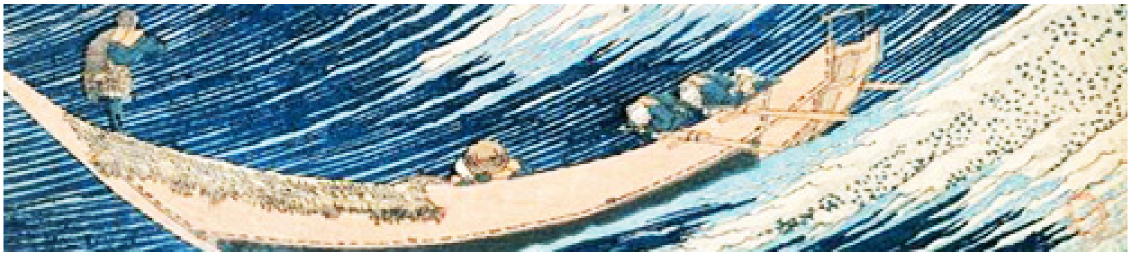
Mathematical Foundations of Data Sciences



Gabriel Peyré
CNRS & DMA
École Normale Supérieure
gabriel.peyre@ens.fr
https://mathematical-tours.github.io
www.numerical-tours.com

September 11, 2024

Chapter 1

Shannon Sampling Theory

Shannon's theory of information, published in 1948/1949, is made of three parts:

1. Sampling: it studies conditions under which sampling a continuous function to obtain a discrete vector is invertible. The discrete real values representing the signal are then typically quantized to a finite precision to obtain a set of symbols in a finite alphabet.

2. Source coding: it studies optimal ways to represent (code) such a set of symbols as a binary sequence. It leverages the statistical distributions to obtain the most possible compact code.

3. Channel coding (not studied here): it studies adding some redundancy to the coded sequence to gain robustness to errors or attacks during transmission (flip of certain bits with some probability). It is often named "error correcting codes theory".

This chapter is focussed on the sampling theory and Chapter **??** is dedicated to the source coding theory. We do not cover channel coding. The main reference for this chapter is [1].

1.1 Analog vs. Discrete Signals

To develop numerical tools and analyze their performances, the mathematical modeling is usually done over a continuous setting (so-called "analog signals"). Such a continuous setting also aims at representing the signal in the physical world, which are inputs to sensor hardware such as microphones, digital cameras, or medical imaging devices. An analog signal is a 1-D function $f_0 \in \mathrm{L}^2([0,1])$ where $[0,1]$ denotes the domain of acquisition, which might for instance be time. An analog image is a 2D function $f_0 \in \mathrm{L}^2([0,1]^2)$ where the unit square $[0,1]^2$ is the image domain.

Although these notes are focused on the processing of sounds and natural images, most of the methods extend to multi-dimensional datasets, which are higher-dimensional mappings

$$f_0 : [0,1]^d \to [0,1]^s$$

where d is the dimensionality of the input space ($d = 1$ for sound and $d = 2$ for images) whereas s is the dimensionality of the feature space. For instance, grayscale images correspond to $(d = 2, s = 1)$, videos to $(d = 3, s = 1)$, color images to $(d = 2, s = 3)$ where one has three channels (R, G, B). One can even consider multi-spectral images where $(d = 2, s \gg 3)$ is made of many channels for different light wavelengths. Figures 1.1 and 1.2 show examples of such data.

1.1.1 Acquisition and Sampling

Signal acquisition is a low-dimensional projection of the continuous signal performed by some hardware device. This is for instance the case for a microphone that acquires 1D samples or a digital camera that

