Quantization noise: theory and techniques

Giovanni Cerretani

September 25th, 2015

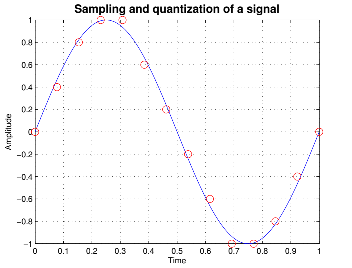
Università di Pisa

Analog-to-digital conversion

Analog signal $\xrightarrow{\textbf{A/D conversion}}$ **Digital signal**

continuous-time $\xrightarrow{\textit{sampling}}$ discrete-time

continuous-amplitude $\xrightarrow{\textit{quantization}}$ discrete-amplitude



Sampling and quantization:

- are mathematically commutable;
- degrade the quality of a signal.

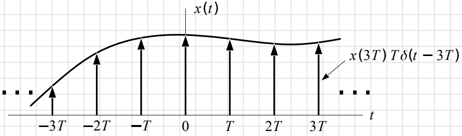
Sampling: time domain

Let's suppose our continuous signal is represented by the time function

$$\boxed{x(t)}$$

and define the Linvill's **impulse carrier**

$$c(t) = \sum_{k=-\infty}^{+\infty} T \cdot \delta(t - kT)$$



If we take samples of the signal at uniform time interval T, then we can express the sampled signal $x_s(t)$ as

$$\boxed{x_s(t) = x(t) \cdot c(t)}$$

Notes

- The factor T in $c(t)$ preserves the integral in the sense that the area of the samples of $x(t)$ approximately equals the area of $x(t)$;

Sampling: frequency domain

In the frequency domain, if we define $X\left(j\omega\right)$ as the Fourier transform of $x\left(t\right)$

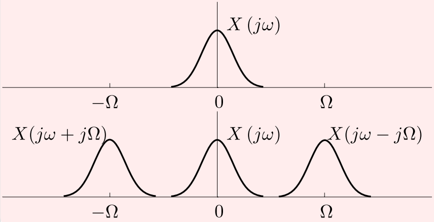
$$\mathcal{F}\left[x\left(t\right)\right] = X\left(j\omega\right)$$

then it can be proved that the Fourier transform of $x_s\left(t\right) = x\left(t\right) \cdot c\left(t\right)$ is

$$\mathcal{F}\left[x_s\left(t\right)\right] = X_s\left(j\omega\right) = \sum_{n=-\infty}^{+\infty} X\left(j\omega + jn\Omega\right)$$

where $\Omega = \frac{2\pi}{T}$ is the sampling radian frequency.

Aliasing



$X_s\left(j\omega\right)$ is a periodic function, sum of an infinite number of displaced replicas of $X\left(j\omega\right)$.
The "aliasing" effect occurs when the original spectrum overlaps its replicas.

Sampling theorem

Now should be pretty easy to understand the renowned theorem:

Nyquist–Shannon sampling theorem

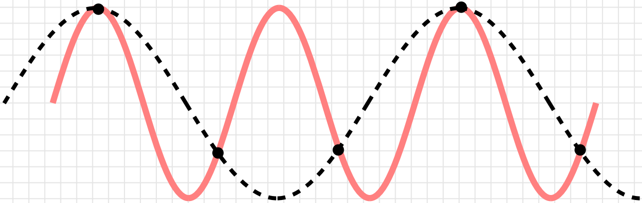
In the sampling process, if Ω is high enough so that

$$X(j\omega) = 0 \text{ for } |\omega| > \frac{\Omega}{2}$$

then

- there is no aliasing;
- $x(t)$ is perfectly recoverable from its samples.

Otherwise, "aliasing" occurs.



PDF and CF

Before to proceed with the quantization, it's useful to introduce these two tools:

Probability Density Function

The Probability Density Function (**PDF**) $f_x(x)$ describes the relative likelihood for a (random) variable x to take on a given value.
Of course

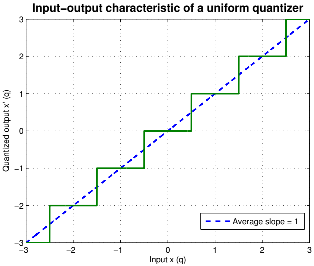
$$\int_{-\infty}^{+\infty} f_x(x)\, dx = 1$$

Characteristic Function

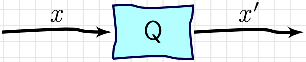
The Characteristic Function (**CF**) $\Phi_x(u)$ is defined as the Fourier Transform of the PDF:

$$\mathcal{F}[f_x(x)] = \Phi_x(u) = \int_{-\infty}^{+\infty} f_x(x)\, e^{jux}\, dx$$

The unform quantizer



It is convenient to define the **quantizer Q** as a nonlinear operator having this input-output staircase relation (rounding to the nearest integer).



