SKA Project Series
Quantization effects and nonlinearities in multi-bit correlation

S. Chiarucci[1], G. Comoretto[1]

[1]INAF - Osservatorio Astrofisico di Arcetri

Arcetri Technical Report N° 4/2018
25-aug-2018

Abstract

An astronomic correlator analyses a signal that has a Gaussian bivariate statistics, quantized with a limited number of bits. The correlator response to the quantized signal is linear only in a limited range of the signal amplitude, but the relation between the quantized and unquantized correlation is well known, and most present day correlators correct for this effects. In large interferometers, with hundreds of antennas and thousands of spectral channels, it can be advantageous to limit the signal amplitude to the linear region, and completely avoid the quantization correction. Nonlinearities have been computed for quantization schemes from 4 to 9 bits. When 4 or 5 bits are used, keeping the signal in the linear region causes a degradation in the quantization efficiency. Autocorrelation is always affected when using quantization with less than 8 bits, and must be corrected. The effect of rounding in requantization is also considered.

1 Introduction

An astronomic correlator computes the correlation product $< x_1 x_2 >$ of two signals x_1 and x_2, e.g. from two antennas in an interferometer, or from two different delay values in an autocorrelation spectrometer. These signals are stochastic in nature, with a Gaussian bivariate distribution. Usually they have zero mean, and their statistics is characterized by their variance σ_1^2, σ_2^2 and correlation ρ. The correlation product in this case is $r(\rho) =< x_1 x_2 >= \rho\, \sigma_1 \sigma_2$.

Due to its stochastic nature, the astronomic signal can be quantized in a very crude way without losing appreciable information and dynamic range. The relationship between the correlation products of the quantized and unquantized signals has been extensively studied and, in the simple case in which the quantization occurs just before the product, the original unquantized correlation product can be retrieved with very good accuracy from the quantized one. The procedure is usually called *Van Vleck correction*, from the 1-bit case [10].

The situation is more complex when multiple quantizations occur at various stages of the signal processing chain. For example in a FX (Fourier transform followed by multiplication) correlator the quantized signal is Fourier transformed, and the result in each frequency channel is then re-quantized and correlated. When a digital receiver is employed, the down-converted signal is also re-quantized. In general a modern digital signal processing system may use three or more requantizations. It is possible to correct for the quantizations others than the last, but the correction can be mathematically accurate only if the whole spectral content of the signal to be corrected is known, that is not the case for example in digital receivers. Approximate correction techniques have been used in this case, for example in the ALMA hybrid FXF correlator [1].

In large interferometers the sheer number of correlation products makes a correction problematic. In particular in the Square Kilometre Array (SKA) Low telescope 512 dual polarization antennas are cross correlated in a FX correlator employing 16384 spectral channels. The total number of (real) correlation products is then 2^{34} (16 million autocorrelations and $\simeq 8$ billion complex cross correlations), that should be corrected every few seconds.

It would therefore be useful to avoid completely the quantization correction, if at all possible. This can be done if a sufficient number of discrete levels are used in the quantization process, and in practice no corrections are used for quantization with 8 or more bits, and even with just 5 or 6 bits. Where high dynamic range is required, however, a quantization scheme of 8 or less bits may introduce significant nonlinearities, and these depend heavily on the signal level. The *away from zero* rounding scheme, also frequently employed to avoid rounding DC biases, may also introduce significant nonlinearities. In this work we will analyze these effects in order to derive limits on the signal amplitude as a function of the quantization scheme and required nonlinearity levels.

In some limiting cases it could be useful to use a simplified correction scheme, requiring less computing resources. Some indications for such a scheme are also outlined.

2 Mathematical definitions

We will consider two stochastic signals, x_1, x_2, that are described by a joint bivariate normal probability function:

$$\begin{aligned} P(x_1, x_2, \rho) &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)} \left(z_1^2 + z_2^2 - 2\rho z_1 z_2 \right) \right) \\ z_1 &= \frac{x_1 - \mu_1}{\sigma_1}, \quad z_2 = \frac{x_2 - \mu_2}{\sigma_2} \end{aligned} \tag{1}$$

where each of the two signals has a Gaussian distribution with mean μ_i and standard deviation σ_i.

The correlation coefficient $\rho =< (x_1 - \mu_1)(x_2 - \mu_2) > /\sigma_1\sigma_2$ ranges from -1 to $+1$. The correlator output, without quantization, would be:

$$r =< x_1 x_2 >= \mu_1 \mu_2 + \rho \sigma_1 \sigma_2 \tag{2}$$

In this report we will assume for simplicity that the two signals have zero mean, $\mu_1 = \mu_2 = 0$. This is usually the case, as the DC component is removed before the quantization.

Quantization of the two signals $x_{1,2}$ means replacing them with a value $X_{1,2}$ in a limited set. The set is represented by the n values, v_i and the corresponding thresholds, l_i and $l_{[i+1]}$, with the index i ranging from 0 to $n-1$. The number of levels n can be odd or even, usually of the form $2^k - 1$ or 2^k. Here we will assume that the input signal is expressed in units of the quantization step, and the quantized value is the center of each quantization interval:

$$\begin{aligned} v_i &= -\frac{n-1}{2}, \ldots, +\frac{n-1}{2} \\ l_i &= v_i - \frac{1}{2}, \quad i = 1, \ldots, n-1 \\ l_0 &= -\infty \quad l_n = +\infty \end{aligned} \tag{3}$$

A particular case occurs for rounding, in which the quantization intervals are not uniform. It will be discussed separately in section 6.

