NATIONAL RADIO ASTRONOMY OBSERVATORY
Charlottesville, Virginia

ELECTRONICS DIVISION TECHNICAL NOTE NO. 213

Word-Boundary Detection in a Serialized, Gaussian-Distributed, White-Noise Data Stream

Matt Morgan
Rick Fisher

October 13, 2009

Word-Boundary Detection in a Serialized, Gaussian-Distributed, White-Noise Data Stream

Matt Morgan
Rick Fisher
10/13/2009

Introduction

In a previously-published white paper [1], the authors introduced the concept of a point-to-point fiber-optic digital data link for radio astronomy receivers with minimal processing overhead on the transmit side. A block diagram of what this might look like is shown in Figure 1.

Unlike conventional digital fiber-optic links, which employ power-intensive formatting, framing, and encoding operations at the transmitter to manage and maintain the link, the unique statistical characteristics of radio astronomy 'noise' renders many of these techniques unnecessary, and allows the maintenance of the link to be performed entirely at the receive-end using the data itself as the diagnostic input.

Bit-scramblers, for example, are not needed to guarantee logic-level transitions for clock-recovery. The noise from the receiver alone will be sufficient to ensure transitions in the data with sufficient frequency.

In this memo, we turn our attention to a somewhat more challenging problem, namely the detection of word boundaries in a continuous, serialized data stream. We assume the size of the word in bits is known a priori.

Since all sample values are theoretically possible, albeit with varying likelihood, examination of a single word-length is not sufficient to reliably determine the offset of the word boundaries in the data. Instead, a large number of word-lengths will have to be processed in order to accumulate statistical certainty before a positive detection is made and the required bit-shift is put in place. The link may then be considered synchronous with respect to word boundaries. Whether re-synchronization is performed continuously, once during startup, or periodically (as in a routine calibration) is an operational detail that is beyond the scope of this memo.

Assumptions

We make two assumptions about the analog signal statistics. First, that it is noisy (random) with a Gaussian-distribution, and second, that it is white, guaranteeing that consecutive samples are uncorrelated. Together, these conditions provide a mathematical basis on which to design word-detection strategies and then to evaluate their performance.

It is worth acknowledging briefly that neither of these conditions is strictly true in an exact sense. The spectrum will inevitably include non-Gaussian components at some level, and there will always be some variation in noise power across the receiver's instantaneous bandwidth due to band-limiting filters, RFI, and the astronomical signal of interest. The impact of these effects needs to be investigated with numerical experiments, but the present analysis is believed to be sufficiently accurate for the great majority of real-world cases to be encountered.

Under these assumptions, the probability distribution of the analog signal voltage for any sample is given by

$$p(v) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(v-\mu)^2}{2\sigma^2}\right) \tag{1}$$

where v is the instantaneous analog voltage, μ is the mean value, and σ^2 is the variance. The probability that the signal will fall between v_1 and v_2 at any one instant is given by

$$\int_{v_1}^{v_2} p(v)dv = \frac{1}{\sigma\sqrt{2\pi}} \int_{v_1}^{v_2} \exp\left(-\frac{(v-\mu)^2}{2\sigma^2}\right) dv \tag{2a}$$

$$= \frac{1}{\sqrt{\pi}} \int_{\frac{v_1-\mu}{\sigma\sqrt{2}}}^{\frac{v_2-\mu}{\sigma\sqrt{2}}} e^{-t^2} dt \tag{2b}$$

$$= \frac{1}{\sqrt{\pi}} \int_{\frac{v_1-\mu}{\sigma\sqrt{2}}}^{\frac{v_2-\mu}{\sigma\sqrt{2}}} e^{-t^2} dt \tag{2b}$$

$$= \frac{1}{2}\,\mathrm{erf}\left(\frac{v_2-\mu}{\sigma\sqrt{2}}\right) - \frac{1}{2}\,\mathrm{erf}\left(\frac{v_1-\mu}{\sigma\sqrt{2}}\right). \tag{2c}$$

These probabilities have been tabulated in Figure 2 for 32 bins using the normalization

$$a = \frac{v_{max}-v_{min}}{2\sigma\sqrt{2}} = \frac{v_0 2^N}{2\sigma\sqrt{2}} = \frac{v_0 2^{N-1}}{\sigma\sqrt{2}} \tag{3}$$

where v_{max}- v_{min} is the analog input range of the sampler, v_0 is the threshold voltage for each level, and N is the number of bits per sample. It has further been assumed that there is no offset from center. Small offset errors, less than one sampler threshold, should not affect the statistics appreciably. On the right side of the figure are the 5-bit binary representations of the sampler output in several binary formats. This diagram is for illustration only, as we make no assumptions at this stage about the number of bits per word.

