Error Analysis in Digital Filters

summary

Prof. Ioannis Pitas
Aristotle University of Thessaloniki
pitas@csd.auth.gr
www.aiia.csd.auth.gr

Version 1.2.1

Artificial Intelligence &
Information Analysis Lab

Error Analysis in Digital Filters

- **Digital Filters**
- Quantization errors
- Statistic Error Analysis in FIR filters
- Statistic Error Analysis in IIR filters

Digital Filters

- A *digital filter* is a system that stores and processes data in a discrete way. Digital hardware and software *components* (internal and external) used to transform *data* into a *digital solution*.
- Digital Filters can be implied by using fixed point arithmetic or floating point arithmetic. The differences will be discussed later on.

Fixed vs floating point arithmetic

The term ‘fixed point’ refers to the corresponding manner in which numbers are represented, with a fixed number of digits after, and sometimes before, the decimal point.

With floating-point representation, the placement of the decimal point can ‘float’ relative to the significant digits of the number. A floating point number does not reserve a specific number of bits for the integer part or the fractional part. Instead it reserves a certain number of bits for the number (called the *mantissa* or *significand*) and a certain number of bits to say *where* within that number the decimal place sits

Quantization

- The deformation of the original signal can be studied as the need of quantization of the original signal so that it can be displayed by a limited number of digits.

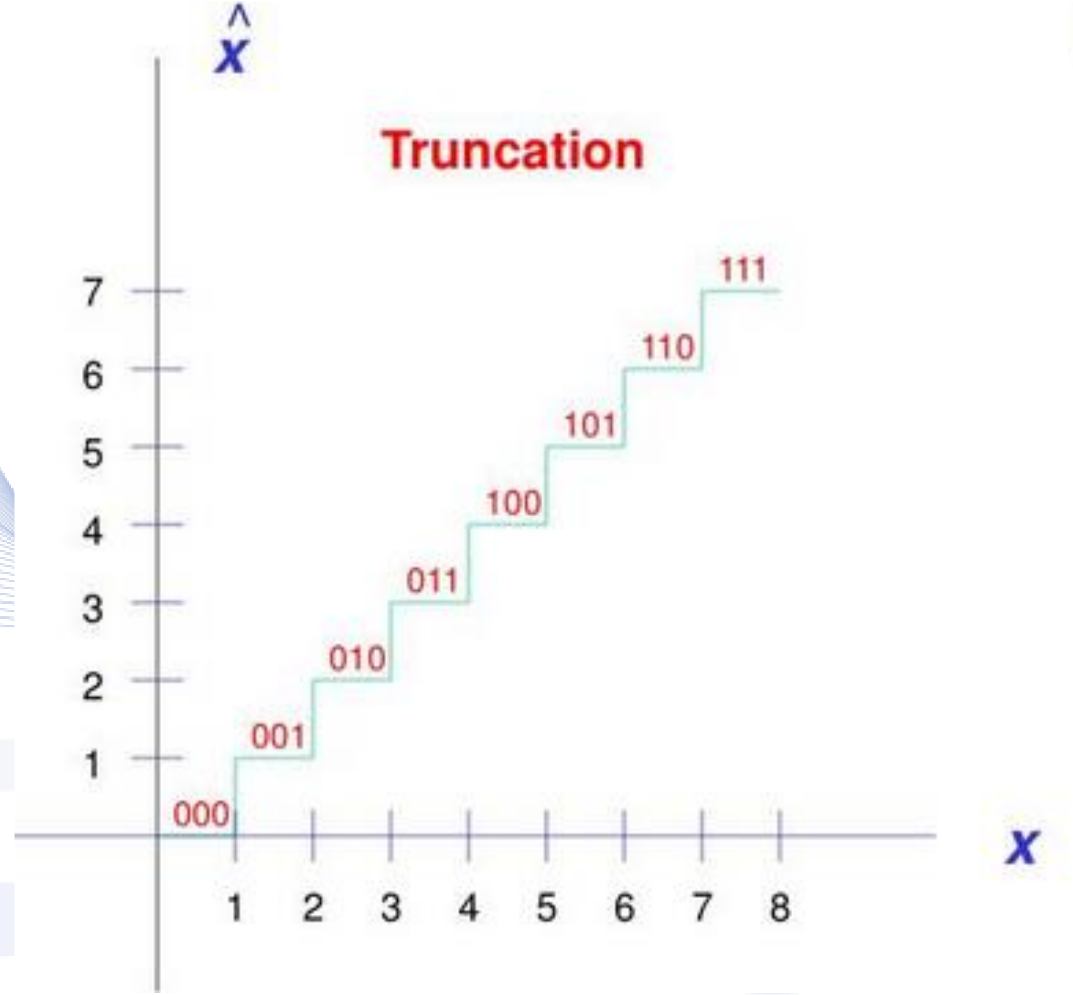
It is applied either by:

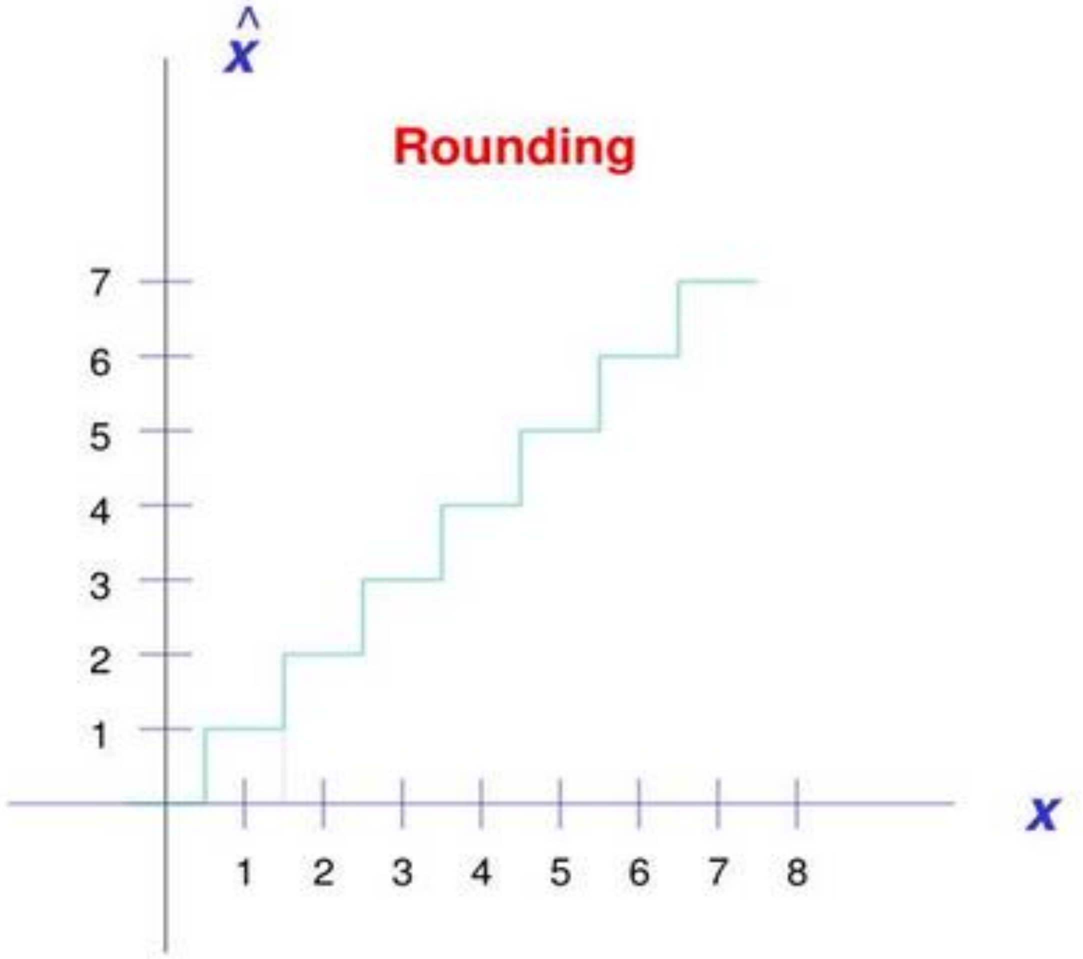
a. Rounding (most preferred as it returns average error value 0)