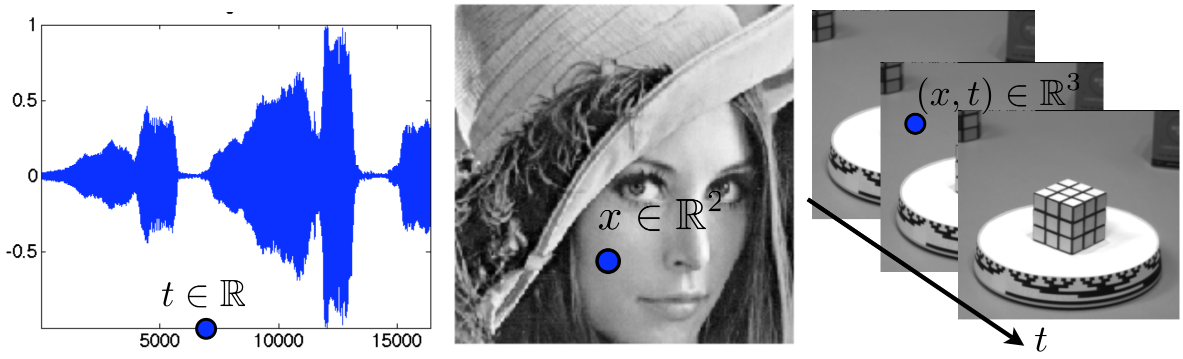


Figure 1.1: Examples of sounds ($d = 1$), image ($d = 2$) and videos ($d = 3$).

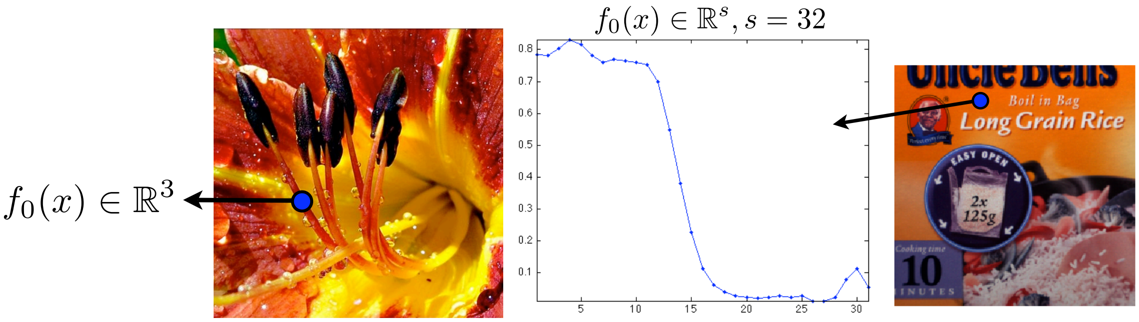


Figure 1.2: Example of color image $s = 3$ and multispectral image ($s = 32$).

acquires 2D pixel samples. The sampling operation thus corresponds to mapping from the set of continuous functions to a discrete finite dimensional vector with N entries.

$$f_0 \in \mathrm{L}^2([0,1]^d) \mapsto f \in \mathbb{C}^N$$

Figure 1.3 shows examples of discretized signals.

1.1.2 Linear Translation Invariant Sampler

A translation-invariant sampler performs the acquisition as an inner product between the continuous signal and a constant impulse response h translated at the sample location

$$f_n = \int_{-S/2}^{S/2} f_0(x) h(n/N - x) \mathrm{d}x = f_0 \star h(n/N). \tag{1.1}$$

The precise shape of $h(x)$ depends on the sampling device and is usually a smooth low-pass function that is maximal around $x = 0$. The size S of the sampler determines the precision of the sampling device and is usually of the order of $1/N$ to avoid blurring (if S is too large) or aliasing (if S is too small).

Section **??** details how to reverse the sampling operation in the case where the function is smooth.

