Chapter 2

Introduction to digital signals

2.1 Sampling and A/D, D/A conversion

2.1.1 Shannon sampling theorem

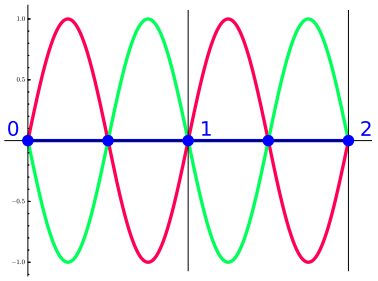
Suppose that we have a continuous time depending signal $f = f(t)$, for a mathematician a nice function $f : \mathbb{R} \longrightarrow \mathbb{R}$, and we want to treat it with magical digital processing tools. We need an A/D converter (analog-to-digital) giving a bunch of bits blocks out of the input signal. Typically it involves firstly to *sample* the signal, a discretization in time that produces values $f(t_n)$, and secondly a discretization of the sampled values, called *quantization*. In this way, the input analog signal is transformed into an ordered list of discrete values, say integers, that can be encoded with bits ans stored in a file or promptly processed digitally. If the required output is analog, for instance sound through loudspeakers, we will have to reverse the process with a D/A converter (digital-to-analog). In a scheme:

$$\begin{array}{cccccccc} & \overset{\text{analog}}{\underset{\text{input}}{\longrightarrow}} & \overbrace{\boxed{\text{sampling}} \rightarrow \boxed{\text{quantization}}}^{\text{A/D}} & \longrightarrow & \boxed{\begin{array}{c}\text{Digital}\\ \text{processing}\end{array}} & \longrightarrow & \boxed{\text{D/A}} & \overset{\text{analog}}{\underset{\text{output}}{\longrightarrow}} \\ & & f(t_n) \rightarrow \quad y_n & \longrightarrow & F(y_n) & \longrightarrow & g(t) & \end{array} \tag{2.1}$$

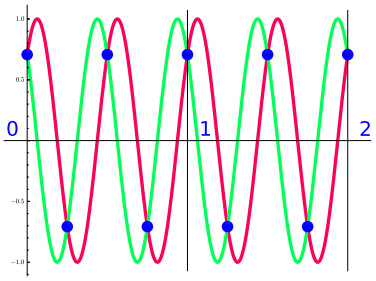
By the way, even this simple scheme is oversimplifying. For instance, an image in principle requires a function of two variables, the horizontal and vertical coordinates x and y. The images stored in our computer usually provide three channels of color (and perhaps an extra α-channel) then f is better modeled in this case as a two variables function with target in a 3-dimensional space.

A natural situation is the *uniform sampling*. In this case, if the time spacing is T the *sampling frequency* is $\nu_s = 1/T$. Again, for a mathematician uniform sampling corresponds to the sequence $\{f(n/\nu_s)\}_{n=-\infty}^{\infty}$ and it goes without saying that the sequence do not characterize the signal, even for pure tones. To convince yourself, look the graphs below. With the indicated sampling we cannot distinguish between the signals 0, $\sin(2\pi t)$

and $-\sin(2\pi t)$ in the first case, and between $\sin(4\pi t + \frac{\pi}{4})$ and $\cos(4\pi t + \frac{\pi}{4})$ in the second case. These signals become "aliases" with our methods. In general the indistinguishability of two signals under sampling is called *aliasing* and it is something that one wants to avoid imposing some conditions.



$\mathsf{T} = 1/2, \quad \nu_{\mathsf{s}} = 2$
$0, \quad \sin(2\pi \mathsf{t}), \quad -\sin(2\pi \mathsf{t})$



$\mathsf{T} = 1/4, \quad \nu_{\mathsf{s}} = 4$
$\sin(4\pi \mathsf{t} + \frac{\pi}{4}), \quad \cos(4\pi \mathsf{t} + \frac{\pi}{4})$

We are going to assume that $\widehat{f}$ is integrable and compactly supported, namely

$$\widehat{f}(\xi) = 0 \qquad \text{when } |\xi| \geq B. \tag{2.2}$$

Especially in the signal processing literature, f is said to be a *band limited signal*. It is not an unnatural hypothesis. Firstly because the common channels have a limit to transmit large frequencies (for instance, even modern fiber-optic cannot transmit at a rate of 1 exabyte per second) and secondly because very often we are not interested in large frequencies and they are already filtered.

For instance, you cannot hear anything beyond $20000\,Hz$ and then for audio applications one can assume $B = 20000$ in (2.2) without losing anything, in fact for adults the upper of audible frequencies is in general smaller and probably few readers over 30 years old could hear[1] $16000\,Hz$ and we all have problems to distinguish nearby frequencies in a much smaller range. On the other hand, according to some authors, the frequency of voice in male adults is around $85 - 180\ Hz$ and $165 - 255\ Hz$ in female adults but it does not mean that we could take safely $B = 300$ because voice is a complicated signal and these values only indicate the fundamental frequency (something like the one with biggest Fourier coefficient). Bigger frequencies are needed to distinguish the different sounds. They actually "form" the voice and the most important of them (let us say, those giving peaks of the Fourier coefficients or Fourier transform) are called *formants* in phonetics and acoustic. To give a figure, $B = 5000\,Hz$ suits for voice without noticeable (by me) loss of quality.