The correlation of the quantized signals R is given by

$$\begin{aligned} R(\rho) = < X_1 X_2 > &= \sum_i \sum_j v_i v_j \, B(i, j; \rho) \\ B(i, j; \rho) &= \int_{l_i}^{l_{i+1}} \int_{l_j}^{l_{j+1}} P(x_1, x_2; \rho) \, dx_1 dx_2 \end{aligned} \tag{4}$$

where the integral is the probability of having the two quantized samples in the respective quantization intervals, and is expressed in terms of the bivariate normal cumulative distribution function.

In the typical radioastronomic case the amplitudes σ_1, σ_2 of the two signals do not vary significantly with time, and can be independently measured, while the correlation varies significantly, e.g. due to interferometric fringes. ρ is usually not large for cross correlation, or for autocorrelation with a nonzero delay, and is identically 1 for autocorrelation with zero delay. We will focus then on these two cases.

In a first approximation the correlation of the quantized signals R is close to the unquantized correlation r. The quantization introduces the following effects:

- The amplitude of R is different from r, but for small ρ they are almost proportional. We can assume $R(\rho) = g\, r(\rho)$, where the gain g is close to (and usually slightly smaller than) unity when using the quantization parameters of equations (3).
- For large ρ the above relation deviates from a purely linear one. We quantify this nonlinearity evaluating the quantity $R(\rho) - g \cdot r(\rho)$, or determining the next element in a Taylor expansion in ρ, proportional to ρ^3.
- The signal-to-noise ratio is worse for the quantized correlation, mostly, but not only, because of added quantization noise. The *correlator efficiency* is measured as the ratio of the signal-to-noise for the quantized vs. non-quantized correlations and it has been computed for the general case by several authors (see for example [9], or formula (11) here).

All these effects depend strongly on the signal amplitude. The gain and quantization efficiency do not depend very much on the correlation coefficient ρ, and therefore we will assume their value for small ρ. The nonlinear effect, being proportional to ρ^3, is important for high values of ρ, usually at least $\rho = 0.3$. In section 4 we will derive quantitative formulas for these effects.

In principle it is possible to correct the quantized correlation both for the gain and the nonlinearities. The two signal amplitudes $\sigma_{1,2}$ are derived from the respective autocorrelation products $< X_1 X_1 >$ and $< X_2 X_2 >$, and the correction derived for example by interpolating pre-tabulated values of $R(\sigma_1, \sigma_2, \rho)$.

This is done routinely for interferometers up to the size of ALMA, but the computation effort is very large, and requires extensive bookkeeping of the signal amplitudes at all stages of the signal processing chain. As we will see, g is relatively independent from signal amplitude in a wide set of conditions, and under these conditions this correction can be completely avoided. In chapter 5 we will derive these conditions, under specifications on gain stability and linearity derived from those of the Square Kilometre Array interferometer.

For $\rho = 1$ (autocorrelation) this is usually not true, and a correction of the autocorrelation products may be necessary even with quantization schemes with 7 or 8 bits. This correction depends however only on a single parameter, the amplitude σ of the single signal being correlated, that can be derived from the autocorrelation itself, and is thus much simpler to implement.

3 Methods

Different approaches are possible to evaluate the relation between $r(\rho)$ and $R(\rho)$.

3.1 Direct correlation of pseudo-random sequences

The more direct method to evaluate this relation is by direct simulation, using pseudo-random partly correlated sequences.

This has the advantage of being usable even for very complex data processing algorithms, involving multiple quantizations. Its precision is limited by the length of the pseudo-random sequence. For accuracies of 10^{-5}, sequences of $\simeq 10^{10}$ samples are required, and this increases by a couple of orders of magnitude when a channelization stage is present in the processing. For this reason it has not been used here and is listed only for completeness.

3.2 Integration of the bivariate distribution

The distribution in equation (1) can be directly integrated for each of the possible values of the two quantized signals. An integral form for the normalized bivariate distribution has been computed using a Legendre approximation by [2], implemented as a C routine by [8] as part of the reduction software for the Green Bank telescope and of the ALMA interferometer. In Matlab the function `mvncdf`, also based on the Drezner algorithm, can be used. This function is then directly used in a modified version of equation (4) that exploits the symmetries in the threshold values. Even with this optimization, the integration requires a summation of about $n^2/2$ evaluations of the `mvncdf` function, that becomes significant for large values of n.

3.3 Application of the Price's theorem

The integral relation of the previous case can be used to compute the derivative $dR/d\rho$, using Price's theorem [7]. This is then integrated over ρ starting from $R(0) = 0$. This approach is described for example for $n = 4$ in the data reduction software for the IRAM interferometer [3] and for an arbitrary number of levels in the GBT documentation [8].

Even in this approach the integrand is a sum over n^2 values of the thresholds, that must be evaluated in hundreds of points by the integration algorithm. Therefore it is advantageous only when the full functional form of $R(\rho)$ is required, and for n small.

3.4 Taylor expansion of the bivariate distribution

Finding a computationally efficient algorithm for equation (4) is a well known problem. For quantizations with a small number of levels, n up to 4–8, an integral relation has been proposed [5]. Other approximate formulas for $n = 2$ to 4 have been proposed for the EVLA [6]. These approaches are not suitable for large n.

