Advanced Material

A Mathematical Tribute to Shannon, Nyquist & Co.

When the Proof of the Sampling Theorem Makes Sense

Reference: P. Bremaud, "Mathematical Principles of Signal Processing", Springer Verlag.

École Polytechnique Fédérale de Lausanne - EPFL

A Mathematical Tribute to Shannon, Nyquist & Co.

The Sampling and Reconstruction Theorem

This is what most of you have learned by heart:

"*A* $[-B/2, B/2]$ *frequency band-limited signal* $x(t)$ *can be completely reconstructed from its samples* $x(nT_s)$ *taken at period* $T_s \leq 1/B$. *That is, reconstruction from the samples is possible when the sampling frequency* $f_s = 1/T_s$ *is greater or equal to twice the maximum frequency of the signal*".

While acknowledged, the result of such a theorem cannot be considered intuitive, nor straightforward. Trying make to make such a result simpler or a straightforward consequence of basic **Fourier** tools, yields to mathematical aberrations and unaesthetic computations that cannot be proven.

Shannon, Nyquist & Co. theorem has two mathematical rigorous formulations, one in the L^1 space, the other in the L^2 space, with very elegant related proofs. The formulation in L^1 shows that the Fourier transformation $X_{\text{sampled}}(f)$ of the sampled signal correspond to the f_s replicas of the Fourier transformation $X(f)$ of the analog signal, *i.e.*

$$X_{\text{sampled}}(f) = \sum_{k \in \mathbb{Z}} X(f - k f_s) \, .$$

Consequently, the analog signal is reconstructed by $[-B/2, B/2]$ low pass filtering the sampled signal. The formulation in L^2 shows that any band limited signal can be decomposed into a basis of cardinal sines, that is, it can be seen as the convolution between a discrete sequence and an ideal low pass filter.

Let us first start by showing what can go wrong when trying to present the sampling theorem as simpler, too intuitive, or too straightforward.

A Mathematical Tribute to Shannon, Nyquist & Co.

A Tempting Model and an Evil Computation

The temptation, leading to sinful developments, is to mathematically model the sampling of $x(s)$ (sampling frequency $f_s = \frac{1}{T_s}$), as the product between $x(t)$ and an infinite sum of shifted Diracs deltas

$$x_{\text{sampled}}(t) = x(t) \sum_{k\in\mathbb{Z}} \delta(t - kT_s)\,.$$

Such a temptation becomes diabolic when trying to derive the spectrum $X_{\text{sampled}}(f)$ of the sampled signal as the Fourier transformation of the above product, pretending to use the multiplication-convolution rule

$$\mathrm{FT}(x_{\text{sampled}}(\cdot)) = \mathrm{FT}(x(\cdot)) * \mathrm{FT}(\sum_{k\in\mathbb{Z}} \delta(\cdot - kT_s))\,,$$

and hoping to derive that

$$X_{\text{sampled}}(f) = \sum_{k\in\mathbb{Z}} X(f - kf_s)\,.$$

In order to prove that the Fourier transformation of a product is the convolution of the corresponding Fourier transformations, one has to express each term of the product as its inverse Fourier transformation. That is, in order to be able to apply the multiplication-convolution rule, $x(t)$ has to be expressed as inverse Fourier transform $x(t) = \int_{\mathbb{R}} X(f)e^{i2\pi ft}df$ and, since it is a periodic signal, $\sum_{k\in\mathbb{Z}} \delta(t - kT_s)$ has to be expressed as a Fourier series expansion $\sum_{k\in\mathbb{Z}} \delta(t - kT_s) = \sum_{k\in\mathbb{Z}} a_k e^{i2\pi kf_s t}$

But when we try to compute the coefficients of the Fourier series expansion we obtain a constant value α: The Fourier series expansion does not exist since $\sum_{k\in\mathbb{Z}} \alpha e^{i2\pi kf_s t}$ is a non convergente series!