The uniform quantizer output x' is a single-valued function of the input x, and the quantizer has an "average gain" of unity. The basic unit of quantization is designated by q.

The output of the quantizer differs from the input by a quantity known as the **roundoff error**

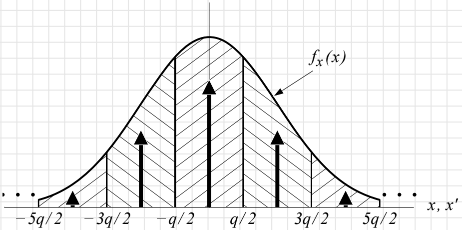
$$\nu = x' - x$$

that is also called the **quantization noise** because, as we'll see, in most cases it can be considered as a white noise term added to the quantizer input.

The quantizer output

Let's consider for example a uniform quantizer and a normally distributed white noise x with PDF $f_x(x)$.

At the output of the quantizer, the PDF $f_{x'}(x)$ of x' consists of a series of Dirac impulses, uniformly spaced along the amplitude axis:



The area under $f_x(x)$ within each quantization box is compressed into a Dirac delta in $f_{x'}(x)$ by the quantizer, as it maps any value in the interval $\left[mq - \frac{q}{2}, mq + \frac{q}{2}\right]$ to the central value mq.

What does it means?

The formation of $f_{x'}(x)$ from $f_x(x)$ is a sampling process: we can see a quantizer as a sampler of the PDF of our signal. Actually, this is an **area sampling**, and differs from the conventional sampling. Let's see how.

Area sampling

Analytically, $f_{x'}(x)$ can be expressed as

$$\begin{aligned} f_{x'}(x) &= \ldots + \delta(x+q) \int_{-\frac{3q}{2}}^{-\frac{q}{2}} f_x(x) + \delta(x) \int_{-\frac{q}{2}}^{+\frac{q}{2}} f_x(x) + \ldots \\ &= \sum_{m=-\infty}^{+\infty} \delta(x - mq) \int_{mq-\frac{q}{2}}^{mq+\frac{q}{2}} f_x(x) \end{aligned}$$

It can be easily proved that this summation is equal to

$$\boxed{f_{x'}(x) = (f_n(x) \star f_x(x)) \cdot c(x)}$$

where:

$$f_n(x) = \begin{cases} q^{-1} & \text{if } x \in [-q/2, +q/2] \\ 0 & \text{elsewhere} \end{cases} \quad \text{and} \quad c(x) = \sum_{m=-\infty}^{+\infty} q \cdot \delta(x - mq)$$

- $f_n(x)$ is the **rectangular pulse function**;
- $c(x)$ is the Linvill's **impulse carrier** scaled by q

Area sampling

Analytically, $f_{x'}(x)$ can be expressed as

$$\begin{aligned} f_{x'}(x) &= \ldots + \delta(x+q)\int_{-\frac{3q}{2}}^{-\frac{q}{2}} f_x(x) + \delta(x)\int_{-\frac{q}{2}}^{+\frac{q}{2}} f_x(x) + \ldots \\ &= \sum_{m=-\infty}^{+\infty} \delta(x-mq)\int_{mq-\frac{q}{2}}^{mq+\frac{q}{2}} f_x(x) \end{aligned}$$

It can be easily proved that this summation is equal to

$$\boxed{f_{x'}(x) = \left(f_n(x) \star f_x(x)\right) \cdot c(x)}$$

where:

$$f_n(x) = \begin{cases} q^{-1} & \text{if } x \in [-q/2, +q/2] \\ 0 & \text{elsewhere} \end{cases} \quad \text{and} \quad c(x) = \sum_{m=-\infty}^{+\infty} q \cdot \delta(x-mq)$$

In conclusion

The quantization is first convolution of the PDF with a rectangular pulse function, then conventional sampling with "period" q. This is the **area sampling**.

Quantization theorem

It's useful to define a "quantization radian frequency" Ψ defined as

$$\Psi = \frac{2\pi}{q}$$

analogous to the sampling radian frequency $\Omega = \frac{2\pi}{T}$, that express **how small is the quantization step**.

