Two-bit quantization for 0-dB gain

Ziad S. Haddad
Scott. Hensley
William T.K. Johnson

Jet Propulsion Laboratory, California Institute of Technology

Abstract

For space missions such as NASA’s Magellan mission to Venus and the Cassini mission to Saturn, the communication channel to Earth has a limited channel capacity. This limitation requires that the output of the radar be encoded. Block-adaptive quantization can achieve the required data rate reduction. However, coding the 8-bit radar samples within a given block using 2 bits affects the accuracy and the gain of the reconstructed signal. In this paper, we consider three different optimization criteria, and we compare the performance of the resulting codes.
1 Introduction

The Magellanspacecraft used a Synthetic Aperture Radar (SAR) to probe and image much of the surface of Venus. 011 successive orbits, reflected echoes from the SAR signal were collected around periapsis, and the raw data was retransmitted to earth during the remainder of the orbit.

The amount of 8-bit raw data collected to produce SAR imagery was enormous. Given the 806 KHz constraint on the data rate, and the nominal transmit time of 37.2 minutes, data compression was required to meet the science objectives. The data compression was accomplished by reducing each raw I/Q sample to 2 bits, one for sign and the other for magnitude. The value of the magnitude bit was determined by the estimated power for a block of samples, and the corresponding value transmitted to earth along with the compressed data samples. Grouping adjacent SAR samples into blocks is justified by the fact that the SAR's signal statistics are Gaussian, the variation in the power of the return echo being a slow function of range, and the fact that the echo power profile is very nearly periodic with respect to transmitted-pulse number. The encoding thresholds and decoding values for every given block of data samples were determined to satisfy certain optimization criteria whose aim was to keep the quantization noise low without introducing any gain to the signal.

For the Cassini mission to Saturn and Titan to begin in 1997, limited resources once again constrain the maximum data rates. To accomplish the science requirements and satisfy the resource constraints, a block-adaptive quantization algorithm similar to the one used for Magellan will be implemented. However, a close exam of the Magellan algorithm showed that its performance could be improved, in a way that would not have significantly affected the overall performance of the Magellan signal processing algorithm but that will be crucial for Cassini.

In this paper, we describe the optimization criteria one is naturally led to consider, and we compare the performance of the resulting quantization algorithms.

2 Mathematical Approach

The input to our A/n converter consists of amplified radar echoes plus receiver and thermal noise. Symbolically, the \( i \)-th radar sample \( S_m(i) \) within a block of \( M \) samples can be represented as a sum

\[
S_m(i) = s_R(i) + s_T(i) + \sum_{k=1}^{K} s_k(i), \quad 1 \leq i \leq M, \tag{1}
\]

where \( s_R \) represents receiver noise, \( s_T \) represents thermal noise, and each \( s_k(i) \) represents the backscattered signal received during the time interval corresponding to the \( i \)-th sample, from the \( k \)-th radar resolution bin. Each of the summands in the right-hand-side of (1) has 0-mean Gaussian
statistics. Therefore, $S_{in}$ itself is 0-mean Gaussian. Assuming that the input to the A/D converter is sampled using 8 bits, we would like to find the "optimal" 2-bit code to represent these samples. There are three distinct optimality criteria.

First, the direct criterion of Max([4]) and Lloyd([3]) can be considered. According to this approach, we must minimize the quantization noise, i.e. the r.m.s. difference

$$\sqrt{\mathbb{E} (S_{in} - S_{out})^2}$$

(2)

between the input 8-bit samples $S_{in}$ and the 4-level (2-bit representation) decoded output samples $S_{out}$. Since, in our case, in addition to minimizing the quantization noise, we need to make sure that the gain remains equal to one, after minimizing (2) we would still need to calculate the gain

$$G = \frac{\mathbb{E}\{S_{out}^2\}}{\mathbb{E}\{S_{in}^2\}}$$

that would result, and divide the output $S_{out}$ by $\sqrt{G}$. Let us call this the modified Max-Lloyd approach.

A somewhat different second approach consists in imposing the 0-dB gain requirement a priori. In this case, we need to find the code that minimizes the quantization noise (2), subject to the condition

$$G = \frac{\mathbb{E}\{S_{out}^2\}}{\mathbb{E}\{S_{in}^2\}} = 1$$

(3)

We call this the conventional-0-gain approach.