For the ALMA correlator one of the authors has proposed a different method based on a series expansion of the bivariate distribution [1]. Equation (1) can be Taylor expanded as a power series of ρ for constant $\sigma_{1,2}$, and the first terms of the expansion integrated directly in equation (4).

$$P(x_1, x_2, \rho) = \frac{1}{2\pi\sigma_1\sigma_2} \sum_k \frac{\rho^k}{k!} Q_k\left(\frac{x_1}{\sigma_1}\right) \exp\left(-\frac{x_1^2}{2\sigma_1^2}\right) Q_k\left(\frac{x_2}{\sigma_2}\right) \exp\left(-\frac{x_2^2}{2\sigma_2^2}\right) \tag{5}$$

where $Q_k(x)$ is a polynomial of degree k in x.

A nice feature of this relation is that terms in x_1 and x_2 are completely separated. As a consequence the integrals are also separated, and must be evaluated only in $2n$ points, compared with $O(n^2)$ points when using the full integral $L(x_1, x_2; \rho)$.

The first terms have been explicitly derived in [4], that also noted the separation of the terms in x_1 and x_2.

$$
\begin{aligned}
Q_0(x) &= 1\\
Q_1(x) &= x\\
Q_2(x) &= x^2 - 1\\
Q_3(x) &= x(x^2 - 3)\\
Q_4(x) &= x^4 - 6x^2 + 3
\end{aligned}
\tag{6}
$$

Successive terms can be derived noting that the polynomials Q_k satisfy the relation:

$$
Q_k(x)\exp\left(-\frac{x^2}{2}\right) = -\frac{d}{dx}\left(Q_{k-1}(x)\exp\left(-\frac{x^2}{2}\right)\right) \tag{7}
$$

Expressing polynomials as $Q_k(x) = \sum_{i=0}^{k} q_{k,i}x^i$, the coefficients $q_{k,i}$ can be derived by the recursive relation:

$$
q_{k,i} = q_{(k-1),(i-1)} - (i+1)q_{(k-1),(i+1)} \tag{8}
$$

If the signal has zero mean, even order terms cancel in equation (4) due to symmetry. Assuming this is the case, substituting expansion (5) in (4), and exploiting relation (7) to perform the integration, we obtain:

$$
\begin{aligned}
R(\rho) = \frac{1}{2\pi}\sum_{k>0}\frac{\rho^k}{k!}\sum_i v_i\left[Q_{k-1}\left(\frac{x_1}{\sigma_1}\right)\exp\left(-\frac{x_1^2}{2\sigma_1^2}\right)\right]_{l_i/\sigma_1}^{l_{i+1}/\sigma_1}\\
\cdot\sum_j v_j\left[Q_{k-1}\left(\frac{x_2}{\sigma_2}\right)\exp\left(-\frac{x_2^2}{2\sigma_2^2}\right)\right]_{l_j/\sigma_2}^{l_{j+1}/\sigma_2}
\end{aligned}
\tag{9}
$$

Collecting terms with the same v_i, v_j, using the quantization scheme in (3), and noting that the exponential vanishes for the thresholds $l_0, l_n = \pm\infty$, we obtain, for zero mean signals:

$$
\begin{aligned}
R(\rho) &= \frac{1}{2\pi}\sum_k \frac{\rho^k}{k!}A_k(\sigma_1)A_k(\sigma_2)\\
A_k(\sigma) &= \sum_{i=-n/2+1}^{n/2-1} Q_{k-1}\left(\frac{i}{\sigma}\right)\exp\left(-\frac{i^2}{2\sigma^2}\right)
\end{aligned}
\tag{10}
$$

where the sum in (9) and the first polynomial contain only odd values of k.

The method is very efficient, as the argument in the summation for A_k is derived just once for each quantization threshold. Only positive thresholds can be considered, due to symmetry. The computation intensity is thus linear with the number of levels, instead of quadratic. The exponentials can be computed once and multiplied element-by-element to the functions Q_k.

The accuracy of the method is rather good, especially with many-bit quantizations. For $n = 15$ and $\sigma = 2.82$ (the optimal quantization level), the 5^{th} and 7^{th} order approximations give residuals below 10^{-6} up to $\rho = 0.45$ and 0.62 respectively. The residual at $\rho = 0.98$ for a 5^{th} order approximation is about $2\,10^{-4}$. This method is suited for fast computation of the Van Vleck correction in correlators using 4 or 5 bits.

4 Derivation of quantization effects

The formulas derived above can be used to determine in a quantitative way the effects described in section 1.

4.1 Quantization gain

The linear term in (10) can be directly used to compute the correlator gain $g_l(\sigma, n)$ is:

$$g_l(\sigma, n) = \frac{R_{lin}(\rho, \sigma, n)}{\rho\,\sigma^2} = \frac{1}{2\pi\sigma^2}\left(\sum_{i=-n/2+1}^{n/2-1} \exp\left(-\frac{i^2}{2\sigma^2}\right)\right)^2 \tag{11}$$

For large n the sum can be approximated by an integral that is exactly the finite integral of a Gaussian distribution, up to the clipping point of the quantizer. Then the gain is approximately the fraction of the signal distribution that is not clipped, and is less than 1 because of the clipping. For small signals, or n small, $g_l(\sigma, n)$ has a more complex behavior, approaching infinity or 0 for n even/odd, respectively.

