Noise
and
Signal
Processing

Extra syllabus for (third-year) course
“Signaalverwerking & Ruis”

Martin van Exter
Universiteit Leiden
August 2003
Contents

1 Introduction 5

2 Statistical analysis of noise 9
  2.1 The variance of the noise: \( \langle N(t)^2 \rangle \) 9
  2.2 Noise strength behind spectral filters: \( S_N(f) \) 12
  2.3 Fourier analysis of noise: Wiener-Khintchine 14

3 Different types of noise 19
  3.1 Noise with different spectral properties 19
  3.2 Thermal white noise 21
  3.3 Shot noise 23
  3.4 1/f noise 26
  3.5 Transfer of signal and noise 28
  3.6 Some quantitative examples 30

4 How to improve S/N 33
  4.1 Four techniques to improve S/N 33
  4.2 Time averaging and low-frequency filtering 35
  4.3 Correction for offset and drift 37
  4.4 Multiple time averaging 39
  4.5 Modulation and phase-sensitive detection 43
  4.6 Noise optimization in parameter fitting 46
  4.7 Some quantitative examples 48

5 Analog-to-digital conversion and sampling 51
  5.1 Hardware for Analog-to-Digital Conversion 51
  5.2 Consequences of sampling: bitnoise & Shannon’s theorem 52

6 FFT and z-transform 55
  6.1 Discrete & Fast Fourier Transform 55
  6.2 Reduction of spectral leakage by windowing 56
6.3 Noise synthesis .......................................... 58
6.4 The z-transform ........................................ 60
Chapter 1

Introduction

What is noise? In our daily life noise of course refers to loud, disagreeable sound without any musical aspirations. In the early days of radio communication the word noise was introduced to describe “any unwanted (electrical) signal within a communication system that interferes with the sound being communicated” (quote from the Webster dictionary), which is thus audible as “noise” on a headphone. In the context of physical experiment the word noise is more general and refers to “any unintentional fluctuations that appear on top of signals to be measured”. Any quantity can exhibit the random fluctuations that we call noise. In electronic circuits we deal with voltage noise and current noise caused by among others the thermal fluctuations of the electronic carriers. In the radio and microwave region we deal with electro-magnetic fluctuations caused by the thermal or spontaneous emission of low-energetic photons. But noise can also refer to unintentional fluctuations in other quantity, like the traffic flow on a highway, or the rhythm of water droplets on a roof.

Noise is omnipresent; whenever one tries to measure a signal there is always some form of noise to be accounted for. Even in physics “there is no free lunch”; high-quality cutting-edge experiments always require serious work, which generally also includes an evaluation of the noise sources and some tricks to reduce the influence of this noise. The importance of noise analysis becomes immediately clear when one realizes that the quality of experimental data is not determined by the absolute strength of the signal, but rather by the ratio of signal strength over noise strength. From research experience I can tell you that a good signal-to-noise ratio is often more easily obtained via a reduction of the noise strength than via an increase of the signal strength.

Now that you have a rough idea of the nature of noise, let me also tell you what I don’t consider to be noise, but would rather rank under a different
name. In my dictionary noise is always an unintentional fluctuation, with a long-time average equal to zero. This means that any static offset should not be ranked under noise, but should just be called “offset”, or “systematic measurement uncertainty or error”. Furthermore, I would also like to make a distinction between noise and perturbations. For me noise is always “truly random”, with an amplitude and phase that is intrinsically unpredictable and only satisfy certain statistical properties. Perturbations, on the other hand, do have a well defined frequency, amplitude, and phase. Although often unintentional and unwanted, perturbations are not really random but predictable instead, and often even avoidable with clever design. A classical example of a perturbation is the electronic pickup of the voltage and current variations in the power grid at frequencies of 50 Hz and multiples thereof.

The distinction between “random noise” and “perturbations” in a time-varying signal $V(t)$ is more or less equivalent to the distinction between the random and systematic errors that can exist in measurements of a single quantity $V$. This syllabus deals specifically with random noise, as this noise often has a clear physical origin (and in many cases even a fundamental lower limit), whereas perturbations depend specifically on the design of the experiment and can in principle be avoided.

This syllabus is ordered as follows: in chapter 2 we discuss the fundamental statistical and dynamic properties of noise and ways to measure these. In chapter 3 we show how these dynamic or spectral properties allows one to divide noise in various classes, like white noise, pink noise, 1/f noise, and drift. Chapter 4 discusses some techniques that can be used to improve the signal-to-noise ratio (S/N) in experimental situation. For an experimental physicist this chapter probably contains the most practical information, but chapters 2-3 are definitely needed to provide for the theoretical framework and understanding. Chapter 4 shows that especially the time-variation or frequency contents of both signal and noise provides for handles to manipulate and improve the S/N ratio. Chapter 5 discusses the coupling between the real world, with its analog signals and noises, and the computer world, with its digital information. Issue to be discussed are bit noise and the sampling rate in relation to the available signal and noise frequencies. Finally, chapter 6 discussed two mathematical tools for the processing of digital information, namely the fast (digital) Fourier transform and the $z$-transform.

This syllabus was written in the summer of 2003. The first part was inspired by old notes from C. van Schooneveld [3]; the second part is a mixture of several books, among others the book “Signal recovery (from noise in electronic instrumentations)” by T.H. Wilmshurst [2]. This syllabus is an essential ingredient of the course with the dutch title “Signaal Verwerking & Ruis (SVR)”, as it treats the various aspects of noise and noise reduction
in a much more extended and advanced way than the book *Instrumentele Elektronica* by P.P.L. Regtien [1], which was used in the first part of that course to teach analog electronics. However, it is still useful to read (or reread) the sections from Regtien’s book that have some connection with the present subject of “noise and signal processing”. These sections are:

- Hoofdstuk 2 *Signalen*
- §5.2 *Het modelleren van stoorsignalen*
- §17.2 *Signaalbewerkingssystemen met synchrone detectie*
- Hoofdstuk 18 *Digitaal-analoog en analoog-digitaal omzetters*
Chapter 2

Statistical analysis of noise

2.1 The variance of the noise: \( \langle N(t)^2 \rangle \)

Random noise is per definition uncontrollable, and its precise realization will differ from experiment to experiment. Random noise should thus be analyzed and characterized in statistical terms. In this chapter we will introduce the basic concepts for such a statistical analysis in relatively loose terms; for a more exact discussion we refer to refs. [4, 5, 6].

We will use the following notations. We will denote the combination of a time-varying signal \( S(t) \) plus noise \( N(t) \) as

\[
V(t) = S(t) + N(t) = V_0 + \Delta V(t) \, ,
\]

but use the second notation only if the signal is a constant voltage so that the noise voltage is the only fluctuating term. We choose to exclude a possible offset or static perturbation from our definition of random noise and take the expectation value \( \langle N(t) \rangle = 0 \). The strength of the noise is now fully described by its variance

\[
\text{var}(N) \equiv N_{\text{rms}}^2 \equiv \langle N(t)^2 \rangle \, .
\]

The notation “rms” stands for root-mean-square, so that \( N_{\text{rms}} \) is read as the “root-mean-square noise (fluctuations)”. The brackets \( \langle \rangle \) denote “ensemble averaging”, i.e., averaging over many different realizations of the same physical system.

In most systems noise has two important general properties: it has a Gaussian probability distribution and its statistical properties are stationary. These properties are a natural consequence of the notion that noise is often generated by the combined action of many microscopic processes. Whenever noise is generated out of a large sum of very small contributions, the central
limit theorem tells us that the probability distribution $p(N)$, which quantifies the probability to find a certain noise amplitude $N$, indeed has a Gaussian shape of the form:

$$p(N) = \frac{1}{N_{\text{rms}}\sqrt{2\pi}} \exp \left( -\frac{N^2}{2N_{\text{rms}}^2} \right).$$

(2.3)

This Gaussian form is generic for random noise, but the probability distribution for other types of perturbation can of course be different. As an example we mention a possible harmonic fluctuation (like 50 Hz pick-up), which has a probability distribution that is strongly peaked around the extreme values (or turning points) of the fluctuation. If these types of perturbations are present we will treat them separately and exclude them from the definition of $N(t)$.

Figure 2.1: Example of voltage noise generated by a simple electrical resistor. It is a law of nature that any dissipative element at non-zero temperature generates what is called thermal noise. This law is called the fluctuation-dissipation theorem (see Chapter 3).

Figure 2.2: Active electronic elements, like transistor amplifiers, also generate electronic noise. The amount and spectral contents of noise from active devices can be quite different from that of passive devices. However, even this kind of noise is generally stationary and Gaussian.
2.1. THE VARIANCE OF THE NOISE: \( \langle N(T)^2 \rangle \)

Noise is called \emph{stationary} whenever its statistical properties are independent of time. This means that its variance \( \langle N(t)^2 \rangle \) should be independent of \( t \). But also other statistical properties, like the autocorrelation function \( \langle N(t)N(t+\tau) \rangle \), which correlates the noise amplitude at time \( t \) with that at time \( t + \tau \), should be independent of \( t \). Noise is stationary whenever it is generated by the combined action of many microscopic processes in a physical system that is stable on a macroscopic scale. In such a system, which is called “ergodic”, the ensemble- and time-averaging can readily be swapped and the property is obvious.

An important example of random noise are the voltage fluctuations over an electrical resistor that result from the combined action of small displacements of many charge carriers inside the resistor (see Fig. 2.1). The strength of this kind of noise, or more specifically its variance within a given frequency bandwidth, is proportional to the temperature. As such voltage fluctuations arise from the combined action of many individual noise sources (= moving carriers) they obey a Gaussian probability distribution and are stationary whenever the temperature is constant. If temperature changes the noise strength will also change and the noise becomes a-stationary. This is demonstrated in Fig. 2.3.

![Diagram showing stationary and a-stationary noise](image)

\footnote{Figure 2.3: \textit{Noise is called stationary (in dutch: stationair) when it’s statistical properties do not change in time; otherwise it is called a-stationary. As an example we mention the (thermally-generated) voltage fluctuations over an electrical resistor; these are stationary when the temperature of the resistor is fixed, but are a-stationary and vary in strength when the temperature changes.}}

In systems with multiple noise source, where the noise amplitude can be written as \( N(t) = N_1(t) + N_2(t) + \ldots \), the key question to be answered is “are all noise sources \( N_i(t) \) independent?” If the answer is yes, the variance of the total noise is naturally given by
\[ \langle N(t)^2 \rangle = \langle N_1(t)^2 \rangle + \langle N_2(t)^2 \rangle + \ldots \]  

(2.4)

as any cross correlation between two independent noise sources, with a form like \( \langle N_1(t)N_2(t) \rangle \), is per definition zero for independent noise sources. A short (but somewhat sloppy) description of this important Eq. (2.4) is that “rms noise amplitudes of independent noise sources add quadratically”. A better description is that “you have to add the noise power (\( \propto \) square amplitudes) of uncorrelated noise sources”.