Consequently, the multiplication-convolution rule does not apply, and it is mathematically impossible to prove, modeling the sampling as a sum of shifted Dirac deltas, that $X_{\text{sampled}}(f) = \sum_{k\in\mathbb{Z}} X(f - kf_s)$.

A Mathematical Tribute to Shannon, Nyquist & Co.

A Witchy Function

Such mathematical complications arise because modeling the sampling as the multiplication with a shifted Diracs deltas is probably intuitive but not an appropriate model

$$x_{\text{sampled}}(t) = x(t) \sum_{k \in \mathbb{Z}} \delta(t - kT_s) \, .$$

By using it one put itself into troubles!

The Dirac delta $\delta(t)$ is indeed not a function: Functions are defined point-wise, while the Dirac delta is defined as $\int_{\mathbb{R}} f(t)\delta(t)dt = f(0)$ for functions $f(t)$ belonging to a particular space.

Notice also that the Riemann integral of a Dirac delta is not defined, while its Lebesgue integral is $\int_{\mathbb{R}} \delta(t)dt = 0$, when $\delta(t)dt$ is interpreted as a Dirac measure.

The use of the sum of shifted Dirac deltas, commonly called a Dirac comb, is surely intuitive, but definitively a wrong mathematical tool for a rigzourous yet simple proof.

Of course, one can consider Dirac deltas in their natural environment, that is, distribution theory, and formulate the sampling theorem within such a theory. But using distribution theory to prove the sampling theorem is like shooting sparrows with a cannon!

A Mathematical Tribute to Shannon, Nyquist & Co.

- **But it is Written in Books!**

  That is right. Unfortunately some books present an "impossible to prove" formulation of the sampling theorem. They all get stuck with the inexistent definition of the Fourier series of a Dirac comb, pretending to be allowed to write

$$\frac{1}{T_s} \sum_{k \in \mathbb{Z}} e^{j2\pi kt/T_s} = \sum_{k \in \mathbb{Z}} \delta(t - kT_s)\,,$$

  which is, as discussed above, a mathematical nonsense. The left term is a divergent series, hence, not defined.

  So is a rigorous approach so complicated?

A Mathematical Tribute to Shannon, Nyquist & Co.

- **Making Things as Simple as Possible, But not Simpler**

  When adopting the right approach things are not complicated at all!

Theorem (Shannon, Nyquist & Co. in L^1)

Let $x(t)$ be a continuos time signal such that:

- $x(t) \in \mathrm{L}^1 \cap C^0$;
- Its Fourier transform X has a finite support $[-B/2, B/2]$;
- $\sum_{n\in\mathbb{Z}} |x(nT_s)| < \infty$ for some T_s such that $0 < T_s < 1/B$.

Then

$$\sum_{n\in\mathbb{Z}} X\left(f + \frac{n}{T_s}\right) = T_s \sum_{n\in\mathbb{Z}} x(nT_s) e^{-2\pi i f n T_s}\,, \quad \forall\ f \in \mathbb{R}\,,$$

and

$$x(t) = BT_s \sum_{n\in\mathbb{Z}} x(nT_s)\, \frac{\sin(2\pi(t - nT_s)B/2)}{2\pi(t - nT_s)B/2}\,, \quad \forall\ t \in \mathbb{R}\,.$$

A Mathematical Tribute to Shannon, Nyquist & Co.

Making Things as Simple as Possible, But not Simpler

Proof (Shannon, Nyquist & Co. in L^1)

The first statement is a direct consequence of the Poisson summation formula (it is the Poisson formula itself!)