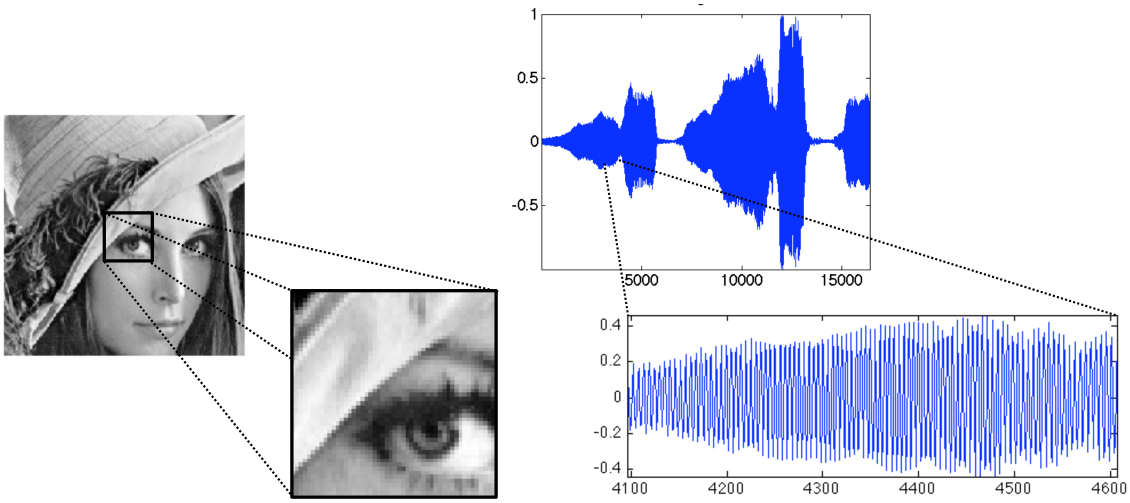


Figure 1.3: Image and sound discretization.

1.2 Shannon Sampling Theorem

Reminders about Fourier transform. For $f \in L^1(\mathbb{R})$, its Fourier transform is defined as

$$\forall\, \omega \in \mathbb{R}, \quad \hat{f}(\omega) \overset{\text{def.}}{=} \int_{\mathbb{R}} f(x) e^{-\mathrm{i}x\omega} \mathrm{d}x. \tag{1.2}$$

One has $\|\hat{f}\|^2 = (2\pi)^{-1}\|f\|^2$, so that $f \mapsto \hat{f}$ can be extended by continuity to $L^2(\mathbb{R})$, which corresponds to computing $\hat{f}$ as a limit when $T \to +\infty$ of $\int_{-T}^{T} f(x)e^{-\mathrm{i}x\omega}\mathrm{d}x$. When $\hat{f} \in L^1(\mathbb{R})$, one can invert the Fourier transform so that

$$f(x) = \frac{1}{2\pi}\int_{\mathbb{R}} \hat{f}(\omega)e^{\mathrm{i}x\omega}\mathrm{d}\omega, \tag{1.3}$$

which shows in particular that f is continuous with vanishing limits at $\pm\infty$.

The Fourier transform $\mathcal{F} : f \mapsto \hat{f}$ exchanges regularity and decay. For instance, if $f \in C^p(\mathbb{R})$ with an integrable Fourier transform, then $\mathcal{F}(f^{(p)})(\omega) = (\mathrm{i}\omega)^p \hat{f}(\omega)$ so that $|\hat{f}(\omega)| = O(1/|\omega|^p)$. Conversely,

$$\int_{\mathbb{R}} (1+|\omega|)^p |\hat{f}(\omega)| \mathrm{d}\omega < +\infty \quad \Longrightarrow \quad f \in C^p(\mathbb{R}). \tag{1.4}$$

For instance, if $\hat{f}(\omega) = O(1/|\omega|^{p+1+\varepsilon})$ for $\varepsilon > 0$, one obtains that $f \in C^p(\mathbb{R})$.

A related, but different way to impose smoothness is by using the Sobolev H^1 norm $\int_{\mathbb{R}}(1+|\omega|^2)|\hat{f}(\omega)|^2\mathrm{d}\omega$, which, when f is smooth, is equal to $\|f\|_2^2 + \|f'\|_2^2$. It is a fundamental space to define weak solutions (non smooth) to PDE's, and also to control compression errors in image process. We will not pursue this here.

Reminders about Fourier series. We denote $\mathbb{T} = \mathbb{R}/2\pi\mathbb{Z}$ the torus. A function $f \in L^2(\mathbb{T})$ is 2π-periodic, and can be viewed as a function $f \in L^2([0, 2\pi])$ (beware that this means that the boundary points are glued together), and its Fourier coefficients are

$$\forall\, k \in \mathbb{Z}, \quad \hat{f}_k \overset{\text{def.}}{=} \frac{1}{2\pi}\int_0^{2\pi} f(x)e^{-\mathrm{i}xk}\mathrm{d}x.$$

