

# EVLA Memo. 88

## Quantization Noise

A. R. Thompson and D. T. Emerson

January 21, 2005

### Abstract

In receiving systems in which the analog signal voltage is sampled and quantized to allow further processing in digital form, the difference between the analog samples and their digital representation gives rise to a component of random quantization noise. The power spectrum of the quantization noise is close to being uniformly level across the receiver passband, even for large variation in the shape of the input spectrum. Thus in cases where the gain of the analog system varies across the passband, the addition of the quantization noise causes a variation in signal-to-noise ratio (SNR). This effect limits the allowable variation of analog gain and is particularly important in wideband receiving systems. The memorandum examines the definition of quantization noise and its relationship to quantization efficiency. Numerical simulation is used to determine the spectrum of quantization noise for a number of commonly used quantization schemes, with Nyquist and higher sampling rates. Examples are given of the limiting values of gain variation within the passband. These are modeled as slopes and sinusoidal ripples and are applicable to the EVLA and ALMA systems. Equations for precise calculation of quantization efficiency based on evaluation of the quantization noise are derived in Appendix A.

### 1. Introduction.

The degradation in signal-to-noise ratio resulting from quantization noise in systems that use digital correlators is well known for the case where the gain is uniform across the passband (see, e.g. Thompson, Moran, and Swenson, 1986, 2000, Ch. 8, and references therein). In the case where the gain varies significantly across the passband as, for example, where there is a linear variation of several decibels from one side to the other, the quantization noise can cause the SNR of the correlated signals to vary across the passband. This occurs even if the SNR before digitization is uniform across the passband. If a variation in gain across the receiver bandwidth results in a slope of the input power spectrum, the corresponding variation of the power of the quantization noise is generally very much smaller, that is, the spectrum of the quantization noise is generally close to flat. Thus the quantization noise, spread approximately uniformly across the passband, degrades the SNR of the correlated data more severely at frequencies where the gain is low, and less severely where the gain is high. This effect has been investigated for three-level<sup>1</sup> quantization by Lamb (2002), and for quantization with larger numbers of levels by Carlson and Perley (2004). In the present memorandum we use numerical analysis to investigate the spectrum of the quantization noise for various quantization schemes.

In the broadband analog stages increasingly used in radio astronomy systems, variation of gain with frequency becomes more difficult to control as the bandwidth increases. Digital filtering in the following stages offers a means for restoration of uniformity in the signal level. However, such compensation has the effect of inducing variation in the spectrum of the quantization noise, and this effect limits the extent to which gain variation in the analog stages can be compensated. It therefore becomes an important consideration

---

<sup>1</sup>The number of levels is equal to the number of divisions into which the input data are partitioned for the assignment of quantized values.

in determination of the tolerance on the variation of gain with frequency across the passband of the analog stages.

To explain what is meant by the quantization noise, we note that for each data point the initial, unquantized value minus the quantized value represents an inequality introduced by the quantization. In general, this inequality in a sequence of data contains a component that is correlated with the unquantized input data and a component of random noise. The first of these components has a spectrum identical to that of the input data, and thus after compensation for the gain variation is uniform across the spectrum. It does not degrade the signal-to-noise ratio. At the digitizer output the second (random and uncorrelated) component is flat across the passband to within about 10%, and thus after compensation for spectral gain variation becomes nonuniformly distributed. This random component is the one that the present memorandum is concerned with, and we refer to it as the quantization noise. To investigate quantization noise we have performed numerical simulations on various quantization schemes, including the effect of slopes and ripples in the IF passband. For the simulations we have primarily made use of the program Mathcad, and a number of Mathcad conventions are used in the equations that follow.

## 2. Two-level Quantization.

We begin by considering two-level quantization. Let  $x$  represent the voltage of the unquantized input data. Two-level quantization can be represented by<sup>2</sup>

$$y_2 = \text{sign}(x), \quad (1)$$

where the Mathcad function  $\text{sign}(x)$  returns 1 if  $x > 0$ , and  $-1$  otherwise. The difference between the unquantized and quantized samples is  $\Delta_2$ , which we shall refer to as the quantization inequality:

$$\Delta_2 = x - \alpha y_2, \quad (2)$$

where  $\alpha$  is a scaling factor, two values of which are found to be of particular importance. Curves illustrating  $y_2$  and  $\Delta_2$  as functions of  $x$  are shown in the top diagram of Fig. 1. Since we are concerned with the response of the quantizer to Gaussian noise, we consider that  $x$  and  $y_2$  are measured in units of  $\sigma$ , the rms level of the signal at the quantizer input.

The quantization inequality  $\Delta_2$  is partially correlated with  $x$ , so we can envisage it as containing a fraction of  $x$  plus a component of random noise that is uncorrelated with  $x$ . To demonstrate the correlation we calculate the correlation coefficient for  $x$  and  $\Delta$ , which, for any quantization scheme, is:

$$\frac{\langle x\Delta \rangle}{x_{rms}\Delta_{rms}} = \frac{\langle x^2 \rangle - \alpha \langle xy \rangle}{x_{rms}\Delta_{rms}}. \quad (3)$$

Here the angle brackets  $\langle \rangle$  indicate the mean value. From Eq. (1),  $y_{2rms} = 1$ . We take  $\sigma = 1$ , and thus  $\langle x^2 \rangle = 1$ . Since  $xy_2 = |x|$ ,

$$\langle xy_2 \rangle = \sqrt{\frac{2}{\pi}} \int_0^\infty x e^{-x^2/2} dx = \sqrt{\frac{2}{\pi}} \quad (4)$$

To evaluate  $\Delta_{2rms}$  we have,

$$\langle \Delta_2^2 \rangle = \sqrt{\frac{2}{\pi}} \int_0^\infty (x - \alpha)^2 e^{-x^2/2} dx = \left(1 - 2\alpha\sqrt{\frac{2}{\pi}} + \alpha^2\right). \quad (5)$$

For example, for an arbitrary value of  $\alpha = 1$ ,  $\Delta_{2rms} = 0.6358$  and from Eq. (3) the correlation coefficient equals 0.3179. Thus  $\Delta_2$  contains a component that is correlated with  $x$ . However, for a value of  $\alpha$  which we

---

<sup>2</sup>We use subscripts 2, 3, 4, 8, N(even), and N(odd) to  $y$ ,  $\Delta$ , and  $q$  to indicate when these symbols refer to specific quantization schemes. In expressions that apply to all quantization schemes these subscripts are omitted.

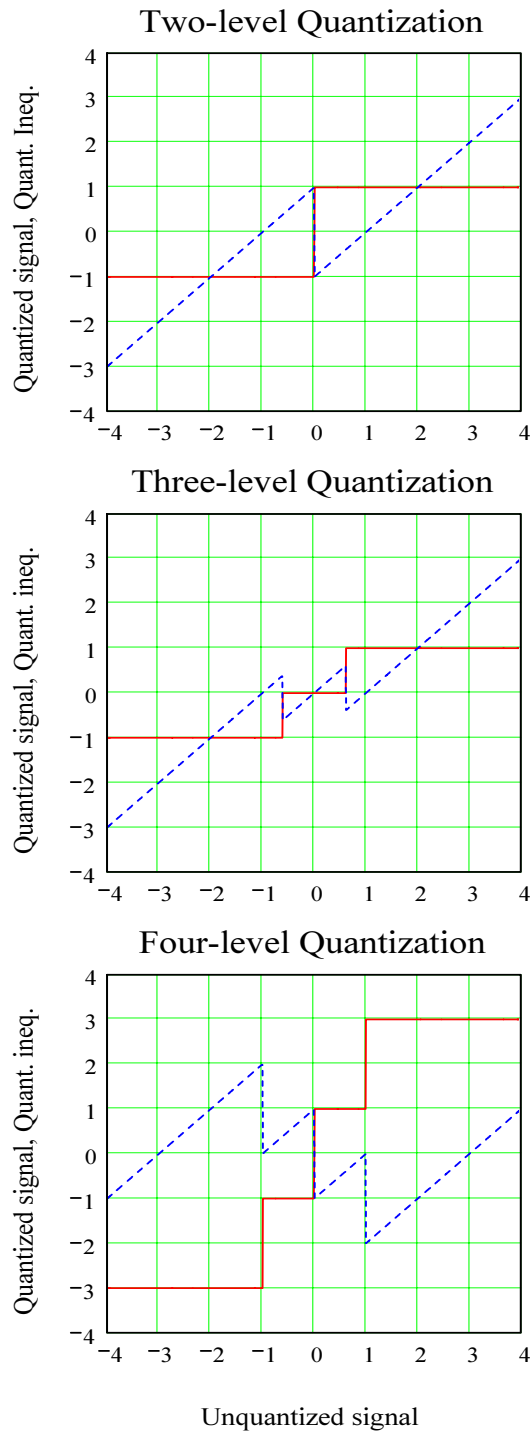


Figure 1: Two-, three-, and four-level quantization. The abscissa is the amplitude of the unquantized signal in units equal to the rms amplitude  $\sigma$ . The full (red) curves show the quantization characteristic. The dotted (blue) curves show the quantization inequality, for which an arbitrary value of  $\alpha = 1$  is used.

designate as  $\alpha_1$ , the correlation coefficient is zero. From Eq. (3) this is given by  $\langle x^2 \rangle - \alpha_1 \langle xy_2 \rangle = 0$ , that is, in general,  $\alpha_1 = \langle x^2 \rangle / \langle xy_2 \rangle$ , and for the two-level case,  $\alpha_1 = \sqrt{\pi/2}$ . Thus the random component of  $\Delta$ , i.e. the quantization noise, is given in general by

$$q = x - \alpha_1 y, \quad \text{and, in the 2-level case, by} \quad q_2 = x - \sqrt{\frac{\pi}{2}} y_2. \quad (6)$$

From Eq. (4) the factor  $\sqrt{2/\pi}$  is the correlation coefficient of a Gaussian function with its own 2-level quantization (i.e.  $x$  with  $y_2$ ). Thus a fraction  $\alpha_1$  of  $y$  is fully correlated with  $x$ . In Eq. (6)  $y$  is weighted by the reciprocal of this fraction, so that the  $x$  component is exactly canceled, leaving only the random component.

