order to have a simpler representation when working in the frequency domain.

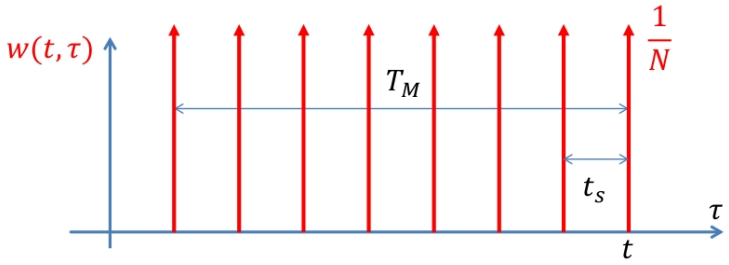


Figure 4.29: Weighting function for an ideal sampler.

In order to deal with the noise, we now have to calculate the temporal autocorrelation of the weighting function with itself:

$$k_{w_{tt}}(\gamma) = \int w(t,\tau)w(t,\tau+\delta)\ d\tau$$

since its value in zero will be useful for determining the mean square value of the output noise. When $\gamma = 0$, all the delta functions of the Dirac comb are overlapped with the corresponding functions of the other comb, thus giving:

$$k_{w_{tt}}(0) = \frac{1}{N^2}\cdot N\cdot\delta(\gamma) = \frac{1}{N}\delta(\gamma).$$

Moving away from this point, for a different value of γ we will not have any overlap between the different combs until they start to overlap again when the shift is equal to t_S. In this case, however, only $N-1$ delta functions are overlapping, thus giving:

$$k_{w_{tt}}(t_S) = \frac{1}{N^2} \cdot (N-1) \cdot \delta(\gamma - t_S).$$

Repeating this reasoning over and over, we can determine the following expression for the autocorrelation for the weighting function:

$$k_{w_{tt}}(\gamma) = \begin{cases} 0, & \gamma \neq nt_S \\ \frac{1}{N^2}(N-n)\delta(\gamma), & \gamma = nt_S \end{cases}, \quad n \in \mathbb{N}$$

and this formula can be rewritten, in order to have a simpler representation when dealing with the frequency domain, as:

$$k_{w_{tt}}(\gamma) = \frac{1}{N}\text{tri}(T_M) \sum_{k=-\infty}^{+\infty} \delta(\gamma - kt_S)$$

that is the product between a unitary amplitude triangular signal and a Dirac comb; this will be the expression of the output noise for this device.

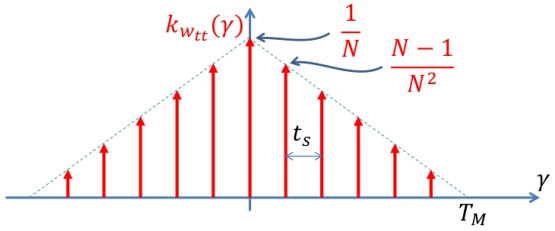


Figure 4.30: Autocorrelation of the weighting function.

Assuming to have a weakly self-correlated input noise with a correlation time T_n that is smaller than the sampling time t_S of the device:

$$T_n < t_S$$

in this case the only relevant part of the autocorrelation of the weighting function for the calculation of the mean square value of the output noise will be the central spike and, therefore, we can write this noise term as:

$$\overline{n_y^2} = \int R_{xx}(\gamma) k_{w_{tt}}(\gamma)\, d\gamma = \frac{\overline{n_x^2}}{N}.$$

Note that this value is equal to the one at the input of the device but reduced of a factor N, thus giving an improvement of $\sqrt{N}$ in the signal-to-noise ratio of the device, exactly as we expected from the previous calculations.

In the frequency domain, we can calculate the Fourier transform of the weighting function, that will be the convolution of the Fourier transform of

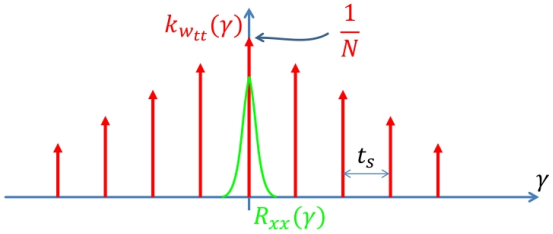


Figure 4.31: Calculation of the mean square value of the output noise.

a rectangle (that is a sinc function[10]) with a series of delta functions (since the Fourier transform of a Dirac comb is a rescaled Dirac comb):

$$\mathcal{F}\left[\mathrm{rect}(T_m)\right] = T_m \,\mathrm{sinc}(\pi f T_m), \quad \mathcal{F}\left[\sum_{k=-\infty}^{+\infty}\right] = \frac{1}{t_S}\sum_{k=-\infty}^{+\infty}\delta\left(f-\frac{k}{t_S}\right).$$

Remembering that the following relationship holds:

$$T_M = N t_S$$

we can write the Fourier transform of the weighting function as:

$$W(t,f) = \frac{T_M}{N}\,\mathrm{sinc}(\pi f T_M) * \frac{1}{t_S}\sum_{k=-\infty}^{+\infty}\delta\left(f-\frac{k}{t_S}\right) =$$
$$= \sum_{k=-\infty}^{+\infty}\mathrm{sinc}\left(\pi T_M\left(f-\frac{k}{t_S}\right)\right)$$

where we have considered that the convolution with a Dirac comb corresponds to a shift of the convolved quantity along the frequency axis. We will thus have a certain number of sinc functions that are centred in integer multiples of $1/t_S$ whose first zero is at $1/T_M$ from the maximum. This Fourier transform is represented in Figure 4.32.

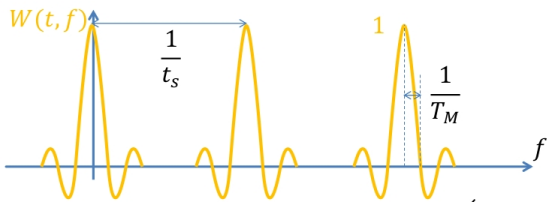


Figure 4.32: Fourier transform of the weighting function.

To calculate the mean square value of the output noise in the frequency domain we can consider the Fourier transform of the autocorrelation of the weighting function, that will correspond to the squared modulus of the Fourier

[10]To calculate the factor in front of it, we have to consider that the value in zero of the Fourier transform will be the integral in the time domain.

transform of the weighting function. Computing this value:

$$\mathcal{F}\left[k_{w_{tt}}(\gamma)\right] = |W(t,f)|^2 = \frac{T_M}{N}\,\mathrm{sinc}^2(\pi f T_M) * \frac{1}{t_S}\sum_{k=-\infty}^{+\infty}\delta\left(f - \frac{k}{t_S}\right) =$$

$$= \sum_{k=-\infty}^{+\infty}\mathrm{sinc}^2\left(\pi T_M\left(f - \frac{k}{t_S}\right)\right).$$

Notice that, in the calculation of this square, we have considered that the set of the shifted sinc functions is an orthogonal set, since it will not lead to any cross-product.

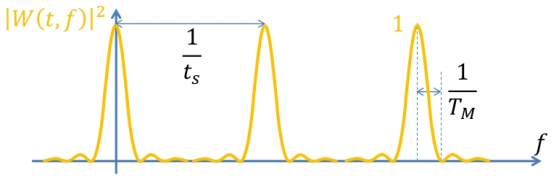


Figure 4.33: Square modulus of the Fourier transform of the weighting function.

From these calculations, we can derive the expression of the mean square value of the output noise in the case of a quasi-white noise in the frequency domain. In general, it can be written as:

$$\overline{n_y^2} = \int S_x(f)|W(t,f)|^2\,df$$

but since we are assuming to have a quasi-white noise, then the power spectral density will not vary significantly over a certain sinc function. This means that the sinc function is acting as if it were a delta function, sampling the power spectral density in certain well defined points in which it is almost constant:

$$\overline{n_y^2} \simeq \frac{1}{T_M}\sum_{k=-\infty}^{+\infty} S_x\left(\frac{k}{t_S}\right)$$

where the factor $1/T_M$ is related to the fact that area of each sinc^2 function is equal to this factor. From the expression of the measurement time:

$$T_M = N t_S$$

we can write:

$$\overline{n_y^2} = \frac{1}{N}\sum_{k=-\infty}^{+\infty}\frac{1}{t_S}S_x\left(\frac{k}{t_S}\right) = \frac{\overline{n_x^2}}{N}.$$

In the last equivalence, we have considered that the term over which we are summing is a series of rectangles that are not overlapping one with the other and that, when summed, give the power spectral density at the input of the device, thus being equal to $\overline{n_x^2}$.

We can now assume to have, at the input of this filter, a correlated input noise. However, if we assume this noise to be not too much correlated, since the

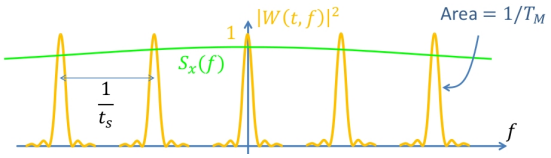


Figure 4.34: Square modulus of the weighting function compared to the power spectral density of the input quasi-white noise.

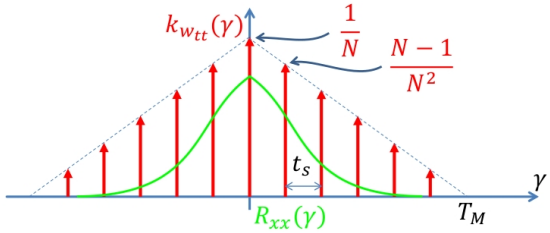


Figure 4.35: Autocorrelation of the weighting function and autocorrelation of a non-white input noise.

mean square value of the output noise can be written as:

$$\overline{n_y^2} = \int R_{xx}(\gamma) k_{w_{tt}}(\gamma)\, d\gamma$$

where $R_{xx}(\gamma)$ is the autocorrelation of the input noise, we can assume this function to be significantly different from zero only over a limited number, with respect to N, of the delta functions that are present in the autocorrelation of the weighting function. Momentarily neglecting the amplitude of these delta functions, this means that we are sampling the autocorrelation of the input noise with a (rescaled) Dirac comb, obtaining the following discrete sum:

$$\overline{n_y^2} \simeq \frac{1}{N} \sum_k R_{xx}(k t_S)$$

where the sum is extended over the number of Dirac delta functions in which the autocorrelation of the input noise is significantly different from zero. It is important to note that also in this case we have the contribution of the central Dirac delta function ($\gamma = 0$) as in the case of the white noise, plus a series of additional terms that represents positive contributions to the output noise. This means that if the noise is self-correlated, even weakly, it is more difficult to obtain a cancellation of the noise contribution between two different samples of the input.
In the frequency domain, this calculation leads to the same result:

$$\overline{n_y^2} = \int S_x(f) |W(t,f)|^2\, df \simeq \frac{1}{N} \sum_k \frac{1}{t_S} S_x\left(\frac{k}{t_S}\right)$$

where again we can consider the sinc^2 functions as if they were delta functions with respect to the power spectral density. This quantity, in fact, even though

it will not be exactly flat due to the fact that we are not dealing with a white noise, can be considered almost constant on the extension of the sinc^2 function. Alternatively, the equivalence of the two expressions for the mean square value of the output noise can be demonstrated as a consequence of the Parseval's theorem.

4.6 Comparison between continuous- and discrete-time filters

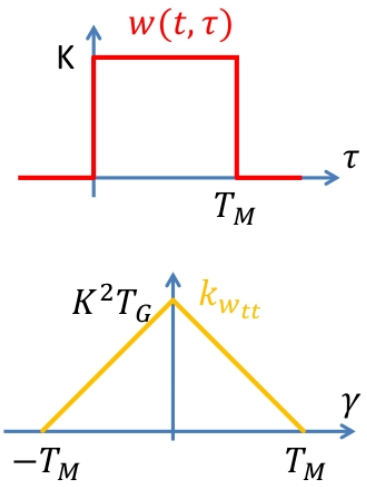


Figure 4.36: Weighting function and associated autocorrelation for a gated integrator.

It this section, we can now compare discrete-time filters with gated integrators. To perform this comparison, we first have to set the gain of the two filters to be equal. Since the gain of a discrete-time filter is unitary:

$$K_{DT} = 1$$

we have to set the temporal duration of the gate to be equal to the measurement time:

$$T_G = T_M$$

and to write that, for a gated integrator:

$$K_{GI} = \frac{1}{T_G}.$$

This means that the peak value of the autocorrelation of the weighting function for a gated integrator will be equal to:

$$k_{w_{tt}}(0) = K^2 T_G = \frac{1}{T_M}.$$

In this condition, the mean square value of the output noise of a gated integrator can be written as:

$$\overline{n_{GI}^2} = \int R_{xx}(\gamma) k_{w_{tt}}(\gamma)\ d\gamma$$

and assuming to have a small self-correlation of the input noise:

$$T_n < T_M$$

we can assume the autocorrelation of the weighting function $k_{w_{tt}}(\gamma)$ to be almost constant and equal to $k_{w_{tt}}(0)$ on the interval in which the autocorrelation of the input noise $R_{xx}(\gamma)$ is different from zero, thus obtaining:

$$\overline{n_{GI}^2} \simeq \frac{1}{T_M} \int R_{xx}(\gamma)\ d\gamma = \frac{1}{T_M} S_x(0)$$

since the integral over the whole time axis is equivalent to the value in the origin of the corresponding function in the frequency domain.
Assuming then to have the same output signal, since we have assumed to have the same gain for both filters, we can compute the ratio between the signal-to-noise ratio of the gated integrator and the one of the discrete-time filter. Since the two output signals cancel out, we can obtain:

$$\frac{\overline{n_{AV}^2}}{\overline{n_{GI}^2}} = \frac{\frac{1}{N}\sum_k R_{xx}(kt_S)}{\frac{1}{T_M}\int R_{xx}(\gamma)\ d\gamma}$$

but since we know that:

$$T_M = N \cdot t_S$$

we obtain:

$$\frac{\overline{n_{AV}^2}}{\overline{n_{GI}^2}} = \frac{t_S \sum_k R_{xx}(kt_S)}{\int R_{xx}(\gamma)\ d\gamma} > 1.$$

This ratio can be evaluated from Figure 4.37, where the continuous line represents the integrand, while the piecewise constant function is the outcome of the discrete sum. As it can be clearly seen, for any rectangle we are more or less approximating the area underlying the continuous curve apart from the central rectangle, that is always above the continuous curve. This means that the numerator in the ratio is bigger than the denominator and thus that the signal-to-noise ratio of the gated integrator is higher than the one of the dicrete-time filter considered, since the noise in this last filter is higher than the noise collected from a gated integrator. In the frequency domain, this can be written as:

$$\frac{\overline{n_{AV}^2}}{\overline{n_{GI}^2}} = \frac{\frac{1}{T_M}\sum_k S_x\left(\frac{k}{t_S}\right)}{\frac{1}{T_M} S_x(0)} = \frac{S_x(0) + \sum_{k\neq 0} S_x\left(\frac{k}{t_S}\right)}{S_x(0)} > 1.$$

Up to now, in our comparison we have only considered, as a discrete-time filter, the uniform average. However, it is also possible to perform a non-uniform average and it is usually used to mimic the behaviour of any continuous time filter or to design filters that were unfeasible in a continuous time perspective. Considering once again a constant input signal, the output signal can be written as:

$$\overline{y} = \overline{\sum_{k=1}^{N} w_k(x_k + n_k)} = \sum_{k=1}^{N} \overline{w_k}\overline{(x_k + n_k)} = \sum_{k=1}^{N} w_k(x_k + \overline{n_k})$$

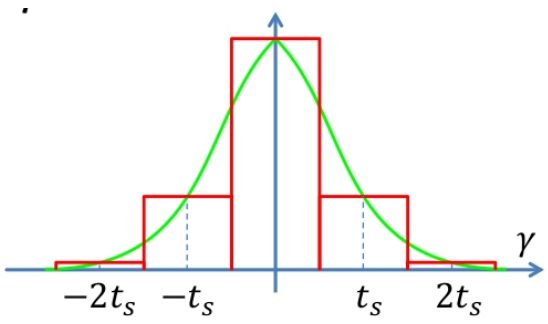


Figure 4.37: Comparison between the integral and the discrete sum in the expression of the mean square value of the output noise for a gated integrator and for a discrete-time filter.

but since the average of the noise is zero:

$$\overline{y} = \sum_{k=1}^{N} w_k x_k = A \sum_{k=1}^{N} w_k.$$

On the other hand, for a non-correlated stationary noise in each sample, thus a noise having a correlation time lower than the distance between two neighbouring samples:

$$T_n < t_S$$

the mean square value of the output noise can be written as:

$$\overline{n_{out}^2} = \sum_{k=1}^{N} w_k^2 \overline{n_k^2} = \overline{n_{in}^2} \sum_{k=1}^{N} w_k^2.$$

Calculating then the signal-to-noise ratio in this case:

$$\left(\frac{S}{N}\right)_{out} = \frac{A \sum_{k=1}^{N} w_k}{\sqrt{\overline{n_{in}^2} \sum_{k=1}^{N} w_k^2}} = \left(\frac{S}{N}\right)_{in} \frac{\sum_{k=1}^{N} w_k}{\sqrt{\sum_{k=1}^{N} w_k^2}}.$$

Note that, if we assume to have constant weights, the improvement with $\sqrt{N}$ that we have previously demonstrated can be recovered.
Another possibility for having a discrete-time filter is to use a power-law weighting, where the weights are equal to a certain basis α to the power of the order k considered:

$$w_k = \alpha^k.$$

In this case, the average value of the output signal can be written as:

$$\bar{y} = A \sum_{k=0}^{N-1} \alpha^k \simeq A \sum_{k=0}^{\infty} \alpha^k = A \frac{1}{1-\alpha}$$

where we have assumed that the number of samples N considered is large, thus allowing us to recognize the presence of a geometric series. Under the same

assumption, the mean square value of the output noise can be written as:

$$\overline{n_{out}^2} = \overline{n_{in}^2} \sum_{k=0}^{N-1} \alpha^{2k} \simeq \overline{n_{in}^2} \sum_{k=0}^{\infty} \alpha^{2k} = \overline{n_{in}^2} \frac{1}{1-\alpha^2}$$

thus giving rise to the following signal-to-noise ratio:

$$\left(\frac{S}{N}\right)_{out} = \left(\frac{S}{N}\right)_{in} \frac{\sqrt{1-\alpha^2}}{1-\alpha} = \left(\frac{S}{N}\right)_{in} \frac{\sqrt{(1-\alpha)(1+\alpha)}}{1-\alpha} = \left(\frac{S}{N}\right)_{in} \sqrt{\frac{1+\alpha}{1-\alpha}}$$

where we have considered the input signal-to-noise ratio:

$$\left(\frac{S}{N}\right)_{in} = \frac{A}{\sqrt{\overline{n_{in}^2}}}.$$

In this case, we can define the equivalent number of samples as:

$$N_{eq} = \frac{1+\alpha}{1-\alpha}$$

consistently with what we have seen before. It is important to see, now, that the equivalent number of samples N_{eq} increases with α, being divergent for:

$$\alpha \ \rightarrow \ 1$$

where, in this limit:

$$\left(\frac{S}{N}\right)_{out} \ \xrightarrow[\alpha \rightarrow 1]{} \ +\infty$$

thus meaning that we do not have any noise. However, reasoning on the meaning of this limit, in this case we are obtaining a uniform average and this assumption, that is clearly unphysical, comes from the fact that we have assumed the number N of samples to be extremely large, in particular:

$$N \ \rightarrow \ \infty$$

in order to be able to write the previous output signal and noise as two geometric series. Not considering this approximation, we could have written:

$$\sum_{k=0}^{N-1} \alpha^k = \frac{1-\alpha^N}{1-\alpha}, \quad \sum_{k=0}^{N-1} \alpha^{2k} = \frac{1-\alpha^{2N}}{1-\alpha^2}$$

for a finite value of N. In this case, the expression of the signal-to-noise ratio is different and it gives the following number of equivalent samples:

$$N_{eq} = \left(\frac{1+\alpha}{1-\alpha}\right) \cdot \left(\frac{1-\alpha^N}{1+\alpha^N}\right)$$

where we can see that also in this case:

$$N_{eq} \propto \alpha$$

and the best case is when:

$$\alpha \rightarrow 1 \ \Rightarrow \ N_{eq} \simeq N$$

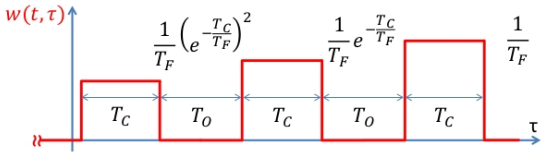


Figure 4.38: Representation of a boxcar averager as an averaging filter.

and we are considering a uniform average. This result is not strange: every sample, in fact, will contain the same signal and the same amount of noise, therefore there is not any reason for treating different samples differently. This means that the best option for filtering an uncorrelated noise is to have a uniform average.

We can now consider the weighting function of a boxcar averager in the following limit:

$$T_C \ll T_F$$

as it is represented in Figure 4.38. In this case, the boxcar averager can be seen as the cascade of two different filters. One filter, in fact, will give the single-pulse rectangular weighting function and, therefore, it will be a gated integrator with a gain equal to $1/T_F$. The other filter, then, will scale the output of the first filter with a power-law average, where we can consider the following parameter:

$$\alpha = e^{-\frac{T_C}{T_F}}.$$

The improvement in the signal-to-noise ratio given by the introduction of the second stage can then be written as:

$$N_{eq} = \frac{1+\alpha}{1-\alpha} = \frac{1+e^{-\frac{T_C}{T_F}}}{1-e^{-\frac{T_C}{T_F}}} \simeq \frac{2T_F}{T_C}$$

under the previously considered hypothesis, thus obtaining exactly the relationship that is linking the signal-to-noise ratio of a boxcar averager to the one of a gated integrator. Considering also the discharge of the device when the switch is open, thus during the time intervals T_O, we can obtain an analogous expression for the ratemeter:

$$\alpha = e^{-\frac{T_C+T_O}{T_F}} \;\Rightarrow\; N_{eq} = \frac{2T_F}{T_C+T_O}$$

where this expression will be valid in the following approximation:

$$T_C + T_O \ll T_F.$$

4.7 Optimum filtering

In the general case, therefore, we have seen how it is possible to filter some types of noise depending on the signal we have. The question now is: what is the best filter, called the optimum filter, that we can apply for filtering a certain noise and a certain signal?

In many physics and engineering applications, the signal has a fixed and known shape, that will depend on several different parameters. The purpose of filtering, then, is to make a precise measurement of these parameters, for example the amplitude or the arrival time of this signal. We want thus to determine the optimum filter, that gives us the best signal-to-noise ratio for this kind of problem. The first example we can consider is the one of a discrete-time filter that is filtering a generic, non-constant signal $Ax(t)$ (where A is the amplitude) affected by white noise. Assuming this discrete-time filter to be sampling the input (both signal and noise) at discrete time instants:

$$t_k = k \cdot t_S$$

we can write the sampled input signal as:

$$A \cdot x_k = A \cdot x(k \cdot t_S)$$

and we can assume this amplitude A to be the parameter of interest. In this case, at the output of the filter we will have:

$$y = \sum_{k=0}^{N-1} w_k (A x_k + n_k)$$

and we can thus write the mean value of the output signal as:

$$\bar{y} = \overline{\sum_{k=0}^{N-1} w_k (A x_k + n_k)} = \sum_{k=0}^{N-1} w_k \overline{(A x_k + n_k)} = \sum_{k=0}^{N-1} w_k (A x_k + \overline{n}_k) =$$
$$= A \sum_{k=0}^{N-1} w_k x_k.$$

The mean square value of the output noise, then, can be written as:

$$\overline{n_{out}^2} = \sum_{k=0}^{N-1} w_k^2 \overline{n_k^2} = \overline{n_{in}^2} \sum_{k=0}^{N-1} w_k^2$$

where we have considered that the signal has a null variance (being deterministic), that all the noise samples are uncorrelated (being a white noise) and thus the variances adds up and that the noise is stationary, thus all the noise samples have the same, constant variance $\overline{n_{in}^2}$. From these considerations, we can write the output signal-to-noise ratio as:

$$\left(\frac{S}{N}\right)_{out} = \left(\frac{S}{N}\right)_{in} \frac{\sum_{k=0}^{N-1} w_k x_k}{\sqrt{\sum_{k=0}^{N-1} w_k^2}}$$

where the input signal-to-noise ratio clearly is:

$$\left(\frac{S}{N}\right)_{in} = \frac{A}{\sqrt{\overline{n_{in}^2}}}.$$

Our goal, now, is to determine the optimum choice for the weights w_k in order to maximize the signal-to-noise ratio. This can be done by differentiating the

output signal-to-noise ratio with respect to a certain k-th weight and imposing this derivative equal to zero; repeating this procedure for any weight, we will obtain the optimum filter:

$$\frac{\partial}{\partial w_n}\left(\frac{S}{N}\right)_{out} = 0 \quad \forall n = 1, \dots, N.$$

Computing one of these derivatives (they are all equal), we obtain:

$$\frac{x_n\sqrt{\sum_{k=0}^{N-1} w_k^2} - \frac{w_n}{\sqrt{\sum_{k=0}^{N-1} w_k^2}}\sum_{k=0}^{N-1} w_k x_k}{\sum_{k=0}^{N-1} w_k^2} = 0$$

where we have neglected the constant term represented by the input signal-to-noise ratio (since it would have multiplied the whole fraction). Solving this expression for the n-th weight w_n we obtain:

$$w_n = x_n \frac{\sum_{k=0}^{N-1} w_k^2}{\sum_{k=0}^{N-1} w_k x_k} \quad \forall n.$$

In this case, we can clearly see that the optimum weights are proportional to the amplitude of the signal. We can immediately observe, in fact, that the ratio between the two sums that is multiplying the signal x_n will be the same for any weight, thus representing a constant factor; the only relevant dependence is the one from the amplitude of the input signal x_n. This is actually reasonable: any sample will have the same amount of uncertainty (represented by the noise) since the noise is stationary. We want thus to give more importance to the samples in which we have more information, that are actually the samples in which the signal is stronger.
A second case that we can study is the one of a continuous-time filter in which again the input signal $Ax(t)$ is affected by a white noise. In this case, from the definition of the output signal, we can write:

$$y = A\int x(\tau)w(t,\tau)\ d\tau$$

and thus, since the autocorrelation of the noise is a delta function by definition of white noise, we can write the mean square value of the output noise as:

$$\overline{n_{out}^2} = \lambda k_{w_{tt}}(0) = \lambda \int w^2(t,\tau)\ d\tau.$$

In this case, the square of the output signal-to-noise ratio can be written as:

$$\left(\frac{S}{N}\right)_{out}^2 = \frac{y^2}{\overline{n_{out}^2}} = \frac{A^2}{\lambda}\frac{\left|\int x(\tau)w(t,\tau)\ d\tau\right|^2}{\int w^2(t,\tau)\ d\tau}.$$

To determine the optimum value of the weighting function, we have now to recall an important inequality coming from Functional Analysis, that is called the Schwartz inequality, in the Lebesgue space L^2:

$$|\langle x, w\rangle| \leq \|x\|_2 \cdot \|w\|_2.$$

Squaring this relationship, since all the terms involved are by definition positive quantities, we can write:

$$|\langle x, w\rangle|^2 \leq \|x\|_2^2 \cdot \|w\|_2^2$$

and therefore, from the definitions of the scalar product and norm in this space, we obtain that:

$$\left| \int x(\tau) w(t,\tau)\ d\tau \right|^2 \leq \int |x(\tau)|^2\ d\tau \cdot \int |w(t,\tau)|^2\ d\tau.$$

It is important to note that, in the case in which:

$$w(t,\tau) \propto x(\tau)$$

this inequality actually becomes an equivalence, from the definition of norm and of inner product. Applying this inequality to the previously computed signal-to-noise ratio:

$$\left(\frac{S}{N}\right)^2_{out} \leq \frac{A^2}{\lambda} \frac{\int |x(\tau)|^2\ d\tau \cdot \cancel{\int |w(t,\tau)|^2\ d\tau}}{\cancel{\int w^2(t,\tau)\ d\tau}} = \frac{A^2}{\lambda} \int |x(\tau)|^2\ d\tau$$

and, as we have previously said, the maximum signal-to-noise ratio, that corresponds to the optimum case, will be obtained when:

$$w(t,\tau) \propto x(\tau)$$

consistently with what we have stated for discrete-time systems. In this condition we are dealing with a so called matched filter, that can be represented as in Figure 4.39.

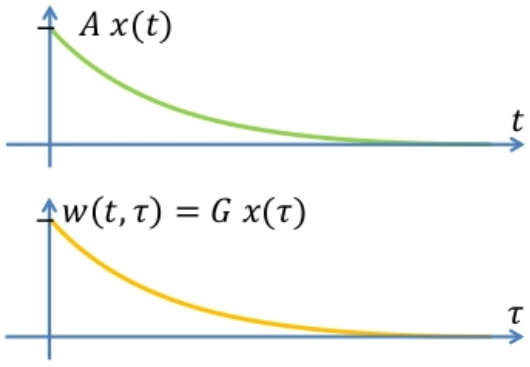


Figure 4.39: A matched continuous-time filter.

It is important to notice that this filter will weight more the regions in which the signal is stronger, while it will discard the regions in which it is negligible. In the first regions, in fact, the signal will be much stronger than the noise, thus having an higher signal-to-noise ratio, while in the other regions the signal will be similar to or lower than the noise, giving a very poor signal-to-noise ratio. This will imply, however, that the designer of the filter knows the shape

(in a more formal way, the time dependence) of the signal, having as the only unknown its amplitude. This is the case in almost any physical signal, but not in everyone of them; it depends on the problem we are dealing with. Assuming a matched filter:

$$w(t,\tau) = G \cdot x(\tau)$$

we can write the output signal as:

$$y = AG \int x^2(\tau)\ d\tau$$

thus obtaining as the mean square value of the output noise:

$$\overline{n_{out}^2} = \lambda k_{w_{tt}}(0) = \lambda G^2 \int x^2(\tau)\ d\tau$$

that will give the following value of the signal-to-noise ratio:

$$\left(\frac{S}{N}\right)_{out} = \frac{A}{\sqrt{\lambda}} \sqrt{\int x^2(t)\ dt} = \sqrt{\frac{E}{\lambda}}$$

where E has been defined as the energy of the signal.
This optimum filtering property can also be determined in the frequency domain, where the previously written signal-to-noise ratio becomes:

$$\begin{aligned}\left(\frac{S}{N}\right)^2 &= \frac{A^2}{\lambda} \cdot \frac{\left|\int X(f)W^*(t,f)\ df\right|^2}{\int |W(t,f)|^2\ df} \leq \\ &\leq \frac{A^2}{\lambda} \cdot \frac{\int |X(f)|^2\ df \cdot \cancel{\int |W(t,f)|^2\ df}}{\cancel{\int |W(t,f)|^2\ df}} = \\ &= \frac{A^2}{\lambda} \int |X(f)|^2\ df\end{aligned}$$

where we have considered the following inequality:

$$\left|\int X(f)W^*(t,f)\ df\right|^2 \leq \int |X(f)|^2\ df \cdot \int |W(t,f)|^2\ df$$

where the square modulus makes $W(t,f)$ equal to its complex conjugate and where the equality holds if and only if:

$$X(f) = W(t,f).$$

Alternatively, the same result could have been derived using the Parseval's theorem.
In the frequency domain, we can now consider the case of a non-white noise, where the signal in the frequency domain is $AX(f)$ and the noise has a certain stationary power spectral density $S_n(f)$. In this case, we can assume to have at our disposal a particular filter, called linear whitening filter $H_w(f)$, that is able to transform any non-white input power spectral density into a white power spectral density:

$$S_n(f)|H_w(f)|^2 = \lambda = \text{const.}$$

From this definition, this whitening filter can be written as:

$$|H_w(f)| = \sqrt{\frac{\lambda}{S_n(f)}}$$

and, after it, we can use a matched filter on the output white signal:

$$X_w(f) = H_w(f)AX(f)$$

as we have just seen in the previous part of this section. It is extremely important to remember that this whitening filter will act both on the noise and on the signal without any distinction between them: this is a possible source of errors during the exercises.

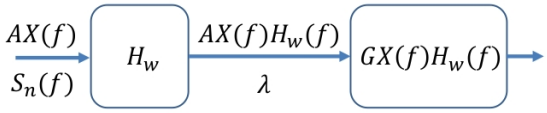


Figure 4.40: Use of a whitening filter and its effect on the signal and on the noise.

In Figure 4.40 we have represented the cascade of a whitening filter and an optimum filter and its effect on the input signal and input noise, in particular considering that for an optimum filtering the weighting function of the filter must be proportional to the signal at the input of this filter multiplied by a suitable gain G:

$$GX(f)H_w(f).$$

At the output of this device (that is usually called, as a whole, thus comprehending also the whitening filter, an optimum filter) we obtain the following signal:

$$\begin{aligned} y(t) &= \int AX(f)H_w(f) \cdot (GX(f)H_w(f))^* \ df = \\ &= \int AX(f) \left(GX(f)|H_w(f)|^2\right)^* \ df = \\ &= \int A|X(f)|^2|H_w(f)|^2 \ df \end{aligned}$$

where thus the global weighting function of the filter, in the frequency domain, is:

$$W(t,f) = GX(f)|H_w(f)|^2.$$

This gives us the following signal-to-noise ratio:

$$\begin{aligned} \left(\frac{S}{N}\right)^2_{out} &= \frac{A^2}{\lambda} \int |X(f)|^2|H_w(f)|^2 \ df = \\ &= A^2 \int \frac{|X(f)|^2}{S_n(f)} \ df. \end{aligned}$$

The optimum filter can then be redefined, in the frequency domain, as:

$$W(f) = GX(f)|H_w(f)|^2 = G' \cdot \frac{X(f)}{S_n(f)}, \quad G' = G \cdot \lambda$$

thus being proportional to the ratio between the spectrum of the signal $X(f)$ and the input, non-white power spectral density of the noise. Alternatively, we could have derived this result from an analytical point of view directly in the frequency domain by using the Schwartz inequality:

$$\begin{aligned}\left(\frac{S}{N}\right)^2_{out} &= A^2 \frac{\left|\int X(f)W^*(t,f)\ df\right|^2}{\int |W(t,f)|^2 S_n(f)\ df} \leq \\ &\leq A^2 \frac{\int \frac{|X(f)|^2}{S_n(f)}\ df \cdot \cancel{\int S_n(f)|W^*(t,f)|^2\ df}}{\cancel{\int |W(t,f)|^2 S_n(f)\ df}} = \\ &= A^2 \int \frac{|X(f)|^2}{S_n(f)}\ df\end{aligned}$$

where, to apply the Schwartz inequality, we have multiplied and divided, in the integral at the numerator, by $\sqrt{S_n(f)}$. We can thus consider that the optimum filter will be then related to the ratio between the spectrum of the signal and the power spectral density of the non-white noise at the input. The fact that we are splitting the description of the optimum filter in a whitening filter cascaded with a matched part is only a matter of convenience and interpretation: in reality the two parts are implemented together in a variety of ways. Last, if the noise is gaussian, the filter will be optimum even with respect to some non-linear alternatives.

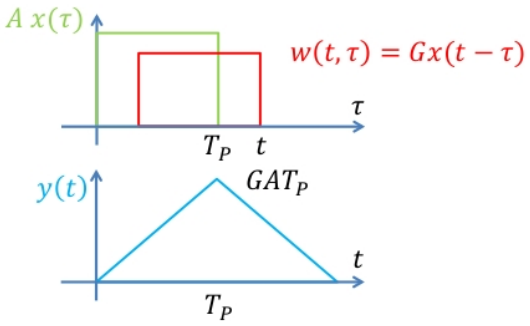


Figure 4.41: A finite time pulse and the corresponding output.

We can now apply these considerations to the case, represented in Figure 4.41, of a finite pulse time. In particular, we can assume $Ax(\tau)$ to be a rectangular signal to which is superimposed a white noise. Since we want to be dealing with an optimum filtering continuous-time case, we can write the weighting function of the filter as:

$$w(t,\tau) = Gx(t-\tau)$$

since in this way it is clearly proportional to the signal and it will have the same

duration T_p of the pulse. In this case, the output signal will be:

$$y(t) = \int Ax(\tau)w(t,\tau)\ d\tau$$

and therefore it will be the integral of the product between two rectangles, that is clearly the triangle represented in the Figure. In this case, the maximum value of the output is equal to:

$$GAT_p$$

and this will be the value to be considered when calculating the signal-to-noise ratio. However, to obtain this value we need to wait a time T_p from the arrival of the signal, therefore a time T_p is needed for processing the pulse. If this time T_p is very long, this might be a problem.

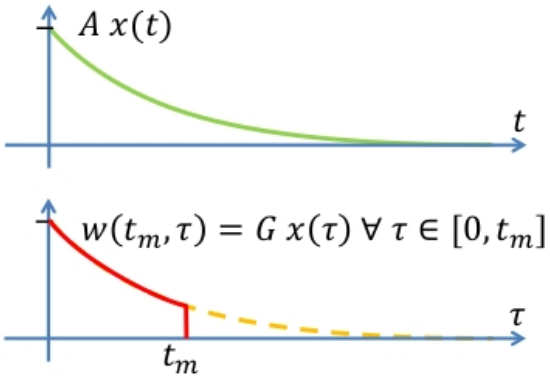


Figure 4.42: Effect of a finite measurement time.

To better understand why a long pulse duration might represent a problem, we can consider the case represented in Figure 4.42. In this case, the input signal is a decaying exponential, that therefore will vanish only in the limit for the time t that tends to infinity. From a practical point of view, however, we can suppose that five or six time constants are enough for the signal to be completely decayed. If another signal is coming to the system before the first one has completely decayed or, alternatively, if we require the response of the system to have a certain speed, this might represent a problem. The question therefore is: what is the maximum time that we can wait for the acquisition of this signal? This time will be finite and it is called the readout time t_m. This means that the weighting function will match the signal only in the interval of interest, thus between 0 (assumed to be the arrival time of the pulse) and the time t_m. In this case, the signal-to-noise ratio that we can obtain will be smaller than the optimum one, since the output signal will be:

$$y(t) = \int_0^{\infty} Ax(\tau)w(t,\tau)\ d\tau = A\int_0^{t_m} w^2(t,\tau)\ d\tau.$$

We are thus loosing part of the precision of our filter, but it will have a faster response.

In the case of an exponential input signal, then, the output will be the one represented in Figure 4.43. In the case of a truncated exponential input

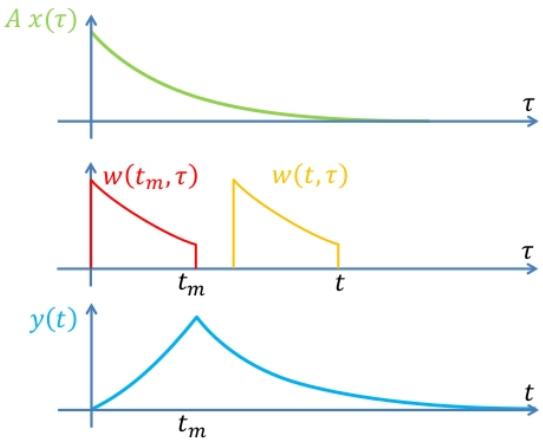


Figure 4.43: Output signal for an exponential input signal.

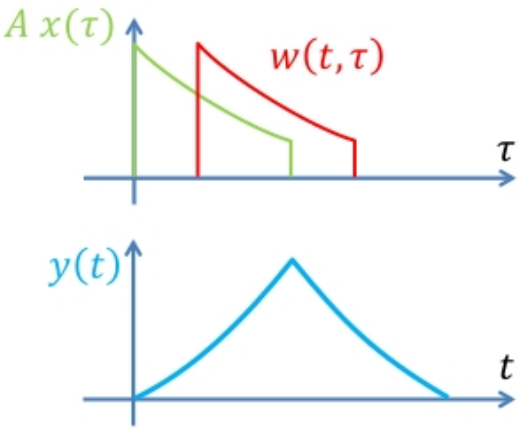


Figure 4.44: Output signal for a truncated exponential input signal.

signal, that is a time-limited input signal, on the other hand, the output will be proportional to the time correlation of the input signal, as it is represented in Figure 4.44. This is the reason why the matched filter is sometimes called the correlator.

We can now study the delta-function response of this filter with a limited measurement time. If the filter is linear and time-invariant, then we have that the delta function response is just the shifted time reversal of the weighting function:

$$w(t, \tau) = h(t - \tau).$$

In principal, however, this filter is difficult or impossible to build, since the weighting function of an optimum filter is almost never a linear and time-invariant filter. In fact, it will rather be a time-variant or digital filter, since it will involve a sampling procedure, therefore some approximations are often used in order to simplify the design of the filter still having an acceptable performance degradation. In this case, the weighting function $w(t, \tau)$ is only an

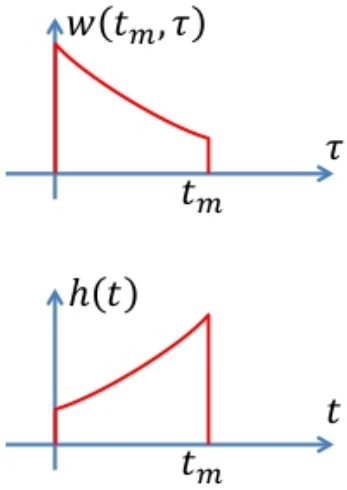


Figure 4.45: Delta function response of this filter.

approximation of the signal $x(t)$, thus leading to some errors that will lower the signal-to-noise ratio; we must then verify if these errors justify or not the increase in the complexity that is needed for implementing a real optimum filter. In the case of non-stationary white noise, the autocorrelation of the noise and its power spectral density can be written as:

$$R_{nn}(t_1, t_2) = \lambda(t_1)\delta(t_2 - t_1), \quad S_n(t, f) = \lambda(t)$$

where the power spectral density is constant but with a time-dependent value, while the noise autocorrelation is a delta function that is changing position with time. This time dependence is the consequence of the fact that we are dealing with a non-stationary noise. In this case, the output noise can be written as:

$$\begin{aligned}
\overline{n_{out}^2}(t) &= \iint R_{nn}(\alpha, \beta) w(t, \alpha) w(t, \beta)\ d\alpha\ d\beta = \\
&= \iint \lambda(\alpha)\delta(\alpha - \beta) w(t, \alpha) w(t, \beta)\ d\alpha\ d\beta = \\
&= \int \lambda(\alpha) \left(\int \delta(\alpha - \beta) w(t, \beta)\ d\beta \right) w(t, \alpha)\ d\alpha = \\
&= \int \lambda(\alpha) w^2(t, \alpha)\ d\alpha.
\end{aligned}$$

Since the output signal is, again:

$$y = A \int x(\tau) w(t, \tau)\ d\tau$$

we can observe that we have obtained exactly the same expressions that we had for a stationary non-white noise, therefore we can draw the same conclusions, obtaining that the optimum filter will be characterized by the following weighting function:

$$w(t, \tau) = G \frac{x(\tau)}{\lambda(\tau)}.$$

In the general case, however, we will not have any simplification and we will need to rely only on the most general formulas.
Last, as an example we can consider the case of the shot noise. Assuming to be dealing, for example, with the signal coming from a photodetector, the output signal will be a certain current $I(t)$ that is affected by a shot noise[11] with the following bilateral power spectral density:

$$S_n = qI(t)$$

thus being a white and non-stationary noise term. In this case, the optimum filter will become, from the previous relation:

$$w(t,\tau) = G \cdot \frac{\cancel{I(t)}}{q\cancel{I(t)}} = \text{const}$$

thus being identical to a gated integrator. Changing the amplitude of the signal, in fact, we are changing also the amplitude of the noise for the whole duration of the signal, under the assumption that the shot noise is the prevailing noise source in our device. If other noise sources are present (for example, the thermal noise), the weighting function must be modified accordingly, depending on the fact that we are dealing with a noise that is white and/or stationary.

4.8 Low-frequency noise

At this point, we have completed the part about the white and high-frequency noise, therefore we can change our perspective and consider the low-frequency noise. In this case, since the noise is at a low frequency, its correlation time will be long with respect to the acquisition time, changing radically the way we are filtering it. If the different noise samples are uncorrelated or weakly correlated, in fact, we can try to average them, but if they are correlated averaging is a totally ineffective operation, since we will obtain almost the same amount of noise. We need then to tackle this problem with different techniques.

4.8.1 High-pass filters

As we have just considered, the previous methods are not effective in reducing a low-frequency noise. This is clear both in the time domain, where the variations of the noise take place on a time-scale that is much larger than the one of the filter, and in the frequency domain, where the power spectral density of the noise is located at low frequencies, where the square modulus of the Fourier transform of the weighting function is different from zero. We want thus to be able to reject the low frequency components and this can be done using a so called high-pass filter (HPF).

This kind of filter is represented in Figure 4.46. It is clearly a linear and time-invariant filter, being the composition of different linear and time-invariant components, and its output is the voltage measured across the resistor R. To determine its delta function response, we can first consider its step function response. When a step is coming to the input of the device, the voltage across

[11] Related to the Poisson fluctuations in the number of carriers, thus making the power spectral density proportional to the signal.

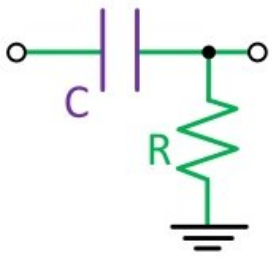


Figure 4.46: An high-pass filter (HPF).

the resistor R will instantaneously increase to the value of the amplitude of the step applied. This is because, for a short time interval, we are considering the behaviour of the capacitor at high frequency, where it can be assumed to be a short-circuit. After that, an exponentially decreasing behaviour starts, since the voltage across the capacitor C will exponentially increase. Deriving this step response, it is then possible to determine the delta function response of this filter, that will be:

$$h(t) = \delta(t) - \frac{1}{T} e^{-\frac{t}{T}} u(t)$$

where T is the time constant of the filter.

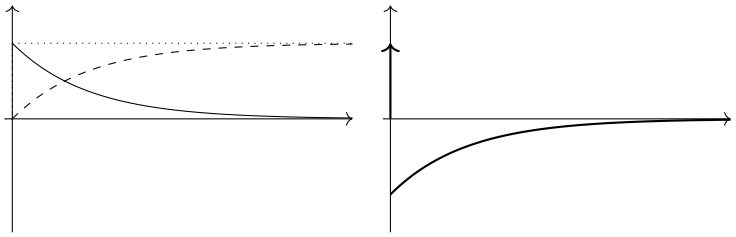


Figure 4.47: Step response (on the left, the input is dotted, the voltage across the capacitor is dashed and the voltage across the resistor is solid) and delta-function response (on the right) of an high-pass filter.

In the frequency domain, from the analysis of the circuit, we can write its transfer function as:

$$H(s) = \frac{sT}{1+sT} = 1 - \frac{1}{1+sT}.$$

It is important to notice that since:

$$H(0) = 0$$

then we will have that:

$$\int_0^{+\infty} h(t)\ dt = 0$$

and therefore the area underlying the Dirac delta must be equal to the area that is above the negative increasing exponential function. The autocorrelation

of the weighting function can then be calculated as:

$$k_{hh}(\tau) = \int h(t)h(t+\tau)\ dt = \\ = \int_0^{+\infty} \left[\delta(t) - \frac{1}{T}e^{-\frac{t}{T}}\right] \cdot \left[\delta(t+\tau) - \frac{1}{T}e^{-\frac{t+\tau}{T}}\right]\ dt$$

but since for positive values of τ the delta function $\delta(t+\tau)$ is sampling the other response $h(t)$ in part in which it is for sure identically equal to zero, we can write:

$$k_{hh}(\tau) = -\int_0^{\infty} \left[\delta(t) - \frac{1}{T}e^{-\frac{t}{T}}\right] \cdot \frac{1}{T}e^{-\frac{t+\tau}{T}}\ dt = \\ = -\frac{e^{-\frac{\tau}{T}}}{T} + \frac{e^{-\frac{\tau}{T}}}{T^2}\int_0^{\infty} e^{-\frac{2t}{T}}\ dt = -\frac{e^{-\frac{\tau}{T}}}{2T}$$

and since this function is defined as an even function (because we do not mind which replica of the delta-function response we are actually shifting in time), we obtain:

$$k_{hh}(\tau) = -\frac{e^{-\frac{|\tau|}{T}}}{2T}.$$

However, since the value in zero of the autocorrelation $k_{hh}(0)$ must be equal to the integral of the two weighting functions that are perfectly overlapping, in this position also the delta-functions that are present in the two replicas of the weighting function are overlapping; this means that one of them will sample the other and it then will be present in the autocorrelation of the weighting function, giving as a correct result:

$$k_{hh}(\tau) = \delta(\tau) - \frac{e^{-\frac{|\tau|}{T}}}{2T}.$$

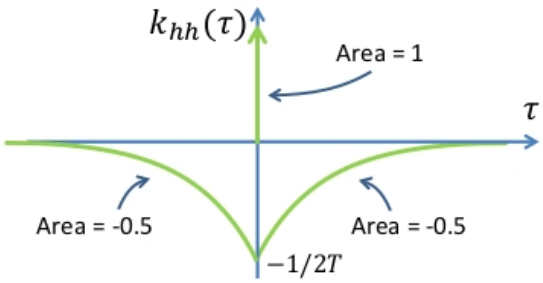


Figure 4.48: Autocorrelation of the weighting function of an high-pass filter.

At the output of this filter, therefore, the mean square value of the noise, given a certain input noise autocorrelation $R_{xx}(\tau)$, will be:

$$\overline{n_y^2} = \int R_{xx}(\tau)k_{hh}(\tau)\ d\tau = R_{xx}(0) - \frac{1}{2T}\int R_{xx}(\tau)e^{-\frac{|\tau|}{T}}\ d\tau.$$

A signal with relatively large correlation time, therefore, will make the contribution of the second term, that is negative, relevant, thus reducing the mean

square value of the output noise, as desired. On the other hand, if the correlation time of the input noise is small, it will be almost exclusively sampled by the delta-function and not significantly reduced by the exponential parts.
To study the effect of this filter on the noise, we can consider an input noise with a rectangular autocorrelation $R_{xx}(\tau)$ with constant amplitude $\overline{n_x^2}$ extended over a certain interval from $-T_n$ to T_n. In this case, since both the autocorrelation of the input noise and the autocorrelation of the weighting function are two even functions, we can write the mean square value of the output noise as:

$$\overline{n_y^2} = \overline{n_x^2} - \frac{1}{T}\int_0^{T_n} \overline{n_x^2} e^{-\frac{\tau}{T}}\, d\tau = \overline{n_x^2} e^{-\frac{T_n}{T}}.$$

For a white or uncorrelated noise, the time T_n will be small; in particular:

$$T_n \ll T \;\Rightarrow\; \overline{n_y^2} \simeq \overline{n_x^2}$$

and thus the filter has little effect on this kind of noise. On the other hand, for a low-frequency noise:

$$T_n \gg T$$

this filter will be effective in reducing this noise. In the frequency domain, therefore, this filter will reject all the components that will be below the following cut-off frequency:

$$\frac{1}{2\pi T}.$$

4.8.2 Effects on pulsed signals

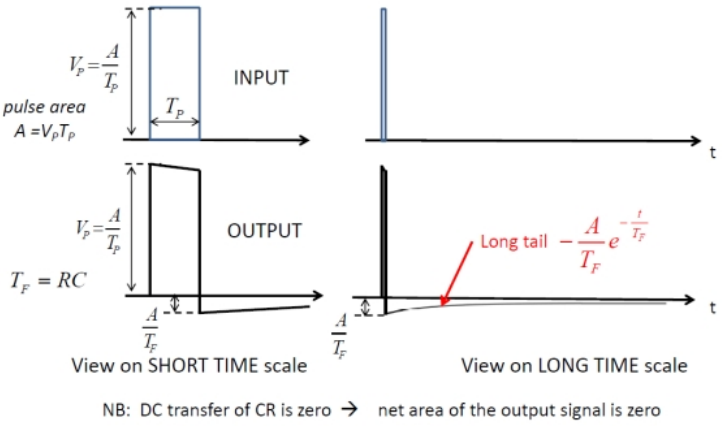


Figure 4.49: Effect of an high-pass filter on a single pulse.

We can now investigate the effect of this device on a pulsed signal, that is the prototype of an high-frequency signal. First of all, we can see that the "step" part of this pulse will be perfectly reproduced at the output, increasing abruptly the output voltage. Then, an exponential discharge of the capacitor will start, decreasing the value of this voltage. At the end of this pulse, we have another step, that will be reflected in a step of the same amplitude at

the output. However, due to the presence of the exponential discharge, we are now at a voltage that is lower than the initial one, therefore this step will bring the output at a certain negative voltage proportional to the amplitude of the exponential discharge that has taken place. Then, the exponential discharge will continue, bring the output from a certain negative voltage again toward zero. The length of this exponential tail will be related to the time constant of the filter T_F (previously indicated with T), while the "undershoot" at the end of this pulse will be related both to this time constant and to the duration T_p of the pulse. Since the DC transfer of this high-pass filter is zero, then the net area of the pulse will be exactly equal to zero and therefore the area of the output signal actually related to the pulse will be equal to the area of the negative exponential tail related to the discharge of the capacitor. This behaviour is represented in Figure 4.49.

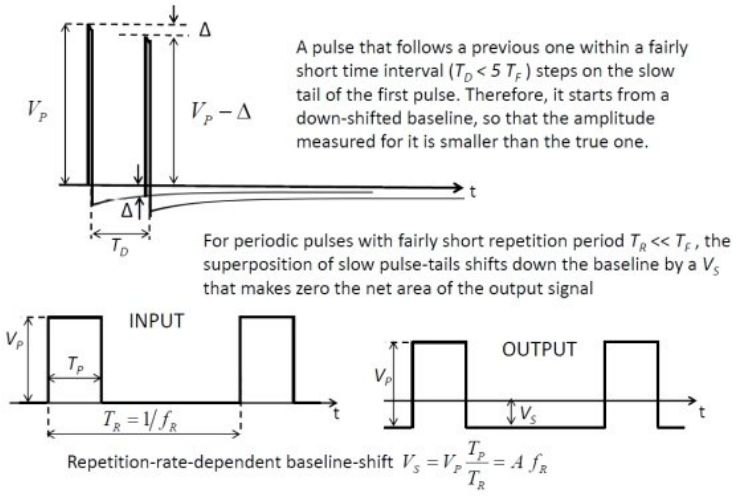


Figure 4.50: Effect of an high-pass filter on two subsequent pulses.

In general, this long exponential tail can be completely neglected, unless we have more than one pulse coming to our filter; in this case, the arrival time of the second pulse is important. In fact, if the second pulse comes when the exponential tail of the first one has completely decayed (thus, the pulses are coming with a small repetition rate), nothing changes in the behaviour of the device: the output voltage is again at zero when the second pulse arrives. However, if the second pulse arrives when the system has not already relaxed (this means that the exponential tail of the first pulse is not again at zero), the system will undergo a positive step of amplitude equal to the initial one but, since this step is starting from a certain negative voltage, it will reach a maximum voltage that is lower than the one reached by the first step. Then, we will have again the exponential decay due to the discharge of the capacitor and, at the end of the second pulse, again a negative step. This step will bring the output at a negative voltage that is in magnitude higher than the previous one, determining an even longer exponential tail that can become relevant for subsequent pulses. This dangerous effect is called the pile-up of the system.
This superposition of the exponential tails of the pulses, at the end, gives prob-

lems when the repetition rate of the pulses is high. The error we commit as a consequence of this effect is in fact proportional to the repetition rate of the signal (assuming to have a periodic input signal). After a certain transient, at the end this will determine a variation of the baseline of the signal as in Figure 4.50. In fact, in an high-pass filter the maximum penalization of the input will be related to the DC component of the input signal, that is associated to its average value. This means that we are bringing the average value of the output signal at zero. As a general rule, therefore, we must be extremely careful in using high-pass filter, since almost any signal will be modified by them.

4.9 Baseline restorers

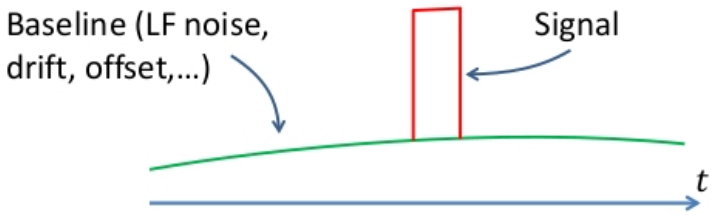


Figure 4.51: A signal superimposed to a slowly varying baseline.

An alternative for measuring a signal superimposed to a slowly varying baseline could be to measure the baseline when the signal is not present and, then, subtract this measurement from the noisy sample of the signal. This is obviously a time-variant filter and it makes sense if and only if the noise is correlated between the two samples. This filter is called a baseline restorer and it can be represented as in Figure 4.52.

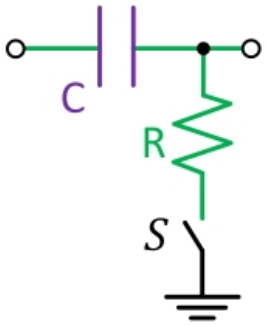


Figure 4.52: A baseline restorer.

The simplest implementation of this filter consists in a capacitor C, a resistor R and a switch S, where this last element is able of controlling the different phases of this device. When we want to measure the baseline of the signal, therefore, the switch S will be closed. In this time interval, there is not any signal coming to the device and therefore we are measuring only the low-frequency

noise at the input. Considering this noise to be, for example, just a constant offset, after a certain transient the voltage across the resistor R will be equal to this offset voltage. The next phase, then, is to subtract this baseline from the signal, and this can be done when the switch is open. In this time interval, at the input of the device we have the signal superimposed to the noise, therefore opening the switch the voltage across the capacitor remains unchanged and equal to the value it had at the end of the previous phase. This makes the output voltage equal to the voltage at the input (determined by the input and the baseline) minus the voltage across the capacitor (exclusively determined by the baseline), giving at the output the input voltage without the baseline.
This device will work perfectly in the case of constant noise, where the noise is completely self-correlated. Decreasing the autocorrelation of the noise, the baseline between the two samples will be increasingly different, thus making this network work worse and worse.

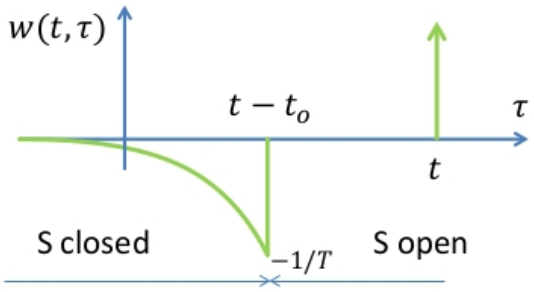


Figure 4.53: Weighting function of a baseline restorer.

To determine the weighting function of a baseline restorer, we must consider that the output voltage of the device can be considered as equal to the difference between the input voltage and the voltage across the capacitor:

$$V_{out} = V_{in} - V_C.$$

Remembering that the weighting function is the response of the system at time t to a delta-function applied at time τ, we can study this behaviour for the voltage V_C across the capacitor and, then, derive from it the behaviour of the output. When the switch is open, the input voltage will be equal to the output voltage:

$$V_{in} = V_{out}$$

and thus it will determine, in the delta-function response of the device, a delta-function; this behaviour holds for any arrival time τ of the input delta-function within the time interval in which the switch S is open. When the switch is closed, this system is identically equal to a low-pass filter, thus giving a decreasing exponential behaviour until the time instant at which the switch opens; after that time interval, the delta-function response is constant. From the definition of weighting function, then, we can derive the behaviour represented in Figure 4.53. Since we have already seen that the output voltage is given by two contributions, one that is positive and that is related to the input voltage and the other that is negative and that is related to the voltage across the capacitor, we can say that the positive delta function in the represented weighting function will come from

the contribution of the input voltage, while the negative exponential behaviour will come from the voltage across the capacitor. This means that the output at a certain time instant is determined by the input at the same time instant minus whatever we have previously stored in the capacitor (weighted by an exponential). It is important to remember that the time constant of this negative exponential must be smaller than the correlation time of the input noise, since we want to have correlated noise samples in order to perform an effective reduction of the noise.
From the weighting function represented in Figure 4.53, we can see that it is different from the one of an high-pass filter since we are decoupling the moment in which we are sampling the input signal (consisting in the pulse superimposed to the baseline) from the moment in which we are storing the baseline (that is actually the noise).