This formula is equivalent to the computation of an inner-product $\hat{f}_k = \langle f,\, e_k \rangle$ for the inner-product $\langle f,\, g\rangle \overset{\text{def.}}{=} \frac{1}{2\pi}\int_{\mathbb{T}} f(x)\bar{g}(x)\mathrm{d}x$ and for $e_k(x) \overset{\text{def.}}{=} e^{\mathrm{i}xk}$. For this inner product, $(e_k)_k$ is orthonormal and is a Hilbert basis, meaning that one reconstructs with the following converging series

$$f = \sum_{n\in\mathbb{Z}} \langle f,\, e_k\rangle e_k \tag{1.5}$$

which means $\|f - \sum_{k=-N}^{N} \langle f,\, e_k\rangle e_k\|_{L^2(\mathbb{T})} \to 0$ for $N \to +\infty$. The pointwise convergence of (1.5) at some $x \in \mathbb{T}$ is ensured if, for instance, f is differentiable. The series is normally convergent (and hence uniform) if for instance f if of class C^2 on $\mathbb{T}$ since in this case, $\hat{f}_k = O(1/n^2)$. If there is a step discontinuity, then there are Gibbs oscillations preventing uniform convergence, but the series still converges to half of the left and right limits.

Poisson formula. The Poisson formula connects the Fourier transform and the Fourier series to sampling and periodization operators. For some function $h(\omega)$ defined on $\mathbb{R}$ (typically the goal is to apply this to $h = \hat{f}$), its periodization reads

$$h_P(\omega) \overset{\text{def.}}{=} \sum_n h(\omega - 2\pi n). \tag{1.6}$$

This formula makes sense if $h \in L^1(\mathbb{R})$, and in this case $\|h_P\|_{L^1(\mathbb{T})} \leqslant \|h\|_{L^1(\mathbb{R})}$ (and there is equality for positive functions). Indeed, one has

$$|h_P(x)| \leqslant (|h|_P)(x), \quad \Longrightarrow \quad \|h_P\|_{L^1} \leqslant \||h|_P\|_{L^1} = \|h\|_{L^1}.$$

The Poisson formula, stated in Proposition 1 below, corresponds to proving that the following diagram

$$\begin{tikzcd} f(x) \arrow[r, "\mathcal{F}"] \arrow[d, "\text{sampling}"'] & \hat{f}(\omega) \arrow[d, "\text{periodization}"] \\ (f(n))_n \arrow[r, "\text{Fourier serie}"] & \sum_n f(n)e^{-\mathrm{i}\omega n} \end{tikzcd}$$

is commutative. Beware that $\sum_n f(n)e^{-\mathrm{i}\omega n}$ is actually a reverse Fourier series (there is a + sign in the reconstruction formula for Fourier series).

Proposition 1 (Poisson formula)**.** *Assume that $\hat{f}$ has compact support and that $|f(x)| \leqslant C(1+|x|)^{-3}$ for some C. Then one has*

$$\forall\, \omega \in \mathbb{R}, \quad \sum_n f(n)e^{-\mathrm{i}\omega n} = \hat{f}_P(\omega). \tag{1.7}$$

Proof. Since $\hat{f}$ is compactly supported, $\hat{f}_P$ is well defined (it involves only a finite sum) and since f has fast decay, using (1.4), $(\hat{f})_P$ is C^1. It is thus the sum of its Fourier series

$$(\hat{f})_P(\omega) = \sum_k c_k e^{\mathrm{i}k\omega}, \tag{1.8}$$

where

$$c_k = \frac{1}{2\pi}\int_0^{2\pi} (\hat{f})_P(\omega)e^{-\mathrm{i}k\omega}\mathrm{d}\omega = \frac{1}{2\pi}\int_0^{2\pi} \sum_n \hat{f}(\omega - 2\pi n)e^{-\mathrm{i}k\omega}\mathrm{d}\omega.$$

One has

$$\int_0^{2\pi} \sum_n |\hat{f}(\omega - 2\pi n)e^{-\mathrm{i}k\omega}|\mathrm{d}\omega = \int_{\mathbb{R}} |\hat{f}|$$

which is bounded because $\hat{f} \in L^1(\mathbb{R})$ (it has a compact support and is C^1), so one can exchange the sum and integral

$$c_k = \sum_n \frac{1}{2\pi}\int_0^{2\pi} \hat{f}(\omega - 2\pi n)e^{-\mathrm{i}k\omega}\mathrm{d}\omega = \frac{1}{2\pi}\int_{\mathbb{R}} \hat{f}(\omega)e^{-\mathrm{i}k\omega}\mathrm{d}\omega = f(-k)$$

where we used the inverse Fourier transform formula (1.3), which is legit because $\hat{f} \in L^1(\mathbb{R})$. □