Investigation of  $\Delta_2$  using numerical simulation as described in Section 4 shows that although the condition  $\alpha = \alpha_1$  results in  $\Delta_2$  containing only the quantization noise  $q$ , this condition does not minimize the rms value of  $\Delta_2$ . As  $\alpha$  is decreased below  $\alpha_1$ , the sum of the variances of  $x$  and  $\alpha y_2$  decreases. We now investigate the value of  $\alpha$  that minimizes the rms value of  $\Delta_2$ . Since the quantized signal  $y$  is partially correlated with  $x$ , as shown for two level quantization by Eq. (4), we shall assume that in the general case  $y$  can be expressed as a scaled version of  $x$  plus a component of the uncorrelated quantization noise,  $q$ . Then Eq. (2) becomes,

$$\Delta = x - \alpha(ax + bq) = x(1 - \alpha a) - \alpha bq, \quad (7)$$

where  $a$  and  $b$  are constants. From Eq. (6), when  $\alpha = \alpha_1$ , the  $x$  components of  $\Delta$  and  $y$  cancel, leaving  $\Delta$  equal to  $q$ . This requires that  $a = 1/\alpha_1$  and  $b = -1/\alpha_1$ . Thus Eq. (7) becomes

$$\Delta = x - \alpha \left( \frac{x - q}{\alpha_1} \right) = x \left( 1 - \frac{\alpha}{\alpha_1} \right) - \frac{\alpha q}{\alpha_1}, \quad (8)$$

The variance of  $\Delta$  (noting that  $x$  and  $q$  are uncorrelated) is

$$\langle \Delta^2 \rangle = \langle x^2 \rangle \left( 1 - \frac{\alpha}{\alpha_1} \right)^2 + \left( \frac{\alpha}{\alpha_1} \right)^2 \langle q^2 \rangle. \quad (9)$$

Differentiating the right-hand side with respect to  $\alpha$ , and equating the derivative to zero, we obtain the value of  $\alpha$  that minimizes  $\Delta^2$ , which we denoted by  $\alpha_2$ . Thus,

$$\frac{\alpha_2}{\alpha_1} = \frac{\langle x^2 \rangle}{\langle x^2 \rangle + \langle q^2 \rangle}, \quad (10)$$

In Eq. (10) the right-hand side is the variance of an analog signal (for which  $q$  is zero) as a fraction of the variance of an equivalent quantized signal. The square root of this ratio is the SNR for the digitized signal divided by the SNR for the unquantized analog signal. Therefore the ratio of the SNR at the correlator output for the correlation of two digital signals, to the SNR for the equivalent analog signals, which is referred to as the quantization efficiency  $\eta_Q$ , is equal to the ratio in Eq. (10). That is,

$$\eta_Q = \frac{\alpha_2}{\alpha_1}. \quad (11)$$

For the 2-level case,  $\alpha_1 = \sqrt{\pi/2}$  from Eq. (6), and the quantization efficiency equals  $2/\pi$ , the well-known result originally derived from the work of van Vleck and Middleton (1966). Thus  $\alpha_2 = \sqrt{2/\pi}$ , which is in agreement with the result from numerical analysis described below. Also we have

$$\frac{\langle q^2 \rangle}{\langle x^2 \rangle} = \left( \frac{\alpha_1}{\alpha_2} \right) - 1 = \frac{1}{\eta_Q} - 1. \quad (12)$$

This is the variance of the quantization noise expressed as a fraction of the variance of the unquantized input data, which provides a useful way of quantifying the power level of the quantization noise.

### 3. Numerical Analysis

To determine how the spectral shape of the input (unquantized) signal affects the spectral shape of the quantization noise, we have performed numerical analyses in which either a power-linear slope, or a sinusoidal ripple is introduced into the input spectrum. In practice, variations in the frequency response of IF systems can largely be modeled in terms of such slopes and ripples. Mathcad was used for these programs. The essential steps are as follows.

- 1) Generate  $2^n$  samples of random noise (Mathcad function *rnorm*) with zero mean and a standard deviation  $\sigma = 1$ : values of 16 to 20 were used for  $n$ .
- 2) Use an FFT (Mathcad function *fft*) to generate the frequency spectrum of the noise,  $(2^{n-1} + 1)$  complex values.
- 3) Insert a slope or ripple into the spectral data. With the slope the power level varies linearly from 1 at the low frequency end of the band to a value varying between 1 and 100 at the high end. With a ripple, the power level is proportional to  $[1 + m \sin(2\pi k f / F)]$  where  $m$  is the modulation index,  $f$  is the spectral frequency, and  $k$  is the number of ripple cycles across the input passband, which extends from zero to  $F$ . The value of  $k$  used was generally about 100 and the value of  $m$  was in the range 0 to 1, that is, 0% to 100% modulation.
- 4) Scale the frequency data so that the rms value remains equal to one, as before the slope or ripple were added.
- 5) Use an FFT (Mathcad function *ifft*) to transform the data back to  $2^n$  real values of the amplitude in the time domain. These data with spectrum modulation and unit rms level provide the input data for the quantization process.
- 6) Apply quantization (e.g. Eqs. (1), (14), etc.) and the appropriate expression for the inequality,  $\Delta$ .
- 7) Use an FFT (Mathcad function *fft*) to transform the inequality data to the frequency domain ( $2^{n-1} + 1$  complex values), and take the squared moduli of these data.
- 8) For the slope investigation, fit a straight line (rms best fit) to the squared moduli as a function of frequency (Mathcad functions *intercept* and *slope*). For the ripple investigation, determine the amplitude and phase of the Fourier component at the frequency of the input-data ripple.
- 9) To obtain the desired accuracy, repeat the sequence of steps above taking the average of the results. For each repetition a different seed value for *rnorm* must be used in step (1) to provide an independent set of random data.

Since the simulations are based on the use of random noise as the input data, the accuracy is limited by the number of data values used. With  $2^{16}$  to  $2^{20}$  samples of white noise, the expected precision of the results is approximately 1 part in  $2^8$  to  $2^{10}$ , or about 0.4% to 0.1%. However, the insertion of the slope or ripple (step 3) results in a non-white noise spectrum, and the noise power becomes concentrated toward the high frequency end of the band, or the positive peaks of the ripple. As a result, the effective bandwidth decreases, and so does the statistical independence of the noise samples. Also, for high slopes, the data at the low edge of the spectrum are very small compared to the overall noise level. As indicated in step (9), computations were repeated with independent data and averaged. The aim was to reach a level of accuracy that would illustrate the behavior of the quantization noise, for which, in general, of a few tenths of a percent was sufficient.

Note that in describing the results, we use the term “slope ratio” to indicate the ratio of the power spectral density at the upper end of the band to that at the lower end. With this definition a slope ratio of

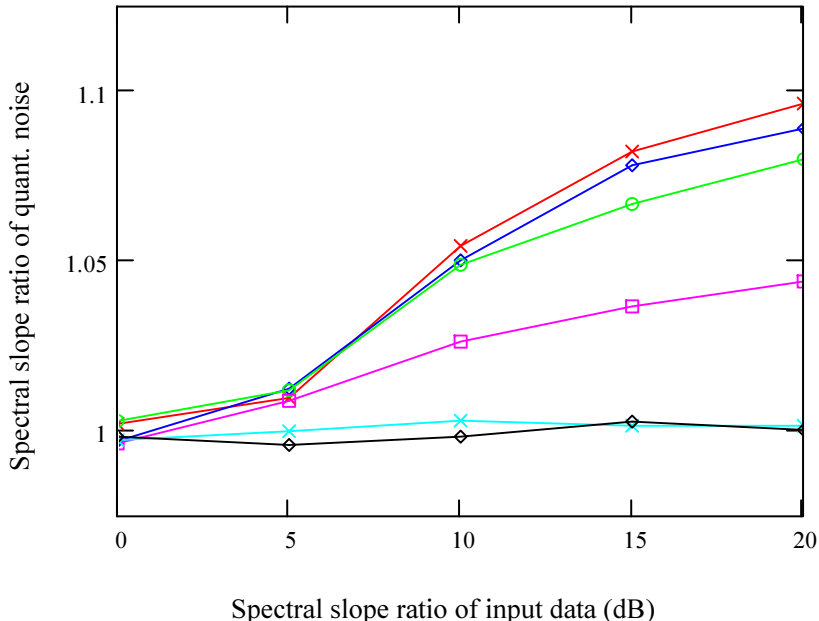


Figure 2: The ordinate is the slope ratio of the quantization noise power spectrum on a linear scale. The abscissa is the slope ratio of the power spectrum of the input data on a dB scale. Moving downward from the top, the curves are for: 2-level (red, crosses), 3-level (blue, diamonds), 4-level (green, circles), 8-level (magenta, squares), N(even)-level (cyan, crosses), N(odd)-level (black, diamonds). The standard deviation of the points is approximately  $3 \times 10^{-3}$ . Thus the values do not differ significantly from unity for all curves at 0 dB on the abscissa, or for all points on the curves for N(even) and N(odd) levels.