4.2 Quantization efficiency

The quantization efficiency η is defined as the ratio of the signal-to-noise before and after the quantization. It has been computed by several authors. A handy formula has been computed by [9]. The same result can be found also by dividing the gain in equation (11) by the variance of the quantized product σ_R, computed in the case $\rho = 0$.

$$\eta(\sigma, n) = \frac{\sigma_r}{r} \cdot \frac{R}{\sigma_R} = \frac{g(\sigma, n) \cdot \sigma^2}{\sigma_R} = \frac{\frac{1}{2\pi}\left(\sum_i \exp\left(-\frac{i^2}{2\sigma^2}\right)\right)^2}{\left(\frac{n-1}{2}\right)^2 - \sum_i (i - \frac{1}{2})\mathrm{erf}\left(\frac{i}{\sigma\sqrt{2}}\right)} \tag{12}$$

where $\sigma_r = \sigma^2$ is the variance of the unquantified correlation for $\rho = 0$ and the sums extend from $i = -(n-1/2)$ to $(n+1/2)$.

4.3 Nonlinearities

The next terms in the equation (10) describe the deviation from the linear case. The dominant term up to large values of ρ is the second, proportional to ρ^3. The ratio of the second to the first term, proportional to ρ^2, is the fractional error in the determination of the correlation function due to quantization nonlinearities. This error directly translates into mapping artifacts, and thus in map dynamic range.

For example assuming that a strong point source is present in the mapped field, with position $\vec{r}$ and peak correlation ρ, artifacts proportional to this nonlinearity will appear at position $3\vec{r}$. Intermodulation products between different strong sources will also show up as ghost sources with similar amplitudes. Further terms will produce ghost images at positions $5\vec{r}$, $7\vec{r}$ and so on, but typically with a much reduced amplitude.

For the case of signals with equal amplitudes σ, the ratio of the ghost to the true image is:

$$N_3(\sigma, \rho) = \frac{\rho^2}{6}\left(\frac{\sum_{i=-n/2+1}^{n/2-1} \exp\left(-\frac{i^2}{2\sigma^2}\right) \cdot \left(\frac{i^2}{\sigma^2} - 1\right)}{\sum_{i=-n/2+1}^{n/2-1} \exp\left(-\frac{i^2}{2\sigma^2}\right)}\right)^2 \tag{13}$$

A strong point source may give a correlation of $\rho = 0.2$ to 0.5, and the ratio (13) must be evaluated up to these values of ρ.

5 Results

In the next sections we compute the correlator gain, the correlator efficiency and the behavior of the nonlinearities for quantization schemes of 4 to 9 bits, as a function of the signal amplitude σ and of the correlation coefficient ρ. All these results have been evaluated in Matlab, using the bivariate distribution to compute the quantization gain, and using the polynomial approximation to evaluate the nonlinear terms.

The goal is to provide limits in the signal amplitude σ for which the quantization process does not degrade the linearity and the noise of the quantized correlation above predefined limits. It is assumed that the correlation products will not be corrected for these effects, with the possible exception for autocorrelation (total power) products, and that therefore all nonlinearities and gain fluctuations would directly impact the imaging quality.

The amplitude is always expressed as a fraction of the maximum quantized value, $(n-1)/2$, to compare in the same plots quantization schemes with very different number of levels.

The limits considered have been derived from system level specifications for the Square Kilometre Array:

- **Correlation gain stability**: SKA gain stability must be better than $4\,10^{-4}$. As signal amplitude varies with time, we assumed this constrain on gain accuracy, for $\rho < 0.9$. The autocorrelation case has been treated separately.
- **Quantization noise**: The total quantization noise in SKA digital signal processing must be less than 2%. A reasonable budget considering multiple quantizations in the system leaves 0.5% quantization noise in the correlation, and 0.5% in possibly other three quantizations along the signal processing chain. We also considered a 1% quantization loss in the correlator and reduced losses elsewhere. For the intrinsically noisy 4 bit correlation, a limit of 1.5% or 2% of added noise in the correlator was assumed.
- **Nonlinearities**: Limits in nonlinearities were derived from those on mapping fidelity. We assumed a relative error due to nonlinear terms equal to 10^{-4}. Higher linearity may be required in previous stages of signal processing, in order to prevent the generation of intermodulation or harmonics in the presence of narrow-band radio frequecy interferences (RFI).

5.1 Correlator gain

The correlator gain for quantization schemes with 4 to 9 bits is shown in figure 1, as a function of the signal root mean square (RMS) amplitude in units of the maximum quantization level. The gain is relatively stable for a RMS level below 1/4 the peak digital value, and deviates again when the level drops below 0.8 times the quantization interval. Outside these bounds the gain changes rather quickly. The gain error is absolutely negligible for $\sigma < 0.22$, and increases to 0.1% for $\sigma \simeq 0.28$–0.3. For 8 bit quantization, the maximum allowable signal amplitude is about 32–35 quantization steps.