The third approach was originally proposed for the SAR on NASA's Magellan Mission ([2]). It is motivated by the observation that if a single given summand $s_k(i)$ in equation (1) is assumed to be deterministic rather than stochastic, say if $s_k(i) = \mu$ a constant (small) value, for all $i$ between 1 and $M$, then one might not be so much interested in enforcing condition (3) as in making sure that the constant value $\mu$ does not get altered by the quantization. Since, under such an assumption, the input signal $S_{in}$ would then be Gaussian with mean $\mu$, a real number relatively close to zero (compared to the variance of $S_{in}$), the third approach would seek to minimize expression (2) subject to the condition that the "linear gain"

$$\frac{\partial}{\partial \mathbb{E}\{S_{in}^2\}} \mathbb{E}\{S_{out}\}$$

(4)

representing the rate of change of the mean of the output with respect to the mean of the input when these means are close to zero, remain equal to one. We shall call this the linear-0-gain approach.

Let us now try to solve each of these three optimization problems. To fix the notation, call $\sigma$ the r.m.s. value of the input signal $S_{in}$, write $T$ for the encoding threshold, and $y_0$ and $y_1$ for the decoded output levels. Thus

$$S_{out} = \begin{cases} -y_1 & \text{if } S_{in} \leq -T \\ y_0 & \text{if } -T < S_{in} < 0 \\ -y_0 & \text{if } 0 \leq S_{in} < T \\ y_1 & \text{if } T \leq S_{in} \end{cases}$$

(5)
In this notation, the noise-to-signal ratio $N(T, y_0, y_1)$, given by the square of expression (2) divided by $\sigma^2$, can be written as

$$N = \left(\frac{y_0}{\sigma}\right)^2 \text{erf}\left(\frac{T}{\sigma\sqrt{2}}\right) + \left(\frac{y_1}{\sigma}\right)^2 \left(1 - \text{erf}\left(\frac{T}{\sigma\sqrt{2}}\right)\right) \cdot \frac{y_0}{\sigma} \sqrt{\frac{8}{\pi}} \left(1 - e^{-T^2/(2\sigma^2)}\right) - \frac{y_1}{\sigma} \sqrt{\frac{8}{\pi}} e^{-T^2/(2\sigma^2)} \tag{6}$$

For each of the three approaches outlined above, the problem is to determine the optimal values of $T$, $y_0$ and $y_1$ as functions of $\sigma$. We show in the appendix that the conventional 0-gain approach requires us to

A1) minimize $N(T, y_0, y_1)$ as represented by (6),

A2) subject to $\left(\frac{y_0}{\sigma}\right)^2 \text{erf}\left(\frac{T}{\sigma\sqrt{2}}\right) + \left(\frac{y_1}{\sigma}\right)^2 \left(1 - \text{erf}\left(\frac{T}{\sigma\sqrt{2}}\right)\right) \cdot 1$.

On the other hand, the linear 0-gain approach requires us to

B1) minimize $N(T, y_0, y_1)$ as represented by (6),

B2) subject to $\frac{y_0}{\sigma} \sqrt{\frac{8}{\pi}} \left(1 - e^{-T^2/(2\sigma^2)}\right) + \frac{y_1}{\sigma} \sqrt{\frac{8}{\pi}} e^{-T^2/(2\sigma^2)}$.

Again, the details are supplied in the appendix. Finally, the modified Max-Lloyd approach requires that we

C1) minimize $N(T, y_0, y_1)$ as represented by (6),

C2) compute $G = \left(\frac{y_0^*}{\sigma}\right)^2 \text{erf}\left(\frac{T^*}{\sigma\sqrt{2}}\right) + \left(\frac{y_1^*}{\sigma}\right)^2 \left(1 - \text{erf}\left(\frac{T^*}{\sigma\sqrt{2}}\right)\right)$, where $T^*$, $y_0^*$ and $y_1^*$ are the optimal values found in C1,

C3) then use $y_0^*/\sqrt{G}$ and $y_1^*/\sqrt{G}$ as our decoding levels.