1.0 indicates a uniform power level.

#### 4. Results for 2-level Quantization.

With  $\alpha = \alpha_1$ , a slope or ripple modulation inserted in the input data should be strongly minimized, since this condition eliminates the component of  $\Delta$  that is correlated with  $x$ . Using the program outlined above, the value of  $\alpha$  that minimizes the modulation was found to be 1.248 from investigation using the slope, and 1.255 from investigation using the ripple. Both of these values are statistically consistent with the theoretical value of  $\sqrt{\pi/2}$  ( $= 1.2533$ ). Note that, as in the example in Fig. 3, the minima are not sharp. Then with  $\alpha$  set to 1.2533 the slope of the quantization noise spectrum as a function of the slope of the input data (step 3 of the program) was investigated. The results are shown by the upper curve in Fig. 2. The spectrum of the quantization noise is flat for a slope ratio of 1 in the input spectrum, but shows a small residual slope with increasing slope of the input data. However, this is not a large effect: an input slope ratio of 100 (20 dB) results in a residual slope ratio of 1.09 in the quantization noise. Similarly, a ripple modulation of 100% in the input data produces a peak-to-peak modulation of only 8% (i.e. 4% modulation depth in the two-level curve of Fig. 4). The value of  $\alpha$  that minimizes the rms value of  $q_2$  was investigated by including an evaluation of this rms in step 6, and was found to be  $\alpha = 0.797$  with an uncertainty of about  $\pm 0.001$ . This is consistent with the theoretical value of  $\alpha_2$  obtained from  $\alpha_1$  and  $\eta_Q$ : see Table 1. In the investigation of the residual modulation resulting from a sinusoidal ripple in the receiver passband, we found that the amplitude of the residual is dependent on the phase of the input ripple. We believe that this is explained by the aliasing back into the passband of higher frequency quantization noise introduced by the digitization

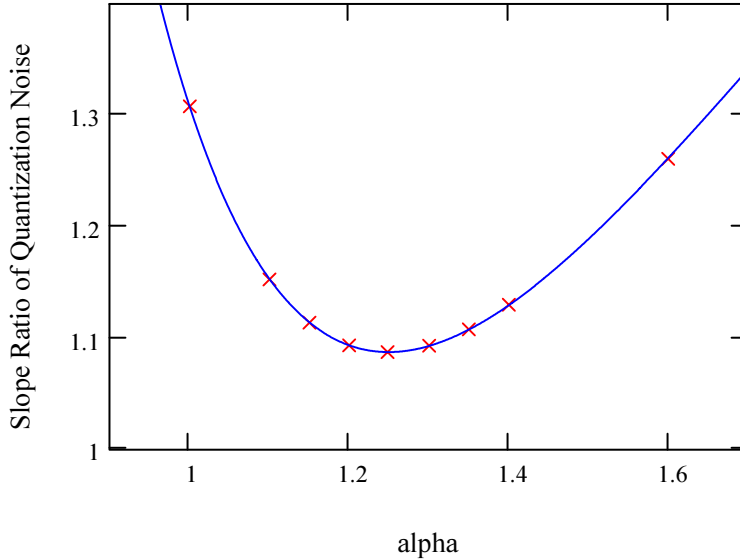


Figure 3: Minimum in slope of the quantization noise spectrum for 2-level sampling as a function of  $\alpha$ . The slope ratio of the input power spectrum is 20 dB. Further points, not shown, provided a value of  $\alpha_1 = 1.248$ .

process. Note that for both slope and ripple input spectral shapes, the envelope of the quantization noise mimics the spectral shape but with much reduced amplitude; however the actual noise voltage under this envelope is completely uncorrelated with the original input noise.

## 5. Aliasing.

For most of the numerical analysis we have used Nyquist-sampled data. The passband of the input data, which may contain gain modulation in the form of a slope or sinusoidal ripples, extends from frequency 0 to  $F$ , and the quantized data are sampled at frequency  $2F$ . However, to investigate the effects of aliasing, some simulations were made using oversampled data. As the oversampling factor is increased the spectrum of the quantized data resembles more closely the spectrum of a continuous (unsampled) but quantized function. Oversampling is simulated by zero-padding the data, that is, inserting zero values in the frequency domain beyond the original non-zero spectrum, before transforming back to an *oversampled* voltage in the time domain. For  $n$ -times oversampling the number of zeros is  $(n - 1)$  times the number of data values within the 0-to- $F$  baseband spectrum. Figure 5 shows a power spectrum of the quantized signal ( $y_2$ ) from 32:1 oversampled data. For frequencies greater than  $F$ , the envelope of the quantization noise power follows approximately an inverse square law. In the graphs in Fig. 5 the envelope shows an extended plateau or tail which we attribute to multiple aliasing of frequencies greater than  $16F$ . Similarly, the detailed shape of the spectrum of the quantization noise in our Nyquist sampled simulations is the result of multiple aliasing back into the input passband of higher frequency components introduced by the quantization process.

In a qualitative way, the effect of aliased quantization noise can be visualized as follows. Consider first the quantization, the sampling being a separate operation performed subsequently. The quantized signal contains abrupt steps as the analog signal moves between digitization thresholds. Each such step can be considered as the addition of a Heaviside step function to the mean signal at the level-crossing instant. A Heaviside function has  $1/f$  frequency spectrum in amplitude, i.e.  $1/f^2$  in power. With purely random noise, the precise instant at which these abrupt steps occur is largely random, so the frequencies generated by individual steps can be considered as random in phase. The power spectrum of the quantization noise

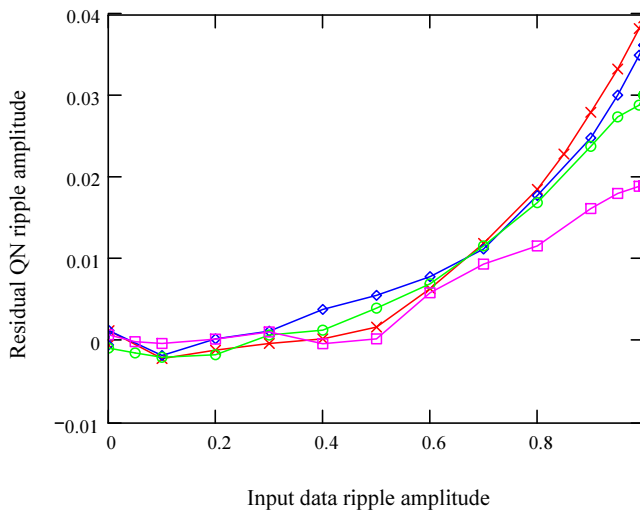


Figure 4: The effect of sinusoidal ripple modulation. The abscissa is the ripple amplitude (modulation depth) for the input data, and the ordinate is the amplitude of the resulting residual modulation of the quantization noise. The curves are for 2-level quantization (red, crosses), 3-level (blue, diamonds), 4-level (green, circles), 8-level (magenta, squares).

resulting from the combination of multiple steps will also have a  $1/f^2$  power slope. This is consistent with the spectral shape seen in Fig. 5. In the subsequent sampling, the high frequency tails of these  $1/f^2$  spectra become aliased back into the baseband frequency range zero to  $F$ .