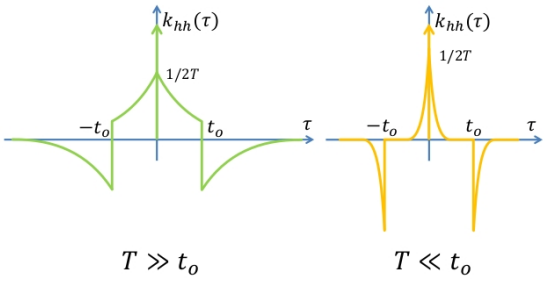


Figure 4.54: Autocorrelation of the weighting function of a baseline restorer in two limiting cases.

By definition, then, the autocorrelation of the weighting function of the baseline restorer can be written as:

$$k_{hh}(\tau) = \int w(t,\gamma)w(t,\gamma+\tau)\ d\gamma.$$

In the time $\tau = 0$, then, the two delta-functions of the weighting function are overlapped and, therefore, we obtain a delta-function plus the integral of the two exponentials. For $\tau > 0$, the two delta-functions are no longer overlapped, therefore we have to calculate only the correlation between the two negative exponentials, that is again a decreasing exponential. If we now assume that the temporal separation between the delta-function and the exponential in the weighting function, that is defined as t_0, is much larger than the time constant T of the filter:

$$t_0 \gg T$$

then this exponential in the autocorrelation will completely decay before having an overlapping between one of the two delta-functions and the other exponential, making the autocorrelation equal to zero for a certain interval. At some point, then, we will have that one of these delta-functions will be overlapping to the decaying exponential, determining a negative exponential contribution to the autocorrelation.
In the other limiting case, the distance between the delta-function and the exponential is shorter than the time constant of the filter:

$$t_0 \ll T$$

and thus the previously described sampling of the negative exponential will take place before the positive exponential decay is completed.
To evaluate the effect of this filter on the noise, we can write the mean square value of the output noise as:

$$\overline{n_y^2} = \int R_{xx}(\tau) k_{hh}(\tau)\, d\tau.$$

In the case of an high-frequency noise, the mean square value of the output noise $\overline{n_y^2}$ will be even larger than the mean square value of the input noise $\overline{n_x^2}$, since we are subtracting uncorrelated samples, thus adding their variances. In this case, in fact, we are near to the delta-function that is present in $k_{hh}(0)$, giving a positive contribution to the output noise.
On the other hand, if we consider a low-frequency noise, for which the correlation time is much longer than the distance between the delta-function and the negative exponential in the weighting function:

$$T_n \gg t_0$$

then the negative exponential parts of the autocorrelation of the weighting function will reduce the mean square value of the output noise with respect to the one of the input noise, making this filter effective. This exclusively due to the fact that the negative exponential tails of the autocorrelation of the weighting function come into play.

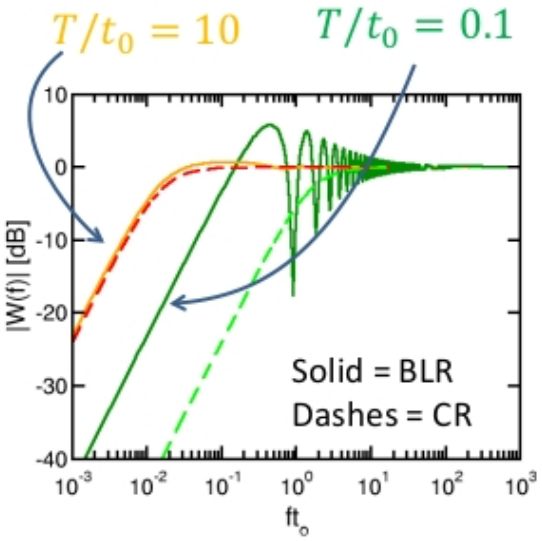


Figure 4.55: Fourier transform of the weighting function for a baseline restorer.

In the frequency domain, the Fourier transform of the weighting function for a baseline restorer can be represented as in Figure 4.55. In this case, setting for the sake of simplicity the time instant of interest equal to zero:

$$t = 0$$

from the time-reversal property we can obtain the following transfer function:

$$W(s) = 1 - \frac{e^{-st_0}}{1 - sT}$$

where the first, constant contribution will come from the positive delta-function, while the second contribution will come from the shifted and time-reversed contribution of the negative exponential. At low frequencies, the exponential that is present in this expression can be simplified as:

$$e^{-st_0} \simeq 1 - st_0$$

thus obtaining the following low-frequency approximation of the filter:

$$W(f) \simeq -j2\pi f(T - t_0).$$

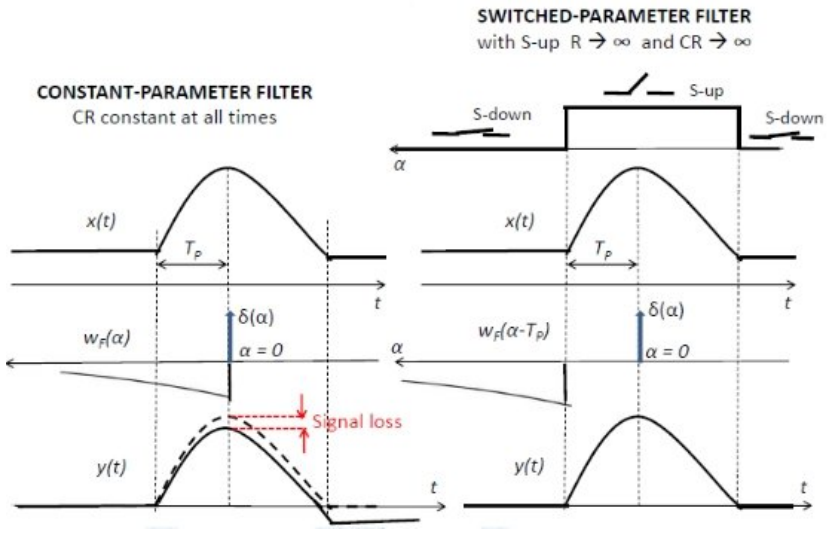


Figure 4.56: Effect on a signal of a high-pass filter (on the left) and of a baseline restorer (on the right).

In Figure 4.56, it is possible to observe a comparison between the behaviour of an high-pass filter, that will determine a little negative overshoot of the signal due to the negative discharge of the filter, and the one of a baseline restorer, that works perfectly in this case of constant baseline. In all real cases, even with DC coupled electronics, the weighting function is generally not extended to zero, because an intrinsic high-pass filter is present in any operation due to the fact that this operation started at some time and from a zero value. This is the reason why we always have to manually set to zero the baseline.

A different type of filter is the correlated double sampling. In this device, we have decided to sample the noise in one point and, then, to subtract it from the sample that we have taken considering the signal and the noise. The weighting function, therefore, will just consist in two delta-functions, one centred in the time in which we have to sample the signal and the noise and the other, with negative amplitude, when we want to sample just the noise. This methods holds if the noise is strongly correlated (thus being a low-frequency or constant noise) and it is generally applicated, for example, in digital cameras. The weighting function, then, can be written as:

$$w(0, \tau) = \delta(\tau) - \delta(\tau + t_S)$$

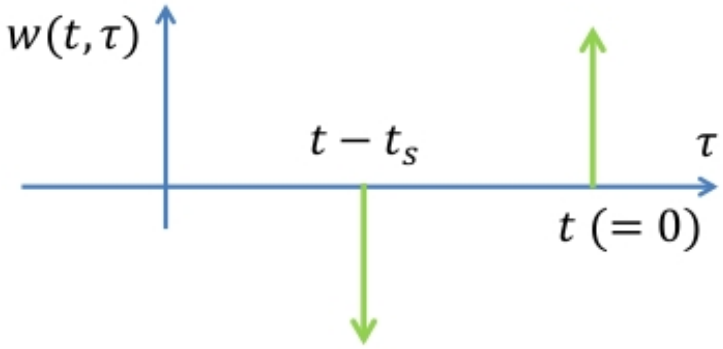


Figure 4.57: A correlated double sampling.

thus its Fourier transform, in the frequency domain, will be:

$$W(s) = 1 - e^{st_S} \;\rightarrow\; W(t,f) = 1 - e^{j2\pi f t_S} = 1 - \cos(2\pi f t_S) - j\sin(2\pi f t_S).$$

This allows us to calculate the square modulus of the transform of the weighting function as:

$$|W(t,f)|^2 = [1 - \cos(2\pi f t_S)]^2 + \sin^2(2\pi f t_S) = 2\,[1 - \cos(2\pi f t_S)]\,.$$

In the low-frequency limit:

$$|W(f)| \simeq 2\pi f t_S \;\Rightarrow\; |W(t,0)|^2 \simeq 0.$$

4.10 Amplitude modulation (AM) and synchronous detection

In this description of the types of noise and the solutions we can adopt to fight it, only one case is missing: the one of a low-frequency incoming signal superimposed to a low-frequency noise. The idea, in this case, is to move the frequency of the signal far from the region of frequencies in which we have the noise and, then, use a filter to improve the signal-to-noise ratio.
If the signal is at the DC level and it is buried into a low-frequency noise, in fact, an high-pass or a band-pass filter become useless. From the field of telecommunications, however, we can consider that if we are able to move the spectrum of the signal to higher frequencies, with an operation that is called modulation (or, for the dual operation, demodulation), the signal-to-noise ratio would improve. An high quality band-pass filter, then, can be used to recover the signal once it has been moved to another frequency region.

The first example of modulation that we can study is the amplitude modulation (AM), that is schematically represented in Figure 4.58. In this case, the amplitude of a carrier wave $c(t)$ is modified using a modulating signal $x(t)$. The resulting amplitude of the modulated signal is therefore the product between the amplitudes of the two incoming signals:

$$m(t) = x(t)c(t)$$

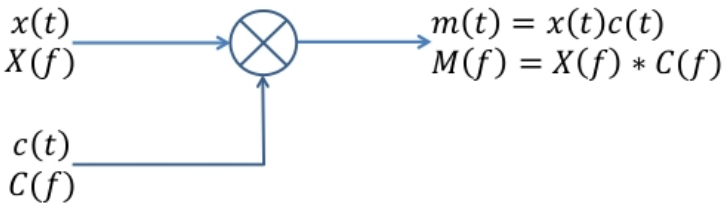


Figure 4.58: Amplitude modulation.

therefore in the frequency domain the resulting signal will be the convolution of the two incoming ones:

$$M(f) = X(f) * C(f).$$

This method was originally developed for telephone and radio communications, since transmitting signal is much simpler if they are moved at higher frequency regions. In our description, we will consider only sinusoidal carriers $c(t)$.

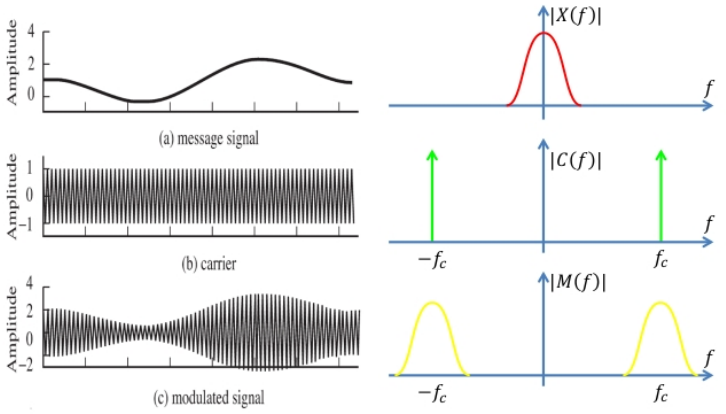


Figure 4.59: Amplitude modulation both in the time and in the frequency domain.

Given therefore a generic (and generally slowly changing) signal waveform and a sinusoidal carrier at much higher frequency, we can obtain a signal in the time domain that can be represented as in Figure 4.59. In the frequency domain, the signal will have a certain bilateral Fourier transform that, being a slowly changing signal, will be mainly concentrated in the low-frequency region and it will have a certain width, while the Fourier transform of the carrier, being a perfectly sinusoidal wave, will be only two delta functions centred at the corresponding frequency. The modulated signal, being the convolution of the two original signals in the frequency domain, will be equal to the spectrum of the original signal but now centred at frequency $\pm f_c$, where f_c is the frequency of the carrier.
We can thus write the sinusoidal carrier in the time domain as:

$$c(t) = A\cos(\omega_c t + \phi_c)$$

while in the frequency domain it will be:

$$C(f) = \frac{A}{2}\left[e^{j\phi_c}\delta(f - f_c) + e^{-j\phi_c}\delta(f + f_c)\right].$$

Given therefore $x(t)$ to be the slowly varying signal and $X(f)$ to be its spectrum, we can write the modulated signal as, in the time domain:

$$m(t) = x(t)c(t)$$

while in the frequency domain it will be:

$$M(f) = \frac{A}{2}\left[e^{j\phi_c}X(f - f_c) + e^{-j\phi_c}X(f + f_c)\right].$$

This property can be clearly understood considering that a multiplication in the time domain corresponds to a convolution in the frequency domain. The carrier therefore will have, in the frequency domain, some complex exponential terms that will be related to the phase of this oscillation while the convolution of the signal with the carrier will determine, in the frequency domain, a shift in the spectrum of the original signal.

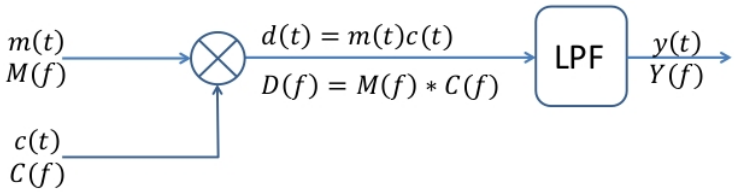


Figure 4.60: Demodulation of a signal.

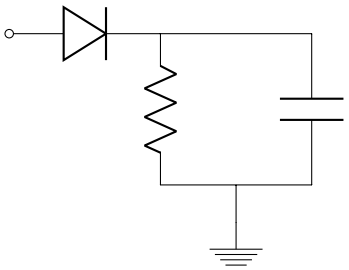


Figure 4.61: Circuit originally used for an incoherent detection.

The inverse operation, needed for recovering the signal in its original bandwidth, is called demodulation. From an historical point of view, this was originally done in a process called incoherent detection with a circuit that is represented in Figure 4.61. This method used a diode and a low-pass filter to remove the harmonics coming from the modulation of the signal. However, this method was quite inaccurate, therefore a coherent approach exactly equal to the modulation process has been developed. We can thus, as represented in Figure 4.60,

multiply the modulated signal $m(t)$ by another reference signal completely identical to the carrier $c(t)$, thus obtaining a demodulated signal:

$$\begin{aligned} d(t) &= m(t)c(t) = x(t)c^2(t) = x(t)A^2\cos^2(\omega_c t + \phi_c) = \\ &= A^2 x(t)\left[\frac{1+\cos(2\omega_c t + 2\phi_c)}{2}\right] = \\ &= \frac{A^2}{2}x(t) + \frac{A^2}{2}x(t)\cos(2\omega_c t + 2\phi_c) \end{aligned}$$

to which we can apply a low-pass filter. This filtering operation will obviously cancel the residual oscillation at twice the frequency of the carrier, thus giving:

$$y(t) = \frac{A^2}{2}x(t)$$

that is clearly proportional to the original signal. In the frequency domain:

$$D(f) = \frac{A^2}{2}X(f) + \frac{A^2}{4}\left[e^{j2\phi_c}X(f-2f_c) + e^{-j2\phi_c}X(f+2f_c)\right]$$

where we have two replicas of the original spectrum that are shifted in frequency and that can be cancelled using the low-pass filter, obtaining:

$$Y(f) = \frac{A^2}{2}X(f).$$

This procedure can be represented as in Figure 4.62.

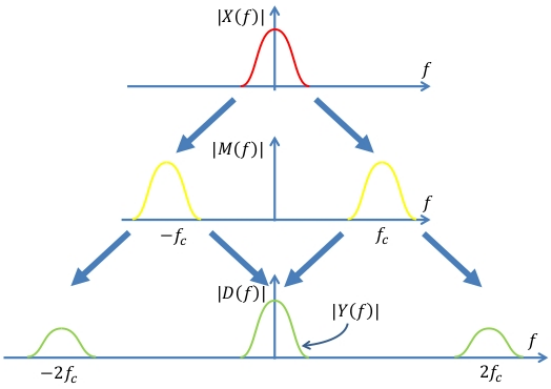


Figure 4.62: Demodulation operation.

It is important to notice that, in this description of the demodulation operation, we are assuming to be using exactly the same reference signal $c(t)$ that we used as a carrier in the modulation operation. What does it happen if this is not the case? We can now consider the demodulation operation with a slightly difference reference signal both in frequency and in phase:

$$w_R(t) = B\cos[(\omega_c + \Delta\omega)t + \phi_c + \Delta\phi].$$

Remembering the following property:

$$\cos(\alpha)\cos(\beta) = \frac{1}{2}\left[\cos(\alpha+\beta) + \cos(\alpha-\beta)\right]$$

we can write the demodulation signal in the time domain as:

$$d(t) = \frac{AB}{2} x(t) \left\{ \cos[(2\omega_c + \Delta\omega)t + 2\phi_c + \Delta\phi] + \cos(\Delta\omega t + \Delta\phi) \right\}$$

and we can immediately notice that the high-frequency oscillation will be filtered by the low-pass filter, thus giving:

$$y(t) = ABx(t) \frac{\cos(\Delta\omega t + \Delta\phi)}{2}.$$

We are thus obtaining two possible source of errors:

- a phase error, related to $\Delta\phi$, that can give a reduction of the amplitude of the signal from $AB/2$ to $AB\cos(\Delta\phi)/2$;
- a frequency error, related to $\Delta\omega$, that gives a residual oscillatory behaviour that, in the frequency domain, will shift the spectrum $X(f)$ symmetrically at higher (in modulus) frequencies.

Both these types of errors are extremely dangerous and, therefore, it is mandatory to reduce them as much as possible. The reference signal. therefore, must be locked both in frequency and in phase to the carrier, thus giving rise to the so called synchronous (or coherent) detection. In principle this is feasible, especially in labs, even though with some difficulties. In telecommunications, it is much more difficult if not impossible, even though some methods have been developed to retrieve the carrier from the signal using suitable techniques. Due to its properties, the demodulator is also called a phase-sensitive detector (PSD[12]) and it will consist in a multiplication stage followed by a low-pass filter.

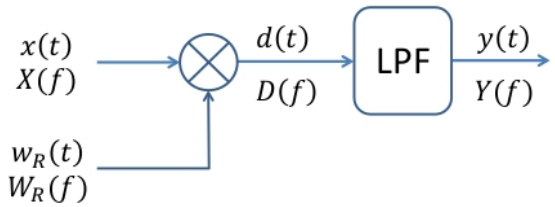


Figure 4.63: Demodulation operation and corresponding signals.

We want now to study in details this recovery stage using a weighting function approach, thus adopting the point of view of the signal. Given the scheme of a phase-sensitive detector as the one represented in Figure 4.63, we can write the signal at the output of the low-pass filter as the integral of the demodulation signal at its input multiplied by the weighting function of the filter:

$$y(t) = \int d(\tau) w_{LP}(t, \tau) \, d\tau.$$

At this level, notice that we are still considering a generic low-pass filter and the related, not further specified, weighting function. However, we know that

[12] This must not be confused with the power spectral density that we have studied before: they are totally different things.

the demodulation signal can be written as the product between the incoming modulated signal $x(t)$ (not to be confused with the slowly oscillating original signal, defined using the same variable, that we described in the modulation stage) and the reference signal:

$$d(t) = x(t)w_R(t)$$

thus obtaining:

$$y(t) = \int x(\tau)w_R(\tau)w_{LP}(t,\tau)\ d\tau = \int x(\tau)w(t,\tau)\ d\tau$$

where we have defined the weighting function of this phase-sensitive detector as:

$$w(t,\tau) = w_R(\tau)w_{LP}(t,\tau).$$

This is a direct consequence of the properties that we have studied for the weighting function of a generic filter. In the frequency domain, from Parseval's theorem, we can write:

$$y(t) = \int x(\tau)w(t,\tau)\ d\tau = \int X(f)W^*(t,f)\ df$$

and since the product in the time domain is the convolution in the frequency domain we have that:

$$W_R(f) = \mathcal{F}\{w_R(\tau)\}, \quad W_{LP}(t,f) = \mathcal{F}\{w_{LP}(t,\tau)\}$$

$$W(t,f) = W_R(f) * W_{LP}(t,f).$$

This weighting function and the associated Fourier transform can be represented as in Figure 4.64.

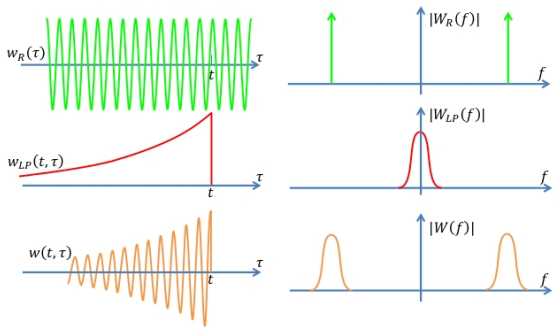


Figure 4.64: Weighting function of the phase-sensitive detector.

In the representation of this weighting function, we have assumed the weighting function of the low-pass filter to be equal to an exponential. We can immediately notice that even though the low-pass filter may be a linear and time-invariant filter, since the weighting function depends on the reference signal that depends on time this detector will act as a time-variant filter, thus giving, for a stationary noise at the input, a non-stationary noise at the output. In the

particular case of a phase-sensitive detector with a linear and time-invariant low-pass filter, in the frequency domain the output signal can be written as:

$$Y(f) = W_{LP}(t,f)D(f) = W_{LP}(t,f)\left[X(f) * W_R(f)\right].$$

Since the reference signal $w_R(t)$ is a periodic signal, then the modulus of its Fourier transform $|W(f)|$ will represent the frequency components that give significant contributions in the baseband.
We can now study the behaviour of a phase-sensitive detector as an optimum filter. In particular, we can take the case of a simple constant initial signal of amplitude A, for which the modulated signal at the input of the phase-sensitive detector will be:

$$x(t) = A\cos(\omega_c t).$$

In this case, if we assume the low-pass filter to be a linear and time-invariant integrator, since the delta-function response of this filter is a step with a certain gain K then the weighting function of this filter will be a time-reversed and shifted step with amplitude K:

$$w_{LP}(t,\tau) = Ku(t-\tau).$$

This means that the weighting function of the overall phase-sensitive detector will be[13]:

$$w(t,\tau) = K'u(t-\tau)\cdot\cos(\omega_R\tau)$$

but since we know the expression of the incoming signal, if the frequency of the carrier is equal to the frequency of the reference signal we have that:

$$\omega_c = \omega_R: \quad w(t,\tau) \propto x(\tau).$$

We are thus in a case of optimum filtering and, for this kind of signal, this phase-sensitive detector with a linear and time-invariant integrator is the optimum filter. In the general case, where the input signal is not constant and the low-pass filter is not a linear and time-invariant integrator, being instead a generic low-pass filter with an exponential weighting function[14], then the phase-sensitive detector will only be the quasi-optimum filter, however giving an higher flexibility.
For the case of a linear and time-invariant integrator, the output can be written as:

$$y(t) = K\int_{-\infty}^{t} x(\tau)w_R(\tau)\,d\tau.$$

We can immediately notice that, in the limit for $t\to\infty$, this would have been the cross-correlation between these two signals in zero, since they are not shifted one with respect to the other:

$$y(t) \simeq K\int_{-\infty}^{+\infty} x(\tau)w_R(\tau)\,d\tau = K_{xw_R}(0).$$

[13]Considering that, in principle, it could have also a different gain K'.

[14]Notice that if the time constant of this exponential is big enough, the exponential decay in the envelope of the weighting function is almost negligible with respect to the period of the oscillations and thus it is very similar to the one we have just described for the case of the linear and time-invariant integrator.

This is only an approximation since in reality this integral will end at t, that is clearly finite. However, from this we can understand the $y(t)$ can be seen as an estimate of the cross-correlation between the input and the reference signals and therefore the maximum output will be achieved when the two signals, that are oscillating, have the same frequency and the same phase. Moreover, if we assume now a certain noise $n_1(t)$ superimposed to the signal $x(t)$ and a certain noise $n_2(t)$ superimposed to the reference signal $w_R(t)$ (note that these will only be realizations of this stochastic process), the we get that:

$$y(t) \simeq K_{(x+n_1)(w_R+n_2)}(0) = K_{xw_R}(0) + K_{xn_2}(0) + K_{n_1w_R}(0) + K_{n_1n_2}(0)$$

and if each noise term is uncorrelated to the signal, to the reference signal and to the each other, then we get again:

$$y(t) \simeq K_{xw_R}(0)$$

thus actually having reduced the noise contribution thanks to this filter.
Assuming, for example, a simple RC filter[15], in this case we have:

$$y(t) = \frac{1}{T_F}\int_{-\infty}^{t} x(\tau)w_R(\tau)e^{-\frac{t-\tau}{T_F}}\,d\tau$$

that is again the cross-correlation $K_{xw_R}(0)$ estimated over a certain time T_F that is equal to the time constant of the filter.
The same discussion can be made also in the frequency domain, where we have that:

$$D(f) = \mathcal{F}\{x(t)w_R(t,\tau)\} = X(f) * W_R(f) = \int X(\nu)W_R(f-\nu)\,d\nu.$$

Assuming then that the output low-pass filter will select only the low-frequency components of this demodulated signal, that will be place around the origin of the frequency axis:

$$Y(f) \simeq D(0) = \int X(\nu)W_R(-\nu)\,d\nu$$

where we have also considered that:

$$W_R(-\nu) = W_R^*(\nu).$$

We have thus found, also in this case, the cross-correlation behaviour of the output signal and thus the maximum output signal will be obtained when the input signal and the reference signal are correlated.

We can now study the behaviour of the noise at the output of this phase-sensitive detector. Considering that a certain realization of the input noise will be multiplied by the reference signal, we will obtain a certain "demodulated" signal that is nothing but a noise realization enveloped in an oscillating function. This means that even if the input noise is stationary, the output noise of this device will be non-stationary, thus exhibiting a time-dependent behaviour. In particular, in this case, since we are enveloping this noise realization in an

[15]That is clearly not a linear and time-invariant integrator, thus having a different weighting function from the signal.

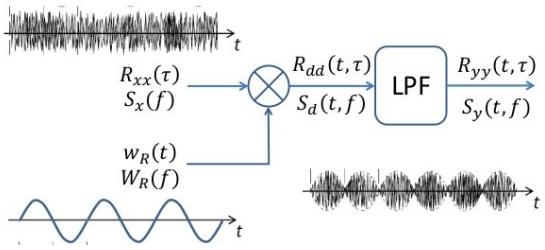


Figure 4.65: Behaviour of the noise in a phase-sensitive detector.

oscillating function, the behaviour of the autocorrelation of this noise will be periodic and thus this noise is called cyclostationary. The autocorrelation of the "demodulated" noise can be written, according to its definition, as the ensemble average of the "demodulated" noise signal on many realizations:

$$R_{dd}(t, t+\tau) = \overline{n_d(t) n_d(t+\tau)}.$$

However, from behaviour of the filter we are considering, we can say that this "demodulated" term will be equal to the product between the input noise (that is a stochastic process) and the reference signal (that is deterministic):

$$n_d(t) = n_x(t) \cdot w_R(t)$$

thus giving the fact that this ensemble average will act only on the input noise:

$$R_{dd}(t, t+\tau) = \overline{n_x(t) n_x(t+\tau)} w_R(t) w_R(t+\tau) = R_{xx}(\tau) w_R(t) w_R(t+\tau)$$

where in the last passage we have recognized the definition of the autocorrelation of the input noise. In particular, therefore, we can write the expression of the "demodulated" noise as:

$$\tau = 0: \quad \overline{n_d^2(t)} = \overline{n_x^2} w_R^2(t).$$

Since the output filter averages over many period of the reference signal[16], we can write the output autocorrelation, that thus will independent from the time t, as the time average of the autocorrelation of the "demodulated" noise:

$$R_{yy}(t, \tau) \simeq R_{yy}(\tau) \simeq \langle R_{dd}(t, t+\tau) \rangle = R_{xx}(\tau) \langle w_R(t) w_R(t+\tau) \rangle$$

where in the last equivalence we have considered that the autocorrelation of the input noise is independent from the time t. From the definition of temporal average, then, we can write that:

$$R_{yy}(\tau) \simeq R_{xx}(\tau) \lim_{T \to +\infty} \frac{1}{2T} \int_{-T}^{T} w_R(t) w_R(t+\tau)\, dt$$

and recognizing the autocorrelation of the reference signal, that is a power signal, we can write:

$$R_{yy}(\tau) \simeq R_{xx}(\tau) \cdot K_{w_R w_R}(\tau).$$

[16] In fact, being a low-pass filter that is used to get rid of the frequencies above twice the frequency of the carrier, it must be averaging over a time that is significantly larger than half a period of the carrier.

Considering now a sinusoidal reference signal and assuming that its phase is identically equal to zero:

$$w_R(t) = B\cos(\omega_c t)$$

we can write its autocorrelation as:

$$\begin{aligned} K_{w_R w_R}(\tau) &= \lim_{T\to+\infty} \frac{B^2}{2T}\int_{-T}^{T} \cos(\omega_c t)\cos(\omega_c(t+\tau))\ dt = \\ &= \lim_{T\to+\infty} \frac{B^2}{2T}\int_{-T}^{T} \cos(\omega_c t)\cdot[\cos(\omega_c t)\cos(\omega_c\tau) - \sin(\omega_c t)\sin(\omega_c\tau)] \end{aligned}$$

but noting that:

$$\int_{-T}^{T} \cos(x)\sin(x)\ dx = \int_{-T}^{T} \frac{\sin(2x)}{2}\ dx = 0$$

due to the symmetry of the sine function, the second of the previous integrals will vanish. This will thus give us:

$$\begin{aligned} K_{w_R w_R}(\tau) &= B^2\cos(\omega_c\tau)\lim_{T\to+\infty}\frac{1}{2T}\int_{-T}^{T}\cos^2(\omega_c t)\ dt = \\ &= B^2\cos(\omega_c\tau)\lim_{T\to+\infty}\frac{1}{2T}\int_{-T}^{T}\frac{1+\cos(2\omega_c t)}{2}\ dt = \\ &= \frac{B^2}{2}\cos(\omega_c\tau). \end{aligned}$$

The time correlation of a sinusoidal signal, therefore, is once again a sinusoidal signal.
In the frequency domain, since $w_R(t)$ is a periodic signal, thus being a power signal, the associated power spectral density $S_{w_R}(f)$ will be an average power spectral density, since we have cut the power signal, obtaining an energy signal between $-T$ and T:

$$w_R^T(t) = w_R(t)\cdot \mathrm{rect}(2T)$$

that will be then averaged for $T\to\infty$. In the frequency domain, therefore, this truncated signal will give[17]:

$$\begin{aligned} W_R^T(f) &= \frac{B}{2}[\delta(f-f_c)+\delta(f+f_c)] * 2T\,\mathrm{sinc}(2\pi f T) = \\ &= \frac{B}{2}\cdot 2T\cdot[\mathrm{sinc}(2\pi(f-f_c)T) + \mathrm{sinc}(2\pi(f+f_c)T)] \end{aligned}$$

and taking the square modulus, since the cross product will vanish in the following limit:

$$|W_R^T(f)|^2 = \frac{B^2}{4}\cdot(2T)^2\cdot\left[\mathrm{sinc}^2(2\pi(f-f_c)T) + \mathrm{sinc}^2(2\pi(f+f_c)T)\right]$$

[17] Considering that:

$$\mathcal{F}[\mathrm{rect}(2T)] = 2T\,\mathrm{sinc}(2\pi f T).$$

we obtain the following expression for the power spectral density[18]:

$$S_{w_R}(f) = \lim_{T\to+\infty} \frac{1}{2T}|W_R^T(f)|^2 = \frac{B^2}{4}\left[\delta(f-f_c)+\delta(f+f_c)\right]$$

that we could have directly obtained by Fourier transforming the expression of the autocorrelation of the reference signal:

$$S_{w_R}(f) = \mathcal{F}\{K_{w_R w_R}(\tau)\}.$$

We can thus can write:

$$R_{dd}(\tau) = R_{xx}(\tau)K_{w_R w_R}(\tau) \;\Rightarrow\; S_d(f) = S_x(f) * S_{w_R}(f).$$

From the autocorrelation of the reference signal or, alternatively, in the frequency domain, from the power spectral density of the reference signal we can write the autocorrelation of the output noise in the demodulation stage as:

$$R_{dd}(\tau) = R_{xx}(\tau)\frac{B^2}{2}\cos(\omega_c\tau)$$

while its power spectral density will be:

$$S_d(f) = \frac{B^2}{4}\left[S_x(f-f_c)+S_x(f+f_c)\right].$$

This means that, in this device, we are modulating the signal but we are also modulating the noise. We have now to take into account the presence of the low-pass filter.

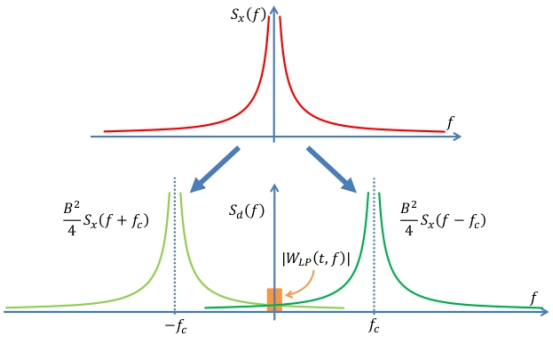


Figure 4.66: From the power spectral density of the input noise to the power spectral density in the demodulation stage: the shadowed (orange) region will be the only region that is allowed to pass in the low-pass filter.

[18]It is important to notice that in this limit the maximum amplitude of the sinc function is in $2T$, that will tend to infinity in the limit, while the first zero of the sinc function is in $1/2T$ that will tend to zero in this limit. This means that the sinc function, in this limit, is tending to a delta-function and we have to impose that the area below the sinc function:

$$\int_{-\infty}^{+\infty} \mathrm{sinc}^2(\pi f\alpha)\; df = \frac{1}{\alpha}$$

is equal to the amplitude of the delta-function.

The effect of the low-pass filter is represented, in the frequency domain, in Figure 4.66. It is important to notice that if we would have placed it directly on the input power spectral density $S_x(f)$, this would have been totally ineffective, since in that region the input power spectral density associated to the noise is quite high. On the other hand, placing it after the demodulation stage, we can obtain a significant reduction of the noise in our device.

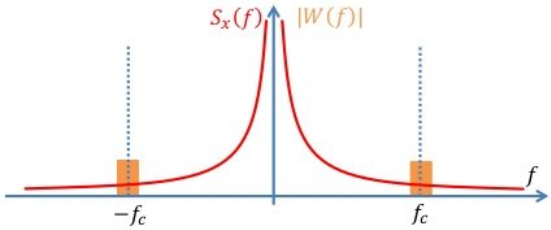


Figure 4.67: Effect of the low-pass filter on the input power spectral density.

From a different perspective, we can evaluate the regions of the input power spectral density that are sampled by the low-pass filter as in Figure 4.67. In this case, instead of moving the centre of the input power spectral density at $\pm f_c$, the demodulation stage is moving the central frequency of the low-pass filter to $\pm f_c$, thus demonstrating that we are sampling regions in which the power spectral density of the noise is much lower. These are, therefore, the only components of the input noise that will contribute to the output noise, stating an important meaning for the weighting function of the filter. The output noise, in fact, will be the correlation between the input noise and the weighting function. The only noise components that can contribute, therefore, will the ones that are correlated to the weighting function, thus being the ones at $\pm f_c$.
We can now consider the equivalent noise bandwidth of the low-pass filter to be extended in the range $[-BW_n, BW_n]$. Since this output low-pass filter has, in general, a quite narrow band, we can approximate the power spectral density in the demodulation stage with its value at zero frequency:

$$S_d(f) \simeq S_d(0) \in [-BW_n, BW_n].$$

Under these approximations, the mean square value of the output noise can be written as:

$$\begin{aligned}\overline{n_y^2} &= \int S_d(f)|W_{LP}(t,f)|^2\, df \simeq \int_{-BW_n}^{BW_n} S_d(0)\, df = \\ &= S_d(0) \cdot 2BW_n = 2BW_n \cdot \frac{B^2}{4}\left[S_x(f_c) - S_x(-f_c)\right] = \\ &= B^2 \cdot BW_n \cdot S_x(f_c)\end{aligned}$$

where, from the expression of $S_d(f)$, we considered that:

$$S_d(0) = \frac{B^2}{4}\left[S_x(f_c) + S_x(-f_c)\right]$$

and that the input power spectral density is an even function:

$$S_x(f_c) = S_x(-f_c).$$

We can thus write the signal-to-noise ratio for a constant input signal. This input signal will give a modulated signal that can be written as:

$$x(t) = A\cos(\omega_c t)$$

thus giving the following output signal (assuming to not have any phase error):

$$y(t) = \frac{AB}{2}$$

while the mean square value of the output noise will be:

$$\overline{n_y^2} = B^2 \cdot BW_n \cdot S_x(f_c).$$

From these expressions, we can write the signal-to-noise ratio as:

$$\frac{S}{N} = \frac{A}{\sqrt{4S_x(f_c)BW_n}}.$$

We can now compare this result with the one that we would have obtained directly applying, to the same signal with the same input noise, a low-pass filter:

$$\frac{S}{N} = \frac{A}{\sqrt{2S_x(0)BW_n}}.$$

Comparing these two expressions, we can see that the synchronous detection is useful if and only if the power spectral density at the carrier frequency is much lower than its DC value:

$$S_x(0) \gg S_x(f_c).$$

This is, for example, the case of the flicker noise, for which this kind of detection will be particularly effective. On the other hand, in the case of white noise, where the power spectral density is constant, it will be only worsening the situation. Moreover, it is important to consider that the modulation stage must be inserted before all the relevant low-frequency noise sources, that are generally represented by the amplifiers. If this is not the case, in fact, this kind of filtering is totally ineffective, since we will be shifting in frequencies with the modulation process both the signal and the noise. As we have already said, most of the low-frequency noise will come from the electronics (in particular, amplifiers), therefore we want to place the modulation stage as early as possible, in particular much earlier than every amplification stage.

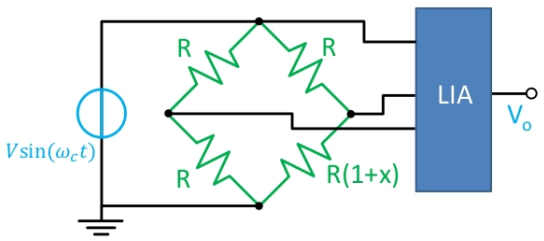


Figure 4.68: Wheatstone bridge with a lock-in amplifier.

These considerations can be applied to the case of a measure with a Wheatstone bridge connected to a lock-in amplifier, as in Figure 4.68. In this case, if

the modulation is done after the instrumentation amplifier, it is almost useless. Therefore, we need to move the modulation before the amplifier, putting it as early as possible in the acquisition chain. For example, since we can have a low-frequency noise also in the resistors that compose the Wheatstone bridge, we can modulate the voltage that is driving the bridge, as it is represented in the Figure.

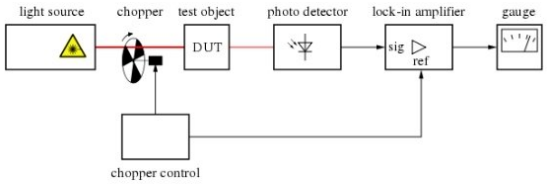


Figure 4.69: Low-light measurement with a lock-in amplifier.

Another possible measurement of this kind is represented in Figure 4.69. In this case, a continuous-wave laser can be used for fluorescence or spectroscopy experiments, impinging on a sample and then collecting the light coming from it with a photodetector. In this acquisition chain, most of the low-frequency noise will come from the amplifier and from the photodiode that are used for the acquisition. To modulate the signal as early as possible in this chain, we want to modulate the optical signal. However, it is not easy to directly modulate the optical signal by modulating the driving current of the laser, therefore we can do it with a chopper. A chopper is a sort of disc with empty sectors that is rotating. The light beam that is impinging on it will then be passed to the sample if we are in an empty space, while it will be blocked if we are in a solid one. Since this disc is rotating, on the sample is impinging a sort of square wave optical signal, whose sinusoidal components can be found by expanding it in a Fourier series: this will allow us to apply all the previous theoretical results.

4.11 Lock-in amplifiers

In the previous section, we have seen from a theoretical perspective the basic behaviour of these instruments, that are called lock-in amplifiers (LIAs). They are powerful tools that make us use a phase-sensitive detector to recover a weak modulated signal that is buried in the noise. However, several modifications are made to the basic scheme that we have studied in the previous section in order to reach high performances.

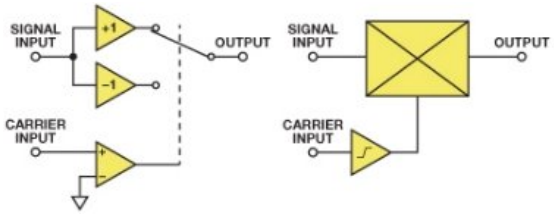


Figure 4.70: On the left, a switching demodulator; on the right, a generic demodulator.

From the previous theory, we know that the multiplication of the modulated signal with the reference signal can be performed using an analog multiplier. These devices, however, are in general complicated and expensive[19] and, since they are non-linear devices, generally introduce distortions. In real systems, therefore, modulators, demodulators and mixers are generally substituted by switching demodulators. In these devices, the input signal is periodically changed of sign (thus being multiplied by $+1$ or -1) depending on an input reference signal that controls a voltage controlled switch. This gives, as an example, the modulation of the input signal with a square wave of frequency f_c that depends on the driving pattern of the switch. These new devices are better because they are simpler than analog multipliers, they are less expensive and they are more precise. In particular, they allow us to neglect almost any noise or amplitude variation of the carrier signal, since it will only control a switch. The noise, in this device, will be then mainly introduced by the switches and it will usually be lower than the one we would have had in multipliers. Last, the fact that they are less expensive is strongly related to the fact that they are easier to design and to manufacture. The only possible drawback is that they will work only with a square wave modulation: does this change significantly the theory that we have already studied?

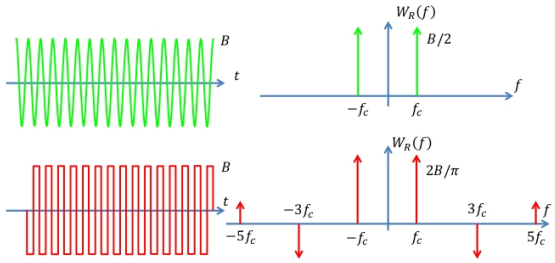


Figure 4.71: Above, a sinusoidal reference signal; below, a square wave one.

To do this, we can consider parallelly the case of a sinusoidal reference signal and of a square wave reference signal that is represented in Figure 4.71. While in the case of a sinusoidal signal we know that we will obtain two delta-functions as a Fourier transform, in the case of a square wave function we can calculate that:

$$W_R(f) = \frac{2B}{\pi} \sum_{k=0}^{\infty} \frac{(-1)^k}{2k+1} \left[\delta(f-(2k+1)f_c) + \delta(f+(2k+1)f_c)\right].$$

It is immediate to notice that we have obtained only odd harmonics of the frequency of the square wave f_c, with a changing sign decaying amplitude. This can be immediately understood if we consider that each cosine function that is represent in the Fourier series of the square wave considered will give rise, in the frequency domain, to a pair of delta-functions.

Assuming now that a sinusoidally modulated constant signal is coming to the previously described demodulation stage, in the case of the sinusoidal reference we can write, from the previous theory, the associated demodulated signal

[19] They are generally realized taking the logarithm of both the signals, adding them together and then calculating the exponential of the obtained signal; this is clearly a non-linear scheme.

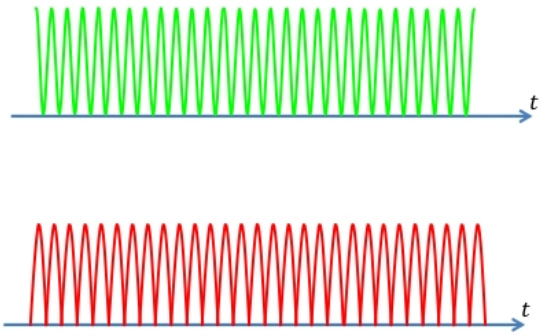


Figure 4.72: Above, a sinusoidal demodulated signal; below, a square wave one.

(before the application of the low-pass filter) as:

$$d(t) = AB\cos^2(\omega_c t)$$

while in the case of of the square wave reference signal it is possible to demonstrate that we will obtain the following demodulated signal:

$$d(t) = AB|\cos(\omega_c t)|$$

as represented in Figure 4.72. In fact, the multiplication of a sinusoidal signal with a square wave will give contributions only at the frequency ω_c of the carrier, while we will obtain contributions also for any other harmonic in the case of the noise. This means that only two "windows", in the frequency domain, are passing the signal, while many other (each pair corresponding to a different harmonic) will be demodulating the noise with decreasing weights.

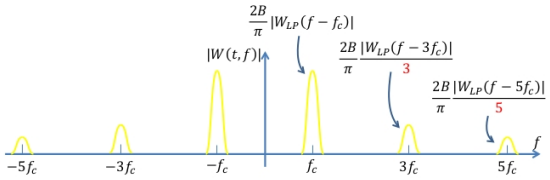


Figure 4.73: Spectral response of a switching phase-sensitive detector.

The spectral response of a switching phase-sensitive detector can thus be represented as in Figure 4.73, where we can clearly observe the "windows" through which the input noise is passing. This spectral response can thus be written, from a theoretical point of view, as:

$$W(t,f) = W_R(f) * W_{LP}(f).$$

From the expression of the reference signal in the time domain, that can be written as the Fourier series, with suitable coefficients[20] B_k, of the square wave:

$$w_R(t) = \sum_{k=0}^{+\infty} 2B_k \cos(k\omega_c t)$$

[20]These coefficients can be obviously algebraically computed, but this is not really important in this theoretical reasoning.

we can transform it, obtaining the Fourier transform of the reference signal:

$$W_R(f) = \sum_{k=0}^{+\infty} B_k \left[\delta(f - kf_c) + \delta(f + kf_c)\right].$$

In an analogous way, we can calculate the autocorrelation of this reference signal, that being a sum of sinusoidal terms will be again a sum of sinusoidal terms without any cross-product:

$$K_{w_R w_R}(\tau) = \sum_{k=0}^{+\infty} 2B_k^2 \cos(k\omega_c t)$$

and calculating the Fourier transform of this quantity we obtain the power spectral density of the reference signal:

$$S_{w_R}(f) = \sum_{k=0}^{+\infty} B_k^2 \left[\delta(f - kf_c) + \delta(f + kf_c)\right].$$

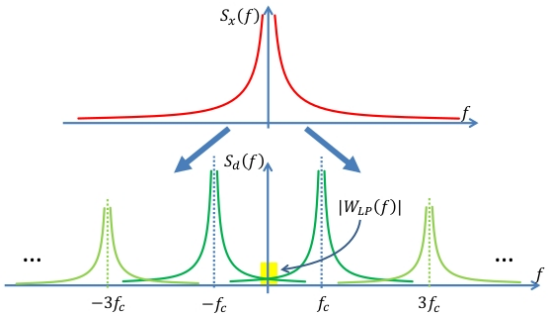


Figure 4.74: Power spectral density of an input flicker noise and corresponding power spectral density of the noise in the demodulated signal; the shadowed (yellow) region is the bandwidth of the low-pass filter.

The power spectral density of the demodulated signal (to which will be then applied the low-pass filter) can thus be written as:

$$S_d(f) = S_x(f) * S_{w_R}(f)$$

and it is represented, in the case of the flicker noise, as in Figure 4.74. It will consist in many different replicas of the flicker noise power spectral density at the output of the multiplication stage in the demodulator. Since the output filter, then, will collect noise from all the replicas, we can write the mean square value of the output noise as:

$$\overline{n_y^2} = \int S_d(f)|W_{LP}(t,f)|^2 \, df$$

and assuming the noise bandwidth of the output low-pass filter $[-BW_n, BW_n]$ to be particularly small and thus the power spectral density of the demodulated signal to be constant over this bandwidth, we can write:

$$\overline{n_y^2} \simeq S_d(0) \cdot 2BW_n.$$

From the expression of the power spectral density of the demodulated signal, then, we can write:

$$\overline{n_y^2} = 2BW_n \sum_{k=0}^{+\infty} B_k^2 \cdot [S_x(-kf_c) + S_x(kf_c)] = 4BW_n \sum_{k=0}^{+\infty} B_k^2 S_x(kf_c) =$$
$$= BW_n \cdot \sum_{k=0}^{+\infty} (2B_k)^2 S_x(kf_c)$$

where we can recognize that $2B_k$ is the amplitude of the various sinusoidal components of the reference signal. From this last observation, we can see that this result is consistent with the one that we have obtained for the case of a sinusoidal modulation and a sinusoidal reference signal.

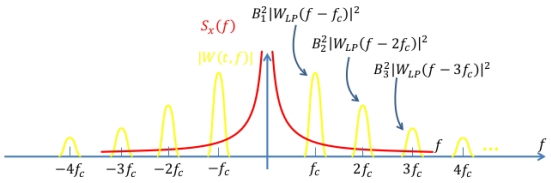


Figure 4.75: Filtering of the input (flicker noise) power spectral density in the frequency domain.

This kind of filtering can be represented as in Figure 4.75. Considering therefore the various "windows" through which is "leaking" the noise, we obtain that the mean square value of the noise can be written as:

$$\overline{n_y^2} = 2 \sum_{k=0}^{+\infty} B_k^2 \cdot 2BW_n \cdot S_x(kf_c).$$

We can notice that in the Figure we have represented the spectral response of the filter that will only give contributions, in frequency, to the mean square value of the output noise.
Considering now a sinusoidally modulated constant signal:

$$x(t) = A\cos(\omega_c t)$$

we can try to calculate the associated signal-to-noise ratio. Assuming that we do not have any phase error between the carrier and the reference signal, the output signal[21] can be written as:

$$y(t) = A\frac{2B}{\pi}$$

and the mean square value of the output noise will be:

$$\overline{n_y^2} = 4BW_n \left(\frac{2B}{\pi}\right)^2 \cdot \sum_{k=1,3,5,\ldots} \frac{S_x(kf_c)}{k^2}$$

[21] By considering that:

$$\langle \cos^2(\omega_c t) \rangle = \frac{1}{2}.$$

where we have considered that, in this last expression, only odd indices are relevant, while the other will have a zero amplitude in the Fourier series. Notice that the quantity of noise collected in this device, that will affect the signal-to-noise ratio, depends on the value of this sum. Calculating then the signal-to-noise ratio, we obtain:

$$\frac{S}{N} = \frac{A}{\sqrt{4BW_n}\sum_{k=1,3,5,\ldots}\frac{S_x(kf_c)}{k^2}}.$$

We can now study in details a few cases, depending on the value of the noise and on the type of modulation we are considering. Remembering the right hand-side graph in Figure 3.29 at page 181, if we are modulating the signal beyond the noise corner frequency, then the input noise power spectral density $S_x(kf_c)$ is white and we can simply evaluate the previous sum as:

$$\sum_{n=0}^{+\infty}\frac{1}{(2n+1)^2} = \frac{\pi^2}{8} \simeq 1.2337$$

from a well-known result on the series. This means that in the case of a sinusoidal carrier and a square wave demodulation, we can write the signal-to-noise ratio as:

$$\left(\frac{S}{N}\right)_{sqd} = \frac{1}{\sqrt{\pi^2/8}}\left(\frac{S}{N}\right)_{sin} = \frac{1}{\sqrt{1.2337}}\left(\frac{S}{N}\right)_{sin} = \frac{1}{1.11}\left(\frac{S}{N}\right)_{sin}$$

thus being clearly lower (even if only for a small factor) than the case of a fully sinusoidal modulation and demodulation. Assuming, instead of a white noise, a fully flicker spectrum (thus with a pure $1/f$ dependency for all the frequencies) this correction factor to be put at the denominator can be calculated to be 1.026.

Alternatively, we can consider the case of a square wave carrier that is demodulated with a square wave reference. In this case, the amplitude of the signal is simply constant and equal to AB and thus the signal-to-noise ratio in this case can be written as:

$$\left(\frac{S}{N}\right)_{fsq} = \frac{\pi/2}{\sqrt{\pi^2/8}}\left(\frac{S}{N}\right)_{sin} = \sqrt{2}\left(\frac{S}{N}\right)_{sin} \simeq 1.41\left(\frac{S}{N}\right)_{sin}.$$

Last, we can consider the case of a square wave carrier demodulated with a sinusoidal reference signal, in this case obtaining, from the evaluation of the signal-to-noise ratio[22]:

$$\left(\frac{S}{N}\right)_{sqs} = \frac{4}{\pi}\left(\frac{S}{N}\right)_{sin} \simeq 1.27\left(\frac{S}{N}\right)_{sin}.$$

A few issues, however, are possible for a switching phase-sensitive detector. First of all, it is possible that the average value of the reference signal is different from zero:

$$\langle w_R(t)\rangle \neq 0.$$

This will add, in the frequency domain, another "window" centred in $f = 0$, thus making the lock-in amplifier to collect also the low-frequency noise thus wasting

[22] The willing student can try to demonstrate this relationship.

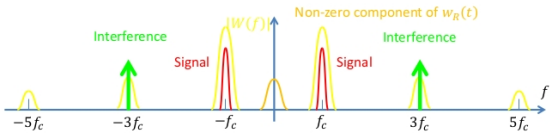


Figure 4.76: Issues concerning the switching phase-sensitive detector.

all the efforts for changing the bandwidth of the signal from the bandwidth of the noise (especially in the case of a sinusoidal signal). Moreover, we know that every noise term that will be placed in the "windows" will then be brought back, by the demodulation process, in the low-frequency band and will then pass through the low-pass filter. This is particularly dangerous when we have an interference that is placed at one of the harmonics of the carrier frequency f_c. To counteract this effect, in general we choose a carrier frequency that is strange enough to exclude all the interferences that may be coming from the harmonics of the more common signals (for example, residuals of the rectification of the AC current in common supply networks).

4.11.1 Analog LIAs

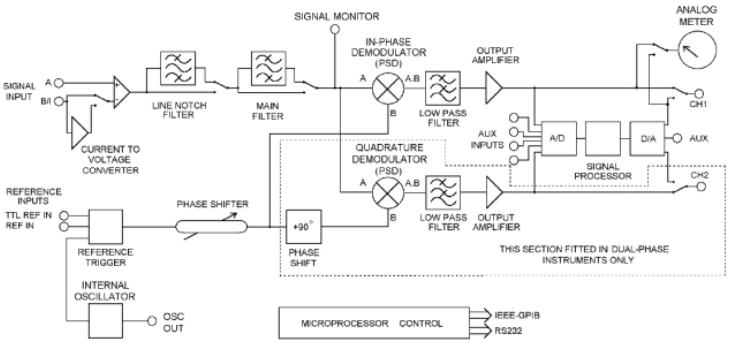


Figure 4.77: Structure of an analog lock-in amplifier.

The structure of an analog lock-in amplifier is represented in Figure 4.77. First of all, we can observe that we have two filters (one is called "line notch filter", the other "main filter") placed on the incoming signal. These filters are used to remove part of the noise that is at a bandwidth different from the one of the signal, thus cleaning as much as possible the incoming signal. Moreover, they are used since the following circuit has, as any real circuit, a limited dynamic and we want to avoid to overdrive it. These filters will be effective also in the case of a square wave modulation. The notch filter will remove the components at 50/60 Hz and/or its second harmonics, while the main filter can in general be either a band-pass filter (cleaning the signal and improving the dynamic), a low-pass filter (in the case of low-frequency carriers) or even may be absent (in non-demanding applications).
The second element that we can analyse is the reference trigger and the internal oscillator. In fact, as we have seen before, if the sinusoidal reference has a non-

zero offset, we are actually demodulating also a noise term at low-frequency, thus making this amplifier ineffective. We want thus to remove any constant offset from the reference signal by using a suitable filter called reference trigger. Moreover, it is possible to use either an external reference (that will be provided to the instrument) or an internal one (that will be coming from a local, or internal, oscillator). The reference trigger will be responsible of generating a trigger signal, for example a square wave signal, that is synchronous with the reference or at twice the frequency of the carrier.
Again in the line that is providing the reference signal to the demodulation stage, we can observe the presence of a phase shifter. This element is needed to make the phase of the reference signal perfectly matched to the one of the carrier, thus increasing the amplitude of the output signal. In modern systems, called dual-phase instruments, this is done parallelly demodulating in two different ways the incoming signal, one with a reference with a certain phase and the other with a reference that is in quadrature (thus having a $\pi/2$ shift) to the first one; from these two signals it is then possible to obtain the correct amplitude of the incoming signal. Alternatively, the phase shifter can be used to adjust the phase of the carrier signal in order to obtain the maximum output amplitude.
Then, the phase-sensitive detector is in general a square-wave mixer, but more refined implementations (for example suppressing the third harmonic response) may be employed.
The last element we can further analyse is the low-pass filter (or the two low-pass filters, in the case of dual-phase instruments). In general, either first order or second order filters can be used. Second order filters, having two poles, are obviously more selective, having an abrupt cut-off frequency. However, in some applications in which a lock-in amplifier is used in a negative feedback system they will be adding a pole, possibly giving stability issues to the whole network. At the end, an output amplifier will provide a simple gain, increasing the sensitivity. Since this last amplifier is DC-coupled, its drift and noise must be controlled.
Last, we can study the main parameters of this device. The frequency range at which it is operating is typically between a fraction of hertz to a few MHz (and up to 200 MHz for radio-frequency lock-in amplifiers). The time constant of the output low-pass filter will be between a few milliseconds and a few hundreds of seconds. Last, we define the dynamic reserve the ratio between the largest "tolerable" noise signal and the full scale signal, expressed in decibels. Typical values for this quantity are lower than or equal to 60 dB, thus meaning that at the input of the lock-in amplifier we can have a noise that is up to 100 times larger than the incoming signal.

4.11.2 Digital LIAs

The digital[23] implementation of a phase-sensitive detector, that is represented in Figure 4.78, is actually quite complicated. In this case, the reference signals are pre-computed with an higher precision, thus giving a better performance of the device, in particular with respect to offsets and harmonics. In particular, they can reach a dynamic reserve that is up to 100 dB, allowing also much lower frequencies for the carriers (up to some mHz). Moreover, their design is more

[23] This part of the program has not been presented during lectures.

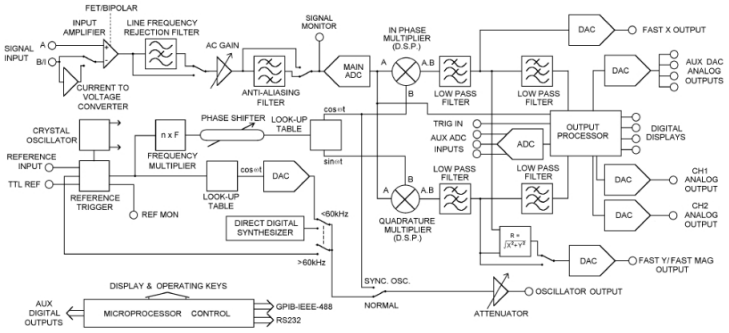


Figure 4.78: Structure of a digital lock-in amplifier.

flexible and not having an output amplifier they are also more stable.
It seems, therefore, that digital lock-in amplifiers are far better than analog ones. However, analog lock-in amplifiers are still used in some cases. For example, they can deal with the demodulation at high frequencies (up to the MHz regime), since the frequency of the carrier is limited to a few MHz in digital devices. Then, they are suitable for all those applications that require short time constants and mid-range frequencies (in the order of 100 kHz) and, also, in feedback loops.