Shannon theorem. Shannon's sampling theorem states a sufficient condition ensuring that the sampling operator $f \mapsto (f(ns))_n$ is invertible for some sampling step size $s > 0$. It require that $\text{supp}(\hat{f}) \subset [-\pi/s, \pi/s]$, which, thanks to formula (1.3), implies that $\hat{f}$ is C^∞ (in fact it is even analytic). This theorem was first proved by Whittaker in 1915. It was re-proved and put in perspective in electrical engineering by Nyquist in 1928. It became famous after the paper of Shannon in 1949, which put forward its importance in numerical communications. Figure 1.4 gives some insight on how the proof works (left) and more importantly, on what happens when the compact support hypothesis fails (in which case aliasing occurs, see also Figure 1.7).

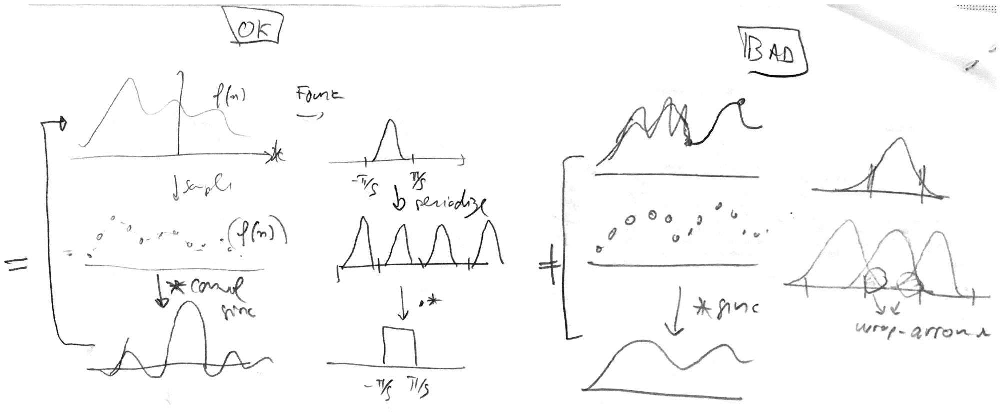


Figure 1.4: Schematic view for the proof of Theorem 1.

Theorem 1. *If* $|f(x)| \leqslant C(1 + |x|)^{-3}$ *for some* C *and* $\text{supp}(\hat{f}) \subset [-\pi/s, \pi/s]$, *then one has*

$$\forall\, x \in \mathbb{R}, \quad f(x) = \sum_n f(ns)\,\text{sinc}(x/s - n) \quad \textit{where} \quad \text{sinc}(u) = \frac{\sin(\pi u)}{\pi u} \tag{1.9}$$

with uniform convergence.

Proof. The change of variable $g \overset{\text{def.}}{=} f(s\cdot)$ results in $\hat{g} = 1/s\hat{f}(\cdot/s)$, indeed, denoting $z = sx$

$$\hat{g}(\omega) = \int f(sx)e^{-\mathrm{i}\omega x}\mathrm{d}x = \frac{1}{s}\int f(z)e^{-\mathrm{i}(\omega/s)z}\mathrm{d}z = \hat{f}(\omega/s)/s,$$

so that we can restrict our attention to $s = 1$. With this change of variable, we thus need to prove that

$$g(x) = \sum_n g(n)\,\text{sinc}(x - n),$$

(and in the following, we keep using the notation $f = g$). The compact support hypothesis implies $\hat{f}(\omega) = 1_{[-\pi,\pi]}(\omega)\hat{f}_P(\omega)$. Combining the inversion formula (1.3) with Poisson formula (1.8)

$$f(x) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \hat{f}_P(\omega)e^{\mathrm{i}\omega x}\mathrm{d}\omega = \frac{1}{2\pi}\int_{-\pi}^{\pi}\sum_n f(n)e^{\mathrm{i}\omega(x-n)}\mathrm{d}\omega.$$