For the Nyquist sampling rate of  $2F$ , signals will be aliased back into the 0-to- $F$  frequency band in such a way that at frequency  $f_0$  within the band there will appear additional components from outside the band at frequencies  $(2nF - f_0)$  and  $(2nF + f_0)$ , where  $n$  is an integer. The aliasing is alternately from initially negative and positive frequencies, respectively. The resultant baseband power at frequency  $f_0$  will be the sum of all the  $(2nF - f_0)$  component powers and the  $(2nF + f_0)$  component powers, corresponding to aliased negative and positive frequencies. If those aliased components themselves have amplitude proportional to  $1/f$ , then an infinite series

$$1/(2nF - f)^2 + 1/(2nF + f)^2, \quad (13)$$

where  $n$  goes from 1 to  $\infty$ , will define the additional power in the 0-to- $F$  baseband due to aliased quantization noise. This is valid for *any* quantization scheme. For a model  $1/f^2$  profile, Fig. 6 shows the power spectral density profiles for the first six orders of components aliased back. Note, however, that in Fig. 5 the profile of the noise spectrum shows some deviation from the  $1/f^2$  profile in the range from channel number 128 to about 300, so the first order (top) curve in Fig. 6 and possibly others do not accurately represent the behavior of the quantization noise derived from numerical simulation. The curves in Fig. 6 help visualize the aliasing back in a qualitative way, but the quantitative details of the effect on the slope of the quantization noise within the passband are not determined. Different numbers of quantization levels will change the amplitude of this aliased noise, but its spectral shape should be the same: see for example the curves for two-level sampling in Fig. 5. However, for Nyquist sampling the individual data points are statistically independent, and for a flat input spectrum this results in a flat spectrum of quantization noise within the passband. Thus in the Nyquist case the components that are aliased back into the baseband spectrum exactly compensate for the roll-off of the oversampled spectrum. Figure 6 illustrates how multiple alias terms are folded back into the original baseband. If the high frequencies contain sinusoidal terms, as is generally true for gain ripples in the passband, then depending on the phase of the original ripple, the structure in the aliased spectrum may

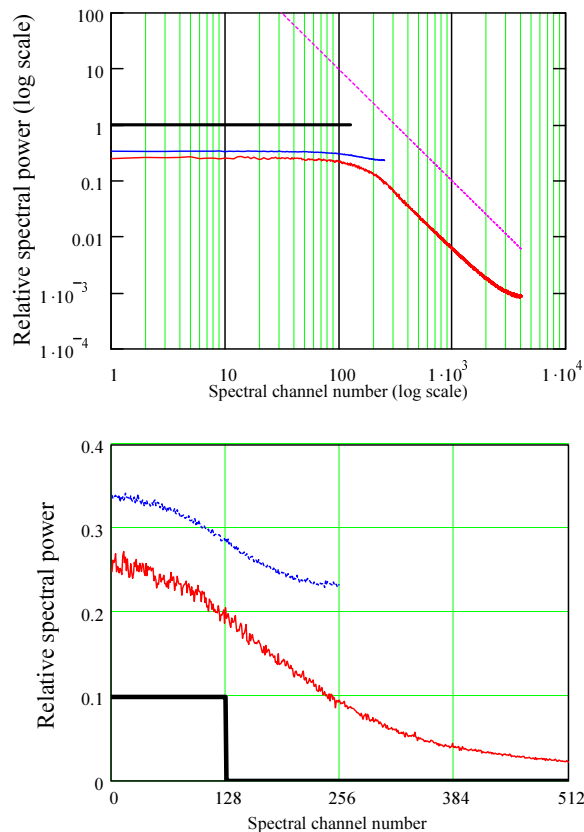


Figure 5: The red curves are plots of the power spectrum of a two-level quantized signal ( $y_2$ ) with 32:1 oversampling. The blue curves show the power spectrum for quantization noise with 2:1 oversampling. Spectral data resulting from  $2^{18}$  oversampled time-domain points were averaged in blocks of 32 to simulate frequency channels. There are 128 channels within the input passband, which is indicated by the black lines. To reduce the noise on the curves, the smoothed spectra were repeatedly averaged using independent random input data for each iteration. The original input spectrum was flat and the noise beyond channel 128 is created by the quantization process. As shown in the upper plot, the power spectrum falls off approximately inversely as the square of the frequency, as predicted by the simple theory described in the text. For comparison the magenta (dashed) line shows a slope of -2 on this log-log plot. The lower plot shows the first 512 channels of the same data plotted on linear scales to reveal the variation of the level within the passband frequency range.

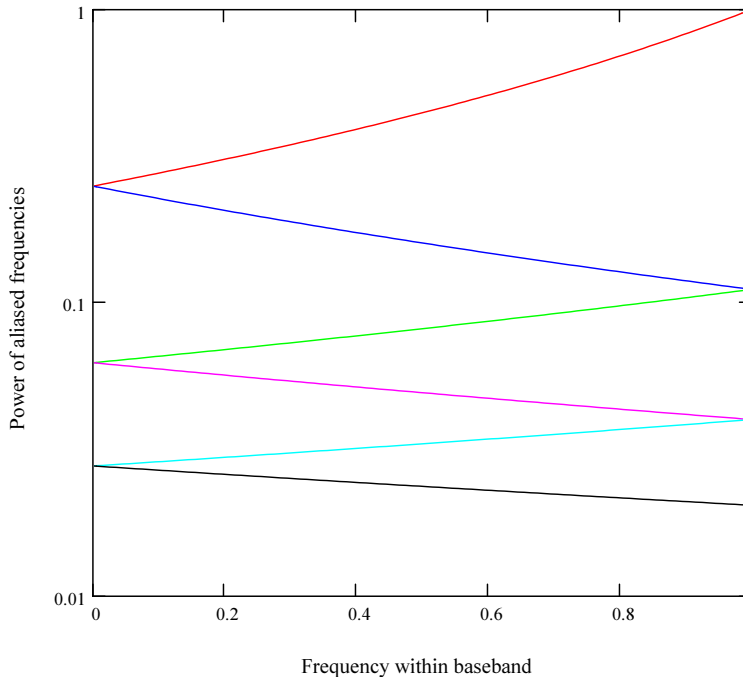


Figure 6: Power spectral density of  $1/f^2$  components aliased back into the baseband spectrum of the input signal, for 2-level sampling. The ordinate is on a log scale. The top curve is for the first-order alias, the next curve down for the second order, etc. down to the bottom curve for the sixth order. They are derived from Eq. (6) for  $n = 1$  to 3. Odd-order components slope upward toward the right and even orders slope downward.

reinforce or partially cancel the residual of the original ripple in the passband. This explains the dependence of the original ripple amplitude on the phase of the IF passband ripple, as mentioned in Section 4.

Oversampling by a factor of two is sometimes used to improve the SNR when using two-level sampling. The quantization efficiency is then increased from 0.6366 to 0.7442, the latter figure being from Thompson, et al. [2001, see Eqs. (8.32) and (8.34)]. Deriving the continuum noise degradation factors involves integrating the variances of the power spectral densities over frequency<sup>3</sup>. Using this method for 2-level sampling with 2-times oversampling, we obtain  $\eta_Q = 0.7443 \pm 0.0004$  by averaging 16 independent runs of  $2^{19}$  data samples. From the two curves in the lower graph of Fig. 5, it is seen that when oversampling is used the quantization noise level varies by about 1.2:1 over the passband of the input signal. Thus, in using oversampling to improve the sensitivity of spectral line observations, the noise will no longer be uniform across the band, but for a baseband IF should be a little better at the high-frequency end. Figure 7 shows the spectrum of oversampled quantization noise for an IF passband in which the low frequency does not extend down to zero but only to half the upper frequency. In this case the quantization noise is highest near the center of the passband and decreases toward the edges. Subsidiary maxima in the spectrum are seen at odd harmonics of the intermediate frequency.

<sup>3</sup>Specifically,  $\eta_Q = \sqrt{(v_1^2 + v_2^2)/v_0^2}$ , where  $v_1$  is the variance of the squared modulus of the frequency-domain data integrated over the passband,  $v_2$  is the variance of the equivalent quantity integrated over frequency outside the original passband, and  $v_0$  is the of the squared modulus of the pre-quantization frequency-domain data.

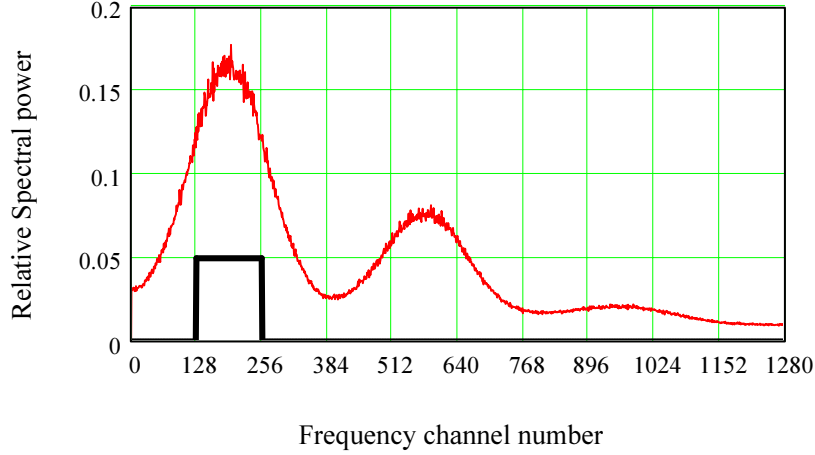


Figure 7: Relative spectral power of a two-level quantized signal with 32:1 oversampling, for the case where the signal (IF) spectrum does not have a baseband characteristic, but has a cutoff at low frequency equal to half that at the high frequency.

### 6. Three- and Four-level Quantization.

Curves representing the quantization characteristic and the quantization noise for 3 and 4 levels are also shown in Fig. 1.