Finally, we assume without loss of generality that the data is serialized in "little-endian" fashion, meaning that the least significant bit in each word, b_0, is transmitted first, followed by the other bits, b_1...b_{N-1}, in order of increasing significance. This is the most common convention for digital serial links, however the analysis that follows applies equally-well to "big-endian" data streams, provided the bits are processed in the reverse order, and where the text makes reference to the "preceding" or "following" word in a sequence, the opposite sense is understood instead.

Algorithm

The algorithm is as follows. The incoming serialized data stream is divided up into N-bit chunks, where N is the number of bits per word. These usually will not represent actual samples, since the word boundaries have not yet been detected. The goal of the algorithm is to determine the offset of the actual N-bit samples within the N-bit chunks. For each bit position within a chunk, a score is given. The bit position receives 1 point if a specified logical test is TRUE, and 0 points if the test result is FALSE. The test will depend on the binary format used. The score is then tallied over some large number of chunks, M, and the bit position with the highest score is declared the most significant bit (MSB) of the actual samples.

The performance of the above algorithm can be evaluated by first considering the probability that a given bit will yield a TRUE in the logical test defined for the given binary format. Denote this probability for bit k in the actual samples as p_k.

The point awarded to bit k in chunk i is then a Bernoulli Distributed random variable, $X_{i,k}$, with mean and variance given by

$$E\{X_{i,k}\} = p_k \tag{4a}$$
$$Var\{X_{i,k}\} = p_k(1-p_k) \tag{4b}$$

(Note that the index k refers to the significance of the bit in the actual samples, where k=0 corresponds to the least significant bit and k=N-1 corresponds to the most significant bit. The position of these bits within chunk, i, is unknown until completion of the algorithm.)

The score, Y_k, for bit position k after tallying M chunks is

$$Y_k = \sum_{i=1}^{M} X_{i,k} \tag{5}$$

By the Central Limit Theorem, we know that for large M the probability distribution of Y is Gaussian, with mean and variance given by

$$E\{Y_k\} = Mp_k \tag{6a}$$
$$Var\{Y_k\} = Mp_k(1-p_k) \tag{6b}$$

The probability of the algorithm failing, P_f, is the probability that the score for the MSB is smaller than for one of the other bits,

$$P_f = P\{Y_{N-1} \le Y_k \mid k < N-1\} \cong \frac{1}{2}\sum_{k=1}^{N-2} \text{erfc}\left(\frac{E\{Y_{N-1}\}-E\{Y_k\}}{\sqrt{2(Var\{Y_{N-1}\}+Var\{Y_k\})}}\right) \tag{7a}$$
$$= \frac{1}{2}\sum_{k=1}^{N-2} \text{erfc}\left(\frac{\sqrt{M}(p_{N-1}-p_k)}{\sqrt{2(p_{N-1}(1-p_{N-1})+p_k(1-p_k))}}\right) \tag{7b}$$

The approximation holds when the probability of failure is small, and is conservative (that is, it overestimates the probability of failure by double-counting the cases where more than one wrong bit has a higher score than the most significant bit.) A useful upper bound for the complementary error function is

$$\operatorname{erfc}(x) < \frac{2e^{-x^2}}{\sqrt{\pi}\left(x+\sqrt{x^2+\frac{4}{\pi}}\right)}, \quad x > 0 \tag{8}$$

We now evaluate the performance of this algorithm for three common binary formats -- sign-magnitude, straight binary, and two's complement.

Sign-Magnitude

Although sign-magnitude notation is rarely used in this context, it is a fairly simple case to analyze, so it will be treated here for completeness.

First, we must define a logical test that preferentially results in TRUE when applied to the sign bit (which is considered the most significant for the purposes of this analysis).