Chapter 5

Exercises

5.1 Review of the Laplace transform, linear circuits and Bode plots

5.1.1 The Laplace transform and its properties

Given a certain signal $f(t)$, we can define its Laplace transform $F(s)$ as:

$$F(s) = \mathcal{L}\left(f(t)\right) = \int_0^{+\infty} f(t) \cdot e^{-st}\, dt$$

where passing a function from the time domain to the frequency domain we can introduce the Laplace operator s, that has the dimension of the inverse of a time:

$$s = [s^{-1}].$$

We can immediately understand that, from the linearity of the integral, the Laplace transform is linear:

$$\mathcal{L}\left(\alpha f(t) + \beta g(t)\right) = \alpha F(s) + \beta G(s)$$

where:

$$\mathcal{L}(f(t)) = F(s), \quad \mathcal{L}(g(t)) = G(s).$$

We can observe that, since the time integral goes from zero to $+\infty$, we are implicitly assuming that before time $t = 0$ the signal is identically zero, thus starting only in the origin of the temporal axis. The advantage of the introduction of the Laplace transform is represented by its properties:

1. time differentiation[1]:

$$\mathcal{L}\left(\frac{d}{dt} f(t)\right) = sF(s) - f(0);$$

[1] Consider for example an exponentially decaying signal starting at time $t = 0$. In this kind of signal, we obviously have a discontinuity at time $t = 0$. Therefore, what is the correct value of the $f(0)$ term we need to use in this formula? It depends on the meaning that we give to the derivative. In the distribution sense, if $u(t)$ is the step function:

$$\frac{d}{dt} u(t)\Big|_{t=0} = \delta(t)$$

and therefore we have that:

$$f(0) = f(0^-) = 0.$$

2. frequency differentiation:

$$(-1)^n \cdot \frac{d^n F}{ds^n} = \mathcal{L}\left(t^n \cdot f(t)\right)$$

and applying it to the first derivative:

$$-\frac{dF}{ds} = \mathcal{L}\left(t \cdot f(t)\right);$$

3. integration in the time domain:

$$\mathcal{L}\left(\int_0^t f(\tau)\ d\tau\right) = \frac{F(s)}{s}$$

and this makes clearer that the advantage connected to the usage of the Laplace transform and the frequency domain is that, in this domain, the differential operations (derivatives and integrals) become algebraic operations;

4. integration in the Laplace domain:

$$\mathcal{L}\left(\frac{f(t)}{t}\right) = \int_s^{+\infty} F(\sigma)\ d\sigma;$$

5. time-shift property[2]:

$$\mathcal{L}\left(f(t-T)\right) = e^{-sT} F(s)$$

and therefore any shift in the time domain adds an exponential term in the frequency domain;

6. frequency-shift property:

$$\mathcal{L}\left(e^{at} f(t)\right) = F(s-a);$$

7. scaling:

$$\mathcal{L}\left(f(at)\right) = \frac{1}{|a|} F\left(\frac{s}{a}\right)$$

and therefore a broadening in the time domain corresponds to a narrowing in the frequency domain and vice versa; if $a = -1$ we obtain the time reversal property:

$$\mathcal{L}\left(f(-t)\right) = F(-s);$$

On the other hand, in the classical sense the distribution of the step function in the origin of time does not exists and we can use as a value in the origin the one from the positive side:

$$f(0) = f(0^+).$$

In general, we will consider the distribution sense.

[2] This is especially useful when we are dealing with signals that do not start at time $t = 0$.

8. convolution:
$$\mathcal{L}\left(f(t) * g(t)\right) = F(s) \cdot G(s)$$
where the convolution is defined as:
$$f(t) * g(t) = \int_{-\infty}^{+\infty} f(\tau) g(t-\tau)\ d\tau$$
while the dual relationship of convolution in the frequency domain is in general not used;

9. initial value theorem:
$$f(0^+) = \lim_{s \to +\infty} sF(s);$$

10. final value theorem:
$$\lim_{t \to +\infty} f(t) = \lim_{s \to 0} sF(s).$$

5.1.2 A few elementary signals

We will make an extensive use of the following signals and of their Fourier transforms, therefore it is worth to recall them:

- Dirac delta function:
$$\delta(t) = 0\ \forall t \neq 0, \int_{-\infty}^{+\infty} \delta(t)\ dt = 1\ \Rightarrow\ F(s) = \int_{0}^{+\infty} \delta(t) e^{-st}\ dt = 1;$$

- step function:
$$u(t) = \begin{cases} 1, & t \geq 0 \\ 0, & t < 0 \end{cases} \Rightarrow\ \mathcal{L}(u(t)) = \mathcal{L}\left(\int_{-\infty}^{+\infty} \delta(t)\ dt\right) = \frac{1}{s}\mathcal{L}(\delta(t)) = \frac{1}{s};$$

- ramp function:
$$f(t) = \begin{cases} 0, & t < 0 \\ t, & t \geq 0 \end{cases} \Rightarrow\ \mathcal{L}\left(f(t)\right) = \frac{1}{s^2};$$

- rectangular function: it can be seen as the subtraction between two different step function, one centred in the origin and the other centred at T, therefore from the linearity of the Laplace transform:
$$f(t) = \begin{cases} 1, & 0 \leq t \leq T \\ 0, & \text{elsewhere} \end{cases}$$
$$\Rightarrow\ \mathcal{L}(f(t)) = \mathcal{L}(u(t) - u(t-T)) = \frac{1}{s} - \frac{1}{s}e^{-sT} = \frac{1}{s}\left(1 - e^{-sT}\right);$$

- decreasing exponential function: observing that the presence of an exponential in the time domain is equivalent to a shift in the frequency domain:
$$f(t) = e^{-\frac{t}{\tau}} u(t)\ \Rightarrow\ \mathcal{L}\left(f(t)\right) = \frac{1}{s + \frac{1}{\tau}} = \frac{\tau}{1 + s\tau};$$

- sine signal:
$$f(t) = \sin(\omega t)u(t) \;\Rightarrow\; \mathcal{L}(f(t)) = \frac{\omega}{s^2 + \omega^2}$$
that is a second order polynomial;

- cosine signal: observing that, apart from a constant term, it is the time derivative of the sinusoidal signal:
$$f(t) = \cos(\omega t)u(t) \;\Rightarrow\; \mathcal{L}(f(t)) = \mathcal{L}\left(\frac{1}{\omega}\frac{d}{dt}\sin(\omega t)\right) = \frac{s}{s^2 + \omega^2}.$$

After this brief review, it is possible to make an example. We have previously demonstrated that, using the properties of the Laplace transform:

$$\mathcal{L}\left(\frac{d}{dt}e^{-\frac{t}{\tau}}u(t)\right) = \frac{s\tau}{1 + s\tau}.$$

A different possibility is to write explicitly the derivative in the time domain:

$$\frac{d}{dt}e^{-\frac{t}{\tau}} = -\frac{1}{\tau}e^{-\frac{t}{\tau}}$$

and then try to calculate the Laplace transform of this signal. Doing this calculation, we will obtain a different result from what we expected. What is missing? Where is the problem?
Needless to say, deriving in the time domain we have neglected the presence of the step function, thus deriving a decreasing exponential over the whole time domain, from $-\infty$ to $+\infty$. If we consider it, we need to add to the time derivative the time derivative of the step function, that is a Dirac delta function of unitary amplitude. This is the reason why calculating the derivative in the time domain we have obtained a Laplace transform that is different from the one we have obtained using the properties of the Laplace transform. Transforming the correct derivative, it is possible to observe that we obtain:

$$-\frac{1}{1 + s\tau} + 1 = \frac{s\tau}{1 + s\tau}$$

that is exactly the expected result. The take home message, then, is to never forget the presence of the step function, since we have assumed that every signal starts at time $t = 0$.

5.1.3 Elementary components

Another important topic for the following part of the course is the behaviour of a few elementary electronic components, that will be needed for understanding more complex circuits. They are:

- resistor: its constitutive relationship, called Ohm's law, in the time domain:
$$v(t) = R \cdot i(t)$$
while in the frequency domain:
$$V(s) = R \cdot I(s)$$

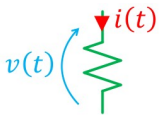


Figure 5.1: A resistor.

thus giving the following complex impedance:

$$\frac{V(s)}{I(s)} = R;$$

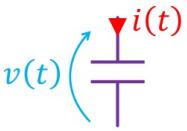


Figure 5.2: A capacitor.

- capacitor: in the time domain:

$$i(t) = \frac{dq(t)}{dt} = \frac{d}{dt}(Cv(t)) = C\frac{dv(t)}{dt}$$

where C is called capacity and, in the frequency domain:

$$I(s) = sCV(s)$$

that gives the following complex impedance:

$$\frac{V(s)}{I(s)} = \frac{1}{sC} = Z_C$$

thus making the voltage and the current in the Laplace domain directly proportional through a constant (the impedance) that depends on the Laplace operator s;

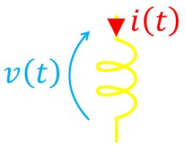


Figure 5.3: An inductor.

- inductor: in the time domain:

$$v(t) = L \cdot \frac{di(t)}{dt}$$

where L is called inductance and, in the frequency domain:

$$V(s) = sLI(s)$$

thus allowing us to define the following complex impedance:

$$\frac{V(s)}{I(s)} = sL = Z_L$$

that again makes the constitutive relationship linear in the Laplace domain.

5.1.4 RC network

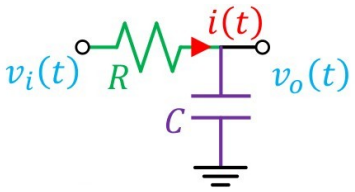


Figure 5.4: An RC network.

A first, important network that we can study is the RC network. It consists in a resistor and a capacitor in series and, in the time domain, the circuit can be studied starting from the constitutive relationship of the capacitor and from the Kirchhoff's law for voltages:

$$\begin{cases} v_i = v_o + iR \\ i = C\frac{dv_o}{dt} \end{cases}.$$

Substituting one in the other, we obtain:

$$RC\frac{dv_o}{dt} + v_o = v_i$$

and we know that the solution of this differential equation will be the solution of the homogeneous equation plus a particular integral. To solve the associated homogeneous equation, we need to find the zeros of the characteristic polynomial:

$$RCz + 1 = 0 \;\Rightarrow\; z = -\frac{1}{RC}$$

and since we know that the homogeneous equation will have an exponentially decaying solution:

$$v_o(t) = ke^{-zt} = ke^{-\frac{t}{RC}}$$

while a particular integral will be the one for constant input and constant output:

$$v_o^* = v_i = \text{const}$$

thus giving the following overall solution:

$$v_o(t) = ke^{-\frac{t}{RC}} + v_i.$$

In the frequency domain, we can replace the capacitor with the associated complex impedance, thus obtaining through a voltage partition:

$$V_o(s) = \frac{Z_C}{Z_C + R}V_i(s) = \frac{\frac{1}{sC}}{\frac{1}{sC} + R}V_i = \frac{1}{1 + sCR}V_i.$$

Observing the denominator of the ratio between output and input, also called transfer function:

$$\frac{V_o}{V_i} = \frac{1}{1+sCR}$$

we can observe that it is exactly identical to the characteristic equation (this time, the variable is represented by the Laplace operator s), therefore it will give the eigenvalue of the problem:

$$s = -\frac{1}{RC}.$$

This will allow us to find the solution of the problem. In practical cases, then, a few difficulties may arise when we need to go back from the Laplace domain to the time domain, anti-transforming it.
As a first example, we can consider the following input signal:

$$v_i(t) = A \cdot \delta(t).$$

In the Laplace domain, this gives:

$$V_i(s) = A$$

and therefore the output in the Laplace domain will be:

$$V_o(s) = \frac{A}{1+sRC} = \frac{A}{1+s\tau}, \quad \tau = RC$$

and we have solved the problem in the Laplace domain. Most of the times, this is enough, since we know to get back to the time domain by using the important signals we have introduced in the previous section. In this case:

$$v_o(t) = \frac{A}{\tau} e^{-\frac{t}{\tau}}.$$

Another example, we can suppose to have a step input:

$$v_i(t) = u(t)$$

thus obtaining in the frequency domain:

$$V_i(s) = \frac{A}{s}.$$

This gives the following output, that can be decomposed as the sum of two polynomials:

$$V_o(s) = \frac{1}{1+s\tau} \cdot \frac{A}{s} = A\left(\frac{1}{s} - \frac{\tau}{1+s\tau}\right)$$

that, recognizing one fundamental signal and the translation property in the frequency domain, gives:

$$v_o(t) = A\left(1 - e^{-\frac{t}{\tau}}\right) \cdot u(t).$$

We can immediately observe, then, that the input signals given in the two previous examples are closely related. In fact, the last one is the integral of the

previous one and, therefore, this property will be valid also at the output, where the second signal is the integral of the first one. The same property holds considering the differentiation of a certain signal and it is one of the reasons for which the Dirac delta function and the step function are among the most important signals used to characterize the response of a certain system.
As an exercise, the willing student is asked to find the output signal of an RC network in which the position of the resistor and of the capacitor are swapped with respect to what is represented in Figure 5.4 when at the input of the network is applied a Dirac delta signal and a step function signal.

5.1.5 Lag network

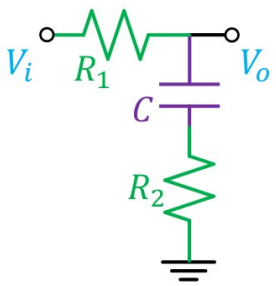


Figure 5.5: A lag network.

In Figure 5.5 it is possible to observe the so called lag network, consisting in the series of two resistors and a capacitor in between them. Solving it in the Laplace domain, due to the linearity property we can sum the resistances and the complex impedance of the capacitor, since they are in series. Applying a voltage partition, then, we obtain:

$$\frac{V_o}{V_i} = \frac{R_2 + Z_C}{R_1 + R_2 + Z_C} = \frac{R_2 + \frac{1}{sC}}{R_1 + R_2 + \frac{1}{sC}} = \frac{1 + sCR_2}{1 + sC(R_1 + R_2)}$$

that means:

$$V_o = \frac{1 + sCR_2}{1 + sC(R_1 + R_2)} V_i.$$

Given a step input:

$$v_i(t) = u(t) \;\Rightarrow\; V_i = \frac{A}{s}$$

we obtain the following output:

$$V_o = \frac{1 + sCR_2}{1 + sC(R_1 + R_2)} \cdot \frac{A}{s} = A\left(\frac{1}{s} - \frac{CR_1}{1 + sC(R_1 + R_2)}\right)$$

and defining the following time constant:

$$\tau = C(R_1 + R_2)$$

in the time domain we obtain:

$$v_o(t) = A\left(1 - \frac{R_1}{R_1 + R_2} e^{-\frac{t}{\tau}}\right).$$

It is therefore important to remember that when we have a linear network the solution will be a sum of exponentials with time constants equal to the poles of the transfer function, that are the solutions of the denominator of the transfer function. Every reactive element (namely, capacitors) in general will add a pole to the transfer function, each one with its own time constant. Moreover, the time constant of a capacitor will be proportional to the product between its capacity and the equivalent resistance seen from the capacitor. This can be found putting the input and the output at zero (and switching off any source) and computing the equivalent resistance seen from the pins of the capacitor.
In this particular case, since we have just one capacitor, we will have just one exponential. A different possibility is then to write the solution by using the initial and final value theorems:

$$v_o(0^+) = \lim_{s\to+\infty} sV_o(s) = \lim_{s\to+\infty} \cancel{s} \cdot \frac{1+sCR_2}{1+sC(R_1+R_2)} \cdot \frac{A}{\cancel{s}} = A \cdot \frac{R_2}{R_1+R_2}$$

$$v_o(+\infty) = \lim_{s\to 0} \cancel{s} \cdot \frac{1+sCR_2}{1+sC(R_1+R_2)} \cdot \frac{A}{\cancel{s}} = A$$

and joining these two asymptotic values with an exponential behaviour with time constant calculated as before.
There are, therefore, different ways of computing the same solution. A final check, then, at least for simple networks, is to consider whether the behaviour we have written is consistent from the physics of the circuit. As time tends to infinity, in fact, we will reach a steady-state condition and therefore the voltage across the capacitor will constant. This means that the capacitor will act as an open circuit and the output will be equally identical to the input, giving a transfer function equals to one. An infinitesimal amount of time after the application of the input signal, on the other hand, by the continuity of physical variables the voltage across the capacitor cannot be changed abruptly (this would have required an infinite amount of current, thus clearly being unphysical), thus being equal to zero. This means that the capacitor, in this limiting condition, will behave as a short-circuit, thus giving:

$$v_o = \frac{R_2}{R_1+R_2} v_i$$

from a voltage partition. This means that our analytical solution is consistent with the physics of the problem.

5.1.6 Sinusoidal signals and Bode plots

Consider now a system with a certain transfer function $T(s)$. If we apply a certain input in the Laplace domain $V_i(s)$, the system will produce an output that can be written as:

$$V_o(s) = T(s) \cdot V_i(s).$$

Assuming now that the input signal is a sinusoidal function of frequency ω:

$$v_i(t) = A\sin(\omega t)$$

in general we can split the response of the system in a transient, exponential behaviour and in a steady state response. In a linear system, the steady state

component of the response will be oscillating at the same frequency but, in general, with a different amplitude and with a different phase from the original one. It is possible to relate the new amplitude and the new phase to the original ones using the characteristics of the system, that are resumed in the transfer function. The output signal, then, can be written as:

$$v_o(t) = AB\sin(\omega t + \phi)$$

where:

$$B = |T(j\omega)|\,, \quad \phi = \angle\left(T(j\omega)\right).$$

Therefore, the transfer function will contain all the needed information. This function can be represented in a peculiar way, using the so called Bode plots. They are logarithmic graphs, in which on the horizontal axis we have the logarithm[3] in base ten of the frequency, while in the amplitude plots on the vertical axis we will have the magnitude of the transfer function expressed in decibels:

$$|T(j\omega)|_{\text{dB}} = 20 \cdot \log\left(|T(j\omega)|\right)$$

and in the phase plots the scale will be linear with the phase of the transfer function.
An immediate example is the representation of the transfer function of the RC network represented in Figure 5.4. In the Laplace domain, it can be written as:

$$T(s) = \frac{1}{1 + s\tau}$$

and if we consider explicitly the meaning of the Laplace operator:

$$s = j\omega \;\Rightarrow\; T(j\omega) = \frac{1}{1 + j\omega\tau}.$$

In this case, we can then write the magnitude in decibels as:

$$|T(j\omega)| = \frac{1}{\sqrt{1 + (\omega\tau)^2}}$$

$$|T|_{\text{dB}} = 20\log\left(\frac{1}{\sqrt{1 + (\omega\tau)^2}}\right) = -20\log\left(\sqrt{1 + (\omega\tau)^2}\right).$$

A simple way of plotting this relationship is to study the asymptotic behaviour of the Bode plots. This means that, for this relationship, we can consider two limiting cases: one for $\omega\tau \ll 1$ and the other for $\omega\tau \gg 1$. Analysing the first one:

$$\omega\tau \ll 1: \quad \omega \ll \frac{1}{\tau} \;\rightarrow\; f \ll \frac{1}{2\pi\tau},\; |T|_{\text{dB}} = 0$$

while for the second one:

$$\omega\tau \gg 1: \quad \omega \gg \frac{1}{\tau} \;\rightarrow\; f \gg \frac{1}{2\pi\tau}$$

$$|T|_{\text{dB}} = -20\log(\omega\tau) = -20\log(2\pi f\tau) = -20\log(f) + 20\log\left(\frac{1}{2\pi\tau}\right)$$

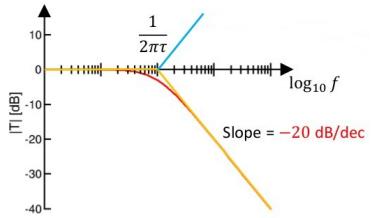


Figure 5.6: Bode plot of the amplitude of the transfer function of an RC circuit (in red), asymptotic behaviour of the same function (in yellow) and Bode plot of a single-zero function (in light blue).

but since $\log(f)$ is the variable on the horizontal axis, the right end of the graph will be represented as a straight line with a slope of -20 dB/dec.

In Figure 5.6, it is possible to observe how we have extended the asymptotic behaviour of the transfer function, thus obtaining a function whose first derivative is discontinuous. Moreover, note that the discontinuity of this single-pole function is placed exactly in the frequency of the pole:

$$s = -\frac{1}{\tau}.$$

The maximum error between the exact transfer function and the asymptotic one will thus be obtained in that point and it will be equal to 3 dB. After the pole, the function continuous as a straight line with a slope of -20 dB. If instead of the pole we had a zero at that frequency, we would have obtained a straight line increasing with a slope of $+20$ dB/dec after that point (while being identically zero before it).
This kind of representation of the transfer functions explains how a sinusoidal signal is affected by the system depending on the frequency of the input signal. If we have, for example, the following transfer function:

$$T(s) = \frac{1}{s\tau}$$

we can observe that it is clearly a single-pole transfer function and that its pole is placed in the origin, for frequency $f = 0$. This means that we are able to represent the transfer function as a straight line always decreasing with a slope of -20 dB/dec.
In the same way, we can calculate the phase of the previous functions. For the transfer function of the RC network:

$$\angle\left(\frac{1}{1+s\tau}\right) = \angle\left(\frac{1}{1+j\omega\tau}\right) = -\angle(1+j\omega\tau) = -\arctan(\omega\tau)$$

and it can be represented as in Figure 5.7.

A zeroth order approximation of the phase function, in this case, is to assume to have a $-90°$ shift every time we reach a pole and a $+90°$ phase shift every time we cross a zero. This approximation, however, is quite brutal and it can lead to quite significant errors. In a more accurate, first order approximation, we

[3] In general, when omitted the base, as in log we will assume base ten.

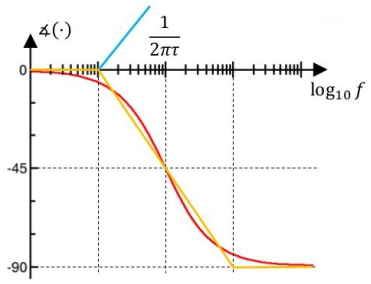


Figure 5.7: Bode plot of the phase of the transfer function of an RC circuit (in red), asymptotic behaviour at the first order of the same function (in yellow) and Bode plot of a single-zero function (in light blue).

can assume this transition to be a little more smooth, with a transient behaviour represented with a straight line with slope $-45°/\text{dec}$ (in the case of a zero, it will change the sign) extending from one decade before the pole (or the zero) to one decade after the pole (or the zero). In the case of a single-zero transfer function:

$$T(s) = 1 + s\tau \;\rightarrow\; \angle\left(T(j\omega)\right) = \arctan(\omega\tau).$$

In general, the transfer function has a certain number of poles and zeros and it can be written as:

$$T(s) = G \cdot \frac{(1+s\tau_{z1}) \cdot (1+s\tau_{z2}) \cdot \;\ldots\; \cdot (1+s\tau_{zn})}{(1+s\tau_{p1}) \cdot (1+s\tau_{p2}) \cdot \;\ldots\; \cdot (1+s\tau_{pn})}$$

and in decibels it can be written as:

$$|T|_{\text{dB}} = |G|_{\text{dB}} + \sum_{i=1}^{n} |1+j\omega\tau_{zi}|_{\text{dB}} - \sum_{k=1}^{n} |1+j\omega\tau_{pk}|_{\text{dB}}$$

$$\angle(T) = \sum_{i=1}^{n} \angle(1+j\omega\tau_{zi}) - \sum_{k=1}^{n} \angle(1+j\omega\tau_{pk}).$$

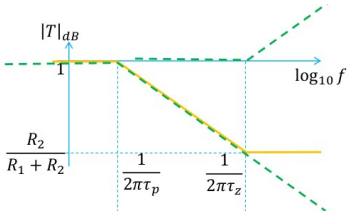


Figure 5.8: Bode plot of the amplitude of the transfer function of the lag network.

Another example is represented by the transfer function of the lag network, that is represented in Figure 5.5. It can be written as:

$$T(s) = \frac{V_o}{V_i} = \frac{1 + sCR_2}{1 + sC(R_1 + R_2)}$$

and therefore we can identify a zero:

$$s = -\frac{1}{CR_2} \;\Rightarrow\; f_z = \frac{1}{2\pi CR_2}$$

and a pole that will be at lower frequency with respect to the zero:

$$s = -\frac{1}{C(R_1+R_2)} \;\Rightarrow\; f_p = \frac{1}{1\pi C(R_1+R_2)} < f_z.$$

The behaviour of the single-pole transfer function and of the single-zero transfer function are represented by the dotted lines in Figure 5.8. By summing them, it is possible to obtain the behaviour of the overall transfer function, that is represented by the yellow line.

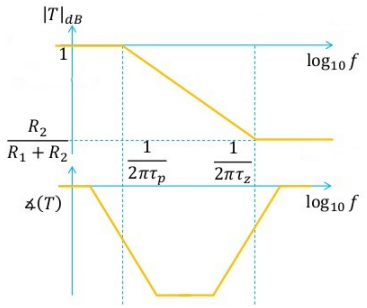


Figure 5.9: Bode plot of the amplitude and of the phase of the transfer function of the lag network.

In general, to represent the Bode diagram of a certain transfer function we identify all the zeros and poles of the function and, then, starting from an horizontal line (if there is not any pole or zero in the origin), we increment of 20 dB/dec the slope of the line when we cross a zero and we diminish it of -20 dB/dec after crossing every pole. The same can be done for the phase adding or subtracting 90°. It is important to note that, especially with respect to the phase diagram, the asymptotic Bode plots will be a good enough approximation of the real behaviour of the transfer function if and only if the zeros and the poles of the function will be well apart one from the other. Also in this simple case, we can investigate whether the Bode plot is consistent with the physical behaviour of the circuit in the limiting conditions, thus verifying the initial and final value of the Bode plot. In fact, in the low frequency approximation, since the capacitor tends to be an open circuit:

$$f \to 0 \;\Rightarrow\; Z_C \to \infty \;\Rightarrow\; V_o = V_i \;\Rightarrow\; |T|_{\text{dB}} = 0$$

while in the high frequency approximation, since the capacitor tends to be a short-circuit:

$$f \to \infty \;\Rightarrow\; Z_C \to 0 \;\Rightarrow\; V_o = \frac{R_2}{R_1+R_2}V_i \;\Rightarrow\; |T|_{\text{dB}} = \left.\frac{R_2}{R_1+R_2}\right|_{\text{dB}}.$$

The study of the limiting cases can thus be useful for checking errors in the Bode plots.

As an exercise, then, the student can represent a network that gives a transfer function in which we have one pole and one zero whose frequencies are related as:

$$f_z < f_p.$$

This kind of network is called lead network.
Final, we can investigate the so called band-pass filter, whose transfer function is:

$$T(s) = A\frac{s\tau_0}{(1+s\tau_1)(1+s\tau_2)}$$

where $\tau_1 \gg \tau_2$. This means that:

$$f_1 = \frac{1}{2\pi\tau_1} \ll f_2 = \frac{1}{2\pi\tau_2}$$

and that we have one zero in the origin. Therefore, we will obtain the Bode diagram represented in Figure 5.10. The student is then ask quote the Bode diagram represented in Figure 5.10 and to calculate the step response of this band-pass filter.

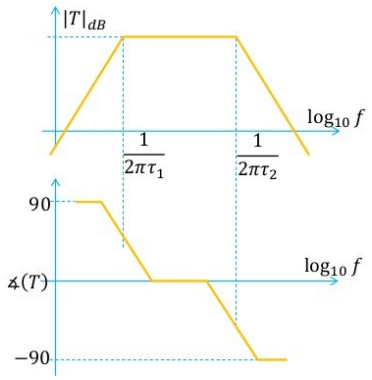


Figure 5.10: Bode plot of the amplitude and of the phase of the transfer function of a pass-band filter.

5.2 Integrator and differentiator circuits

5.2.1 The integrator

In Figure 5.11 it is represented the circuit of an ideal integrator. Since in such circuit the operation amplifier is an ideal operation amplifier, we will not have any current flowing through the negative input pin and, since it is a negative feedback system, we will have:

$$V^- = V^+ = 0 \text{ V}.$$

Therefore, we can write the current flowing through the resistance R and the capacitor C as:

$$I = \frac{V_i}{R}$$

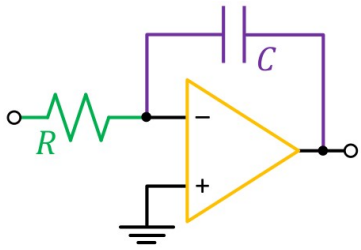


Figure 5.11: An ideal integrator.

and thus the voltage drop across the capacitor, in the Laplace domain:

$$V_C = Z_C I = \frac{1}{sC} I.$$

From Kirchhoff's voltage law, then, we can write the output voltage:

$$V_C + V_o = 0 \;\Rightarrow\; V_o = -V_C = -\frac{I}{sC} = -\frac{V_i}{sCR}$$

thus obtaining the following transfer function:

$$T(s) = \frac{V_o}{V_i} = -\frac{1}{sCR}$$

where we recognize that $1/s$ is the Laplace integral operator.

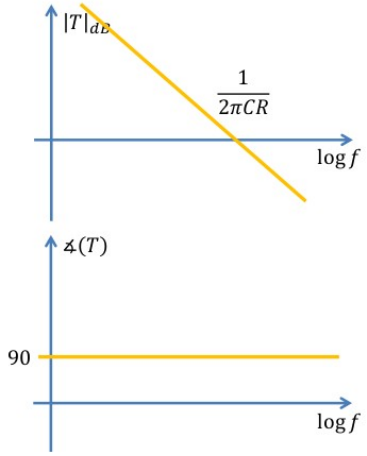


Figure 5.12: Bode plots of the transfer function of an ideal integrator.

In Figure 5.12 are represented the Bode plots of the magnitude and of the phase of the previous transfer function. For the magnitude, since:

$$T(j\omega) = -\frac{1}{j\omega CR} = \frac{j}{\omega CR}$$

we obtain:

$$|T| = \frac{1}{\omega CR}$$

and we can calculate the frequency at which the magnitude crosses the zero decibels axis as:

$$|T(j\omega_0)| = 1 \;\Rightarrow\; \omega_0 = \frac{1}{CR} \;\Rightarrow\; f_0 = \frac{1}{2\pi CR}.$$

For the phase, on the other hand, we can observe that we are dealing with a purely imaginary quantity, therefore the phase will be 90° at every frequency.
In practice, however, the configuration represented in Figure 5.11 is never adopted, while it more common to find the configuration represented in Figure 5.13. The first configuration, in fact, may give rise to some problems when we are dealing with a non-ideal operation amplifier. For example, we can consider the presence of an offset voltage, that leads to having a voltage V_{os} at the positive input pin of the operation amplifier. Since we are always dealing with a negative feedback system, we can write:

$$V^- = V^+ = V_{os}$$

and therefore in this case the current flowing through the resistance R and the capacity C will be:

$$I = \frac{V_{os}}{R}$$

even without having any input. This current, in particular, will come from the ideal voltage source that is inside the operation amplifier and will pass through the feedback capacitance C, thus giving the following output voltage:

$$V_{os} + V_C = V_o \;\Rightarrow\; V_o = V_{os} + Z_C I = V_{os} + \frac{V_{os}}{sCR}.$$

Assuming the offset voltage to be constant, we can transform it in the time domain:

$$v_o(t) = v_{os}(t) + \frac{1}{CR}\int_0^t v_{os}(t)\; dt$$

and observe that it will give rise to the following behaviour:

$$v_o = v_{os} + \frac{v_{os}}{CR} \cdot t.$$

Assuming the constant offset voltage to be positive[4], we obtain that the output will continuously increase with time until it reaches an upper limit v_{cc} represented by the maximum value of the output swing.
The willing student, then can demonstrate that the same problem arises when we have a bias current i_B through the pins of the operation amplifier or when there is even a small offset current.

The only, really working configuration is therefore the one represented in Figure 5.13. Adding a resistance R_C in parallel to the capacitor C, we can write a complex impedance from this parallel:

$$Z = R_C \| \frac{1}{sC} = \frac{R_C}{1 + sCR_C}$$

[4]Nothing changes having a negative offset voltage, only the output will be decreasing with time instead of increasing.

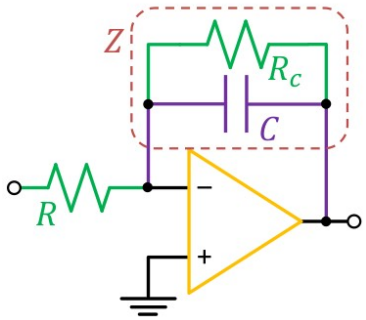


Figure 5.13: A practical configuration for an integrator.

thus obtaining the following transfer function that closely resembles the one of an inverting amplifier:

$$T(s) = \frac{V_o}{V_i} = -\frac{Z}{R} = -\frac{R_C}{R} \cdot \frac{1}{1 + sCR_C}.$$

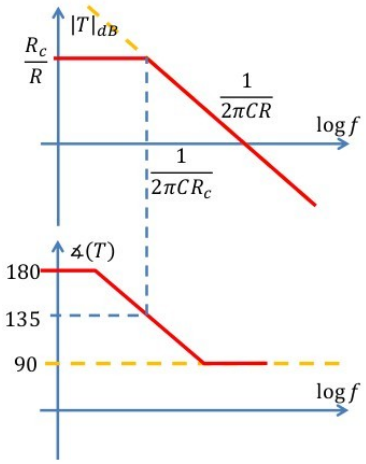


Figure 5.14: Bode plot of the transfer function of the practical configuration for an integrator.

The Bode plot of the magnitude[5] and of the phase of this transfer function is represented in Figure 5.14. From the plot of the magnitude, it is possible to observe that this system is not a perfect integrator: it behaves like an integrator only at high enough frequencies, after the pole in:

$$f_p = \frac{1}{2\pi CR_C}.$$

Moreover, we can calculate the zero frequency gain (by substituting $s = 0$ in the expression for the transfer function) obtaining the value represented on the

[5]The willing student is asked to calculate the value of the frequency at which the magnitude crosses the zero decibel axis.

graph of R_C/R. The phase, on the other hand, will start at 180° since at low frequency we have a positive and constant quantity and it will pass to 90° after that pole.
Adding to this circuit the effect of an offset, we can demonstrate that without any input the output will be:

$$V_o = -\frac{R_C}{R} V_{os} \quad \text{for } s = 0$$

considering a DC offset voltage. In this case in fact, since we are considering a continuous signal, the capacitor will act as an open circuit and therefore the overall behaviour will be identical to the one of an inverting amplifier. Considering the circuit in Figure 5.11, on the other hand, the fact that the capacitor could be replaced by an open circuit made the network equivalent to an operation amplifier used in an open-loop configuration, thus leading to an infinite output. However, in this second case the circuit at low frequency does not act as an integrator and we can study its temporal behaviour as a step response.
Given therefore a step of amplitude A as an input:

$$v_i(t) = Au(t) \;\Rightarrow\; V_i = \frac{A}{s}$$

we can write the output of the circuit as:

$$V_o = -\frac{R_C}{R} \frac{A}{1 + sCR_C} \cdot \frac{1}{s} = -\frac{AR_C}{R} \left(\frac{1}{s} - \frac{CR_C}{1 + sCR_C} \right)$$

and recognizing that the first term is a step of unitary amplitude while the second is an exponential, we can return to the temporal domain and write:

$$v_o(t) = -\frac{AR_C}{R} \left(1 - e^{-\frac{t}{\tau}}\right) u(t), \quad \tau = R_C C.$$

This behaviour is represented in Figure 5.15.

In this temporal step response, we can recognize two different behaviours, one after a time much shorter than the time constant τ, the other after a time much longer than that time constant:

$$v_o(t) \simeq \begin{cases} -\frac{AR_C}{R}, & t \gg \tau \\ -\frac{AR_C}{R} \left[1 - \left(1 - \frac{t}{\tau}\right)\right] = -\frac{AR_C}{R} \cdot \frac{t}{\tau} = -\frac{A}{RC} t, & t \ll \tau \end{cases}.$$

It is important to note that at high enough frequencies, when the circuit is acting as an integrator:

$$T(s) = -\frac{1}{sCR} \;\rightarrow\; v_o(t) = -\frac{A}{RC} t$$

consistently with the behaviour for $t \ll \tau$. This is because short time means high frequencies (from the Fourier theory); the same consideration can be done regarding the long-time behaviour and the low-frequencies region of the Bode plot.

Back to the ideal integrator, we can now study the network assuming to have a finite gain of the operation amplifier. We have previously calculated the ideal gain of this network:

$$G_{id} = -\frac{1}{sCR}$$

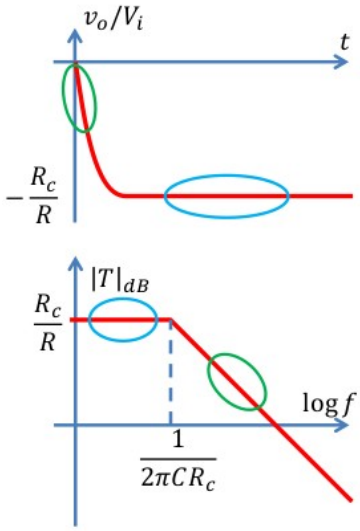


Figure 5.15: Temporal step response of a real integrator and corresponding behaviour on the Bode diagram of the magnitude.

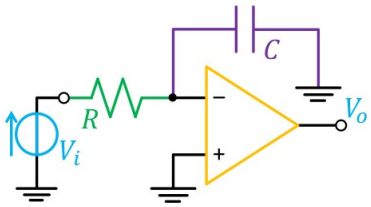


Figure 5.16: Open-loop gain calculation for an ideal integrator.

and we can now compute the open-loop gain. Considering the network represented in Figure 5.16, we can write the output as:

$$V_o = A(V^+ - V^-) = -AV^- = -A(s)\frac{1}{1+sCR}V_i$$

thus obtaining the following open-loop gain:

$$G_{ol} = \frac{V_o}{V_i} = -\frac{A_0}{1+s\tau}\cdot\frac{1}{1+sCR}$$

thus obtaining a function with two poles. The magnitude of this transfer function can be represented as in Figure 5.17 where, in general, the low frequency pole comes from the fact that we are dealing with a non-ideal operation amplifier.

On the same Figure, we can then represent also the ideal gain G_{id} and then study the loop gain G_{loop}, that can be approximated as equal to the lower function between the open-loop gain and the ideal gain. This means that, when we are considering a real operation amplifier, the ideal integrator will actually act as an integrator only on a finite bandwidth between a low frequency f_L and an high frequency f_H. To find the low frequency, we can observe that it

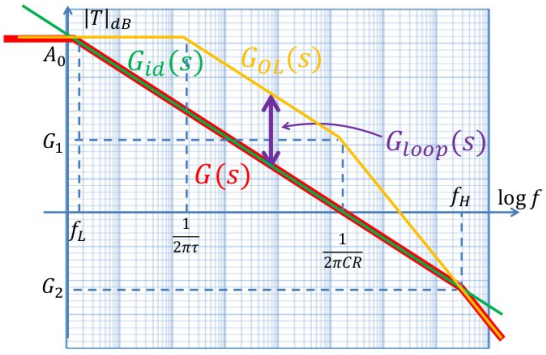


Figure 5.17: Bode plot of the magnitude for the open-loop gain.

corresponds to the point where the ideal gain is identical to the zero-frequency gain of the operation amplifier:

$$G_{id} = A_0 \;\Rightarrow\; \frac{1}{\omega CR} = A_0 \;\Rightarrow\; f_L = \frac{1}{2\pi CRA_0}.$$

Once we have found it, we can calculate the gain G_2 in Figure 5.17. To do this, we need to remember that if we have that a certain transfer function, on a certain interval is decreasing with a slope of $-n \cdot 20$ dB/dec, then the product:

$$G \cdot f^n = \text{const}$$

will be constant[6]. This allows us to write the following system of equations:

$$\begin{cases} A_0 f_L = G_2 f_H \quad \rightarrow \quad A_0 f_L = \frac{1}{2\pi CR} \\ G_1 \left(\frac{1}{2\pi CR}\right)^2 = G_2 \left(f_H\right)^2 = A_0 f_L f_H = \frac{1}{2\pi CR} f_H \end{cases}$$

thus obtaining:

$$f_H = \frac{G_1}{2\pi CR}.$$

However, we have also that:

$$A_0 \frac{1}{2\pi\tau} = G_1 \frac{1}{2\pi CR} \;\rightarrow\; G_1 = A_0 \frac{CR}{\tau}$$

thus obtaining:

$$f_H = A_0 \frac{CR}{\tau} \cdot \frac{1}{2\pi CR} = \frac{A_0}{2\pi\tau} = GBWP$$

and:

$$G_2 = A_0 \frac{f_L}{f_H} = \frac{1}{2\pi CR} \cdot \frac{1}{f_H} = \frac{1}{2\pi CR} \cdot \frac{2\pi\tau}{A_0} = \frac{\tau}{CRA_0}.$$

Alternatively, we could have started from the expression of the open-loop gain:

$$G_{ol} = \frac{A_0}{1 + s\tau} \cdot \frac{1}{1 + sCR}$$

[6]Obviously, if the function is increasing with a slope of $m \cdot 20$ dB/dec, the correct relationship will be:

$$\frac{G}{f^m} = \text{const}.$$

and since we have that the high frequency f_H is well beyond both poles:

$$G_{ol} \simeq \frac{A_0}{s\tau sCR} = \frac{A_0}{s^2\tau CR} \rightarrow \frac{A_0}{4\pi^2 f_H^2 \tau CR}$$

while for the ideal gain:

$$G_{id} = \frac{1}{sCR} \rightarrow \frac{1}{2\pi f_H CR}$$

and we obtain, since the two quantities are equal:

$$\frac{A_0}{4\pi^2 f_H^2 \tau CR} = \frac{1}{2\pi f_H CR} \quad \Rightarrow \quad f_H = \frac{A_0}{2\pi\tau}.$$

Both the previous methods are valid. A third possibility comes from the following observation:

$$G_{ol}|_{\text{dB}} - G_{id}|_{\text{dB}} = G_{loop}|_{\text{dB}}$$

and it can be calculated from a network similar to the one we had in Figure 5.16 but where the resistance is grounded and the test signal is imposed where we have cut the loop, between the capacitor and the output. From it, we can write:

$$V^- = V_T \frac{R}{R + \frac{1}{sC}} = V_T \frac{sCR}{1 + sCR}$$

thus obtaining the following loop gain:

$$G_{loop} = -A(s)\frac{sCR}{1 + sCR}.$$

In the high-frequency regime, therefore, the capacitor is a short-circuit and this means that the loop gain tends to be equal to $A(s)$, that crosses the zero decibel axis in the gain-bandwidth product $GBWP$.

5.2.2 The differentiator

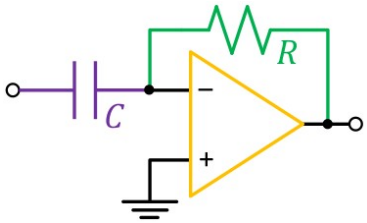


Figure 5.18: The differentiator.

The network of a differentiator is represented in Figure 5.18. From a direct inspection of this circuit, assuming an ideal operation amplifier, we can write its transfer function as:

$$T(s) = -sCR$$

thus obtaining the Bode diagrams for the magnitude and the phase represented in Figure 5.18. Investigating the Bode plot of the magnitude, we can determine

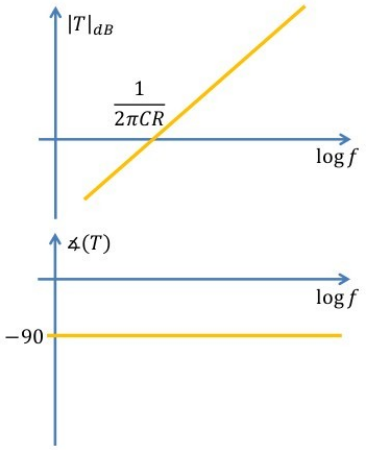


Figure 5.19: Bode diagrams for the magnitude and the phase of a differentiator.

the frequency of the point where the transfer function crosses the zero decibel axis:

$$|T| = 1 \;\Rightarrow\; 2\pi f_0 CR = 1 \;\Rightarrow\; f_0 = \frac{1}{2\pi CR}.$$

On the other hand, for the phase, we can observe that at every frequency we have a negative and purely imaginary quantity at every frequency:

$$\angle T = -90°.$$

In general, however, this ideal circuit cannot be used. In fact, we can observe that in the high-frequency regime its gain must tend to infinite and this is obviously an unphysical characteristic.

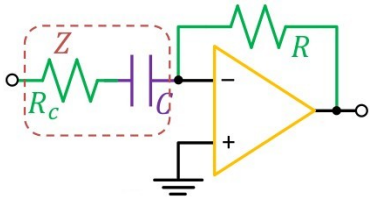


Figure 5.20: A real differentiator.

A real differentiator can be realized adopting the network represented in Figure 5.20. In this case, we can see that the capacity C and the additional resistance R_C are in series and therefore we can define the following complex impedance:

$$Z = R_C + \frac{1}{sC} = \frac{sCR + 1}{sC}$$

thus obtaining the following transfer function:

$$T(s) = -\frac{R}{Z} = \frac{sCR}{1 + sCR_C}.$$

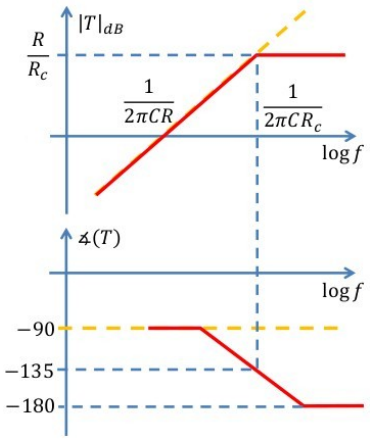


Figure 5.21: Bode diagrams of the magnitude and phase of a real differentiator.

Usually, we have that the compensation resistance is much lower than the other one:

$$R_C \ll R$$

due to the fact that it is, as we have said, just a compensation element. The Bode diagrams of the magnitude and the phase of this transfer function are represented in Figure 5.21. It is possible to observe that, in the high-frequency limit:

$$\lim_{s\to\infty} |T(s)| = \frac{R}{R_C}$$

consistently with the fact that the capacity behaves as a short-circuit at infinite frequency.

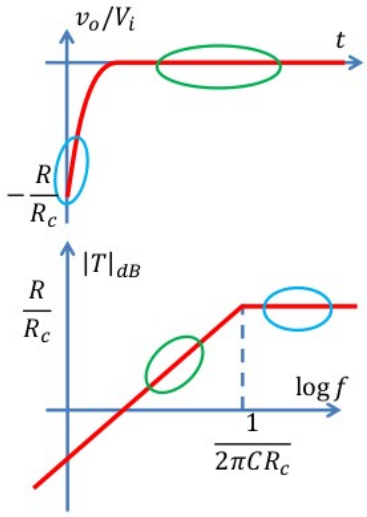


Figure 5.22: Temporal step response of a real differentiator.

Assuming a step input signal:

$$v_i(t) = Au(t) \;\rightarrow\; V_i = \frac{A}{s}$$

we can obtain the following output in the Laplace domain:

$$V_o(s) = \frac{sCR}{1+sCR_C}\frac{A}{s} = -\frac{AR}{R_C}\cdot\frac{CR_C}{1+sCR_C}$$

where we can recognize the Laplace transform of an exponential:

$$v_o(t) = -\frac{AR}{R_C}e^{-\frac{t}{CR_C}}.$$

From the representation of this temporal response in Figure 5.22, we can observe that the circuit will behave as a differentiator only for long times $t \gg \tau$ and, therefore, in the low-frequency region, while for short times $t \ll \tau$ and therefore in the high-frequency limit it will be a negative step.
We can now study the various gain terms considering a real operation amplifier with a single-pole transfer function $A(s)$. Considering for example:

$$C = 100 \text{ nF},\;\; R = 16 \text{ k}\Omega,\;\; R_C = 470\ \Omega$$

$$GBWP = 70 \text{ MHz},\;\; A_0 = 10^6$$

we can obtain that the ideal gain is:

$$G_{id} = -\frac{sCR}{1+sCR_C}.$$

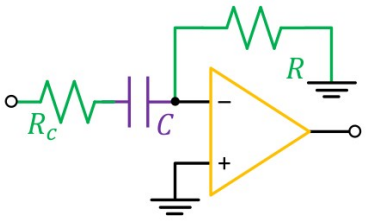


Figure 5.23: Calculation of the open-loop gain for a differentiator.

Considering the network represented in Figure 5.23, we can calculate now the open-loop gain. The voltage at the inverting pin of the operation amplifier will be:

$$V^- = V_T\frac{R}{R+Z} = V_T\frac{R}{R+R_C+\frac{1}{sC}} = V_T\frac{sCR}{1+sC(R+R_C)}$$

thus giving a pole with time constant $C(R+R_C)$, as we could have expected studying the circuit, observing the presence of the capacity C and calculating its equivalent resistance $R+R_C$. Then, the output voltage will be:

$$V_o =_A V^-$$

and therefore the open-loop gain:

$$G_{ol} = -A(s)\frac{sCR}{1+sC(R+R_C)} = -\frac{A_0}{1+s\tau}\cdot\frac{sCR}{1+sC(R+R_C)}.$$

We can then calculate the frequency of the pole in the ideal gain:

$$f_{p,id} = \frac{1}{2\pi CR_C} = 3.54 \text{ kHZ}$$

the frequency of the first pole in the open-loop gain:

$$f_{p1} = \frac{1}{2\pi\tau} = \frac{GBWP}{A_0} = 7 \text{ Hz}$$

and the frequency of the second pole in the open-loop gain:

$$f_{p2} = \frac{1}{2\pi C(R+R_C)} = 97 \text{ Hz}.$$

We can then calculate the following gain value:

$$G_2 = \frac{R}{R_C} = 35.5 = 31 \text{ dB}.$$

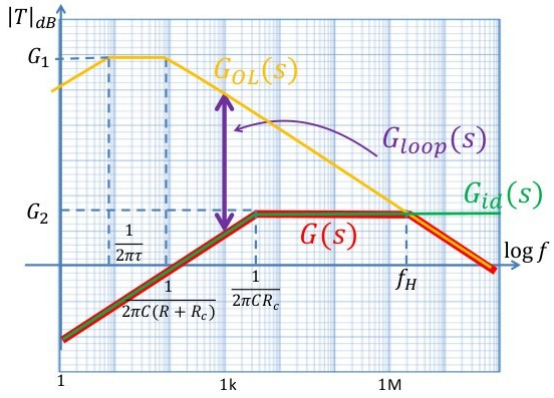


Figure 5.24: Bode plots of the gains for the differentiator.

In the low frequency limit, then, we have that:

$$G_{ol} \propto -A_0 sCR$$

therefore the frequency at which it will cross the zero decibel axis will be:

$$|G_{ol}| = 1 \;\Rightarrow\; f_0 = \frac{1}{2\pi A_0 CR}$$

and this allows us to find the remaining value of gain:

$$\frac{1}{f_0} = \frac{G_1}{\frac{1}{2\pi\tau}}$$

that gives:

$$G_1 = \frac{1}{2\pi\tau f_0} = \frac{2\pi A_0 CR}{2\pi\tau} = \frac{A_0 CR}{\tau} = 7 \cdot 10^5 \simeq 117 \text{ dB}.$$

At high-frequencies, the circuit is not an ideal circuit and we can calculate the frequency where:

$$|G_{loop}| = 0 \text{ dB}$$

that will be:

$$G_1 \frac{1}{2\pi C(R + R_C)} = G_2 f_H \;\Rightarrow\; f_H = 97\frac{G_1}{G_2} = 1.9 \text{ MHz}.$$

An alternative method for finding it consists in considering the expression of the open-loop gain well beyond the second pole:

$$G_{ol} = \frac{A_0}{1 + s\tau} \frac{sCR}{1 + sC(R + R_C)} \propto \frac{A_0 sCR}{s\tau sC(R + R_C)} = \frac{A_0 R}{s\tau(R + R_C)}$$

and at frequency f_H it must be equal to R/R_H:

$$G_{ol}(f_H) = \frac{A_0 R}{2\pi f_H \tau(R + R_C)} = \frac{R}{R_C}$$

thus obtaining:

$$f_H = \frac{A_0 R_C}{2\pi\tau(R + R_C)} = GBWP \frac{R_C}{R + R_C}$$

and it is consistent with what we have written before. An alternative, third methods consists in considering that at very high frequency the capacitor can be assumed to be similar to a short-circuit, thus giving a network similar to the one of an inverting amplifier, where:

$$G_{ol} \simeq \frac{GBWP}{1 + \frac{R}{R_C}}.$$

Considering the representation of the open-loop and ideal gain in Figure 5.24, then, we can again observe that the loop gain will be equal to the difference between the open-loop gain and the ideal gain, both expressed in decibels. To compute the loop gain, then, we can assume a network similar to the one represented in Figure 5.23 but where the input is grounded and to the breaking point, between the feedback resistance and the output, has been applied the test signal. We can then write the voltage at the inverting pin of the operation amplifier as:

$$V^- = V_T \frac{R_C \cdot \frac{1}{sC}}{R_C + R + \frac{1}{sC}} = V_T \frac{1 + sCR_C}{1 + sC(R + R_C)}$$

and the loop gain therefore is:

$$G_{loop}(s) = -A(s) \frac{1 + sCR_C}{1 + sC(R + R_C)} = -\frac{A_0}{1 + s\tau} \frac{1 + sCR_C}{1 + sC(R + R_C)}.$$

This transfer function will have a pole in:

$$f_p = \frac{1}{2\pi C(R + R_C)}$$

and a zero in:

$$f_z = \frac{1}{2\pi C R_C}$$

and we can represent it in a Bode plot[7]. Again we can find the gain G_1:

$$A_0 \frac{1}{2\pi\tau} = G_1 f_p \;\rightarrow\; G_1 = \frac{A_0}{f_p}\frac{1}{2\pi\tau} = A_0 \frac{2\pi C(R + R_C)}{2\pi\tau}$$

the gain G_2:

$$G_1 f_p^2 = G_2 f_z^2 \;\rightarrow\; G_2 = G_1 \frac{f_p^2}{f_z^2} = A_0 \frac{C(R + R_C)}{\tau} \cdot \frac{(2\pi C R_C)^2}{(2\pi C(R + R_C))^2} =$$
$$= \frac{C R_C^2}{\tau(R + R_C)}$$

and the high-frequency limit f_H at which the loop gain is unitary:

$$G_2 f_z = f_H \;\rightarrow\; f_H = \frac{A_0 C(R + R_C)}{\tau}\frac{1}{2\pi C R_C} = \frac{A_0(R + R_C)}{2\pi\tau R_C}.$$

5.2.3 The phase shifter

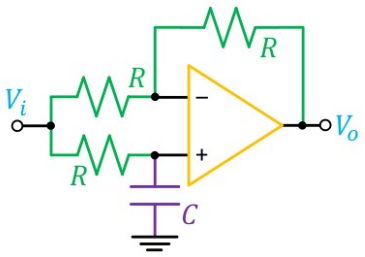


Figure 5.25: The phase shifter.

We can now consider another circuit, the phase shifter, represented in Figure 5.25. We can assume the following values:

$$R = 2\ \text{k}\Omega,\ C = 10\ \text{nF},\ A_0 = 10^6,\ GBWP = 10\ \text{MHz}.$$

To compute the ideal gain, we can use the superposition principle, splitting the input V_i as two inputs, one connected to the positive pin and the other connected to the negative pin of the operation amplifier. Grounding the input of the positive pin, the output will be:

$$V_o = V_i \;\rightarrow\; G = -1$$

[7]The willing student is invited to draw it.

while switching off the other input and recognizing the network of a non-inverting amplifier[8]:

$$G = 1 + \frac{R}{R} = 2.$$

Superimposing these two outputs with respect to a common input, we obtain:

$$V_o = -V_i + 2V^+ = -V_i + 2\frac{1}{1+sCR}V_i = \frac{1-sCR}{1+sCR}V_i$$

thus observing that we have obtained the transfer function of an all-pass filter:

$$T(s) = \frac{1-sCR}{1+sCR}.$$

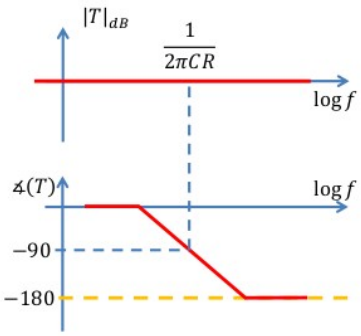


Figure 5.26: Bode diagrams of the magnitude and phase of a phase shifter.

The ideal gain of this network, therefore, will be an horizontal line at 0 dB, while the phase can be calculated as:

$$\begin{aligned}\angle\frac{1-j\omega CR}{1+j\omega CR} &= \angle(1-j\omega CR) - \angle(1+j\omega CR) = \\ &= -\arctan(\omega CR) - \arctan(\omega CR) = \\ &= -2\arctan(\omega CR) = \begin{cases} 0, & \omega \ll \omega_C \\ -180°, & \omega \gg \omega_C \end{cases}.\end{aligned}$$

This is the reason why this circuit is called phase shifter.
In the time domain, assuming a step input:

$$v_i(t) = Au(t) \;\rightarrow\; V_i(s) = \frac{A}{s}$$

we can write the step response in the Laplace domain as:

$$V_o(s) = \frac{1-sCR}{1+sCR}\frac{A}{s} = A\left(\frac{1}{s} - \frac{2CR}{1+sCR}\right)$$

thus obtaining:

$$v_o(t) = A\left(1 - 2e^{-\frac{t}{\tau}}\right)\cdot u(t)$$

[8]It is extremely important to be able to isolate pieces of known networks, recognizing their topology, in order to not need to solve them from scratch.

where the time constant is:

$$\tau = \frac{1}{CR}.$$

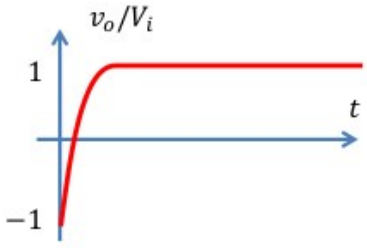


Figure 5.27: Step response of a phase shifter.

We can immediately observe that at high frequency, therefore for short times, the gain is one and the phase is $-\pi$, thus determining a change of sign in response to the step input. On the other hand, at low frequencies, therefore for long times, the gain is again one but the phase is zero and the output signal is identical to the input one.
As an alternative, we could have used the initial value theorem:

$$v_o(0^+) = \lim_{s\to\infty} sV_o(s) = \lim_{s\to\infty} s\frac{1-sCR}{1+sCR}\frac{A}{s} = -A$$

and the final value theorem:

$$\lim_{t\to+\infty} v_o(t) = \lim_{s\to 0} sV_o(s) = \lim_{s\to 0} s\frac{1-sCR}{1+sCR}\frac{A}{s} = A$$

for computing these quantities.
Another alternative was to compute them from a physical interpretation of the circuit, noting that at low frequencies (and therefore for long times) the capacitor is an open circuit while at high frequencies (and therefore for short times) the capacitor is a short-circuit.

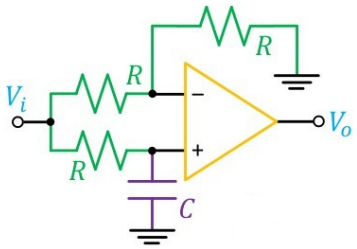


Figure 5.28: Calculation of the open-loop gain for a phase shifter.

Considering the network in Figure 5.28, we can try to calculate the open-loop gain. The output will be:

$$V_o = A(s)(V^+ - V^-)$$

where we have:

$$V^- = \frac{V_T}{2}, \quad V^+ = \frac{1}{1+sCR}V_T.$$

Therefore, the open-loop gain is:

$$G_{ol}(s) = A(s)\left(\frac{1}{1+sCR} - \frac{1}{2}\right) = A(s)\left(\frac{1-sCR}{1+sCR}\right)\frac{1}{2} =$$
$$= \frac{A_0}{1+s\tau}\left(\frac{1-sCR}{1+sCR}\right)\frac{1}{2}$$

and it can be represented as in Figure 5.29.