b. Truncation

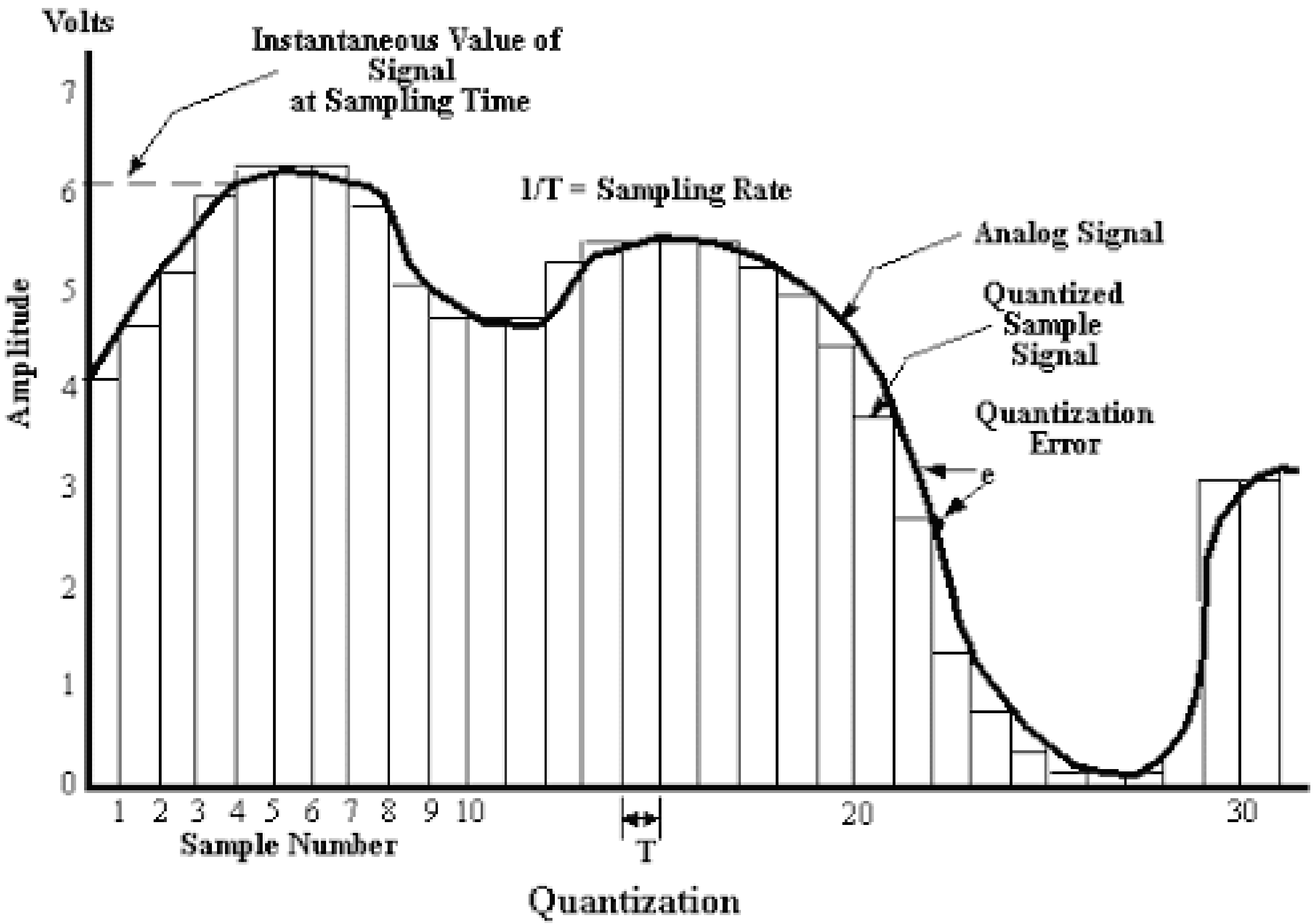
An application example follows.

Truncation & rounding





Quantization and $e(n)$



Non-Uniform quantization

Nonuniform Quantization is the type of quantization in which the quantization levels are unequal.

- The quantization step size varies with the amplitude of the input signal
- SNR ration can be maintained constant
- Reduces some quantization error

Limit cycle oscillation

By definition: A limit cycle oscillation is a periodic low-level oscillatory disturbance (useless signal) that may exist in an otherwise stable filter.

The equation that describes the ideal first order IIR filter is:

$$y(n) = a \cdot y(n-1) + x(n)$$

And for the non-linear system:

$$w(n) = Q[a \cdot w(n-1)] + x(n)$$

It occurs due to non-linearities that derive from inherent quantization in the system. It is observed only in IIR filters.

Basically, in a stable IIR filter excited by finite sequence, the output will ,eventually , be zero. However, due to non-linearities, the issue of limit cycle will keep some oscillations going in the output.

There are two types:

- Zero limit cycle oscillations,
- Overflow limit cycle oscillations

Dead bands

- During the limit cycle oscillations, the output of the filter oscillates between a finite positive and negative value. This range of values is called the Dead band of the filter. When we are in the dead band the system acts as is it’s poles are on a unit cycle (effective poles).

Signal To Noise Ratio

SNR reflects the relationship between signal strength and error Quantization and $e(n)$ strength due to rounding/truncation.

$$\frac{\sigma_x^2}{\sigma_e^2} = (12 \cdot 2^{2b}) \cdot \sigma_x^2$$

And in decibel:

$$SNR = 10\, log_{10}(\frac{\sigma_x^2}{\sigma_e^2})\ \text{dB}$$

FIR Filters

In signal processing, a **finite impulse response** (**FIR**) **filter** is a filter whose impulse response (or response to any finite length input) is of finite duration, because it settles to zero in finite time.