It is evident from inspection of Figure 2 that the most significant bit after the sign bit is almost always zero near the center of the voltage scale, in the sample words with the highest probability of occurrence. The latter bits take on the value zero with lesser frequency. We might then decide to use "b_{k-1}=0" as our logical test for scoring each bit, however since all the high order bits assume a value of zero with high probability, that test doesn't discriminate between them very well, especially at low analog power levels where only the inner most sample codes are ever used. Several bit positions could easily end up with almost identical scores. Instead, we add to the test the criterion that the current bit, b_k, must be one. Although the sign bit fails this test roughly half the time, it fails the first few high order bits with greater regularity. Therefore,

$$X_{i,k}^{sm} = \begin{cases} 1, & b_k = 1, b_{k-1} = 0 \\ 0, & else \end{cases} \tag{9}$$

where the superscript "sm" refers to the sign-magnitude format. Throughout this document, subscripts shall be understood to be cyclic, so that

$$b_{k<0} = b_{k+N} \tag{10}$$

The bits which pass the test in Equation (9) are shaded in the figure, red if the bit pattern is contained within the word, and blue if it crosses over a word boundary. The probability of a given bit receiving a point is

$$p_k^{sm} = P\{b_k = 0\} \tag{11}$$

and can be read off of Figure 2 for particular cases as follows

$$p_{N-1}^{sm} = \tfrac{1}{2}\operatorname{erf}\left(\tfrac{1}{2}a\right) \tag{12a}$$
$$p_{N-2}^{sm} = \operatorname{erf}\left(\tfrac{3}{4}a\right) - \operatorname{erf}\left(\tfrac{1}{2}a\right) \tag{12b}$$
$$p_{N-3}^{sm} = \operatorname{erf}\left(\tfrac{7}{8}a\right) - \operatorname{erf}\left(\tfrac{3}{4}a\right) + \operatorname{erf}\left(\tfrac{3}{8}a\right) - \operatorname{erf}\left(\tfrac{1}{4}a\right) \tag{12c}$$

By recognizing the above pattern, we can write a formula for the more general case as

$$p_{N-1-s}^{sm} = \begin{cases} \tfrac{1}{2}\operatorname{erf}\left(\tfrac{1}{2}a\right) & s = 0 \\ \sum_{i=1}^{2^{s-1}}\left[\operatorname{erf}\left(2^{-s}\left(2i-\tfrac{1}{2}\right)a\right) - \operatorname{erf}(2^{-s}(2i-1)a)\right] & 1 \le s \le N-2 \\ \tfrac{1}{2} - \tfrac{1}{2}\sum_{i=1}^{2^{N-2}}\left[\operatorname{erf}(2^{-N}(4i-2)a) - \operatorname{erf}(2^{-N}(4i-4)a)\right] & s = N-1 \end{cases} \tag{13}$$

where the first case, s=0, corresponds to the sign bit. This is plotted in Figure 3 as a function of signal strength.

Note that when the signal strength is very high, the probability of the sign bit scoring a point drops off rapidly. This corresponds to the saturation of the sampler, in which case the outermost sample codes begin to occur even more frequently than those in the center. The algorithm will fail in that situation. This is not a useful operating point for the sampler, however, and should almost never occur in practice. The nominal signal strength for radio astronomy is usually optimized for quantization noise, which in most cases will put the operating point on the left side of the plot, well below the saturation crossover point, where there is a strong statistical bias for the sign bit. A notable exception is two-bit sampling (N=2), which will be discussed later.

Straight Binary

Let us now consider serialized data streams in straight binary (or offset binary) format, shown in the second column of Figure 2. The patterns of 0's and 1's have odd symmetry about the center of the probability distribution, so any bit, no matter what it's significance, will assume both values with equal probability. Therefore, the method described above for the sign-magnitude case will not apply.

Instead, we note that in the most likely sampler outputs, those in the center of the sampler range, the two most significant bits differ, whereas the next few bits are the same. Therefore, the logical test we use for the straight binary case is that the current bit differs from the preceding bit,

$$X_{i,k}^{sb} = \begin{cases} 1, & b_k \neq b_{k-1} \\ 0, & b_k = b_{k-1} \end{cases} \tag{14}$$