The second statement in a consequence of

$$X(f) = 1_{[-\frac{B}{2},\frac{B}{2}]}(f) \sum_{k\in\mathbb{Z}} X\left(f + \frac{n}{T_s}\right), \quad \text{and} \quad x(t) = \int_{\mathbb{R}} e^{i2\pi ft} X(f) df\,.$$

Indeed

$$\begin{aligned}
x(t) &= \int_{\mathbb{R}} e^{i2\pi ft} 1_{[-\frac{B}{2},\frac{B}{2}]}(f) \sum_{n\in\mathbb{Z}} X\left(f + \frac{n}{T_s}\right) df \\
&\overset{\text{Poisson}}{=} \int_{\mathbb{R}} e^{i2\pi ft} 1_{[-\frac{B}{2},\frac{B}{2}]}(f) T_s \sum_{k\in\mathbb{Z}} x(nT_s) e^{-i2\pi fnT_s} df \\
&= T_s \int_{\mathbb{R}} \sum_{n\in\mathbb{Z}} x(nT_s) e^{-i2\pi fnT_s} e^{i2\pi ft} 1_{[-\frac{B}{2},\frac{B}{2}]}(f) df \\
&\overset{\text{Fubini}}{=} T_s \sum_{n\in\mathbb{Z}} x(nT_s) \int_{\mathbb{R}} e^{i2\pi f(t-nT_s)} 1_{[-\frac{B}{2},\frac{B}{2}]}(f) df \\
&= T_s \sum_{n\in\mathbb{Z}} x(nT_s) \frac{\sin(2\pi(t-nT_s)\frac{B}{2})}{\pi(t-nT_s)} = BT_s \sum_{n\in\mathbb{Z}} x(nT_s) \frac{\sin(2\pi(t-nT_s)\frac{B}{2})}{2\pi(t-nT_s)\frac{B}{2}}\,.
\end{aligned}$$

A Mathematical Tribute to Shannon, Nyquist & Co.

Making Things as Simple as Possible, But not Simpler

In the L^1 version of the sampling theorem one has to cope with the condition $\sum_{k\in\mathbb{Z}} |x(nT_s)| < \infty$, which is restrictive and difficult to be verified *a priori*. Formulation of the sampling theorem in the L^2 space is much more elegant and simple!

Theorem (Shannon, Nyquist & Co. in L^2)
Let $\mathrm{L}^2_{\mathbb{C}}(\mathbb{R}\,;B)$ be the Hilbert subspace of $\mathrm{L}^2_{\mathbb{C}}(\mathbb{R})$ consisting of the finite energy signals $x(t)$ with a Fourier transform $X(f)$ having finite support $[-B/2, B/2]$. Let T_s be such that $0 < T_s < 1/B$.

The sequence

$$\left\{ BT_s \frac{\sin(2\pi(t-nT_s)B/2)}{2\pi(t-nT_s)B/2} \right\}_{n\in\mathbb{Z}},$$

is an orthonormal basis of $\mathrm{L}^2_{\mathbb{C}}(\mathbb{R}\,;B)$ and we have

$$\lim_{N\to\infty} \int_{\mathbb{R}} \left| x(t) - \sum_{n=-N}^{N} \alpha_n BT_s \frac{\sin(2\pi(t-nT_s)B/2)}{2\pi(t-nT_s)B/2} \right|^2 dt = 0\,,$$

where $\alpha_n = \int_{-B/2}^{B/2} X(f) e^{i2\pi f nT_s}\, df$.

A Mathematical Tribute to Shannon, Nyquist & Co.

- **No Big Deal! The Result is the Same!**

  You may argue that a rigorous mathematical approach is not needed since, at the end, the result is the same when using the Dirac comb model and improperly applying Fourier transformations.

  In this particular case, as mentioned, the Dirac formalism can benefit of a rigorous framework given by the distribution theory (which, for the sampling theorem, is like shooting sparrows with a cannon!). Luckily enough, by mishandling the Dirac comb (and closing our eyes on the mathematical proofs) we end up with the same result obtained using the distribution theory. But this is not always the case! Actually, is rarely the case!

  So, in order not to get into mathematical troubles

  "Everything Should Be Made as Simple as Possible, But Not Simpler" (A. Einstein),

  because

  "No notice is taken of a little evil, but when it increases it strikes the eye" (Aristotle).

Markov Chains

Reference: P. Bremaud, "Markov Chains ", Springer Verlag.