We can give various, somewhat related, extra arguments to explain why independent noise sources add up in the way they do. First of all, the mathematical argument given above can be extended by evaluating the probability distribution \( P(N_1 + N_2) \) as a convolution of the individual distributions \( P(N_1) \) and \( P(N_2) \). If these individual distributions have Gaussian shapes the combined distribution is easily shown to be Gaussian as well, and to have a variance that obeys Eq. (2.4). Also for more general shapes does the combined distribution obey the addition law of Eq. (2.4). A second (and very simple) argument why noise amplitudes do not add up in a straightforward way is that the noise \( N_i(t) \) is random and can be either positive or negative; one really needs Eq. (2.4) for a correct additional of the noise powers! As a third argument for Eq. (2.4) one could say that noise is a fluctuating quantity that behaves as if it were composed of many sinusoidal component of different frequencies with random amplitudes and phases that obey certain statistical properties. If you add two noise sources, it is like adding lots of sinusoidal components with different amplitudes and phases (and possibly even different frequencies). For the addition of such “random sine functions” it is again easily shown that the variance of the combined noise is equal to the sum of the variances of the individual noise sources.

### 2.2 Noise strength behind spectral filters: \( S_N(f) \)

The variance \( \text{var}(N) = N_{\text{rms}}^2 \) quantifies the noise strength (at any moment in time), but doesn’t contain any information yet on its time-dependence or frequency contents. This information is contained in the autocorrelation function of the noise

\[ R_{NN}(\tau) = \langle N(t)N(t + \tau) \rangle , \]  

(2.5)

which correlates the noise amplitude at time \( t \) with the noise amplitude at some earlier or later time \( t + \tau \). For stationary noise a simple substitution
$t' = t + \tau$ immediately shows that the autocorrelation function is symmetric in the time delay $\tau$, as $R_{NN}(-\tau) = R_{NN}(\tau)$.

Although the autocorrelation function $R_{NN}(\tau)$ already gives a complete description of the time-dependence of the noise, it is often more convenient to work in the frequency domain instead of the time domain. As a simple introduction to the frequency analysis of noise, we will first consider the situation depicted in Fig. 2.4, where noise $N(t)$ is passed through a narrow-band filter that is centered around frequency $f_0$ and has a spectral bandwidth $\Delta f$.

The key point to realize is that the variance of the noise signal behind the filter $\langle y(t)^2 \rangle = \langle N_{\Delta f}(t)^2 \rangle$ is proportional to the filter bandwidth $\Delta f (= B)$, at least for small filter bandwidths, making the r.m.s noise amplitude proportional to the square root of the filter bandwidth. This scaling is a natural consequence of the “quadratic addition” mentioned earlier in combination with the fact that noise sources from different frequency bands act independently; for non-overlapping frequency bands the product $N_1(t)N_2(t)$ will oscillate at the approximate difference frequency $f_1 - f_2$ and average out to zero.

Figure 2.4: The amount of voltage noise around a frequency $f_0$ can be determined by passing this noise through a filter that is centered around $f_0$ and has a bandwidth $\Delta f \ll f_0$. The r.m.s. voltage noise behind the filter is proportional to $\sqrt{\Delta f}$.

With the scaling mentioned above, it is convenient to define the noise spectral density $S_N(f_0)$ as the variance of the noise per unit bandwidth around a center frequency $f_0$. In equation form this looks like

$$S_N(f_0) \equiv \lim_{\Delta f \to 0} \frac{\langle N_{\Delta f}(t)^2 \rangle}{\Delta f} \quad [\text{V}^2/\text{Hz}]$$

(2.6)

Note that the dimension of the noise spectral density is $\text{V}^2/\text{Hz}$ if the noise
amplitude given in Volts. The noise spectral amplitude $\sqrt{S_N(f_0)}$ is thus given in units of $V/\sqrt{\text{Hz}}$. Although these units might look strange, the $\sqrt{\text{Hz}}$ symbol is directly related to the scaling; it just shows that the rms voltage fluctuation observed behind a filter of bandwidth $\Delta f$ centered around frequency $f_0$ is given by

$$N_{\text{rms}} = \sqrt{S_N(f_0) \Delta f} \ [V] \quad (2.7)$$

Next we will consider the situation where the noise $N(t)$ is routed through a linear filter that is more general than the narrow pass-band filter discussed above. If we write the amplitude transfer function of the filter as $H(f)$, the variance of the noise $y(t)$ as measured behind this more general filter is given by

$$\langle y(t)^2 \rangle = \int_0^\infty |H(f)|^2 S_N(f) \, df \ . \quad (2.8)$$

In the absence of any spectral filtering, i.e. for $H(f) \equiv 1$, we recover

$$\langle N(t)^2 \rangle = \int_0^\infty S_N(f) \, df \ . \quad (2.9)$$

Note that the integral is limited to positive frequencies only; this choice was already made in the definition of $S_N(f)$ and should be followed all the way through to avoid double counting. This integral expression again shows that the noise spectral density $S_N(f)$ is nothing more than a functional representation of the division of the noise variance over frequencies; this is depicted in Fig. 2.5. In systems with multiple independent noise source, where the noise amplitude can be written as $N(t) = N_1(t) + N_2(t) + \ldots$, the total noise spectral density $S_N(f)$ is obviously just the sum of the individual noise spectral densities $S_{N_i}(f)$, each having its own frequency dependence.

Examples of four different noise source, with equal noise strength but different frequency contents are shown in Fig. 2.6. Note how peak-peak fluctuations are easily four times larger than the rms noise amplitude, which is 75 mV in all cases (vertical scale is 100 mV per unit). Also note how Fig. 2.6 a and b look quite similar after a re-scaling of the time axis by a factor 5.

### 2.3 Fourier analysis of noise: Wiener-Khintchine

The clever reader must have guessed already that the time domain description of noise via the autocorrelation function $R_{NN}(\tau)$ and the frequency domain description via the noise spectral density $S_N(f)$ are related via a Fourier
2.3. FOURIER ANALYSIS OF NOISE: WIENER-KHINTCHINE

Figure 2.5: Noise sources from different frequency bands are “independent” and add in a power-like fashion, making the overall noise power (or mean-square noise amplitude) equal to the frequency integral of noise spectral density $S_N(f)$.
Figure 2.6: Examples of four different noise spectra: (a) low-frequency noise from 0-1000 Hz, (b) low-frequency noise from 0-5000 Hz, (c) broadband noise between 2500-5000 Hz, (d) narrow-band noise between 4700-5300 Hz. In the oscilloscope traces each horizontal unit corresponds to 1 ms. The noise strengths, or mean-square noise amplitudes, are approximately equal in all four situations.
2.3. FOURIER ANALYSIS OF NOISE: WIENER-KHINTCHINE

The specific form of this transformation is given by the so-called Wiener-Khintchine theorem as

\[
R_{NN}(\tau) = \int_0^\infty S_N(f) \cos(2\pi f \tau) \, df , \quad (2.10)
\]

\[
S_N(f) = 4 \int_0^\infty R_{NN}(\tau) \cos(2\pi f \tau) \, d\tau . \quad (2.11)
\]

Note that the frequency and time integration ranges are over positive values only. The factor of 4 in the lower expression is an immediate consequence of this conventional choice. It is very important to remember that the noise spectral density \( S_N(f) \) is generally defined for positive frequencies only. As an aside we note that the mentioned factor of 4 could have been removed completely by redefining the noise spectral density \( S'_N(f) = \frac{1}{4} S_N(f) \) as a symmetric function in \( f \) and taking the Fourier integrations in Eq. (2.10) and Eq. (2.11) over both positive and negative times and frequencies. Mathematically, this seems to be the better choice, but physically the earlier definition (positive frequencies only) makes more sense; we will therefore stick to this definition.

One might wonder why the Fourier relation between the time and frequency domain description of noise goes via the Wiener-Khintchine theorem and not via ordinary Fourier relations. The reason for this is that ordinary Fourier relations are formulated specifically for the set of mathematically “well-behaved” functions that are “quadratically integrable” and thus belong to \( L^2 \). In contrast, the signal and noise function \( S(t) \) and \( N(t) \) that we study have no natural start or ending and are therefore not inside \( L^2 \). It is exactly for this reason that our Fourier relations are based on the autocorrelation function \( R_{NN}(t) \), which does belong to \( L^2 \) even for random noise. For voltage noise, the autocorrelation function \( R_{NN}(\tau) \) has dimensions \( V^2 \), while its Fourier transform, the noise spectral density \( S_N(f) \), has dimensions \( V^2/Hz \).

Above we stated that the Fourier analysis of noise, or any signal outside the group \( L^2 \), should formally proceed via the intermediate step of the autocorrelation function, which generally does belong to \( L^2 \). There is, however, also a more direct, albeit less rigorous, approach. This approach is based on the direct Fourier transform of the noise \( N(t) \), but now truncated over a finite time interval \([-T, T]\) and power-normalized via

\[
F_{N,T}(f) = \frac{1}{\sqrt{2T}} \int_{-T}^{T} N(t) \exp(j2\pi ft) \, dt . \quad (2.12)
\]

As the noise \( N(t) \) will be different from run to run, its Fourier transform \( F_{N,T}(f) \) will be just as “noisy”; only the statistical properties of this Fourier
transform have any significance. This “noisy” character is also the reason why the normalization goes via \( \sqrt{2T} \) instead of \( 2T \); it is based on the addition of noise powers over consecutive time intervals.

In the limit \( T \to \infty \) one can show that the expectation value of the absolute square Fourier amplitude, i.e. of \( |F_{N,T}(f)|^2 \), becomes equal to the double-sided noise power spectrum \( S'_N(f) \). The rough “proof” of this statement is based on a rewrite of the expression

\[
|F_{N,T}(f)|^2 = \frac{1}{2T} \int_{-T}^{T} \int_{-T}^{T} N(t_1)N(t_2) \exp(j2\pi f(t_1 - t_2)) dt_1 dt_2 . \tag{2.13}
\]

By introducing the average time \( \bar{t} \equiv (t_1 + t_2)/2 \) and the time difference \( \tau \equiv (t_1 - t_2) \) one can rewrite this expression as

\[
|F_{N,T}(f)|^2 = \frac{1}{2T} \int_{-T}^{T} \int_{-T}^{T} N(\bar{t} + \frac{1}{2}\tau)N(\bar{t} - \frac{1}{2}\tau) \exp(j2\pi f\tau) d\bar{t} d\tau . \tag{2.14}
\]

If we neglect the subtle differences in integration area, which will disappear anyhow for \( T \to \infty \), we recognize the autocorrelation function \( R_{NN}(\tau) \) as

\[
R_{NN}(\tau) = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} N(\bar{t} + \frac{1}{2}\tau)N(\bar{t} - \frac{1}{2}\tau) d\bar{t} . \tag{2.15}
\]

The step to Eq. (2.11) is now quickly made and yields the relation \( |F_{N,T}(f)|^2 = \frac{1}{2}S_N(f) = S'_N(f) \).