As above, the probability of given bit scoring a point can be read off the figure

$$p_{N-1}^{sb} = \operatorname{erf}\left(\tfrac{1}{2}a\right) \tag{15a}$$
$$p_{N-2}^{sb} = \operatorname{erf}\left(\tfrac{3}{4}a\right) - \operatorname{erf}\left(\tfrac{1}{4}a\right) \tag{15b}$$
$$p_{N-3}^{sb} = \operatorname{erf}\left(\tfrac{7}{8}a\right) - \operatorname{erf}\left(\tfrac{5}{8}a\right) + \operatorname{erf}\left(\tfrac{3}{8}a\right) - \operatorname{erf}\left(\tfrac{1}{8}a\right) \tag{15c}$$

or in other words,

$$p_{N-1-s}^{sb} = \sum_{i=1}^{2^s}(-1)^{i-1}\,\mathrm{erf}\left(\left(1-2^{-s}\left(i-\tfrac{1}{2}\right)\right)a\right) \tag{16}$$

where

$$0 \le s \le N-2. \tag{17}$$

When s=N-1, the two bits being compared extend over a word boundary between the current word and the previous word (using little-endian bit order). The probability of a the two bits differing in this case is simply one half, owing to the fact that 0's and 1's are equally likely in all bit positions and subsequent words in the data stream are uncorrelated. Therefore,

$$p_{N-1-s}^{sb} = \begin{cases} \sum_{i=1}^{2^s}(-1)^{i-1}\,\mathrm{erf}\left(\left(1-2^{-s}\left(i-\tfrac{1}{2}\right)\right)a\right) & 0 \le s \le N-2 \\ 0.5 & s = N-1 \end{cases} \tag{18}$$

This is plotted in Figure 4, which clearly shows a strong statistical bias for the most significant bit (k=N-1) to the left of the saturation crossover point.

Two's Complement

The sample codes for two's complement notation are shown in the final column of Figure 2. Unlike straight binary, the first two bits are equal in the most common samples rather than different. In fact, that is the only difference between the straight binary and two's complement sample codes. Our logical test for two's complement then will be that the current bit matches the preceding bit, and does not match the following bit,

$$X_{i,k}^{tc} = \begin{cases} 1, & b_k = b_{k-1} \ne b_{k+1} \\ 0, & else \end{cases} \tag{19}$$

Once again, the probability of a bit scoring is easiest to read off the figure,

$$p_{N-1}^{tc} = \tfrac{1}{2}\mathrm{erf}\left(\tfrac{1}{2}a\right) \tag{20a}$$
$$p_{N-2}^{tc} = 1-\mathrm{erf}\left(\tfrac{3}{4}a\right) \tag{20b}$$
$$p_{N-3}^{tc} = \mathrm{erf}\left(\tfrac{5}{8}a\right)-\mathrm{erf}\left(\tfrac{3}{8}a\right) \tag{20c}$$
$$p_{N-4}^{tc} = \mathrm{erf}\left(\tfrac{13}{16}a\right)-\mathrm{erf}\left(\tfrac{11}{16}a\right)+\mathrm{erf}\left(\tfrac{5}{16}a\right)-\mathrm{erf}\left(\tfrac{3}{16}a\right) \tag{20d}$$

The general equation for s≥2 is,

$$p_{N-1-s}^{tc} = \begin{cases} \sum_{i=1}^{2^{s-2}}\left[\mathrm{erf}\left(2^{-s}\left(4i-\tfrac{3}{2}\right)a\right)-\mathrm{erf}\left(2^{-s}\left(4i-\tfrac{5}{2}\right)a\right)\right] & 2 \le s \le N-2 \\ \tfrac{1}{2}\sum_{i=1}^{2^{N-3}}\left[\mathrm{erf}\left(2^{-N}(8i-2)a\right)-\mathrm{erf}\left(2^{-N}(8i-6)a\right)\right] & s = N-1 \end{cases} \tag{21}$$

These probabilities are plotted in Figure 5.

Reliability

The algorithm works reliably so long as the logical tests provide a strong statistical bias for the MSB, which is always the case for nominal signal levels in radio astronomy. As an example, the probability distributions of scores for 8-bit, two's-complement data, with $\sigma=5v_0$, and M=255 words counted are shown in Figure 6. The bell curve for the MSB (k=7) is well to the right of all the others. In this case, the chances of failure, as calculated using Equation (7), are less than $7x10^{-7}$.

The reliability improves exponentially with the number of words counted, as shown in Figure 7 for a number of common-use cases. In all of these cases, the nominal signal level for optimum quantization noise is well below the saturation cross-over point.