Coming back to the topic of this section, Shannon sampling theorem asserts that a band limited signal can be recovered with uniform sampling if the sampling frequency is large enough. It is rather a lemma or a clever observation than a theorem unquestionably

[1] If you want to try, there is an applet in `https://www.echalk.co.uk/Science/biology/hearing/HowOldIsYourHearing/resource.html` that guesses your age in terms of the result. Let me add that I am not very confident with these online tests because your computer and your loudspeakers or headphones play a role here.

deserving this name and probably the name has been popularized by engineers. Judge by yourself. We do not pursue the best regularity hypotheses. In connection with this, note that if f is continuous, and we implicitly assume so, using the inversion formula we conclude from (2.2) that $f \in C^\infty$. In fact, it has a entire extension that is characterized by *Paley-Wiener theorem* as any real-complex variable lover knows.

Theorem 2.1.1 (Shannon sampling theorem)**.** *Let f be a function with $\widehat{f} \in C^2$ satisfying* (2.2) *with $2B \leq \nu_s$. Then*

$$f(t) = \sum_{n=-\infty}^{\infty} f(n/\nu_s)\,\mathrm{sinc}(\nu_s t - n) \tag{2.3}$$

where sinc *is as in* (1.60)*.*

In other words, given a sampling frequency ν_s, the maximal frequency that can be contained in our signal to be fully determined is $\nu_s/2$. It is called the *Nyquist frequency*.

Proof. Let g be the ν_s-periodic extension of $\widehat{f}$ restricted to $I = [-\nu_s/2, \nu_s/2]$. The Fourier expansion of g is

$$g(\xi) = \nu_s^{-1} \sum_{n=-\infty}^{\infty} e(n\xi/\nu_s) \int_I \widehat{f}(x) e(-nx/\nu_s)\, dx. \tag{2.4}$$

Note that I includes the support of $\widehat{f}$ and consequently the integral can be extended to $\mathbb{R}$ and evaluated as $f(-n/\nu_s)$ by the inversion formula. Substituting this expression for g into $f(t) = \int_I g(\xi) e(t\xi)\, d\xi$ we have

$$f(t) = \sum_{n=-\infty}^{\infty} f(-n/\nu_s)\nu_s^{-1} \int_I e((t + n/\nu_s)\xi)\, d\xi = \sum_{n=-\infty}^{\infty} f(-n/\nu_s)\,\mathrm{sinc}(\nu_s t + n) \tag{2.5}$$

and only remains to rename $n \mapsto -n$. □

There are several generalizations of this theorem [Mar91], [Wal96]. For instance, Papoulis generalized sampling theorem says that we can relax the condition $2B \leq \nu_s$ to $2B \leq N\nu_s$ if we can sample a vector of N filtered instances of the original signal (see the details in [Mar91, §4.2]). Essentially, if we multiply our knowledge of the signal by N, we can divide the sample rate by N.

In practice there is some uncertainty when measuring $f(n/\nu_s)$ plus a quantization error if we store it by digital means. Even if we reduce this source of errors to negligible levels, the application of Theorem 2.1.1 requires infinitely many samples to recover f and it does not seem very practical. Perhaps we can increase a lot ν_s but still the number of samples must stop at some point.

Let us say that we only sample in a certain finite interval $[-T, T]$ and we do not assume T to be very large (imagine that we have time limited access to the signal). If our technology allows us to take many samples, interpolating them, we can assume that

we know quite well the signal in this interval. The mathematical question that arises is if $f|_{[-T,T]}$ determines f satisfying (2.2). The answer is "yes" because f is in fact an analytic function but the real question is how to construct such f.

The *Papoulis-Gerchberg algorithm* [Pap75] is an iterative scheme that addresses this problem. Let us denote χ_A the characteristic function of $[-A, A]$. The starting point is the known function $f_0 = f\chi_T$. Of course $\widehat{f_0}$ cannot be compactly supported then we force the condition (2.2) considering $\widehat{f_0}\chi_B$. The inverse Fourier transform of this function is band limited but fails to coincide with the known values of f, then we modify it by hand in $[-T, T]$ to obtain f_1 and we repeat the process. In one line

$$f_{n+1} = f_0 + (1 - \chi_T)(\widehat{f_n}\chi_B)^{\check{}}. \tag{2.6}$$

We know that the convolution turns into a product under the Fourier transform hence we can write this as

$$f_{n+1} = f_0 + (1 - \chi_T)(f_n * h) \qquad \text{with} \quad h(t) = 2B\,\mathrm{sinc}(2Bt). \tag{2.7}$$

It is known that the algorithm converges but it is slow for some practical applications. There are several techniques to speed it up that in some way parallel those employed for ill-conditioned linear systems (cf. [Byr04]). The discretization of the algorithm allows to apply it into the setting of genuine finite samples [Byr15], [Xia93, (9)].