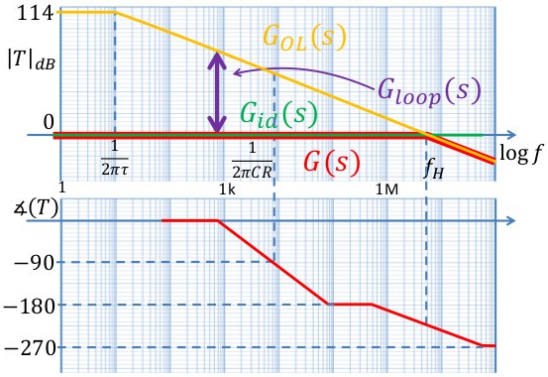


Figure 5.29: Bode diagrams of the magnitude and the phase of gains for the phase shifter.

We can calculate the frequency of the poles and zeros:

$$\frac{1}{2\pi\tau} = \frac{GBWP}{A_0} = 10 \text{ Hz}, \quad \frac{1}{2\pi CR} \simeq 8 \text{ kHZ}$$

the initial gain:

$$\frac{A_0}{2} = 5 \cdot 10^5 = 114 \text{ dB}$$

and the frequency at which the magnitude of the open-loop gain is zero decibel:

$$f_H = \frac{A_0}{2}\frac{1}{2\pi\tau} = \frac{GBWP}{2} = 5 \text{ MHz}.$$

To calculate then the loop gain, we can assume to have again the network in Figure 5.28 but with grounded input and a test signal applied at the breaking point, before the feedback resistance. In the high frequency limit, this will give:

$$V^- = \frac{V_T}{2}, \quad V^+ = 0$$

and therefore the output:

$$V_o = -A(s)\frac{1}{2}V_T$$

thus giving the following loop gain:

$$G_{loop}(s) = -\frac{A(s)}{2} = -\frac{1}{2}\frac{A_0}{1+s\tau}$$

where, to find the frequency at which it crosses the zero decibel axis, its high-frequency behaviour can be approximated as:

$$\frac{A_0}{2(1+s\tau)} \simeq \frac{A_0}{2s\tau} = 1$$

consistently with what we have found before.
From an analogous point of view we can study the phase of the gain G, obtaining the behaviour represented in Figure 5.29.

5.3 Input and output impedances and gain calculations

5.3.1 Choice of the test source

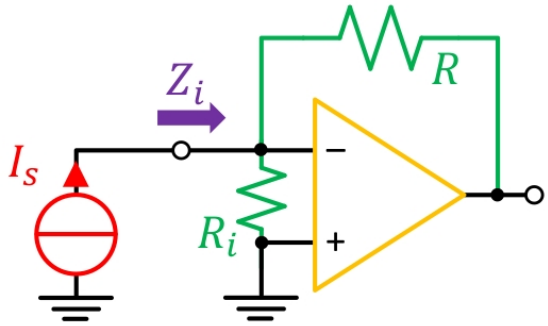


Figure 5.30: Circuit considered with an applied current source.

Consider now the circuit represented in Figure 5.30; compute the input impedance of this circuit. We can suppose, as in Figure, to apply a current source to the inverting pin and to try to calculate the input impedance as:

$$Z_i = \frac{V_S}{I_S}.$$

Applying a certain current I_S, we get that:

$$V_S = V^-$$

and, in the ideal case, it will give:

$$V^+ = V^- = 0 \;\Rightarrow\; V_S = 0 \;\Rightarrow\; Z_{id} = 0.$$

This means that the input impedance can be written as:

$$Z = \frac{Z_{ol}}{1 - G_{loop}}$$

and the feedback will tend to decrease the value of the open-loop impedance.
The open-loop impedance can be calculated by imposing:

$$G_{loop} = 0$$

but since the loop gain is proportional to the gain A of the operation amplifier, we can impose:

$$A = 0 \;\Rightarrow\; V_o = 0.$$

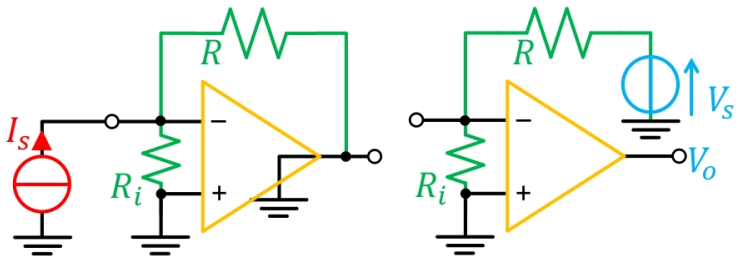


Figure 5.31: Calculation of the open-loop impedance (on the left) and of the loop gain (on the right).

In this condition, the open-loop impedance is equal to:

$$Z_{ol} = R_i \| R = \frac{R_i R}{R_i + R}.$$

Last, we can compute the loop gain when the amplifier has a certain gain $A(s)$ and we have set $I_S = 0$, thus having an open circuit connected to the inverting pin of the operation amplifier:

$$G_{loop} = -A(s) \cdot \frac{R_i}{R_i + R}.$$

Putting all these considerations together, we obtain the following input impedance:

$$Z_i = \frac{Z_{ol}}{1 - G_{loop}} = \frac{R_i R}{R_i + R} \cdot \frac{1}{1 + A(s)\frac{R_i}{R_i + R}} =$$
$$= \frac{R_i R}{R_i(1 + A(s)) + R} = \frac{R R_i}{R + R_i + R_i A(s)}.$$

Alternatively, we could have considered the whole circuit and solve it, noting that the resistance R_i is in parallel with the rest of the circuit:

$$Z = R_i \| Z'$$

where Z' is the input impedance of the second stage and, connecting a suitable source, we can calculate:

$$Z' = \frac{Z'_{ol}}{1 - G'_{loop}}$$

obtaining that:

$$Z'_{ol} = R,\; G'_{loop} = -A(s) \;\Rightarrow\; Z' = \frac{R}{1 + A(s)}.$$

Substituting back, then, we can write the value of the impedance Z:

$$Z = \frac{R_i Z'}{R_i + Z'} = \frac{\frac{R_i R}{1+A}}{R_i + \frac{R}{1+A}} = \frac{R_i R}{R + R_i(1 + A)} = \frac{R R_i}{R + R_i + R_i A}$$

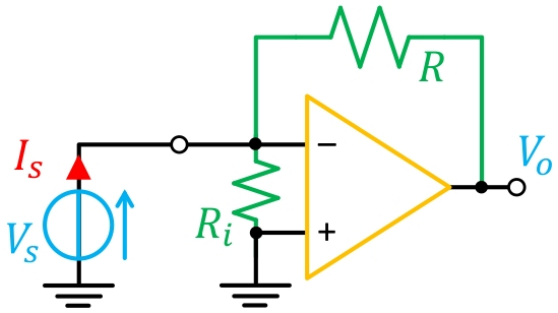


Figure 5.32: Circuit considered with an applied voltage source.

analogously to what we had in the previous case.

A third way of studying this circuit is to neglect all the feedback theory that we have studied, apply a voltage source as in Figure 5.32 and solve it. From an analysis of the circuit, we can observe that:

$$V^- = V_S,\ V^+ = 0 \ \Rightarrow\ V_o = -AV_S$$

and this gives the following current:

$$I_S = I_1 + I_2 = \frac{V_S}{R_i} + \frac{V_S(1+A)}{R}$$

where I_1 is the current passing through the input resistance R_i and I_2 is the current flowing through the feedback resistance R. From this, we get:

$$Z = \frac{V_S}{I_S} = \frac{1}{\frac{1}{R_i} + \frac{1+A}{R}} = \frac{R_i R}{R + R_i + R_i A}$$

and this is consistent with what we had before. In general, the application of the feedback theory takes more time but it makes the solution of the circuits easier.

Assume now to be using the first method, based on feedback theory, but we want to apply it to the circuit in Figure 5.32, where we have a voltage source. In this case, we need to calculate the ideal impedance but, doing this, we immediately have to face a problem. In fact, in the following limit:

$$A \ \to\ \infty$$

this means that both pins are set at the same voltage:

$$V^- = V^+$$

but since we have a voltage source:

$$V^- = V_S, \quad V^+ = 0$$

this condition seems to not be respected. The only way of satisfying these conditions is to have:

$$V_o \ \to\ -\infty$$

but this is an awkward situation. Moreover, this leads the current I to be infinite (at least from an ideal point of view), therefore:

$$I_S \rightarrow \infty$$

and this gives the following ideal impedance:

$$Z_{id} = \frac{V_S}{I_S} = 0$$

that is consistent with what we had in the previous case. Therefore, we are still recovering the right behaviour even if we are dealing with a clearly unphysical situation. Since the input impedance can be written as:

$$Z = \frac{Z_{ol}}{1 - G_{loop}}$$

we need now to calculate these two quantities.

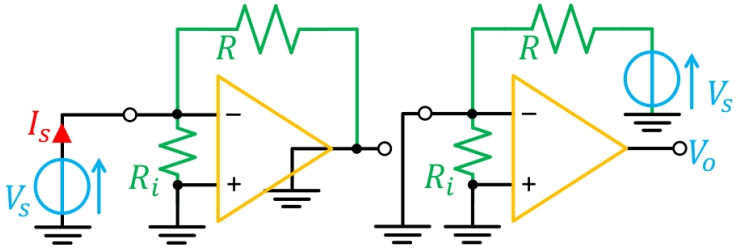


Figure 5.33: Calculation of the open-loop impedance (on the left) and of the loop gain (on the right) in the case of a test voltage source.

The open loop impedance, from Figure 5.33, can be immediately recognized to be equal to:

$$Z_{ol} = R_i \| R$$

and calculating the loop gain we obtain that:

$$V^+ = V^- = 0 \Rightarrow G_{loop} = 0.$$

However, this means that there is not any feedback and the circuit is working in open loop conditions, thus giving the following input impedance:

$$Z = Z_{ol}.$$

Applying the feedback theory, therefore, we must pay attention to not be breaking the loop and therefore of choosing the best test source. A wrong choice, in fact, can lead to break the loop, thus forcing a different behaviour of the circuit. In this case, if we do not choose a current source, we obtain some quantities that tend to infinity in the computation of the ideal impedance, thus leading to:

$$G_{loop} = 0.$$

Therefore, when we have a current flowing through the impedance that is equal to zero we need to use a voltage source, while if the voltage of the input pin is

set to zero we will have to use a current source. In the majority of the cases, the inverting pin of the operation amplifier will have a zero ideal impedance and, therefore, if it is the input pin of the operation amplifier the best choice is probably represented by a current source. The same, then, will happen when we will consider the output pin of the operation amplifier, when trying to calculate the associated impedance. For the positive pin, generally, its impedance tends to infinite and therefore we must drive it with a voltage source.
Alternatively, we could have used Blackman's formula:

$$Z = Z_{ol}\frac{1 - G_{loop}|_{sc}}{1 - G_{loop}|_{oc}}$$

where the loop gain is computed twice, one when the input is short-circuited to the ground and the other when it is left floating. It is important to note that if there is not anything connected in series or in parallel to the considered impedance of the circuit, then one of the two loop gains will be equal to zero.

5.3.2 Differential stage

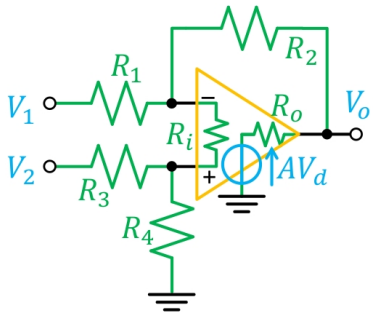


Figure 5.34: Differential stage considered.

Consider the differential stage represented in Figure 5.34, with the following values:

$$R_1 = 2\ \text{M}\Omega, \quad R_o = 75\ \Omega, \quad A_0 = 126\ \text{dB}$$

and find the values of the resistors R_1, R_2, R_3 and R_4 such that the ideal gain is:

$$G_{id} = 10$$

and then calculate the input impedance Z_1 and Z_2 respectively seen from V_1 and V_2.
Since we are dealing with the ideal gain, we can apply the superposition principle and, switching off the input at the positive pin:

$$V_2 = 0 \ \Rightarrow \ V_{o1} = -\frac{R_2}{R_1}V_1$$

while switching off the input at the inverting pin:

$$V_1 = 0 \ \Rightarrow \ V_{o2} = \frac{R_4}{R_3 + R_4} \cdot \frac{R_1 + R_2}{R_1}V_2.$$

Superimposing these two effects, we obtain that:

$$V_o = \frac{R_2}{R_1}\left(-V_1 + V_2 \cdot \frac{1 + R_1 \| R_2}{1 + R_3 \| R_4}\right)$$

from which we can obtain the usual condition for having a subtractor:

$$\frac{R_1}{R_2} = \frac{R_3}{R_4}$$

and the ideal gain sets this ratio to:

$$G = 10 = \frac{R_2}{R_1}.$$

Ideally, these resistances will be between 1 kΩ and 100 kΩ, since lower resistances will give too high currents while higher resistances will produce more noise, therefore we can choose, for example:

$$R_1 = R_3 = 10 \text{ k}\Omega, \quad R_2 = R_4 = 100 \text{ k}\Omega.$$

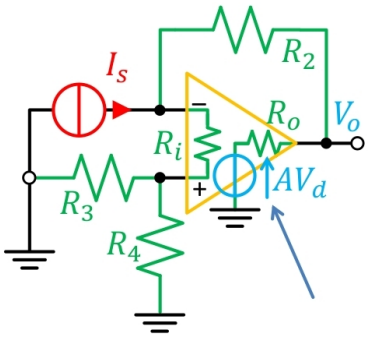


Figure 5.35: Computation of the first input impedance.

We can now compute the input impedance starting from the one corresponding to the input V_1. Considering the circuit in Figure 5.35, we can observe that the input impedance will be given by the following series:

$$Z_1 = R_1 + Z_1'$$

since R_1 is in series with respect to the rest of the circuit. Since now the test source must be connected to the inverting input pin, we can expect to have a low impedance at this node and therefore we can drive it using a current source I_S. Studying the network:

$$V^+ = 0, \; V^+ = V^- \;\Rightarrow\; V_S = 0 \;\Rightarrow\; Z'_{id,1} = 0$$

and this gives:

$$Z_1' = \frac{Z'_{ol,1}}{1 - G'_{loop,1}}.$$

To calculate the open-loop impedance, we need to shut off the voltage controlled voltage source that is indicated with the blue arrow in Figure, obtaining:

$$A = 0 \;\Rightarrow\; Z_{ol,1} = (R_i + R_3 \| R_3) \| (R_2 + R_o) = 95.2 \text{ k}\Omega.$$

The calculation of the loop gain gives:

$$G'_{loop,1} = \frac{R_i}{R_1 + R_2 + R_o + R_3 \| R_4} \cdot -A(s) = -0.95A(s).$$

In the low frequency limit, the loop gain is:

$$G'_{loop,1} \simeq -2 \cdot 10^6 \cdot 0.95$$

and therefore the input impedance is:

$$Z'_1 = \frac{95.2 \text{ k}\Omega}{1 + 2 \cdot 10^6 \cdot 0.95} \simeq 50 \text{ m}\Omega.$$

This value is very low, even lower than the impedance of the ground, therefore the following approximation holds:

$$Z_1 = R_1 + Z'_1 \simeq R_1 = 10 \text{ k}\Omega.$$

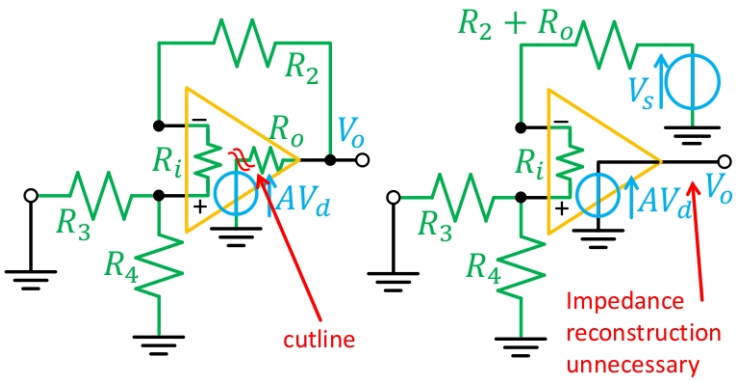


Figure 5.36: Computation of the loop gain and impedance reconstruction in this case.

To compute the second input impedance, we can note that the resistance R_3 is in series and R_4 in parallel to the impedance of the rest of the network, thus giving:

$$Z_2 = R_3 + R_4 \| Z'_2.$$

Driving the network with a voltage source, in the ideal case:

$$V^+ = V^- \;\Rightarrow\; I_S = 0 \;\Rightarrow\; Z'_{id} = \infty$$

and we can observe that choosing a current source to drive this node we would have obtained a clearly wrong result. Therefore:

$$Z'_2 = Z'_{ol,2}(1 - G_{loop,2})$$

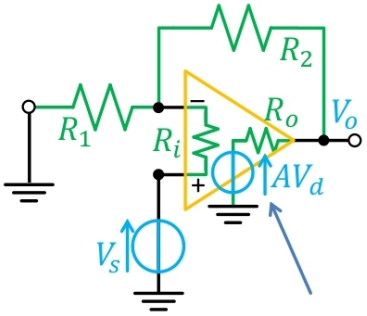


Figure 5.37: Computation of the second input impedance.

and we can calculate the open-loop impedance as:

$$Z'_{ol,2} = R_i + R_1 \| (R_2 + R_o) = 2 \text{ M}\Omega$$

and the loop gain as:

$$G_{loop,2} = -A(s) \frac{R_i \| R_1}{R_i \| R_1 + R_2 + R_o} = -0.09 A(s).$$

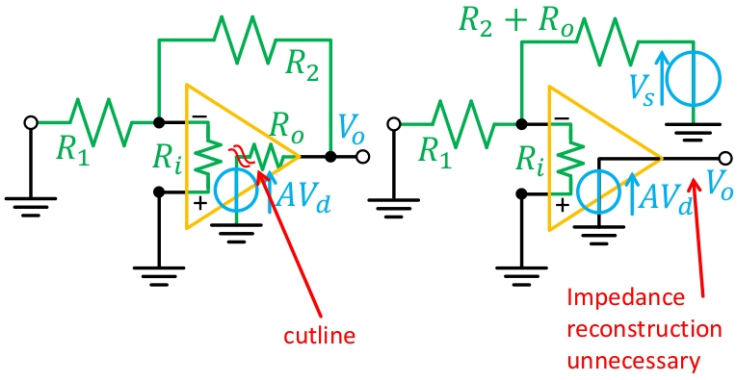


Figure 5.38: Computation of the loop gain and impedance reconstruction in this second case.

In the low frequency limit, this gives:

$$Z'_2 = 2 \text{ M}\Omega \cdot (1 + 0.09 \cdot 2 \cdot 10^6) \simeq 360 \text{ G}\Omega$$

and at the end we obtain the following input impedance:

$$Z_2 = R_3 + R_4 \| Z'_2 \simeq R_3 + R_4.$$

5.3.3 Buffer stage

Consider the buffer stage represented in Figure 5.39, in which we have the following values:

$$R = 900 \ \Omega, \ R_i = 1 \text{ M}\Omega, \ C = 20 \text{ nF}, \ R_o = 10 \ \Omega$$

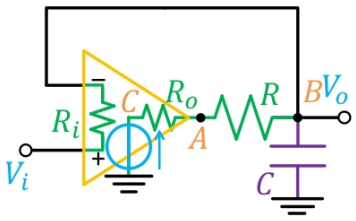


Figure 5.39: Buffer stage considered.

$$GWBP = 1 \text{ MHz}, \ A_0 = 100 \text{ dB}.$$

Calculate the loop gain by breaking in the three different points A, B and C and try to choose a compensation scheme if the phase margin of this network is not enough.

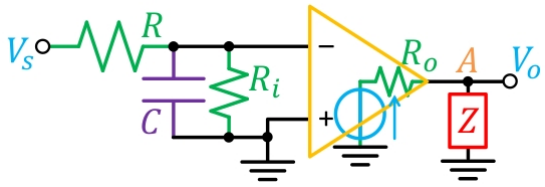


Figure 5.40: Buffer stage when the loop is broken in point A.

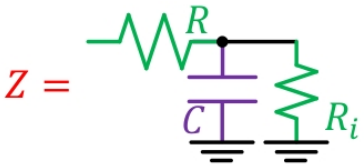


Figure 5.41: Reconstruction impedance when the loop is broken in point A.

The first point in which it is possible to break the loop is point A, obtaining the network represented in Figure 5.40. In this scheme, it is possible to add a compensation impedance Z as in Figure and, from the network, we can observe that it will be equal to:

$$Z = R + \frac{1}{sC} \| R_i = R + \frac{R_i}{1 + sCR_i} = \frac{R + R_i + sCRR_i}{1 + sCR_i}.$$

Applying a test source (no matter which one), we can obtain:

$$V^- = \frac{R_i \| \frac{1}{sC}}{R_i \| \frac{1}{sC} + R} V_S$$

therefore the output voltage will be:

$$V_o = -A(s) \frac{R_i \| \frac{1}{sC}}{R_i \| \frac{1}{sC} + R} \cdot \frac{Z}{Z + R_o} V_S$$

thus giving the following loop gain:

$$G_{loop,A} = -A(s)\frac{R_i \| \frac{1}{sC}}{R + R_o + R_i \| \frac{1}{sC}}.$$

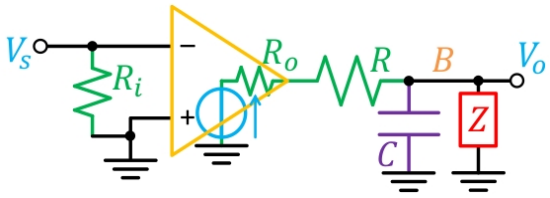


Figure 5.42: Buffer stage when the loop is broken in point B.

Breaking in the point B, the reconstruction impedance will be:

$$Z = R_i.$$

But since:

$$V^- = V_S$$

we obtain:

$$V_o = -A(s)V_S\frac{R_i \| \frac{1}{sC}}{R_o + R + R_i \| \frac{1}{sC}}$$

thus giving the following loop gain:

$$G_{loop,B} = -A(s)\frac{R_i \| \frac{1}{sC}}{R_o + R + R_i \| \frac{1}{sC}}$$

as in the previous case.

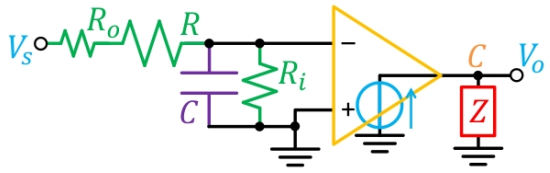


Figure 5.43: Buffer stage when the loop is broken in point C.

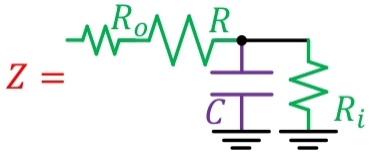


Figure 5.44: Reconstruction impedance when the loop is broken in point C.

In the last case, when the loop is broken in point C, the reconstruction impedance can be neglected since the output voltage V_o is directly set by the ideal voltage source. In fact:

$$V^- = \frac{R_i \| \frac{1}{sC}}{R + R_o + R_i \| \frac{1}{sC}} V_S$$

and the output voltage will be:

$$V_o = -A(s)V^-$$

thus giving the following loop gain:

$$G_{loop,C} = -A(s)\frac{R_i \| \frac{1}{sC}}{R + R_o + R_i \| \frac{1}{sC}}$$

This is a consequence of the fact that we are dealing with an ideal voltage controlled voltage source. Plugging in the numbers, we can obtain that:

$$\begin{aligned} G_{loop} &= -A(s)\frac{R_i \| \frac{1}{sC}}{R + R_o + R_i \| \frac{1}{sC}} = -\frac{A_0}{1+s\tau}\frac{\frac{R_i}{1+sCR_i}}{R + R_o + \frac{R_i}{1+sCR_i}} = \\ &= -\frac{A_0}{1+s\tau}\frac{R_i}{R + R_o + R_i + sCR_i(R+R_o)} = \\ &= -\frac{A_0}{1+s\tau}\frac{R_i}{R_i + R + R_o}\frac{1}{1+sC[R_i \| (R+R_o)]} \simeq \\ &\simeq -\frac{A_0}{1+s\tau}\frac{1}{1+sC[R_i \| (R+R_o)]} \end{aligned}$$

since we have noticed that:

$$\frac{R_i}{R_i + R + R_o} \simeq 1.$$

The time constant of the pole that we have obtained, therefore, is equal to the product between the capacity C and the equivalent resistance. Moreover, we can check the consistency of our result by observing that, without any capacitor, the output would be determined by the partition $R_i/(R_i+R+R_o)$. Substituting the numbers, it is possible to demonstrate that the pole of the operation amplifier is set in 10 Hz, while the other pole will be at 8.74 Hz.

Defining G_1 and G_2 as in the Bode plot reported in Figure 5.45, we can write:

$$G_1 f_{p1} = G_2 f_{p2} \;\rightarrow\; G_2 = G_1 \frac{f_{p1}}{f_{p2}} = 114 \simeq 41 \text{ dB}$$

and we can determine the crossover frequency as:

$$G_2 f_{p2}^2 = f_c^2 \;\rightarrow\; f_c = \sqrt{G_2} f_{p2} = 93.2 \text{ kHz}.$$

The phase margin will be surely quite low, since we are crossing the zero decibel axis with a negative slope of two:

$$\phi_m = 180^\circ - 90^\circ - \arctan\left(\frac{f_c}{f_{p2}}\right) \simeq 5^\circ$$

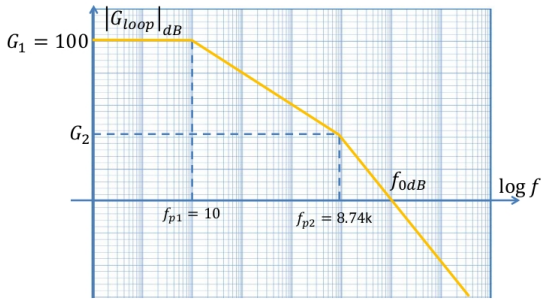


Figure 5.45: Bode plot of the loop gain that we have obtained.

where the first $-90°$ contribution is related to the presence of the pole within the operation amplifier, that is at a very low frequency. Note that this phase margin is obviously not enough, so we need to compensate for it. For the sake of simplicity, we can neglect the input resistance R_i and we can try to find a place where we can put a compensation capacitor. A first idea could be to put it between the positive and the negative input pins of the operation amplifier: however in this way the compensation capacitor will be in parallel to the other capacitor C and this compensation scheme will surely not work. A possible alternative, then, is to put it with one connected between R_o and R and the other one connected to ground. To study whether a compensation capacitor is in a good position for adding a zero, we need to study the behaviour of the network in the high-frequency limit, where the capacitor is equal to a short-circuit. In this condition, we want to retrieve the same transfer function (since capacitors should not change the behaviour of the network in the high-frequency limit), therefore also this position is not suitable for placing it, since it will lead to a zero transfer in the high-frequency limit that is different from what we had before.

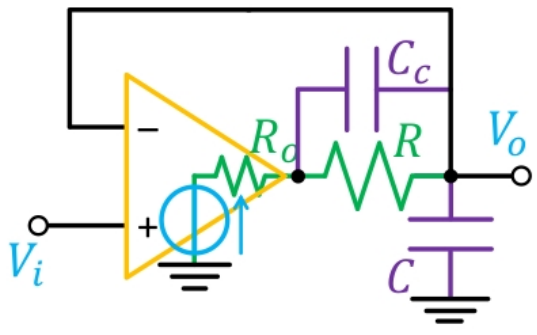


Figure 5.46: Compensated network.

The only residual possibility is to place it in parallel to the resistor R, as in Figure 5.46. In this position, studying the high-frequency limit, we can see that it will not give a zero transfer function, thus providing a correct behaviour. The

loop gain, then, can be calculated as:

$$\begin{aligned} G_{loop} &= -A(s)\frac{\frac{1}{sC}}{R_o + R\|\frac{1}{sC_c} + \frac{1}{sC}} = \\ &= -A(s)\frac{\frac{1}{sC}}{R_o + \frac{R}{1+sC_cR} + \frac{1}{sC}} = \\ &= -A(s)\frac{1+sC_cR}{sCR_o(1+sC_cR) + sCR + 1 + sC_cR} = \\ &= -A(s)\frac{1+sC_cR}{1 + s(CR + C_cR + CR_o) + s^2CC_cRR_o} \end{aligned}$$

where we can immediately recognize the presence of three additional terms that depend on the compensation capacitor C_c. Using this compensation capacitor, therefore, we have obtained an additional zero and an additional pole, since the two capacitors of the network are one independent from the other. However, the contributions coming from these two capacitors are mixing in the position of the two poles. Studying the circuit, we can try to solve the second order equation that we have obtained at the denominator of the loop gain and find the position of the two poles. In a design problem, however, at least the compensation capacitance C_c is not known and, therefore, we need to find some approximations for the position of the poles.
Assuming that the two poles are well separated in the frequency domain, since in general:

$$C_c \ll C$$

the compensation capacitor will probably be responsible for the high-frequency pole, while the other capacitor C will be responsible for the low-frequency one. At low frequencies, the compensation capacitor C_c can be approximated with an open circuit and we can calculate the equivalent resistance of the second capacitor C, thus obtaining the following position for the associated pole:

$$f_{p,LF} = \frac{1}{2\pi C(R + R_o)}.$$

In the high-frequency limit, we will probably be beyond the frequency of the pole given by the capacitor C and thus it can be approximated with a short-circuit, thus giving:

$$f_{p,HF} = \frac{1}{2\pi C_c(R\|R_o)}.$$

We have thus found that the pole of the capacitor C is not changing, we are only adding a pole and a zero through the compensation capacitor C_c. In this case, we want to have this zero exactly at the crossover frequency of the previous Bode diagram, while the high-frequency pole must be at a frequency higher than the crossover frequency. From their expressions:

$$f_z = \frac{1}{2\pi C_cR}, \quad f_{p,HF} = \frac{1}{2\pi C_c(R\|R_o)}$$

and this is consistent with the following consideration:

$$R\|R_o \ll R \;\Rightarrow\; f_{p,HF} \gg f_z.$$

We can thus calculate:

$$f_z = \frac{1}{2\pi C_c R} = 93.2 \text{ kHz}$$

thus obtaining the following value for the compensation capacity:

$$C_c = \frac{1}{2\pi f_z R} = 1.88 \text{ nF}$$

that gives the following position for the high-frequency pole:

$$f_{p,HF} \simeq 8.5 \text{ MHz}.$$

A last, alternative possibility was to place the compensation capacitor in series both to R and R_o; this will lead to the introduction of a zero. However, this kind of compensation can never be performed. To understand why, it is enough to study the low-frequency limit of the device, in which the capacitors are open circuits. In this condition, in fact, the open-loop gain is zero and the loop is open. We are therefore completely compromising the DC behaviour of the device, that should be preserved, otherwise we are completely changing the functionality of the circuit. We can thus observe that we always have to place the compensation capacitors in parallel with other elements of the network, never in series with them.
A second alternative for finding the position of the poles of the compensated circuit under the assumption of well-separated poles is to write the denominator of the loop gain as:

$$(1+s\tau_1)(1+s\tau_2) = 1+s(\tau_1+\tau_2)+s^2\tau_1\tau_2 = 1+s[CR+C_cR+CR_o]+s^2CC_cRR_o.$$

However, if the poles are well-separated:

$$\tau_1 \gg \tau_2$$

and therefore this product can be rewritten as:

$$\begin{aligned}(1+s\tau_1)(1+s\tau_2) &\simeq 1+s\tau_1+s^2\tau_1\tau_2 = \\ &= 1+s[CR+C_cR+CR_o]+s^2CC_cRR_o\end{aligned}$$

thus obtaining:

$$\tau_1 = CR+C_cR+CR_o, \quad \tau_2 = \frac{CC_cRR_o}{\tau_1} = \frac{CC_cRR_o}{CR+C_cR+CR_o}.$$

This is an alternative way of avoiding the solution of the previous second order equation. In this case, the positions of the poles will be:

$$f_{p,LF} = \frac{1}{2\pi\tau_1} = \frac{1}{2\pi(CR+C_cR+CR_o)} \simeq 8 \text{ kHz}$$

$$f_{p,HF} = \frac{1}{2\pi\tau_2} = \frac{CR+C_cR+CR_o}{2\pi CC_cRR_o} \simeq 9.4 \text{ MHz}$$

and we can note that they are not exactly equal to the previous ones. These last two positions, in fact, are slightly less approximated solutions of the second order equation, but they reduce to the first ones under the following assumptions:

$$C_c \ll C, \quad C_c \to 0.$$

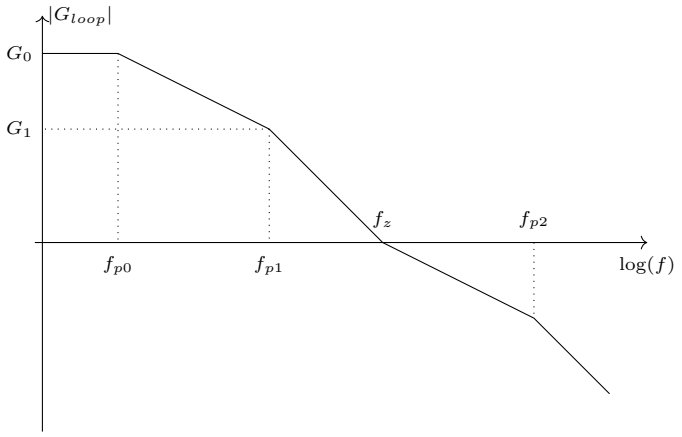


Figure 5.47: Bode diagram of the magnitude of the loop gain in the compensated circuit.

Referring to the Bode diagram of the compensated network that is reported in Figure 5.47, we can calculate:

$$G_0 f_{p0} = G_1 f_{p1}, \quad G_1 f_{p1}^2 = f_z^2$$

obtaining:

$$f_z^2 = (G_1 f_{p1}) f_{p1} = G_0 f_{p0} f_{p1} = A_0 \frac{1}{2\pi\tau} \frac{1}{2\pi(CR + CR_o + C_c R)}$$

and solving for the frequency of the zero we can calculate, from it, the value of the compensation capacitance. Alternatively, in the denominator of the last equation we can neglect the presence of the compensation capacitor C_c, obtaining:

$$\left(\frac{1}{2\pi C_c R}\right)^2 = \frac{A_0}{4\pi^2 \tau C(R + R_o)}$$

thus obtaining[9]:

$$C_c = \frac{1}{R}\sqrt{\frac{\tau C(R + R_o)}{A_0}} \simeq 198 \text{ nF}$$

that is almost equal to the value that we have obtained before. This is, therefore, a way of solving, at least in a first approximation, a problem that would have required, to be solved exactly, more complicated simulations. In general, with these approximate methods it is difficult to obtain an accuracy that higher than a factor of two. Back to the expression of the loop gain:

$$G_{loop} = \frac{A_0}{1 + s\tau} \frac{1 + sC_c R}{1 + s(CR + CR_o + C_c R) + s^2 C C_c R R_o}$$

an alternative possibility is to directly work on the Bode diagram that is represented in Figure 5.47. In fact, at a certain frequency the Bode diagram is an

[9]The result that we have reported have been calculated without this last assumption of neglecting the compensation capacitance in the denominator of the right hand-side of the last equation.

approximation of the magnitude of the loop gain that depends exclusively on the poles and zeros that are placed at a frequency that is lower than the considered one. In this case, before the frequency of the zero, it will not contribute to the loop gain and we can approximate the associated term as:

$$1 + sC_cR \simeq 1.$$

From an analogous reasoning, since we are beyond the pole of the operation amplifier:

$$1 + s\tau \simeq s\tau$$

and considering the denominator, assuming again that the two poles are well separated:

$$\tau_1 \gg \tau_2 : \quad (1 + s\tau_1)(1 + s\tau_2) \simeq 1 + s\tau_1 + s^2\tau_1\tau_2.$$

Near to the frequency of the zero, however, we are beyond the low-frequency pole a before the high-frequency pole, therefore we can write their approximated contributions as:

$$1 + s\tau_1 \simeq s\tau_1, \quad 1 + s\tau_2 \simeq 1$$

thus obtaining:

$$(1 + s\tau_1)(1 + s\tau_2) \simeq s\tau_1.$$

However, from the exact expression of the loop gain, we can observe that the term depending on the first power of s in the denominator will be:

$$s\tau_1 \simeq s(CR + CR_o + C_cR)$$

and therefore the first section of the Bode diagram in which the slope is equal to minus two can be approximated as:

$$G_{loop} \simeq \frac{A_0}{s^2\tau(CR + CR_o + C_cR)}.$$

Setting this approximated loop gain to be equal to one, it is possible to find the frequency of the zero considered:

$$\omega^2 = \frac{A_0}{\tau(CR + CR_o + C_cR)} \;\rightarrow\; f_z = \frac{1}{2\pi}\sqrt{\frac{A_0}{\tau(CR + CR_o + C_cR)}}$$

that is exactly equal to the value that we have found in the previous approximation.

We have thus obtained several different ways of finding an approximate position of the poles and zeros.

Another possible alternative, one might think, was to place the compensation capacitance in parallel to the output resistance R_o. However, it is important to remember that this is impossible in real devices, where the output resistance R_o is actually inside the operation amplifier, thus these pins are not accessible on the device. The same identical reasoning applies in the case we want to place something in series to the input impedance.

A last question might be: is it possible to compensate this circuit by adding a resistor? A possibility, in this case, is to add a compensation resistor R_c between the output node and the capacitor C. In this way, we are slightly changing the

position of the pole of the network and we are adding a zero, that will be placed at:

$$f_p = \frac{1}{2\pi C(R_c + R + R_o)}, \quad f_z = \frac{1}{2\pi C R_c}.$$

In this case, this is actually the simplest possible compensation scheme.

5.4 Multiple feedback loops

In this section, we will deal with the problem of studying the stability of networks in which we have many different feedback loops. In general, this is a quite complicated problem to address, unless we are dealing with two particular cases:

- the two loops are in parallel: there is a common node that can be used as breaking point;
- the two loops are "nested": there are smaller loop that are part of a bigger one and they can be replaced using the associated closed-loop transfer function, thus solving the biggest one.

5.4.1 High-pass amplifier

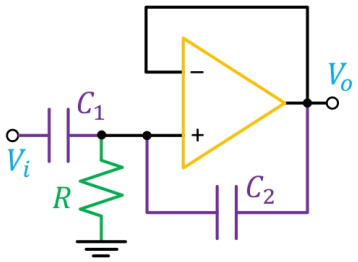


Figure 5.48: The high-pass filter considered.

Considering the high-pass filter represented in Figure 5.48:

$$R = 8\ \text{k}\Omega,\ C_1 = 1\ \text{nF},\ C_2 = 10\ \text{nF}$$

$$GBWP = 10\ \text{MHz},\ A_0 = 120\ \text{dB}$$

calculate the ideal gain, the stability, the loop gain and the closed-loop gain. From the circuit, in an ideal situation, we can write:

$$V^+ = V_i \frac{sC_1R}{1 + sC_1R} = V^- = V_o$$

and therefore, in an ideal circuit, we will not have any current flowing through the capacitor C_2, therefore we can eliminate it. The ideal gain of this network, then, can be written as:

$$G_{id} = \frac{sC_1R}{1 + sC_1R}.$$

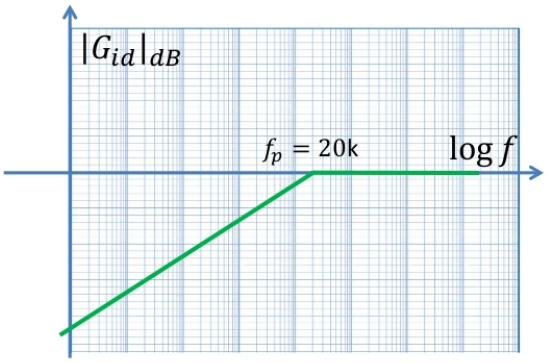


Figure 5.49: Ideal gain of the high-pass filter considered.

The frequency of the pole of the network, then, is equal to:

$$f_p = \frac{1}{2\pi C_1 R} = 20 \text{ kHz}$$

and thus, from an ideal point of view, we have obtained an high-pass filter.

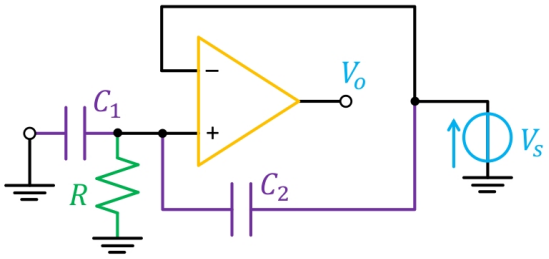


Figure 5.50: Open-loop network for the ideal high-pass filter.

To calculate the loop gain, the capacitor C_2 comes into play. In particular, we can notice the presence of two different feedback loops, one going from the output to the negative input pin, the other going from the output to the positive input pin. The stability of the whole network, therefore, will be ensured when both loop are stable. Since there is a common node for both networks, we can cut the loop at this node, opening both the loops at the same time. Cutting in this point, we obtain that:

$$V^- = V_S$$

and defining the following impedance:

$$Z = C_1 \| R = \frac{\frac{R}{sC_1}}{R + \frac{1}{sC_1}} = \frac{R}{1 + sC_1 R}$$

we obtain that:

$$V^+ = V_S \frac{Z}{Z + \frac{1}{sC_2}} = V_S \frac{R}{1 + sC_1R} \frac{sC_2(1 + sC_1R)}{sC_2R + 1 + sC_1R} =$$
$$= V_S \frac{R}{\cancel{1 + sC_1R}} \frac{sC_2\cancel{1 + sC_1R}}{1 + s(C_1 + C_2)R} =$$
$$= V_S \frac{sC_2R}{1 + s(C_1 + C_2)R}.$$

We can immediately observe that if the source term V_S is grounded then the two capacitors are in parallel:

$$C_1 \| C_2$$

and this is consistent with the result that we have obtained. This gives the following output voltage:

$$V_o = A(s)(V^+ - V^-) = -A(s)\left(1 - \frac{sC_2R}{1 + s(C_1 + C_2)R}\right) V_S =$$
$$= -A(s) \frac{1 + sC_1R}{1 + s(C_1 + C_2)R} V_S$$

thus giving the following loop gain:

$$G_{loop} = -A(s) \frac{1 + sC_1R}{1 + s(C_1 + C_2)R}.$$

In our loop gain, therefore, we have two poles, one that is related to the operation amplifier and the other that comes from the network, and one zero:

$$f_{p0} = \frac{GBWP}{A_0} \simeq 10 \text{ Hz}, \quad f_{p1} = \frac{1}{2\pi(C_1 + C_2)R} \simeq 1.8 \text{ kHz}$$

$$f_z = \frac{1}{2\pi C_1 R} \simeq 20 \text{ kHz}.$$

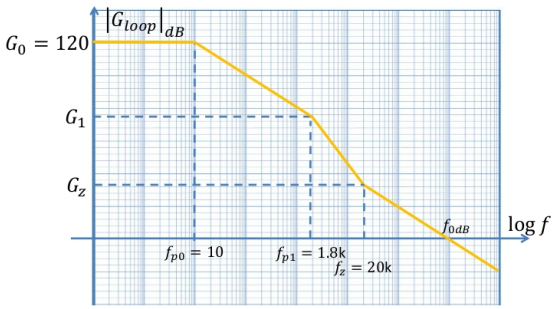


Figure 5.51: Loop gain of the high-pass filter considered.

Considering the Bode diagram of the loop gain that is represented in Figure 5.51, we can calculate:

$$A_0 f_{p0} = G_1 f_{p1} \;\Rightarrow\; G_1 = A_0 \frac{f_{p0}}{f_{p1}} = 5530 \simeq 75 \text{ dB}$$

and analogously:

$$G_1 f_{p1}^2 = G_2 f_z^2 \;\Rightarrow\; G_2 = G_1 \frac{f_{p1}}{f_z^2} = 45 \simeq 33 \text{ dB}.$$

Note that in this case we are crossing the zero decibels axis with a negative and unitary slope, therefore the phase margin will probably be satisfactory. The crossover frequency can be evaluated by writing:

$$f_{0dB} = G_2 f_z \;\Rightarrow\; f_{0dB} = G_2 f_z = 900 \text{ kHz}.$$

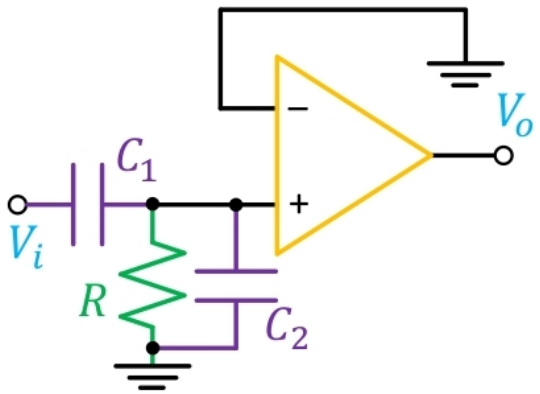


Figure 5.52: Calculation of the open-loop gain of the high-pass filter considered.

Before calculating the closed-loop gain of the network, then, we need to calculate the open-loop gain G_{ol}. To do this, we need to simultaneously disconnect both the loop and we need to ground them, as represented in Figure 5.52. Considering the following impedance:

$$Z = R \| C_2 = \frac{\frac{R}{sC_2}}{R + \frac{1}{sC_2}} = \frac{R}{1 + sC_2 R}$$

this gives the following output voltage:

$$V_o = A(s) \frac{Z}{\frac{1}{sC_1} + Z} V_i = A(s) \frac{\frac{R}{1+sC_2R}}{\frac{1}{sC_1} + \frac{R}{1+sC_2R}} V_i =$$
$$= A(s) \frac{sC_1 R}{1 + s(C_1 + C_2)R} V_i$$

thus giving the following open-loop gain:

$$G_{ol} = A(s) \frac{sC_1 R}{1 + s(C_1 + C_2)R}.$$

Note that since the following relation holds:

$$G_{ol} = -G_{loop} \cdot G_{id}$$

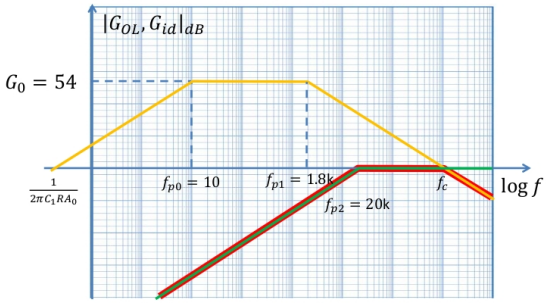


Figure 5.53: Open-loop and ideal gain of the high-pass filter considered.

we could have determined this gain just considering the two previously computed ones.

In this case, we have a zero in the origin and two poles at the following frequencies:

$$f_{p0} = \frac{GBWP}{A_0} = 10 \text{ Hz}, \quad f_{p1} = \frac{1}{2\pi(C_1 + C_2)R} = 18 \text{ kHz}.$$

To compute the gain indicated with G_0 in Figure, we can write an approximation of the open-loop gain when we are far below the first pole:

$$f \ll f_{p0}: \quad G_{ol} = \frac{A_0}{1 + s\tau} \frac{sC_1R}{1 + s(C_1 + C_2)R} \simeq sC_1RA_0$$

and imposing it equal to one we can determine the frequency in which the open-loop gain crosses the zero decibels axis for the first time:

$$sC_1RA_0 = 1 \ \rightarrow \ f_0 = \frac{1}{2\pi C_1 R A_0}.$$

From this value, we can write:

$$\frac{1}{f_0} = \frac{G_0}{f_{p0}}$$

thus obtaining:

$$G_0 = \frac{f_{p0}}{f_0} = \frac{GBWP}{A_0} 2\pi C_1 R A_0 = 2\pi CR \cdot GBWP = 503 \simeq 54 \text{ dB}.$$

An analogous calculation can be written for finding the second crossover frequency:

$$f_c = G_0 f_{p1} = \cancel{2\pi} C_1 \cancel{R} \frac{GBWP}{\cancel{2\pi}(C_1 + C_2)\cancel{R}} \simeq \frac{C_1 GBWP}{C_1 + C_2} \simeq 905 \text{ kHz}.$$

Representing on the same graph also the ideal gain, we can immediately observe that the closed-loop gain G will be the minimum between them.
Considering again the initial circuit, we can try to calculate the input impedance. First of all, we can immediately recognize that the capacitor C_1 is in series

with the rest of the network. Moreover, we will not have any current flowing through the capacitor C_2, therefore we will neglect it. This means that the input impedance, in the ideal case, can be written as:

$$Z_{in} = R + \frac{1}{sC_1}.$$

This is the starting point for the calculation of the real input impedance (that we will not calculate), recognizing that the capacitor C_1 is in series and the resistance R is in parallel to the actual value of the input impedance. Removing these two elements and studying the remaining network, we can immediately observe that it will have an infinite ideal input impedance.

5.4.2 Low-pass filter

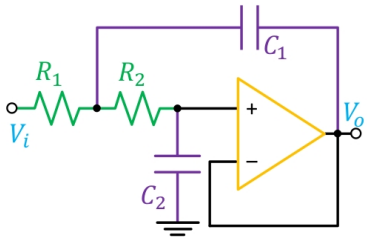


Figure 5.54: The low-pass filter considered.

Considering the low-pass filter in Figure 5.54:

$$R_1 = R_2 = 1\ \text{k}\Omega,\ C_1 = C_2 = 1\ \text{nF},\ GBWP = 10\ \text{MHz},\ A_0 = 100\ \text{dB}$$

calculate the ideal gain, the stability, the loop gain and the closed-loop gain. Studying this circuit, we can immediately observe that:

$$V^+ = V^- = V_o$$

and labelling V_1 the voltage at the node between R_1 and R_2, I_1 the current flowing from C_1 to that node, I_2 the current flowing from R_1 to that node and I_3 the current flowing from that node to R_2 (and then through C_2), we can immediately write:

$$I_1 + I_2 = I_3$$

that gives:

$$(V_o - V_1)sC_1 + \frac{V_i - V_1}{R_1} = \frac{V_1 - V_o}{R_2} = \frac{V_1}{R_2 + \frac{1}{sC_2}} = sC_2V_o$$

or, alternatively:

$$V_o = V_1 \frac{\frac{1}{sC_2}}{R_2 + \frac{1}{sC_2}} = V_1 \frac{1}{1 + sC_2R_2}.$$

In our case:

$$sC_1(V_o - V_1) + \frac{V_i - V_1}{R_1} = sCV_o,\ V_1 = V_o(1 + sC_2R_2)$$

and since:

$$C_1 = C_2 = C,\ R_1 = R_2 = R$$

we can write:

$$sCV_o - sCV_o(1 + sCR) + \frac{V_i}{R} - \frac{V_o}{R}(1 + sCR) = sCV_o$$

$$\frac{V_i}{R} = V_o\left(2sC + \frac{1}{R} + s^2C^2R\right)$$

that gives the following output voltage:

$$V_o = V_i\frac{1}{1 + 2sCR + s^2C^2R^2}$$

from which we can get the following ideal gain:

$$G_{id} = \frac{1}{(1 + sCR)^2}.$$

This means that the network we are studying is a second order low-pass filter with two coincident poles in:

$$f_p = \frac{1}{2\pi CR} \simeq 160 \text{ kHz}.$$

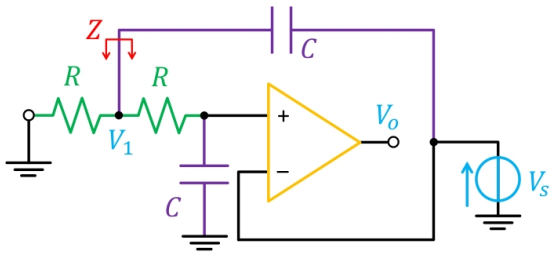


Figure 5.55: Calculation of the loop gain of the low-pass filter considered.

To calculate the loop gain and to study its stability, we can turn off the input source and break the two loops. As in the previous case, the two loop have a common node and we will break the network exactly there, as it is represented in Figure 5.55. Studying the network, we can immediately note that:

$$V^- = V_S$$

and defining the following impedance:

$$Z = R\| \left(R + \frac{1}{sC}\right) = \frac{R^2 + \frac{R}{sC}}{2R + \frac{1}{sC}} = \frac{R(1 + sCR)}{1 + 2sCR}$$

we can write the following previously defined voltage:

$$V_1 = V_S\frac{R(1 + sCR)}{\cancel{1 + 2sCR}}\frac{sC\cancel{(1 + 2sCR)}}{sCR(1 + sCR) + 1 + 2sCR} =$$

$$= V_S\frac{sCR(1 + sCR)}{1 + 3sCR + (sCR)^2}.$$

From this, we can obtain the following value of the voltage at the positive input pin:

$$V^+ = V_1 \frac{\frac{1}{sC}}{R + \frac{1}{sC}} = V_1 \frac{1}{1 + sCR} = V_S \frac{sCR}{1 + 3sCR + s^2C^2R^2}$$

thus giving the following output voltage:

$$V_0 = (V^+ - V^-)A(s) = A(s)V_S \left(\frac{sCR}{1 + 3sCR + s^2C^2R^2} - 1 \right) =$$
$$= -A(s) \frac{1 + 2sCR + s^2C^2R^2}{1 + 3sCR + s^2C^2R^2} = -A(s) \frac{(1 + sCR)^2}{1 + 3sCR + s^2C^2R^2}$$

from which comes the following loop gain:

$$G_{loop} = -A(s) \frac{(1 + sCR)^2}{1 + 3sCR + s^2C^2R^2}.$$

Note that the overall loop gain is negative, since to achieve the stability in general we need to deal with negative feedback loops. If this were not the case, the loop gain G_{loop} should have been much smaller than one to achieve stability. The two capacitors in this network are not interacting one with the other and, since they have the same value, it may seem to be a bad idea to assume the associated poles to be well separated in frequency. However, trying to exploit this assumption anyway, we can first neglect the second order term at the denominator and then neglect the zeroth order term at the denominator, obtaining:

$$f_{p1} \simeq \frac{1}{6\pi CR} \simeq 53 \text{ kHz}, \quad f_{p2} \simeq \frac{1}{\frac{2\pi CR}{3}} \simeq 477 \text{ kHz}.$$

Basically, between the two poles there is a factor of nine, therefore the error in the position of the poles will be in the order of 1/9, that means an 11% of the correct value. These two correct values can be found by computing:

$$\frac{-3 \pm \sqrt{9 - 4}}{2} \cdot \frac{1}{2\pi CR} = \begin{cases} 60.8 \text{ kHz} \\ 416 \text{ kHz} \end{cases}.$$

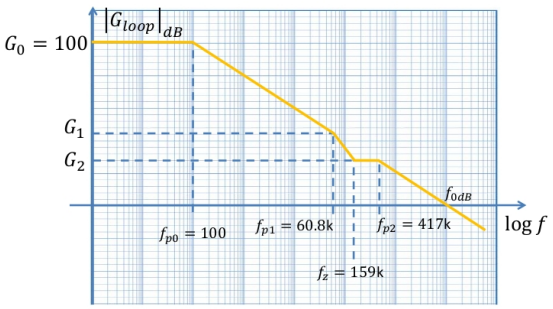


Figure 5.56: Bode plot of the loop gain of the low-pass filter considered.

The two zeros of the loop gain are coincident and they are placed in:

$$f_z = \frac{1}{2\pi CR} \simeq 160 \text{ kHz}.$$

On the Bode plot, we can calculate the frequency of the first zero, that comes from the operation amplifier, as:

$$f_{p0} = \frac{GBWP}{A_0} = 100 \text{ Hz}$$

thus obtaining the following gain:

$$G_0 f_{p0} = G_1 f_{p1} \;\rightarrow\; G_1 = G_0 \frac{f_{p0}}{f_{p1}} = 164 = 44 \text{ dB}.$$

Analogously, for finding the second gain:

$$G_1 f_{p1}^2 = G_2 f_z^2 \;\rightarrow\; G_2 = G_1 \frac{f_{p1}^2}{f_z^2} = 24 = 28 \text{ dB}.$$

Last, we can find the crossover frequency as:

$$G_2 f_{p2} = f_{0dB} \;\rightarrow\; f_{0dB} = G_2 f_{p2} = 100 \text{ MHz} = GBWP.$$

Considering the loop gain of the operation amplifier, we can immediately observe that it can be factorized as the gain of the operation amplifier and a factor:

$$\frac{1 + 2sCR + (sCR)^2}{1 + 3sCR + (sCR)^2}$$

that will be equal to one both in the low-frequency limit and in the high-frequency limit, therefore it is not surprising that we are crossing the zero decibels axis exactly in the gain-bandwidth product of the operation amplifier. The phase margin, since we are crossing with a negative and unitary slope and there is more than a decade between the last pole and the crossover frequency, will be approximately 90°.

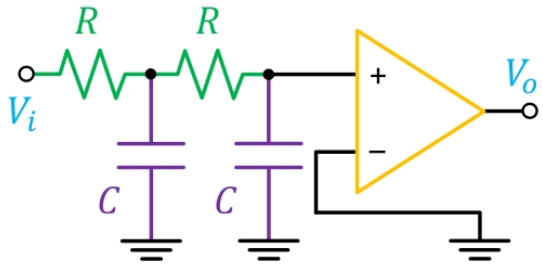


Figure 5.57: Calculation of the open-loop gain of the low-pass filter considered.

We can now calculate the open-loop gain. Instead of considering the network represented in Figure 5.57, we can obtain the loop gain from the following expression:

$$G_{ol} = -G_{loop} G_{id} = A(s) \frac{1 + 2sCR + (sCR)^2}{1 + 3sCR + (sCR)^2} \cdot \frac{1}{(1 + sCR)^2} =$$
$$= A(s) \frac{1}{1 + 3sCR + (sCR)^2}.$$

In the open-loop gain, therefore, we will have three poles in the previously defined positions and, from the associated Bode diagram, we can calculate the following gain terms:

$$G_0 f_{p0} = G_1 f_{p1} \;\rightarrow\; G_1 = G_0 \frac{f_{p0}}{f_{p1}} = 164 = 44 \text{ dB}$$

$$G_1 f_{p1}^2 = G_2 f_{p2}^2 \;\rightarrow\; G_2 = G_1 \frac{f_{p1}^2}{f_{p2}^2} = 24 = 28 \text{ dB}.$$

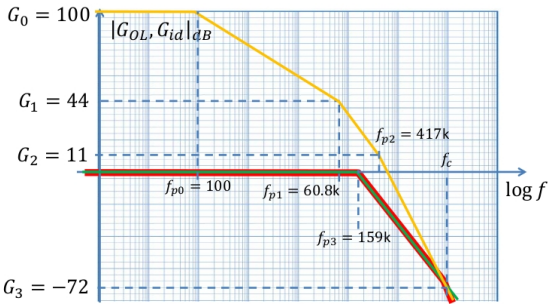


Figure 5.58: Bode plot of the open-loop gain of the low-pass filter considered.

From the expression of the ideal gain, it is then possible to express the overall gain G. The frequency at which the ideal gain is equal to the open-loop gain is the crossover frequency f_{0dB} and, at this frequency, the loop gain will be identically equal to zero. This gives:

$$f_{p3}^2 = G_3 f_{0dB}^2 \;\rightarrow\; G_3 = 2.5 \cdot 10^{-4} = -72 \text{ dB}.$$

5.4.3 Current source

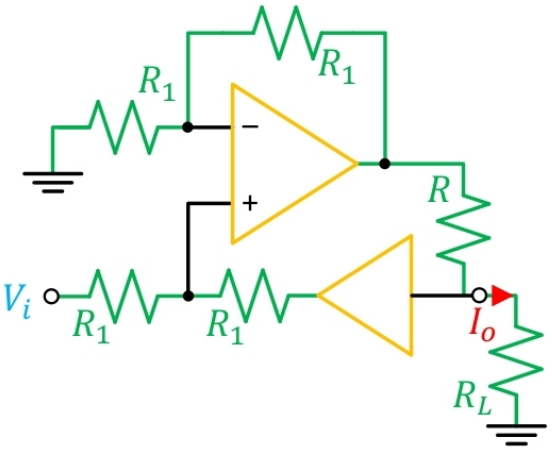


Figure 5.59: Network considered.