The above “proof”, with it’s somewhat tricky limit \( T \to \infty \), is quite rough. In fact, there is still a subtle difference between \( |F_{N,T}(f)|^2 \) and \( S_N(f) \). On first sight the amount of noise in these two functions is completely different. The function \( |F_{N,T}(f)|^2 \) is very noisy, even in the limit \( T \to \infty \) as each Fourier component is generated from a normalized integral over noise sources of the form \( N(t) \exp(j2\pi ft) \). Each Fourier component is in fact randomly distributed according to a complex Gaussian distribution so that the \( |F_{N,T}(f)|^2 \) values are randomly distributed according to a single-side exponential probability distribution

\[
P(|F|^2) \propto \exp(-|F|^2/C) .
\]

The function \( S'_N(f) \) on the other hand is generally very smooth as it is just the Fourier transform of the autocorrelation function \( R_{NN}(\tau) \), which is already averaged over time. The answer to this paradox is that \( |F_{N,T}(f)|^2 \) can obtain a similar smooth appearance by averaging over a finite frequency interval \( \Delta f \). As the noise in \( |F_{N,T}(f)|^2 \) changes on a frequency scale \( \approx 1/T \) such a frequency integration corresponds to averaging over some \( \Delta f \times T \) independent variables and the result becomes smooth in the limit \( T \to \infty \). This solves the paradox.
Chapter 3

Different types of noise

3.1 Noise with different spectral properties

Noise is best characterized on the basis of its time and frequency dynamics. Figure 3.1 shows how different types of noise can be distinguished on the basis of their noise spectral density $S_N(f)$. In this syllabus we will distinguish the following most common four types of noise (although other divisions are also possible):

- Spectrally white noise (or “pink noise” in a more realistic description)
- Harmonics perturbations
- $1/f$ noise
- drift

In this chapter we will discuss the origin of each of these noise sources, with a strong emphasis on the most important one, being the spectrally white noise.

Spectrally white noise is per definition noise with a noise spectral density that is independent of frequency. This is of course an idealized description, as frequency integration over a constant $S_N(f)$ would result in an infinite variance $\langle N^2(t) \rangle$. In practical systems noise is thus never truly white, but rather “spectrally pink”, which means that the noise spectral density is relatively constant up to a certain cutoff frequency, but decreases beyond this cutoff frequency, to keep the variance finite. In practical systems the cutoff frequency is often large enough, and the noise spectral density below cutoff is often constant enough, to validate the white noise model. Spectrally white noise is the most fundamental and physical noise source in the list given
CHAPTER 3. DIFFERENT TYPES OF NOISE

Figure 3.1: Different forms of noise can be distinguished on the basis of their noise spectral density (see text).

above. Whereas most other noise sources can (in principle) be removed by clever design, the spectrally white noise often poses a fundamental limit as it has deep physical roots. Spectrally white noise comes in two flavors: thermal noise and shot noise. These will be discussed extensively in Section 3.2 and Section 3.3.

Harmonic perturbations are not really random noise, but rather harmonic fluctuations originating from (nearby) coherent sources. These perturbations, which are sometimes also referred to as pick-up, have well-defined frequencies and can in principle be avoided by proper design. Tricks to use in the experiment are: shielding, proper grounding, and/or removal by balancing (to make the system less sensitive to pick-up). As harmonic perturbations have well defined frequencies, they give rise to persisting (= non-decaying) oscillations in $R_{NN}(t)$ and delta-functions in the power spectral density. This singular behavior makes these perturbations quite different from ordinary (random) noise.

$1/f$-noise is the name for a special type of low-frequency noise, having a noise spectral density that scales with the inverse frequency as $S_N(f) \propto 1/f$. It is often present in semiconductor components and is generally attributed to so-called “deep traps”, which trap charged carriers for an extended period of time, although the precise noise-generating mechanism is not always known. As a consequence of these deep and slow traps, the macroscopic de-
vice properties will change slowly in time and a strong low-frequency noise component will result. The strength of the 1/f noise depends strongly on the production technique and can even differ from device to device. It is obviously only important at relatively low frequencies. For typical semiconductor devices the turn-over point, where the strengths of the 1/f noise and white noise are equal, is somewhere in the wide range between 1 Hz and 100 kHz.

Previously, we have assumed the noise to be offset-free, with $\langle N(t) \rangle = 0$. This is not always the case. In some systems the offset might even change gradually with time. When this change occurs in a steady linear way we call it drift. Although the statement is not rigorously valid, it does make some sense to call drift very low-frequency noise, as a Fourier transform of a sloping line corresponds to something like (the derivative of) a delta function around $f = 0$. It is best to reserve the word “drift” for linear variations in the offset; if the offset does not vary linearly in time, but instead fluctuates somewhat within the timescale of the experiment, it should not be called drift, as a Fourier transformation will also yield frequencies different from $f = 0$.

3.2 Thermal white noise

The most important and fundamental form of noise is spectrally white noise of thermal origin. Thermal (white) noise, sometimes also denoted as Johnson noise, is a natural consequence of the so-called fluctuation-dissipation theorem. This theorem states that any lossy element will spontaneously generate fluctuations with a strength proportional to the dissipation that this element would produce under an external driving field. The physical origin for both fluctuations and dissipation is the coupling between the element and its environment; a strong coupling leads to both strong dissipation and strong fluctuations. Furthermore, at frequencies $f \ll kT/h$, where $k$ and $h$ are Boltzmann’s and Planck’s constant respectively, the spectral power of the thermal fluctuations is constant at a value that is directly proportional to the temperature $T$; hence the name “thermal (white) noise”.

An example of thermal noise are the voltage fluctuations over an electrical resistor with resistance $R$. Under the influence of an external voltage $V$ this resistor will dissipated a power $P = V^2/R$ due to collisions of the charge carrying elements with the atomic lattice, which turn electrical energy into heat. In the absence of an external voltage the same electron-atom collisions will, however, “shake” the individual electrons around and produce a fluctuating voltage over the leads of the resistor. The variance of this thermal noise voltage over a resistance $R$ is
\[ \langle \Delta V^2 \rangle = 4kTR\Delta f \, , \]  

(3.1)

making the noise spectral power \( S_{\Delta V}(f) = 4kTR \). At room temperature the thermal fluctuations over a 1 kΩ resistor amounts to \( V_{\text{rms}} = 0.9 \text{ nV}/\sqrt{\text{Hz}} \), making the integrated noise in a 1 MHz bandwidth 0.9 µV. Please try to remember these simple numbers: 1 kΩ, 1 MHz, thermal noise slightly less than 1 µV.

The fluctuation-dissipation theorem in general, and the above expression for \( \langle \Delta V^2 \rangle \) in specific, are intimately related to the equipartition principle of statistical physics, which requires that at thermal equilibrium each degree of freedom contains on average \( kT \) of energy (equally distributed over a potential and kinetic part). By considering the voltage fluctuations in a simple electronic circuit, consisting of a resistor \( R \) connected to a capacitor \( C \), it is relatively easy to derive Eq. (3.1). The argument is based on Fig. 3.2 and goes as follows: (i) Replace the noisy resistor by an ideal (= noiseless) resistor in parallel with a current source that produces spectrally white current noise with a mean-square amplitude \( i_n^2 = B\Delta f \), where the constant \( B \) is not yet known. (ii) Calculate the frequency-integrated voltage fluctuations over the RC combination, by letting this current noise pass in parallel through both resistor and capacitor, where the resistor takes care of the low-frequencies, while the capacitor short-circuits the high-frequency components. The frequency-dependent voltage fluctuations over the circuit are equal to \( V_n^2 = |Z(f)|^2i_n^2 \), where the equivalent impedance has the well-known form \( Z = R/(1+j2\pi fRC) \). Integration of \( |Z(f)|^2 \) over all frequencies shows that the equivalent noise bandwidth of the RC-circuit, being defined as the width of an equivalent flat-top filter with the same area and peak transmission, is \( 1/(4RC) \). The frequency-integrated voltage fluctuations are thus \( \langle \Delta V^2 \rangle = i_n^2 \times R/(4C) \). For completeness we note that this noise bandwidth \( 1/(4RC) \) is larger than the 3 dB bandwidth \( 1/(2\pi RC) \). (iii) The equipartition principle states that the “average potential energy” of the capacitor should equal \( \frac{1}{2}C \langle \Delta V^2 \rangle = \frac{1}{2}kT \). The frequency-integrated voltage fluctuations should therefore also be equal to \( \langle \Delta V^2 \rangle = kT/C \) (iv) Finally, we combine step (ii) and (iii) to find that the strength of the noise current is \( i_n^2 = B\Delta f = (4kT/R)\Delta f \). This makes the thermal voltage noise over the resistor (without capacitor) equal to \( V_n^2 = 4kTR\Delta f \).

For completeness we note that the equipartition principle is an argument from classical statistical mechanics, which needs some adaption to be valid in the quantum domain. Quantum effects are to be expected only at very high frequencies, where a single quant can contain energies up to and above \( kT \) (\( hf \geq kT \) corresponds to \( gg \) 100 GHz even at \( T = 4 \text{ K} \)). The transition from
3.3. SHOT NOISE

classical to quantum noise is well known from Planck’s expression for black-body radiation, which is nothing more than an expression for the thermal (noise!) power emitted by the object. Planck showed that the average (noise) power \( E \) per degree of freedom is not \( kT \), but

\[
\langle E \rangle = \frac{hf}{\exp(hf/(kT)) - 1}, \tag{3.2}
\]

which expands to the familiar \( kT \) only for \( hf \ll kT \).

\[\text{Figure 3.2: By considering the current and voltage fluctuations in a simple RC circuit, and replacing the experimental (and noisy) resistor by an ideal noiseless resistor in parallel with a current noise source } i_n, \text{ we can easily derive an expression for the thermal (spectrally white) current and voltage noise of the resistor.}\]

3.3 Shot noise

Shot noise arises when discrete elements (=quanta) pass a certain barrier in an independent way. The archetype of shot noise is the water flow produced by rain hitting a roof (see Fig. 3.3). As the rain consists of rain drops that fall more or less independently the water flow will vary erratically. If the water droplets, or even better hail stones, make some sound when they hit the roof and when the rhythm is fast enough to smear out the individual hits you can even hear the shot noise! Other examples of shot noise are (i) the variations in the electrical current produced by electrons emitted from a cathode, and (ii) the variations in the optical flux c.q. light intensity produced by photons hitting a photosensitive element.

Below we will show that the fluctuations produced by shot noise are evenly spread over all frequencies, at least up to a certain cut-off frequency that corresponds to the inverse time duration of the individual events; in other words shot noise is spectrally white, just as thermal noise is spectrally white. More specifically, the spectral density of the shot noise in an electrical current of average strength \( i_0 \) is

\[
S_i(f) = 2qi_0 \left[ A^2/\text{Hz} \right], \tag{3.3}
\]
CHAPTER 3. DIFFERENT TYPES OF NOISE

Figure 3.3: A typical example of shot noise is the noise in the water current (or water flow) of rain. If the rain droplets fall down in an uncorrelated way the noise spectral density is constant up to frequencies that correspond to the inverse of the average “hit time”.