The only case in radio astronomy which comes close to saturating the sampler is when two-bit sampling is used. The probability of bits scoring for N=2 is shown in Figure 8. This plot is exactly the same for all binary formats. The optimum level for quantization noise using two-bit sampling is approximately $\sigma=v_0$ (on the plot, $\sigma/2v_0 = 0.5$). The algorithm will work in this situation, but the statistical margin has been reduced relative to all the cases discussed so far, so a somewhat larger number of samples will have to be counted to achieve the same level of reliability.

Figure 8 shows that as the signal level increases beyond that point, the long tail of the Gaussian curve for the analog signal builds up the probability of occurrence of the outer two sample codes until they become even more likely to occur than the innermost codes. These codes are identical to the innermost codes except shifted by 1 bit. Under these conditions, the algorithm will fail by locking on the wrong bit.

In practice, if there is any fear of the sampler being in saturation when the word-boundary detection algorithm is running, an easy fix would be to simply bias down the front-end LNAs before doing so, dropping the gain of the system and pushing the analog signal level to the far left side of the plot. Once word-lock is established, the gain could be turned back on.

Validation

To validate the theory, as well demonstrate the effects of RFI and other non-Gaussian, non-white effects on the algorithm, actual data from laboratory receivers was analyzed. The data was scored according to the logical tests prescribed by the algorithm and the results averaged for each bit. Multiple signal levels were simulated from the same data set by clipping the waveform and truncating the bits. The results are shown in Figure 9 through Figure 12. Each data point represents the average over 100,000 samples.

Figure 9 is the scoring probability for the data without any CW components, as shown in the spectrum in the upper-right corner. The markers, representing real data, fall on top of the theoretical curves over most of the dynamic range of the plot. Only the k=0 bit deviates slightly at the right side of the plot where the waveform is beginning to clip. This is the bit for which the logical test ($b_k \neq b_{k-1}$) crosses over a word boundary. The theoretical curve for k=0 is based on the assumption that the noise is white, so there is no correlation between successive samples. The deviation we see in the measurement is probably due to the "color" of the spectrum, or in other words the gain slope at the high end of the band. With the higher-frequency components dropping off in amplitude, the correlation between adjacent samples is small but positive (so the chances of a 'mismatch' across the word boundary are less).

Figure 10 shows the scoring probability for a spectrum which contains a strong, high-frequency, CW tone. The strength of the CW tone roughly doubles the total integrated power in the spectrum. In this case, the k=0 bit has a higher probability of scoring at large signal levels than predicted, due to the small but negative correlation between adjacent samples. The remaining bits also drop off in scoring probability at the far right of the plot somewhat faster than expected, but overall the agreement between measurement and theory is quite good.

Figure 11 shows the scoring probability for a spectrum which contains a strong, low-frequency CW tone. Again, the strength of the CW tone is roughly equal to the noise power in the spectrum, and therefore doubles the total integrated power in the data stream. Here, the strong low-frequency component creates a positive correlation coefficient between successive samples, and the k=0 bit deviates toward lower-probability at high signal levels.

Finally, Figure 12 shows the theoretical and measured scoring probability for two's complement data streams taken from a number of different spectra, including some with low-, mid-, and high-band CW injected tones, and some using the L-Band front-end on the Green Bank Telescope for which high levels of broadband RFI are present. In all of these cases, however, the non-ideal components were too weak to cause a statistically significant deviation from the theoretical prediction.

Overall, despite some measurable effects due to very strong non-Gaussian components and non-white bandpass shape, the agreement between measurement and theory is excellent, especially over the dynamic range for which the algorithm operates.

Conclusions

In summary, an algorithm has been presented which can reliably detect the word boundaries in serialized, Gaussian-distributed, white-noise data using any of sign-magnitude, straight-binary, or two's-complement formats. The algorithm consists of scoring the bits in the data stream with a periodicity of N, where N is the word-length in bits, and then selecting as the most significant bit the one which receives the highest score after some large number of samples. The condition under which bit b_k receives a point depends on the binary format, and is as follows,

sign-magnitude:	$b_k=1$, $b_{k-1}=0$
straight binary:	$b_k \neq b_{k-1}$

two's complement: $b_k = b_{k-1} \neq b_{k+1}$

The algorithm has been tested using real-world data and was found to be robust in the presence of very strong non-idealities, such as large CW tones, gain slope, and RFI.