Consider the network represented in Figure 5.59:

$$R_1 = 25\ \text{k}\Omega,\ R = 2.5\ \text{k}\Omega,\ GBWP = 5\ \text{MHz}$$

in which we have a certain load R_L and a current as an output; calculate the ideal gain, the loop gain and the output impedance of the network.
The buffer, that is represented in Figure as a triangle with one input and one output, has a unitary gain, an infinite input impedance and a zero output impedance. The ideal gain, then, can be calculated by considering that:

$$V^+ = \frac{V_i + V_o}{2} \ \rightarrow\ V_2 = V_i + V_o$$

and this gives the following current:

$$I_o = \frac{V_2 - V_o}{R} = \frac{V_i + V_o - V_o}{R} = \frac{V_i}{R}$$

from which we get the following ideal gain:

$$G_{id} = \frac{1}{R}.$$

Therefore, the output current I_o is independent, at least from an ideal point of view, from the load resistance R_L and we are dealing with a transconductance amplifier.

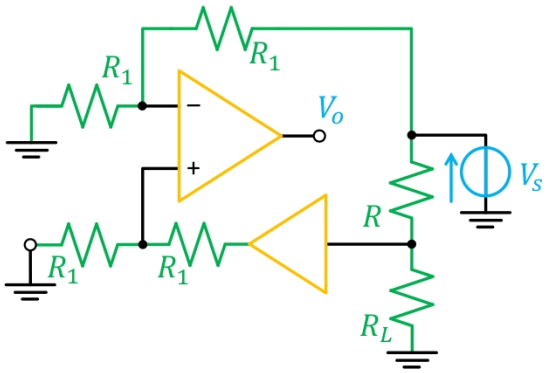


Figure 5.60: Calculation of the loop gain in the amplifier considered.

To calculate the loop gain, we need to cut the network at the end of the operation amplifier as in Figure 5.60, obtaining[10]:

$$V^- = \frac{V_S}{2},\ V_3 = V_S \frac{R_L}{R + R_L},\ V^+ = \frac{V_S}{2}\frac{R_L}{R + R_L}$$

thus obtaining the following output voltage:

$$V_o = A(s) \cdot \left(V^+ - V^-\right) = A(s) \left(\frac{R_L}{2(R + R_L)} - \frac{1}{2}\right) V_S$$

[10]We have defined V_3 the voltage at the input of the buffer stage.

that gives the following loop gain:

$$G_{loop} = -A(s)\frac{R_L}{2(R+R_L)}.$$

Again, the loop gain is negative and this is good for the stability of the network. It is possible to note that there is only one pole that is related to the operation amplifier, therefore the phase margin will be surely 90° and the closed-loop system is stable. If we assume the loop resistance to be small:

$$R_L \ll R \;\Rightarrow\; G_{loop} \simeq -\frac{A(s)}{2}$$

while if it is large:

$$R_L \gg R \;\Rightarrow\; G_{loop} \simeq -\frac{A(s)}{2}\frac{R}{R_L}.$$

This means that everything is fine with respect to the stability of the network, the only drawback is that when we have an high resistive load this gives a reduction of the bandwidth and a reduction of the gain of the loop gain. It is important to note that since $|G_{loop}|^{-1}$ is proportional to the error between the ideal gain and the real one, a reduction in the loop gain gives an increase in the error of the network.

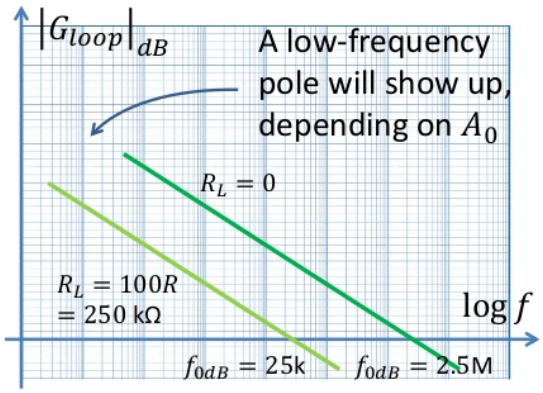


Figure 5.61: Loop gain in the amplifier considered depending on the gain of the operation amplifier.

To calculate the impedance, since we are dealing with the output impedance, we have to get rid of the load resistance R_L and substitute it with a test source. First, we need to obtain the ideal value of the impedance. Using a current source, this gives:

$$V^+ = \frac{V_S}{2},\; V^- = \frac{V_S}{2}$$

and therefore at the input of the buffer:

$$V_3 = \frac{R_1 + R_1}{R_1}V^- = V_S.$$

Therefore, the current is not flowing neither in the buffer nor through the R resistor, thus giving:

$$Z_{id} = \infty.$$

This value makes us understand that it is better to use, as a test source, a voltage source, thus giving the network represented in Figure 5.63.

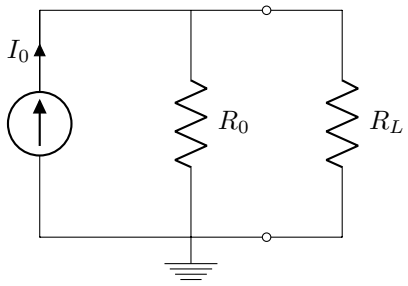


Figure 5.62: Norton equivalent of the network considered as seen from the load resistance.

In an ideal case, we could have considered the Norton equivalent of the circuit as seen from the load resistance R_L, but since we have just calculated that the current flowing in the load resistance is independent from its value:

$$I_0 \frac{R_0}{R_0 + R_L} = I_0$$

this is possible if and only if:

$$R_0 \rightarrow \infty.$$

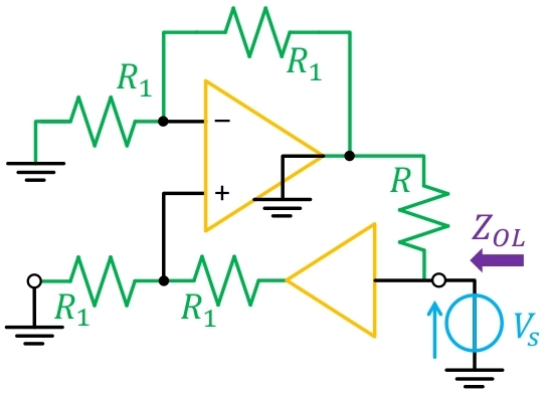


Figure 5.63: Calculation of the output impedance in the amplifier considered.

This means that the output impedance can be written as:

$$Z_{out} = Z_{ol}(1 - G_{loop})$$

where the loop gain comes from the circuit represented in Figure 5.64.

Studying this circuit, we obtain:

$$G_{loop} = -\frac{A(s)}{2}$$

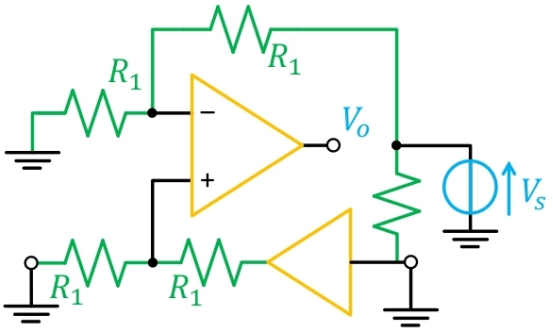


Figure 5.64: Calculation of the loop gain in the amplifier considered.

while setting the voltage controlled voltage source in the operation amplifier equal to zero we can obtain the open-loop resistance:

$$A = 0: \quad Z_{ol} = R.$$

Putting these two results together, we obtain the output impedance:

$$Z_{out} = Z_{ol}(1 - G_{loop}) = R\left(1 + \frac{A(s)}{2}\right).$$

At low frequencies, this gives:

$$Z_{out} \simeq R\left(1 + \frac{A_0}{2}\right) \simeq 10 \text{ G}\Omega.$$

Alternatively, we could have used the Blackman's formula, in which we have to calculate two loop gains, one "short-circuited" in which the load resistance is replaced by a short-circuit, the other "open circuited" in which the load resistance is replaced by an open circuit. In the short-circuit case:

$$G_{loop}|_{sc} = -\frac{A(s)}{2}$$

as before, while in the open loop case:

$$V^+ = \frac{V_S}{2}, \; V^- = \frac{V_S}{2} \;\rightarrow\; V_o = 0 \;\rightarrow\; G_{loop}|_{oc} = 0.$$

Therefore, using this formula we obtain again the previous result:

$$Z = Z_{ol}\frac{1 - G_{loop}|_{sc}}{1 - G_{loop}|_{oc}} = R\left(1 + \frac{A(s)}{2}\right).$$

It is important to remember that, when using this formula, one of the two loop gains is always expected to be equal to zero, otherwise there is something wrong. At first sight, this formula seems to be easier; however, it will work only if one of the two loop gains will have an ideal value of zero or infinity. If there is something in parallel or in series to the network, it will not work.

We can now replace the buffer with the equivalent non-ideal network, as in Figure 5.65. In particular, it is important to note that the loop in this second

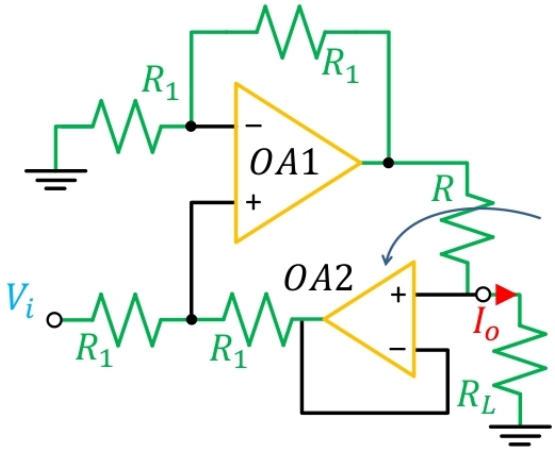


Figure 5.65: Network with a non-ideal buffer stage.

operation amplifier (identified as $OA2$) is not in parallel to the other loops of the network. To study its stability, therefore, we need to consider all the loops that are present. First of all, we can consider the first operation amplifier as an ideal amplifier and study the stability of the second one. Cutting the loop at the end of the second operation amplifier, we can observe that[11]:

$$V_2^- = V_S,\ V_1^+ = \frac{V_S}{2},\ V_1^- = \frac{V_S}{2},\ V_3 = V_S$$

where we have considered that the network having as an input the positive input pin of the first operation amplifier and as an output the output pin of the first operation amplifier can be considered as an ideal stage with gain two. This gives:

$$V_2^+ = V_S \frac{R_L}{R + R_L}$$

thus giving the following output:

$$V_o = A(s) \cdot \left(\frac{R_L}{R + R_L} - 1 \right) V_S = -A(s) \frac{R}{R + R_L} V_S$$

from which we obtain the following loop gain for the second operation amplifier:

$$G_{loop,2} = -A(s) \frac{R}{R + R_L}.$$

It is important to note that this loop gain is different from the one that we have obtained previously when studying the other loop when the second operation amplifier was considered as an ideal one. Moreover, also this second loop is stable.

However, the fact that we are considering alternatively one operation amplifier or the other as an ideal stage is not completely correct. Considering again the circuit represented in Figure 5.65, we want now to study the effects of the fact

[11] Where again V_3 is referred to the output voltage of the first operation amplifier.

that both operation amplifiers are not ideal and then to calculate the impedance of the first operation amplifier[12]. To study the effect of having a non-ideal operation amplifier, we can replace the buffer with its closed-loop transfer function (or gain):

$$G = \frac{G_{ol}}{1 - G_{loop}} = \frac{A(s)}{1 + A(s)} \xrightarrow[A(s)\to+\infty]{} 1.$$

Considering the network, this gives:

$$V_1^- = \frac{V_S}{2},\ V_1^+ = \frac{R_L}{R + R_L} T(s) \frac{1}{2} V_S,\ T(s) = \frac{1}{1 + s\tau_2}$$

and therefore the output voltage of the first operation amplifier can be written as:

$$V_o = \frac{A(s)}{2} \cdot \left(1 - \frac{\frac{R_L}{R+R_L}}{1 + s\tau_2} \right) = \frac{A(s)}{2} \cdot \frac{\frac{R}{R+R_L} + s\tau_2}{1 + s\tau_2}.$$

From this expression it is then possible to obtain the loop gain G_{loop} when also the second operation amplifier is not an ideal one. The willing student can try to do the opposite.

5.5 Different configurations of the Wheatstone bridge

5.5.1 Wheatstone bridge and instrumentation amplifier

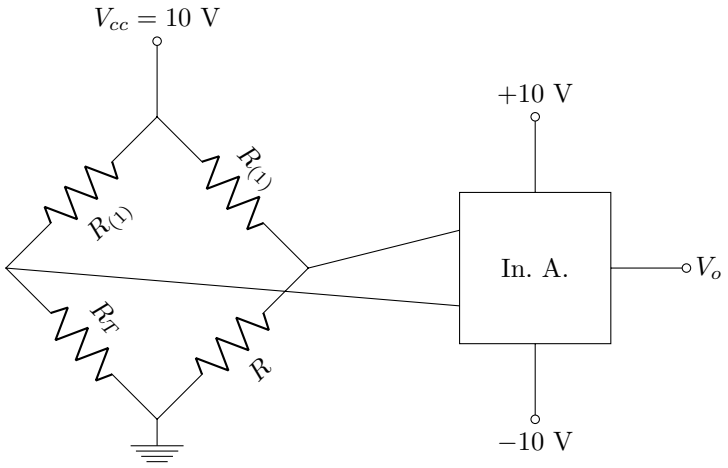


Figure 5.66: Wheatstone bridge connected to an instrumentation amplifier.

Considering the Wheatstone bridge represented in Figure 5.66:

$$R = R_{(1)} = 30\ \Omega,\ R_T = 30 + 0.15T = 30\left(1 + \frac{0.15}{30}T\right) = 30(1 + xT)$$

[12] This study has been done in the solution to the exam of November 14th, 2011.

where we have the following temperature and output voltage range (or dynamic of the sensor):

$$T = 0 - 100^{\circ}\text{C},\ V_o = 0 - 10\ \text{V}$$

calculate the gain of the instrumentation amplifier and try to keep the dissipated power below the following threshold value:

$$P_{dis} < 1\ \text{mW}$$

in each one of the resistive elements of the network.
The input of the instrumentation amplifier can be written as:

$$V_i = \frac{V_{cc}}{4} x$$

where we have defined:

$$x = \frac{0.15}{30} T$$

therefore we obtain:

$$V_i = \frac{V_{cc}}{4} \alpha T,\ \alpha = \frac{0.15}{30} = 5 \cdot 10^{-3}\ ^{\circ}\text{C}^{-1}.$$

The maximum value of the input of the instrumentation amplifier then can be written as:

$$T = 100^{\circ}\text{C}\ \rightarrow\ V_i = \frac{V_{cc}}{4} \alpha \cdot 100 = 1.25\ \text{V}$$

and therefore the gain that is needed is:

$$G = \frac{V_{o,max}}{V_{i,max}} = \frac{10}{1.25} = 8.$$

The maximum value of the relative variation of the resistance will be:

$$x_{max} = 5 \cdot 10^{-3} \cdot 100 = 0.5$$

and we can observe that this is a pretty large value. In this case, since the relative variation is so large, we have to check the linearity error that is involved in this approximation:

$$\epsilon = \frac{x}{2} \simeq 0.25 = 25\%$$

and this is a quite big error.
In a more correct way, therefore, we can use the exact value for the output voltage of the Wheatstone bridge. At the maximum possible temperature:

$$R_{T,max} = 30 + 0.15 \cdot 100 = 45\ \Omega$$

and therefore:

$$V_i = V_{cc} \left(\frac{R_{T,max}}{R + R_{T,max}} - \frac{1}{2} \right) = V_{cc} \left(\frac{45}{30 + 45} - \frac{1}{2} \right) = 1\ \text{V}$$

and thus the gain needed is:

$$G = 10.$$

We can immediately observe that this is different from what we have obtained in the previous case, where the voltage V_i at the input of the operation amplifier was 1.25 V, thus being off of 25%.
To study the power dissipation, we can observe that the worst case for the power dissipation takes place at 0°, where the variable resistors R_T has its maximum value and therefore we have the maximum current flow. In this condition, the power dissipated over each resistive element can be written as:

$$P = I^2 R = \left(\frac{V_{cc}}{2R}\right)^2 = 833 \text{ mW} \gg 1 \text{ mW}.$$

We are therefore quite far from the requirement on the power dissipation. To match it, the simplest possibility is to reduce the bias voltage of the Wheatstone bridge V_{cc}:

$$P = \frac{V_{cc}^2}{4R} \leq 1 \text{ mW}$$

thus obtaining:

$$V_{cc} \leq 200 \text{ mV}.$$

The disadvantage, in this case, is that the output of the bridge is directly proportional to the bias voltage of the bridge, therefore we are also reducing the signal by a factor of 50. This means that we have to increase the gain of the instrumentation amplifier, since the new input voltage to this device will be:

$$V_{i,max} = \frac{1 \text{ V}}{50} = 20 \text{ mV} \;\Rightarrow\; G_{max} = 500.$$

However, this might not be possible due to the presence of the noise, that could lead to a complete cancellation of the signal.
Alternatively, it is possible to observe that the power dissipation is controlled by the current flowing in the resistors, that can thus be reduced by increasing the series $R + R_T$ in the worst temperature condition, at 0°C. Since the variable resistor R_T is fixed from the problem, we can unbalance the bridge by increasing the values of the two upper resistance R (that in this second part of the problem will be called $R_{(1)}$). In fact, the lower right resistor cannot be changed, otherwise the bridge will be unbalanced. We can thus write the input voltage of the instrumentation amplifier as:

$$V_i = V_{cc}\left(\frac{R(1+x)}{R(1+x)+R_1} - \frac{R}{R+R_1}\right)$$

where in this case:

$$R_1 > R.$$

Defining the following coefficient:

$$\frac{R}{R+R_1} = k$$

we can write the following input voltage of the instrumentation amplifier:

$$V_i = V_{cc}\left(\frac{R(1+x)}{R(1+x)+R_1} - k\right) = V_{cc}\left(\frac{R(1+x)}{(R+R_1)\left(1+\frac{Rx}{R+R_1}\right)} - k\right) =$$
$$= V_{cc}\left(\frac{k(1+x)}{1+kx} - k\right) = V_{cc}\left(\frac{\cancel{k} + kx - \cancel{k} - k^2x}{1+kx}\right) =$$
$$= V_{cc}\left(\frac{kx(1-k)}{1+kx}\right) \simeq V_{cc}kx(1-k)$$

where in the last equivalence we have assumed that both k and x are small. From the constraint on the power dissipation:

$$P_T = \left(V_{cc}\frac{R}{R+R_1}\right)^2 \frac{1}{R} = \frac{V_{cc}^2}{R}k^2 = 1 \text{ mW}$$

we can obtain:

$$k \simeq 1.7 \cdot 10^{-2} \;\Rightarrow\; \frac{R}{R_1} \simeq 57.$$

The maximum value of the output of the bridge, in this condition, can be written as:

$$V_{i,max} = V_{cc}x_{max}k(1-k) \simeq 85 \text{ mV}$$

thus giving the following requirement on the gain:

$$G \simeq \frac{10 \text{ V}}{85 \text{ mV}} \simeq 117.$$

This is a better solution with respect to the previous one, since the power dissipation is low enough to satisfy the requirement but the gain needed for the instrumentation amplifier is smaller than in the previous case.

5.5.2 Wheatstone bridge and operation amplifiers

Considering the network represented in Figure 5.67, where:

$$R = 250\ \Omega,\ R_T = R(1+\alpha T),\ \alpha = 2 \cdot 10^{-4}\ ^\circ\text{C}^{-1}$$

choose the best output of the network between V_1 and V_2 for measuring a variation of the temperature T in the variable resistor in the bridge and then find the bias voltage V_{cc} of the Wheatstone bridge that keeps the power dissipation over a resistor R limited:

$$P_R < 1.5 \text{ mW}.$$

First of all, from the network we can note that:

$$V_1^+ = 0 = V_2^- = V_2^+$$

and therefore we will have:

$$V_1^- = 0$$

thus giving the following current:

$$I = \frac{V_{cc}}{R}.$$

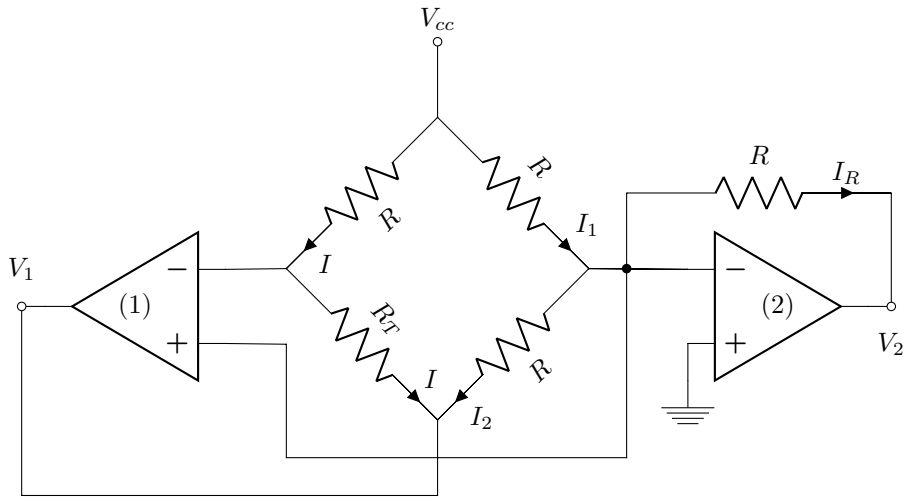


Figure 5.67: Network considered.

From this, recognizing that the first operation amplifier is in an inverting configuration where R_T is the feedback resistor and R is the resistor between the inverting pin and the input voltage V_{cc}, we can obtain the output voltage of this stage as:

$$V_1 = -IR_T = -\frac{R_T}{R}V_{cc} = -V_{cc}(1+\alpha T).$$

From this, we can obtain the following currents:

$$I_1 = \frac{V_{cc}}{R},\ I_2 = \frac{V_{cc}(1+\alpha T)}{R} = -\frac{V_1}{R}$$

thus obtaining:

$$I_R = I_1 - I_2 = \frac{V_{cc}}{R} - \frac{V_{cc}}{R}(1+\alpha T) = -\frac{V_{cc}}{R}\alpha T.$$

Therefore, the output of the second operation amplifier can be written as:

$$V_2 = -RI_R = V_{cc}\alpha T.$$

From the expression of the two output voltages V_1 and V_2, we can see that both are linear with respect to the temperature and, moreover, they have the same sensitivity. However, the output of the first operation amplifier V_1 has a certain bias, constant signal that is superimposed to the variation related to the temperature, therefore it is better to choose V_2 as the output of the network:

$$V_o = V_2.$$

Note that the result we have obtained is perfectly linear: the two operation amplifiers linearise the output of the bridge and they increase of a factor four the sensitivity of a standard Wheatstone bridge.
The power dissipated on a resistor will be maxima at 0°C, where:

$$R_T = R$$

and thus it can be written as:

$$P_d = \frac{V_{cc}^2}{R} < 1.5 \text{ mW}$$

thus giving the following limitation for the bias voltage of the bridge:

$$V_{cc} < 612 \text{ mV}.$$

If we want, for example, to be able to discriminate the following variation in temperature:

$$\Delta T_{min} = 0.1^\circ\text{C}$$

the we must have:

$$V_{2,min} = V_{cc}\alpha\Delta T = 12 \ \mu\text{V}.$$

5.5.3 Strain gauge

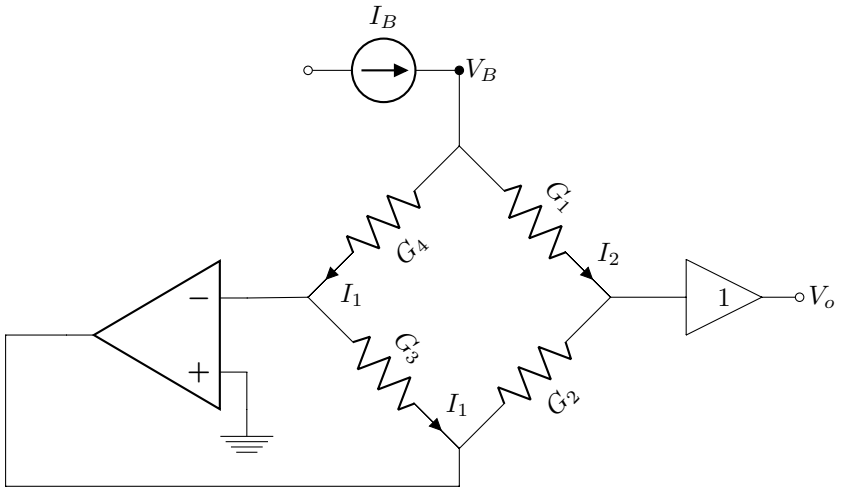


Figure 5.68: Circuit considered.

Consider the strain sensor represented in Figure 5.68, where the variable resistors G_1 and G_2 are place on different sides of the sensor, thus experiencing a strain with different sign (one is elongating, the other is contracting). Assuming the following values to be known:

$$G = 2, \ R(0^\circ\text{C}) = 150 \ \Omega, \ \alpha = 3 \cdot 10^{-3} \ ^\circ\text{C}^{-1}$$

consider what happens when we have a change ΔT in the temperature of the device and what happens when we have a deformation of G_1 and G_2.
To evaluate the temperature variation ΔT, we can assume to not have any mechanical deformation and that the temperature is uniform over the whole bridge. In this case, any resistor can be written as:

$$R = R_0(1 + \alpha T)$$

under the assumption of uniformity of the temperature over the bridge. We have thus four equal elements and, if V_B is the bias voltage in the upper part of the bridge, due to the same considerations of the previous exercise we can set the bottom node of the bridge and the output of the amplifier at $-V_B$. This allows us to evaluate the following current:

$$I_1 = \frac{V_B}{R}$$

and this implies:

$$I_2 = \frac{V_B}{R}.$$

However, since:

$$I_1 = I_2 = \frac{I_B}{2}$$

we can obtain the bias voltage:

$$V_B = \frac{RI_B}{2}$$

and this condition gives:

$$V_o = V_B - RI_2 = 0.$$

This circuit, therefore, is automatically compensating the temperature variations.
When G_1 and G_2 changes due to strains:

$$\epsilon = \frac{\Delta l}{l}$$

we can write the variation in the resistance as:

$$R = R_0(1 \pm x) = R_0(1 \pm G \cdot \epsilon)$$

where the sign depends on the direction of the strain considered (elongation or compression). Since the bias voltage V_B is not changed, the whole left hand-side arm of the bridge is not changed. Studying the variations on the other arm:

$$I_2 = \frac{2V_B}{R_{G1} + R_{G2}} = \frac{2V_B}{R_0(1+x) + R_0(1-x)} = \frac{2V_B}{2R_0} = \frac{V_B}{R_0}.$$

Therefore, also the current flowing in this arm is equal due to the fact that the bias voltage V_B has not changed. Then, the value of the output voltage:

$$V_o = V_B - R_{G1}I_2 = \frac{R_0 I_B}{2} - \frac{I_B}{2}R_0(1-x) = \frac{R_0 I_B}{2}G\epsilon.$$

also in this case, therefore, the output is fully linear with respect to the quantity that we want to measure.

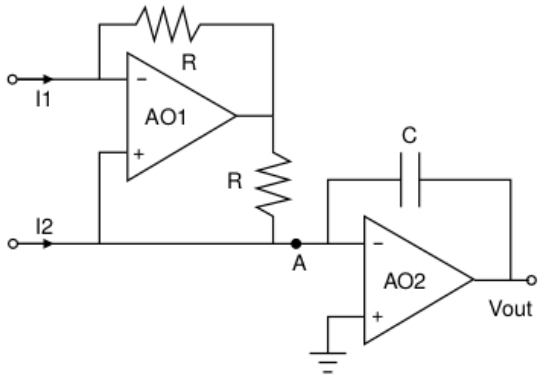


Figure 5.69: Network considered.

5.5.4 An exercise on multiple feedback loops

This exercise comes from the exam of September 9th, 2009. Consider the voltage controlled network represented in Figure 5.69, where:

$$R = 1\ \text{k}\Omega,\ C = 160\ \text{nF},\ A_0 = 10^5,\ f_{oa} = 10\ \text{Hz}$$

and calculate the ideal gain and the loop gain of the network[13].
From the circuit, we can immediately observe that:

$$V_2^- = 0 = V_2^+$$

and this gives:

$$V_1^+ = V_1^- = 0.$$

Therefore, since the current I_1 can only flow through the feedback resistance R of the first operation amplifier, defining V_1 to be the output of the first operation amplifier we will have:

$$V_1 = -I_1 R.$$

Studying now where the current I_2 can flow, we can immediately observe that across the resistor R connecting the output of the first operation amplifier to the inverting pin of the second one we will have a current equal to I_1:

$$\frac{V_1}{R} = I_1$$

and thus the current I_c flowing through the feedback capacitor of the second operation amplifier will be:

$$I_c = I_2 + \frac{V_1}{R} = I_2 - I_1.$$

This allows us to calculate the output voltage of the second operation amplifier:

$$V_o = -\frac{I_c}{sC} = -\frac{I_2 - I_1}{sC} = \frac{I_1 - I_2}{sC}.$$

[13]Before starting, it is necessary a hint: never try to compute a current balance at the output of an operation amplifier. In fact, it is determined by a voltage controlled voltage source that, being an ideal voltage source, can sustain every possible current.

Since we are interested in the difference between the two input currents in the time domain, we can see that this network will give an output that is proportional to the integral of that quantity.
The loop gain, then, have to be discussed depending on the different loops that we are considering, checking first the stability of the loops around the first operation amplifier (assuming the second one to be ideal) and then the stability of the loops around the second one (assuming the first one to be ideal).

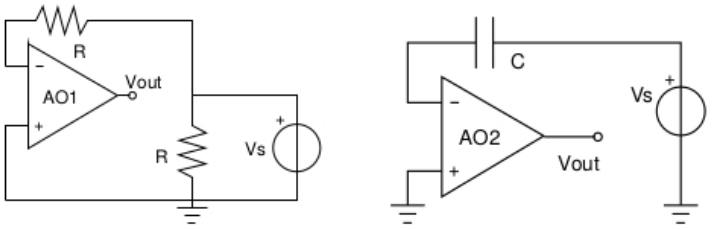


Figure 5.70: Computation of the loop gain of the first operation amplifier and of the second one when the other is considered to be an ideal operation amplifier.

To study the loop gain of the first operation amplifier when the second one is an ideal operation amplifier, we can consider that the only effect of the second operation amplifier is to set the inverting pin of the second operation amplifier at ground. According to this consideration, then, we can break the first loop and study the equivalent network that is represented in the left hand-side part of Figure 5.70. In this case:

$$V_1^- = V_S,\ V_1^+ = 0$$

and the output voltage will be:

$$V_o = -A(s)V_S$$

thus giving the following loop gain:

$$G_{loop,1} = -A(s).$$

In the second case, when the first operation amplifier is assumed to be an ideal one, then both the input pins of the first operation amplifier will have an infinite input impedance and thus we will not have any current flowing through them. This gives:

$$V_1^+ = V_1^- = V_1 = V_2^-$$

and therefore the equivalent network is the one represented in the right hand-side of Figure 5.70. However, since we cannot have any current flowing through the capacitor, we obtain that:

$$V_2^- = V_S$$

that gives the following output voltage:

$$V_o = -A(s)V_S$$

from which we obtain the following loop gain:

$$G_{loop,2} = -A(s).$$

We can now try to remove these ideal characteristics of the operation amplifiers, thus calculating the two loop gains when the other operation amplifier is not an ideal one. Using the hint given in the exam, we first can try to compute the equivalent impedance of the non-ideal closed-loop operation amplifier.

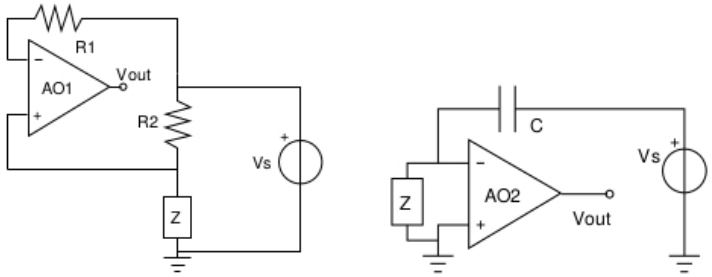


Figure 5.71: Computation of the loop gain of the first operation amplifier and of the second one when the other is not considered to be an ideal operation amplifier and it is replaced by the associated equivalent impedance.

First of all, we can try to compute the loop gain of the second operation amplifier when the first one is represented by its equivalent impedance, as it is represented in the right hand-side part of Figure 5.71. Before computing the loop gain, therefore, we have to study the value of this equivalent impedance. In the ideal case, as we have seen previously, the input impedance of the first operation amplifier is infinite, there will not be any current flowing from the negative pin of the second operation amplifier toward the rest of the network and therefore the associated ideal impedance is:

$$Z_{id} = \infty.$$

This means that the equivalent impedance can be written as:

$$Z = Z_{ol}(1 - G_{loop}).$$

Considering therefore only the left hand-side part of the network and connecting a test source where once we had the negative input pin of the second operation amplifier, we can shut off the voltage controlled voltage source inside the first operation amplifier and calculate the open-loop impedance as:

$$Z_{ol} = R.$$

Cutting then the loop at the output of the operation amplifier, applying a test source and grounding the negative input pin of the second operation amplifier, we can immediately observe that the loop gain of this network will be:

$$G_{loop} = -A(s).$$

Therefore, the equivalent impedance for the first loop when the first operation amplifier is not assumed to be ideal is equal to:

$$Z = R(1 + A(s)).$$

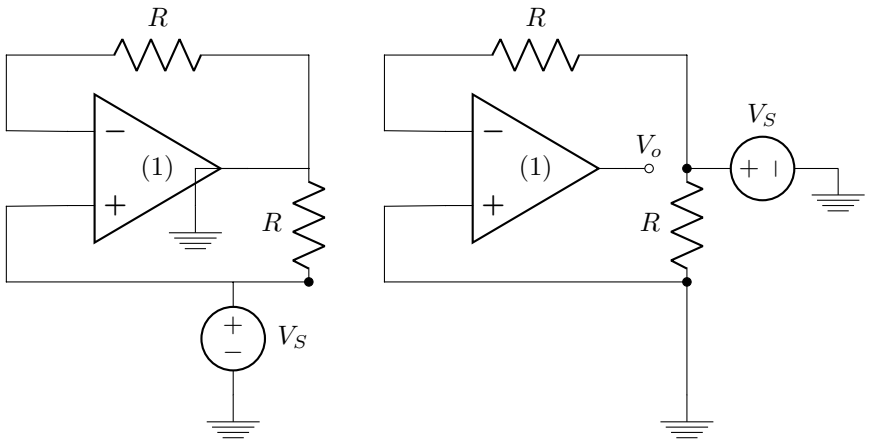


Figure 5.72: On the left, computation of the open-loop impedance equivalent to the first loop; on the right, computation of the loop gain for the first loop.

Studying the remaining circuit, that is represented in the right hand-side of Figure 5.71 we obtain:

$$V_2^- = V_S \frac{Z}{Z + \frac{1}{sC}} = \frac{sCZ}{1 + sCZ} V_S = \\ = \frac{sCR(1 + A(s))}{1 + sCR(1 + A(s))} V_S$$

that leads to the following loop gain[14]:

$$G_{loop,2} = -A(s) \frac{sCR\left(1 + \frac{A_0}{1+s\tau}\right)}{1 + sCR + \frac{sCRA_0}{1+s\tau}} = \\ = -\frac{A_0}{1 + s\tau} \cdot \frac{sCRA_0\left(1 + s\frac{\tau}{A_0}\right)}{1 + sCRA_0 + s^2\tau CR}.$$

From this expression, we can immediately observe that we will have two zeros (one of them in the origin) and three poles. Aside from the pole given by the operation amplifier, we can calculate the other two in an approximate way assuming them to be well separated. In this case, for the low-frequency one:

$$1 + sCRA_0 \simeq 0 \ \rightarrow \ f_{pL} = \frac{1}{2\pi CRA_0} \simeq 10^{-2} \text{ Hz}$$

while for the high-frequency one:

$$sCRA_0 + s^2\tau CRA_0 \simeq 0 \ \rightarrow \ f_{pH} = \frac{A}{2\pi\tau} = GBWP = 1 \text{ MHz}.$$

We can immediately observe that their approximate positions are very well separated, therefore this approximation is meaningful. The frequency of the

[14] A lot of calculations, that should be trivial algebra, have been skipped.

zero can then be calculated as:

$$f_z = \frac{1}{2\pi \frac{\tau}{A_0}} = \frac{A_0}{2\pi\tau} = GBWP = 1 \text{ MHz}.$$

Therefore, this zero and the high-frequency pole will be one near to the other.

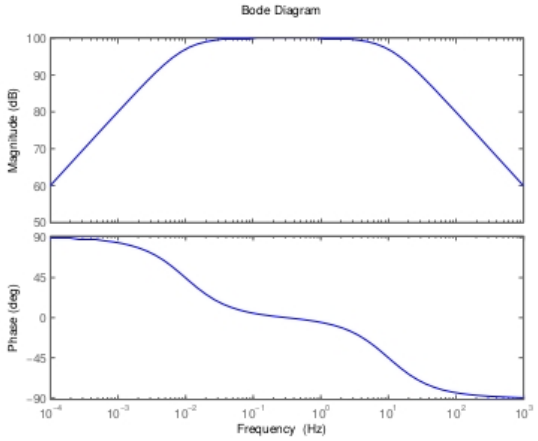


Figure 5.73: Bode diagram of the loop gain obtained when we are not considering any operation amplifier as an ideal one.

The low-frequency behaviour of this loop gain can be approximated as:

$$G_{loop} \simeq sCRA_0^2$$

therefore this loop gain will be crossing the zero decibels axis for the first time at:

$$sCRA_0^2 = 1 \;\rightarrow\; f_{0dB} = \frac{1}{2\pi CRA_0^2} = 10^{-7} \text{ Hz}.$$

We can then calculate the gain at which we have the *plateau*:

$$\frac{1}{f_{0dB}} = \frac{G_1}{f_{pL}} \;\rightarrow\; G_1 = \frac{f_{pL}}{f_{0dB}} = 10^5 = 100 \text{ dB}.$$

Alternatively, we could have considered that in the flat region we are far before the high-frequency poles and the zero, thus approximating the loop gain as:

$$G_{loop} \simeq \frac{sCRA_0^2}{1 + sCRA_0}.$$

Moreover, at higher frequencies but again before the high-frequency pole and zero, the loop gain will be similar to the gain of the operation amplifier, therefore it will cross the zero decibels axis exactly in the gain-bandwidth product $GBWP$.

Replacing now the second operation amplifier with its equivalent impedance, we can immediately calculate the ideal value of this impedance:

$$Z_{id,2} = 0$$

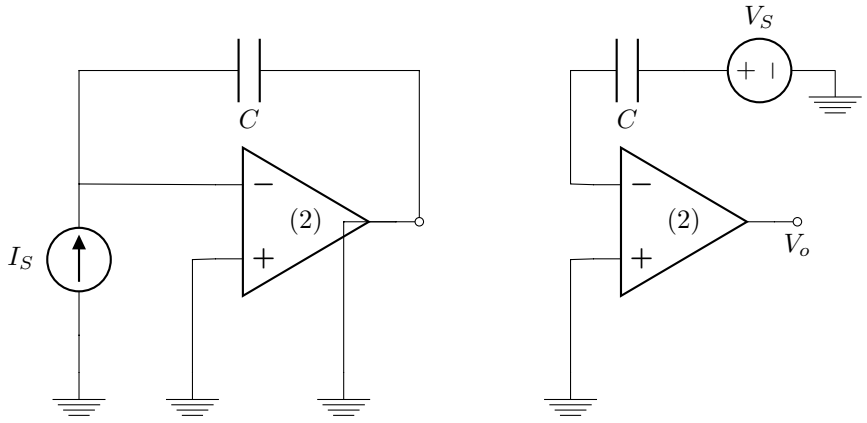


Figure 5.74: On the left, computation of the open-loop impedance for the second part of the network; on the right, computation of the loop gain for that part of the network.

thus giving the following expression for the equivalent impedance:

$$Z_2 = \frac{Z_{ol}}{1 - G_{loop}}.$$

Since we have obtained that the ideal impedance is equal to zero, to calculate the open-loop impedance we need to drive the device with a current source, as represented in Figure 5.74. Therefore, grounding the output of the operation amplifier we can obtain:

$$Z_{ol} = \frac{1}{sC}$$

while the loop gain will be:

$$G_{loop} = -A(s).$$

This gives the following equivalent impedance for this network:

$$Z_2 = \frac{\frac{1}{sC}}{1 + A(s)}.$$

Alternatively, we could have used Blackman's formula to derive this expression. Considering now the loop around the first operation amplifier, that is represented in the left hand-side of Figure 5.71, we can write:

$$V^- = V_S,\ V^+ = V_S \frac{Z_2}{Z_2 + R}$$

thus obtaining as an output voltage:

$$V_o = -A(s)\left(1 - \frac{Z_2}{R + Z_2}\right) V_S = -A(s)\frac{R}{R + Z_2} V_S$$

and this gives the following loop gain:

$$G_{loop,1} = -A(s)\frac{R}{R + Z_2} = -A(s)\frac{R}{R + \frac{1}{sC(1+A(s))}}.$$

Expanding this calculation, we can find that:

$$G_{loop,1} = G_{loop,2}$$

but it is important to remember that this is not a general result: it holds only for this circuit.

5.5.5 Another exercise on multiple feedback loops

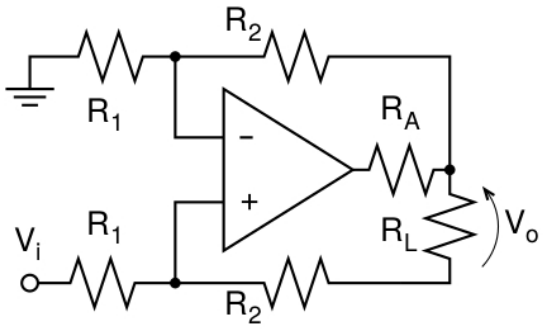


Figure 5.75: Multiple loop network considered.

Considering the network represented in Figure 5.75, where:

$$R_A = 10\ \Omega,\ R_1 = 1\ \text{k}\Omega,\ R_2 = 10\ \text{k}\Omega$$

compute the ideal gain of the network, its output impedance and the condition on the load resistance that ensures a static precision better than 1%.
Assuming initially the operation amplifier to be ideal, we can study the circuit and observe that:

$$V^- = V_1 \frac{R_1}{R_1 + R_2} = V^+$$

and therefore this gives the following current:

$$I = \frac{V_i - V^+}{R_1} = \frac{V_i - V_2}{R_1 + R_2}.$$

This allows us to calculate:

$$V^+ = V_2 + R_2 I = V_2 + \frac{R_2}{R_1 + R_2} V_i - \frac{R_2}{R_1 + R_2} V_2 =$$
$$= \frac{R_2}{R_1 + R_2} V_i + \frac{R_1}{R_1 + R_2} V_2 = V_1 \frac{R_1}{R_1 + R_2}$$

and therefore we obtain that:

$$V_i \frac{R_2}{R_1 + R_2} = \frac{R_1}{R_1 + R_2}(V_1 - V_2) = \frac{R_1}{R_1 + R_2} V_o.$$

This gives therefore the following output voltage:

$$V_o = \frac{R_2}{R_1 + R_2} \frac{R_1 + R_2}{R_1} V_i = \frac{R_2}{R_1} V_i$$

that corresponds to the following ideal gain:

$$G_{id} = \frac{R_2}{R_1}.$$

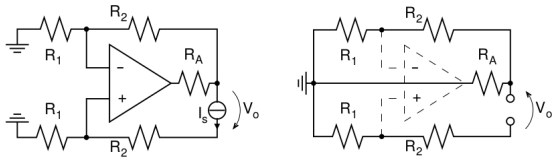


Figure 5.76: Computation of the ideal output impedance and of the open-loop output impedance.

To compute the output impedance of the circuit, we first have to consider that the load resistance is not part of the network and thus we have to calculate the ideal output impedance. From the circuit represented in Figure 5.76, we can obtain that:

$$V_2 = I_S(R_1 + R_2),\ V^+ = I_S R_1 = V^-$$

and since the current I_S is flowing also in the upper resistors R_1 and R_2 we have that:

$$V_1 = I_S(R_1 + R_2) \ \rightarrow \ V_S = V_2 - V_1 = 0$$

and therefore the ideal impedance will be:

$$Z_{id} = 0.$$

This means that the output impedance of this network will be written as:

$$Z = \frac{Z_{ol}}{1 - G_{loop}}.$$

Calculating the open-loop impedance, again from Figure 5.76 we obtain that:

$$Z_{ol} = [(R_1 + R_2) \| R_A] + R_1 + R_2 = R_1 + R_2 + \frac{R_A(R_1 + R_2)}{R_1 + R_2 + R_A} \simeq$$
$$\simeq R_1 + R_2 + R_A \simeq R_1 + R_2 \simeq 11 \text{ k}\Omega.$$

To calculate the loop gain, we will not have any load resistance R_L and breaking the loop at the output of the operation amplifier and applying a test signal to the breaking point, we can obtain that since V_2 is floating we will not have any current flowing through the lower resistors R_1 and R_2, thus giving:

$$V^+ = 0.$$

In the upper part of the network, on the other hand:

$$V^- = V_S \frac{R_1}{R_1 + R_2 + R_A}$$

that gives the following output voltage:

$$V_o = -A(s) \frac{R_1}{R_1 + R_2 + R_A} V_S$$

from which we obtain the following loop gain:

$$G_{loop} = -A(s)\frac{R_1}{R_1 + R_2 + R_A} \simeq -A(s)\frac{R_1}{R_1 + R_2}.$$

From this we can then obtain the impedance as:

$$Z = \frac{R_1 + R_2}{1 + A(s)\frac{R_1}{R_1 + R_2}}.$$

Alternatively, we could have used the Blackman's formula for obtaining the same result.
Last, we have to study the static precision of this network. Neglecting the resistor R_A since it is small, we can observe that the error between the ideal gain and the one that we can obtain is equal to $1/G_{loop}$. This means that the requirement that we have formulated on the static precision of this network is equivalent to the following one:

$$G_{loop} > 100.$$

Before starting all the calculations, we can observe that if R_L is extremely small we are compromising the behaviour of the loop, while if it tends to infinity it will not give any problem; therefore, we expect to find a lower bound for the load resistance. Cutting the loop at the output of the operation amplifier and applying a test signal V_S, we can immediately obtain that:

$$V^+ = V_S\frac{R_1}{R_1 + R_2 + R_L},\ V^- = V_S\frac{R_1}{R_1 + R_2}$$

and this gives:

$$\begin{aligned} G_{loop} &= -A(s)\left(-\frac{R_1}{R_1 + R_2 + R_L} + \frac{R_1}{R_1 + R_2}\right) = \\ &= -A(s)R_1\frac{R_L}{(R_1 + R_2 + R_L)(R_1 + R_2)} \end{aligned}$$

that can be evaluated in the following two limits:

$$G_{loop} \xrightarrow[R_L \to \infty]{} -A(s)R_1, \quad G_{loop} \xrightarrow[R_L \to 0]{} 0.$$

In static conditions, if the load resistance is small:

$$G_{loop} \simeq -A_0\frac{R_1 R_L}{(R_1 + R_2)^2} > 100$$

and this gives:

$$R_L > 100\frac{(R_1 + R_2)^2}{R_1 A_0} = 1.2\ \text{k}\Omega.$$

5.6 Noise transfer and OAs

5.6.1 Exercise 1

Given the circuit represented in Figure 5.77, where:

$$R_1 = 1\ \text{k}\Omega,\ R_2 = 100\ \text{k}\Omega,\ A_0 = 120\ \text{dB},\ GBWP = 10^7\ \text{Hz}$$

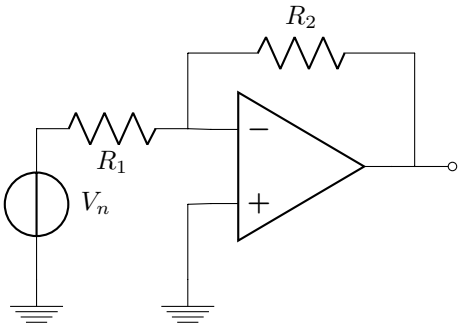


Figure 5.77: Circuit considered.

with the following noise sources:

$$\sqrt{S_v} = 18 \text{ nV}/\sqrt{\text{Hz}}, \quad \sqrt{S_i} = 10 \text{ pA}/\sqrt{\text{Hz}}$$

discuss the noise transfer of this network, computing the output total noise. Recognizing that this network is an inverting amplifier, we can write the ideal gain of this network as:

$$G = -100.$$

Actually, then, we have many different noise sources that we need to take into account. First of all, we have the noise sources associated to the thermal noise in the resistors whose power spectral density is, for example for the first resistor:

$$S_{v1} = 4k_B T R_1.$$

We can thus consider the noise as if it were a signal, finding the associated transfer function and calculating then the associated square modulus. In this case, considering the voltage equivalent noise source, the transfer function will be:

$$V_o = -\frac{R_2}{R_1} V_n$$

and therefore, considering the associated power spectral density:

$$S_{vo} = \left(-\frac{R_2}{R_1}\right)^2 S_v = 10^4 S_v.$$

Since we then want to compute the mean square value of the output noise:

$$\overline{V_o^2} = \int_0^{+\infty} S_{vo} \, df$$

since in circuits we will always be dealing only with unilateral power spectral densities. The problem, now, is that if we consider the constant power spectral density at the input this integral is clearly diverging. In this case, this means that we have to remove an approximation: the fact that we are dealing with the ideal gain of the circuit, neglecting the fact that the operation amplifier is a real one. The real operation amplifier, therefore, will include a pole and therefore also the real gain will include this pole that will make the previous integral finite.

Since, by definition, the real gain is the minimum between the ideal gain and the open-loop one, we can cut the loop at the output of the operation amplifier and inject a test signal to the inverting pin of the operation amplifier, obtaining the following open-loop gain:

$$G_{ol} = -A(s)\frac{R_2}{R_1 + R_2} \simeq -A(s) = -\frac{A_0}{1 + s\tau}.$$

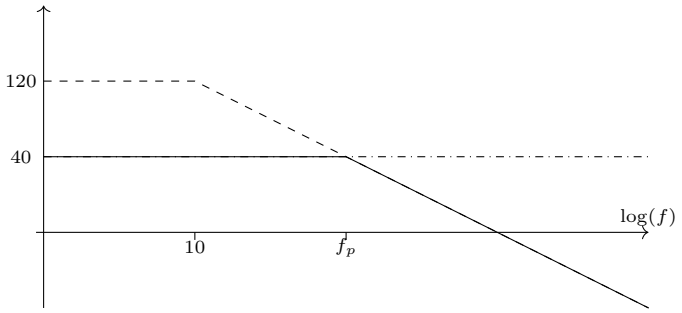


Figure 5.78: Ideal gain (dashed and dotted line), open-loop gain (dashed line) and real gain (solid line) for the circuit considered.

The pole in the real gain, then, will be place at:

$$f_p G_{id} = GBWP \;\rightarrow\; f_p = \frac{GBWP}{G_{id}} = 10^5 = 100 \text{ kHz}$$

therefore we can write the noise transfer function as:

$$V_o = -\frac{100}{1 + s\tau_c} V_n, \quad \tau_c = \frac{1}{2\pi f_p},\ f_p = 100 \text{ kHz}.$$

Alternatively, we could directly have calculated the loop gain and the associated crossover frequency f_p. Squaring the modulus of this transfer function, we obtain:

$$S_{vo} = \frac{10^4 S_v}{1 + (\omega\tau_c)^2}$$

and thus we obtain:

$$\begin{aligned}\overline{V_o^2} &= \int_0^{+\infty} \frac{10^4 S_v}{1 + (\omega\tau_c)^2}\,\frac{d\omega}{2\pi} = \frac{10^4 S_v}{2\pi}\left(\arctan(\omega\tau_c)|_0^{+\infty}\right) = \\ &= 10^4 S_v \frac{\pi}{2} f_p \simeq 2.6 \cdot 10^{-8} \text{ V}^2\end{aligned}$$

and this is the contribution of the first resistance R_1.

For the second resistance, considering the circuit represented in Figure 5.79, form an analogous reasoning we can observe that the ideal gain from this noise source is unitary:

$$G_{id} = 1.$$

Also in this case, therefore, we have to compute the real gain of the network, otherwise this result will diverge. Disconnecting the output and using the voltage equivalent noise source as a test source, we can find the open-loop gain as:

$$G_{ol} = -\frac{R_1}{R_1 + R_2} A(s) \simeq -\frac{R_1}{R_2} A(s)$$

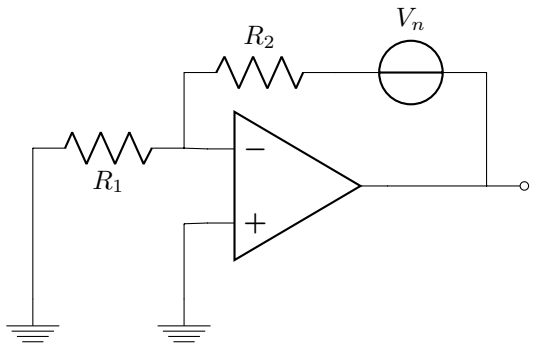


Figure 5.79: Noise contribution for the second resistance.

and this can be represented in a Bode plot as in Figure 5.80.

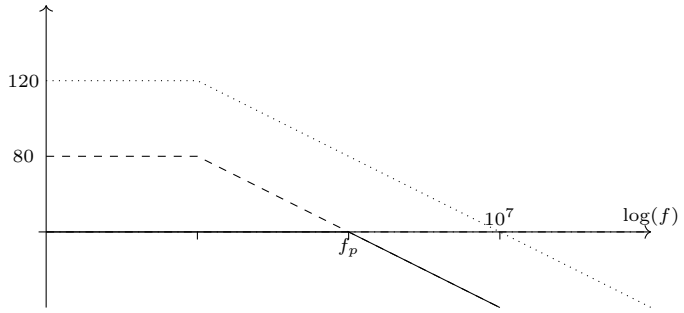


Figure 5.80: Ideal gain (dashes and dots), open-loop gain (dashes), gain of the OA (dots) and real gain (solid line) for the circuit considered.

As it is expected, also in this case the pole of the real gain is placed exactly in the same position of the pole of the previous real gain; in fact, it identifies the frequency at which the loop gain is unitary and we know that the loop gain is independent from the position of the input considered. In this case, for example, the loop gain will be:

$$G_{loop} = -\frac{R_1}{R_1 + R_2} A(s)$$

and thus the transfer function will be:

$$V_o = \frac{V_n}{1 + s\tau_c} \;\Rightarrow\; S_{vo} = \frac{S_v}{|1 + s\tau_c|^2}.$$

Integrating this quantity we can get the mean square value of the output noise associated to this noise source:

$$\overline{V_o^2} = \int_0^{+\infty} S_{vo}\, df = S_v \frac{\pi}{2} f_p = 2.6 \cdot 10^{-10}\ \mathrm{V}^2.$$

Notice that there are two orders of magnitude between the two noise terms. In particular, even though the two resistors are different by two order of magnitudes:

$$R_1 = 10^{-2} R_2$$

the noise terms will have an inverse ratio because of the amplification of the thermal noise in resistor R_1 with respect to the one coming from resistor R_2. The factor 10^2, therefore, will come from a 10^4 amplification multiplied by the 10^{-2} difference in the two input noise power spectral densities.

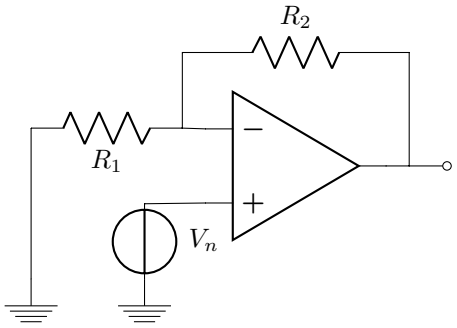


Figure 5.81: Noise equivalent voltage source for the noise in the operation amplifier.

To consider the noise equivalent voltage source for the operation amplifier we need to study the circuit represented in Figure 5.81. In this case, we can immediately see that the output voltage can be written as:

$$V_o = \frac{R_1 + R_2}{R_1} V_n$$

and therefore computing the square modulus of this transfer function:

$$S_{vo} = \left(\frac{R_1 + R_2}{R_1} \right)^2 S_v.$$

Also in this case, this is a white power spectral density and therefore we need to calculate the real gain in order to not have an infinite result. Also in this case, from this computation we will obtain an additional pole placed at f_p since the loop gain of the network is always the same regardless of the position of the input:

$$S_{vo} = \frac{\left(\frac{R_1+R_2}{R_1} \right)^2}{|1 + s\tau_c|^2} S_v$$

thus obtaining the following expression for the mean square value of the output noise:

$$\overline{V_o^2} = S_v \left(\frac{R_1 + R_2}{R_1} \right)^2 \frac{\pi}{2} f_p = 1.6 \cdot 10^{-7} \text{ V}^2.$$

To consider the noise equivalent current sources, we have to study the two networks represented in Figure 5.82. This noise equivalent current source can be placed either at the positive input pin or at the negative input pin of the amplifier. As we can clearly see from the Figure, the noise equivalent current source placed at the positive input pin will not give any contribution to the overall output noise due to the fact that the input impedance of the operation amplifier is infinite. Considering the noise equivalent current source at the

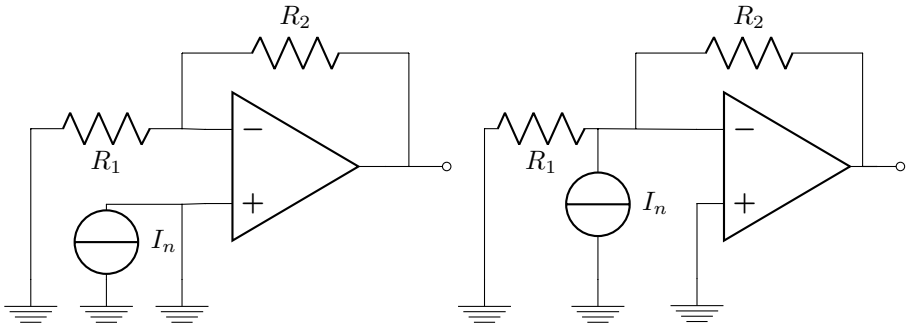


Figure 5.82: Noise equivalent current sources for the noise in the operation amplifier.

negative input pin, the associated transfer function will be:

$$V_o = -R_2 I_n$$

and thus we obtain:

$$S_{vo} = (-R_2)^2 S_I.$$

Also in this case the noise is white, therefore we have to consider the real transfer function, that will be:

$$S_{vo} = \frac{R_2^2}{|1 + s\tau_c|^2} S_I$$

thus giving the following mean square value of the output noise:

$$\overline{V_o^2} = \int_0^{+\infty} S_{vo}\, df = S_I R_2^2 \frac{\pi}{2} f_p \simeq 1.6 \cdot 10^{-7}\ \mathrm{V}^2.$$

By linear superposition, since the noise processes are random processes and they are uncorrelated one with respect to the other, we can sum the associated variances, obtaining:

$$\overline{V_o^2} = 2.6 \cdot 10^{-8} + 2.6 \cdot 10^{-10} + 1.6 \cdot 10^{-7} + 1.6 \cdot 10^{-7} \simeq 3.46 \cdot 10^{-7}\ \mathrm{V}^2.$$

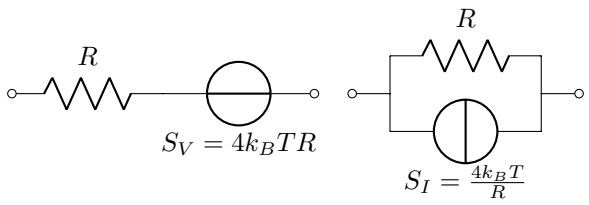


Figure 5.83: Thermal noise equivalent sources.