Figure 3.4: Shot noise is the result of the random and uncorrelated arrival of the individual quanta (shots) that add up to the total (average) flux \( \langle dN/dt \rangle \) (see upper time trace). After time integration over a time window \( T \) the relative fluctuations in the time-integrated flux are approximately \( \sqrt{N}/\bar{N} = 1/\sqrt{\bar{N}} \), where \( \bar{N} = \langle dN/dt \rangle T \) is the average number of quanta within the time window (see lower time-trace).
where \( q \) is the elementary charge, sometimes also denoted by \( e \). As a result of this shot noise the variance of the current fluctuations \( \Delta i \) observed behind a filter of bandwidth \( B \) is equal to \( \text{var}(\Delta i) = 2qi_0B \). As the variance is proportional to the average current, the absolute rms (= root-mean-square) current fluctuations increase only as \( \sqrt{\text{i}_0} \) and the relative current fluctuations \( (\Delta i/\text{i}_0) \) decrease with current as \( 1/\sqrt{\text{i}_0} \). This is the natural scaling law for shot noise: large electron fluxes consist of many quanta per unit time and are thus relatively more stable than small fluxes that contain less quanta.

One possible derivation of Eq. (3.3) is based on a counting of the number of quanta within a *square time window* of width \( T \) (see Fig. 3.4). Within this time window the average number of detected quanta is \( \bar{N} = \langle dN/dt \rangle T = (i_0/q)T \) with a corresponding rms shot noise of \( \sqrt{\bar{N}} \). This makes the absolute current shot noise equal to \( \text{var}(\Delta i) = \bar{N}(q/T)^2 = i_0q/T \). As a final ingredient we perform a Fourier transformation of the square time window and note that the equivalent noise bandwidth of the transformed function \( [\sin (\pi Tf)/f]^2 \) is \( \Delta f = 1/(2T) = B \), making \( \text{var}(\Delta i) = 2i_0qB \).

An alternative derivation of Eq. (3.3) considers the variations in the electron flux as measured behind a (more realistic) *first-order low-pass frequency filter*, instead of the square time window discussed above. Such a filter transforms the ideal \( i(t) = q\delta(t-t_0) \) responses of the individual quanta into \( i(t) = q/(RC) \exp \left( -(t-t_0)/(RC) \right) \Theta(t-t_0) \) responses, where the time-integrated “area” \( \int i(t)dt = q \) is of course the same in both cases. To generate an average current \( \text{i}_0 \) an average number of \( \bar{N} = i_0/q \) quanta have to pass per unit time. As these quanta are assumed to be uncorrelated the second moment of the current is

\[
\langle i(t)^2 \rangle = \left\langle \left( \sum_i \frac{q}{RC} \Theta(t-t_i) e^{-t_i/t_0} \right)^2 \right\rangle \\
= \left\langle \sum_i \frac{q^2}{(RC)^2} \Theta(t-t_i) e^{-2(t-t_i)/t_0} + i_0 \sum_i \frac{q}{RC} \Theta(t-t_i) e^{-t_i/t_0} \right\rangle \\
= \frac{q^2}{(RC)^2} \frac{RC}{2} \bar{N} + i_0 q \bar{N} = \frac{q^2 i_0}{2RC} + i_0^2. \tag{3.4}
\]

The first term originates in a way from “the overlap of the \( i \)-th current pulse with itself”, while the second term results from the overlap of the \( i \)-th current pulse with all the other uncorrelated pulses. Subtraction of \( i_0^2 \) from the above result, in combination with the result from the previous section that the equivalent noise bandwidth of an RC filter is \( B = 1/(4RC) \), leads to the result of Eq. (3.3), being \( \text{var}(\Delta i) = 2qi_0B \).
As a final remark on shot noise, we note that the crucial assumption for its appearance is that the individual quanta arrive in an uncorrelated way. Whenever correlations occur, the noise can (in principle) be lower than the shot noise limit. This is for instance the case for the current fluctuations over a resistor, which contains very many charged carriers that feel each other all the time; resistors indeed don’t exhibit shot noise, although they are of course plagued by thermal noise. Other cases of altered shot noise are (i) the fluctuation in cathode current in the presence of space charge effects, which can suppress current fluctuations [6], and (ii) the intensity fluctuations in some laser that operate very far above the lasing threshold [8].

At first sight it might seem strange that lasers can potentially emit sub-shot noise light as the photons themselves do not really interact. Still it has been demonstrated [8] that the emission process in semiconductor laser provides for some back action and results in photon correlations and sub-shotnoise emission; the emitted light is said to have squeezed amplitude fluctuations and possess sub-Poissonian statistics. Some other “non-classical” light sources can provide such correlations as well, with the optical parametric oscillator, which uses nonlinear optics to convert single photons into photon pairs, as the most potent example. Some physicist have even made it a sport to go as much as possible below the shot-noise limit. The record is about 6 dB, i.e. a factor 4 in noise power. However, these low noise levels degrade easily; any type of optical loss will bring the light closer to the shot-noise level as it corresponds to a random destruction of photons which will obviously randomize the particle flow.

3.4 1/f noise

All systems contain some form of white noise, but many practical systems are also plagued by additional low-frequency noise components. When this low-frequency noise has the common form in which the noise spectral density increases as the inverse frequency one speaks about 1/f noise (see Fig. 3.5). The relative importance of 1/f noise is best specified via the transition frequency $f_k$: white noise dominates for frequencies above $f_k$, 1/f noise dominates below $f_k$.

The presence of 1/f noise can be quite bothersome in the experimental analysis, as will be discussed in more detail in Chapter 4. For now it suffices to point at Fig. 3.6 and state that the dominance of low-frequency components makes time integration practically useless for 1/f noise.

The physical origin of 1/f noise in the electronic conduction mechanism in semiconductor devices generally lies in impurities in the semiconductor...
3.4. 1/F NOISE

Figure 3.5: Noise spectral density in the presence of both white noise and 1/f noise. In this plot the transition frequency \( f_k \approx 300 \) Hz.

Figure 3.6: Time traces of (a) spectrally white noise, and (b) 1/f noise. Note the much stronger presence of the low-frequency components in the righthand curve. The dominance of these low-frequency components makes time integration practically useless for 1/f noise.
material and imperfection in the production process. As these imperfections and impurities are more likely to occur at the surface than in the bulk, 1/f noise is generally more prominent in devices that are small and contain lots of surface (like MOSFETs). As technology improves the amount of 1/f noise is expected to steadily decrease up to a point were it might become practically irrelevant.

### 3.5 Transfer of signal and noise

The question whether thermal noise can be used to generate energy is addressed in Fig. 3.7, which depicts how the thermal noise that is generated in resistor $R_1$ is partially dissipated in resistor $R_2$ and vice versa. To calculate the amount of power that is transferred from left to right we replace the noisy resistor $R_1$ by a noiseless one in series with a voltage noise source. With a calculation of the loop current as intermediate step we then find that the transfer of noise power from left to right is

$$P_{\text{left} \rightarrow \text{right}} = kT \frac{4R_1R_2}{(R_1 + R_2)^2}.$$  \hfill (3.5)

This transfer reaches a maximum for matched resistors ($R_1 = R_2$) at a value of $kT$ per Hz bandwidth; it decreases below this value when $R_1 \neq R_2$. For the energy transfer from right to left we of course find a similar expression, now containing $T_2$ instead of $T_1$.

The fact that the maximum “available noise power” is $kT$ per unit bandwidth is (again) rooted in the statistical mechanics: equipartition requires the sum of “kinetic and potential energy” to be $kT$ per degree of freedom, where the degrees of freedom are now the electro-magnetic waves (or voltage and current fluctuations) travelling between the two resistors. Note that the net energy transfer between the two resistors is proportional to their temperature difference $T_1 - T_2$. Even the available noise power can not be used to violate the second-law of thermodynamics: it always flows from the hot region to the cold region.

Now that we have discussed the transfer of noise power from one system to another, it is equally interesting to consider the transfer of signal power. It doesn’t take much imagination (and mathematical computation) to show that the transfer of signal power obeys similar rules as found for the transfer of noise power. In general the transfer of signal power is optimum when the input impedance at the receiving end, which is often a (pre-)amplifier, matches the output impedance of the signal source. As this receiving end is likely to produce it’s own noise, it can be quite important to aim for
impedance matching, as this will generally give the lowest degradation of the signal-to-noise ratio S/N.

In many practical systems the receiving end is a matched pre-amplifier that boosts both signal and noise up to a level where the noise generated in possible other elements in the detection chain becomes relatively unimportant. That’s why the first (pre-) amplifier should be “the best one” in the chain, adding the least noise to the signal.

One parameter that is often used to quantify the noise performance of an amplifier is its effective noise temperature \( T_{\text{eff}} \), which is defined as

\[
T_{\text{eff}} = \frac{P_{\text{amplifier}}}{k},
\]

where \( P_{\text{amplifier}} \) is the extra noise power added by the amplifier (and calculated back to the amplifier input). In a similar way we can define the noise temperature of the signal source as \( T_{\text{source}} = \frac{P_{\text{source}}}{k} \). The amplifier hardly adds noise and behaves practically ideal when \( T_{\text{eff}} \ll T_{\text{source}} \).

Another parameter that specifies the quality of the amplifier is the \( F \) number. This number specifies the amplifier quality in terms of the noise it adds on top of the original (unamplified) noise. It is defined as

\[
F \equiv \frac{(S/N)_{\text{out}}}{(S/N)_{\text{in}}},
\]

where the \((S/N)\) ratios in the numerator and denominator are power ratios at the output and input of the amplifier, respectively. The advantage of this definition is that the \( F \) number directly specifies the amplifier noise in relation to the input noise. This is also a disadvantage, as \( F \) does not only depend on the quality of the amplifier, but also on the actual amount of input noise.
3.6 Some quantitative examples

As a first example we consider the noise in a quick/fast voltage measurement; a series of such quick measurements comprise the time sampling of a certain signal $V(t)$ and form the basic input for the display of a digital oscilloscope.

**Question:** what is the accuracy of a quick voltage measurements if we are limited by thermal noise only, and if we measure within a time span of 10 ns over a 1 MΩ resistor with an ideal noiseless oscilloscope?

**Answer:** The conversion from a 10 ns wide square time window to an equivalent frequency bandwidth is our first concern. Fourier transformation of a square time window with width $T$ yields a power spectral of the form $(\sin (\pi T f)/f)^2$, having an equivalent frequency width of $\Delta f = 1/(2T)$ (just for remembrance: the first zero is at $f = 1/T$). Insertion of the calculated bandwidth of 50 MHz and the resistance of 1 MΩ into Eq. (3.1) immediately yields an rms noise voltage of 0.2 mV. This value shows that fast voltage measurements over large resistors are intrinsically noisy. As the rms noise scales with $\sqrt{\Delta f}$ or $1/\sqrt{T}$ a decrease of the time window leads to an increase in the sampling noise: with a 0.1 ns the rms voltage noise would already be 2 mV!

Of course the above values are rather large mainly because we took a large resistor. Within the same 0.1 ns time window the rms voltage noise over a 50 Ω resistor is only 14 μV. This example shows the importance impedance matching: if we measure on a low-impedance source the results will be much less noisy if we use a matched (= low-impedance) meter.

As a second example we consider the noise in optical detection with a (semiconductor) photodiode. Photodiodes measure the light intensity by converting the incident photons into charge carriers (electrons and holes) in some semiconductor material like silicon. When the incident photons arrive in an uncorrelated manner both the photon and electron flux suffer from shot noise. As practical photodiodes are never ideal, there is generally already some small leakage current even in the absence of light.  