Three-level can be represented by:

$$y_3 = if[|x| > 0.612\sigma, \text{sign}(x), 0], \quad (14)$$

where the Mathcad function  $if()$  indicates that if  $|x| > 0.612\sigma$  is true,  $y_3$  is assigned the value  $\text{sign}(x)$ , and if it is false,  $y_3$  is assigned the value 0. Threshold levels of  $\pm 0.612\sigma$  provide optimum SNR for the correlated signal. The quantization noise is given by:

$$q_3 = x - \alpha_1 y_3. \quad (15)$$

The mean value of  $xy_3$  is

$$\langle xy_3 \rangle = \sqrt{2/\pi} \int_{0.612}^{\infty} x e^{-x^2/2} dx = 0.6616, \quad (16)$$

where the integral is evaluated numerically using Mathcad. Then following the equivalent discussion for the 2-level case the value of  $\alpha_1$  for three-level quantization is  $1/0.6616 = 1.5114$ .

Four-level quantization can be represented by:

$$y_4 = if[|x| > 0.996, 3 \text{sign}(x), \text{sign}(x)], \quad (17)$$

where, for optimum SNR, the quantization thresholds are equal to  $\pm 0.996\sigma$  (Schwab 2005), and the inner and outer levels are assigned values of  $\pm 1$  and  $\pm 3$  respectively. The quantization noise is:

$$q_4 = x - \alpha_1 y_4. \quad (18)$$

The mean value of  $xy_4$  is

$$\langle xy_4 \rangle = \sqrt{2/\pi} \left[ \int_0^{0.996} x e^{-x^2/2} dx + 3 \int_{0.996}^{\infty} x e^{-x^2/2} dx \right] = 1.7696, \quad (19)$$

where the integrals are again evaluated numerically using Mathcad. We thus obtain a value for  $\alpha_1$  of  $1/1.7696 = 0.5651$ .

As in the case of two-level quantization, the values of  $\alpha_1$  and  $\alpha_2$  for three- and four-level quantization derived by numerical simulation are in good agreement with the analytical values, as shown in Table 1. The slope of the quantization noise as a function of the slope of the input data is shown in Fig. 2, and the ripple modulation of the quantization noise is shown in Fig. 3. For both slopes and ripples, the resulting modulation on the quantization noise for a given modulation level of the input data is relatively small and decreases as the number of quantization levels increases.

## 7. N-Level Quantization

When the number of quantization levels  $N$  is larger than four, the levels are usually uniformly spaced<sup>4</sup> in voltage at intervals  $\epsilon$  (in units of the rms level  $\sigma$ ). An input sample that falls between levels  $m\epsilon$  and  $(m+1)\epsilon$  is assigned a value  $(m+1/2)\epsilon$ , and one that falls between  $-m\epsilon$  and  $-(m+1)\epsilon$  is assigned  $-(m+1/2)\epsilon$ . The partitioning of the samples into level intervals and the assignment of values can be thought of as separate operations<sup>5</sup>. In this section we consider cases where the number of levels is large enough that the range of the quantization levels extends over several times the rms level. Thus the probability of occurrence of values that lie outside the range of the quantization levels is very small and such values can often be ignored<sup>6</sup>. Figure 8 shows the characteristic curves for both even and odd numbers of levels. For an even number of levels, one can also describe the system as one in which  $x/\epsilon$  is truncated toward the more negative level, and then  $\epsilon/2$  is added. The quantization scheme can be expressed in terms of the Mathcad functions  $\text{trunc}(x)$ , which returns the integer part of  $x$ , and  $\text{sign}(x)$ :

$$y_{N(\text{even})} = \epsilon \left[ \text{trunc}\left(\frac{x}{\epsilon}\right) + \frac{\text{sign}(x)}{2} \right] \quad (20)$$

and

$$q_{N(\text{even})} = x - \alpha_1 y_{N(\text{even})} \quad (21)$$

The corresponding scheme for an odd number of quantization levels can be represented by:

$$y_{N(\text{odd})} = \epsilon \text{trunc}\left(\frac{x}{\epsilon} + \frac{\text{sign}(x)}{2}\right), \quad (22)$$

and

$$q_{N(\text{odd})} = x - \alpha_1 y_{N(\text{odd})} \quad (23)$$

In applying the program to the  $N(\text{even})$  and  $N(\text{odd})$  cases in Fig. 8, a value of  $\sigma/2$  is used for the interval between levels (i.e.  $\epsilon = 0.5$ ), as proposed for the 256-level (8-bit) quantizer of the EVLA. For both even and odd cases a value of 1.0 is used for  $\alpha_1$ . From the curves for the inequality in Fig. 8, it is seen that this quantity changes sign as the amplitude of the input data  $x$  varies through each level increment  $\epsilon$ . This variation in the sign of  $\Delta$  can be expected to greatly reduce any correlation with the input data and thus  $\Delta$  consists mainly of the quantization noise component. As a result, there is a broad minimum in the rms value of  $\Delta$  which is centered on  $\alpha_1 = 1.0$ . As noted in Section 2,  $\alpha_1 = \langle x^2 \rangle / \langle xy \rangle$ , so in higher level schemes in which the individual input data values  $x$  are more accurately represented by the quantized values  $y$ ,  $\alpha_1$  approaches unity.

<sup>4</sup>Small improvements (of order a few tenths of one percent) in quantization efficiency can be obtained by varying the level spacings (see e.g. Jenet and Anderson, 1998; Schwab 2005), but the simpler case of uniform increments is satisfactory for investigation of the main features of the quantization noise.

<sup>5</sup>Partitioning and assignment of values may, in practice, take place in different locations. If the signals are digitized at the antennas, the level partitions are first encoded in a manner that is most efficient for transmission of the large data streams to the correlator, and numerical values then assigned as part of the correlation process.

<sup>6</sup>It is becoming a common practice in radio astronomy systems to allow for a level of RFI (radio-frequency interference), in addition to the noise, within the operating range of the quantizer. For some bands of the EVLA, 256-level (8-bit) quantization is being used for this reason. The noise alone is then well within the range of the quantization levels.

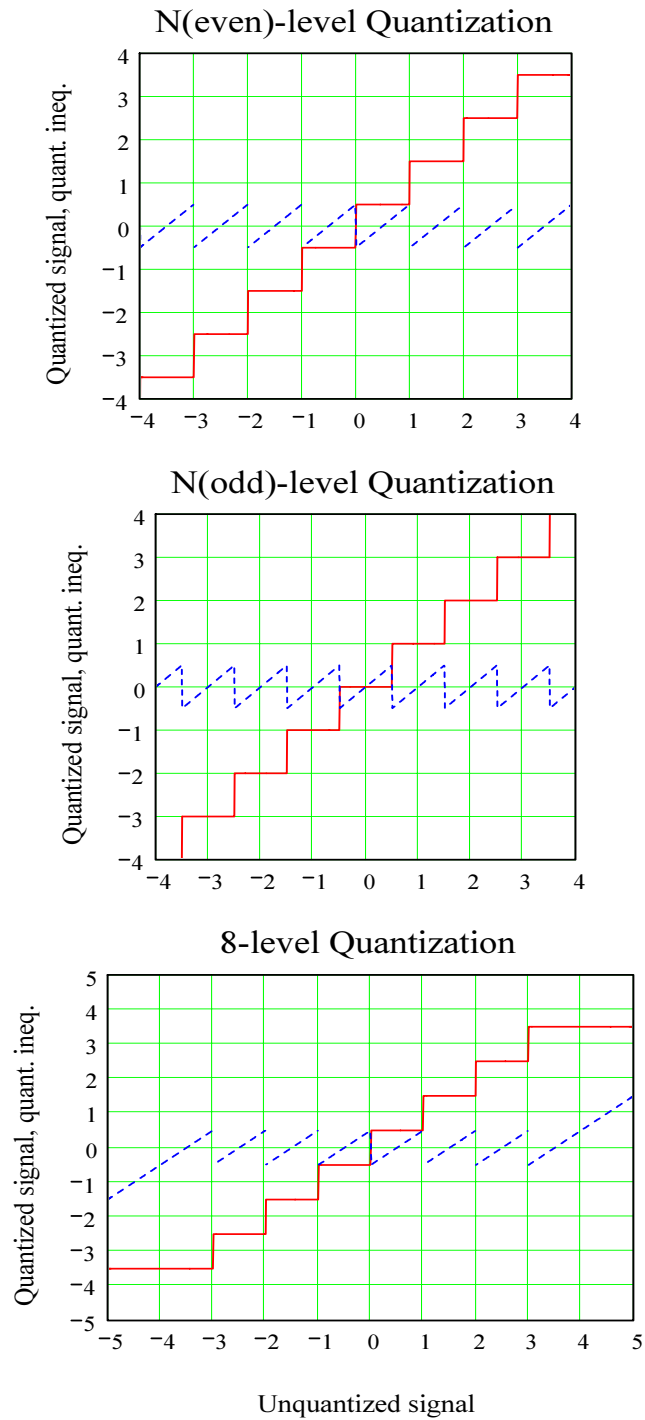


Figure 8: Quantization characteristics, full (red) curves, and quantization inequality, dotted (blue) curves. The abscissa is the amplitude of the unquantized signal in units equal to the rms amplitude  $\sigma$ , i.e.  $\epsilon = 1$ . An arbitrary value of  $\alpha = 1$  is used for the quantization inequality.