Before ending this exercise, it is worthy to make a few considerations. First of all, we can note that the thermal noise is a resistor can be represented, as in Figure 5.83, either with its Thévenin equivalent (thus through a voltage source)

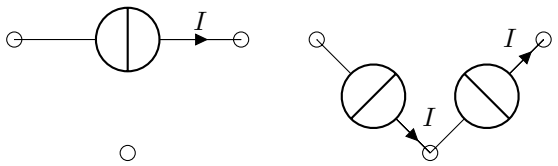


Figure 5.84: Equivalence between circuits.

or through its Norton equivalent (thus using a current source). This depends on our particular choice and the convenience in the circuit we are dealing with.

Considering now three different nodes in a generic network and a current source that is connecting two of them, we can observe that we can draw an equivalent scheme of the network as in Figure 5.84. This means that we can break any current source as the series of two different current sources without affecting any property of the network.

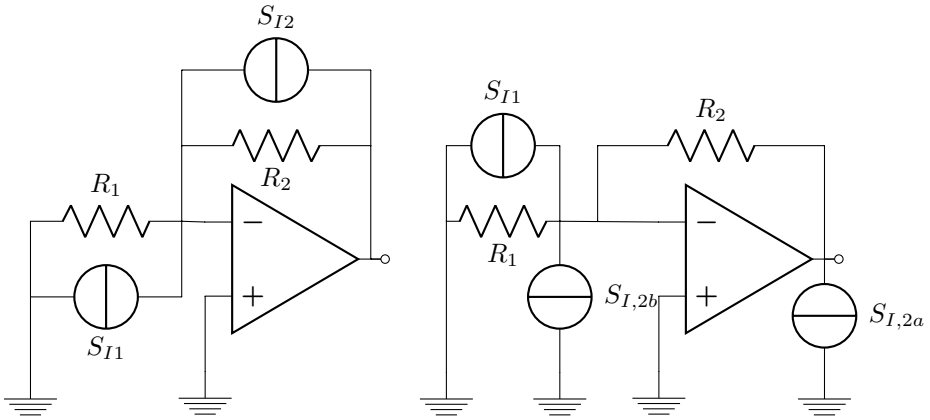


Figure 5.85: Equivalent representation of the circuit considering the thermal noise equivalent current sources in both the resistors.

We can now apply these properties as in Figure 5.85. In this drawing, we have considered the thermal noise equivalent current sources for both the resistors and we have split the source for the resistor R_2 in two different source that have, as a common pin, the ground. Applying the linear superposition, we can first consider only the current source whose power spectral density has been indicated with $S_{I,2a}$ and observe that the output will always remain a zero: it will not give any contribution. Considering now the current source with power spectral density $S_{I,2b}$, we can see that it will be in parallel to the noise equivalent current source for the thermal noise in resistor R_1. We can then note that also the noise equivalent current source for the negative input pin of the operation amplifier will be in parallel to these sources, thus giving an equivalent circuit that is represented in Figure 5.86.

From this representation, we can immediately obtain that:

$$V_o = (I_{T_1} R_2 + I_{T_2} R_2 + I_{OA} R_2)$$

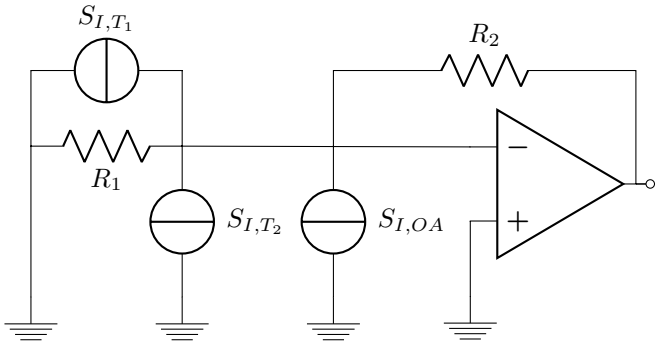


Figure 5.86: Equivalent noise representation of the source for the thermal noise in both the resistors and for the current equivalent noise source in the operation amplifier.

from which comes:

$$S_{vo} = S_{I,T_1} R_2^2 + S_{I,T_2} R_2^2 + S_{I,OA} R_2^2 = \left(S_{I,T_1} + S_{I,T_2} + S_{I,OA}\right) R_2^2$$

where we have neglected the double products since the noise terms are uncorrelated one with the other.

5.6.2 Exercise 2

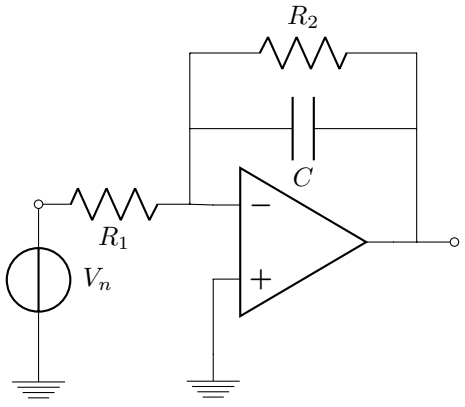


Figure 5.87: Circuit considered (where we have already considered the noise equivalent voltage source for the thermal noise in the first resistor).

Consider the circuit represented in Figure 5.87, where:

$$R_1 = 1\ \text{k}\Omega,\ R_2 = 1\ \text{M}\Omega,\ C = 100\ \text{nF},\ A_0 = 120\ \text{dB}$$

$$GBWP = 10\ \text{MHz},\ \sqrt{S_v} = 100\ \text{nV}/\sqrt{\text{Hz}},\ \sqrt{S_i} = 10\ \text{pA}/\sqrt{\text{Hz}}$$

and evaluate the associated noise.
This circuit is an approximated integrator and the first noise term that can

be studied is the thermal noise in the resistor R_1. From the given circuit, the associated transfer function will be:

$$V_o = -V_n \frac{R_2 \| \frac{1}{sC}}{R_1} = -V_n \frac{R_2}{R_1} \cdot \frac{1}{1 + sCR_2}$$

thus obtaining the following output power spectral density:

$$S_{vo} = S_v \left(\frac{R_2}{R_1} \right)^2 \frac{1}{|1 + sCR_2|^2}.$$

We can immediately note that, due to the presence of a capacitor, this power spectral density has already a pole and, therefore, its integral will be finite. In this case, therefore, we do not need to calculate the real gain of the circuit, since it will only add another high-frequency pole, that will give only a small correction to this calculation. Since we do not want to add useless complexity to this already complex problem, we can evaluate the mean square value of the output noise for the thermal noise in the resistor R_1 as:

$$\overline{V_o^2} = 4k_B T R_1 \cdot \left(\frac{R_2}{R_1} \right)^2 \frac{\pi}{2} \frac{1}{2\pi C R_2} \simeq 4 \cdot 10^{-11} \text{ V}^2.$$

The same result could have been obtained by using a noise equivalent current source, since the transfer function would have been:

$$V_o = -I_n \cdot \left(R_2 \| \frac{1}{sC} \right) = -I_n \frac{R_2}{1 + sCR_2}$$

and thus the output power spectral density:

$$S_{vo} = S_i \frac{R_2^2}{|1 + sCR_2|^2} = \frac{4k_B T}{R_1} \cdot \frac{R_2^2}{|1 + sCR_2|^2}.$$

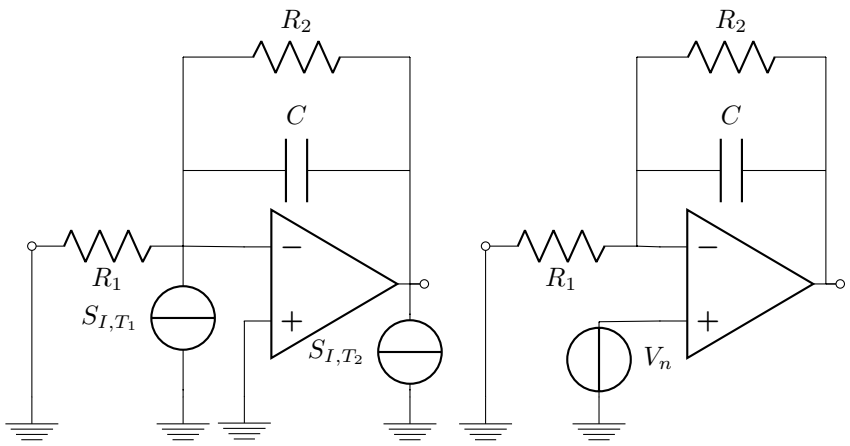


Figure 5.88: On the left, noise equivalent current sources for the thermal noise in R_2; on the right, noise equivalent voltage source for the operation amplifier.

Considering now the thermal noise in R_2 as in Figure 5.88, since the current source whose power spectral density has been indicated with S_{I,T_2} is not relevant

(since it will not change the output voltage), we can focus only on the first power spectral density, obtaining, as in the previous case:

$$S_{Vo} = S_{I,T_1} \frac{R_2^2}{|1 + sCR_2|^2}$$

and since its integral is again finite we do not have to consider the real gain of the network, obtaining the following mean square value of the output:

$$\overline{V_o^2} = S_{I,T_2} \frac{\pi}{2} R_2^2 f_p = \frac{4k_B T}{R_2} R_2^2 \frac{1}{4CR_2} \simeq 4 \cdot 10^{-14} \text{ V}^2.$$

For the current source of the amplifier, the only contributing term will be the one connected to the negative input pin (since the positive one will see only the infinite input impedance of the operation amplifier) and its contribution will be formally identical (apart from the power spectral density) to the previous one:

$$\overline{V_o^2} = S_i R_2^2 \frac{1}{4CR_2} \simeq 2.5 \cdot 10^{-10} \text{ V}^2.$$

Last, for the noise equivalent voltage source for the operation amplifier, we can write it as:

$$\begin{aligned} V_o &= V_n \left(1 + \frac{R_1 \| \frac{1}{sC}}{R_1}\right) = V_n \left(1 + \frac{R_2}{R_1(1 + sCR_2)}\right) = \\ &= V_n \left(\frac{R_1 + R_2 + sCR_1R_2}{R_1(1 + sCR_2)}\right) = \frac{R_1 + R_2}{R_1} \frac{1 + sC(R_1 \| R_2)}{1 + sCR_2} V_n \end{aligned}$$

and evaluating that:

$$\frac{R_1 + R_2}{R_1} \simeq 10^3, \quad R_1 \| R_2 \simeq R_1$$

we could square this relationship and evaluate the output power spectral density. However, since we have more than one singularity in this expression, this integral is diverging. We need thus to consider the real gain, that will add an extra pole at high frequency. We know, in fact, that the transfer function of the noise source must always go to zero in the high frequency limit, otherwise leading to an unphysical behaviour. We can evaluate the frequency of the additional pole f_c by computing it as the crossover frequency for the loop gain:

$$G_{loop}(f_c) = 1$$

and therefore eliminating every source term from the network, cutting the loop at the output of the operation amplifier and adding a voltage test source, we can write it as:

$$G_{loop} = -A(s) \frac{R_1}{R_1 + \frac{R_2}{1 + sCR_2}} = -A(s) \frac{R_1(1 + sCR_2)}{R_1 + R_2 + sCR_1R_2}$$

and since we are way above the position of the pole and of the zero, we can say that the crossover frequency will be at:

$$f_c = GBWP.$$

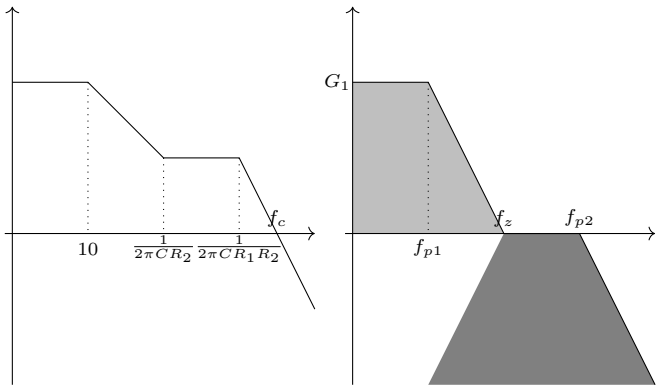


Figure 5.89: On the left, Bode diagram of the loop transfer function; on the right, Bode diagram of the real gain and its representation as piecewise single-pole transfer functions.

The real gain transfer function can thus be evaluated as in Figure 5.89 and approximating it with the area underlying the constant parts[15], we can write the mean square value of the output noise as:

$$\overline{V_o^2} = S_v G_1^2 \frac{\pi}{2} f_{p1} + S_v 1^2 \frac{\pi}{2}(f_{p2} - f_z)$$

where this approximation holds if the poles and the zeros are well separated one from the other. In this case, therefore, the mean square value of the output noise will be:

$$\overline{V_o^2} = 2.5 \cdot 10^{-8} + 1.5 \cdot 10^{-7} = 1.75 \cdot 10^{-7} \text{ V}^2.$$

The final mean square value of the output noise can then be written as the sum of all the previous different contributions:

$$\overline{V_o^2} = 4.1 \cdot 10^{-11} + 4.1 \cdot 10^{-14} + 2.5 \cdot 10^{-10} + 2.5 \cdot 10^{-8} + 1.5 \cdot 10^{-7} \simeq 1.75 \cdot 10^{-7} \text{ V}^2.$$

The root mean square value of the output noise, then, will be:

$$\sqrt{\overline{V_o^2}} \simeq 425 \ \mu\text{V}.$$

5.6.3 Exercise 3

This exercise comes from the exam of July 20th, 2010. Considering the network represented in Figure 5.90, where:

$$A_0 = 10^6, \ GBWP = 1 \text{ MHz}, \ G = 100, \ R = 10 \ \Omega, \ R_L \leq 1 \text{ k}\Omega$$

calculate the ideal gain, the loop gain and discuss the stability of the network. Then calculate the output impedance of the network.
From a direct inspection of this network, we can immediately write that:

$$V^- = V_i$$

[15]Remember that any time you study the area underlying a transfer function you can approximate it with its flat parts, considering them just as if they were single-pole portions.

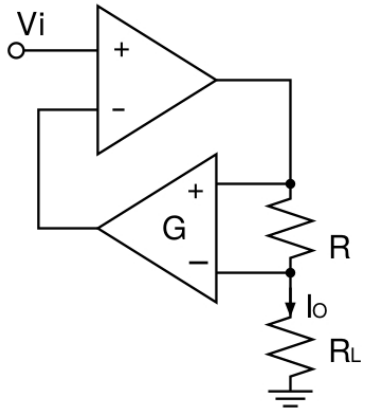


Figure 5.90: Representation of the circuit considered.

and therefore:

$$V_i = GI_0R \;\rightarrow\; I_0 = \frac{V_i}{GR}.$$

This gives that the ideal gain is:

$$G_{id} = \frac{1}{GR} = \frac{1}{100R} = 10^{-3}\ \Omega^{-1}.$$

To compute the loop gain, we can cut the loop at the output of the operation amplifier and apply, to the same node, a test voltage source V_S. In this case, again investigating the network, we obtain that:

$$V^- = G\frac{R}{R+R_L}V_S$$

from which the output voltage:

$$V_o = -A(s)\frac{R}{R+R_L}GV_S$$

that gives the following loop gain:

$$G_{loop} = -A(s)\frac{R}{R+R_L}G = -A(s)\frac{100R}{R+R_L}.$$

For the stability, the operation amplifier $A(s)$ has one pole while the other term is constant, therefore the system is expected to be stable with a phase margin of 90°. However, since we have a range of values in which we can select the output resistance:

$$R_L \in [0,1]\ \text{k}\Omega$$

we can evaluate the loop gain for these two extreme values:

$$R_L = 1\ \text{k}\Omega:\ G_{loop} \simeq -A(s), \quad R_L = 0:\ G_{loop} \simeq -100A(s)$$

and in both cases the system will be stable. However, from the gain-bandwidth product of the operation amplifier:

$$GBWP = 1\ \text{MHz}$$

we can immediately see that if the loop gain is higher (as in the case $R_L = 0$) we can move the crossover frequency at a much higher frequency, as it can be represented in a suitable Bode diagram. From the theory of the dominant pole compensation, we know that above the gain-bandwidth product of an operation amplifier we will have a lot of other poles whose contribution is, in general, not relevant. However, if the loop gain is too high, we are actually moving the crossover frequency in higher frequency regions and this will possibly lead an unstable system, since one or more of these poles can come into play. We are actually making the leading pole compensation not effective. As a general rule, the crossover frequency of the loop gain should not exceed the gain-bandwidth product of the operation amplifier. In principle, in fact, this network seems to be stable, but nobody knows what is happening above the dominant pole of the operation amplifier.

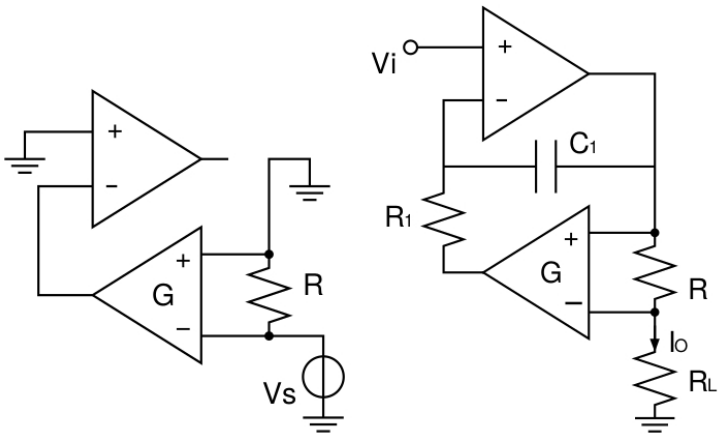


Figure 5.91: On the left, circuit needed for the computation of the open-loop impedance; on the right, compensated circuit.

For the computation of the output impedance, the first thing that we can immediately do is to replace the load resistance R_L, that will not be part of the network, with a test source. Since this device is, actually, a current source, we expect to have, in an ideal case, an infinite output impedance, therefore we can try first to apply a voltage test source. In the ideal case, the two pins of the operation amplifier will be at the same voltage and equal to zero, therefore the two input pins of the gain stage will be at the same voltage and thus there will not be any current flowing through the resistor R. Since we cannot neither have a current flowing through the input pins of the gain stage (that being an ideal stage has an infinite output impedance), we conclude that in the ideal case:

$$I_S = 0$$

and thus that the ideal output impedance is equal to:

$$Z_{id} = \infty.$$

This means that the closed-loop output impedance will be written in the following form:

$$Z = Z_{ol}(1 - G_{loop})$$

where the loop gain has to be calculated without the load resistance R_L. To compute the open-loop impedance, we can cut the loop at the output of the operation amplifier and observe that the only resistance seen from the voltage test source is:

$$Z_{ol} = R.$$

For the computation of the loop gain, we can replace this test source with a short-circuit to the ground and cutting the loop and applying a test voltage source in the direction of the loop we obtain that:

$$V^- = V_S \;\rightarrow\; V_o = -GA(s)V_S \;\rightarrow\; G_{loop} = -GA(s).$$

This means that we can write the closed-loop output impedance as:

$$Z = R(1 + GA(s))$$

and, at zero frequency, it will be equal to:

$$Z = 10(1 + 10^2 \cdot 10^6) = 1 \text{ G}\Omega.$$

Alternatively, we could have used the Blackman's formula, where the short-circuit loop gain is the one that we have just calculated:

$$G_{loop}|_{sc} = -GA(s)$$

while the open circuit loop gain can be compute leaving the pin at which we have R_L floating, obtaining:

$$G_{loop}|_{oc} = 0$$

since we cannot have any current flowing through R_S and thus the output of the gain stage is zero. This formula, therefore, would have lead to exactly the same result of the previous one.
We can now introduce an additional pole in the gain G, that will now have the following gain-bandwidth product:

$$GBWP' = 50 \text{ MHz}.$$

We want now to discuss the stability and, eventually, compensate this circuit. As in the previous case, the loop gain can be written as:

$$G_{loop}(s) = -A(s)G(s)\frac{R}{R + R_L}$$

but now the gain is itself a single-pole transfer function:

$$G(s) = \frac{100}{1 + s\tau_g}, \quad f_g = \frac{1}{2\pi\tau_g} = 500 \text{ kHz}.$$

We can immediately observe, from the position of this pole that it will be at a lower frequency with respect to the crossover frequency that we obtained in the previous, ideal gain case, thus affecting the stability of the circuit. Representing in a suitable Bode diagram for both the limiting case of R_L this transfer function, we can obtain that in the case in which $R_L = 1 \text{ k}\Omega = R_{L,max}$ the phase margin is decreased until $\phi_m \simeq 35°$, while in the case in which the load resistor is

identically equal to zero it is even worse than this. We need thus to find a way to compensate this circuit.
The simplest idea, in this case, could be to add a capacitor in parallel to one of the elements that are present, for example R or R_L. Putting it in parallel to R we will not compensate the circuit, since it will only add a pole, thus resulting not effective (and, on the other hand, worsening the phase margin), while putting it in parallel to R_L is also useless. Another idea, then, could be to put the capacitor in the feedback loop, thus somehow resulting in parallel to the gain stage. However, in this way we are not changing any voltage of the network in the steady-state case, therefore we have to add also a resistor R_1 that can be placed as in Figure 5.91. In this case, the output of the gain stage, whose voltage we can call $V_{out,g}$, can be calculated (in the computation of the loop gain, that is the quantity relevant for the stability) to be equal to:

$$V_{out,g} = V_S \frac{R}{R + R_L} G(s).$$

In this way, we have reduced to the equivalent circuit represented in Figure 5.92.

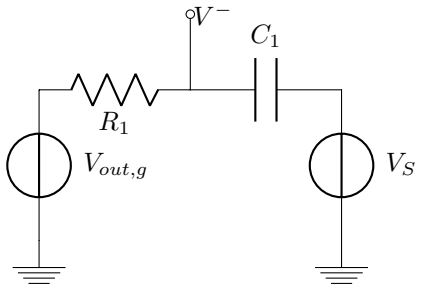


Figure 5.92: Equivalent circuit that we have to solve.

In this case, we can compute the voltage at the negative input pin of the operation amplifier as:

$$\begin{aligned} V^- &= V_S G(s) \frac{R}{R + R_L} \frac{1}{1 + sC_1R_1} + V_S \frac{sC_1R_1}{1 + sC_1R_1} = \\ &= \frac{V_S}{1 + sC_1R_1} \left(\frac{100}{1 + s\tau_g} \frac{R}{R + R_L} + sC_1R_1 \right) = \\ &= \frac{V_S}{1 + sC_1R_1} \cdot \frac{K + sC_1R_1 + s^2C_1R_1\tau_g}{1 + s\tau_g} \end{aligned}$$

where we have defined:

$$K = 100 \frac{R}{R + R_L}.$$

We can thus write the loop gain as:

$$G_{loop} = -\frac{A_0}{1 + s\tau} \cdot \frac{K + sC_1R_1 + s^2C_1R_1\tau_g}{(1 + s\tau_g)(1 + sC_1R_1)}$$

where we have that:

$$K \propto R_L, \quad K \in [1, 100].$$

We can immediately determine the position of the three poles:

$$f_{p0} = 1 \text{ Hz}, \quad f_{p1} = \frac{1}{2\pi C_1 R_1}, \quad f_{pg} = f_{p2} = 500 \text{ kHz}$$

and of the two zeros in an approximate way:

$$f_{z1} \simeq \frac{K}{2\pi C_1 R_1}, \quad f_{z2} \simeq \frac{C_1 R_1}{C_1 R_1 \tau_g 2\pi} \simeq \frac{1}{2\pi\tau_g} = f_{pg}$$

thus obtaining a cancellation. We are thus left only with one zero and two poles. When we have that:

$$K = 1 \ \rightarrow \ R_L = R_{L,max}$$

we have another pole-zero cancellation:

$$f_{p1} = f_{z1}$$

and thus the closed-loop compensated system that we have obtained is clearly asymptotically stable. On the other hand, when we have that:

$$K = 100 \ \rightarrow \ R_L = 0$$

we need to further study the behaviour of the loop gain as in Figure 5.93.

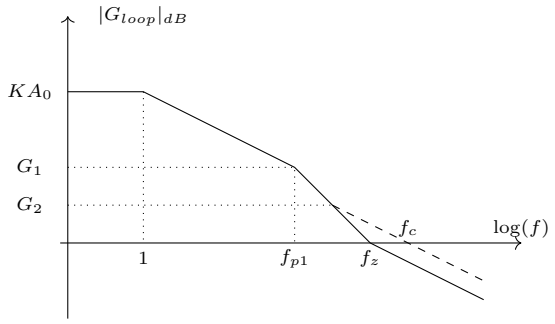


Figure 5.93: Bode diagram of the loop gain for $K = 100$.

We need now to have both f_{p1} and f_z to be before the frequency identified by the gain-bandwidth product. Since we have that:

$$KA_0 f_{p0} = G_1 f_{p1}, \quad G_1 f_{p1}^2 = G_2 f_{p2}^2$$

we can write this second gain as:

$$G_2 = G_1 f_{p1} \frac{f_{p1}}{f_{p2}^2} = KA_0 f_{p0} \frac{f_{p1}}{f_z^2} = KA_0 f_{p0} \frac{f_{p1}}{K^2 f_{p1}^2} = \frac{A_0 f_{p0}}{K f_{p1}}$$

and thus the crossover frequency will be:

$$f_c = G_2 f_z = K f_{p1} G_2 = A_0 f_{p0} = GBWP.$$

Alternatively, we could have taken the second pole and brought it to the right of the crossover frequency. From the Bode plot of the magnitude of the loop gain

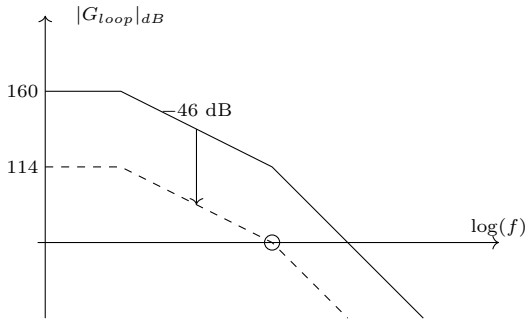


Figure 5.94: Bode diagram of the loop gain for different values of K.

that is represented in Figure 5.94, it means that we have to reduce the loop gain of a factor 200, that is equal to 46 dB.

This decrement can be easily obtained, considering the circuit that is represented in Figure 5.90, by adding a compensation resistor between the positive input pin of the gain stage and the output of the operation amplifier. In the computation of the loop gain, this will involve a partition for computing the voltage across the resistor R, therefore we have to set:

$$R_C = 200R.$$

This solution will always work, but it is only the last resource when we are not able to compensate the circuit in any other way, since it has two important drawbacks. First of all, a decrease in the loop gain corresponds to an increment of the error, that is in general bad (even though, in this case, we are not excessively decreasing it). Second, it will limit the dynamic of the circuit. In fact, if we assume to have a zero compensator resistor, the maximum current that can flow in the circuit will be:

$$R_C = 0 \;\rightarrow\; I_{0,max} = \frac{V_i}{R} = \frac{10\text{ V}}{10\ \Omega} = 1\text{ A}$$

while if we are using a compensation resistance:

$$R_C = 200R = 200\text{ k}\Omega \;\rightarrow\; I_{0,max} = \frac{V_i}{R_C} = 5\text{ mA}$$

therefore the maximum value of the current that can come from the same input is much lower. We can now compute the total output power spectral density for the noise as the following sum:

$$S_{io} = S_v \frac{1}{G^2R^2} + S_v \frac{1}{R^2} + 0 + S_i + \frac{4k_BT}{R} = 4 \cdot 10^{-18}\text{ A}^2/\text{Hz}$$

where in the order we have considered the noise equivalent voltage source for the operation amplifier, the noise equivalent voltage source for the gain stage, the noise equivalent current source for the operation amplifier, the noise equivalent current source for the gain stage and, last, the thermal noise in the resistor R.

5.7 Signal conditioning

5.7.1 Exercise 1

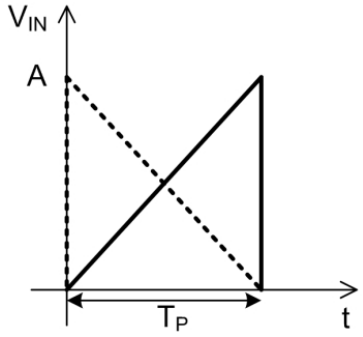


Figure 5.95: Signal considered.

This exercise comes from the exam of February 25th, 2011. Consider the signal represented with a solid line in Figure 5.95:

$$T_p = 100\ \mu\text{s}, \quad \sqrt{S_v} = 20\ \text{nV}/\sqrt{\text{Hz}}$$

and evaluate the output of a low-pass filter with the following characteristics:

$$R = 1\ \text{k}\Omega, \quad C = 100\ \text{nF}$$

and of a gated integrator with a suitable integration time T_g.
First of all, we have to remember that since we are dealing with a real signal then S_v is a unilateral power spectral density and thus we want to find the minimum detectable signal A that we can detect at the output of the low-pass filter. By definition, the minimum detectable signal is the one for which we can have the following signal-to-noise ratio:

$$\left(\frac{S}{N}\right)_{out} = 1.$$

From the circuit of the low-pass filter, we can calculate the time constant of this filter as:

$$RC = 10^{-4} = 100\ \mu\text{s} = T_p$$

and observe that it is equal to the duration of the pulse. Since the input is not constant, we have to explicitly compute the output signal as:

$$y = \int x(\tau)h(t-\tau)\ d\tau$$

where the signal can be written as:

$$x(\tau) = \begin{cases} \frac{A\tau}{T_p}, & 0 < \tau < T_p \\ 0, & \text{elsewhere} \end{cases}$$

while the response of the filter can be written as:

$$h(t-\tau) = \frac{1}{RC} e^{-\frac{t-\tau}{RC}}.$$

Computing this convolution:

$$\begin{aligned} y(t) &= \frac{A}{T_p}\int_0^t \frac{\tau}{RC} e^{-\frac{t-\tau}{RC}}\ d\tau = \frac{A}{T_p} e^{-\frac{t}{RC}} \int_0^t \frac{\tau}{RC} e^{\frac{\tau}{RC}}\ d\tau = \\ &= \frac{A}{T_p} e^{-\frac{t}{RC}} \left[\left(\tau e^{\frac{\tau}{RC}} \Big|_0^t \right) - \int_0^\tau e^{\frac{\tau}{RC}}\ d\tau \right] = \\ &= \frac{A}{T_p} e^{-\frac{t}{RC}} \left[te^{\frac{t}{RC}} - \left(RCe^{\frac{\tau}{RC}} \Big|_0^t \right) \right] = \\ &= \frac{A}{T_p} e^{-\frac{t}{RC}} \left[(t-T_p)e^{\frac{t}{RC}} + RC \right] = A\frac{t-T_p}{T_p} + Ae^{-\frac{t}{RC}} = \\ &= A\left(\frac{t}{T_p} - 1 + e^{-\frac{t}{RC}} \right). \end{aligned}$$

This is the standard way of computing the output signal. Alternatively, we could have considered the step response of this filter and, since the input signal, being a ramp, is the integral of the step, the response of the filter to the input is the integral of the step response of the filter:

$$\int_0^t \frac{A}{T_p} \left(1 - e^{-\frac{t'}{RC}} \right)\ dt'$$

after having suitably rescaled it. After the end of the signal, then, the capacitor will start to discharge, thus giving an exponentially decaying behaviour. The maximum amplitude is thus obtained at:

$$y(T_p) = y_M$$

thus giving the following signal-to-noise ratio:

$$\left(\frac{S}{N} \right)_{out} = \frac{y(T_p)}{\sqrt{\overline{n_y^2}}}.$$

From the expression of the output of the filter, we can write:

$$y(T_p) = Ae^{-1} = \frac{A}{e}, \quad T_p = RC$$

and we need thus to calculate the mean square value of the low-pass filter, that from the theory can be written as:

$$\overline{n_y^2} = \frac{\pi}{2} S_x f_p = \frac{\pi}{2} S_x \frac{1}{2\pi CR} = \frac{S_v}{4RC} = (1\ \mu\text{V})^2.$$

Imposing the requirement on the signal-to-noise ratio:

$$\left(\frac{S}{N} \right)_{out} = 1 \ \rightarrow\ \frac{A}{e} = \sqrt{\frac{S_v}{4RC}}$$

and we obtain that:

$$A = e\sqrt{\frac{S_v}{4RC}} = 2.7\ \mu\text{V}$$

is the minimum amplitude required for the signal.
It is important to notice that another possible source of noise is the resistance R of the low-pass filter, therefore we need to make sure that the input noise is the dominating source of noise in this device:

$$\sqrt{4k_BTR} = 4\ \text{nV}/\sqrt{\text{Hz}} \ll \sqrt{S_v}$$

and thus we can neglect it.
Considering now a gated integrator, we need to define the start and the stop of the window over which we are integrating. Obviously, there is not any sense in extending the window beyond the end of the signal. A possible choice is to take $t = 0$ as the start of the integrating window and $t = T_p$ as the end, but since we are asked to optimize the signal-to-noise ratio we need to further study it.
A first possible choice is to start from zero and end at:

$$T_g < T_p.$$

An even better choice, however, for the same size of the window, is to put it where the signal is stronger, thus starting at $T_p - T_g$ and ending at T_p. In fact, in both cases we are collecting the same amount of noise, therefore we want to maximize the signal collected.
The output signal, therefore, will be:

$$\begin{aligned} y(t) &= \int x(\tau)w(t,\tau)\ d\tau = G\int_{T_p-T_g}^{T_p} \frac{At}{T_p}\ dt = \\ &= \frac{GA}{T_p}\int_{T_p-T_g}^{T_p} t\ dt = \frac{GA}{2T_p}\left[T_p^2 - (T_p - T_g)^2\right] = \\ &= \frac{GA}{2T_p}T_g(T_p - T_g). \end{aligned}$$

From the theory on the gated integrator, we know that the mean square value of the output noise can be written as:

$$\overline{n_y^2} = \lambda G^2 T_g$$

where λ is the bilateral power spectral density. Since in this case it is given the unilateral power spectral density, the two can be related by writing:

$$\lambda = \frac{S_v}{2}$$

and therefore the mean square value of the output noise can be written as:

$$\overline{n_y^2} = \frac{S_v}{2}G^2T_g = \lambda\int |W(t,f)|^2\ df = \lambda\,(GT_g)\,\frac{1}{T_g}$$

where we have considered that:

$$|W(t,f)|^2 = (GT_g)^2\,\text{sinc}^2(\pi f T_g).$$

From this value, we can calculate the signal-to-noise ratio:

$$\left(\frac{S}{N}\right)_{out} = \frac{\frac{GA}{2T_g}(2T_pT_g - T_g^2)}{\sqrt{\frac{S_v}{2}T_gG^2}} = \frac{A}{2T_g}\sqrt{\frac{2}{S_v}}\frac{2T_pT_g - T_g^2}{\sqrt{T_g}}$$

where we can immediately see that, as expected, any constant gain is cancelled. Maximizing this quantity with respect to the integration time T_g, we can obtain:

$$\frac{\partial}{\partial T_g}\left(\frac{2T_pT_g - T_g^2}{\sqrt{T_g}}\right) = 0 \;\rightarrow\; \frac{d}{dT_g}\left(2T_p\sqrt{T_g} - (T_g)^{\frac{3}{2}}\right) = \frac{2T_p}{2\sqrt{T_g}} - \frac{3}{2}\sqrt{T_g} = 0$$

that gives:

$$T_g = \frac{2}{3}T_p.$$

This is a well-known result for the gated integrator. Replacing it in the signal-to-noise ratio, we can set it equal to one:

$$\left(\frac{S}{N}\right)_{out} = \frac{3A}{4T_p}\sqrt{\frac{2}{S_v}}\frac{\frac{4}{3}T_p^2 - \frac{4}{9}T_p^2}{\sqrt{\frac{2}{3}T_o}} = 1$$

we obtain the minimum amplitude for the signal to be measurable:

$$A \simeq 2.6\ \mu\ \text{V}.$$

Notice that this value is quite similar to the previous result that we have obtained for a low-pass filter.
At this point, we can consider the signal that is represented as a dashed line in Figure 5.95 and repeat the same calculation for the low-pass filter and for the gated integrator.
Starting from the case of the gated integrator, the signal-to-noise ratio will be exactly the same if we place the integration window from time $t = 0$ to:

$$T_g = \frac{2}{3}T_p.$$

In the case of the low-pass filter, on the other hand, we need to repeat the computation of the integral that leads to the solution. The output signal, in this case, will be:

$$\begin{aligned} y(t) &= \int x(\tau)h(t-\tau)\ d\tau = \int_0^t x(\tau)\frac{1}{RC}e^{-\frac{t-\tau}{RC}}\ d\tau = \\ &= \int_0^t A\left(1 - \frac{\tau}{T_p}\right)\frac{1}{RC}e^{-\frac{t-\tau}{RC}}\ d\tau = e^{-\frac{t}{RC}}\frac{A}{RC}\int_0^t\left(1 - \frac{\tau}{T_p}\right)e^{\frac{\tau}{RC}}\ d\tau = \\ &= Ae^{-\frac{t}{RC}}\left(\int_0^t \frac{1}{RC}e^{\frac{\tau}{RC}}\ d\tau - \int_0^t \frac{\tau}{T_p}\frac{1}{RC}e^{\frac{\tau}{RC}}\ d\tau\right) \end{aligned}$$

and considering that the second integral has already been calculated, we can solve only the first one, at the end obtaining:

$$y(t) = A\left(2 - 2e^{-\frac{t}{RC}} - \frac{t}{T_p}\right).$$

Since in the signal-to-noise ratio we want to find the maximum amplitude for this signal, we can set:

$$\frac{dy(t)}{dt} = 0 \;\rightarrow\; \frac{2}{RC}e^{-\frac{t}{RC}} - \frac{1}{T_p} = 0$$

obtaining the time instant at which it is maximum:

$$e^{-\frac{t}{RC}} = \frac{1}{2} \;\rightarrow\; t = RC\log(2).$$

This means that the maximum signal will be:

$$y(t) = A\left(2 - 2e^{-\log(2)} - \log(2)\right) = A\left(1 - \log(2)\right)$$

thus giving, since the mean square value of the output noise will be of the form of the previously computed one, the following signal-to-noise ratio:

$$\left(\frac{S}{N}\right) = \frac{A\left(1 - \log(2)\right)}{\sqrt{\frac{S_v}{4RC}}} = 1$$

that gives the following minimum amplitude:

$$A = 3.25\ \mu\text{V}.$$

This value is higher than the previous one, therefore this signal acquisition is worse than the previous case.

5.7.2 Exercise 2

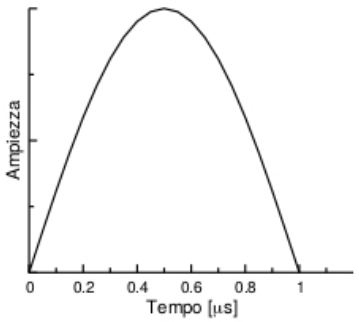


Figure 5.96: Signal considered.

Given the signal, defined as the portion of a sinusoid of amplitude A, represented in Figure 5.96, where[16]:

$$A = 10\ \text{mV}, \quad \lambda = 10^{-12}\ \text{V}^2/\text{Hz}$$

compute the signal-to-noise ratio at the output of a gated integrator keeping in mind that the following equation holds:

$$\tan(x) = 2x \;\rightarrow\; x \simeq 1.1655.$$

[16]Remember that λ is the bilateral power spectral density of the signal.

In this case, a possible choice is to set the integration interval symmetrically with respect to the peak of the signal considered. Changing the reference system in order to have the time instant $t = 0$ exactly in the position of the maximum of the signal, we can consider a gate that is extending from $-T_g$ to T_g, thus obtaining an output signal equal to:

$$y(t) = \int x(\tau)w(t,\tau)\ d\tau = G\int_{-T_g}^{T_g} A\cos\left(\frac{\pi t}{t_0}\right)\ dt =$$
$$= A\frac{Gt_0}{\pi}\sin\left(\frac{\pi t}{t_0}\right)\Bigg|_{-T_g}^{T_g} = \frac{2AGt_0}{\pi}\sin\left(\frac{\pi T_g}{t_0}\right)$$

where we have defined t_0 as the pulse duration in the original time reference system. From the theory of the gated integrator, the mean square value of the output noise can be written as:

$$\overline{n_y^2} = \lambda G^2(2T_g)$$

and therefore the signal-to-noise ratio at the output of this filter can be written as:

$$\frac{S}{N} = \frac{2A\cancel{G}t_0}{\pi}\sin\left(\frac{\pi T_g}{t_0}\right)\frac{1}{\cancel{G}\sqrt{2\lambda T_g}} = \frac{A\sqrt{2}t_0}{\pi\sqrt{\lambda}}\cdot\frac{\sin\left(\frac{\pi T_g}{t_0}\right)}{\sqrt{\frac{\pi T_g}{t_0}}}\cdot\sqrt{\frac{\pi}{t_0}} =$$
$$= A\sqrt{\frac{2t_0}{\pi\lambda}}\cdot\frac{\sin(x)}{\sqrt{x}}$$

where we have defined:

$$x = \frac{\pi T_g}{t_0}.$$

Optimizing the size of the gate in order to have the maximum possible signal-to-noise ratio:

$$\frac{\partial}{\partial T_g}\left(\frac{S}{N}\right) = 0\ \rightarrow\ \frac{\partial}{\partial x}\left(\frac{S}{N}\right) = 0$$

and we get:

$$\frac{\cos(x)\sqrt{x} - \sin(x)\frac{1}{2\sqrt{x}}}{x} = 0\ \rightarrow\ \tan(x) = 2x$$

that can be solved, from the hint, as:

$$x = 1.1655 = \frac{\pi T_g}{t_0}.$$

This gives a signal-to-noise ratio of:

$$\frac{S}{N} \simeq 6.8.$$

5.7.3 Exercise 3

Considering again the signal and the situation considered in the first exercise, assume now that the triangular waveform is part of a repetitive signal $x(t)$ with a repetition frequency of:

$$f_{rep} = 100 \text{ Hz}.$$

The amplitude A of this triangular signal can be considered as constant over a time interval of 5 s; find a filter to improve its signal-to-noise ratio can adjust its parameters in order to obtain the best performance.
Since we are dealing with a repetitive series of pulses, we can use as a filter a boxcar averager. In this case, the signal-to-noise ratio at the output of the boxcar can be related to the signal-to-noise ratio of the single pulse (denoted by the subscript *sp*) by the following formula:

$$\left(\frac{S}{N}\right)_{BA} = \left(\frac{S}{N}\right)_{sp} \cdot \sqrt{N_{eq}}$$

where we have defined the number of equivalent pulses as:

$$N_{eq} = \frac{2T_F}{T_C}.$$

In general, we can hopefully assume that:

$$T_F \gg T_C$$

and this hypothesis will allow us to neglect the exponential discharge of the capacitor, thus significantly simplifying the calculations. For each single pulse, therefore, the weighting function can be approximated with the weighting function of a gated integrator and, therefore, from the first exercise, we can assume again the duration of each gate to be equal to the size that we determined imposing the maximization of the single-pulse signal-to-noise ratio:

$$T_C = \frac{2}{3}T_p \simeq 67\ \mu\text{s}.$$

Notice now that the time T_F must be large but not too large, since we want to maintain the amplitude of the rectangular signal A to be constant; this means that we have to perform the whole measurement in a time that is smaller than 5 s. Using an equivalent time description of the boxcar averager, we can obtain a new weighting function that is equivalent, both from the viewpoint of the noise and of the signal, to the correct one. Assuming this exponential (that is the equivalent time weighting function) to be completely finished after five time constants, this means that the measurement will last for $5T_F$, that must be smaller than 5 s in the real time representation. In this time interval, we can study in the real time representation the following number of pulses:

$$5\ \text{s} \cdot 100\ \text{Hz} = 500\ \text{pulses}$$

and therefore in the equivalent time representation we will have 500 piecewise exponential integration intervals with a duration equal to T_C:

$$500T_C = 5T_F \ \rightarrow\ T_F = 100T_C = 6.7\ \mu\text{s}.$$

This is therefore the time constant that is needed for satisfying this requirement. The number of equivalent pulses therefore will be:

$$N_{eq} = \frac{2T_F}{T_C} \simeq 200$$

and thus we will improve the previously calculated single-pulse signal-to-noise ratio (that will come from the calculation for the gated integrator) of a factor $\sqrt{N_{eq}}$.

5.7.4 Exercise 4

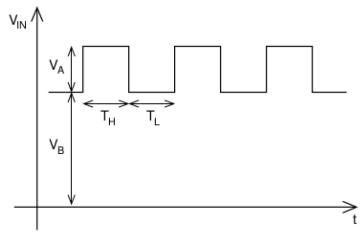


Figure 5.97: Signal considered.

It is given the square wave repetitive signal represented in Figure 5.97, where:

$$T_H = T_L = 1\ \mu\text{s},\ V_A \simeq 10\ \mu\text{V},\ \sqrt{S_v} = 50\ \text{nV}/\sqrt{\text{Hz}},\ S/N = 10.$$

Propose a way of measuring the square wave signal V_A superimposed over an unknown offset signal V_B using only boxcar averagers. Find then all the relevant parameters for this experiment and calculate the equivalent number of samples N_{eq}.
Quite intuitively, it is not possible to use only one boxcar averager to measure the signal V_A. We need thus to use two differently synchronized boxcar averagers, one measuring the periods of length T_H in which the signal is at the high level and the other measuring the signal during the periods of length T_L in which it is at the low level. Notice that, in this way, one boxcar averager is measuring a signal proportional to V_A+V_B, while the other is measuring a signal proportional to V_B. Calculating, using a suitable electronic circuit, the difference between these two signals, we can obtain the desired signal V_A.

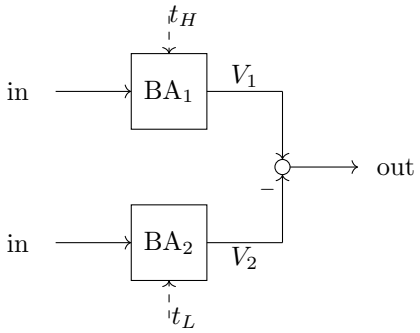


Figure 5.98: The proposed scheme for this measurement.

In this case, the weighting function for these two boxcar averagers will be made of windows each one with a width of T_L or T_H depending on the boxcar averager considered[17], in order to obtain the maximum signal. In the equivalent time representation for the first boxcar averager, we obtain a signal that is proportional to V_A+V_B and the weighting function can be considered equivalent

[17]From a practical point of view, however, they are equal.

to the one of a low-pass filter, allowing us to calculate the output signal and the mean square value of the output noise:

$$V_1 = V_A + V_B, \quad \overline{n_1^2} = \frac{S_v}{4T_F}$$

where we have indicated with T_F the equivalent time constant of the filter. Analogously, at the output of the second boxcar averager we have that:

$$V_2 = V_B, \quad \overline{n_2^2} = \frac{S_v}{4T_F}$$

since the noise will be the same at the two different inputs and where we have assumed both the filters to have the same equivalent time constant. At the output of the whole device, then, we will have:

$$V_o = V_1 - V_2 = V_A + V_B - V_B = V_A$$

from the viewpoint of the signal, while for the noise:

$$\overline{n_o^2} = \overline{n_1^2} + \overline{n_2^2} = \frac{S_v}{4T_F}$$

where we have considered that the variances of uncorrelated[18] stochastic processes adds up. Imposing the required value for the signal-to-noise ratio:

$$\frac{S}{N} = \frac{V_A}{\sqrt{\frac{S_v}{4T_F}}} = 10 \ \rightarrow \ T_F = 1.25 \text{ ms}.$$

At this point, we can consider V_B to be a sinusoidal interference at frequency $3f_0$, where f_0 is the frequency of the square wave signal:

$$3f_0 = 1.5 \text{ MHz}.$$

How is it possible to improve this acquisition?
In this case, the signal V_B can be written as:

$$V_B = V_B 0 \sin(3\omega_0 t + \phi)$$

and it will be integrated over a time window T_C. In this case, the output signal can be written as:

$$V_o = \int_0^{T_H} \frac{1}{T_F} V_{B0} \sin(3\omega_0 t + \phi) \ dt = \frac{V_{B0}}{T_F 3\omega_0} \cos(3\omega_0 t + \phi)|_{T_H}^0 =$$
$$= \frac{V_{B0}}{T_F 3\omega_0} [\cos(\phi) - \cos(3\omega_0 T_H + \phi)]$$

but since we have that:

$$\omega_0 T_H = \pi$$

we obtain the following output signal:

$$V_o = \frac{V_{B0}}{3\omega_0 T_F} [\cos(\phi) - \cos(3\pi + \phi)] = \frac{V_{B0}}{3\omega_0 T_F} [\cos(\phi) - \cos(\pi + \phi)] =$$
$$= \frac{V_{B0}}{3\omega_0 T_F} 2\cos(\phi) = \frac{2V_{B0}}{3\omega_0 T_F} \cos(\phi).$$

[18] Since we have assumed the input noise to be white.

The output signal is thus proportional to a cosine function that depends on the initial phase ϕ of the disturbance. In the worst case:

$$\phi = 0 \ \rightarrow \ V_o = \frac{2V_{B0}}{3\omega_0 T_F}.$$

Notice that we cannot always cancel out this noise in two consecutive measurements: on the first window corresponding to T_H we will have a period and a half of the disturbance obtaining a positive contribution, while in the other window corresponding to T_L it will have a negative contribution. Taking the difference of these two signals, at the end, we are summing up the different sinusoidal contributions, thus increasing the noise at the end of the two boxcar averagers. At the end, this results in a noise equal to:

$$V_o = \frac{4V_{B0}}{3\omega_0 T_F}$$

that is twice the value of the previous result. Studying the weighting function of the boxcar averager, we can see that the contributions corresponding to each different pulse will give the same amount of noise with the same phase to the total noise due to the fact that the frequency of the interference is an integer multiple of the f_0. However, these contributions will have a decaying amplitude, therefore we can write them as a series:

$$\frac{4V_{B0}}{3\omega_0 T_F}\left(1 + e^{-\frac{T_H}{T_F}} + e^{-2\frac{T_H}{T_F}} + \dots\right) = \frac{4V_{B0}}{3\omega_0 T_F}\sum_{n=0}^{\infty}\left(e^{-\frac{T_H}{T_F}}\right)^n =$$

$$= \frac{4V_{B0}}{3T_F\omega_0}\frac{1}{1 - e^{-\frac{T_H}{T_F}}} \simeq \frac{4V_{B0}}{3T_F\omega_0}\frac{T_F}{T_H} = \frac{4V_{B0}}{3T_H\omega_0} = \frac{4V_{B0}}{3\pi}.$$

To get rid of this disturbance, we can set the integration time for a single pulse (that was previously equal to $T_H = T_L$) to an integer multiple of the period of the sinusoidal disturbance. In frequencies, this means to set to zero the sinc^2 function that is within the square modulus of the Fourier transform of the weighting function $|W(t, f)|^2$ in the position corresponding to the frequency of the disturbance. This will slightly reduce the signal but for sure it will completely cancel out the disturbance.

5.7.5 Exercise 5

This exercise comes from the exam of September 5th, 2011. Consider a rectangular signal with amplitude A and duration T to which it is superimposed a white noise with given power spectral density:

$$A \simeq 5\ \mu\text{V}, \quad T = 100\ \mu\text{s}, \quad \sqrt{S_v} = 40\ \text{nV}/\sqrt{\text{Hz}}$$

that is at the input of a buffer stage with unitary gain presenting a pole at the following frequency:

$$f_p = 100\ \text{KHz}.$$

Compute the signal-to-noise ratio at the output of the buffer stage and the improvement deriving from the use of a gated integrator.
The bandwidth of this rectangular signal can be evaluated as:

$$BW \simeq \frac{1}{T} \simeq 10\ \text{KHz}$$

and therefore it will not be affected by the fact that the bandwidth of the amplifier is limited. This means that, at the output of the buffer stage, the signal-to-noise ratio can be written as:

$$\frac{S}{N} = \frac{A}{\sqrt{S_v \frac{\pi}{2} f_p}} \simeq 0.32.$$

Adding a gated integrator at the output of the buffer stage, we are actually integrating over the whole rectangular signal, thus having a size of the gate equal to:

$$T_g = T = 100 \ \mu\text{s}.$$

In this case, from our previous theoretical description of the gated integrator, we can write the signal-to-noise ratio at the output of this filter as:

$$\frac{S}{N} = \frac{GAT_g}{\sqrt{\frac{S_v}{2} G^2 T_g}} = \frac{AT_g}{\sqrt{S_v T_g 2}} = A\sqrt{\frac{2T_g}{S_v}} = 1.77.$$

We have thus improved the signal-to-noise ratio of a factor that is proportional to the ratio between the bandwidth of the noise at the input and the noise bandwidth of the gated integrator. At the input, the noise bandwidth can be evaluated by writing:

$$f_n = \frac{\pi}{2} f_p = 157 \text{ kHz}$$

while the noise bandwidth of the gated integrator is equal to:

$$BW_n = \frac{1}{2T_g} = 5 \text{ kHz}$$

and thus the signal-to-noise ratio can be written as:

$$\left(\frac{S}{N}\right)_{out} = \left(\frac{S}{N}\right)_{in} \sqrt{\frac{f_n}{BW_n}}.$$

This obviously holds when the noise can be assumed to be white, as in this case, where the bandwidth of the noise is much larger than the bandwidth of the filter; if this were not the case, we would need to take into account also the correlation function.

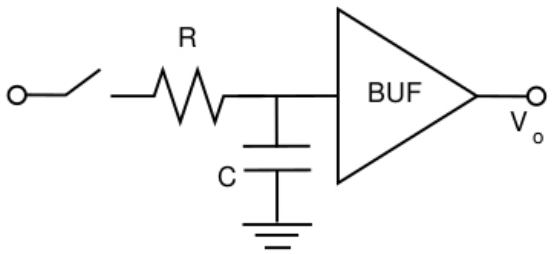


Figure 5.99: Filter considered.

Assume now to have a repetitive signal that is given by the periodic repetition of the previous rectangular signal with a duration of T_C and at a distance, in time, equal to T_A:

$$T_A = 5 \text{ ms}, \quad T_C = 0.1 \text{ ms}.$$

This signal is assumed to be at the input of the filter represented in Figure 5.99, that is clearly a boxcar averager. We are asked to find the value of the resistor R and of the capacitor C such that the signal-to-noise ratio is equal to:

$$S/N = 10.$$

Remembering the value of the signal-to-noise ratio that we have obtained at the input of the device and in the gated integrator:

$$\left(\frac{S}{N}\right)_{in} = 0.32, \quad \left(\frac{S}{N}\right)_{GI} = 1.77$$

since the behaviour of the boxcar averager for a single pulse is exactly identical to the one of a gated integrator, we can write:

$$\left(\frac{S}{N}\right)_{BA} = \left(\frac{S}{N}\right)_{GI} \sqrt{N_{eq}}$$

and from the requirement on the signal-to-noise ratio we can calculate the equivalent number of samples required:

$$N_{eq} \simeq 32.$$

However, from the definition of this parameter:

$$N_{eq} = \frac{2T_F}{T_C} \;\rightarrow\; T_F = \frac{T_C N_{eq}}{2} = 1.6 \text{ ms}$$

and thus we obtain:

$$RC = 1.6 \text{ ms}.$$

At this point, we can choose among various possible reasonable values of R and C such that the thermal noise on the resistor R is negligible with respect to the given noise source:

$$4k_B T R \ll S_v \;\Rightarrow\; R \ll 96 \text{ k}\Omega$$

and we can for example choose:

$$R = 10 \text{ k}\Omega.$$

We can now assume this second buffer to have a finite input impedance equal to:

$$R_{in} = 1 \text{ M}\Omega$$

and we want to calculate what is changing in the behaviour of the circuit with respect to the noise in this situation. When the switch is closed, the filter is completely identical to a low-pass filter with the following time constant:

$$T_F = C \cdot (R \| R_{in})$$

and we can thus write the transfer function of this filter as:

$$\frac{R_{in}}{R+R_{in}} \cdot \frac{1}{1+sC\left(R \| R_{in}\right)} \simeq \frac{1}{1+sCR}$$

thus obtaining that almost nothing changes. When the switch is open, on the other hand, we can immediately observe that the capacitor is discharging over this finite input impedance R_{in}. This means that now the weighting function can be described as the product between a series of rectangles (that are defining the integration windows) and a truly exponentially decaying behaviour. Notice that, when the switch is closed, this exponentially decaying behaviour will have the following time constant:

$$T_F = RC$$

while when the switch is open (thus, outside the rectangles) the time constant will be different:

$$T'_F = R_{in}C.$$

This means that this circuit is not a ratemeter, due to the fact that the time constant of the open-switch sections is much larger than the closed-switch ones. Again, neglecting the discharge over one single rectangle:

$$T_C \ll T_F$$

we can obtain that the amplitude of the various piecewise constant parts of this exponentially decaying weighting function. At the end, therefore, the output signal will be proportional to the following sum:

$$\begin{aligned}
&\sum_{n=0}^{\infty}\left[e^{-\left(\frac{T_C}{T_F}+\frac{T_O}{T'_F}\right)}\right]^n = \sum_{n=0}^{\infty}\left[e^{-\left(\frac{T_C}{RC}+\frac{T_O}{R_{in}C}\right)}\right]^n = \\
&= \frac{1}{1-e^{-\left(\frac{T_C}{RC}+\frac{T_O}{R_{in}C}\right)}} \simeq \frac{1}{1-1+\frac{T_C}{RC}+\frac{T_O}{R_{in}C}} = \\
&= \frac{1}{\frac{T_C}{RC}+\frac{T_O}{R_{in}C}} = \frac{RC}{T_C+T_O\frac{R}{R_{in}}} = \frac{T_F}{T_C+T_O\frac{R}{R_{in}}}.
\end{aligned}$$

If the input impedance, as in the ideal case, tends to infinity:

$$R_{in} \to \infty: \quad \sum_{n=0}^{\infty}\left[e^{-\left(\frac{T_C}{T_F}+\frac{T_O}{T'_F}\right)}\right]^n \to \frac{T_F}{T_C}$$

as in the previous case, while if:

$$R_{in} = R: \quad \sum_{n=0}^{\infty}\left[e^{-\left(\frac{T_C}{T_F}+\frac{T_O}{T'_F}\right)}\right]^n = \frac{T_F}{T_C+T_O}$$

as in the ratemeter. In our case:

$$\frac{R}{R_{in}} = \frac{1}{100}$$

and therefore the equivalent number of samples N_{eq} is smaller, but not excessively small as in the case of the ratemeter, thus decreasing the signal-to-noise

ratio. In order to obtain again the same signal-to-noise ratio as before, we need to impose the following time constant for the filter:

$$T_F = 2.4 \text{ ms}.$$

At this point, we can consider also the presence of the bias currents of the buffer:

$$I_B = 2.5 \text{ pA}$$

and evaluate its impact, finding the value of the resistor R that is needed to limit the error to the 10% of the correct value.
When the switch is closed, the signal is constant and thus we are integrating it, charging the capacitor. On the other hand, when the switch is open, the current will tend to discharge this capacitor. In steady-state conditions, the charging due to the signal when the switch is closed is equal to the discharging due to the bias current when the switch is open. The average value of this periodic behaviour, therefore, must be 10% smaller than the desired steady-state value. When the switch is closed, thus during T_C, the output voltage is exponentially increasing:

$$V_o = A\left(1 - e^{-\frac{t}{RC}}\right)$$

and in order to obtain a linear approximation of this curve we can calculate the first derivative:

$$\frac{dV_o}{dt} = \frac{A}{RC}e^{-\frac{t}{RC}}.$$

Since this approximation must hold near to the steady-state value of the output (that was A when we neglected the presence of the bias current), we can write an approximation in steady-state condition of the behaviour of the output voltage as:

$$V_o = A - Ae^{-\frac{t}{RC}}.$$

We can thus write the derivative as:

$$\frac{dV_o}{dt} = \frac{A - V_o}{RC}$$

and therefore the variation of the output voltage over a time T_C must be equal to:

$$\Delta V_o = \frac{dV_o}{dt} \cdot T_C = \frac{A - V_o}{RC}T_C.$$

Now, since we have a bias current I_B flowing through the capacitor, the variation of the voltage of the capacitor during the time interval in which the switch is open will be:

$$\Delta V_o = \frac{I_B}{C}T_O$$

and at the equilibrium the two variations will be equal one to the other:

$$\frac{A - V_o}{RC}T_C = \frac{I_B}{C}T_O.$$

Notice that in this relation we can simplify the capacitance C, since this parameter will affect both the charge and the discharge of the device. We have thus obtained the following condition

$$R = \frac{A - V_o}{I_B}\frac{T_C}{T_O}$$

but since we want an error lower than or equal to the 10% of the correct value:

$$A - V_o \leq 10\% \cdot A$$

we can write this requirement as:

$$R \leq 0.1 \frac{T_C}{T_O} \frac{A}{I_B} = 4 \text{ k}\Omega.$$

This means that we must decrease the value of R to increase the charging current and thus to decrease the time in which the capacitor is charged.
It is important to notice that in this theory we have completely neglected the effect of the bias current I_B during the charge of the capacitor. What is the error that we are committing with this assumption? The willing student can try to calculate it.