Since f has fast decay, $\int_{-\pi}^{\pi}\sum_n |f(n)e^{\mathrm{i}\omega(x-n)}|\mathrm{d}\omega = \sum_n |f(n)| < +\infty$, so that one can exchange summation and integration and obtain

$$f(x) = \sum_n f(n)\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{\mathrm{i}\omega(x-n)}\mathrm{d}\omega = \sum_n f(n)\,\text{sinc}(x - n).$$

□

One issue with this reconstruction formula is that it uses slowly decaying and very oscillating sinc kernels. In practice, one rarely uses such a kernel for interpolation, and one prefers a smoother and more localized kernel. If $\text{supp}(\hat{f}) \subset [-\pi/s', \pi/s']$ with $s' > s$ (i.e. have a more compact spectrum), one can re-do the proof of the theorem, and one gains some degree of freedom to design the reconstruction kernel, which now can be chosen smoother in Fourier and hence have exponential decay in time.

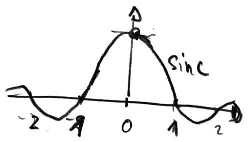


Figure 1.5: sinc kernel

Spline interpolation are defined by considering $\varphi_0 = 1_{[-1/2,1/2]}$ and $\varphi_k = \varphi_{k-1}\star\varphi_0$ which is a piecewise polynomial of degree k and has bounded derivative of order k (and is of class C^{k-1}) with compact support on $[-(k+1)/2, (k+1)/2]$. The reconstruction formula reads $f \approx \tilde{f} \overset{\text{def.}}{=} \sum_n a_n \varphi(\cdot - n)$ where $(a_n)_n$ is computed from the $(f(n))_n$ by solving a linear system (associated to the interpolation property $\tilde{f}(n) = f(n)$). It is only in the cases $k \in \{0, 1\}$ (piecewise constant and affine interpolations) that one has $a_n = f(n)$. In practice, one typically use the cubic spline interpolation, which corresponds to $k = 3$.

`Associated code: test_sampling.m`

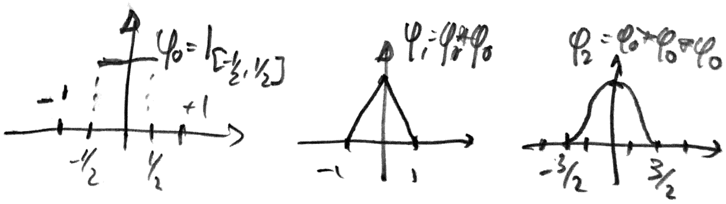


Figure 1.6: Cardinal splines as basis functions for interpolation.

This theorem also explains what happens if $\hat{f}$ is not supported in $[-\pi/s, \pi/s]$. This leads to aliasing, and high frequency outside this interval leads to low frequency artifacts often referred to as "aliasing". If the input signal is not bandlimited, it is thus very important to pre-filter it (smooth it) before sampling to avoid these phenomena (of course this kills the high frequencies, which are lost), see Figure 1.7.

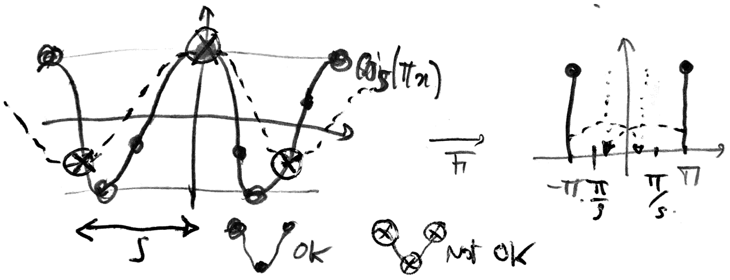


Figure 1.7: Aliasing in the simple case of a sine wave (beware however that this function does not have compact support).

Quantization. Once the signal have been sampled to obtain a discrete vector, in order to store it and transmit it, it is necessary to quantize the value to some finite precision. Section **??** presents transform coding, which is an efficient family of compression schemes which operates the quantization over some transformed domain (which corresponds to applying a linear transform, usually orthogonal, to the sampled values). This is useful to enhance the performance of the source coding scheme. It is however common to operate directly the quantization over the sampled value.