As we show in the appendix, it turns out that the optimal value for the encoding threshold $T$ is the same in all three cases: it is given by $T = \sqrt{2}\sigma x$, where $x$ is that positive real number which solves the equation

$$x \frac{e^{-x^2}}{1 - \text{erf}(x)} - x \frac{1 - e^{-x^2}}{\text{erf}(x)} + \frac{1}{2\sqrt{\pi}} \left(\frac{1 - e^{-x^2}}{\text{erf}(x)}\right)^2 - \frac{1}{2\sqrt{\pi}} \left(\frac{e^{-x^2}}{1 - \text{erf}(x)}\right)^2 = 0.$$

The solution is $x \simeq 0.6941$, so that $T \simeq 0.9816\sigma$. 
As to the decoding levels, we show in the appendix that the optimal levels minimizing $N$, i.e. satisfying C1, are given by

\[ y_0^* = \frac{2 \sqrt{\frac{2}{\pi}} \left( 1 - e^{-x^2} \right)}{\pi \text{erf}(x)} \sigma \simeq 0.4528 \sigma \]

\[ y_1^* = \sqrt{\frac{2}{\pi}} \left( 1 - e^{-x^2} \right) \sigma \simeq 1.5104 \sigma \]  \hspace{1cm} (8)

Since the gain is then given by $G = 2 \left( \frac{1 - e^{-x^2}}{\text{erf}(x)} \right) \sigma \simeq 0.8825$, the optimal decoding levels for the modified Max-Lloyd approach (i.e. satisfying C3) are given by

\[ y_0 = \frac{y_0^*}{\sqrt{G}} \left( 1 - e^{-x^2} \right) \sigma \simeq 0.482 \sigma \]  \hspace{1cm} (9)

\[ y_1 = \frac{y_1^*}{\sqrt{G}} \left( 1 - e^{-x^2} \right) \sigma \simeq 1.6078 \sigma \]  \hspace{1cm} (10)

We show in the appendix that the decoding levels given by formulas (10) and (11) are also the optimal levels for the conventional 0-gain approach. Thus these two approaches lead to the same encoding and decoding values. The optimal decoding levels for the linear 0-gain approach, however, are different. It turns out that B1 and B2 imply that, in that case, the optimal levels $y_0'$ and $y_1'$ are given by

\[ y_0' = \frac{y_0^*}{G} \left( 1 - e^{-x^2} \right) \sqrt{\frac{2}{\pi}} \left( \frac{1 - e^{-x^2}}{\text{erf}(x)} \right) \sigma \simeq 0.513 \sigma \]  \hspace{1cm} (11)

\[ y_1' = \frac{y_1^*}{G} \left( 1 - e^{-x^2} \right) \sqrt{\frac{2}{\pi}} \left( \frac{1 - e^{-x^2}}{\text{erf}(x)} \right) \sigma \simeq 1.7115 \sigma \]  \hspace{1cm} (12)

Thus the conventional 0-gain approach implies lower decoding levels (formulas (10) and (11)) than the linear 0-gain gain approach (formulas (12) and (13)). In the case of a SAR signal, it is clear from (1) that the samples $S_m(i)$ to be coded have 0-mean Gaussian statistics. The performance of the coding algorithm will therefore depend on the variance of $S_m$, not on the variance or the possibly non-0 mean of any individual summand $s_k$, precisely because the coding is a non-linear process. Thus, in our case, the linear 0-gain criterion is of little relevance: the optimal approach is to use the conventional 0-gain constraint. In the following section, we make these considerations more quantitatively precise.
3 Results