**Question:**
(1) How large is the shot noise that is associated with a typical leakage current of $i_0 = 1$ nA?

(2) The photodiode is illuminated with light at a wavelength of 633 nm. At this wavelength the conversion from photon to electron is not ideal, but occurs with a so-called quantum efficiency of 60 %. What is the extra noise that is measured under 1 μW of illumination (with “shot-noise-limited” light)?

**Answer:**
(1) The shot-noise of $i_0 = 1$ nA has a spectral power of $S_i = 2q i_0 =$
3.6. SOME QUANTITATIVE EXAMPLES

$3.2 \times 10^{-28} A^2/\text{Hz}$; the rms noise current is $\sqrt{i_n^2} = 1.8 \times 10^{-14} A/\sqrt{\text{Hz}}$. With an energy of 1.96 eV per 633 nm photon and a quantum efficiency of 60% this corresponds to a so-called noise equivalent power (NEP) of $1.8 \times 10^{-14} \times 1.96/0.6 = 5.9 \times 10^{-14} \text{W}/\sqrt{\text{Hz}}$.

For completeness we note that it is very difficult to realize this low noise level, as any practical amplifier, which is certainly needed to amplify the sub-pico ampere currents to more detectable levels, will add an overwhelming amount of current noise. If the dark current is low enough, the noise level of experimental photodiodes is generally limited by the current noise of the amplifier circuit.

On top of this current noise, amplifiers also produce voltage noise that has to be accounted for. This voltage noise can be especially bothersome at high frequencies, where the complex impedance $Z$ of the photodiode drops rapidly due to its capacitive action. The total current noise over the Op-Amp circuit is given by the (incoherent) sum $I^2 = I_n^2 + V_n^2/|Z|^2$.

(2) Illumination with 1 $\mu$W at a wavelength of 633 nm and a quantum efficiency 60% will generate an average current of $(1/1.96)*0.6 = 0.306 \mu A$. If the light is shot-noise limited, i.e., if it consists of uncorrelated photons, the associated current shot noise is easily calculated to be $3.1 \times 10^{-13} A/\sqrt{\text{Hz}}$, being much larger than the shot noise mentioned under (1). Note that division of the rms current noise by the average current shows that even for an optical source as weak as 1 $\mu$W we can measured the relative intensity fluctuations with a (shot-noise-limited) accuracy as low as $\Delta P/P_0 = \Delta i/i_0 \approx 10^{-6}/\sqrt{\text{Hz}}$. 
CHAPTER 3. DIFFERENT TYPES OF NOISE
Chapter 4

How to improve S/N

4.1 Four techniques to improve S/N

It has been said before, but I want to say it again: “the quality of experimental data is not determined by the absolute signal strength, but rather by the attainable signal to noise ratio S/N”. It is therefore of utmost importance to know the enemy (noise) and find ways to defeat him. As stated in the book of Wilmshurst [2]: we sometimes need a form of signal recovery, to recover our signal out of a background of unwanted noise.

There are four basic tricks for signal recovery. In this section we will list these tricks and discuss them very shortly. We will then spend one section on each of these tricks and finish the chapter again with a couple of quantitative examples. The four techniques to improve the S/N ratio are:

1. Low-frequency filtering and/or visual averaging.
2. Correction for offset and drift.
3. Multiple time averaging (MTA).

The first item on the list “low-frequency filtering and/or visual averaging” indicates that an increase of the measurement time generally leads to an improvement of S/N. This trick works because signals are coherent, i.e., they have a stable amplitude (and a stable frequency and phase, if modulated), while noise is incoherent, i.e., noise has a fluctuating amplitude and phase so that it tends to “average out during time integration”.

The second item on the list “correction for offset and drift” is in a way the least important. This trick is almost as obvious as the first one.
If the signal to be measured is accompanied by an extra (unwanted) DC signal (= offset), which might even change gradually in time (= drift), it is important to find ways to correct for this offset and drift.

The third and fourth item on the list (“multiple time averaging” and “modulation techniques”) require more thinking. These tricks are both based on the very important notion that the amount of noise is generally not evenly distributed over all frequencies, so that the ratio S/N can potentially increase when the signal is shifted to frequencies of relatively low noise. These tricks are thus useless in systems that posses only spectrally white noise. However, as such systems are practically non-existent these two tricks are often needed to improve S/N. Both work specifically against relatively narrow-band noise sources like: 1/f noise, time-varying drift, and “pick-up” or other perturbing signals at well-defined frequencies.

In many systems low frequency noise, such as 1/f noise or time-varying offsets, are the limiting factor. 1/f noise is a serious troublemaker in signal recovery, because it is practically immune to simple time-averaging in the form of low-frequency filtering (item 1). An increase of the measurement time averages out some of the noise, but also makes the system more sensitive to noise at lower frequencies and it is exactly this type of noise that dominates in 1/f noise. In this case Multiple time averaging (MTA), which refers to a different way of averaging, can help. Instead of taking a single long measurement we now perform a series of short measurements that are averaged afterwards. As the individual measurements are short MTA is an effective trick to avoid extra low frequency noise and still perform sufficient averaging.

The use of “modulation techniques”, as our fourth trick to improve the signal-to-noise ratio, can be most easily explained from a spectral point of view. Through modulation we can shift the signal from DC ($f = 0$) to any frequency. In practice we will choose the modulation frequency $f$ such that is sufficiently high to avoid 1/f noise, and far away from any frequencies of possible external perturbations (pick-up). Ideally, the noise at $f$ is dominated by spectrally white noise, which is per definition the same at all frequencies. Even though modulation will reduce the signal strength (at some times there is even no signal at all!), it generally still pays off as the noise at the modulation frequency can really be orders of magnitude smaller than that at the low frequencies that one probes without modulation.
4.2 Time averaging and low-frequency filtering

Time averaging and low-frequency filtering are effective against the most fundamental forms of noise, being spectrally white noise either in the form of thermal noise or shot noise. I call these noise sources fundamental as they are practically impossible to remove; thermal noise can only be reduced by serious cooling; shot noise can only be reduced by removal of the criterium of “independent arrival of the individual quanta” that formed the basis in the derivation of shot noise, but such reduction is very difficult and mainly academic.

At this point it is convenient to introduce the two timescales in the problem. The first time scale is the system response time $T_{\text{res}}$, being the fastest time scale at which the signal output can vary “sizeably”. This response time can for instance be set by the speed of the detector and detection electronics. The system response time determines the maximum frequency $f_{\text{max}} = C/T_{\text{res}}$ present in the output, where the constant $C \approx 1$ although it’s precise value depends on the type of low-frequency filtering. The second important time scale is the integration time $T_{\text{av}}$ of the time averaging or (equivalently) the low-frequency filtering that is applied after detection.

The effect of time average on a signal in the presence of white noise can be explained in both time and frequency domain. In the time domain the explanation is based on the separation of the integration window into $T_{\text{av}}/T_{\text{res}}$ more or less independent time slots. In all these time slots the DC signal is the same, whereas each noise contribution is a more or less independent random variable that obeys Gaussian statistics and has zero mean. As a result, the summation or integration over these time slots will increase the signal-to-noise ratio by a factor $\sqrt{T_{\text{av}}/T_{\text{res}}}$ as compared to the S/N of the individual time slots. Averaging is obviously ineffective on time scales shorter than the system’s response, i.e., for $T_{\text{av}} < T_{\text{res}}$.

In the frequency domain the increase of S/N ratio is based on the reduction of the detection bandwidth. A long integration time removes the noise at frequency beyond $f = C/T_{\text{av}}$, but will not affect the DC signal strength at $f = 0$. As the rms noise is proportional to $\sqrt{f}$ we again find the scaling law $S/N \propto \sqrt{T_{\text{av}}}$.

When we just look at a noisy curve, such as Fig. 4.1(c), we already often perform a different kind of time averaging or low-frequency filtering, which we will denote by “visual averaging”. The eye and mind automatically look for average values within the noisy curves: if spectrally white noise dominates and if we can observe a large time span $T \gg T_{\text{res}}$ we can estimate this average...
CHAPTER 4. HOW TO IMPROVE S/N

Figure 4.1: The effect of time-averaging. Curve (a) shows the original signal; figure (b) shows the square time window that was used for the averaging; curve (c) shows the time-averaged curve. Time averaging makes the step-wise transition triangular in shape, but also leads to a strong noise reduction. An objective measure for this noise reduction is the change in the rms fluctuation $\Delta V_{\text{rms}}$. This reduction is quite sizeable, but still doesn’t look so impressive, because we automatically perform visual averaging on the upper curve; an effect that is strengthened by the two straight lines that represent the noise-free signal.
4.3. **CORRECTION FOR OFFSET AND DRIFT**

In many practical cases the signal to be measured resides on top of a constant offset or even on top of an offset that changes linearly in time and that we call drift. This is for instance the case in many response measurements, which try to quantify the response of one variable in a system due to a (sudden) change in another variable. To properly quantify the signal strength in such a measurement one has to separate it from a (possibly sloping) background.

The basic trick to correct for offset and drift is to measure not only during the application of the stimulus, but also before and after the stimulus is applied. These pre-scan and post-scan periods provide the necessary information on the offset and drift. In the most simple form we correct only for the offset, by subtraction of the pre-scan value from the measurement. In a more extensive form we correct for both offset and drift, by performing a linear interpolation between the values measured during the pre-scan and post-scan period and subtracting these interpolated values from the measured signal (see Fig. 4.2). The corrected signal thus becomes

\[
x_{\text{corrected}}(t) = x(t) - \frac{(T - t)x(0) + tx(T)}{T},
\]

where \( t = 0 \) and \( t = T \) correspond to the borders of the mentioned periods.

You have encountered a practical example of such a correction already during the first-year physics “practicum” in the “calorie meter (W4)” experiment, an experiment that also involved an extensive error analysis.

Next we will answer the question: “How long should we integrate in the pre-scan and post-scan period?”. On the one hand, it seems wise to integrate over at least \( T_b \gg T_{av} \) to ensure that the noise in these periods is less than that in the actual experiment. This is a good investment as there is only one pre-scan and one post-scan data point that will be subtracted (and add noise) to each of the many data points that make up the actual scan. On the other hand, it is useless to increase \( T_b \) too much and we should certainly keep \( T_b \ll T_{scan} \). Combining these two arguments, we find that it is best to choose \( T_b \) somewhat larger than \( T_{av} \) and obey \( T_{av} \ll T_b \ll T_{scan} \).