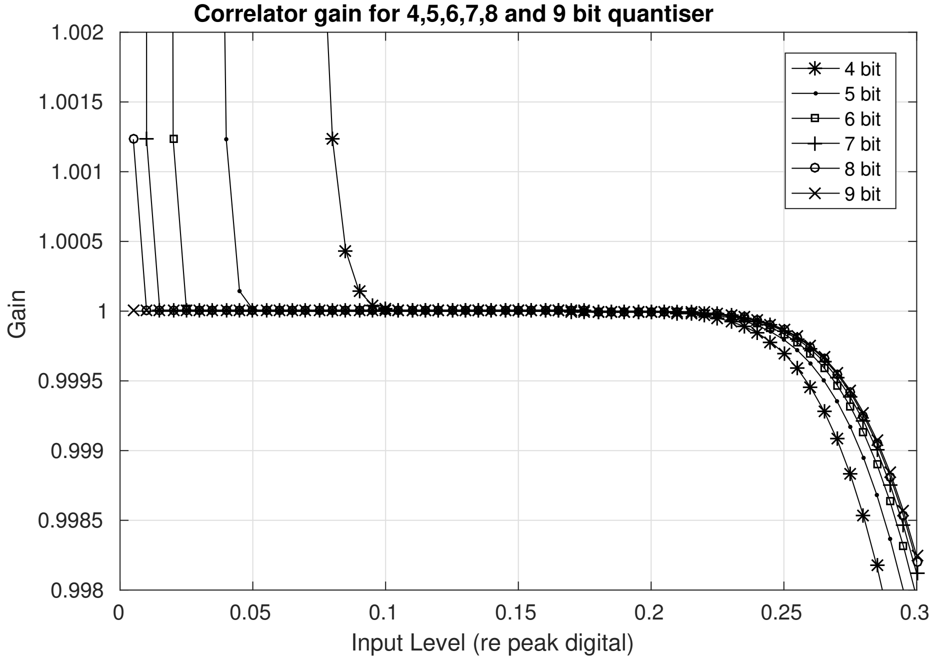

Figure 1: Correlator gain versus the input level for quantized cross-correlation, quantization with 4 to 9 bits

Autocorrelation gain ($\rho = 1$) is shown in figure 2. All quantization schemes with up to 7 bits produce gain variations above 0.1%, and even with 9 bits the stable region is much reduced with respect to the cross correlation case. A correction procedure is thus necessary in post-processing if high accuracy is required. For example, a table of $r(R)$ can be precomputed, and applied to the autocorrelation products using a spline interpolator. The computing overhead is not particularly high, as in an interferometer the autocorrelations are just a small fraction of the total correlation products, and as the function $r(R)$ is univariate, monotone and fixed for a fixed quantization scheme.

These errors affect also total power measurements obtained from quantized signals. If accuracies of the order of 10^{-4} are required, a quantization correction must be used for total power measurements derived even from 8 bit quantized signals.

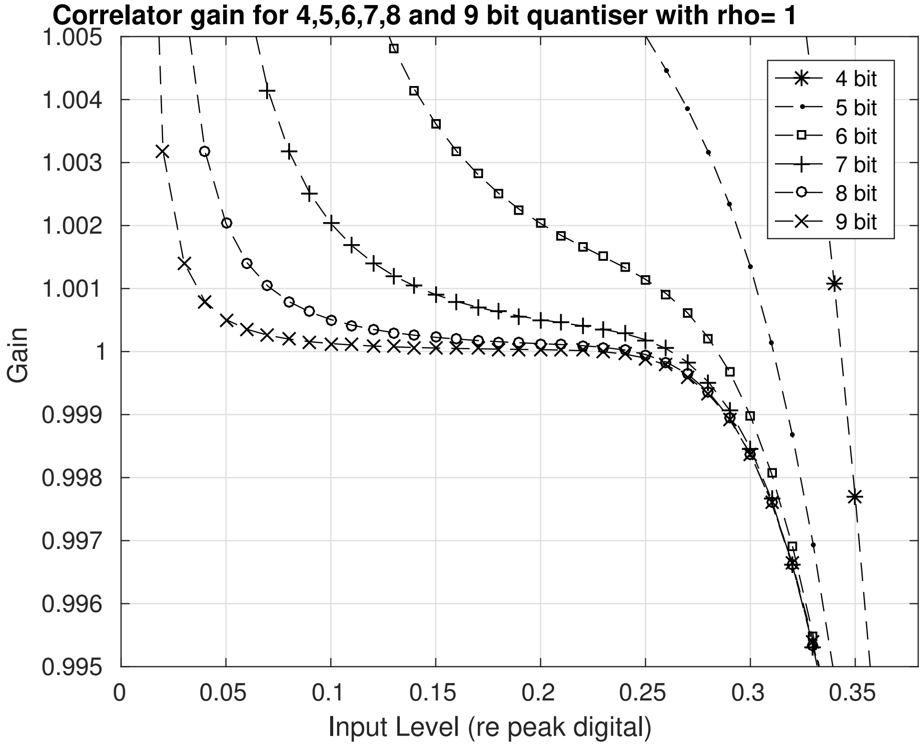

Figure 2: Autocorrelation gain versus input level for quantization with 4 to 9 bits

5.2 Correlation efficiency and quantization losses

Figure 3 shows the quantization losses as a function of the input level for cross correlation. The correlation efficiency is affected at low signal levels by the quantization noise, while at high signal levels is affected by clipping. The curves for 7, 8 and 9 bits show that the effect due to clipping becomes apparent at an input level of 0.3 and this result does not change for 10 or more bits. At low input levels, an excess noise of 0.5% is introduced when the signal amplitude is $\simeq 4$ quantization steps.

As the number of bits decreases, the minimum of the noise is found at larger signal amplitudes. For 4 or 5 bits the quantization losses in the linear region ($\sigma < 0.25$ of the clipping value) are significantly higher than at the optimum level. If no quantization corrections are used, a minimum SNR degradation of 2% and 0.6% must be considered for 4 and 5 bits quantization, instead of the canonical values 1.15% and 0.35% respectively, due to the non optimal signal levels.