5.8 Optimum filtering

5.8.1 Exercise 1

This exercise comes from the exam of June 29th, 2009. Considering the signal represented in Figure 5.96 at page 372, where:

$$A \simeq 10 \text{ mV}, \quad \lambda = 10^{-12} \text{ V}^2/\text{Hz}$$

we have obtained, from the previous points:

$$\left(\frac{S}{N}\right)_{GI} = 6.8.$$

Assuming that this signal can be represented as:

$$x(t) = A \sin(\omega t)$$

for the case of white noise, the associated optimum filter can be written as:

$$w(t, \tau) \propto x(\tau) = G \sin(\omega \tau).$$

The output signal, therefore, will be:

$$y(t) = \int x(\tau) w(t, \tau) \, d\tau = AG \int_0^T \sin^2(\omega \tau) \, d\tau =$$
$$= AG \int_0^T \frac{1 - \cos(2\omega \tau)}{2} \, d\tau = \frac{AG}{2} T.$$

The mean square value of the output noise can then be written as:

$$\overline{n_y^2} = \lambda \int w^2(t, \tau) \, d\tau = \frac{\lambda T G^2}{2}.$$

This means that the optimum signal-to-noise ratio for the case of white noise is:

$$\left(\frac{S}{N}\right)_{opt} = \frac{A}{\sqrt{\lambda}} \cdot \sqrt{\frac{T}{2}} = \left(\frac{S}{N}\right)_{in} \sqrt{\frac{T}{2}} \simeq 7.1.$$

This is only a small improvement from the case of the gated integrator, therefore it is probably not worth to increase the complexity using an optimum filter instead of the simpler gated integrator.

5.8.2 Exercise 2

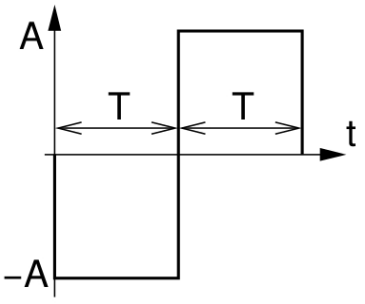


Figure 5.100: Signal considered.

This exercise comes from the exam of July 8th, 2014. Given the current signal represented in Figure 5.100, find the signal-to-noise ratio at the output of a simple RC filter and choose the time (T or $2T$) that is more suitable for the acquisition of this signal. The noise at the input is white with a bilateral power spectral density equal to S_I.

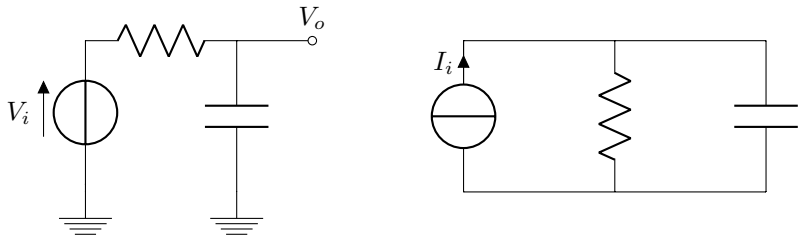


Figure 5.101: A low-pass filter for a voltage signal (on the left) and for a current signal (on the right).

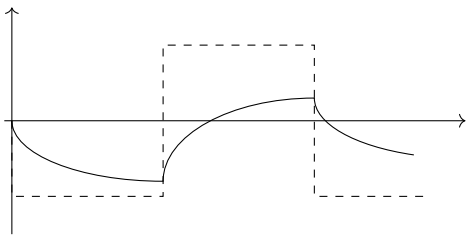


Figure 5.102: Behaviour of the output of the device (solid line) with an input square wave (dashed line).

Before starting, the first thing that we have to consider is that we are dealing with a current signal, therefore the low-pass filter will make a current partition between the resistor R and the capacitor C, giving the circuit that is represented in Figure 5.101. The output of this square wave input current can therefore be represented as in Figure 5.102. Since, to calculate the signal-to-noise ratio, we

want to choose the highest signal, the best situation for measuring it is to sample the output at time T; every other replica of the output exponential waveform will start from a positive or negative voltage, thus giving a reduced maximum value. The signal-to-noise ratio can then be written as:

$$\frac{S}{N} = \frac{A\left(1 - e^{-\frac{T}{RC}}\right)}{\sqrt{\frac{S_V}{2RC}}}$$

where S_V is the unilateral input power spectral density. It is important to note that, in this case, we are not asked the optimum value of the signal-to-noise ratio, therefore its optimization is not needed. A reasonable value of the time constant RC of the filter can be observed to be different from the limit $RC \ll T$ and also from the limit $RC \gg T$, since in both cases the signal-to-noise ratio is degraded. A reasonable choice (even though clearly not optimal) therefore will be:

$$RC \simeq T$$

that will give:

$$\frac{S}{N} = \frac{A\left(1 - e^{-1}\right)}{\sqrt{\frac{S_v}{2T}}}.$$

We are then asked to consider the optimum filter for this input white noise, sketch the behaviour of its output and find the associated signal-to-noise ratio. Since the noise is white, the weighting function of the optimum filter will be:

$$w(t, \tau) \propto x(\tau)$$

thus being a square wave, and the output of the device can be written as:

$$y(t) = \int x(\tau) w(t, \tau)\ d\tau.$$

Evaluating graphically this integral (that will be the integral of the superposition of rectangles) we obtain the behaviour that is represented in Figure 5.103.

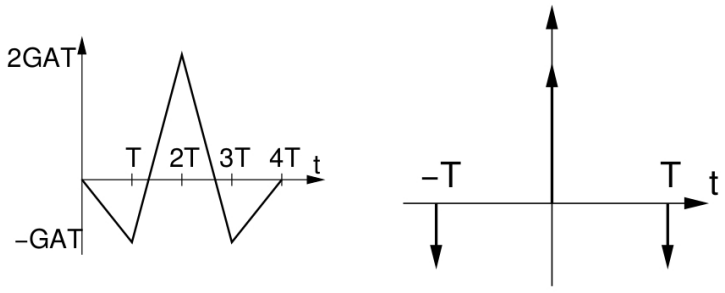


Figure 5.103: On the left, output signal of the optimum filter; on the right, weighting function of the optimum filter in the case of Flicker noise.

We can clearly see that the output signal of this optimum filter will be proportional to the autocorrelation function of the signal $x(t)$. The maximum

of this signal will therefore be in the time instant in which the weighting function is perfectly overlapped to the signal and this happens for a time instant that is equal to the pulse duration $2T$. The signal-to-noise ratio at the output of this filter, from the theory, can then be written as:

$$\frac{S}{N} = \frac{A}{\sqrt{S_I}}\sqrt{\int x^2(\tau)\ d\tau} = A\sqrt{\frac{2T}{S_I}}$$

where $x(t)$ is the unitary amplitude version of the signal. Assuming now to have also a shot noise at the input of the device, we can again discuss the properties of an optimum filter. In this case, we know that the weighting function of the optimum filter can be written as:

$$w(t,\tau) \propto \frac{x(t)}{\lambda(t)}$$

but in the case of shot noise:

$$\lambda(t) \propto qI(t) \propto x(t)$$

therefore the ratio in the weighting function will be constant and the optimum filter will be equal to the gated integrator. Its weighting function, therefore, will be constant over the whole gate. If we assume this gate to be extended from time equal to zero to a time equal to $2T$, the output signal-to-noise ratio will be equal to zero, thus not giving a suitable solution. There must be something wrong in our way of reasoning.
The problem is that we are dealing with a current signal and currents can also be negative, while the power spectral density of the shot noise is always a positive quantity. Therefore, a better expression for the power spectral density of the noise will be:

$$\lambda(t) \propto q|I(t)| \propto |x(t)|$$

that is a constant power spectral density, as in the previous case. This means that the correct answer is, again, the one that we have given in the previous case.
Last, we can find the optimum filter for dealing with Flicker noise. Its power spectral density can be written as:

$$S_f = \frac{k}{f}$$

and thus the optimum weighting function will not be proportional to the signal. In this case, as we have seen in the theoretical part, we can add a whitening filter H_w (where we can find at the output a constant power spectral density λ and a signal $I_w(t)$) and, then, an optimum filter for the white noise, whose weighting function will be proportional to the signal at the output of the whitening filter:

$$I(f)H_w(f).$$

The whitening filter can be written as:

$$\frac{k}{f}|H_w(f)|^2 = \lambda = \text{const} \ \Rightarrow\ |H_w(f)|^2 = \frac{\lambda f}{k} \ \Rightarrow\ |H_w(f)| \propto \sqrt{f}$$

and for the signal:

$$I_w(f) = I(f) \cdot H_w(f).$$

At the output of the whitening filter we will have an optimum filter for the white noise, that will thus be proportional to the output signal of the whitening filter. The overall output filter will thus be equal to the cascade of the whitening filter and the optimum filter for the white noise:

$$W_{opt}(t,f) = I(f) \cdot |H_w(f)|^2 \propto I(f) \cdot f.$$

Since the only quantity we are interested in is the magnitude of the output signal, we can shift it in order to be time-symmetric with respect to the time instant $t = 0$, obtaining the following output signal:

$$I(f) \propto \text{sinc}(\pi f T) \cdot \left[e^{-j\pi fT} - e^{j\pi fT}\right] \propto \text{sinc}(\pi f T)\sin(\pi f T) = \frac{\sin^2(\pi f T)}{\pi f T}.$$

Therefore, the Fourier transform of the weighting function of the optimum filter will be:

$$|W_{opt}(t,f)| \propto \sin^2(\pi f T).$$

Alternatively, we could have used the fact that:

$$|W_{opt}(t,f)| \propto \frac{I(f)}{S_n(f)} = \frac{I(f)}{\frac{k}{f}}$$

directly obtaining the previous result. In this case, the Fourier transform of the weighting function can then be written as:

$$W(t,f) = k\sin^2(\pi f T) = k \cdot \frac{1-\cos(2\pi f T)}{2} \propto 1 - \cos(2\pi f T) = \\ = 1 - \frac{e^{j2\pi fT} - e^{-j2\pi fT}}{2}$$

and therefore in the time domain:

$$w(t,\tau) = \delta(\tau) - \frac{1}{2}\left[\delta(\tau - T) + \delta(\tau + T)\right].$$

5.8.3 Exercise 3

This exercise comes from the exam of September 10th, 2010. It is given a rectangular signal with pulse duration T_S and amplitude A whose leading front arrives at $t = 0$. Consider the two following cases:

- high-frequency noise with delta-like noise autocorrelation;
- low-frequency noise with exponential autocorrelation;

and suppose that, to measure the signal, we want to take two different samples, one before the arrival of the rectangular signal, thus at time $t = 0^-$, to sample the noise, and the other at time T while we have the rectangle, therefore for $0 < T < T_S$, and calculate the difference between them:

$$x(T) - x(0^-).$$

Find the weighting function and the Fourier transform of this kind of filter and calculate the mean square value of the output noise.
Since we know that every sampling operation will give a delta-function in the weighting function, we can write it as:

$$w(t,\tau) = \delta(\tau - T) - \delta(\tau)$$

and in the frequency domain it will be:

$$W(t,f) = -1 + e^{-j2\pi fT}.$$

The mean square value of the output noise can then be written, from a time-domain perspective, as:

$$\overline{n_y^2} = \int R_{xx}(\tau) k_{ww}(\tau)\ d\tau$$

therefore we first have to calculate the autocorrelation of the weighting function of this filter. To do this, we can consider that at time $\tau = 0$ we have four delta functions (two positive and two negatives) that are sampling each other, thus giving a positive delta-function with area equal to two. For positive times $\tau > 0$, this correlation is again equal to zero until we have performed a shift equal to $\tau = T$, where one of the positive delta-functions is overlapped to one of the negative ones. The autocorrelation of the weighting function that we have obtained can therefore be represented as in Figure 5.104.

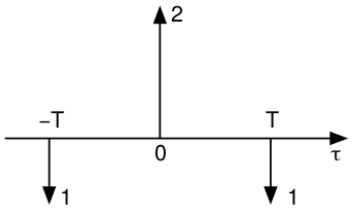


Figure 5.104: Autocorrelation of the weighting function of the filter.

In the case of white noise, then, since we it has a short correlation time with respect to the duration of the pulse:

$$T_n \ll T \ \Rightarrow\ \overline{n_y^2} = \int R_{xx}(\tau) 2\delta(\tau)\ d\tau = 2R_{xx}(0) = 2\overline{n_{in}^2}$$

since we have considered to have taken two uncorrelated samples, thus summing their variances.
In the case of the low-frequency noise, the three delta-functions that are present in the autocorrelation of the weighting function are individually sampling the autocorrelation of the input noise, thus giving:

$$\overline{n_y^2} = 2R_{xx}(0) - R_{xx}(T) - R_{xx}(T)$$

and since the autocorrelation of the noise is an even function:

$$\overline{n_y^2} = 2\left[R_{xx}(0) - R_{xx}(T)\right]$$

and considering that this autocorrelation is an exponentially decaying function whose amplitude is $\overline{n_{LF}^2}$, we can write:

$$\overline{n_y^2} = 2\left[1 - e^{-\frac{T}{T_n}}\right]\overline{n_{LF}^2}.$$

Obviously, this kind of filtering is not working in the white noise case, since the samples are totally uncorrelated. For the low-frequency noise, assumed that we are in the following limit:

$$T \ll T_n$$

this is working well, because the two noise samples that we are subtracting are strongly correlated.
The question now is: how is it possible to modify this filter to improve this performance? Assuming that we want to maintain a discrete-time filtering system, we can try to average these samples: as a consequence of the central limit theorem, the noise contributions will tend to vanish. It is important to notice that this procedure has to be performed for both the sampling procedures, thus obtaining a weighting function that is the sum of two Dirac combs, one with negative amplitude that is sampling only the noise and the other with positive amplitude that is sampling the noise and the signal. Averaging over each single Dirac comb, we are getting rid of the high-frequency noise contributions, since they will uncorrelated between one sample and the other, thus vanishing. Subtracting then the average of the negative Dirac comb to the average of the positive Dirac comb, we are thus obtaining an output signal in which the noise contribution is reduced in both the spectral regions. Assuming each Dirac comb to be composed by N elements, it is important to take care that the whole measurement time T_m is (significantly) smaller than the correlation time of the low-frequency noise, in order to be able to reduce it.

5.8.4 Exercise 4

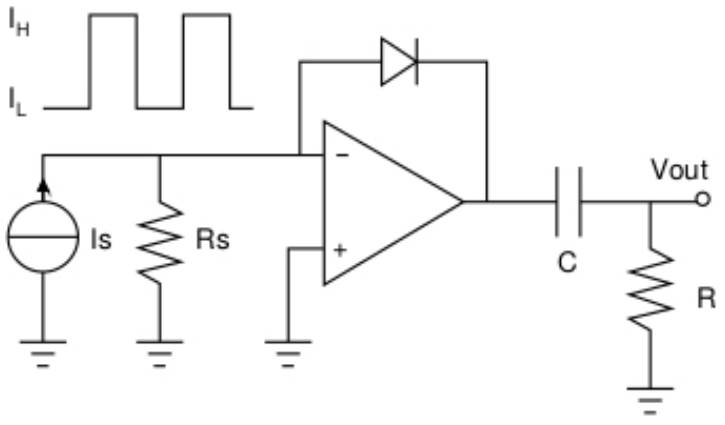


Figure 5.105: Circuit considered.

This exercise comes from the exam of July 20th, 2009. Given the circuit

represented in Figure 5.104 where at the input we have a square wave signal:

$$R = R_S = 1\ \text{M}\Omega,\ I_0 = 10\ \text{nA},\ C = 0.1\ \mu\text{F}$$

$$I_L = 10\ \text{mA},\ I_H = 100\ \text{mA},\ f = 500\ \text{Hz}$$

where we know that for the diode we can use the following relationship:

$$I = I_0 e^{\frac{V_D}{V_t}}, \quad V_t = \frac{k_B T}{q},\ \frac{k_B}{q} = 8.6 \cdot 10^{-5}\ \text{V/K}$$

consider that the setup is generally used to measure the temperature and compute the output signal before and after the CR filter that is connected at the output of the operation amplifier; then find the associated sensitivity as a function of the temperature T.
Studying the network, we can see that at the voltage at the output of the operation amplifier is:

$$V_1 = -V_D = -V_t \ln\left(\frac{I}{I_0}\right)$$

thus being a square wave of amplitude A ranging between two negative voltages V_H and V_L. Considering for example a room temperature condition:

$$T = 300\ \text{K}: \quad V_H = -356.4\ \text{mV},\ V_L = -415.8\ \text{mV}.$$

At the output of the CR filter, since it introduces a pole at:

$$f_p = \frac{1}{2\pi CR} \simeq 1.6\ \text{Hz}$$

we can consider valid the following approximation:

$$f \gg f_p$$

and since we are dealing with an high-pass filter we will obtain at the output a square wave without any DC component. The amplitude A of this square wave output signal can then be written as:

$$A = V_t \ln\left(\frac{I_H}{I_0}\right) - V_t \ln\left(\frac{I_L}{I_0}\right) = V_t \ln\left(\frac{I_H}{I_L}\right) = \frac{k_B T}{q} \ln(10).$$

The sensitivity of this device is, therefore:

$$\frac{k_B}{q} \ln(10) \simeq 200\ \mu\text{V/K}.$$

Assuming now that the output of the high-pass filter is connected to an amplifier with the following characteristics:

$$G = 30, \quad GBWP = 1\ \text{MHz}$$

we can try to calculate the mean square value of the noise at the output. In this case, the noise introduced by the operation amplifier is:

$$\sqrt{S_V} = 20\ \text{nV}/\sqrt{\text{Hz}}, \quad \sqrt{S_I} = 1\ \text{pA}/\sqrt{\text{Hz}}$$

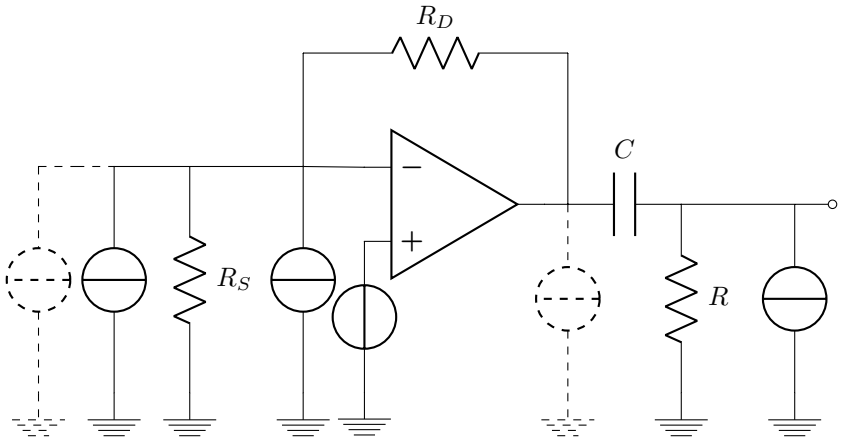


Figure 5.106: Noise equivalent sources for the circuit considered; the dashed generators are equivalent to the noise equivalent current source for the thermal noise in the resistor R_D.

and approximating the diode with an equivalent resistor:

$$R_D = 1\ \Omega$$

we can have the following noise sources represented in Figure 5.105, where the noise equivalent current source for the thermal noise in the resistor R_D has been split in two different equivalent sources.

In this case, the power spectral density at the output of the operation amplifier can be written as:

$$S_{V_1} = \left(\frac{4k_BT}{R_D} + \frac{4k_BT}{R_S} + S_I\right) \cdot R_D^2 + S_V \left(1 + \frac{R_D}{R_S}\right)^2$$

but since in the last term the ratio between the resistances is negligible with respect to the unitary term:

$$S_{V_1} \simeq 4k_BTR_D + \frac{4k_BT}{R_S} R_D + S_I R_D^2 + S_V.$$

Now, we can notice that:

$$\frac{4}{R_D} \gg \frac{4}{R_S}$$

and therefore:

$$S_{V_1} \simeq 4k_BTR_D + S_I R_D^2 + S_V \simeq 1.6 \cdot 10^{-20} + 10^{-24} + 4 \cdot 10^{-16} \simeq$$
$$\simeq 4 \cdot 10^{-16}\ \mathrm{V}^2/\mathrm{Hz}.$$

Since the transfer function of the noise for this circuit will be a band-pass filter with a pole in 1.6 rad/s and the other pole in 10^6 rad/s, we can give an approximate evaluation of the mean square value of the noise at the output of this device as:

$$\overline{V_{out}^2} = S_{V_1} \cdot G^2 \frac{\pi}{2} (f_{p2} - f_{p1}) \simeq 4 \cdot 10^{-16} \cdot 30^2 \cdot \frac{\pi}{2} \left(10^6 - 1.6\right) \simeq (750\ \mu\mathrm{V})^2.$$

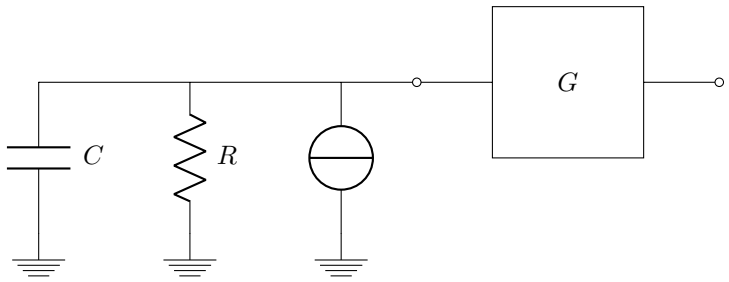


Figure 5.107: Noise contribution of the high-pass filter.

We can then evaluate the noise contribution of the high-pass filter, that is related to the thermal noise in the resistor R, as in Figure 5.107. In this case, the power spectral density can be written as:

$$S_V = S_I \left| \frac{R}{1 + sCR} \right|^2$$

where the input current power spectral density is:

$$S_I = \frac{4k_B T}{R}.$$

This means that we can write:

$$S_V = \frac{4k_B T}{R} \cdot \frac{R^2}{1 + (\omega CR)^2}$$

therefore the mean square value of the output noise can be written as:

$$\overline{n_y^2} = 4k_B T R \frac{1}{4RC} = \frac{k_B T}{C} \simeq (6\ \mu\text{V})^2$$

and we can note that it depends exclusively on the capacitor and that it is negligible.
We have thus determined that we are dealing with a root mean square value of the noise that is equal to 750 μV, while the amplitude of the signal is:

$$A = 200\ \mu\text{V/K} \cdot T.$$

At the output of the amplifier, we will obtained an amplified signal:

$$GA = 6\ [\text{mV/K}] \cdot T\ [\text{K}]$$

and we thus want to measure a difference in the temperature:

$$\Delta T = 0.1^\circ\text{C}$$

with a signal-to-noise ratio that is equal to ten. To do this, we are sampling the square wave and then we are averaging over N different samples. We can now try to find what is the number N of samples that is needed. For this small signal, the amplitude at the end of the amplifier will be:

$$A_m = 6\ [\text{mV/K}] \cdot 0.1\ [\text{K}] = 600\ \mu\text{V}$$

while the root mean square value of the noise will be:

$$\sqrt{\overline{n_y^2}} = 750\ \mu\text{V}$$

thus giving a bad signal-to-noise ratio. Averaging over N samples, the signal-to-noise ratio we expect to obtain is:

$$\left(\frac{S}{N}\right) = \frac{600\ \mu\text{V}}{750\ \mu\text{V}} \cdot \sqrt{N} = 10$$

since the noise correlation is very small and thus we can assume it to be able to approximate it with a white noise.
In this reasoning, however, we are committing an error. In fact, sampling the signal we are considering only half of the amplitude of the signal, therefore two sampling operations are needed to take into account the difference between them and, thus, the whole amplitude of the signal. Since these two sampling operations are distinct from the viewpoint of the noise, the variances of the two noise samples add up, giving an increase of $\sqrt{2}$ in the denominator of the signal-to-noise ratio:

$$\frac{S}{N} = \frac{600\ \mu\text{V}}{750\ \mu\text{V}} \cdot \frac{\sqrt{N}}{\sqrt{2}} = 10 \ \Rightarrow\ N \simeq 313.$$

5.8.5 Exercise 5

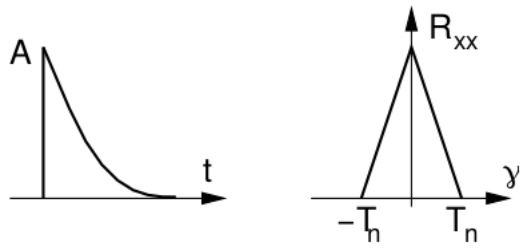


Figure 5.108: Signal and autocorrelation of the noise considered.

This exercise comes from the exam of July 15th, 2013. Consider the signal represented in Figure 5.108, that is exponential with a time constant equal to:

$$\tau = 1\ \mu\text{s}$$

and where we have the following unilateral power spectral density:

$$S_V = 4 \cdot 10^{-16}\ \text{V}^2/\text{Hz}, \quad T_n = 10\ \text{ns}.$$

Suppose we are sampling this signal at certain time instants t_S and, then, we are averaging over these samples.
In the first case, we are performing a uniform average and we have to find the best values for the sampling time t_S and for the number of samples N. The first

requirement that we have to impose is that the distance between two samples is larger than the correlation time of the noise:

$$t_S \geq T_n$$

since we want to be collecting uncorrelated samples. In this condition, the mean square value of the output noise can be written as:

$$\overline{n_{out}^2} = \frac{\overline{n_{in}^2}}{N}.$$

The signal at the output of this device will be:

$$x(t) = Ae^{-\frac{t}{\tau}}$$

therefore its sampled version will be:

$$x(k \cdot t_S) = Ae^{-\frac{kt_S}{\tau}}.$$

At the output, therefore, we have that:

$$y = \sum_{k=0}^{N-1} \frac{Ae^{-\frac{kt_S}{\tau}}}{N} \simeq \frac{A}{N} \cdot \frac{1}{1 - e^{-\frac{t_S}{\tau}}}$$

and thus the signal-to-noise ratio can be written as:

$$\left(\frac{S}{N}\right) = \frac{A}{\sqrt{\overline{n_{in}^2}}} \cdot \frac{1}{\sqrt{N} \cdot \left(1 - e^{-\frac{t_S}{\tau}}\right)}$$

and this will improve for small values of the sampling time t_S. The best choice, in this case, is therefore:

$$t_S = T_n$$

and since according to this choice we have that:

$$t_S \ll \tau$$

we can obtain the following expression for the signal-to-noise ratio:

$$\left(\frac{S}{N}\right) = \frac{A}{\sqrt{\overline{n_{in}^2}}} \cdot \frac{\tau}{\sqrt{N t_S}}.$$

The measurement time therefore is:

$$T = N \cdot t_S \;\rightarrow\; N = \frac{T}{t_S}$$

and thus the signal-to-noise ratio can be written as:

$$\left(\frac{S}{N}\right) = \frac{A}{\sqrt{\overline{n_{in}^2}}} \cdot \frac{\tau}{\sqrt{T t_S}}.$$

From the given data, we can write the mean square value of the input noise as:

$$S_V = 2\overline{n_{in}^2}T_n \;\rightarrow\; \overline{n_{in}^2} = \frac{S_V}{2T_n} = \frac{S_V f_n}{2}$$

where we have considered that this property holds for the bilateral power spectral density, but we were given the unilateral one. In this case, the signal-to-noise ratio can be written as:

$$\frac{S}{N} = \frac{A}{\sqrt{\overline{n_{in}^2}}} \cdot \frac{\tau}{\sqrt{Tt_S}} = \frac{A}{\sqrt{S_V}} \cdot \frac{\tau\sqrt{2t_S}}{\sqrt{Tt_S}} = A\tau\sqrt{\frac{2}{TS_V}} = 1$$

and therefore the minimum detectable amplitude, in this case, is:

$$A_{min} = \frac{1}{\tau}\sqrt{\frac{TS_V}{2}} \simeq 31.6\ \mu\text{V}$$

assuming:

$$T \simeq 5\tau.$$

On the other hand, if we had assumed:

$$T = \tau$$

we would have obtained:

$$A_{min} \simeq 22\ \mu\text{V}$$

while taking just one sample:

$$\left(\frac{S}{N}\right) = \frac{A}{\sqrt{\overline{n_{in}^2}}} \;\rightarrow\; A_{min} = 141\ \mu\text{V}$$

and therefore it is useful to extend this average over different samples.
At this point, we can find the optimum weight and compute the new optimum signal-to-noise ratio. Assuming to have a white noise at the input, the weights must be proportional to the signal, therefore:

$$w_k = e^{-\frac{kt_S}{\tau}}.$$

This means that the output signal can be written as:

$$y = \sum_{k=0}^{N-1} w_k x_k$$

where we have that:

$$x_k = x(k \cdot t_S) = Ae^{-\frac{kt_S}{\tau}}.$$

This gives the following output signal under the assumption of having a large number of samples:

$$y = \sum_{k=0}^{N-1} Ae^{-\frac{2kt_S}{\tau}} \simeq \frac{A}{1 - e^{-\frac{2t_S}{\tau}}}$$

but since:

$$t_S \simeq T_n \ll \tau$$

it gives:

$$y \simeq A\frac{\tau}{2t_S}.$$

The mean square value of the output noise, since we are considering a set of uncorrelated samples:

$$\overline{n_{out}^2} = \overline{n_{in}^2} \cdot \sum_{k=0}^{N-1} w_k^2 = \overline{n_{in}^2} \sum_{k=0}^{N-1} e^{-\frac{2kt_S}{\tau}} \simeq \overline{n_{in}^2} \cdot \frac{\tau}{2t_S}.$$

This gives thus the following signal-to-noise ratio:

$$\left(\frac{S}{N}\right) = \frac{A}{\sqrt{\overline{n_{in}^2}}} \cdot \sqrt{\frac{\tau}{2t_S}}$$

but since:

$$\frac{S_V}{2T_n} = \frac{S_V}{2t_S}$$

it gives:

$$\left(\frac{S}{N}\right) = A\sqrt{\frac{\tau}{S_V}} = 1$$

thus obtaining the following minimum detectable amplitude for the signal:

$$A_{min} = \sqrt{\frac{S_V}{\tau}} \simeq 20\ \mu\text{V}.$$

At this point, we can suppose to be dealing also with an optimum filter in continuous time and study what is the minimum detectable signal. For this filter, the weighting function is proportional to the amplitude of the signal:

$$w(t,\gamma) = e^{-\frac{t}{\tau}}$$

therefore the signal-to-noise ratio can be written as:

$$\left(\frac{S}{N}\right) = \frac{A}{\sqrt{\lambda}} \cdot \sqrt{\int x^2(t)\ dt} = A\sqrt{\frac{2}{S_V}}\sqrt{\int_0^\infty e^{-\frac{2t}{\tau}}\ dt} = A\sqrt{\frac{\tau}{S_V}}$$

where we have considered the unilateral power spectral density as:

$$\lambda = \frac{S_V}{2}.$$

In this case, the minimum detectable amplitude for the signal is:

$$A_{min} = \sqrt{\frac{\lambda}{\tau}} \simeq 20\ \mu\text{V}$$

that is the same value that we obtained for the discrete time optimum filter. This is because the number of samples is so large that basically there is not any difference between the continuous time and the discrete time case.
We can now consider the case of a non-white noise, for which:

$$T_n \ll \tau$$

and we can try to calculate its signal-to-noise ratio remembering that the following integral is valid:

$$\int_0^k xe^{-x}\ dx = 1 - ke^{-k} - e^{-k}.$$

In this case, the associated optimum filter can be written as:

$$w(t,\gamma) = Ge^{-\frac{\gamma}{\tau}}.$$

The mean square value of the output noise, therefore, can be written as:

$$\overline{n_y^2} = \int R_{nn}(\gamma)k_{ww}(\gamma)\ d\gamma$$

and thus we have first to compute the autocorrelation of the weighting function $k_{ww}(\gamma)$. This calculation has already been performed, obtaining the following even exponential behaviour:

$$k_{ww}(\gamma) = G^2\tau e^{-\frac{|\gamma|}{\tau}}.$$

We need therefore to compute this signal as:

$$\begin{aligned}
\overline{n_y^2} &= 2\int_0^{T_n} R_{nn}(\gamma)k_{ww}(\gamma)\ d\gamma = \int_0^{T_n} e^{-\frac{\gamma}{\tau}}G^2\tau\overline{n_{in}^2}\left(1-\frac{\gamma}{T_n}\right)\ d\gamma = \\
&= \overline{n_{in}^2}G^2\tau\int_0^{T_n} e^{-\frac{\gamma}{\tau}}\left(1-\frac{\gamma}{T_n}\right)\ d\gamma - \overline{n_{in}^2}G^2\tau^2\int_0^{\frac{T_n}{\tau}}\left(1-\frac{\tau}{T_n}x\right)e^{-x}\ dx = \\
&= \overline{n_{in}^2}(G\tau)^2\left[1-\cancel{e^{-\frac{T_n}{\tau}}} - \frac{\tau}{T_n}\left(1-\frac{T_n}{\cancel{\tau}}\cancel{e^{-\frac{T_n}{\tau}}} - e^{-\frac{T_n}{\tau}}\right)\right] = \\
&= \overline{n_{in}^2}(G\tau)^2\left(1-\frac{\tau}{T_n}+\frac{\tau}{T_n}e^{-\frac{T_n}{\tau}}\right).
\end{aligned}$$

The output signal, on the other hand, can be written as:

$$y = \int x(\gamma)w(t,\gamma)\ d\gamma = GA\int e^{-\frac{2\gamma}{\tau}}\ d\gamma = \frac{GA\tau}{2}$$

and thus the signal-to-noise ratio will be:

$$\left(\frac{S}{N}\right) = \frac{A}{2\sqrt{\overline{n_{in}^2}}}\cdot\frac{1}{\sqrt{1-\frac{\tau}{T_n}+\frac{\tau}{T_n}e^{-\frac{T_n}{\tau}}}}.$$

5.8.6 Exercise 6

This exercise comes from the exam of September 23rd, 2016. It is given a filter whose weighting function is represented in Figure 5.109 and an input noise with a certain unknown correlation function $R_{nn}(\tau)$. From the graph of the weighting function, we can see that at time $t = 0$ we are sampling the noise with negative weight k, while at time $t = T$ we are sampling the signal and the noise with a unitary positive weight. We want to find the mean square value of the output

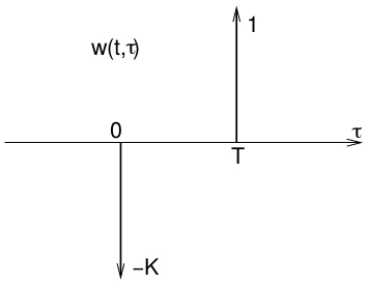


Figure 5.109: Weighting function of the filter considered.

noise and the optimum value of the coefficient k.
From the theory, the mean square value of the output noise can be written as:

$$\overline{n_{out}^2} = \int R_{nn}(\tau) k_{ww}(\tau)\ d\tau$$

where the autocorrelation of the weighting function can be calculated to be:

$$k_{ww}(\tau) = (1+k^2)\delta(\tau) - k\delta(\tau - T) - k\delta(\tau + T).$$

This means that the previous integral can be simply calculated as:

$$\overline{n_{out}^2} = (1+k^2)R_{nn}(0) - 2kR_{nn}(T) = R_{nn}(0)\left[1 + k^2 - 2k\frac{R_{nn}(T)}{R_{nn}(0)}\right].$$

To find the optimum value of k, we have to impose that the mean square value of the noise is minimum:

$$\frac{\partial \overline{n_{out}^2}}{\partial k} = 0 \ \rightarrow\ 2k - 2\frac{R_{nn}(T)}{R_{nn}(0)} = 0$$

thus obtaining:

$$k = \frac{R_{nn}(T)}{R_{nn}(0)}.$$

From this technique, therefore, the mean square value of the output noise can be calculated as:

$$\overline{n_{out}^2} = \overline{n_{in}^2} \cdot (1 - k^2).$$

5.9 Flicker noise and LIAs

5.9.1 Exercise 1

This exercise comes from the exam of June 29th, 2009. It is given the signal represented in Figure 5.96 at page 372, where for the gated integrator we have previously obtained that in the case of an optimum filtering for the white noise the gate time was:

$$T_g = 0.74\ \mu\text{s}$$

where the white noise had the following bilateral power spectral density:

$$\lambda = 10^{-12}\ \text{V}^2/\text{Hz}.$$

Now, we have an additional flicker noise with the following corner frequency:

$$f_{nc} = 10 \text{ kHz}$$

and we know that the system is running for a maximum of eight hours and that the gain of the gated integrator is such that for the following constant input signal we obtain:

$$V_{in} = 10 \text{ mV} \ \rightarrow \ V_{out} = 1 \text{ V}.$$

We have now to calculate the contribution of the white noise and of the flicker noise at the output of this device, stating in particular whether it is convenient to eliminate the flicker noise from this system.
Since we are dealing with a gated integrator, the output of the device will be the amplified integral of the input, thus being

$$V_{out} = G \int_0^{T_g} V_i(t) \ dt = GT_g \cdot 10 \text{ mV} = 1 \text{ V}$$

thus obtaining the following gain at the output of the device:

$$G = \frac{100}{T_g} \quad \left[\frac{1}{\text{s}}\right].$$

It is important to notice that the dimensions of this gain is the inverse of a time. The white noise contribution, therefore, will be:

$$\overline{n^2_{y,WN}} = \lambda T_g G^2 = \lambda T_g \cdot \frac{10^4}{T_g^2} = \frac{10^4 \cdot \lambda}{T_g} = (116 \text{ mV})^2.$$

For the case of the flicker noise, the associated power spectral density will be:

$$S_n = \frac{k}{f}$$

and from the definition of noise corner frequency as the frequency at which the power spectral density of the white noise is equal to the one of the flicker noise:

$$\frac{k}{f_{nc}} = \lambda \ \rightarrow \ k = \lambda \cdot f_{nc} = 10^{-8} \text{ V}^2.$$

Therefore, the mean square value of the output flicker noise:

$$\overline{n^2_{y,FN}} = \int \frac{k}{f} |W(t,f)|^2 \ df$$

but since the weighting function of the gated integrator is rectangular:

$$|W(t,f)| = \text{sinc}(\pi f T_g) \cdot GT_g$$

we obtain:

$$\begin{aligned} \overline{n^2_{y,FN}} &= \int_0^{+\infty} \frac{k}{f} \text{sinc}^2(\pi f T_g) \cdot G^2 T_g^2 \ df = \\ &= k(GT_g)^2 \int_0^{+\infty} \frac{1}{f} \text{sinc}^2(\pi f T_g) \ df. \end{aligned}$$

It is important to notice that we have assumed the flicker noise to have a unilateral power spectral density and that this integral is diverging when it is performed between $f = 0$ and $f \to +\infty$. In the equivalent rectangle approximation, however, the sinc^2 function is transformed into a rectangle that ends either in the first zero of the sinc^2 function[19] or at a certain frequency f_h:

$$\overline{n_{y,FN}^2} \simeq \int_0^{f_h} k(GT_g)^2 \frac{1}{f}\, df.$$

In the zero frequency, this corresponds to an infinitely extended time interval over which we are considering the signal. This is clearly not meaningful, therefore we can replace this lower limit with the minimum frequency that we can detect, that corresponds to the maximum time interval over which we can study the signal:

$$f_l = \frac{1}{T_{tot}} = \frac{1}{8 \text{ hours}} \simeq 34.7\ \mu\text{Hz}.$$

Sinusoidal components at a frequency that is lower than this value, therefore, will only give a constant offset in our signal. This means that our signal can be written as:

$$\overline{n_{y,FN}^2} = k(GT_g)^2 \ln\left(\frac{f_h}{f_l}\right) \simeq (48.7\ \text{mV})^2.$$

It is important to notice that we have evaluated the logarithmic term as:

$$\ln\left(\frac{f_h}{f_l}\right) \simeq 23.7$$

and assuming, for example, that we are off of a factor 2 on one of the frequencies, this will result in an error that is equal to:

$$\pm \ln(2)$$

thus being actually very small. This is due to the fact that this dependence is logarithmic, thus being a very weak dependence. Comparing now this value with the one that we have previously obtained, we can see that the dominating noise term is the white noise one, thus not making really meaningful any additional effort for eliminating the flicker noise:

$$\sqrt{\overline{n_{tot}^2}} = \sqrt{\overline{n_{FN}^2} + \overline{n_{WN}^2}} \simeq 125\ \text{mV} \simeq \sqrt{\overline{n_{WN}^2}}.$$

5.9.2 Exercise 2

This exercise comes from the exam of March 05th, 2012. Again, we are considering the signal represented in Figure 5.97 at page 375 that is coming at the input of an amplifier with gain G and input noise power spectral density S_v:

$$f_0 = 500\ \text{Hz}, \quad S_v = 50\ \text{nV}/\sqrt{\text{Hz}}, \quad G = 100.$$

Assuming to have as a reference signal a square wave signal between -1 and $+1$ that is synchronous with the signal, find a filter for detecting V_A with the following signal-to-noise ratio:

$$\frac{S}{N} = 10.$$

[19] Since we are obtaining a result that depends logarithmically on the frequency, this is an acceptable approximation since it will only give a small error.

To do this, we can try to use a lock-in amplifier in two different cases:

- in the first case, the reference is directly multiplied by the signal and then both are passed to the low-pass filter;
- in the second case, a certain filter (for example a band-pass filter or a selective filter) are added both on the incoming reference and on the incoming signal before the multiplication and then they are passed to the low-pass filter.

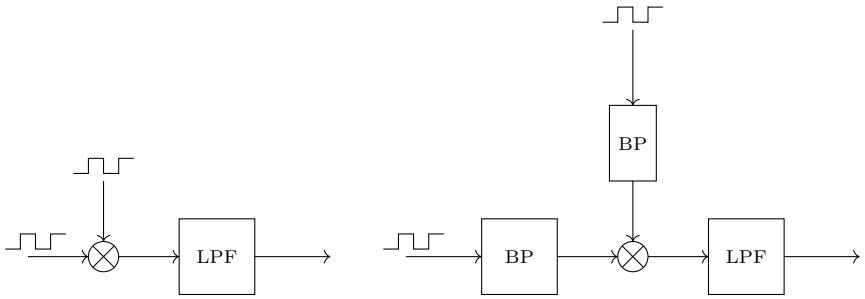


Figure 5.110: On the left, lock-in amplifier without the selective filters; on the right, same scheme but with the addition of these elements.

In the case in which we are using a band-pass filter, only the fundamental oscillation at 500 Hz will pass through this filter, reaching the multiplication stage. In this case, the signal-to-noise ratio:

$$\frac{S}{N} = \frac{GV_{in}}{\sqrt{G^2 4BW_n S_x(f_r)}} = \frac{V_{in}}{\sqrt{4BW_n S_x(f_r)}}$$

where V_{in} is the input amplitude, BW_n is the noise bandwidth of the low-pass filter and $S_x(f_r)$ is the bilateral power spectral density at the reference frequency. However, since the problem is giving us a unilateral power spectral density:

$$\frac{S}{N} = \frac{V_{in}}{\sqrt{2BW_n S_v(f_r)}}$$

where the input noise is white. The amplitude of the fundamental harmonic of the square wave, from the expression of this square wave as a Fourier series:

$$\frac{4B}{\pi}\left[\cos(\omega t) - \frac{\cos(3\omega t)}{3} + \dots\right]$$

can be written as:

$$V_{in} = \frac{V_A}{2} \cdot \frac{4}{\pi} = \frac{2V_A}{\pi}$$

and thus we get the following signal-to-noise ratio:

$$\frac{S}{N} = \frac{2V_A/\pi}{\sqrt{2BW_n S_v}} = 10.$$

From this expression we can obtain the noise bandwidth of the low-pass filter:

$$BW_n = 81 \text{ Hz}.$$

In the other case, where we do not have any filtering on the input signal and reference, we are calculating the product between the two previously defined synchronous square waves, thus obtaining a square wave whose high level is at $V_B + V_A$ and whose low level is at $-V_B$. The low-pass filter, passing only the low-frequency components, will give the average value of this amplitude, that is:

$$V_{out} = \frac{V_B + V_A - V_B}{2} = \frac{V_A}{2}.$$

Since the reference signal can be written, in the frequency domain, as:

$$W_r(f) = \sum_{k=1,3,5,\ldots} B_k \left[\delta(f - kf_r) + \delta(f + kf_r)\right]$$

and therefore the associated frequency response will be:

$$S_{w_r(f)} = \sum_{k=1,3,5,\ldots} B_k^2 \left[\delta(f - kf_r) + \delta(f + kf_r)\right].$$

This gives the following expression of the demodulation signal:

$$\begin{aligned} S_d(f) = S_v(f) * S_{w_r}(f) &= \sum_{k=1,3,5,\ldots} B_k^2 \left[S_v(f - kf_r) + S_v(f + kf_r)\right] \simeq \\ &\simeq 2 \sum_{k=1,3,5,\ldots} B_k^2 S_v(f - kf_r) = 2S_v \cdot \sum_{k=1,3,5,\ldots} B_k^2 = \\ &= 2S_v B_1^2 \cdot \sum_{k=1,3,5,\ldots} \left(\frac{B_k}{B_1}\right)^2 = 2S_v \left(\frac{2}{\pi}\right)^2 \cdot \sum_{n=0}^{+\infty} \frac{1}{(2n+1)^2} = \\ &= 2S_v \left(\frac{2}{\pi}\right)^2 \cdot \frac{\pi^2}{8} = S_v. \end{aligned}$$

This result is consistent with the theory we have studied, since we are multiplying the signal by a square wave whose amplitude ranges from -1 to $+1$, therefore calculating the square modulus we will not change anything on the output with respect to the noise. At the output, therefore, the mean square value of the noise can be calculated as:

$$\overline{n_{out}^2} = \sqrt{S_v BW_n G^2}$$

and thus the requirement on the signal-to-noise ratio:

$$\frac{S}{N} = \frac{V_A/2}{\sqrt{S_v BW_n}} = 10$$

will give the noise bandwidth required:

$$BW_n \simeq 100 \text{ Hz}.$$

Now, we can consider the presence of a sinusoidal interference that is coming to this lock-in amplifier with a frequency equal to $3f_0$. In the case in which we are filtering with a band-pass filter the incoming signal, we can see that this oscillation will not pass through the filter that is centred around f_0, thus allowing us to reject this interference or, at least, to strongly attenuate it in

the demodulation stage, being ultimately rejected in the low-pass filter. On the other hand, in the case in which we are not adopting any band-pass filter, since this is one of the odd harmonics of the fundamental that will be preserved in the demodulation stage (since it corresponds to one of the Fourier components of the square wave), it will be passed by the low-pass filter, determining a decrement in the signal-to-noise ratio. In poor words, we can say that this interference is "leaking" through one of the possible "frequency windows".

5.9.3 Exercise 3

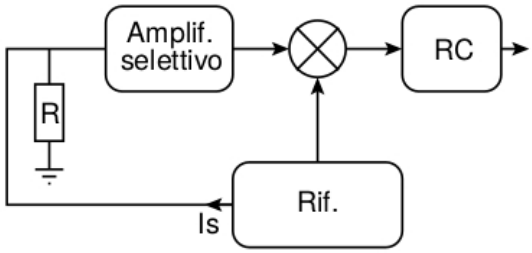


Figure 5.111: Network considered.

This exercise comes from the exam of September 25th, 2009. It is given the network represented in Figure 5.111, where:

$$R \simeq 1 \text{ k}\Omega, \quad BW = 5 \text{ Hz}, \quad S_v = \frac{K}{f}, \; K = 6 \cdot 10^{-8} \text{ V}^2$$

where the power spectral density is unilateral. We would like to measure the resistance R with a resolution of 10 Ω with a signal-to-noise ratio and a maximum signal equal to:

$$\frac{S}{N} = 10, \quad I_S = 10 \; \mu\text{A}$$

and we are ask to find the bandwidth of the low-pass filter and the frequency of the reference signal.
At the input of the low-pass filter, we will have again a baseband signal, therefore we want the bandwidth of this filter to be slightly bigger than the bandwidth of the signal:

$$BW_{LPF} = 10 \cdot BW_{sig} = 50 \text{ Hz}.$$

From this value of the bandwidth of the low-pass filter, we can calculate the associated signal-to-noise ratio:

$$\frac{S}{N} = \frac{V_{in}}{\sqrt{2S_v(f_c)BW_n}} = 10$$

where the noise bandwidth and the flicker noise power spectral density are:

$$BW_n = \frac{\pi}{2} BW_{LPF}, \quad S_v = \frac{K}{f_r}.$$

From these values, we can obtain the required amplitude for the input signal:

$$V_{in} = I_s \cdot \Delta R = 100\ \mu\text{V}$$

where ΔR is the resolution required. This means that the frequency of the reference signal will be:

$$f_r \simeq 93.6\ \text{kHz}\ \rightarrow\ 100\ \text{kHz}.$$

Since in our network we have a selective, bandpass filter, only the sinusoidal component at the fundamental frequency will pass.
Consider now the case in which instead of a resistance R we have a complex impedance:

$$Z = R + jX = R + Xe^{j\frac{\pi}{2}}$$

where we want to contemporarily measure both the real and the imaginary part of this impedance.

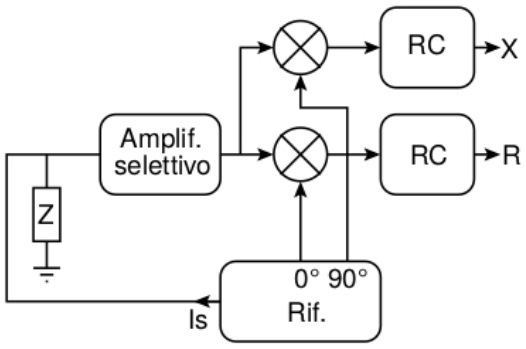


Figure 5.112: Setup for the measurement of the real and imaginary part of the complex impedance.

This can clearly be done as in Figure 5.112 by using two different demodulating networks, one in quadrature with the other, thus having each one of these two reference signals to be synchronous either to the real part of the signal or to the imaginary part (that will have an additional phase of 90°).
Last, we are asked to find the output noise for the network in Figure 5.111 when there is not any selective filter for the signal. In this case, the reference signal in the frequency domain can be written as:

$$W_R(f) = \sum_{k=1,3,5,\ldots} B_k[\delta(f - kf_r) + \delta(f + kf_r)]$$

and therefore the spectral response of the filter can be written as:

$$S_{w_R}(f) = \sum_{k=1,3,5,\ldots} B_k^2[\delta(f - kf_r) + \delta(f + kf_r)].$$

This gives, therefore, the following expression for the power spectral density of the demodulated signal:

$$S_d(f) = S_v * S_{w_R}(f) = \sum_{k=1,3,5,\ldots} B_k^2[S_v(f - kf_r) + S_v(f + kf_r)]$$

thus giving the following power spectral density for the output noise:

$$S_{out}(f) = \sum_{k=1,3,5,\ldots} B_k^2[S_v(-kf_r) + S_v(kf_r)]$$

and considering that this power spectral density ought to be unilateral:

$$S_{out}(f) = \sum_{k=1,3,5,\ldots} B_k^2 \cdot S(kf_r) = B_1^2 \sum_{k=1,3,5,\ldots} \left(\frac{B_k}{B_1}\right)^2 S_v(kf_r)$$

where we have the following amplitude term:

$$B_1 = \frac{4A}{\pi}.$$

In the case of the white noise:

$$\sum_{k=1,3,5,\ldots} \left(\frac{B_k}{B_1}\right)^2 = \sum_{n=0}^{+\infty} \frac{1}{(2n+1)^2} = \frac{\pi^2}{8}$$

while in the case of the flicker noise:

$$\sum_{k=1,3,5,\ldots} \left(\frac{B_k}{B_1}\right)^2 S_v(kf_r) = \sum_{n=0}^{+\infty} \frac{1}{(2n+1)^2} \cdot \frac{K}{(2n+1)f_r} \simeq 1.05\frac{K}{f_r}.$$

Since we have that:

$$\sqrt{1.05} \simeq 1.026$$

we have an improvement of the signal-to-noise ratio in this second case. As in the previous case, we can see that the presence of the selective filter is useful for removing all the possible interferences that are caused by the harmonics of the reference frequency f_r.

5.9.4 Exercise 4

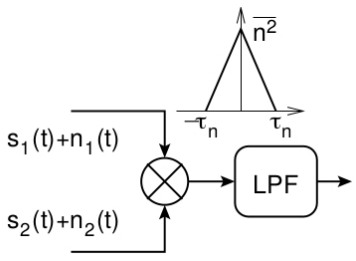


Figure 5.113: The network considered and the temporal autocorrelation of the first noise signal.

This exercise comes from the exam of February 16th, 2015. It is given the network represented in Figure 5.113, where we have that:

$$S_1 = A\cos(\omega_r t) \simeq S_2, \quad f_r = 200 \text{ Hz}$$

and where the autocorrelation of the first noise signal is given in Figure where:

$$\tau_n = 2 \text{ ms}, \quad \overline{n_1^2} = 4 \cdot 10^{-9} \text{ V}^2.$$

Assuming now the following amplitude for the signals and for the second noise:

$$A_1 \simeq A_2 = 10 \ \mu\text{V}, \quad n_2 = 0$$

find the bandwidth of the low-pass filter in order to obtain a signal-to-noise ratio on the first signal that is equal to one.
In this case, the signal S_2 is actually a reference signal and we are dealing with the usual lock-in amplifier. The signal-to-noise ratio can thus be immediately written as:

$$\frac{S}{N} = \frac{A_1}{\sqrt{4S_{n_1}(f_r)BW_n}}$$

where we can write the power spectral density associated to the first noise as:

$$S_{n_1}(f_r) \simeq \text{sinc}^2(\pi f_r \tau_n) \cdot \overline{n_1^2} \tau_n.$$

Therefore, imposing the requirement on the signal-to-noise ratio, we can immediately the noise bandwidth of the device:

$$\frac{S}{N} = 1 \ \Rightarrow \ BW_n \simeq 5 \text{ Hz}$$

and therefore the pole of the low-pass filter will be placed in $2BW_n/\pi$.
Now, the noise n_2 is a white noise with the following power spectral density:

$$\lambda = 10^{-14} \text{ V}^2/\text{Hz}$$

that is totally uncorrelated with the first noise. The system is now used for measuring the cross-correlation in zero for the two signals neglecting the noise terms determine the new signal-to-noise ratio.
In this case, the output signal can be written as:

$$\begin{aligned} x(t) &= (s_1 + n_1) \cdot (s_2 + n_2) = s_1 s_2 + n_1 s_2 + n_2 s_1 + n_1 n_2 \simeq \\ &\simeq s_1 s_2 + s_1 n_2 + s_2 n_1 \end{aligned}$$

where the first term will represent the signal that we would like to measure, while all the other terms will represent noise. From the expression of the signals, we can write:

$$s_1 s_2 = A_1 A_2 \cos^2(\omega_r t) = \frac{A_1 A_2}{2} \left[1 + \cos(2\omega_r t)\right]$$

and considering that at the output of the device we will have a low-pass filter, we will obtain as an output signal:

$$\frac{A_1 A_2}{2} \simeq \frac{A_1^2}{2}.$$

For the noise contribution at the output, we can consider that:

$$s_1 \simeq s_2 \ \Rightarrow \ s_1 n_2 + s_2 n_1 \simeq s_1 (n_1 + n_2)$$

and therefore the output noise of the lock-in amplifier when the two noise sources are uncorrelated can be written as:

$$\overline{n_{out}^2} = A_1^2 \cdot BW_n \cdot S_n(f_r)$$

where the power spectral density of the noise will be the sum of the power spectral densities of the two individual noise contributions:

$$S_n(f_r) = \overline{n_1^2}\tau_n \,\mathrm{sinc}^2(\pi f_r \tau_n) + \lambda.$$

The new signal-to-noise ratio can thus be written as:

$$\frac{S}{N} = \frac{A_1}{\sqrt{4BW_n(\overline{n_1^2}\tau_n \,\mathrm{sinc}^2(\pi f_r \tau_n) + \lambda)}} \simeq 1$$

that is almost identical to the one we had in the previous case since the dominating noise term is the noise on the first signal n_1:

$$\lambda \simeq 10^{-14}\ \mathrm{V}^2/\mathrm{Hz} \ll \overline{n_1^2}\tau_n \simeq 8 \cdot 10^{-12}\ \mathrm{V}^2/\mathrm{Hz}.$$

Now, we have to consider explicitly the term that we previously neglected that was related to the product of the two realizations of the noise processes, thus studying the correlation between these two noises. Defining therefore this term as:

$$n_x(t) = n_1(t)n_2(t)$$

we can calculate its autocorrelation as:

$$R_{n_x n_x}(t, t+\tau) = \overline{n_x(t)n_x(t+\tau)} = \overline{n_1(t)n_1(t+\tau)n_2(t)n_2(t+\tau)}$$

but since the two noise contributions are totally uncorrelated we can write:

$$R_{n_x n_x}(t, t+\tau) = \overline{n_1(t)n_1(t+\tau)} \cdot \overline{n_2(t)n_2(t+\tau)}.$$

This last equivalence can be demonstrated starting from the definition of ensemble average:

$$R_{n_x n_x}(t, t+\tau) = \iint n_1(t)n_1(t+\tau)n_2(t)n_2(t+\tau)\mathbb{P}(n_1, n_2, t, \tau)\ dn_1\ dn_2$$

but since the noise processes are uncorrelated, their joint probability can be written as the product of the single probabilities:

$$\mathbb{P}(n_1, n_2, t, \tau) = \mathbb{P}(n_1, t, \tau) \cdot \mathbb{P}(n_2, t, \tau)$$

thus allowing us to write the ensemble average as the product of the two different ensemble averages for the two noise processes. In this case, since the second noise term is white:

$$R_{n_2 n_2}(\tau) = \lambda \cdot \delta(\tau)$$

and therefore:

$$R_{n_x n_x}(t, t+\tau) = R_{n_1 n_1}(\tau) \cdot \lambda\delta(\tau) = \lambda\overline{n_1^2}\delta(\tau)$$

and we thus obtain a new white noise term with the following power spectral density:

$$S_{n_x}(f) = \lambda \overline{n_1^2} = 4 \cdot 10^{-23} \text{ V}^2/\text{Hz}.$$

This power spectral density is much smaller than any other given one, thus this term is surely negligible with respect to the others.
Last, we can consider the addition of a delay in one of the two branches and study how this affects the noise term in the measurement of the correlation. If we assume that this delay is inserted in the second branch of the device, we can surely say that it will not affect the output noise since the noise term n_2, being white, is stationary, thus not changing with time. The same conclusion, however, could be obtained also inserting this delay in the first branch, since we have just demonstrated the whole output noise to be again stationary (or white). The presence of a delay, regardless of its position, will not affect in any way the output noise term.

5.9.5 Exercise 5

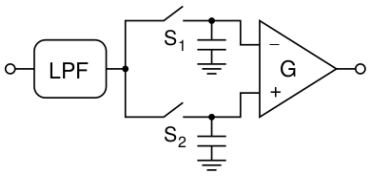


Figure 5.114: The network considered.