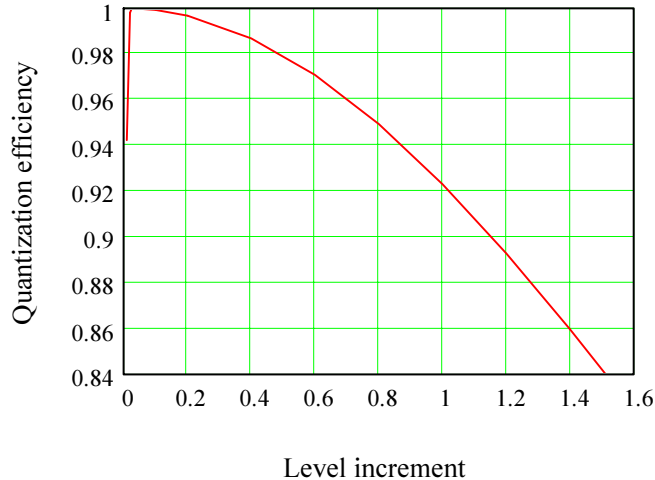


Figure 9: The quantization efficiency  $\eta_Q$  as a function of the increment between quantization levels  $\epsilon$ , for 256-level (8-bit) quantization. Note that  $\eta_Q$  starts to decrease when  $\epsilon$  is decreased below 0.03.

Results for the slope of the quantization noise spectrum as a function of the slope of the input data spectrum are shown in Fig. 2. For the N-level cases, with both even and odd numbers of levels, the rms amplitude of the quantization noise is consistent with a constant level across the band (i.e. slope ratio = 1.0), for values of the input slope ratio of up to 20 dB. For large numbers of levels (e.g. 256), as in the EVLA, the requirement for a linear response to interfering signals, in addition to considerations of SNR, is a major consideration in the choice of  $\epsilon$ . Figure 9 shows the quantization efficiency  $\eta_Q$  as a function of  $\epsilon$  for the range over which  $\eta_Q$  remains relatively high, calculated using Eqs. (31) and (34) derived in Appendix A. Over a range of  $\epsilon$  up to approximately 1.5, the quantization noise spectrum remains essentially flat for input slopes up to 20 dB. As  $\epsilon$  is further increased, the abscissa scales in Fig. 8 are expanded so that an increasingly large part of the input spectrum falls between the two quantization thresholds on either side of zero. For even values of N, the quantization action then begins to resemble that for three-level quantization. For odd values of N, the data falling between the two inner thresholds are quantized to zero and only the outlying values remain.

### 8. 8-Level Quantization

Eight-level (3-bit) quantization, which is used in the EVLA, and in ALMA for digitization at the antenna,<sup>7</sup> is a case of N(even) in which values of the input signal that lie outside the range of the quantization levels cannot be ignored. The quantized values run from  $-3.5\epsilon\sigma$  to  $+3.5\epsilon\sigma$ . The quantization characteristic can be represented by

$$y_8 = if\left[\left(\frac{|x|}{\epsilon}\right) < 4, y_{N(even)}, 3.5\epsilon \text{sign}(x)\right]. \quad (24)$$

Note that when using Eq. (24) it is necessary also to use Eq. (20). The quantization noise is represented by

$$q_8 = x - \alpha_1 y_8. \quad (25)$$

<sup>7</sup>In ALMA the digitized signal may be passed through a FIR filter before being correlated. The output of this filter is quantized to either 4 or 16 levels, with Nyquist or twice-Nyquist sampling at that point. A later memo will address this issue.

Also we can write,

$$\langle xy_8 \rangle = \sqrt{\frac{2}{\pi}} \left[ 0.5 \epsilon \int_0^\epsilon x e^{-x^2/2} dx + 1.5 \epsilon \int_\epsilon^{2\epsilon} x e^{-x^2/2} dx + 2.5 \epsilon \int_{2\epsilon}^{3\epsilon} x e^{-x^2/2} dx + 3.5 \epsilon \int_{3\epsilon}^\infty x e^{-x^2/2} dx \right] = 0.9670, \quad (26)$$

from which  $\alpha_1 = 1/0.9670 = 1.0341$ . The quantization characteristics corresponding to Eqs. (24) and (25) are shown in Fig. 8. For a rectangular passband, the signal to noise ratio for 8-level correlated signals is optimized with a level increment very close to<sup>8</sup>  $\epsilon = 0.60\sigma$  (Thompson et al. 2001: see Table 8.2), and we use this value in computation of the slope of the quantization noise. This slope is shown by a curve in Fig. 2, and for 20 dB input slope ratio the slope ratio for the quantization noise is approximately 1.05.

### 9. Accuracy of the quantized signals

With respect to the accuracy of the representation of the spectral slope, it is also of interest to examine briefly the response of the *quantized signal data* to the slope of the unquantized spectrum. For this we use the quantized data  $y$  rather than the inequality  $\Delta$  as the input to step 7 in the program. Results are shown in Fig. 10. The quantized data reproduce the slope of the unquantized data to an extent that increases with the number of quantization levels. The top curve shows the result obtained from the program when the quantization is omitted, that is, by going directly from step 3 to step 8 in the program, and fitting a straight line to the unquantized spectral data. Ideally, this should be a straight line from the lower left to the upper right corner of the graph, but instead the response has fallen by about 1 dB for a 20 dB input slope ratio. As mentioned above, the accuracy of the analysis is expected to decrease as the slope increases. This reduction of the actual slope values in the quantized representation of the data is, however, is not important in the investigation of the quantization noise, for which the slope ratios from the line-fitting do not exceed 1.1 in the results shown in Fig. 2.

1	2	3	4	5	6	7
Quantization type	$\eta_Q$ (anal.)	$\alpha_1$ (anal.)	$\alpha_1$ (sim.)	$\alpha_2$ (anal.)	$\alpha_2$ (sim.)	$\alpha_2/\alpha_1$ (sim.)
2-level	0.6366	1.2533	1.251	0.7980	0.797	0.637
3-level	0.8098	1.5114	1.510	1.2239	1.223	0.810
4-level	0.8825	0.5651	0.567	0.4987	0.499	0.880
8-level	0.9626	1.0341	1.035	0.9954	0.996	0.962

Table 1. Column 2, values of  $\eta_Q$  from Schwab (2005). Col. 3, analytical values of  $\alpha_1$ , from equations for  $\langle xy \rangle$  in the text. Col. 4,  $\alpha_1$  from numerical simulation, based on results using both slope and ripple modulation. Col. 5, analytical values of  $\alpha_2$  (product of values in Cols. 2 and 3). Col. 6,  $\alpha_2$  from numerical simulation. Col. 7, ratio of values in Cols. 4 and 6 from numerical simulation: compare with values in col. 2.

### 10. Values of $\alpha_1$ and $\alpha_2$

We have shown the relationship between the factors  $\alpha_1$  and  $\alpha_2$  to the quantization efficiency  $\eta_Q$  (Eq. (11) and how the relative variance of the quantization noise can be expressed in terms of these quantities (Eq. (12). Theoretical values of  $\alpha_1$  are derived for Gaussianly-distributed input data from equations for  $\langle xy \rangle$ , and theoretical values of  $\eta_Q$  are available from Schwab (2005) and from various analyses summarized in Thompson et al.(2001). From these data, and Eq. (11), theoretical values of  $\alpha_2$  can be obtained. Table 1

<sup>8</sup>A more precise value for maximization of  $\eta_Q$  is  $\epsilon = 0.586$ : see Table 3.

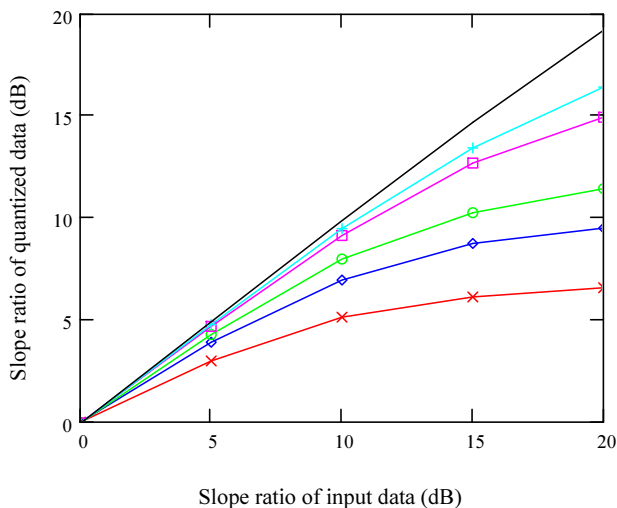


Figure 10: The ordinate is the slope ratio of the power spectrum determined from application of the line-fitting step of the simulation program to the quantized data (that is, the quantized signal  $y$ ). The abscissa is the slope ratio of the input data  $x$ . Moving upward from the lowest, the curves are for 2-level (red,  $\times$  crosses), 3-level (blue, diamonds), 4-level (green, circles), 8-level (magenta, squares) and N(even)-level (cyan, + crosses). The top curve (black, no symbols) shows the result obtained when quantization is omitted.

shows both theoretical values and values from numerical simulation for  $\alpha_1$  and  $\alpha_2$ . Values of the ratio  $\alpha_2/\alpha_1$  from the numerical simulation can be compared with the theoretical values of  $\eta_Q$ . In all cases the agreement is within the expected statistical uncertainties of the simulations and better than 0.5%. We take this as verification of the assumption in Eq. (7) that the quantized data can be expressed as a linear combination of the unquantized data and a component of random quantization noise.