5.3 Nonlinearities

Nonlinearities are evaluated computing the normalized residual $(R(\rho) - g\sigma^2\rho)/R(\rho)$, using relations (11) for g and (10) for $R(\rho)$. Results using equation (13) give very similar results. These quantities are plotted in figure 4

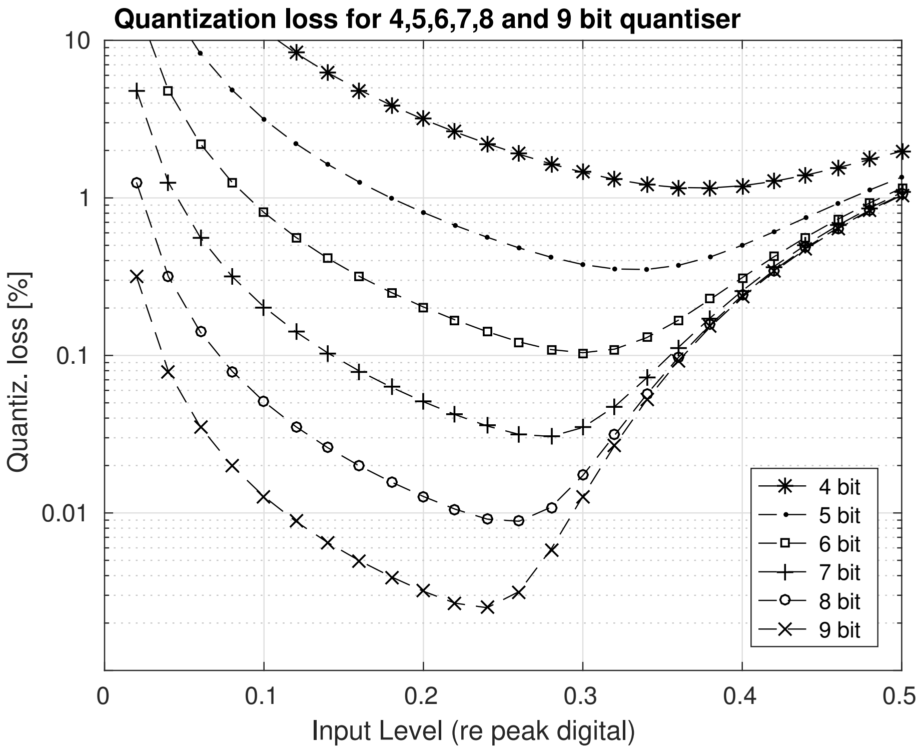

Figure 3: Correlation losses, in percent, for quantizations of 4 to 9 bits as a function of the input level

for 4 to 9 bits quantizations. The curves are computed for $\rho = 0.3$. i.e. for strongly correlated signals.

Even at these high correlation values, nonlinear terms are below 10^{-4} for $\sigma < 0.35$. In the linear gain region nonlinear terms are below 10^{-7}.

5.4 Dependence of the range on the correlation coefficient

All the above limits can be graphically represented in a single graph, as an allowed region in the amplitude-correlation plane. An example of these graphs are provided, for quantizations using 4 and 5 bits, in figure 5.

From these figures is possible to infer the following informations:

- the convex curve shows the quantization loss (right scale). Vertical limits correspond to a quantization loss of 0.5% (dash-dot) and 1% (dashed), for the 5 bit case, and to 1.5% and 2%t for the 4 bits case
- The narrower shaded area corresponds to a gain error of less than 10^{-4}
- The wider shaded area corresponds to a nonlinear term of less than 10^{-4}.

The intersection of these three conditions determine the allowable region for the signal amplitude.

We can see that for 4 bits of quantization the gain condition can be satisfied only degrading the quantization losses to 2%. For 5 bits, a 0.5% quantization loss is compatible with gain and linearity constrains only for a specific amplitude value, $\sigma = 4$ (0.25 of the clipping value).

In general:

- the lower limit is always determined by the correlation efficiency which is the main source of noise at low signal level;
- the upper limit is always given by the correlation gain degradation and is mainly caused by clipping. This occurs for a signal amplitude, normalized to the maximum representable value, of approximately 0.25 for 4 bit quantization, slightly increasing to 0.27 for 6 or more bits;

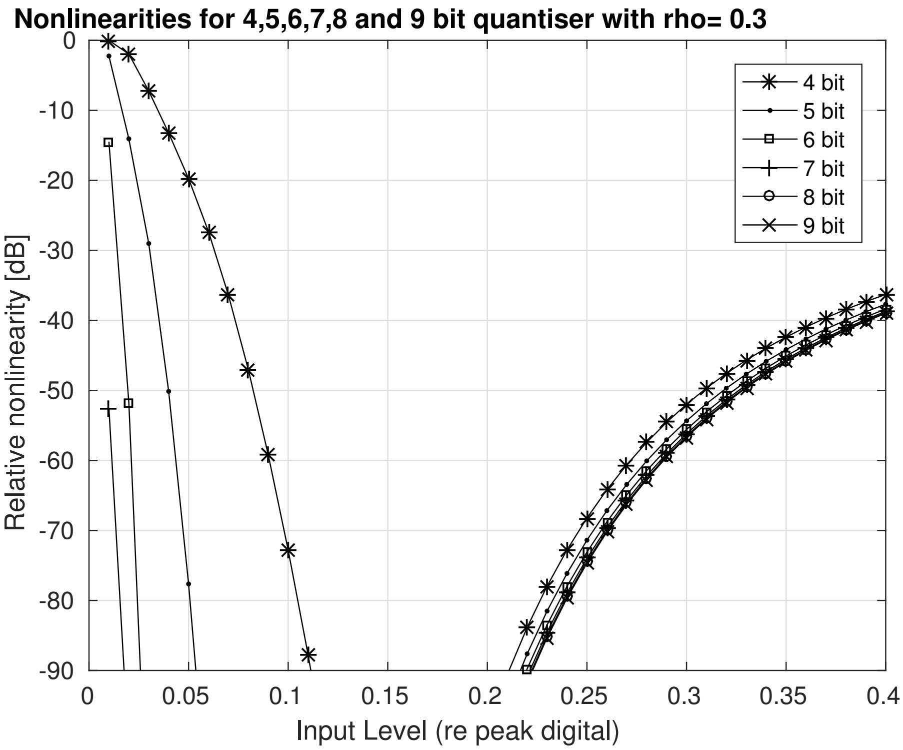

Figure 4: Relative nonlinearity of the quantized correlation for 4 to 9 bit quantization for $\rho = 0.3$