This exercise comes from the exam of September 26th, 2014. It is given the network represented in Figure 5.114, where the output at time t is the difference between two samples of the input signal, one taken at time t by the switch S_2 and the other taken at time $t - t_s$ by the switch S_1. The input signal is a rectangular pulse with amplitude A and duration equal to 1 μs. For the sake of simplicity we can assume the gain of this network to be equal to one and we are asked, first, to calculate the weighting function of this filter.
To calculate it, we can consider the output of this device at time t when a delta-function has arrived at a certain time τ. If a delta function arrives between time t and time $t - t_s$, where the only the first switch is closed, then it will determine an exponential decay of the output, that will be maximum at the arrival time of the delta-function and then it will tend toward zero. Between time t and time $t - t_s$, therefore, the weighting function of the filter will be equal to the one of a low-pass filter. If a delta-function is coming before time $t - t_s$, it will start an exponential decay until time $t - t_s$, where it will be sampled, thus remaining constant. This means that the corresponding weighting function will be a negative exponential with its maximum in $t - t_s$. This behaviour can thus be represented as in Figure 5.115.

Now, we have the given rectangular signal on top of a white noise and we are asked to find the parameters of the filter, in particular the time constant T_F. To maximize the signal-to-noise ratio obtained, we can take the first sample just before the arrival of the rectangular signal and the second sample when the signal is present. To have a good measurement, probably it is not good to have

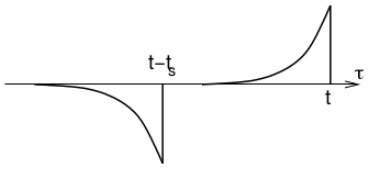


Figure 5.115: Weighting function of the filter considered.

a time constant of the filter that is much larger than the duration of the pulse:

$$T_F \gg T_P \simeq 1\ \mu\text{s}$$

since it will significantly reduce the signal acquired[20]. We can thus impose:

$$T_F < T_P$$

and since five time constants of the filter are needed for reaching the steady-state condition required for the best acquisition of the signal, we can impose:

$$T_F \simeq \frac{T_P}{10} = \frac{1\ \mu\text{s}}{10} = 100\ \text{ns}.$$

This situation is represented in Figure 5.116.

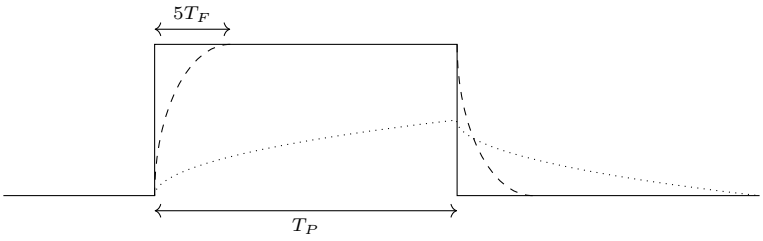


Figure 5.116: Incoming pulse (solid line) sampled with a larger (dotted) or smaller (dashed) time constant of the filter.

Taking then a time interval between the two samples that is:

$$t_s \geq 500\ \text{ns}$$

we are actually safe, thus sampling the signal only when the amplitude of the voltage across the capacitor is at its maximum amplitude A:

$$V_o = A\left(1 - e^{-\frac{t_s}{T_F}}\right) \simeq A.$$

We need now to compute the output noise, that will be the noise at the output of the low-pass filter:

$$\overline{n^2} = \frac{S_v}{4T_F}.$$

[20] This can be seen also considering that the exponential weighting function would experience a fast exponential decay.

At the output of the device, taking the difference of the two samples, if they were uncorrelated we would be increasing the variance of the noise of a factor two. However, the low-pass filter is introducing a correlation in the noise: are thus the two samples uncorrelated or not? If the distance between the two samples is lower than the correlation time

$$t_s > \tau_n$$

then the two samples are uncorrelated and the solution we have just described is fine. At the output of a low-pass filter, the correlation of the noise is the time correlation of an exponential weighting function, that is a bilateral exponential whose maximum value is $\lambda/2T_F$ and whose behaviour can be described by the following function:

$$\frac{\lambda}{2T_F} e^{-\frac{|\tau|}{T_F}}.$$

In this way, therefore, we are relating the correlation time to the time constant of the filter. However, since we know that:

$$t_s > 5T_F$$

the two samples will be indeed not correlated and thus the mean square value of the noise can be written as:

$$\overline{n_{out}^2} = 2 \cdot \overline{n^2} = \frac{S_v}{2T_F}.$$

In the general case, we should have written:

$$\overline{n_{out}^2} = \int R_{xx}(\tau) k_{xx}(\tau)\ d\tau$$

but since the autocorrelation of the noise can be written as:

$$R_{xx}(\tau) = \lambda \cdot \delta(\tau)$$

we obtain:

$$\overline{n_{out}^2} = \lambda k_{xx}(0) = \lambda \int w^2(t,\tau)\ d\tau$$

and explicitly doing this calculation it is possible to obtain the correct value of the mean square value of the noise:

$$\overline{n_{out}^2} = \frac{S_v}{2T_F}\left(1 - e^{-\frac{t_s}{T_F}}\right).$$

Indeed, however, the previous approximation was acceptable.
Now, we can repeat these calculations in the case of a flicker noise. Since the expression of the flicker noise is particularly easy in the frequency domain, we can make use of the Parseval's theorem and write

$$\overline{n_{out}^2} = \int S_v(f) |W(t,f)|^2\ df.$$

Remembering that, from the properties of the Fourier transform, the transform of a reversed exponential centred in $t = 0$ was:

$$\frac{1}{1 - sT_F}$$

while for a reversed exponential centred in t_s we had:

$$\frac{e^{j\omega t_s}}{1 - sT_F}$$

then we can write the spectral response of the filter as:

$$|W(t,f)|^2 = \frac{1}{1 + (2\pi f T_F)^2} \cdot \left|1 - e^{j\omega t_s}\right|^2 .$$

From the well-known properties of the complex exponential, we can write:

$$\left|1 - e^{j\omega t_s}\right|^2 = [1 - \cos(\omega t_s)]^2 + \sin^2(\omega t_s) = 2 - 2\cos(\omega t_s)$$

and thus the spectral response at the end can be written as:

$$|W(t,f)|^2 = 2\,[1 - \cos(\omega t_s)] \cdot \frac{1}{1 + (2\pi f T_f)^2}.$$

Evaluating thus the mean square value of the noise, assuming valid the following approximation:

$$\cos(x) \simeq 1 - \frac{x^2}{2}$$

we can obtain:

$$\begin{aligned}
\overline{n_{out}^2} &= \int_0^{+\infty} \frac{K}{f} \cdot \frac{2(1 - \cos(2\pi f t_s))}{1 + (2\pi f T_F)^2}\, df \simeq \\
&\simeq \int_0^{+\infty} \frac{K}{f} \cdot \frac{(2\pi f t_s)^2}{1}\, df = (2\pi t_s)^2 \cdot K \int_0^{f_h} f\, df = \\
&= K(2\pi t_s)^2 \frac{f_{LP}^2}{2}
\end{aligned}$$

where we have considered that the maximum frequency that we can find at the output is the one of the pole of the low-pass filter:

$$f_{LP} = \frac{1}{2\pi T_F}.$$

We can now comment the effect of this filter on the noise of the amplifier. Assuming a noise generator that, for example, is connected to the positive input pin of the operation amplifier of the network, after the capacitor, we can clearly observe that the input and the output of the low-pass filter in this situation will be placed at zero. Moreover, also the capacitors are set at zero, since this noise source is after them, therefore the filter is completely ineffective in removing this kind of noise, that will completely found at the output of the amplifier.

5.9.6 Exercise 6

This exercise comes from the exam of February 17th, 2014. Given the network represented in Figure 5.117, initially neglecting the presence of the operation amplifier and considering the classical configuration of the Wheatstone bridge (where the negative pin of the operation amplifier is actually directly connected

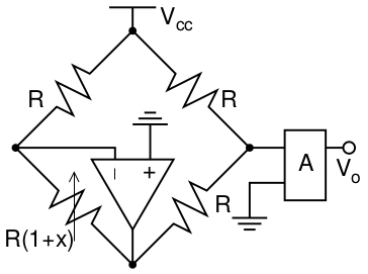


Figure 5.117: The network considered.

at the grounded input of the amplification stage and where the lowest node of the bridge is set to ground) with:

$$V_{cc} = 5 \text{ V}, \quad \Delta V_{cc} = 1\%, \quad R = 350\ \Omega, \quad BW = 10 \text{ Hz}$$

$$x_{max} = 5 \cdot 10^{-3}, \quad 12 \text{ bit of resolution}, \quad CMRR(A) = 60 \text{ dB}$$

calculate the required common-mode rejection ratio neglecting the DC offset of the device.
In this case, the bias voltage of the Wheatstone bridge is fluctuating with an amplitude of 1%:

$$\Delta V_{cc} = 1\% = 50 \text{ mV}.$$

Due to this fluctuation, we have a common-mode variable[21] signal that is equal to:

$$\frac{\Delta V_{cc}}{2} = 25 \text{ mV}.$$

From the expression of the output voltage of a Wheatstone bridge with a single active element, the differential voltage can be written as:

$$\frac{V_{cc}}{4} x_{max} = 6.3 \text{ mV}.$$

In the worst case, the minimum differential voltage that we need to discriminate from the noise will be represented by the least significant bit (LSB) signal:

$$V_{dm,LSB} = \frac{V_{cc}}{4} x_{max} \cdot \frac{1}{2^{12}} \simeq 1.53\ \mu\text{V}$$

and therefore the minimum common-mode rejection ratio required will be:

$$CMRR > \frac{\frac{\Delta V_{cc}}{2}}{V_{dm,LSB}} = \frac{25 \text{ mV}}{1.53\ \mu\text{V}} = 84 \text{ dB}.$$

Since the given common-mode rejection ratio was only 60 dB, this is clearly not enough to reach the required resolution on this measurement.
To improve this common-mode rejection ratio, we can change the circuit as the one actually represented in Figure 5.117. Studying the network, we can

[21] In general, when we are measuring a signal the problems are represented by fluctuations, since constant signals can always be filtered out.

see that the operation amplifier that we have added is working in closed-loop negative feedback configuration with the positive pin grounded. This means that its negative pin will be at virtual ground and, therefore, the output of the operation amplifier, that corresponds to the lowest node of the Wheatstone bridge, will be at a voltage equal to:

$$-V_{cc}(1+x).$$

This means that the only non-grounded pin of the amplifier A will be at a voltage equal to:

$$\frac{V_{cc}}{2} - \frac{V_{cc}(1+x)}{2} = \frac{1}{2}\left[V_{cc} - V_{cc}(1+x)\right] = -V_{cc}\frac{x}{2}$$

and thus we have actually improved the differential signal of a factor two. In this case, the differential mode least significant bit signal can thus be written as:

$$V_{dm,LSB} = \frac{V_{cc}}{2} x_{max} \frac{1}{2^{12}} \simeq 3.06\ \mu\text{V}.$$

Moreover, another advantage is present in this kind of network. In the equilibrium case, in fact, we have that the relative variation of the active element is zero:

$$x = 0$$

and therefore the common-mode voltage is clearly equal to zero. Since we do not have any common-mode voltage, this will not give any problem about the common-mode rejection ratio. The disadvantage of this network is that the introduction of an operation amplifier will increase the complexity of the network and, moreover, since it will give a negative output, now a dual power supply is needed and this feature may or may not be present in our original circuit.
If we now consider the fluctuations on the bias signal of the Wheatstone bridge, assuming ΔV_{cc} to be a residual of the rectification of the power supply at 50 Hz, we can think to add a filter at the output of the amplifier A to improve the signal-to-noise ratio. What is the best filter that we can choose? Obviously, a gated integrator with an integration time that is multiple of the period of the interference at 50 Hz will ensure that this interference is completely rejected. We can thus choose, for example, this integration time to be equal to:

$$T_g \simeq 20\ \text{ms}.$$

In principle, we could have chosen also multiples of this time, however this time interval must not be too long otherwise we end filtering out also the signal we are trying to measure. From a theoretical point of view, also a low-pass filter with a quite step cut-off could have been adopted, thus not, for sure, a single-pole low-pass filter.

5.10 A complete exam test

These exercises come from the exam of July 6th, 2017.

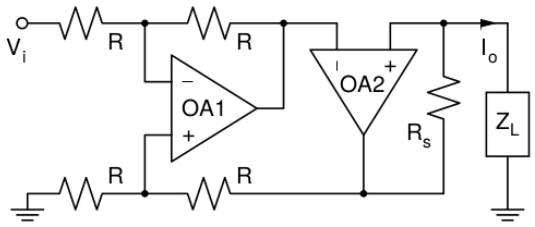


Figure 5.118: The voltage-controlled current source considered.

5.10.1 Exercise 1

It is given the voltage-controlled current source represented in Figure 5.118, where:

$$A_0 = 120 \text{ dB}, \quad GBWP = 1 \text{ MHz}.$$

The first point requires the calculation of the ideal gain of the network. Since we are dealing with a current source or, alternatively, a transconductance amplifier, we can call V_o, that is the output voltage of the network, the voltage at the positive input pin of the second operation amplifier:

$$V_o = V_2^+.$$

Since the second amplifier can be assumed to be in a negative feedback closed-loop configuration[22], defining I as the current flowing at the input of the network and thus also through the feedback resistance of the first operation amplifier, we can write:

$$V_2^+ = V_2^- = V_o \;\Rightarrow\; I = \frac{V_i - V_o}{2R}.$$

This means that the voltage at the negative input pin of the first operation amplifier and, thus, the voltage at the positive input pin of the first operation amplifier can be written as:

$$V_1^- = V_i - RI = V_i - \frac{V_i}{2} + \frac{V_o}{2} = \frac{V_i + V_o}{2} = V_1^+$$

and this could have been demonstrated also considering the superposition principle at these nodes from the input of the circuit V_i and from its output V_o. Since the current flowing in the lower branch of the first operation amplifier will be identical in both the resistors, we can obtain that the voltage at the output of the second operation amplifier will be:

$$V_{out,2} = V_i + V_o.$$

This means that the output current I_o, that is the current flowing through the resistor R_S, will be:

$$I_o = \frac{V_i + V_o - V_o}{R_S} = \frac{V_i}{R_S}$$

[22]This statement will be demonstrated in the second point of the exercise; in principle, we are not sure that this is the right behaviour, but it is likely to be it since it is the most common one.

and this is thus the ideal gain of the amplifier:

$$G_{id} = \frac{1}{R_S}.$$

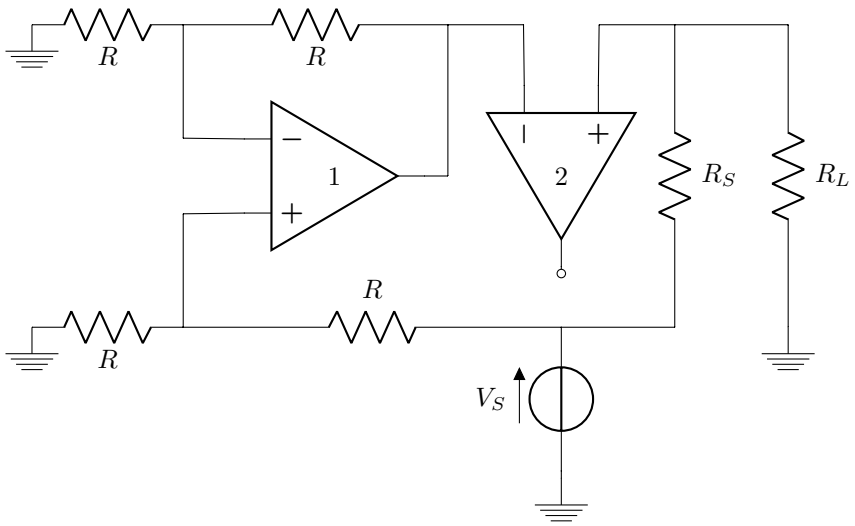


Figure 5.119: The circuit needed for the calculation of $G_{loop,2}$.

The second point consists in calculating the loop gain of the second operation amplifier considering a resistive load R_L and, in a first approximation, considering the first operation amplifier as an ideal one. To do this, we can ground the input V_i and cut the loop at the only common point between the various loops that are involving the second operation amplifier. This common point, that as a side effect will make useless the reconstruction of the impedance, is the output of the second operation amplifier. Imposing then a test source V_S at the node that was previously connected to the output of the second operation amplifier, we can obtain that:

$$V_2^+ = V_S \frac{R_L}{R_L + R_S}$$

while, for the first operation amplifier:

$$V_1^+ = V_1^- = \frac{V_S}{2}.$$

This will give:

$$V_2^- = V_S$$

and from this we can write the output voltage of the second operation amplifier as:

$$V_o = A(s)\left(V_2^+ - V_2^-\right) = A(s)\left(V_S - V_S \frac{R_L}{R_L + R_S}\right) =$$
$$= -A(s)V_S \frac{R_S}{R_L + R_S}.$$

This gives the following expression of the loop gain of the second operation amplifier:

$$G_{loop,2} = -A(s)\frac{R_S}{R_S + R_L} \simeq \begin{cases} -A(s), & \text{if } R_L \ll R_S \\ -A(s)\frac{R_S}{R_L}, & \text{if } R_L \gg R_S \end{cases}.$$

Notice that, in this way, we have demonstrated the second operation amplifier to be in a negative feedback configuration.

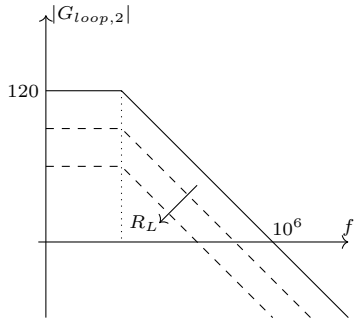


Figure 5.120: Bode diagram of the loop gain of the second operation amplifier in the different limiting cases: the solid line is for $R_L \ll R_S$, the dashed ones for $R_L \gg R_S$.

Since this loop gain is a single pole transfer function, as it is represented in Figure 5.120, from an ideal point of view it should not give any stability issue.

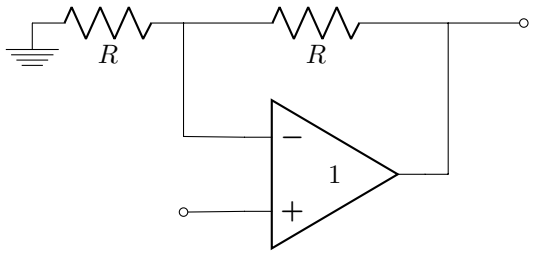


Figure 5.121: A portion of the previous network where we can recognize an ideal amplification stage.

Now, we can refine our analysis by considering also the presence of a pole in the first operation amplifier. Considering that the part of the network that is represented in Figure 5.121 has an ideal gain that is equal to two, we can replace its ideal gain with the real one, thus taking into account the pole that is present in the operation amplifier:

$$G_1 = \frac{G_{id,1}}{1 - \frac{1}{G_{loop,1}}}, \quad G_{id,1} = 2$$

therefore we can obtain:

$$G_1 = \frac{2}{1 + s\tau_c}$$

where the time constant of the pole will be:

$$\frac{1}{2\pi\tau_c} = \frac{GBWP}{2} = 500 \text{ kHz}.$$

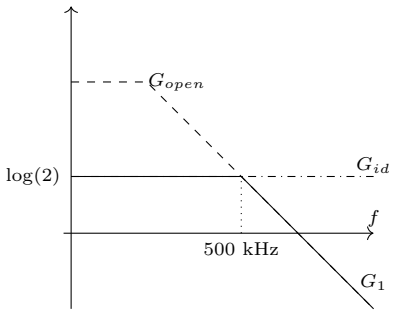


Figure 5.122: Calculation of the real gain of the previous stage.

Therefore, in the calculation of the loop gain of the second operation amplifier we can consider that:

$$V_2^- = V_S \frac{1}{1+s\tau_c}, \quad V_2^+ = V_S \frac{R_L}{R_S+R_L}$$

and therefore we can obtain:

$$G_{loop,2} = A(s)\left(\frac{R_L}{R_S+R_L} - \frac{1}{1+s\tau_c}\right) = -A(s)\frac{R_L+R_S-R_L-s\tau_c R_L}{(R_S+R_L)(1+s\tau_c)} =$$
$$= -A(s)\frac{R_S}{R_S+R_L}\frac{1-s\tau_c\frac{R_L}{R_S}}{1+s\tau_c}.$$

Taking into account the presence of the pole in the first operation amplifier, therefore, we will have an additional pole and an additional non-minimum phase zero in the expression of the loop gain of the second operation amplifier. The frequency of the pole will be completely determined by the frequency of the stage that we have previously considered, while the frequency of the zero will be determined by the value of the load:

$$f_p = \frac{1}{2\pi\tau_c} = 500 \text{ kHz}, \quad f_z = \frac{R_S}{2\pi\tau_c R_L} = \frac{R_S}{R_L}\cdot\frac{GBWP}{2}.$$

Noticing that:

$$\begin{cases} R_L \ll R_S & \Rightarrow \quad f_z \gg f_p \\ R_L \gg R_S & \Rightarrow \quad f_z \ll f_p \end{cases}$$

since we have a non-minimum phase zero we would like it to not came into play in the calculation of the phase, thus being far on the right of the crossover frequency. This means that, for values of R_L since we have two poles on the right of the crossover frequency the phase margin will be smaller than 45°, giving rise to stability issues[23]:

$$\phi_m = 180^\circ - 90^\circ - \arctan\left(\frac{f_{cut}}{f_p}\right) = 90^\circ - \arctan(\sqrt{2}) \simeq 35^\circ.$$

[23]In this case, the crossover frequency can be calculated considering the Bode diagram of the magnitude of the loop gain.

In the other case, when the non-minimum phase zero comes into play, the stability troubles are even worse. In this case, in fact, the non-minimum phase zero is even making the system unstable, as it can be seen considering the associated Bode diagram of the phase in this case.

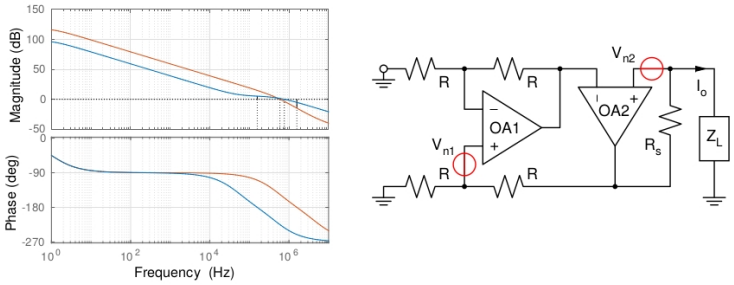


Figure 5.123: On the left, Bode diagram of the loop gain in the two limiting cases; on the right, the voltage-controlled current source considered with the addition of the noise sources.

Now, we are asked to calculate the output noise power spectral density related to the noise equivalent voltage sources S_V in the two identical operation amplifiers. Since the output of this network is a current, we need to find the noise equivalent current power spectral density. We can thus add the noise equivalent voltage source for the operation amplifiers as in Figure 5.123. In this case, dealing with them as if they were normal voltage sources, we can obtain:

$$V_2^- = V_o + V_{n2}$$

from which:

$$V_1^- = \frac{V_o + V_{n2}}{2} = V_1^+$$

and therefore at the output of the second operation amplifier:

$$V_{out,2} = 2\left(\frac{V_o + V_{n2}}{2} - V_{n1}\right) = V_o + V_{n2} - 2V_{n1}.$$

This gives the following output current:

$$I_o = \frac{V_{n2} - 2V_{n1}}{R_S}$$

and thus we can write the associated power spectral density, squaring the modulus of the transfer from the noise sources to the output current, as:

$$S_{I_o} = \frac{|V_{n2}|^2 + 4|V_{n1}|^2}{R_S^2} = \frac{S_V + 4S_V}{R_S^2} = \frac{5S_V}{R_S^2}.$$

Now, we are asked to compensate the second operation amplifier. Since from the previous parts we have found that the loop gain of the second operation amplifier can be written as:

$$G_{loop,2} = -A(s)\left(\frac{1}{1+s\tau_c} - \frac{R_L}{R_L+R_S}\right) = -A(s)\frac{R_S}{R_S+R_L}\frac{1-s\tau_c\frac{R_S}{R_L}}{1+s\tau_c}$$

we can notice that the non-ideality of the first operation amplifier introduces a non-minimum phase zero and a pole: both represent a problem from the stability point of view. To compensate this network, then, different approaches are possible.

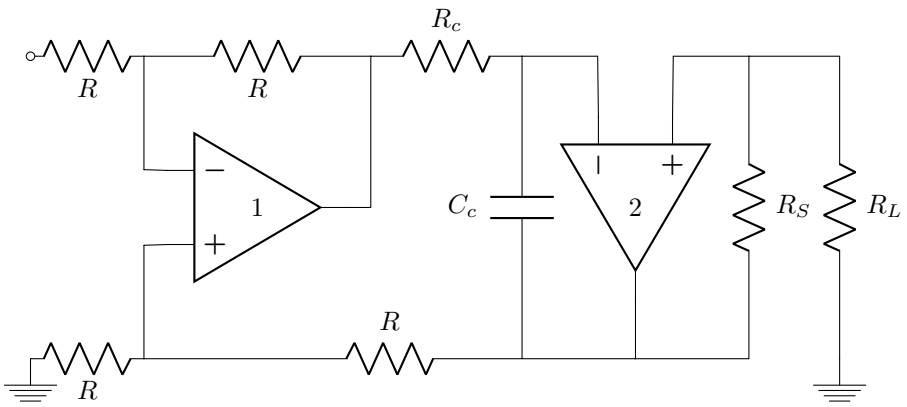


Figure 5.124: A first example of compensation.

For example, as in Figure 5.124, we can bypass the first operation amplifier introducing a capacitor and a resistor. Alternatively, we can observe that the problem of the zero is relevant when we have that:

$$R_L \gg R_S \;\Rightarrow\; \frac{R_L}{R_S + R_L} \simeq 1$$

thus giving, in the high-frequency limit:

$$\frac{1}{1+s\tau_c} - \frac{R_L}{R_L + R_S} \xrightarrow[s\to\infty]{} \frac{1}{s\tau_c} - \frac{R_L}{R_S + R_L} \simeq 0.$$

We can thus try to reduce the term $R_L/(R_S + R_L)$ and therefore to increase the loop gain or, equivalently, pushing the zero to higher frequencies. To reduce the load impedance in the high-frequency limit, we can thus add a capacitor in parallel to it, as in Figure 5.125.

Studying this circuit, we can obtain that:

$$G_{loop,2} = -A(s)\left[\frac{1}{1+s\tau_c}\frac{R_L(1-sC_cR_L)}{\frac{R_L}{1+sC_cR_L}+R_S}\right] = -A(s)\left[\frac{1}{s+\tau_c} - \frac{R_L}{R_L+R_S+sC_cR_LR_S}\right] =$$
$$= -A(s)\left[\frac{1}{1+s\tau_c} - \frac{R_L}{R_L+R_S}\cdot\frac{1}{1+sC_c(R_L\|R_S)}\right]$$

and defining:

$$k = \frac{R_L}{R_L + R_S}$$

we can obtain:

$$G_{loop,2} = -A(s)\left(\frac{1}{1+s\tau_c} - \frac{k}{1+sC_cR_Sk}\right) = -A(s)\frac{1+sC_cR_Sk-k-s\tau_ck}{(1+s\tau_c)(1+sC_cR_Sk)} =$$
$$= -A(s)\frac{1-k+sk(C_cR_S-\tau_c)}{(1+s\tau_c)(1+sC_cR_Sk)}.$$

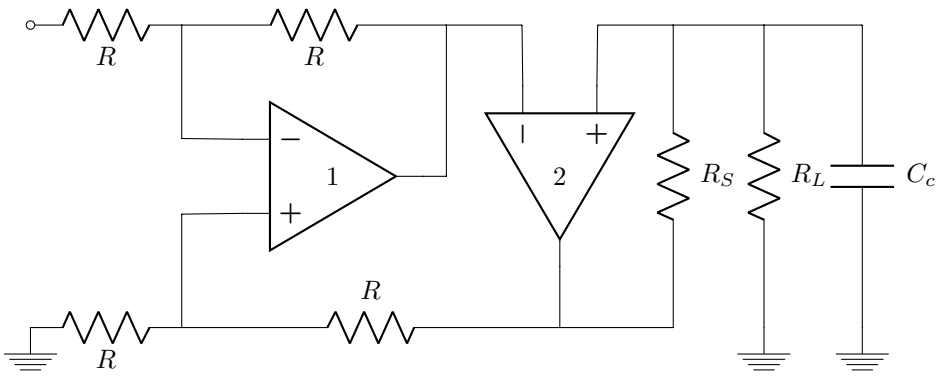


Figure 5.125: Another example of compensation.

Observing that in this case the frequency of the zero is determined by $C_cR_S - \tau_c$ we can move it to the left hand-side of the complex plane by a proper selection of the compensation capacitance:

$$C_cR_S > \tau_c.$$

The frequencies of the zeros and of the poles therefore will be:

$$f_{p1} = \frac{1}{2\pi\tau_c}, \quad f_{p2} = \frac{1}{2\pi C_cR_Sk}, \quad f_z = \frac{1-k}{2\pi k(C_cR_S - \tau_c)}.$$

In the following limiting case:

$$R_L \gg R_S \;\rightarrow\; k \simeq 1 \;\rightarrow\; f_z < f_{p2}$$

the zero will be placed on the left of the second pole, thus increasing the phase margin. With a proper selection of the compensation capacitance, it is also possible to cancel out the second pole. The drawback of this scheme is that we are actually affecting the gain, since at high-frequencies the current will flow through the capacitor C_c, changing the behaviour of the network.
Focusing now on the other limiting condition, in which:

$$R_L \ll R_S$$

we still have a stability issue: we need thus to increase the phase margin.

Considering the network represented in Figure 5.126, we know that without the compensation capacitor that we have just introduced the loop gain of this stage will be:

$$G_{loop,1} = -\frac{A(s)}{2}$$

and it will limit the phase margin of $G_{loop,2}$. To increase this factor 1/2 in the first loop gain, we can change the partition between the two resistances of the network, making the feedback resistance smaller by putting something in parallel to it. In this case, doing all the calculations, we are adding a zero and a pole, obtaining that the loop gain of this stage will cross the zero decibel axis again in the gain-bandwidth product. This means that, in $G_{loop,2}$, we are placing

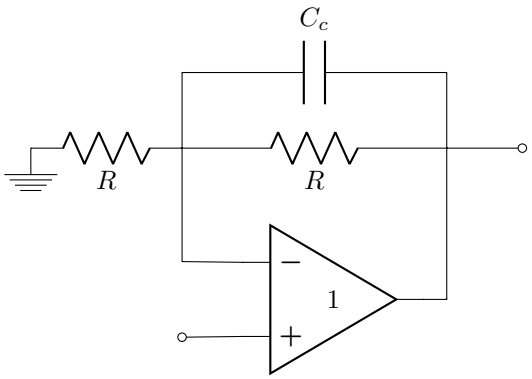


Figure 5.126: Compensation of a portion of the previous network.

an additional pole exactly in the gain-bandwidth product, where it crosses the zero decibel axis, thus increasing the phase margin. This solution, however, in the high-frequency limit will reduce the ideal gain.

Are we able to make again the transfer from the output of the second operation amplifier to the output of the loop of the first operation amplifier again equal to one? To obtain this transfer, we can add another, identical capacitor in parallel to the resistance R that is placed between the positive input pin of the first operation amplifier and the output of the second operation amplifier, as it is shown in Figure 5.127. This capacitor will preserve the symmetry of the circuit and it will improve its performances[24].

5.10.2 Exercise 2

Consider the network represented in Figure 5.128, where we have four different gated integrators that work sequentially. First of all, we have to find the weighting function of this filter and the associated Fourier transform. Since we know that the weighting function of a gated integrator is a rectangle over the integration time T, assuming some undetermined and randomly chosen weights for each one of the various gates, we can obtain a weighting function that can be represented as in Figure 5.129.

In the frequency domain, all these rectangles will give four shifted sinc functions:

$$W(t,f) = w_1 T \operatorname{sinc}(\pi f T) e^{-j2\pi f \frac{T}{2}} + w_2 T \operatorname{sinc}(\pi f T) e^{-j2\pi f \frac{3T}{2}} + \\ + w_3 T \operatorname{sinc}(\pi f T) e^{-j2\pi f \frac{5T}{2}} + w_4 T \operatorname{sinc}(\pi f T) e^{-j2\pi f \frac{7T}{2}}.$$

Now, we can consider the case of a constant input signal superimposed on a white noise; we want to find the optimum value of the weights for this filter in order to obtain the maximum signal-to-noise ratio. Since we know that the signal is constant and that, in the case of white noise optimum filtering, the weighting function must be proportional to the input signal, we can clearly choose:

$$w_1 = w_2 = w_3 = w_4.$$

[24] From the point of view of the solution of this exam, this last consideration was not required.

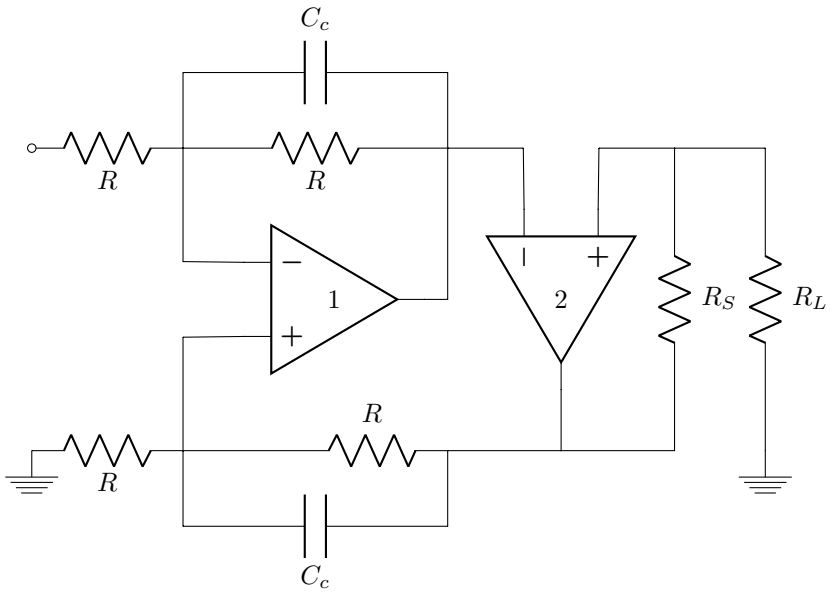


Figure 5.127: A compensation scheme that will not reduce the ideal gain.

In this case, from the expression of the signal-to-noise ratio for a gated integrator, observing that the filter that we have obtained in this case is nothing but a gated integrator with a gate equivalent to $4T$, we can write the associated signal-to-noise ratio as:

$$\frac{S}{N} = \frac{A}{\sqrt{\lambda}}\sqrt{4T}$$

where λ is the power spectral density associated to the white noise.
Now, in addition to this input white noise we have also a constant or very low frequency varying offset; also in this case we want to find the best filter for the signal. To get rid of this offset and contemporarily to filter the white noise, the only possibility is to place some of the gating windows before the arrival of the signal and some of them after the arrival of the signal. Since we want the offset to be subtracted from the signal, we can say that the gates before the arrival time of the signal will have a negative amplitude, while the gates after the arrival of the signal will have a positive one. In this case, depending on how many gates we have before the arrival of the signal (either one, two or three) we can change the weight of these gates and of the remaining one. Assuming to have n negative gates before the arrival of the signal, then we will have $4 - n$ positive gates, as it is represented in the right hand-side of Figure 5.129. In this case, integrating the offset signal, since we want it to not give any contribution at the output:

$$-V_{os}nTw_n + V_{os}(4-n)Tw_p = 0$$

we can obtain the following condition on the weights w_n and w_p associated, respectively, to the negative and to the positive weights:

$$nw_n = (4-n)w_p.$$

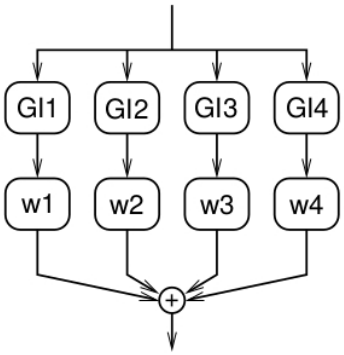


Figure 5.128: The network considered.

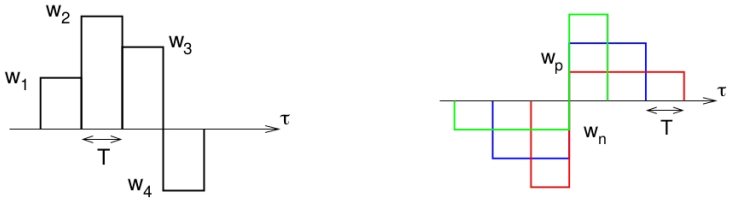


Figure 5.129: On the left, generic weighting function for this filter; on the right, the four different possible weighting function for the case of a constant offset.

This requirement therefore will be needed for not having any output offset signal. The signal-to-noise ratio, in this case, obviously considering at the output only the white noise, will be written as:

$$\frac{S}{N} = \frac{A(4-n)Tw_p}{\sqrt{\lambda w_n^2 nT + \lambda(4-n)Tw_p^2}}$$

where we considered that the white noise, being uncorrelated in every integration gate, will be always added. Substituting the condition for the absence of the offset signal that we have previously obtained, this gives the following signal-to-noise ratio:

$$\frac{S}{N} = \frac{A(4-n)T\cancel{w_p}}{\sqrt{\lambda T}\sqrt{\frac{(4-n)^2}{n}\cancel{w_p^2} + (4-n)\cancel{w_p^2}}} = \frac{A(4-n)T}{\sqrt{\lambda T}\sqrt{\frac{(4-n)^2}{n} + (4-n)}} =$$
$$= \frac{A}{\sqrt{\lambda}} \cdot \frac{\sqrt{(4-n)T}}{\sqrt{\frac{4-n}{n} + 1}} = \frac{A}{\sqrt{\lambda}} \cdot \sqrt{\frac{n(4-n)T}{4}}$$

and therefore maximizing the signal-to-noise ratio we obtain:

$$\max_{n=1,2,3}[n(4-n)] = 4 \;\rightarrow\; n = 2 \;\rightarrow\; \frac{S}{N} = \frac{A}{\sqrt{\lambda}}\sqrt{T}$$

where therefore we have that all the weights have the same modulus but different sign:

$$w_n = w_p.$$

Now, we can consider also the presence of a flicker noise term at the input:

$$S_V = \frac{K}{f}.$$

We have thus to compute the signal-to-noise ratio knowing the value of the following integral:

$$\int_0^{+\infty} \frac{\sin^4(x)}{x^3}\ dx \simeq 0.7.$$

As usual, the mean square value of the output noise will be:

$$\overline{n_y^2} = \int \frac{K}{f}|W(t,f)|^2\ df$$

and computing the square modulus of the Fourier transform of the weighting function:

$$\begin{aligned}|W(t,f)|^2 &= \left|2Tw\,\mathrm{sinc}(2\pi fT)e^{-j2\pi fT} - 2Tw\,\mathrm{sinc}(2\pi fT)e^{j2\pi fT}\right|^2 = \\ &= |2Tw\,\mathrm{sinc}(2\pi fT)\cdot 2j\sin(2\pi fT)|^2 = 4w^2\frac{\sin^4(2\pi fT)}{\pi^2 f^2}.\end{aligned}$$

Computing this integral, therefore:

$$\begin{aligned}\overline{n_y^2} &= \int_0^{+\infty} \frac{K}{f}\cdot 4w^2\frac{\sin^4(2\pi fT)}{\pi^2 f^2}\ df = \int_0^{+\infty} 32w^2T^3K\pi\frac{\sin^4(2\pi fT)}{(2\pi fT)^3}\ df = \\ &= 16w^2T^2K\int_0^{+\infty}\frac{\sin^4(x)}{x^3}\ dx \simeq 16\cdot 0.7(wT)^2K\end{aligned}$$

from which we get[25]:

$$\frac{S}{N} = \frac{A}{\sqrt{16\cdot 0.7K}}.$$

5.11 Another exam test

These exercises come from the exam of September 12th, 2014.

5.11.1 Exercise 1

It is given the circuit represented in Figure 5.130, where we have that:

$$A_0 = 10^5, \quad R_1 = R_3 = 10\ \mathrm{k\Omega}, \quad R_2 = 100\ \mathrm{k\Omega}$$

and where the poles of the operation amplifiers are placed in 10 Hz and 1 MHz. Since this circuit is clearly symmetric, we can study it considering first a common mode input (thus an input that is equal to the two input pins) and then a differential mode input (that an input that is equal in modulus but different in sign at the two pins). In the case of the common mode input:

$$V_{i1} = V_{i2} = V_c$$

[25] In this calculation, there is a wrong, constant numerical coefficient, but it is not so important.

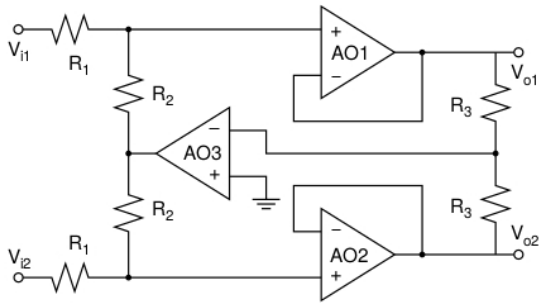


Figure 5.130: Circuit considered.

and observing that:

$$V_1^+ = V_{o1}, \quad V_2^+ = V_{o2}$$

while, thanks to the negative feedback configuration of the third operation amplifier:

$$V_3^- = 0$$

we can state that the current flowing through the two resistors R_3 is identical, thus giving:

$$V_{o2} = -V_{o1}.$$

However, from the analysis of the circuit we can write:

$$V_{o1} = V_c \cdot \frac{R_2}{R_1 + R_2} + V_3 \cdot \frac{R_1}{R_1 + R_2}$$

while for the other output:

$$V_{o2} = V_c \cdot \frac{R_2}{R_1 + R_2} + V_3 \cdot \frac{R_1}{R_1 + R_2}$$

and thus, imposing the previous symmetry condition we obtain that:

$$V_{o2} = -V_{o1} \;\Rightarrow\; V_{o1} = V_{o2} = 0.$$

This means that the common mode output is equal to zero and, thus, we do not have any common mode amplification.
In the case of a differential input, on the other hand:

$$V_{i1} = \frac{V_d}{2}, \quad V_{i2} = \frac{V_d}{2}$$

and thus we expect, again, from the fact that the third operation amplifier is in a negative feedback configuration:

$$V_3 = 0.$$

Also in this case, this means that:

$$V_{o1} = -V_{o2}$$

and from the circuit we can obtain that:

$$V_{o1} = -V_{o2} = \frac{V_d}{2}\frac{R_2}{R_1 + R_2} \simeq \frac{V_d}{2}$$

where we have obtained that the ratio $R_2/(R_1 + R_2)$ can be approximated with one from the given values for these resistors. The differential mode amplification can thus be written as:

$$A_d = \frac{V_{o1} - V_{o2}}{V_d} \simeq 1.$$

Now, we need to compute all the three different loop gains that we can obtain considering one of the operation amplifiers as a real one and the others as ideal ones. We can immediately notice that, from the symmetry of the network, the loop gain of the first operation amplifier will be completely identical to the one of the second operation amplifier:

$$G_{loop,1} = G_{loop,2}.$$

Starting from the third loop gain, we can cut the loop at the output of the third operation amplifier (in a place where there is no need for impedance reconstruction) and we can set a voltage test signal V_S between the two R_2 resistors. In this case, we can immediately write, through a voltage partition, the voltage at the positive input pins of the two remaining ideal operation amplifiers:

$$V_1^+ = V_2^+ = V_S\frac{R_1}{R_1 + R_2}$$

but since, from the negative feedback configuration, they will be to the related output voltages:

$$V_{o1} = V_{o2} = V_1^+ = V_2^+ = V_S\frac{R_1}{R_1 + R_2}$$

and this means that there is not any current flowing through the R_3 resistors, making also the voltage at the input of the third operation amplifier equal to this value:

$$V_3^- = V_{o1} = V_{o2} = V_S\frac{R_1}{R_1 + R_2}.$$

Therefore, the loop gain associated to the third operation amplifier can be written as:

$$G_{loop,3} = -A(s)\frac{R_1}{R_1 + R_2} \simeq \frac{A(s)}{10}.$$

For the loop gain of the first operation amplifier, we can immediately notice that:

$$V_1^- = V_S$$

and since the same current is again flowing through both the resistances indicated with R_3:

$$V_3^- = 0 \;\rightarrow\; V_2^- = -V_S \;\rightarrow\; V_2^+ = -V_S$$

and in the same way, from the symmetry of the circuit:

$$V_1^+ = -V_S.$$

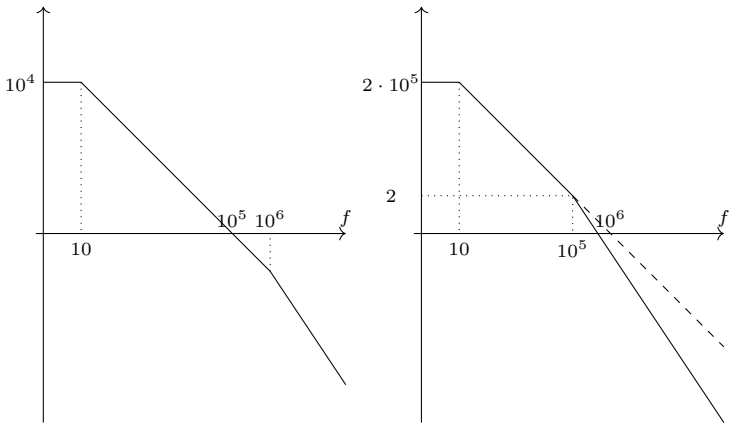


Figure 5.131: Bode diagrams of the magnitude of $G_{loop,3}$ and $G_{loop,2} = G_{loop,1}$ respectively.

This means that we have obtained:

$$G_{loop,1} = -2A(s)$$

and the same must hold for the symmetric loop gain:

$$G_{loop,2} = -2A(s).$$

From the fact that the gain of the operation amplifier is raised by a factor 2, the phase margin is lower than 45°, thus possibly giving some stability problems.

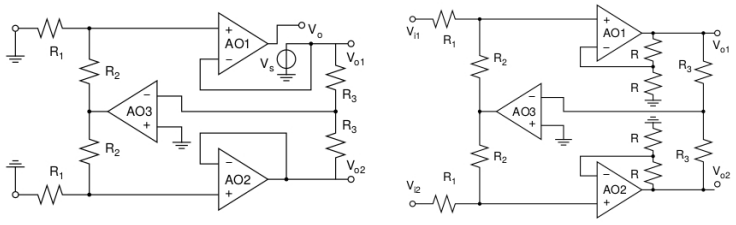


Figure 5.132: On the left, computation of the first loop gain $G_{loop,1}$; on the right, a possible compensation scheme.

We need thus to find a possible way of compensating this scheme. This task is particularly difficult due to the symmetry of the circuit: in fact, we need to take into account that whatever modification we want to do on one side of the circuit must be applied also to the other side. A first possibility is to add the series between a capacitor C_c and a resistor R_c between the positive and the negative input pin of the first and of the second operation amplifiers. In this way, the symmetry of the circuit is preserved and, doing the calculations, it is possible to show that we have actually added a pole and a zero at the following frequencies:

$$f_z = \frac{1}{2\pi C_c R_c}, \quad f_p = \frac{1}{2\pi C_c (R_c + R_1 \| R_2)} < f_z.$$

This solution is good also because when we are dealing with the loop gain of the first operation amplifier, thus assuming the second operation amplifier as an ideal one, since the two input pins of the second operation amplifier will be kept at the same voltage because of the fact that we are dealing with a negative feedback network the associated compensating elements will not come into play. Another possibility, in this case, is to lower the loop gain associated to the first and to the second operation amplifier until the second pole is placed exactly at the crossover frequency or at an higher frequency. In general, this is not the correct way of compensating a network but, in this case, we can adopt it since it requires a reduction of the loop gain only of a factor two, that is affordable. To reduce the loop gain of this quantity, we can place two additional resistors at the output of the first and of the second operation amplifiers, obtaining the network that is represented in Figure 5.132. In this way, we are actually changing the configuration of the network involving the first or, equivalently, the second operation amplifier from the one of a buffer stage (thus with unitary gain) to the one of a non-inverting amplifier with a gain equal to two. Since the gain-bandwidth product of the operation amplifier is constant, increasing the gain we are thus reducing the bandwidth of the operation amplifier. A problem, however, may arise. In fact, in this case, under a purely differential signal we are obtaining a differential gain that is equal to two, thus being different from the one we had in the previous, uncompensated network. Since this might be an unwanted effect, to restore the unitary gain we have to impose the following condition:

$$R_1 = R_2.$$

From the viewpoint of the noise, we now have to find the differential noise output power spectral density for the case of a white noise voltage equivalent source with power spectral density:

$$S_V = 10 \text{ nV}/\sqrt{\text{Hz}}$$

in each operation amplifier, thus neglecting all the noise contributions that are coming from resistances and noise equivalent current sources. Considering the original circuit, then, we can add a noise equivalent voltage source in series to the positive input pin of each operation amplifier of the circuit and we can ground the usual inputs. Considering now only the noise contribution the first operation amplifier, we can obtain that since the third operation amplifier is in a negative feedback configuration:

$$V_3^- = 0 \ \rightarrow \ V_1^+ = V_{o1}$$

and therefore, defining V_4 the only common node between the resistors R_1 and R_2 in the upper part of the network, we have that:

$$V_4 = V_{o1} - V_{n1}$$

where V_{n1} is the noise equivalent voltage source for the first operation amplifier. However, from the symmetry of the circuit and from the inspection of the lower part of the network, this implies that:

$$V_{o2} = V_{o1} - V_{n1}$$

and therefore the differential output noise voltage will be V_{n1}. The same result can be obtained considering a noise equivalent voltage source for the second operation amplifier. Considering now the third operation amplifier, the noise equivalent voltage source will give the fact that:

$$V_3^- = V_{n3}$$

while from the rest of the network we have that:

$$V_1^+ = V_{o1} = V_2^+ = V_{o2}.$$

In this case, therefore, the differential output noise voltage for the third operation amplifier will be identically equal to zero. This is consistent with the fact that, in this last case, we are biasing the network in a symmetry point, therefore there will not be any way of breaking this symmetry obtaining a differential output voltage different from zero.
In the last part of this exercise, we are now asked to discuss the common-mode rejection ratio of this network. From an ideal point of view, we have already seen that the common-mode amplification is equal to zero. However, we can now assume to be in a real case and, therefore, all the resistors in the lower part of the network could be slightly different from the ones in the upper part of the network; we will thus rename all the resistors placed in the lower half with a prime apex. Then, we have to take into account that a slight variation of in resistors is possible:

$$R \ \rightarrow \ R(1 \pm x)$$

and we will always take into account the worst case. From the fact that the third operation amplifier is a negative feedback configuration, we can say that:

$$V_3^- = 0$$

and therefore, since the same current is flowing through R_3 and R_3', we will have:

$$V_{o2} = -V_{o1}\frac{R_3'}{R_3}.$$

However, since also the other operation amplifier are in a negative feedback configuration:

$$V_1^+ = V_{o1}, \quad V_2^- = V_{o2}$$

and thus we get, defining V_3 the output of the third operation amplifier and considering a common-mode input voltage V_c:

$$V_{o1} = V_c\frac{R_2}{R_1+R_2} + V_3\frac{R_1}{R_1+R_2}$$

and:

$$V_{o2} = V_c\frac{R_2'}{R_1'+R_2'} + V_3\frac{R_1'}{R_1'+R_2'}.$$

Solving the system that we have obtained with the previous four equations, we obtain that:

$$V_3 = V_{o1}\left(1+\frac{R_2}{R_1}\right) - V_c\frac{R_2}{R_1} = V_{o2}\left(1+\frac{R_2'}{R_1'}\right) - V_c\frac{R_2'}{R_1'}.$$

Solving the second and the third member of this expression for the first output voltage:

$$V_{o1}\left(1+\frac{R_2}{R_1}+\frac{R_3'}{R_3}+\frac{R_3'}{R_3}\frac{R_2'}{R_1'}\right)=V_c\left(\frac{R_2}{R_1}-\frac{R_2'}{R_1'}\right).$$

We can immediately observe that the term between brackets in the left hand-side of the equation contains only higher order terms and, therefore, it will not be identically equal to zero in a first order approximation where we assume:

$$R_1'\simeq R_1,\quad R_2'\simeq R_2,\quad R_3'\simeq R_3.$$

On the other hand, the term between brackets in the right hand-side of the equation will vanish in this limiting case, therefore we have to keep all these resistors slightly different one from the other. This therefore gives:

$$V_{o1}\left(1+\frac{R_2}{R_1}+\frac{R_3}{R_3}+\frac{R_3}{R_3}\frac{R_2}{R_1}\right)=2V_{o1}\left(1+\frac{R_2}{R_1}\right)\simeq V_c\left(\frac{R_2}{R_1}-\frac{R_2'}{R_1'}\right)$$

and therefore, including the possibility of these resistors to vary, in the worst case we will obtain:

$$2V_{o1}\left(1+\frac{R_2}{R_1}\right)=V_c\left(\frac{R_2(1+x)}{R_1(1-x)}-\frac{R_2(1-x)}{R_1(1+x)}\right)=V_c\frac{R_2}{R_1}(1+2x-1+2x)=$$
$$=V_c\frac{R_2}{R_1}4x$$

and, as it is expected, this result is proportional to x. Therefore, we obtain that:

$$\frac{V_{o1}}{V_c}=\frac{R_2}{R_1}\cdot 2x\cdot\frac{1}{1+\frac{R_2}{R_1}}=\frac{2xR_2}{R_1+R_2}\simeq 2x$$

and the common-mode amplification factor is:

$$A_c=\frac{\frac{V_{o1}+V_{o2}}{2}}{V_c}.$$

From the expression of the second output voltage, however:

$$V_{o2}=-\frac{R_3'}{R_3}V_{o1}=-V_{o1}(1\pm 2x)$$

and thus we get:

$$\frac{V_{o1}+V_{o2}}{2}=\pm xV_{o1}=2x^2V_c$$

from which, at the end, we obtain:

$$A_c=2x^2.$$

We can immediately notice that this amplification factor is of the second order with respect to x. This has two important consequences: the first one is that the common-mode amplification factor is actually very low and the second one is that this result is, actually, wrong. In fact, we have previously neglected all the second order terms, therefore we must trust our result only as a first order approximation, where it only says that the common-mode amplification factor is lower than the first order, thus giving a common-mode rejection ratio that is quite high.

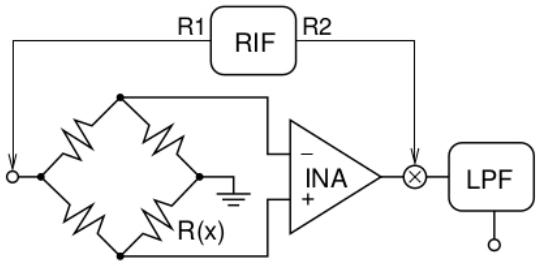


Figure 5.133: Network considered.

5.11.2 Exercise 2

It is given the network represented in Figure 5.133 that we want to use to measure a temperature variation between 0°C and 100°C. We know that the active element of the Wheatstone bridge is characterized by the following coefficient:

$$TCR = 4 \cdot 10^{-3} \ /^{\circ}\mathrm{C}$$

while the bandwidth of the signal is 1 Hz and the output dynamics, equal to the range between 0 and 5 V, is sampled with an 8-bit analog to digital converter (ADC). The system is affected by a flicker noise with unilateral power spectral density:

$$S_V = \frac{K}{f}, \quad K = 10^{-8} \ \mathrm{V}^2$$

and the reference signal can be written as:

$$R_1, \ R_2: \quad A\cos(\omega t + \varphi), \quad A = 1 \ \mathrm{V}.$$

The first task is to set the correct parameters for this acquisition system in order to obtain a signal-to-noise ratio equal to 10. From the theory on the Wheatstone bridge, we know that the voltage that is coming to the instrumentation amplifier will be:

$$V_{in} = V_{R1}\frac{x}{4} = \frac{V_{cc}}{4}TCR \cdot \Delta T = \frac{A}{4}TCR \cdot \Delta T$$

and therefore the maximum possible input signal will be, for the maximum temperature that is 100°C:

$$V_{in,max} = \frac{A}{4}TCR \cdot \Delta T_{max} = 100 \ \mathrm{mV}.$$

From the dynamic of the analog to digital converter, since at the output of the low-pass filter we are obtaining the continuous component of the square of a sinusoidal waveform, that is constant and equal to the input amplitude multiplied by 1/2, we can set the required gain of the instrumentation amplifier to match the full dynamic of the converter:

$$G = 100.$$

The bandwidth of the low-pass filter, then, will be related to the one of the signal and, to be in a safe condition, we can assume it to be a decade larger than the one of the signal:

$$BW_{LPF} = 10 \cdot BW_s = 10 \ \mathrm{Hz}.$$

Last, we have to impose the requirement on the signal-to-noise ratio in order to get the frequency of the reference signal, assuming that it does not have any phase. From the theory of the lock-in amplifiers, we can write the associated signal-to-noise ratio as:

$$\frac{S}{N} = \frac{V_{min}}{\sqrt{2S_V(f_r)BW_n}}$$

where the least significant bit of the converter gives the minimum signal that we want to discriminate:

$$V_{min} = \frac{V_{in,max}}{2^8}.$$

From the expression of the flicker noise at the reference frequency and from the one of the noise bandwidth, that is related to the bandwidth of the filter:

$$S_V(f_r) = \frac{K}{f_r}, \quad BW_n = \frac{\pi}{2}BW_{LPF}$$

and imposing the requirement on the signal-to-noise ratio we obtain the following frequency:

$$\frac{S}{N} = 10 \;\Rightarrow\; f_r \simeq 130 \text{ Hz}.$$

Now, we have to consider a step change in the temperature from 0°C to 20°C and study the behaviour of the signal. At the input of the instrumentation amplifier, the signal can be written as:

$$V_{in} = V_{R1}\frac{x}{4} = V_{R1}TCR\Delta T$$

and therefore it is constant and equal to zero before the step, while it is oscillating with a sinusoidal behaviour of amplitude equal to 20 mV after it. Then, at the output of the instrumentation amplifier, before the demodulation stage, it will be identical to the previous signal, thus showing after the step an oscillating behaviour with amplitude equal to 2 V (since the gain of this stage has been set at $G = 100$). After the demodulation stage, where we are multiplying the oscillating signal with a sinusoidal reference, we obtain a zero constant signal before the arrival of the step, while after we have an oscillating behaviour corresponding to the square of a sinusoidal function and whose average value is 1 V. At the output of the low-pass filter, then, we are obtaining only this continuous component of the oscillation, therefore the signal will be equal to zero before the arrival of the step, while after it it will increase with an exponential behaviour toward 1 V with a time constant that is determined by the time constant of the low-pass filter. These signals are represented in Figure 5.134, where in the signal at the output of the low-pass filter can be found also a residual, small-amplitude oscillation at twice the frequency of the reference that is related to this demodulation stage. The remaining part of the exercise is left to the willing student.

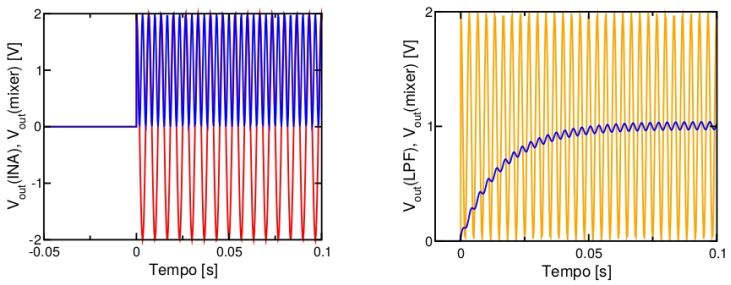


Figure 5.134: Signals in the various parts of the network: on the left we have the signals at the output of the instrumentation amplifier and of the demodulation stage; on the right, again the signal at the output of the demodulation stage and the signal at the output of the low-pass filter.