The effect of offset and baseline correction can also be explained in the frequency domain. For this we will interpret the “correction for offset” or the more extended “correction for offset and drift” as linear operations that transform an input signal \( x(t) \) into an output signal \( y(t) \). We will first con-
Figure 4.2: Correction for offset and drift can be performed based on information obtained before and after the actual measurement in a “pre-scan” and “post-scan” period. There are three relevant time scales: the integration time $T_b$ used in the “pre-scan” and “post-scan” period, the integration time $T_{av}$ used during the actual measurement, and the measurement or scan time $T_{scan}$, which is also the typical time separation between the three periods (see text).
consider the time- and frequency characteristics of a simple offset correction of the form \( y(t) = x(t + t') - x(t') \), with additional time integration or low-pass filtering over a time \( T_b = T_{av} \). The input-output operation of this correction in both time and frequency domain is

\[ y(t) = \frac{1}{T_b} \int_{-\frac{1}{2}T_b}^{\frac{1}{2}T_b} (x(t + t') - x(t')) \, dt' , \quad (4.2) \]

\[ \hat{y}(\omega) = (e^{-i\omega t} - 1) \frac{\sin(\frac{1}{2}\omega T_b)}{\frac{1}{2}\omega} \hat{x}(\omega) . \quad (4.3) \]

The two factors in the relation between \( \hat{y}(\omega) \) and \( \hat{x}(\omega) \) describe the action of the offset correction and the time-integration, respectively. The first factor approaches zero for frequencies \( \omega \ll 1/t \), thus showing how very low frequency noise (and offset) is fully removed by subtraction; the contribution of these low-frequency components to \( x(t) \) is simply not yet changed during such a relatively short time span. This is the frequency-domain explanation of the effect of offset correction. The second factor simply shows how high frequency components, with \( \omega \gg 1/T_b \), disappear due to the low-pass filtering that is associated with time-integration.

The trick used above to give a frequency-domain picture of the effect of offset correction (and time-integration) can also be used to demonstrate the extra advantage of a full drift correction of the form given by Eq. (4.2), as compared to the above simple offset correction, which basically used only \( x_{\text{corrected}}(t) = x(t) - x(0) \). After Fourier transformation the effect of these two different types of correction result in a multiplication in the frequency domain by the following two pre-factors

\[ e^{-i\omega t} - 1 \approx -i\omega t \quad (4.4) \]

\[ e^{-i\omega t} - [(1 - \frac{t}{T}) + \frac{t}{T}e^{-i\omega T}] \approx \frac{1}{2}\omega^2 t(T - t) . \quad (4.5) \]

This comparison shows that a simple offset correction works like a low-pass filter, transmitting low-frequency components only as \( \propto \omega \). The full offset and drift correction, which uses both a pre-scan and post-scan period, performs much better as it transmits low-frequency components only \( \propto \omega^2 \).

### 4.4 Multiple time averaging

In Section 4.2 we already mentioned that time integration and visual averaging are only effective if the noise is dominantly spectrally white and hardly
contains excess low-frequency noise. The reason why integration does not work in the presence of 1/f noise is best appreciated from Fig. 4.3, which shows the noise spectral density of 1/f noise. The two black areas labelled 1 and 2 show how much of this noise is picked up in the experiment. After offset correction the experiment is mainly sensitive to noise at frequencies above $1/T_{\text{scan}}$ and below $1/T_{\text{av}}$ (see Eq. (4.3) with $t \approx T_{\text{scan}}$). Integration of the 1/f shaped noise spectral density over this frequency range via

$$\int_{1/T_{\text{av}}}^{1/T_{\text{scan}}} \frac{C}{f} df = C \ln \left( \frac{T_{\text{scan}}}{T_{\text{av}}} \right)$$

(4.6)

shows that the total noise power does not contain any of these times explicitly, but only the ratio $T_{\text{scan}}/T_{\text{b}}$. In the presence of 1/f noise we thus reach the (possibly somewhat) surprising result that an increase of the integration time $T_{\text{b}}$ only results in an increase in the signal-to-noise ratio if we keep the measurement time $T_{\text{scan}}$ fixed, i.e., if we reduce the “useful number of (independent) data points”; it has no effect if we simultaneously increase $T_{\text{scan}}$ by the same factor.

The important observation that time integration or visual averaging provides only for an efficient suppression of white noise, but is not effective against 1/f noise is also visualized in Fig. 4.4. This figure shows the relative error in the average voltage as a function of the integration or scan time $T_{\text{sc}}$. At short integration times we are mainly sensitive to high frequencies; white noise dominates and the relative error scales as $1/\sqrt{T_{\text{sc}}}$. At intermediate times, where 1/f noise becomes important, this scaling is lost and we reach the situation discussed above where the relative noise is basically independent of integration time. A further increase of the integration time can even harm the attainable noise level as the measurement becomes more and more susceptible to slow variations in offset and drift.
4.4. MULTIPLE TIME AVERAGING

Figure 4.4: If the integration (or measurement) time is increased, the relative noise (or error in the average voltage) generally first decreases, in the regime where white noise is dominant, then stabilized, in the regime where 1/f noise dominates, and finally even increases due to slow variations in offset and drift (see text).

Figure 4.5: Multiple time averaging (MTA) comprises the averaging of a number of fast measurements ($v_1, v_2, v_3$ in the figure) taken quickly one after the other. MTA reduces the noise level and is resistant against low-frequency noise.
There is a trick around the integration dilemma sketched above. This trick is called *multiple time averaging* (MTA) and consists of averaging over a (possibly large) number of fast measurement runs that are taken quickly after each other (see Fig. 4.5). The secret of MTA lies in the speed at which the individual measurement runs are performed. If offset correction is applied to each run or (even easier) to the calculated average run, this correction will remove any (noise) frequencies \( f < 1/T_{\text{scan}} \). To sufficiently remove the influence of 1/f noise we should therefore perform the individual measurements fast enough, using a scan time \( T_{\text{scan}} \ll 1/f_k \), where \( f_k \) is the transition frequency below which 1/f noise dominates (see Fig. 4.6).

Figure 4.6: Graphical representation of the effect of multiple time averaging (MTA) on the observed noise. When we combine MTA with offset correction, we are sensitive only to noise frequencies between \( 1/(2\pi T_{\text{sc}}) \) and \( 1/(2\pi T_{\text{res}}) \), \( T_{\text{sc}} \) and \( T_{\text{res}} \) being the scan (or measurement) time and resolution (or integration) time respectively. In spectrum (b) the scan rate was sufficiently fast, in spectrum (c) the scans were too slow for MTA to work efficiently.
4.5 Modulation and phase-sensitive detection

Excess low-frequency noise is the main problem for accurate detection of static or slowly-varying signals. One of the most obvious things to try is to modulate the signal, i.e. make it go on-and-off, thereby shifting it spectrally to some modulation frequency \( f_0 \) where the noise spectral density is lower (see Fig. 4.7). Ideally, this modulation frequency should be large enough to avoid the aforementioned low-frequency noise and reach the limit set by the white noise; higher modulation frequencies are not needed. Frequency selective detection around frequency \( f_0 \) is generally performed by demodulation via multiplication with a sine or square-wave function followed by low-pass filtering. Modulation will obviously also lead to some reduction in signal power (the signal is generally on only half the time), but this reduction is often overshadowed by the noise reduction, which can be orders of magnitude.

![Figure 4.7: Frequency-domain explanation why phase-sensitive detection avoids 1/f noise. (a) combined noise spectrum, (b) signal spectrum, (c) required frequency response (centered around the modulation frequency \( f_0 \) with a full equivalent bandwidth of \( \approx 1/(2T) \), \( T \) being the RC-type integration time.]

The operation of a phase-sensitive detector or lock-in amplifier is depicted in Fig. 4.8 and Fig. 4.9 is as follows: (i) We start with some DC signal \( S \) that resides on top of noise \( N(t) \) as \( x(t) = S + N(t) \). (ii) Modulation of the
signal can be performed in many ways; for convenience we will first consider a sinusoidal modulation of the form

\[ y(t) = \left( \frac{1}{2} + \frac{1}{2} \sin(\omega_0 t) \right) S + N(t). \] (4.7)

(iii) This combined signal and noise \( y(t) \) is fed into one port of the phase-sensitive detector, were it is first multiplied by a (possibly phase-shifted) reference signal and then integrated over a time interval \( \gg 1/\omega_0 \) to

\[ \int dt \left( y(t) \sin(\omega_0 t + \phi) \right) = \int dt \left( (\frac{1}{4} \cos \phi) S + \sin(\omega_0 t + \phi)N(t) \right). \] (4.8)

The crucial reason why modulation and demodulation helps is that the time integration over \( \sin(\omega_0 t + \phi)N(t) \) is sensitive only to fluctuations in \( N(t) \) around the modulation frequency \( f_0 = \omega_0/(2\pi) \), which might be orders of magnitude lower than the noise at very low frequencies.

If the noise has the unusual property that it is truly spectrally white, even at very low frequencies, modulation is not effective and even leads to a deterioration of the S/N ratio, as the signal is present only part of the time, while the noise strength is generally unaffected by the modulation. A more quantitative argument to estimate the reduction in S/N due to modulation and demodulation is as follows: in the above case of sinusoidal modulation and demodulation, were the modulated signal and the reference were already in-phase, the time-integrated signal reaches a maximum of \( \frac{1}{4} S \) at the phase \( \phi = 0 \). In other situations, time delays or phase shifts could be important so that we might have to tune \( \phi \) for optimum signal. The time integrated value of the noise in Eq. (4.7) is also reduced as compared to it’s original non-modulated value \( \int dt N(t) \). The reduction of the rms amplitude is a factor \( \frac{1}{2} \), as spectrally white noise can equally well be written as \( N(t) \) or \( N'(t) \sin(2\pi f_0 t) + N'(t) \cos(2\pi f_0 t) \), where \( N \) and \( N' \) have the same rms amplitude, and as only the “in-phase-component” survives the time-integration of Eq. (4.7) in the form of a noise amplitude \( \frac{1}{2} N'(t) \). The final balance of our sinusoidal modulation and demodulation scheme is thus a reduction of the amplitude S/N ratio by a factor \( \sqrt[4]{\frac{1}{2}} \) and of the power S/N ratio by \( \frac{1}{\sqrt{16}} \).

The decrease of S/N due to modulation is smaller if we don’t use sinusoidal modulation, but instead use square-wave (on/off) modulation. For this type of modulation the signal amplitude at frequency \( f_0 \) increases by a factor \( 4/\pi \) as compared to sinusoidal modulation. This makes the cost of square-wave modulation only a factor \( 2/\pi \) in amplitude S/N and \( 4/\pi^2 \) in power S/N. The later result can also be understood in a more direct way, as the product of a factor \( \frac{1}{2} \) power loss due to the on/off duty cycle with a factor \( 8/\pi^2 \) for
4.5. MODULATION AND PHASE-SENSITIVE DETECTION

Figure 4.8: Schematic drawing of a phase-sensitive detector. Through modulation the signal is transformed from DC to the modulation frequency $f_0$. This signal is fed into a lock-in amplifier (right-hand side of dotted line) together with a reference that is modulated at the same frequency. After phase shifting, the signal and reference are multiplied and the result is passed through a low-pass filter. The final output is sensitive only to signal and noise components around frequency $f_0$.

Figure 4.9: Signals at various stages of the demodulation process that takes place in a phase-sensitive detector: (a) the original sinusoidally-modulated signal, (b) the square-wave reference signal, which has been aligned with the signal via its phase $\phi$, (c) the demodulated signal is the product of (a) $\times$ (b); the solid line shows the smoothed signal after moderate time integration.
46

CHAPTER 4. HOW TO IMPROVE S/N

the relative power in the fundamental \( f = f_0 \) frequency band. Square-wave modulation is thus somewhat better than sinusoidal modulation, and often easier as well.