Markov Chains

Markov Chain

Let $\{X[n]\}_{n\in\mathbb{Z}}$ be a discrete time stochastic process with values in a countable set $\mathcal{D}$. Denote as i, j, k,... the elements of $\mathcal{D}$. We shall refer to these elements as states. Therefore, if $X[n] = i$, $i \in \mathcal{D}$, the process is said to be in state i at time n (or to visit state i at time n).

If for all integers $n > 0$ and all states $j, i, i_{n-1}, \ldots, i_0 \in \mathcal{D}$

$$\mathrm{P}\left(X[n+1] = j \mid X[n] = i, X[n-1] = i_{n-1}, \ldots, X[0] = i_0\right) = \mathrm{P}\left(X[n+1] = j \mid X[n] = i\right) ,$$

(whenever both sides are well-defined), then $\{X[n]\}_{n\in\mathbb{Z}}$ is called **Markov chain**.

Homogeneous Markov Chain

If in addition $\mathrm{P}\left(X[n] = j \mid X[n-1] = i\right)$ is independent of n, *i.e.*, $\mathrm{P}\left(X[n] = j \mid X[n-1] = i\right) = p_{ij}$, then $\{X[n]\}_{n\in\mathbb{Z}}$ is called **homogenous Markov chain**.

$\boldsymbol{P} = \{p_{ij}\}_{i,j\in\mathcal{D}}$, where

$$p_{ij} = P(X[n+1] = j \mid X[n] = i) , \quad i, j \in \mathcal{D} ,$$

is the **transition matrix** of the homogenous Markov chain.

Notice that

$$0 \le p_{ij} \le 1 , \quad \sum_{k\in\mathcal{D}} p_{ik} = 1 , \quad \text{for all } i \in \mathcal{D} .$$

Markov Chains

Initial State & Initial Distribution

Let $\{X[n]\}_{n\in\mathbb{Z}}$ be a homogeneous Markov chain with transition matrix $\boldsymbol{P}$. The random variable $X[0]$ is called **initial state**, and its probability distribution $\boldsymbol{\nu}(n) = \{\nu_i(0)\}_{i\in\mathcal{D}}$, where $\nu_i(0) = \mathrm{P}\left(X[0] = i\right)$, $i \in \mathcal{D}$, is the **initial distribution**.

Applying Bayes's rule and considering the Markov property, we have

$$P(X[0] = i_0, X[1] = i_1, \ldots, X[k] = i_k) = \\ P(X[0] = i_0)P(X[1] = i_1 \mid X[0] = i_0) \ldots P(X[k] = i_k \mid X[k-1] = i_{k-1}, \ldots, X[0] = i_0)\,.$$

For homogeneous Markov chains and the definition of the transition matrix, we have

$$P(X[0] = i_0, X[1] = i_1, \ldots, X[k] = i_k) = \nu_{i_0}(0)p_{i_0 i_1} \ldots p_{i_{k-1} i_k}\,.$$

The above equation constitute the **probability law** or **probability distribution** of the homogeneous Markov chain, which is determined by the transition matrix and the initial distribution.

Distribution at Time n

The distribution at time n of the chain is the vector $\boldsymbol{\nu}(n) = \{\nu_i(n)\}_{i\in\mathcal{D}}$, where $\nu_i(n) = \mathrm{P}\left(X[n] = i\right)$.

In particular $\nu_j(n+1) = \sum_{i\in\mathcal{D}} \nu_i(n)p_{ij}$ (Bayes's rule of exclusive and exhaustive causes), that is, in matrix form

$$\boldsymbol{\nu}(n+1)^t = \boldsymbol{\nu}(n)^t \boldsymbol{P}\,.$$

Iteration of this equality yields

$$\boldsymbol{\nu}(n)^t = \boldsymbol{\nu}(0)^t \boldsymbol{P}^n\,.$$

Markov Chains

Stationary Distribution of a Homogenous Markov Chain

A probability distribution $\boldsymbol{\pi} = \{\pi_i\}_{i \in \mathcal{D}}$ satisfying

$$\boldsymbol{\pi}^t = \boldsymbol{\pi}^t \boldsymbol{P}$$

is called a **stationary distribution** of the homogeneous Markov chain with transition matrix $\boldsymbol{P}$.