Widrow quantization theorem

Analogous to the Nyquist theorem, it states that if the Fourier transform of the PDF of x, i.e. the CF, is

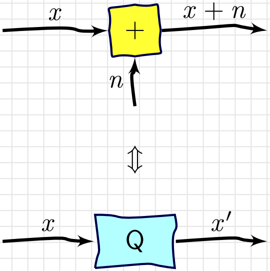
$$\Phi_x(u) = 0 \text{ for } |u| > \frac{\Psi}{2}$$

then:

- the replicas contained in $\Phi_{x'}(u)$ will not overlap (i.e. no "aliasing");
- $f_x(x)$ is perfectly recoverable from $f_{x'}(x)$.

The quantization grain size q must be made small enough: making q smaller raises the "quantization frequency," spreads the replicas, and tends to reduce their overlap.

Pseudo Quantization Noise



Let's define an independent noise n, uniformly distributed between $\pm q/2$. Its PDF is the rectangular pulse function f_n.

If we add it to our variable x, the result is $x+n$ whose PDF is the convolution

$$f_{x+n}(x) = f_n(x) \star f_x(x)$$

Now it's easy to see that there is a fundamental relation between the PDF of $x+n$ and that of the quantizer output $f_{x'}(x)$

$$f_{x'}(x) = \underbrace{\left(f_n(x) \star f_x(x)\right)}_{f_{x+n}(x)} \cdot c(x)$$

Validation of the PQN noise

It can be proved that, when Widrow quantization theorem applies, the quantization noise can be analyzed as a uniformly distributed noise with zero mean and standard deviation $\sigma = q/\sqrt{12}$.

Some examples

#1: a constant signal

Let our signal $x(t)$ to be constant:

$$x(t) = k \Longrightarrow f_x(x) = \delta(x-k)$$

Then the CF is

$$\Phi_x(u) = \int_{-\infty}^{+\infty} \delta(x-k)\, e^{jux}\, dx = e^{juk}$$

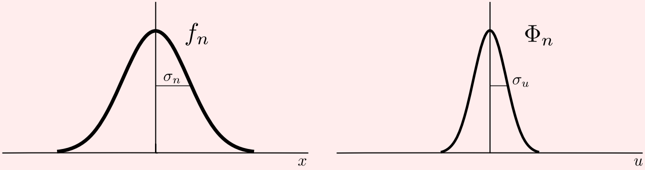
that means

$$\boxed{|\Phi_x(u)| = 1 \qquad \forall u}$$

The quantization theorem obviously does not apply.
It's easy to see that $f_{x'}(x)$ would be the same for all the values of $k \in \left[mq - \frac{q}{2}, mq + \frac{q}{2}\right]$. So, the PDF of x is not recoverable from the quantizer output.

Some examples

#2: Gaussian distributed noise



The CF of a Gaussian noise distributed n with standard deviation σ_n is a Gaussian with standard deviation $\sigma_u = 1/\sigma_n$. Of course this CF is not bandlimited. However, if σ_n is big enough so that

$$\sigma_u = \frac{1}{\sigma_n} \ll \frac{\Psi}{2}$$

that means also

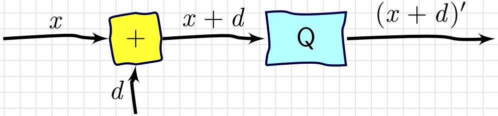
$$\boxed{\sigma_n \gg q}$$

then the quantization theorem is almost satisfied.
This means that if the signal is spread over at least 3 or 4 quantization boxes, then the PQN model applies decently.

Dither

Definition

Dither is an intentionally applied form of noise used to alleviate the effects of nonlinearity and quantization in the A/D and D/A conversions. It is applied at the input of the quantizer.



It is very simple: if the CF of the dither $\Phi_d(u)$ alone is bandlimited and satisfies the QT, then the CF of the quantizer input $x + d$ will be bandlimited and will satisfy the QT, being

$$\boxed{\Phi_{x+d}(u) = \Phi_x(u) \cdot \Phi_d(u)}$$

even if $\Phi_x(u)$ is not bandlimited.

The "anti-alias" effect

The addition of a dither signal d to the input x works as "anti-alias filtering" for the quantization.

Dithering applied to an image

Note

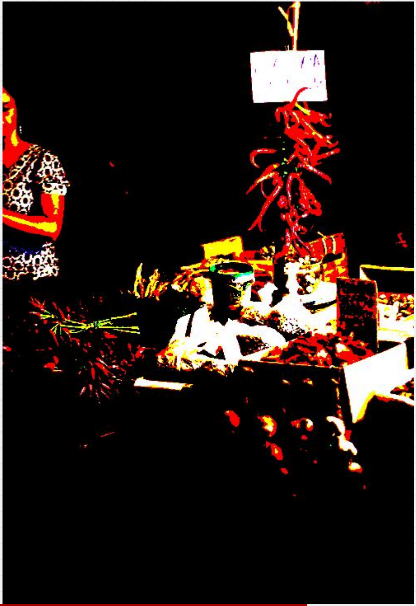
The picture has been taken with a digital camera, and actually is already a quantization of the “real world”. The digital camera is a quantizer that produces an output with 8 bits per color. For our purpose, we can consider it almost analog.
The photo is also sampled: it contains 500×750 pixels.