In some systems one would like to simultaneously measure both the amplitude and phase of the modulated signal. This is easily done with a so-called vector lock-in amplifier, which just contains a double set of mixers and amplifiers and allows for a demodulation of the noise and the modulated signal in both the in-phase and out-of-phase quadrants, via integration over both \( y(t) \cos(\omega t) \) and \( y(t) \sin(\omega t) \). Such vector lock-ins generally have two displays, which either show the in-phase and out-of-phase amplitudes \( X \) and \( Y \), or the total amplitude \( R = \sqrt{X^2 + Y^2} \) and “phase” \( \phi = \arctan(Y/X) \). In a so-called digital lock-in the signal and reference are sampled at the input. After conversion to a stream of numbers all other operations, like mixing and time-integration or other types of low-pass filtering, are performed in a digital form. This make digital lock-ins more flexible than their analog counterparts.

4.6 Noise optimization in parameter fitting

Suppose you want to accurately measure the amplitude of a pulse, with known shape and time of arrival, i.e., suppose you want to know the amplitude \( A \) of a signal of the form \( x(t) = Af(t) \), with known \( f(t) \). The book of Wilmshurst [2] discusses this and related issues in quite some detail. Some of his arguments are depicted in Fig. 4.10 and summarized below. A practical example of the depicted situation is the measurement of the fluorescence intensity of a sample that is illuminated with a laser pulse of known duration and timing. This pulsed type of excitation and detection has, by the way, clear advantages over continuous operation: during the short on-period the fluorescence can be quite strong relative to the noise background, which in principle (after time gating) allows us to reach much higher signal-to-noise ratios than with continuous excitation.

In order to find the required pulse amplitude, one could imagine taking a “sample” around the maximum of the pulse over a time interval \( T_b \) that is much shorter that the pulse duration \( T_{\text{pulse}} \). This is not a good idea as integration over too short a time interval removes much of the potential signal and makes the measurement sensitive to noise of particularly high frequencies (up to \( \approx 1/T_b \)), making \( S/N \propto \sqrt{T_b} \). The opposite approach, being integration over a time interval that is much longer than the pulse duration, is not wise either; outside the pulse window there is no signal anymore but we still integrate over the noise, so that \( S/N \propto 1/\sqrt{T_b} \). This
Figure 4.10: How can we accurately determine the amplitude of a noisy pulse (as in b) if we already know its shape (as in d)? One option would be to integrate the noisy signal over a finite time interval that for example covers the region between the points of half maxima (as in c). Mathematically, one can show that the optimum approach in the presence of white noise only is to first multiply the noisy signal with the expected waveform and then perform a full time integration (see text).
line of reasoning indicates that the best strategy is to integrate the measured signal $x(t)$ over a time window that is comparable to the pulse duration. A more sophisticated analysis shows that it is even better to introduce a \textit{weight function} $w(t)$ that defines a smooth time window and weights the signal $x(t)$ in such a way that the “central components” around the pulse maximum get a larger weight in the integration than the components away from this maximum. This procedure gives the best (= less noisy) estimate of the pulse amplitude via

$$A = \frac{\int x(t)w(t)dt}{\int f(t)w(t)dt}.$$  \hspace{1cm} (4.9)

Wilmshurst [2] shows that one should take $w(t) \propto f(t)$ if the noise is dominantly spectrally-white; this specific weighting procedure is called “matched filtering”. In the presence of 1/f noise or other low-frequency perturbations, like offset or drift, it is better to remove the DC component from the weight function $w(t)$, by using for instance a weight function of the form $w(t) \propto df(t)/dt$.

Apart from these weighting procedures, there are other techniques to obtain low-noise estimates of the pulse amplitude. Possible techniques include (i) least-square fitting, and (ii) optimization of the cross-correlation between measured signal and theoretical prediction. The weighted integration and these new techniques can also be used to obtain low-noise estimates of the pulse durations and pulse positions. We’ll leave you with that statement and wouldn’t go into further details.

### 4.7 Some quantitative examples

Working in optics, the ultimate noise level to reach is generally the shot-noise limit. Most high-quality lasers and even thermal light sources can reach this limit, but only at sufficiently high frequency; at low frequency the optical noise is always dominated by some perturbations, which can for instance be thermal fluctuations or voltage fluctuations over laser or lamp. That the stability requirements are almost impossible to reach at low-frequency is obvious once one realizes that even an optical source of 1 $\mu$W emits more than $10^{12}$ photons per second and thus has a relative shot-noise level of only $10^{-6}/\sqrt{\text{Hz}}$. Even such a weak optical source thus already operates above the shot-noise level when the (relative) change in the driving voltage is more than $\approx 10^{-6}$ on a second timescale.

My PhD research involved some type of pump-probe experiment, where a short laser pulse excited (pumped) a material system after which another
laser pulse monitored (probed) the induced changes as a function of the time delay between the “pump” and “probe” pulse. In our experiment the pump pulse induced only very minor changes in the reflection of the probe laser. These were extremely difficult to measure, the more so because these changes were situated on top of a large base reflection and because the pump and probe beam had to be neatly separated. A simple baseline correction was out of the question as we were looking for relative changes in the order of $10^{-5} - 10^{-7}$. Modulation of the pump beam intensity in combination with phase-sensitive (lock-in) detection of the probe beam intensity seemed the best option. This intensity modulation is most easily performed by passing the beam through a rotating chopper wheel, producing a on/off type modulation with frequencies up to a few kHz. Unfortunately, the relative intensity noise at these frequencies was as much as $10^{-5} / \sqrt{\text{Hz}}$, which might not sound too bad but was still at least two orders of magnitude above the shot-noise level expected for a 1 mW beam. The experimental results obtainable with this type of modulation were not impressive.

A detailed study of the intensity noise in our beam showed that it originated from so-called plasma oscillation in the electrical discharge of the Ar-ion laser that was the heart of our laser system. As this plasma noise is intrinsic and unavoidable we had to work around it. We noticed that the spectral density of the plasma noise was almost constant up to frequencies around 200 kHz but dropped down rapidly beyond that frequency. As a side remark we note that this spectral behavior was consistent with the observed relative intensity noise of $10^{-5} / \sqrt{\text{Hz}}$ in the frequency domain and the 0.5 \% rms intensity fluctuations observed in the time domain. To avoid the plasma noise we increased our modulation frequency to 8.864 MHz (some funny number that was certainly not a multiple of any known pick-up frequency and was still much smaller than the repetition rate of our laser pulses). At this modulation frequency we could and did reach the shot-noise limit. To avoid spurious signals we had to use a double modulation scheme, which involved a 8.864 MHz modulation of the pump and a 200 Hz modulation of the probe, followed by a sequential demodulation at these two frequencies.

The effort certainly paid off; with an integration time as low as 0.1 s we could still observe change in the reflectivity of the order of $10^{-7}$ and the results were impressive. Even though the signal strength suffered seriously, the double modulation worked well as it reduced the noise by about two orders of magnitude (to practically the shot-noise limit) and removed several spurious signals.
Chapter 5

Analog-to-digital conversion and sampling

5.1 Hardware for Analog-to-Digital Conversion

Conversion of an analog signal $V(t)$ into a series of digital values involves discretization in both time (sampling) and voltage (analog-to-digital conversion). In Section 5.2 we will discuss the effect of the time discretization to show that no information is lost when the sampling time $T$ is sufficiently short. In this section we will discuss the practical implementation of the voltage discretization, in the form of analog-to-digital converters (ADCs). The attainable resolution in $V$ is eventually limited by the quality of the ADCs, with "bit noise" or "rounding-off errors" as the "final limit", although you can sometimes even go somewhat below this bit-noise limit by jittering, i.e., by artificially adding a small modulation on the signal.

The general scheme for analog-to-digital conversion is sketched in Fig. 5.1. This scheme of course involves an ADC, but generally also a low-pass filter and a so-called "sample-and-hold" unit. The function of the low-pass filter is to remove high-frequency noise from the analog signal; noise that would be easily picked up in a fast and non-integrating conversion process. The 3-dB frequency of the low-pass filter should be chosen somewhat above the Nyquist frequency $1/(2T)$ (see Section 5.2), but should not be orders of magnitude away. The sample-and-hold circuit does nothing more than rapidly taking a sample on its input channel and holding the result for an extended time at its output channel. This circuit thus ensures that the voltage at the ADC input does not change during the analog-to-digital conversion, which is particularly important for high-frequency signals. Technologically, it is very
Figure 5.1: The conversion from an analog voltage to a digital number in an analog-to-digital convertor (=ADC) can often be improved by the inclusion of a low-pass filter (to filter out any high-frequency noise) and a sample & hold unit (to fix the signal and allow for sufficient conversion time).

5.2 Consequences of sampling: bitnoise & Shannon’s theorem

The Fourier relations are such that a functional product in the time domain corresponds to a functional convolution in the frequency domain. Regular time-domain sampling with a time period $T$ (as in Fig. 5.2) thus corresponds to a periodic repetition of the (signal and noise) spectrum over a frequency period $1/T$ (as in Fig. 5.3).

Shannon’s theorem, which is sometimes also called the Nyquist theorem, states that the sampled signal is as “complete” as the original if this original contains no frequencies beyond the so-called Nyquist frequency $1/(2T)$; from the (low-frequency) power spectrum of the sample data we can then fully reconstruct the original unsampled time trace. If frequencies beyond $1/(2T)$
5.2. CONSEQUENCES OF SAMPLING: BITNOISE & SHANNON’S THEOREM

Figure 5.2: Periodic sampling and discretization of an analog signal reduces the continuous curve $V(t)$ into a set of digital values that are evenly spaced in time at integer multiples of the sampling time $T$.

Figure 5.3: Time-domain sampling with a time period $T$ leads to a repetition of the (signal and noise) spectrum over frequency multiples of $1/T$. Shannon’s theorem states that the sampled signal is as complete as the original if this original contains no frequencies beyond the so-called Nyquist frequency $1/(2T)$, i.e., if the repeating power spectra do not overlap.
do exist, the repeating power spectra will have some spectral overlap (not shown) and some high frequency components in the original will pop up at much lower frequencies in the sampled signal. This effect is called *aliasing* and is based on the notion that repetitive time domain sampling with a period $T$ does not allow one to distinguish between frequencies of the form $f - N/T$, with integer $N$. 
Chapter 6

FFT and z-transform

6.1 Discrete & Fast Fourier Transform

The discrete Fourier is the discrete version of the Fourier transform discussed in Chapter 2. Given a series of data points $x_n$, the discrete Fourier transform and its inverse are defined as:

\[
\tilde{x}_m = \frac{1}{\sqrt{N}} \sum_{n=0}^{N} x_n \exp\left(\pm i2\pi n.m/N\right), \quad (6.1)
\]

\[
x_n = \frac{1}{\sqrt{N}} \sum_{m=0}^{N} \tilde{x}_m \exp\left(-i2\pi n.m/N\right). \quad (6.2)
\]

In the above formulation we chose to distribute the normalization evenly over the forward and backward discrete Fourier transformations, which both contain a pre-factor $1/\sqrt{N}$. Other choices are also possible, like a factor 1 in the forward and a factor $1/N$ in the backward transformation, or vice versa. Even the sign in the exponent might differ, but the idea of “forward and backward transformation” is always the same.