### 11. Limitations on Variation of Gain with Frequency for EVLA and ALMA Systems.

Numerical simulation shows that modulation of the power spectrum of the unquantized input data in the form of slopes and sinusoidal ripples results in only minor, residual, reproductions of these features in the quantization noise spectrum. For two-level quantization, a slope ratio of 10 in the power spectral density of the input data results in a slope ratio of only 1.05 (see Fig. 2) in the quantization noise power, and this residual slope becomes smaller as the number of quantization levels increases. Uncertainties in these figures resulting from the random-noise nature of the input are of order 0.3%.

Consider a system in which the SNR is constant across the passband. Let  $P$  be the power spectral density at any point within the passband before quantization. Then, from Eq. (12), and with the assumption that the quantization noise is uniformly distributed in frequency, the noise level after quantization is  $P + \langle P \rangle (1/\eta_Q - 1)$ , where  $\langle P \rangle$  is the mean power level across the band. The SNR at the same frequency is proportional to  $P / [P + \langle P \rangle (1/\eta_Q - 1)]$ . For values  $P_1$  and  $P_2$  of the power spectral density at two points in the passband, the ratio of the values of SNR is

$$R_{SNR} = \left( \frac{P_1}{P_2} \right) \frac{P_2 + \langle P \rangle (\frac{1}{\eta_Q} - 1)}{P_1 + \langle P \rangle (\frac{1}{\eta_Q} - 1)}. \quad (27)$$

In the case of a linear slope in the power spectrum, one can put  $P_1 / \langle P \rangle = 1 + \delta$  and  $P_2 / \langle P \rangle = 1 - \delta$ , where  $\delta = (r - 1)/(r + 1)$ ,  $r$  being the slope ratio. For sinusoidal ripple across the passband, Fig. 11 shows the variation of  $R_{SNR}$ , in which  $P_1$  represents the levels at the ripple maxima and  $P_2$  the levels at the minima.

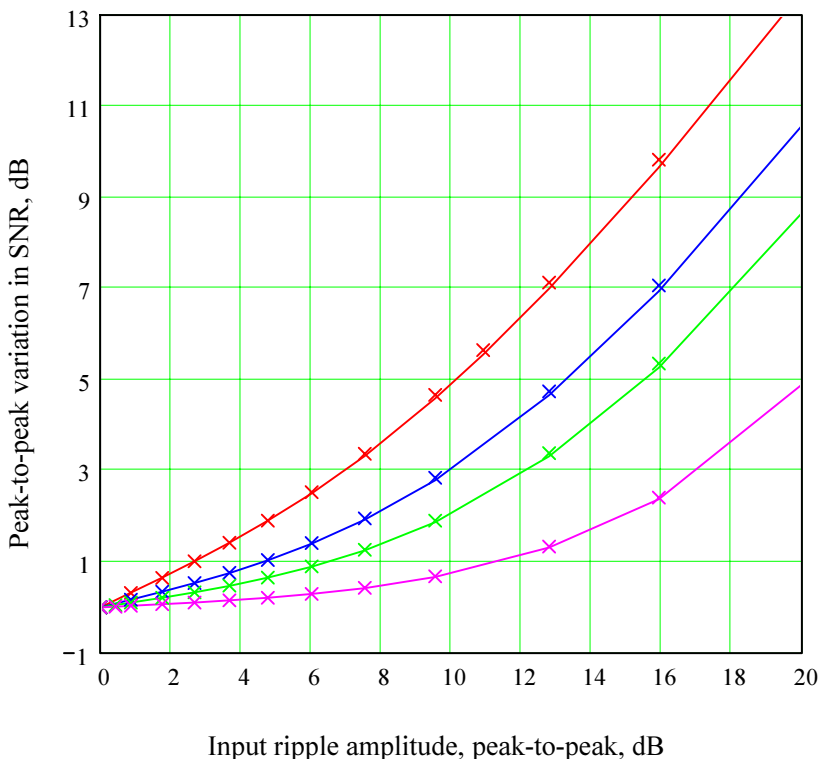


Figure 11: The abscissa is the ratio of the input power levels at the peaks and troughs of the ripple modulation, on a dB scale. The ordinate is the ratio of the highest to the lowest SNR in the power level of the resulting quantized signal, again on a dB scale. From the top down, the curves are for 2-level quantization (red), 3-level (blue), 4-level (green), and 8-level (magenta). The curves include the residual slope of the quantization noise determined from numerical simulation. The crosses are the corresponding values calculated with the assumption that the spectrum of the quantization noise is flat.

The solid curves show the result when residual ripple in the quantization noise is taken into account, based on the results of numerical simulation shown in Fig. 4. The crosses show the equivalent values for which the residual modulation is ignored, that is, the spectrum of the quantization noise is assumed to be flat. The crosses give slightly higher values, as can be seen for a few of the points. However, the differences are so small that for estimation of the tolerable slope or ripple in the data at the quantizer input, the assumption of a flat spectrum for the quantization noise is clearly justified. Although Fig. 11 was calculated for a gain variation in the form of a sinusoidal ripple, it can with reasonable precision be applied to any shape of variation across the receiver passband. As noted above, the amplitude of variations in quantization noise with frequency is dependent on the relative phase of gain variations across the passband, and whether multiple aliasing of higher frequency noise into the baseband enhances or partially cancels the variations. For cases in which cancellation occurs, the assumption of a flat quantization noise spectrum becomes an even better approximation than is indicated in Fig. 11.

The question of how the quantization noise relates to the tolerable variations in the passband response at the input to the quantizer is complicated by the wide range of measurements encountered in radio astronomy. The search for weak spectral lines is one of the activities in which good SNR across the full observing band is most important. A drop in sensitivity by a factor of  $\sqrt{2}$  (-1.5 dB) in some part of the spectrum could

require doubling of the observing time, which in practice is not always possible. Here we take 1 dB variation in SNR relative to the mean value as the maximum tolerable. If  $P_2$  in Eq. (27) represents the mean value  $\langle P \rangle$ , and  $P_1$  represents the minimum value  $P_{min}$ , then

$$R_{SNR} = \frac{\frac{P_{min}}{\langle P \rangle}}{\eta_Q \left( \frac{P_{min}}{\langle P \rangle} + \eta_Q^{-1} - 1 \right)}. \quad (28)$$

Here  $R_{SNR}$  is the ratio of the minimum SNR to the mean SNR, and

$$\frac{P_{min}}{\langle P \rangle} = \frac{1 - \eta_Q}{R_{SNR}^{-1} - \eta_Q}. \quad (29)$$

For 1 dB decrease in SNR relative to the mean,  $R_{SNR} = 0.7943$ , and the corresponding values of  $P_{min}/\langle P \rangle$  are given in Table 2 for different quantization schemes. These values represent the minimum tolerable level, with respect to the mean level, of the power spectral density at the quantizer input. They can be applied directly to a single-dish instrument. In the case of an array, the receiving channels for different antennas may contain both individual and common features of their deviations from the ideal flat response. For example, if signals are transmitted from the antennas to the correlator in analog form individual variation resulting from different transmission path lengths may be important, but these are avoided if the signals are digitized at the antennas. The individual deviations may tend to average down when all of the visibility data are combined. The values in Table 2 are more directly applicable to the common deviations, or to the mean frequency response of the different receiver channels. Thus, the application of the tolerances resulting from quantization noise to the individual frequency responses of an array is difficult to assess, until data on both the individual and common variation of the receiver channels from the ideal spectral response can be determined. A proposed hardware design for reduction of unwanted slopes in the analog frequency response is described by Hayward, Morgan, and Saini (2004).

Quantization type			
2-level	3-level	4-level	8-level
0.584	0.423	0.312	0.126
-2.34dB	-3.73 dB	-5.06 dB	-8.99 dB

Table 2. Lower limits by which the level at the correlator input can differ from the mean value across the passband to prevent variations in which the SNR is decreased by more than 1 dB. The upper figure in each column is the minimum power level as a fraction of the mean level and the lower figure is the same quantity in decibels. Both the values in this table and the points in Fig. 11 are based on Eqs. (27) and (28), but the data in the table and the figure are not directly comparable because one is based on minimum-to-mean ratios and the other on minimum-to-maximum ratios of the power spectral density.

**Acknowledgment** We particularly wish to thank F. R. Schwab for providing us with a pre-publication copy of his paper on quantization functions, which contains the first precise calculations of quantization efficiency for a wide range of quantization schemes that we are aware of, and for much helpful discussion.