Iteration of the above equation (multiplying both sides on the right by $\boldsymbol{P}$) gives

$$\boldsymbol{\pi}^t = \boldsymbol{\pi}^t \boldsymbol{P}^n \,, \quad \forall\, n \geq 0 \,.$$

Consequently, given the expression of the distribution $\boldsymbol{\nu}(n)$ at time n, if $\boldsymbol{\nu}(0) = \boldsymbol{\pi}$, then $\boldsymbol{\nu}(n) = \boldsymbol{\pi}$ for all $n \geq 0$: A chain started with a stationary distribution keeps the same distribution forever.

In addition, the law of the chain

$$\begin{aligned} &P(X[n] = i_0, X[n+1] = i_1, \ldots, X[n+k] = i_k) \\ &\qquad = P(X[n] = i_0) p_{i_0 i_1} \cdots p_{i_{k-1} i_k} = \pi_{i_0} p_{i_0 i_1} \cdots p_{i_{k-1} i_k} \,, \end{aligned}$$

does not depend on n. In this sense the chain is **stationary** (one also says that the chain is in a **stationary regime**, or in **equilibrium**, or in **steady state**).

A chain started with a stationary distribution is stationary!

Numerical Analysis

Errors

Error Propagation in the Solution of a Linear System

Consider a vector induced norm of a matrix, *i.e.*, $\|\boldsymbol{A}\| = \max_{\boldsymbol{y}} \frac{\|\boldsymbol{A}\boldsymbol{y}\|}{\|\boldsymbol{y}\|}$, where $\|\boldsymbol{I}\| = 1$, which is consistent with vector norms, *i.e.*, $\|\boldsymbol{A}\boldsymbol{x}\| \leq \|\boldsymbol{A}\|\|\boldsymbol{x}\|$. For instance $\|\boldsymbol{A}\|_2 = \sigma_{\max}(\boldsymbol{A})$. Let

$$K(\boldsymbol{A}) = \|\boldsymbol{A}\|\|\boldsymbol{A}^{-1}\|,$$

be the conditioning number of the matrix $\boldsymbol{A}$, where $K(\boldsymbol{A}) = \|\boldsymbol{A}\|\|\boldsymbol{A}^{-1}\| \geq \|\boldsymbol{A}\boldsymbol{A}^{-1}\| = \|\boldsymbol{I}\| = 1$. Notice that by taking $\|\boldsymbol{A}\| = \|\boldsymbol{A}\|_2$ we have the standard definition $K(\boldsymbol{A}) = \sigma_{\max}(\boldsymbol{A})/\sigma_{\min}(\boldsymbol{A})$.

Theorem (Relative Error Range)
Let $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ be a linear system to be solved in $\boldsymbol{x}$. Call $\boldsymbol{x}_0$ the exact solution of the system, *i.e.* $\boldsymbol{A}\boldsymbol{x}_0 = \boldsymbol{b}$ and $\widehat{\boldsymbol{x}}$ the computed solution of the system. Then

$$\frac{1}{K(\boldsymbol{A})} \frac{\|\boldsymbol{A}\widehat{\boldsymbol{x}} - \boldsymbol{b}\|}{\|\boldsymbol{b}\|} \leq \frac{\|\widehat{\boldsymbol{x}} - \boldsymbol{x}_0\|}{\|\boldsymbol{x}_0\|} \leq K(\boldsymbol{A}) \frac{\|\boldsymbol{A}\widehat{\boldsymbol{x}} - \boldsymbol{b}\|}{\|\boldsymbol{b}\|}.$$