With respect to the sign in the exponent, we will take a short detour. It is interesting to note that engineers generally work with the complex number $j$ ($j^2 = -1$), as they reserve the symbol $i$ for the electrical current, and generally define the Fourier relation from frequency to time via the evolution $\exp(j\omega t)$. This was also our choice when we discussed the complex impedance of inductances and capacitors in terms of $+j\omega L$ and $+1/(j\omega C)$; in the relation $V = ZI$, the voltage oscillation over the inductance is (1/4 period) ahead to the current oscillation, whereas it is delayed (again by 1/4 period) for the capacitor. Mathematicians (and most physicist), however,
generally work with the complex number $i (i^2 = -1)$ and rather discuss time evolution in terms of $\exp(-i\omega t)$; this preference for righthanded revolution in the complex plane at positive frequency $\omega$ is linked to the sign choice made for the Schrödinger equation: $ihd\Phi/dt = E\Phi$.

When the number of data points $N$ is an exact power of 2, the discrete Fourier transform can be solved relatively fast with a computation trick that makes the transform into what is called a Fast Fourier Transform = FFT. A standard discrete Fourier transform requires $N^2$ operations, being $N$ summations (one for each $m$ in the set 0,1,...,$N-1$), each containing $N$ terms of the form $x_n \exp(i2\pi n.m/N)$. The trick that is used in a FFT is that one looks for repetitive patterns in these $N$ summations, thereby using the multiplicative properties of the exponential function as $\exp(a+b) = \exp(a) \cdot \exp(b)$. Specifically, one uses the notion that the factor $\exp(i2\pi m.n/N)$ can take on at most $N$ different values. Furthermore one performs a binary expansion of either $n$ or $m$, starting with a division in odd and even, followed by a division in multiples of four plus some integer, etc. Clever cuts in half and reshuffles of the $N$ summations thereby reduce the total number of operations from order $N \times N = N^2$ to order $N \times 2\ln(N)$. This speeds up the Fourier discrete Fourier transform tremendously (already about a factor 100 for 1024 data points) and validates the name Fast Fourier Transform. This speed-up procedure is sometimes denoted as the “decimation-in-time” of the “butterfly computation”, names that refer to the mentioned binary expansion and shuffle, respectively.

### 6.2 Reduction of spectral leakage by windowing

The discrete Fourier transform discussed in section 6.1 differs in two ways from the continuous Fourier transform that was discussed in chapter 2. The first difference is of course that we now use only a finite set of data points $x_n \equiv x(nT)$, taken at discrete time intervals $T$, and forget all intermediate values $x(t)$. Mathematically, this reduction is equivalent to a multiplication in the time domain by a comb function

$$\delta_T(t) \equiv \sum_{n=-\infty}^{n=\infty} \delta(t - nT) \quad (6.3)$$

at least when we forget about the normalization. In the frequency domain this operation corresponds to a convolution with a similar comb function of
6.2. REDUCTION OF SPECTRAL LEAKAGE BY WINDOWING

The form
\[ \delta_{f_0}(f) \equiv \sum_{n=-\infty}^{\infty} \delta(f - nf_0), \]  
(6.4)
where \( f_0 = 1/T \). The consequence of this convolution is that we can’t distinguish between frequencies \( f \) and \( f - nf_0 \) (\( n \) integer) anymore. The phenomenon that slow sampling can produce ghost images of some harmonic signal is called aliasing. The statement that the frequency-shifted images will remain well separated from the original spectrum when this spectrum contains no frequencies beyond \( 1/(2T) \) is called Shannon’s theorem (see also Chapter 5).

A second difference between the discrete and continuous Fourier transform is that the former is always taken over a finite time window \( T_w \equiv NT \). Mathematically, this truncation can be accomplished by multiplying the full signal with a time window \( w(t) \). Such a multiplication in the time domain again corresponds with a convolution in the frequency domain, but now with the Fourier-transform of this window, which we’ll denote as \( \tilde{w}(f) \). This convolution will smear out the true spectral information and give rise to so-called spectral leakage; the truncation by the finite time window produces spectral components at frequencies different from the original frequency.

In its most simple form the truncation window \( w(t) \) is rectangular, making it’s Fourier transform equal to \( \tilde{w}(f) \propto \sin(2\pi T_w f)/(2\pi f) \). Although this sinc-function is quite sharp in the frequency domain it has sizeable side peaks and wings having a spectral amplitude that decays only as \( 1/\Delta f \). The reason for this slow spectral decay is basically the discontinuity that occurs in the time domain as a result of the truncation. As the Fourier series are calculated only at discrete frequencies \( f = m/T_w \), the time domain signal is implicitly assumed to repeat itself periodically with a period equal to the length \( T_w = NT \) of the time window. It is the discontinuity between the signal at the end of one time window and the start of the next window that results in the mentioned spectral leakage. From this argument it should be clear that the amount of spectral leakage depends on the exact relation between the observed oscillations and the length of the time window; we expect no spectral leakage at all when an integer number of oscillations fit exactly within the time window, but serious leakage when this integer relation is violated.

Spectral leakage can be reduced considerably by using more subtle and smoother truncation schemes than blunt time-domain cutting. Popular forms of truncation are the Hanning (or Hann) window \( w(t) = 0.5 - 0.5 \cos(2\pi n/N) \) and the Hamming window \( w(t) = 0.54 - 0.46 \cos(2\pi n/N) \). The Hanning window starts and ends at \( w(t) = 0 \) and is thus very smooth in time; its spectral width is relatively limited and its wings drop quickly below that of
the sinc-function mentioned above. The Fourier transform of the Hamming window is somewhat wider, but this window is optimised for minimum height of its first harmonic sidebands. Other, more complicated, windows are also in use and can by tried in the practicum experiment SVR4. As before, the windows are generally nothing more than a sum of cosine functions with some well-chosen amplitudes (first-order filter have only one cosine term, second-order filters also contain a cosine term of the double frequency) and the basic idea is always the same: multiplication by these windows make the transition from the end of one time frame to the start of the next repetitive frame as smooth as possible. Any such action will necessarily shorten the useful time span and thus lead to a reduction in spectral resolution, i.e., even more unwanted spectral components close to oscillation frequency. However, it will also lead to a strong reduction of the spectral wings, i.e., to the reduction of spectral leakage that we were after.

### 6.3 Noise synthesis

Now that you know how the (fast) Fourier transform works, you are also able to synthesize your own noise source. You just have to create a time-varying statistical signal $x(t)$, with the proper strength and the proper time dynamics (= autocorrelation function $R(\tau) = \langle x(t)x(t+\tau) \rangle$) or noise spectral density $S_x(f) \propto \{F(R)\}(f)$. A convenient signal to start from is a string of random numbers taken from the (normalized) normal distribution that is produced by many specialized random generator. In principle, such a string $x_i$ already represent a spectrally white noise source, with a autocorrelation function that peaks around a single point, as $\langle x_i x_j \rangle = \delta_{ij}$.

The noise trace that is produced by a set of random uncorrelated points looks very erratic and spiky. To smoothen this noise trace and create the proper time dynamics we could convolute the string $x_i$ with the appropriate time response function $R(\tau)$. However, such a convolution is quite time consuming. The noise synthesis is often much faster in the frequency domain, where a similar string of complex! random numbers $\hat{x}_i$ now represents the noise amplitude spectrum. In the frequency domain spectral shaping can be done easily through multiplication with the filter function $\{F(R)\}(f)$, being the Fourier transform of $R(\tau)$. After this multiplication a (fast) Fourier transformation of the string $\{F(R)\}, \hat{x}_i$ yields a new string of complex values of which both the real and imaginary parts are noise traces with the spectral properties that we wanted to synthesize.

The noise traces shown in Fig. 6.1 have been synthesized in MATLAB with the procedure described above: we started with a string of random
complex numbers and multiplied by spectral filters with either a Lorentzian (transmission) profile or a square (= “top-hat”) profile. More specifically, we started with as much as 8192 random complex numbers and used frequency filters with equivalent noise bandwidths of only 16 points. This large number of data points is definitely overdone, but at least makes the curves look continuous, as they contain 8192/16 = 512 points per (normalized) time unit. The lower curve shows the output behind the top-hat filter; this output contains only 16 discrete frequency components and could therefore also have been synthesized with only 16 frequency points. The top curve shows the result of Lorentzian filtering; although this curve contains quite a few additional high-frequency components (the Lorentzian filter has quite extended wings of the form $1/(1 + i\omega/\omega_0)$), we still recognize the same low-frequency components as in the lower curve.

There are many different methods of noise synthesis. Although most of these methods are based on the addition of a set of sinusoidal functions
with statistically random amplitudes and phases, the requirement to generate noise with a specific autocorrelation function of average noise spectral density is not sufficient to completely pinpoint the noise statistics of these Fourier components. It is for instance possible to fix all amplitudes and introduce the randomness only in the phases \( \phi(f) \), making the Fourier amplitude \( a(f) = \sqrt{S(f)} \exp(i\phi(f)) \). On the other hand, it is also possible to use the freedom in the choice of the statistical of \( a(f) \) to manipulate the probability distribution of the time trace \( N(t) \) of the noise. The natural approach would be to aim for a Gaussian probability distribution, which we get when we take uncorrelated random values for \( a(f) \). It is, however, also possible to generate noise \( N(t) \) with an almost uniform probability distribution. In the practicum experiment SVR4 you can play around with some of these special noise sources.

6.4 The z-transform

Although the Fourier transform is the most popular, there are other data transformations that have their own specific applications. One of these is the Laplace transform, which uses real-valued negative exponents instead of the complex-valued exponents of the Fourier transform. Another option is the so-called \( z \)-transform, which transforms a series of (real or complex) data points \( x_i \) into a single function \( f(z) \) of the complex variable \( z \), by using them as the coefficients of the Taylor expansion of this function, via

\[
f(z) \equiv \sum_{0}^{N} x_i z^i.
\]  

Note that this \( z \)-transform turns into a fast Fourier transform for \( z \) values of the form \( z = \exp(2\pi i n/N) \), laying at equal distance on the unit circle. The computational strength of the \( z \)-transform, just as with the Fourier transform, is that some operations are much easier in the domain of the transformed function \( f(z) \) than in the original domain \( x_i \). As a simple example we take a multiplication by a factor \( (1 - z) \) on \( f(z) \), which corresponds to a subtraction of the shifted signal \( x_{i-1} \) from the original \( x_i \) or to a (discrete) differentiation. Likewise the inverse operation \( f(z) \rightarrow f(z)/(1 - z) \) is equivalent to a (discrete) integration of \( x_i \), where we note that the similarity with the response \( H(\omega) = 1/(1 + i\omega \tau) \) of a low-pass filter is no coincidence. Another example is the double differentiation \( f(z) \rightarrow (1 - z)^2 \ast f(z) \). The discrete form of the second derivation apparently looks like \( x_i - 2x_{i-1} + x_{i-2} \). Now that you have seen a couple of examples, I hope that you understand why \( z \)-transformations can sometimes be convenient.
Bibliography