With block-adaptive quantizers, the r.m.s. value of the input samples is computed assuming all samples within a given block have identical statistics. In practice, one can expect consistent fluctuations within a block, especially when the block falls at the leading or trailing edge of a radar pulse window. Figure 1a shows the noise-to-signal ratio $N$ when the r.m.s. level $\sigma_{ac}$ of the actual input signal is different from the design r.m.s. level $\sigma$. Assuming that the quantization is performed according to the classical Max-Lloyd algorithm with 110 gain adjustment. Figure 1b shows the gain as a function of $\sigma_{ac}/\sigma$ in this case. Note that the lowest noise level, -9.31 dB, is achieved when $\sigma_{ac}$ is exactly equal to the design $\sigma$, as expected. The corresponding gain is -0.53.5 dB. Figure 2a shows a plot of the noise-to-signal ratio $N$ as a function of $\sigma_{ac}/\sigma$ when the quantization is optimized according to our conventional 0-gain criterion (or, equivalently, according to the modified Max-Lloyd criterion). Figure 2b shows a plot of the gain as a function of $\sigma_{ac}/\sigma$ in that case. As expected, the gain when $\sigma_{ac} = \sigma$ is exactly 0 dB. The corresponding noise level, -9.173 dB, has slightly degraded. Figures 3a and 3b show the noise-to-signal ratio and the gain, respectively, when the quantization is optimized according to the linear-0-gain criterion. Note that, in that case, when $\sigma_{ac}$ is exactly equal to $\sigma$, the gain is 0.55 dB and the quantization-noise-to-signal ratio is -8.76 dB.

In practice, efficiency dictates that a discretization of the function which pairs a particular computed input-level r.m.s. value $\sigma = \mathcal{E}\{S_{in}^2\}$ with the appropriate encoding threshold $T$ (and hence to the decoding levels $y_0$ and $y_1$) should be easily computed and stored in the spaceborne instrument. In fact, rather than using the estimated root-mean-squared value of the input, one can use the more readily computable estimated Inset-all absolute value $\varsigma = \mathcal{E}\{|S_{in}|\}$. For a 0-mean Gaussian input signal $S_{in}$ sampled using 8 bits, one can easily check that $\sigma$ and $\varsigma$ are related by (12)

$$\varsigma = 127.5 - \sum_{n=1}^{127} \text{erf}\left(\frac{n}{\sigma \sqrt{2}}\right)$$

(14)

The noise-to-signal ratio curves allow one to compute the best discretization of the $\varsigma \to T$ correspondence. Figure 4 shows the discretized $\varsigma \to T$ pairing used for the Magellan mission's SAR, and figure 5 shows the pairing to be used for the Cassini mission's SAR. For Magellan, $T$ was represented as a 7-bit word. In the case of Cassini, $T$ is represented as a (7-integer-bits)+1-fractional-bit word. While the different optimization criteria for Magellan and Cassini produce the same thresholds, the noise-to-signal curves are slightly different. Hence the resulting $\varsigma \to T$ correspondence is slightly different, mostly at the upper end of the input dynamic range, where the input level is well into the saturation region.
4 Acknowledgements

This work was performed at the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration.

5 Appendix

If $S_{in}$ is assumed 0-mean Gaussian with variance $\sigma^2$, the r.m.s. quantization noise (\(\sigma\)) produced by the algorithm described by equation (5) is given by

$$\left( \int_{-\infty}^{\infty} (-y_1 S)^2 p(S) dS + \int_{-T}^{0} (-y_0 - S)^2 p(S) dS \cdot \int_{-T}^{0} (y_0 - S)^2 p(S) dS \cdot \int_{T}^{\infty} (y_1 - S)^2 p(S) dS \right)^{1/2}$$

where $p(S) = \frac{1}{\sqrt{2\pi}\sigma^2} e^{-0.5S^2/\sigma^2}$. When the integrals are calculated, and the result squared and divided by $\sigma^2$, one obtains formula (6) for the Noisc-to-signs ratio $N$.

According to our conventional 0-gain approach, we then need to minimize $N$ subject to the requirement that the gain (3) be equal to 1. The gain is given by

$$G = \frac{1}{\pi}\cdot \left( \int_{-\infty}^{T} (-y_1)^2 p(S) dS + \int_{-T}^{0} (-y_0)^2 p(S) dS + \int_{T}^{\infty} (y_0)^2 p(S) dS + \int_{T}^{\infty} (y_1)^2 p(S) dS \right)$$

whence requirements A1 and A2. To solve the optimization problem in this case, write $x$ for $T/(\sigma\sqrt{2})$, $y_0/\sigma$, and $z_1/y_1/\sigma$. We must then minimize

$$z_0^2 \text{erf}(x) + z_1^2 (1 - \text{erf}(x)) - z_0 \sqrt{\frac{8}{\pi}} (1 - e^{-x^2}) - z_1 \sqrt{\frac{8}{\pi}} e^{-x^2}$$