- excluding the autocorrelation case, the nonlinearity gives always less restrictive limits for the level signal compared to those derived by the correlator gain and the correlation efficiency: the nonlinear term in $R(r)$ is negligible in the range allowed for the signal amplitude up to $\rho < 0.9$
- The dynamic range for the input signal is basically null for quantization with 5 bits. It is extended by 6 dB for every increment of one bit and by a further 3 dB if 1% added noise is acceptable;

In table 1 the limits on signal amplitude and the derived dynamic range are shown. The first three rows list the minimum level for three different values of the added noise. The fourth row lists the upper bounds. All values are expressed in terms of the quantization step. The last two rows report the dynamic range $20\log(Max/min)$.

Table 1: First 3 rows: lower bounds for the input amplitude, in quantization steps; row 4: Upper bound for the signal level; last 2 rows: signal dynamic range

	Added Noise	$4bit$	$5bit$	$6bit$	$7bit$	$8bit$	$9bit$
min	0.5%	-	4.1	4.1	4.1	4.2	4.3
	1%	-	2.9	2.9	2.95	3.1	3.3
	2%	2.02					
Max		2.02	4.2	8.4	17.0	34.2	68.6
DR[dB]	0.5%	-	0.2	6.3	12.2	18.4	23.9
	1%	-	3.2	9.3	15.4	21.7	27

6 Nonlinearities due to rounding

Rounding a digital signal, by discarding the least n_r significant bits, is very similar to a re-quantization, and the same approach can be used. The rounding operation introduces both a quantization noise, that can be

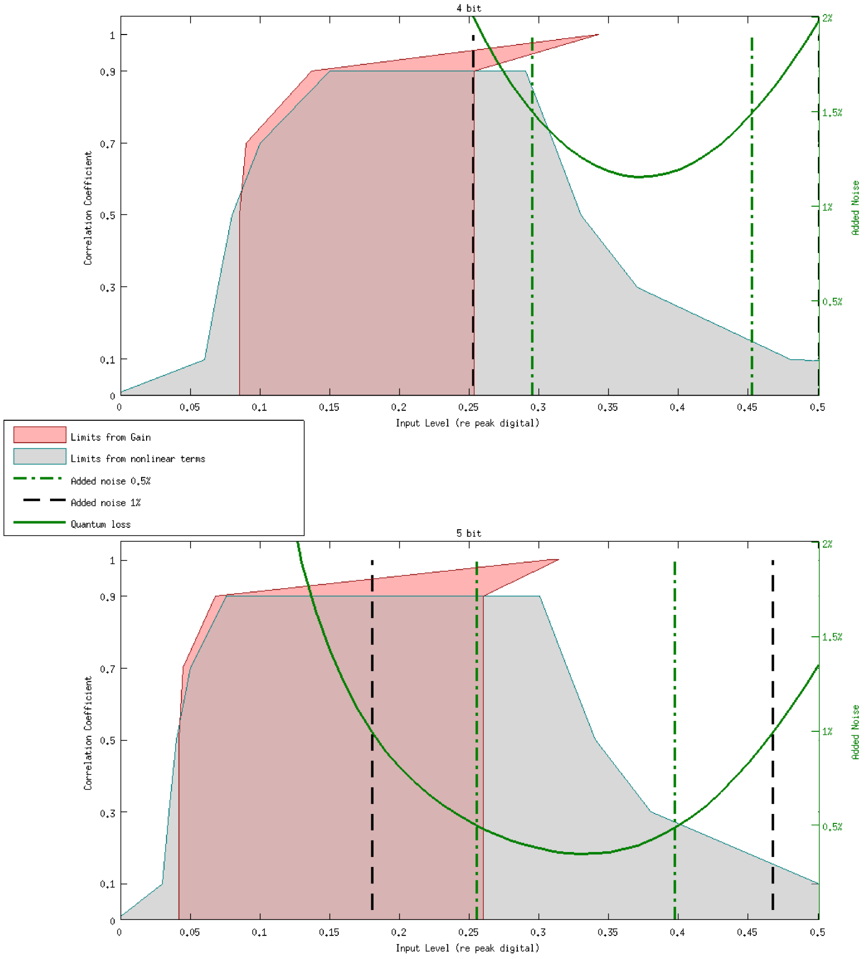

Figure 5: Gain error, quantization noise and nonlinearity level as a function of ρ and σ: 4 and 5 bit quantization

estimated using equation (11), and a bias, due to the discrete nature of the signal being quantized.

Offset rounding, consisting in summing 0.5 to the number to be rounded and then taking the integer part, produces a bias due to the fact that the fraction 0.5 is always rounded up. The bias is 2^{-n_r-1}, i.e. half the quantization step before rounding. To prevent this, if the fraction is exactly 0.5 the rounding is always performed away form zero. This however causes a discontinuity in the rounding function, as the level (or the two levels, for n even) close to zero are slightly smaller than the others. The effect can be modeled using the correct levels in equations (4) and (9).