Considering for instance a step size $s = 1/N$, one samples $(u_n \overset{\text{def.}}{=} f(n/N))_{n=1}^N \in \mathbb{R}^N$ to obtain a finite dimensional data vector of length N. Note that dealing with finite data corresponds to restricting the function f to some compact domain (here $[0,1]$) and is contradictory with the Shannon sampling theorem since a function f cannot have compact support in both space and frequency (so perfect reconstruction never holds when using finite storage).

Choosing a quantization step T, quantization $v_n = Q_T(u_n) \in \mathbb{Z}$ rounds to the nearest multiple of T, i.e.

$$v = Q_T(u) \quad \Leftrightarrow \quad v - \frac{1}{2} \leqslant u/T < v + \frac{1}{2},$$

see Fig. **??**. De-quantization is needed to restore a signal, and the best reconstruction (in average or in worse case) is defined by setting $D_T(v) \overset{\text{def.}}{=} Tv$. Quantizing and then de-quantizing introduce an error bounded by $T/2$, since $|D_T(Q_T(u)) - u| \leqslant T/2$. Up to machine precision, quantization is the only source of error (often called "lossy compression") in Shannon's standard pipeline.

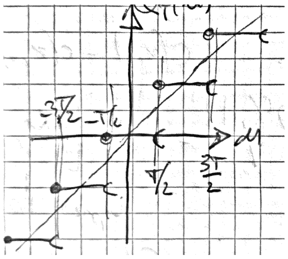


Figure 1.8:

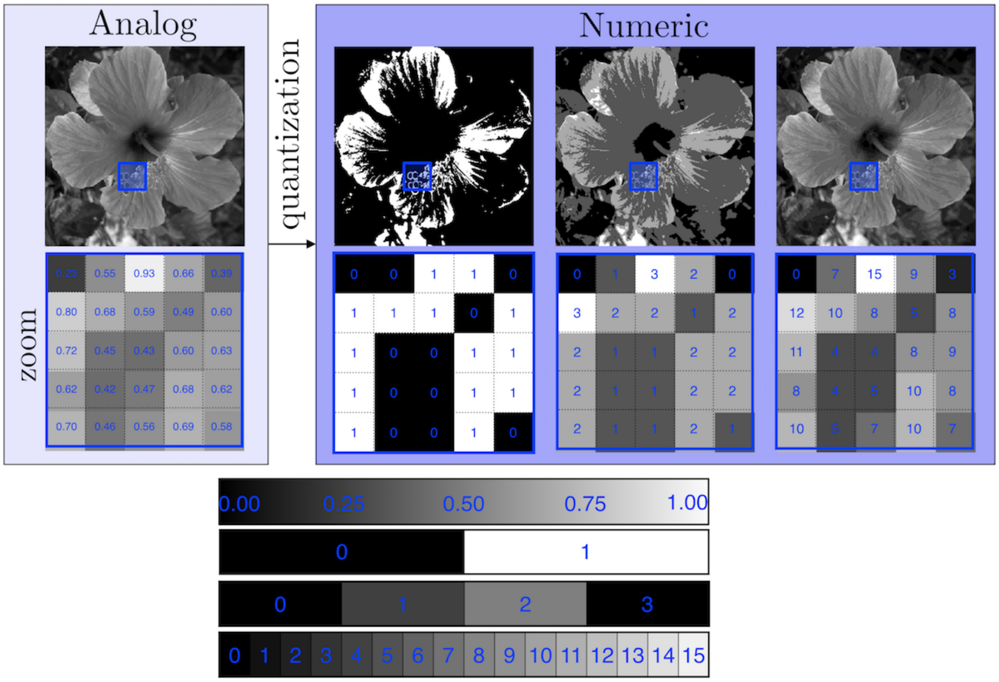


Figure 1.9: Quantizing an image using a decaying $T = 1/K$ where $K \in \{2, 3, 4, 16\}$ is the number of graylevels and the original image is normalized so that $0 \leqslant f_0 < 1$.

Bibliography

[1] Stephane Mallat. *A wavelet tour of signal processing: the sparse way.* Academic press, 2008.