subject to the condition

$$z_0^2 \text{erf}(x) + z_1^2 (1 - \text{erf}(x)) = 1$$

Using the Lagrange multiplier $\lambda$, the condition that the partial with respect to $z_0$ be 0 gives

$$z_0 \left( 1 - e^{-x^2} \right) - \frac{2}{\lambda} \text{erf}(x) \left( 1 - \frac{1}{\lambda} \right)$$

the condition that the partial with respect to $z_1$ be 0 gives

$$z_1 = \frac{e^{-x^2}}{2(1 - \text{erf}(x)) \lambda}$$
the condition that the partial with respect to \( x \) be 0 gives

\[
x - \frac{e^{-x^2}}{1 - \text{erf}(x)} - x \left[ \frac{1 - e^{-x^2}}{\text{erf}(x)} \right] + \frac{1}{2} \frac{1 - e^{-x^2}}{\sqrt{\pi}} \left\{ \frac{1}{2} \sqrt{\pi} \left( \frac{e^{-x^2}}{1 - \text{erf}(x)} \right)^2 \right\} = 0,
\]

(7)

and condition (17) becomes

\[
\sqrt{2} \frac{1}{\text{erf}(T)} + \frac{1}{1 - \text{erf}(x)} = \lambda
\]

(20)

It is now easy to solve (18), (19), (7) and (20) simultaneously. In fact, equation (7) is independent of the others and directly gives \( x \), hence \( T \). And when expression (20) is used for \( \lambda \) in (18) and (19), one finds that

\[
z_0 = \frac{1 - e^{-x^2}}{\text{erf}(x)} \cdot \left( \frac{1 - e^{-x^2}}{\text{erf}(x)} \right)^{-1/2}
\]

(21)

\[
z_1 = \frac{1 - e^{-x^2}}{1 - \text{erf}(x)} \cdot \left( \frac{1 - e^{-x^2}}{\text{erf}(x)} \right)^{-1/2}
\]

(22)

(see equations (10) and (11) section 2).

On the other hand, according to the linear 0-gain approach, we need to minimize \( N \) subject to the requirement that the “linear gain” (4) be equal to 1. With \( p_\mu(S) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-0.5(S - \mu)^2/\sigma^2} \), and calling \( \mu_{out} \) the mean value \( E\{S_{out}\} \) of the output signal \( S_{out} \), this “gain” is given by

\[
\frac{\partial \mu_{out}}{\partial \mu} = \frac{\partial}{\partial \mu} \left( \int_{-\infty}^{-T} (-y_1)p_\mu(S)dS + \int_{-T}^{0} (-y_0)p_\mu(S)dS + \int_{0}^{T} y_0 p_\mu(S)dS + \int_{T}^{\infty} y_1 p_\mu(S)dS \right)
\]

(23)

\[
= z_0 \sqrt{\frac{8}{\pi}} \left( 1 - e^{-x^2} \right) + z_1 \sqrt{\frac{8}{\pi}} e^{-x^2} \text{ when } \mu = 0,
\]

(24)

whence requirements B1 and B2 (we are still writing \( x \) for \( T'/(0/2) \), \( z_0 \) for \( y_0/\sigma \), and \( z_1 \) for \( y_1/\sigma \)). To solve the optimization problem in this case, we must minimize

\[
z_0^2 \text{erf}(x) + z_1^2 \left( 1 - \text{erf}(x) \right) - z_0 \sqrt{\frac{8}{\pi}} \left( 1 - e^{-x^2} \right) - z_1 \sqrt{\frac{8}{\pi}} e^{-x^2}
\]

(25)

subject to the condition

\[
z_0 \sqrt{\frac{8}{\pi}} \left( 1 - e^{-x^2} \right) + z_1 \sqrt{\frac{8}{\pi}} e^{-x^2} = 1.
\]

(26)

Using the Lagrange multiplier \( \nu \), the condition that the partial with respect to \( z_0 \) be 0 gives

\[
z_0 = \frac{1 - e^{-x^2}}{2 \text{erf}(x)} (-\nu).
\]