### Appendix A. Precise Expressions for Quantization Efficiency.

The quantization efficiency  $\eta_Q$  can be defined as the ratio of the signal to noise amplitudes (SNR) at the output of the correlator of a digital system, divided by the same quantity for an analog system. The quantization efficiency is also equal to the variance of the noise at one input of the correlator for an analog system divided by the equivalent quantity for a digital system. A formula for  $\eta_Q$  for eight and higher numbers of levels, based on this ratio of the variances, can be found in Thompson (1998) and Thompson et al. (2001, see section 8.3). In this formula, however, the quantization inequality with  $\alpha = 1$  was used as

an approximation for the quantization noise, as defined here. This is a good approximation if the number of quantization levels is not too small. Also, a piecewise linear approximation of the Gaussian probability function was used to eliminate the need for numerical evaluation of a number of integrals. Using the definition of quantization noise in Eq. (6) we can now revise the earlier treatment to avoid any approximations.

Consider the case for an even number of equally spaced levels as discussed in Sections 7 and 8. It is first necessary to determine  $\alpha_1 = \langle xy \rangle^{-1}$ . Note that we use  $\langle x^2 \rangle = 1.0$ . The 8-level expression for  $\langle xy \rangle$  in Eq. (26) provides an example. For the general case it is convenient to define  $\mathcal{N} = N/2$ , i.e. half the number of levels. The values of  $x$  that fall within the quantization level between  $m\epsilon$  and  $(m+1)\epsilon$  are assigned values  $y = (m+1/2)\epsilon$  and their contribution to  $\langle xy \rangle$  is

$$\frac{1}{\sqrt{2\pi}} \int_{m\epsilon}^{(m+1)\epsilon} (m + \frac{1}{2})\epsilon x e^{-x^2/2} dx. \quad (30)$$

The contribution from the level between  $-m\epsilon$  and  $-(m+1)\epsilon$  is the same as the expression above, so we sum the integrals for the positive levels, include a factor of two, and take the reciprocal of the whole expression:

$$\alpha_1 = \langle xy \rangle^{-1} = \sqrt{\frac{\pi}{2}} \left[ \left( \sum_{m=0}^{\mathcal{N}-2} \int_{m\epsilon}^{(m+1)\epsilon} (m + \frac{1}{2})\epsilon x e^{-x^2/2} dx \right) + \int_{(\mathcal{N}-1)\epsilon}^{\infty} (\mathcal{N} - \frac{1}{2})\epsilon x e^{-x^2/2} dx \right]^{-1}, \quad (31)$$

The summation term contains one integral for each positive quantization level except the highest one. The integral on the right covers the highest level and the range of  $x$  above it, for both of which the assigned value is  $y = (\mathcal{N} - 1/2)\epsilon$ .

To evaluate the quantization noise, again consider first the contribution from values of  $x$  that fall within the quantization level between  $m\epsilon$  and  $(m+1)\epsilon$ . For this level the quantized data  $y$  all have the value  $(m+1/2)\epsilon$ , and the amplitude of the quantization noise is  $[x - \alpha_1(m+1/2)\epsilon]$ . The variance of the quantization noise for all values of  $x$  within this level is

$$\frac{1}{\sqrt{2\pi}} \int_{m\epsilon}^{(m+1)\epsilon} [x - \alpha_1(m + \frac{1}{2})\epsilon]^2 e^{-x^2/2} dx. \quad (32)$$

Since the variance for the level  $-m\epsilon$  to  $-(m+1)\epsilon$  is also equal to the expression above, we again include a factor of 2, sum over all positive quantization levels except the highest, and add a term for the highest level and the range of  $x$  above it. Thus the total variance of the quantization noise  $\langle q^2 \rangle$  is:

$$\langle q^2 \rangle = \sqrt{\frac{2}{\pi}} \left[ \sum_{m=0}^{\mathcal{N}-2} \left( \int_{m\epsilon}^{(m+1)\epsilon} [x - \alpha_1(m + \frac{1}{2})\epsilon]^2 e^{-x^2/2} dx \right) + \int_{(\mathcal{N}-1)\epsilon}^{\infty} [x - \alpha_1(\mathcal{N} - \frac{1}{2})\epsilon]^2 e^{-x^2/2} dx \right] \quad (33)$$

Since we are considering the case in which the variance of the input data  $x$  is equal to one, the total variance for the digitized data is  $1 + \langle q^2 \rangle$ , and  $\eta_Q = 1/(1 + \langle q^2 \rangle)$ . Thus,

$$\eta_Q = \left\{ 1 + \sqrt{\frac{2}{\pi}} \left[ \sum_{m=0}^{\mathcal{N}-2} \left( \int_{m\epsilon}^{(m+1)\epsilon} [x - \alpha_1(m + \frac{1}{2})\epsilon]^2 e^{-x^2/2} dx \right) + \int_{(\mathcal{N}-1)\epsilon}^{\infty} [x - \alpha_1(\mathcal{N} - \frac{1}{2})\epsilon]^2 e^{-x^2/2} dx \right] \right\}^{-1} \quad (34)$$

Equations (31) and (34) provide values of  $\eta_Q$  from starting values of  $\epsilon$  and  $\mathcal{N}$ , and can be evaluated rapidly in Mathcad. For the upper limit on the right-hand integrals in these equations it is satisfactory to use  $20\mathcal{N}\epsilon$ . The integrals can also be expressed in terms of the error function and exponential functions and Eq. (31) then reduces to:

$$\alpha_1 = \sqrt{\frac{\pi}{2}} \epsilon^{-1} \left[ \left( \sum_{m=1}^{\mathcal{N}-2} e^{-m^2\epsilon^2/2} \right) + e^{-(\mathcal{N}-1)^2\epsilon^2/2} + \frac{1}{2} \right]^{-1}. \quad (35)$$

Reduction of Eq. (34) results in a slightly more lengthy expression:

$$\eta_Q = \left\{ 1 + \sqrt{\frac{2}{\pi}} \left[ -2\alpha_1 \epsilon \left( \sum_{m=1}^{\mathcal{N}-2} e^{-m^2 \epsilon^2 / 2} \right) - 2\alpha_1 \epsilon e^{-(\mathcal{N}-1)^2 \epsilon^2 / 2} - \alpha_1 \epsilon \right] - 2\alpha_1^2 \epsilon^2 \left( \sum_{m=1}^{\mathcal{N}-2} m \operatorname{erf} \left( \frac{m\epsilon}{\sqrt{2}} \right) \right) + \left( 1 + \alpha_1^2 \epsilon^2 (\mathcal{N} - 1.5)^2 \right) \operatorname{erf} \left[ \frac{(\mathcal{N} - 1)\epsilon}{\sqrt{2}} \right] + \left( 1 + \alpha_1^2 \epsilon^2 (\mathcal{N} - 0.5)^2 \right) \operatorname{erfc} \left[ \frac{(\mathcal{N} - 1)\epsilon}{\sqrt{2}} \right] \right\}^{-1}. \quad (36)$$

Since no approximations are made, the same method can be used for cases where the number of quantization levels is not small. For example, for 3-level quantization the expression for  $\alpha_1$  is the reciprocal of Eq. (16), and the expression for  $\eta_Q$  is

$$\eta_Q = \left\{ 1 + \sqrt{\frac{2}{\pi}} \left[ \int_0^{0.612} x^2 e^{-x^2/2} dx + \int_{0.612}^{20} (x - \alpha_1)^2 e^{-x^2/2} dx \right] \right\}^{-1}. \quad (37)$$

Values of  $\alpha_1$ ,  $\epsilon$ , and  $\eta_Q$  derived using Eqs. (16), (31), (34), and (35) are shown in Table 3. In each case the values of  $\epsilon$  are chosen empirically to maximize  $\eta_Q$ . The values of  $\eta_Q$  are in exact agreement with the equivalent values derived by Schwab (2005).

Table 3

No. of Levels	$\mathcal{N}$	$\epsilon$	$\alpha_1$	$\eta_Q$
3			1.51144	0.809826
8	4	0.586	1.0389	0.962560
16	8	0.335	1.0117	0.988457
32	16	0.188	1.0035	0.996505
256	128	0.0308	1.000088	0.999912

## References

- Carlson, B. and Perley, R., Quantization Loss for a Sloped Passband, EVLA Memo. 83, NRAO, Aug. 30, 2004.
- Hayward, R., Morgan, M., and Saini, K., A gain Slope Correction Scheme for the EVLA Receiver System, EVLA Memorandum 80, NRAO, June 30, 2004.
- Jenet, F. A. and Anderson, S. B., The Effects of Digitization on Nonstationary Stochastic Signals with Applications to Pulsar Signal Baseband Recording, Pub. Astron. Soc. Pacific, 110, 1467-1478, 1998.
- Lamb, J. W., Bandslope Effects on Sensitivity in Interferometers with Digital Correlators, Alma Memo. 407, NRAO, Jan. 15, 2002.
- Schwab, F. R., Optimal Quantization Functions for Multi-Level Digital Correlators, in preparation, 2005.
- Thompson, A. R. MMA Memorandum 220 (ALMA Memo. series), NRAO, July 9, 1998.
- Thompson, A. R., Moran, J. M., and Swenson, G. W., Interferometry and Synthesis in Radio Astronomy, John Wiley, N.Y., 1986, 2001.
- Van Vleck, J. H. and Middleton, D., The Spectrum of Clipped Noise, Proc. IEEE, 54, 2-19, 1966.