The main effect is a slight increment on the gain, due to the systematic increment of the signal amplitude. If the re-quantized signal has a sufficient number of bits, the sums in equation (10) can be approximated with an integral. The term A_1 in equation (10) is then

$$A_1(\sigma) \simeq \sqrt{2\pi}\sigma + s_r \tag{14}$$

where $s_r = 2^{-n_r}$ is the quantization step before rounding and $\sigma_{1,2}$ are the signal amplitudes. The rounding gain g_r is

$$g_r(\sigma_1, \sigma_2) \simeq 1 + \frac{s_r}{\sqrt{2\pi}} \frac{\sigma_1 + \sigma_2}{\sigma_1 \sigma_2} \tag{15}$$

For equal signal amplitudes, the gain error $g_r - 1$ is inversely proportional to the RMS amplitude of the signal to be rounded. For a signal that is in the linear region of an integer representation with 10–12 bits the requantization introduces a gain error that is around 1%. The error drops to 0.1% for an unquantized signal of 13–15 bits, and to 10^{-4} if the unrounded RMS amplitude is at least 8000, i.e. for signals of at least 16 bits.

The nonlinear term in the Taylor expansion for $R(\rho)$ can be computed in the same way. When other nonlinear effects are negligible, the relative amplitude of the cubic term with respect to the linear term, $N_3(\rho, \sigma)$ (eq. 13), is given by:

$$N_3(\rho, \sigma_1, \sigma_2) \simeq \rho^2 \frac{1}{12\pi} \frac{s_r^2}{\sigma_1 \sigma_2} \tag{16}$$

i.e. is inversely proportional to the squared amplitude of the unrounded signal. For unrounded signal amplitudes as low as 16, $N_3 < 10^{-4}$, even for large values of ρ. Nonlinearity due to rounding is then negligible in most practical cases.

7 Conclusion

The polynomial expressions found in equations (10) for the relation $R(\rho)$ allows for a relatively simple and accurate evaluation of the quantization effects. This is useful in large interferometers, with hundreds of elements and thousands of frequency points, as is the case for the SKA.

For correlation schemes using at least 5 bits in the sample representation, there is a relatively wide range of input signal levels for which the correlation process is highly accurate and linear. In this case it may be advantageous to completely eliminate the quantization correction by keeping the signal amplitude in the linear range computed above. For quantization with 4 or 5 bits this range does not correspond to the optimal value for low quantization losses. This range depends slightly on the maximum expected correlation coefficient.

The quantization introduces a nonlinearity in the autocorrelation for digitizations of up to 7–8 bits. This effect is present also in total power measurements of digitized samples, and amounts to $\simeq$ 0.1% for 7 bit quantization. A post-correlation correction, e.g. using a spline approximation, is advisable when high accuracy is required.

Both gain stability and nonlinear terms become significant when the digitized signal level exceeds 0.25–0.27 times the clipping level. Gain is extremely stable for signal levels below 0.22 times the clipping level, down to a RMS amplitude of 0.8 times the quantization step. The lower limit in the signal amplitude is usually imposed by the added quantization noise. This is around 0.5% or 1% for $\sigma = 2.9$ and 4.1 quantization steps, respectively, independently from the quantization scheme.

Nonlinearities in the correlation process are negligible in this range of signal levels, allowing a mapping dynamic range of 10^6 up to correlation coefficients of $\rho = 0.9$. For smaller correlation coefficients ($\rho < 0.1$) the linearity is always better than 10^{-10}.

Rounding away form zero introduces a mild compression (gain increasing at low amplitudes) in most cases. The gain error is about 1% when the unrounded signal has a RMS amplitude of 80, and goes below 10^{-4} for rounding of signals with RMS > 8000. If a high signal accuracy is required, other forms of symmetric rounding must be used to round signals represented with less than 14 bits. Nonlinearities induced by rounding are usually well below 60 dB.

References

[1] G. Comoretto. Algorithms and formulas for hybrid correlator data correction. Technical Report Memo 583, ALMA, 2008. `http://library.nrao.edu/public/memos/alma/memo583.pdf`.

[2] Z Drezner and G.O. Wesolowskyi. On the computation of the bivariate normal integral. *Journal of Statist. Comput. Simul.*, 35:101–107, 1989.

[3] S. Guillotteau. Clipping correction for 4-level quantization. In *IRAM Millimeter Interferometry Summer School*, page 4.5.2, Institut de Radio Astronomie Millimtrique, Saint Martin d'Hres Cedex, France, 2000. `http://www.iram.fr/IRAMFR/IS/IS2002/html_1/node41.html`.

[4] C.R. Gwinn. Correlation statistics of quantized noiselike signals. *PASP*, 116:84–96, 2004.

[5] J.B. Hagen and D.T. Farley. Digital correlation techniques in radio science. *Radio Science*, 8:775–784, 1973.

[6] Kogan. *Radio Science*, 33(5):1289, 1998.

[7] R. Price. A useful theorem for nonlinear devices having gaussian inputs. *JRE Transactions on Information Theory*, 4:69–72, 1958.

[8] F.R Schwab. Van vleck correction for the gbt correlator. Technical report, NAIC, 2002. `http://www.naic.edu/~jeffh/fschwab_vanvleck.ps`.

[9] A.R. Thompson, D.T. Emerson, and F.R Schwab. Convenient formulas for quantisation efficiency. *Radio Science*, 42:RS3022, 2007.

[10] J. H. Van Vleck and D. Middleton. *Proceedings of the IEEE*, 54:2, 1966.

Contents

List of Figures