(27)
the condition that the partial with respect to \( z_1 \) be 0 gives
\[
z_1 = \frac{c^{-z^2}}{2(1 - \text{erf}(x))}, \quad (28)
\]
the condition that the partial with respect to \( x \) be 0 gives
\[
x \frac{c^{-z^2}}{1 - \text{erf}(x)} - x \frac{1 - c^{-z^2}}{\text{erf}(x)} + \frac{1}{2\sqrt{\pi}} \left( \frac{1 - c^{-z^2}}{\text{erf}(x)} \right)^2 - \frac{1}{2\sqrt{\pi}} \left( \frac{c^{-z^2}}{1 - \text{erf}(x)} \right)^2 = 0, \quad (7)
\]
and condition (26) becomes
\[
- \frac{1}{\sqrt{2}\pi} \left( \frac{c^{-z^2}}{1 - \text{erf}(x)} \right) = \frac{(c^{-z^2})^2}{1 - \text{erf}(x)} = \frac{1}{\nu} \quad (29)
\]
It is now easy to solve (27), (28), (7) and (29) simultaneously. Again, equation (7) is independent of the others and directly gives \( x \), hence \( 7' \). And when expression (29) is used for \( \nu \) in (27) and (28), one finds that
\[
z_0' = \frac{1 - c^{-z^2}}{\text{erf}(x)} \cdot \int \frac{\pi}{2} \left( \frac{1 - c^{-z^2}}{\text{erf}(x)} \right)^2 \left( \frac{c^{-z^2}}{1 - \text{erf}(x)} \right) \quad (30)
\]
\[
z_1' = - \frac{c^{-z^2}}{1 - \text{erf}(x)} \sqrt{\frac{\pi}{2}} \left( \frac{1 - c^{-z^2}}{\text{erf}(x)} \right)^2 \left( \frac{c^{-z^2}}{1 - \text{erf}(x)} \right) \quad (31)
\]
where equations (12) and (13) in section 2.

Finally, if we use the classic Max-Lloyd approach, we need to minimize \( N \), then calculate the gain and normalize by its square-root (requirements C1, C2 and C3). Still using \( x \) for \( f'(\sigma\sqrt{2}) \), \( z_0 \) for \( y_0/\sigma \), and \( z_1 \) for \( y_1/\sigma \), the condition that the partial of \( N \) with respect to \( z_0 \) be 0 gives
\[
z_0^* = \sqrt{\frac{2}{\pi}} \frac{1 - c^{-z^2}}{2\text{erf}(x)} \quad (32)
\]
the condition that the partial of \( N \) with respect to \( z_1 \) be 0 gives
\[
z_1^* = \sqrt{\frac{2}{\pi}} \frac{c^{-z^2}}{2(1 - \text{erf}(x))} \quad (33)
\]
(see equations (8) and (9) in section 2), and the condition that the partial of \( N \) with respect to \( x \) be 0 gives once again
\[
x \frac{c^{-z^2}}{1 - \text{erf}(x)} - x \frac{1 - c^{-z^2}}{\text{erf}(x)} + \frac{1}{2\sqrt{\pi}} \left( \frac{1 - c^{-z^2}}{\text{erf}(x)} \right)^2 - \frac{1}{2\sqrt{\pi}} \left( \frac{c^{-z^2}}{1 - \text{erf}(x)} \right)^2 = 0. \quad (7)
\]
The resulting gain is given by
\[ G = \frac{2}{\pi} \left( \frac{(1 - e^{-x^2})^2}{\text{erf}(x)} + \frac{(e^{-x^2})^2}{1 - \text{erf}(x)} \right) \] (34)

hence the decoding levels in this case are
\[ z_0 = \frac{1 - e^{-x^2}}{\text{erf}(x)} \left( \frac{(1 - e^{-x^2})^2}{\text{erf}(x)} + \frac{(e^{-x^2})^2}{1 - \text{erf}(x)} \right)^{-1/2} \] (35)
\[ z_1 = \frac{e^{-x^2}}{1 - \text{erf}(x)} \left( \frac{(1 - e^{-x^2})^2}{\text{erf}(x)} + \frac{(e^{-x^2})^2}{1 - \text{erf}(x)} \right)^{1/2} \] (36)

whence equations (10) and (11) in section 2.

References


Figure 2b
Figure 3a


Figure 3b