In connection with these ideas, note that if $\nu_s = 2B$ then

$$F(x) = \sum_{n=1}^{N} f(n/\nu_s)\,\mathrm{sin}(\nu_s x - n) \tag{2.8}$$

satisfies $F(n/\nu_s) = f(n/\nu_s)$ for $n = 1, 2, \ldots, N$ and it is always a band limited function with $\widehat{F}$ supported in $[-B, B]$ but $\widehat{F}$ is discontinuous in general.

Shannon sampling theorem is closely related to the *Poisson summation formula* which is obtained integrating the formula (1.37) for δ_P against a function f defined on $\mathbb{R}$. The Dirac comb $\sum \delta(x - n)$ evaluates f at the integers and we get

$$\sum_{n=-\infty}^{\infty} f(n) = \sum_{n=-\infty}^{\infty} \widehat{f}(n). \tag{2.9}$$

The actual rigorous proof requires to consider the 1-periodic function $F(x) = \sum_{k\in\mathbb{Z}} f(k+x)$ and expand it into Fourier series as

$$F(x) = \sum_{n=-\infty}^{\infty} \int_0^1 F(t)e^{-2\pi int}dt\; e^{2\pi inx} = \sum_{n=-\infty}^{\infty} \int_{-\infty}^{\infty} f(t)e^{-2\pi int}dt\; e^{2\pi inx}. \tag{2.10}$$

Taking $x = 0$ we get (2.9). To dub this proof as rigorous we must include some properties assuring that F is well defined and that it can be Fourier expanded. Clearly this holds if f is in the Schwartz class. Of course this is overkilling, a less restrictive condition is for instance $f \in C^2$ with $f(x), f'(x), f''(x) = O(|x|^{-2})$ as $x \to \infty$. In [Zyg88, II.13], the regularity conditions are relaxed a lot as stated in the following result that we do not prove here.

Theorem 2.1.2. *If $f : \mathbb{R} \longrightarrow \mathbb{R}$ is an integrable function of bounded variation such that $f(x+h) + f(x-h) \to 2f(x)$ as $h \to 0$ for every $x \in \mathbb{R}$, then (2.9) holds true.*

Indeed, with some modifications in (2.9) it is possible to cover not integrable cases like $f(x) = x^{-\alpha}$ [Gui41]. Poisson summation formula has striking consequences in a broad range of topics [CR17], for instance, believe or not the best known results on sphere packing are based on it [Coh17].

If we apply (2.9) to $f(x) = g(qx + a)$, we get the following generalization

$$\sum_{n=-\infty}^{\infty} g(qn + a) = \frac{1}{q} \sum_{n=-\infty}^{\infty} e(an/q)\widehat{g}(n/q). \tag{2.11}$$

For $q = 1$, $a = 0$ we recover (2.9). In principle one could derive Shannon sampling theorem from here taking $g(x) = \operatorname{sinc}(\nu_s x) f(t + x)$ with $q = \nu_s^{-1}$ and $a = -t$. In this way the left hand side of (2.11) is the right hand side of (2.3).

Note also that for $q = 1$, $a = x$ we get the Fourier expansion of F in (2.10). In general, Poisson summation formula can be used instead of Fourier expansion when the coefficients can be easily interpolated to a smooth function.

One of the most famous and interesting applications of the Poisson summation formula is the θ modular relation

$$\sum_{n=-\infty}^{\infty} e^{-2\pi\alpha n^2} = \frac{1}{\sqrt{2\alpha}} \sum_{n=-\infty}^{\infty} e^{-\pi n^2/2\alpha} \qquad \text{for} \quad \alpha > 0. \tag{2.12}$$

If α is very small the first sum is expensive from the computational point of view while the second gives readily the approximation $1/\sqrt{2\alpha}$ with exponential gain.

Suggested Readings. The whole book [Mar91] is devoted to topics around Shannon sampling theorem. In [Byr15] there are several discussions about sampling and reconstruction scattered along several chapters. See [CR17] for more about the Poisson summation formula.

2.1.2 Basic quantization

Recall the scheme (2.1). The sampled values $f(t_n)$ are real numbers and we want to approximate them by discrete quantities. This process is called *quantization*.

The function mapping each real number into the nearest integer, sometimes called "round", is given by the formula $x \mapsto \lfloor x + 1/2 \rfloor$ where $\lfloor x \rfloor$ is the integral part defined in (1.56). If instead of the integers $\mathbb{Z}$ we want the output to be Δ multiples of integers to get more (or less precision), we re-scale the argument and the value of this function to find the so called *uniform quantizer*

$$Q(x) = \Delta \lfloor \frac{x}{\Delta} + \frac{1}{2} \rfloor \qquad \text{with } \Delta > 0. \tag{2.13}$$

This is the simplest and more common quantization. The geometric idea is that $Q(f)$ is the approximation of f by a step function such that the spacing between steps is always a multiple of Δ. With Δ small the approximation will be good.