Statistic Error Analysis in FIR- in fixed point arithmetic

FIR filters do not show the Limit cycle oscillation effect.

$$y(n) = \sum_{k=0}^{N-1} h(k)(n-k),$$

where $h(k)$ is unit sample response

IIR Filter

- IIR filters are implemented using the recursion described in this chapter to produce filters with sharp frequency cutoff characteristics. This can be achieved using a relatively small number of coefficients.
- These filters have nonlinear phase behavior and because of this characteristic, phase response must be of little concern to the DSP programmer. Amplitude response must be the primary concern when using these filters.

IIR Filter

- This nonlinear phase characteristic makes IIR filters a poor choice for applications such as speech processing and stereo sound systems.

Statistic Error Analysis in IIR

- When the registers are in the form of fixed arithmetic (positive and negative numbers) of *b* bits size, then for the rounding error it is valid that:

$$\frac{-2^{-b}}{2} < e(n) < \frac{2^{-b}}{2}$$

Statistic Error Analysis in IIR- in fixed point arithmetic

Error due to register overflow:

$$e(n) = Q[a \cdot w(n-1)]a \cdot w(n-1)$$

For $e(n)$ it can be valid that (if everything is random):

a) $e(n)$ is white noise

b) $e(n)$ is uniformly distributed on the quantization space.

c) $e(n)$ is unrelated to x(n) signal and $a \cdot w(n-1)$

CPUs vs GPUs

- The main difference between CPU and GPU architecture is that a CPU is designed to handle a wide-range of tasks quickly (as measured by CPU clock speed), but are limited in the concurrency of tasks that can be running. A GPU is designed to quickly render high-resolution images and video concurrently. Because GPUs can perform parallel operations on multiple sets of data, they are also commonly used for non-graphical tasks such as machine learning and scientific computation. Designed with thousands of processor cores running simultaneously, GPUs enable massive parallelism where each core is focused on making efficient calculations.

Arithmetic used in GPUs

- FP16- Half precision

Half precision (sometimes called FP16) is a binary floating point computer number format that occupies 16 bits (two bytes in modern computers) in computer memory. They can express values in the range ±65,504, with precision up to 0.0000000596046.

Arithmetic used in GPUs

- FP32

FP32 is a number format, that uses 32 bit (4 byte) per number. You basically have one signbit. Then you have two to the power of an 8 bit number (-127 to 127) and then you have 24 bits for some fraction.

Bibliography

[OPP2013] A. Oppenheim, A. Willsky, Signals and Systems, Pearson New International, 2013.

[MIT1997] S. K. Mitra, Digital Signal Processing, McGraw-Hill, 1997.

[OPP1999] A.V. Oppenheim, Discrete-time signal processing, Pearson Education India, 1999.

[HAY2007] S. Haykin, B. Van Veen, Signals and systems, John Wiley, 2007.

[LAT2005] B. P. Lathi, Linear Systems and Signals, Oxford University Press, 2005.

[HWE2013] H. Hwei. Schaum's Outline of Signals and Systems, McGraw-Hill, 2013.

[MCC2003] J. McClellan, R. W. Schafer, and M. A. Yoder, Signal Processing, Pearson Education Prentice Hall, 2003.

Bibliography

[PHI2008] C. L. Phillips, J. M. Parr, and E. A. Riskin, Signals, Systems, and Transforms, Pearson Education, 2008.

[PRO2007] J.G. Proakis, D.G. Manolakis, Digital signal processing. PHI Publication, 2007.

[DUT2009] T. Dutoit and F. Marques, Applied Signal Processing. A MATLAB-Based Proof of Concept. New York, N.Y.: Springer, 2009

Bibliography

[PIT2000] I. Pitas, “Digital Image Processing Algorithms and Applications”, J. Wiley, 2000.

[PIT2021] I. Pitas, “Computer vision”, Createspace/Amazon, in press.

[PIT2017] I. Pitas, “Digital video processing and analysis” , China Machine Press, 2017 (in Chinese).

[PIT2013] I. Pitas, “Digital Video and Television” , Createspace/Amazon, 2013.

[NIK2000] N. Nikolaidis and I. Pitas, “3D Image Processing Algorithms”, J. Wiley, 2000.

Q & A

Thank you very much for your attention!

More material in
http://icarus.csd.auth.gr/cvml-web-lecture-series/

Contact: Prof. I. Pitas
pitas@csd.auth.gr