Suppose to quantize the picture with a uniform quantizer with 1 bit per color:

- red
- green
- blue

So, we have one 1-bit quantizer per color channel.

Dithering applied to an image



The output is very ugly!

Dithering applied to an image

The output is very ugly!

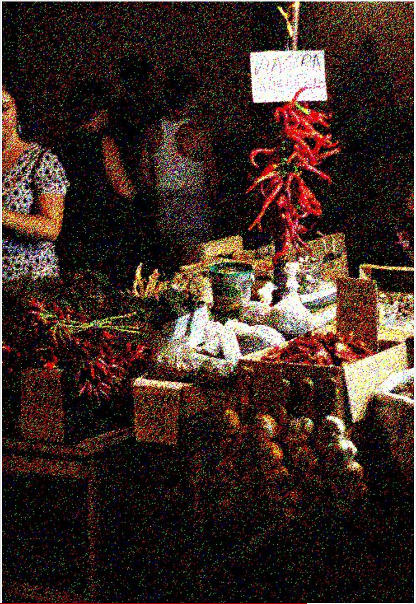
- come back to the original, “analog” image

Dithering applied to an image

The output is very ugly!

- come back to the original, "analog" image
- add a uniform noise in range $[-q/2, q/2]$, high-pass filtered.

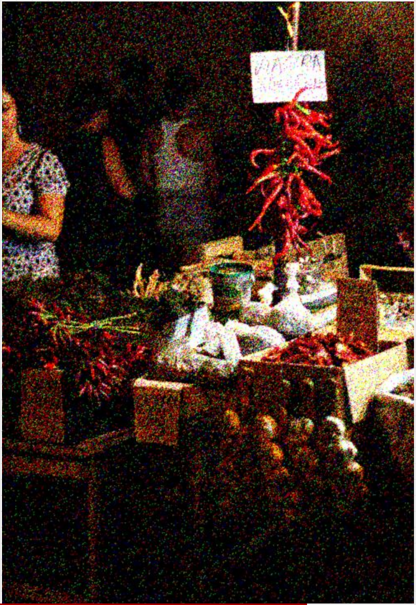
Dithering applied to an image



The output is very ugly!

- come back to the original, “analog” image
- add a uniform noise in range $[-q/2, q/2]$, high-pass filtered.
- 1-bit quantization per color

Dithering applied to an image

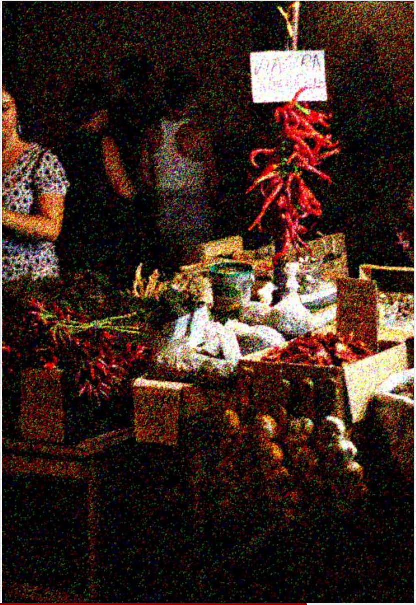


The output is very ugly!

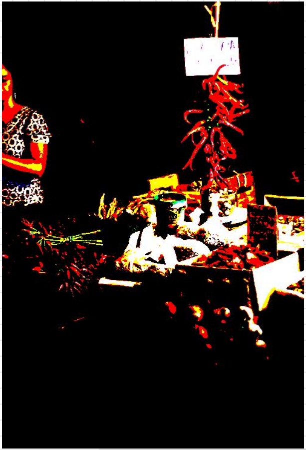
- come back to the original, "analog" image
- add a uniform noise in range $[-q/2, q/2]$, high-pass filtered.
- 1-bit quantization per color
- apply a low-pass filter

Done...

Dithering applied to an image



... and it's much better than without the dither!



Conclusion and references

- It is important to choose the number of bits of the ADC of our experiment to keep under control the quantization noise.
- Floating-point quantization has not been presented here, but it has very interesting features due to its strange nonlinearity.

For further reading:

- B. Widrow and I. Kollár, *Quantization Noise: Roundoff Error in Digital Computation, Signal Processing, Control, and Communications,* Cambridge University Press (2008)