Notice that $\|\boldsymbol{A}\widehat{\boldsymbol{x}} - \boldsymbol{b}\|/\|\boldsymbol{b}\|$ represents the relative error of the given data, while $\|\widehat{\boldsymbol{x}} - \boldsymbol{x}_0\|/\|\boldsymbol{x}_0\|$ is the relative error of the solution. The theorem states that the relative error of the solution is bounded by the relative error of the given data multiplied by $1/K(\boldsymbol{A})$ on the lower bound, and $K(\boldsymbol{A})$ on the higher bound. For a low condition number ($K(\boldsymbol{A}) \approx 1$), the relative errors of the given data and the solution have the same order of magnitude. For a high condition number, the relative error of the solution can be of a much higher order of magnitude with respect to the relative error of the given data.

Errors

Error Propagation in the Solution of a Linear System

Proof

1) $\boldsymbol{A}\left(\widehat{\boldsymbol{x}}-\boldsymbol{x}_0\right)=\boldsymbol{A}\widehat{\boldsymbol{x}}-\boldsymbol{b}$, then:
   - $\left(\widehat{\boldsymbol{x}}-\boldsymbol{x}_0\right)=\boldsymbol{A}^{-1}\left(\boldsymbol{A}\widehat{\boldsymbol{x}}-\boldsymbol{b}\right)$ and $\left\|\widehat{\boldsymbol{x}}-\boldsymbol{x}_0\right\| \leq \left\|\boldsymbol{A}^{-1}\right\|\left\|\boldsymbol{A}\widehat{\boldsymbol{x}}-\boldsymbol{b}\right\|$;
   - $\|\boldsymbol{A}\|\left\|\widehat{\boldsymbol{x}}-\boldsymbol{x}_0\right\| \geq \left\|\boldsymbol{A}\widehat{\boldsymbol{x}}-\boldsymbol{b}\right\|$ and $\left\|\widehat{\boldsymbol{x}}-\boldsymbol{x}_0\right\| \geq \frac{\left\|\boldsymbol{A}\widehat{\boldsymbol{x}}-\boldsymbol{b}\right\|}{\|\boldsymbol{A}\|}$;

   implying $\frac{\left\|\boldsymbol{A}\widehat{\boldsymbol{x}}-\boldsymbol{b}\right\|}{\|\boldsymbol{A}\|} \leq \left\|\widehat{\boldsymbol{x}}-\boldsymbol{x}_0\right\| \leq \left\|\boldsymbol{A}^{-1}\right\|\left\|\boldsymbol{A}\widehat{\boldsymbol{x}}-\boldsymbol{b}\right\|$.

2) $\boldsymbol{A}\boldsymbol{x}_0=\boldsymbol{b}$, then:
   - $\boldsymbol{x}_0=\boldsymbol{A}^{-1}\boldsymbol{b}$ and $\|\boldsymbol{x}_0\| \leq \left\|\boldsymbol{A}^{-1}\right\|\|\boldsymbol{b}\|$;
   - $\|\boldsymbol{A}\|\|\boldsymbol{x}_0\| \geq \|\boldsymbol{b}\|$ and $\|\boldsymbol{x}_0\| \geq \frac{\|\boldsymbol{b}\|}{\|\boldsymbol{A}\|}$;

Combining $\frac{\left\|\boldsymbol{A}\widehat{\boldsymbol{x}}-\boldsymbol{b}\right\|}{\|\boldsymbol{A}\|} \leq \left\|\widehat{\boldsymbol{x}}-\boldsymbol{x}_0\right\| \leq \left\|\boldsymbol{A}^{-1}\right\|\left\|\boldsymbol{A}\widehat{\boldsymbol{x}}-\boldsymbol{b}\right\|$, $\|\boldsymbol{x}_0\| \leq \left\|\boldsymbol{A}^{-1}\right\|\|\boldsymbol{b}\|$, and $\|\boldsymbol{x}_0\| \geq \frac{\|\boldsymbol{b}\|}{\|\boldsymbol{A}\|}$ leads to the theorem inequality.