References

[1] M. Morgan and J. Fisher, *Next Generation Radio Astronomy Receiver Systems*, Astro2010 Technology Development White Paper, March 2009.

Figures

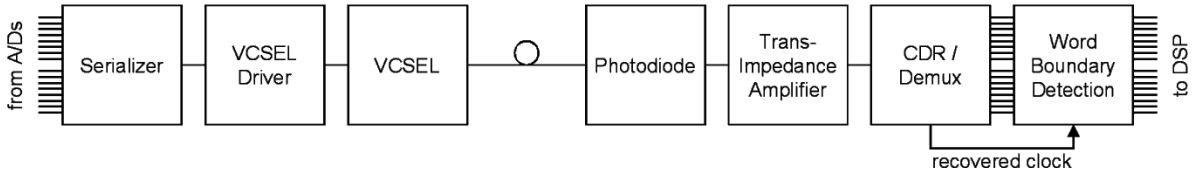

Figure 1. Simplified block diagram of a minimal-transmit-overhead photonic link for radio astronomy receivers.

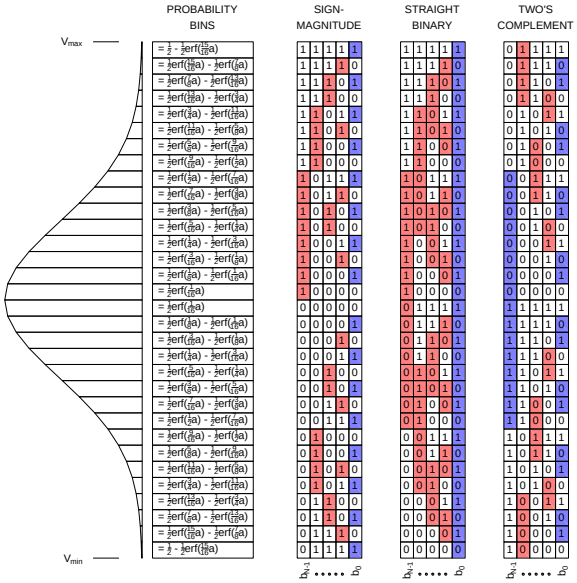

Figure 2. Diagram of the probability distribution for an analog signal and the corresponding 5-bit sampled output words using sign-magnitude, straight binary, and two's complement format.

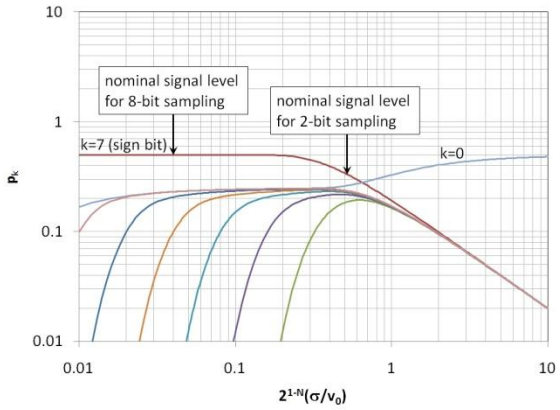

Figure 3. Plot of the scoring probability for sign-magnitude format. N=8 bits. The abscissa may be interpreted as the average voltage swing of the signal (2σ) divided by the full-scale range of the sampler ($2^N v_0$).

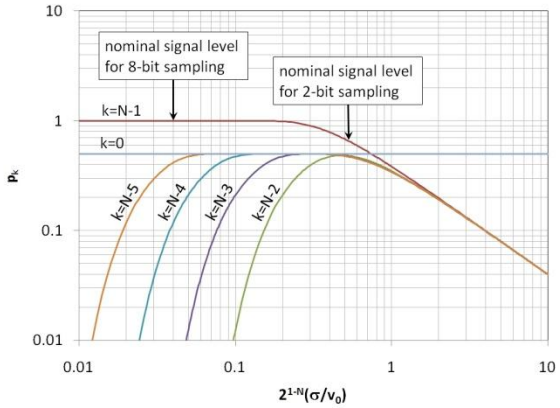

Figure 4. Plot of the scoring probability for straight binary format. N is arbitrary. The abscissa may be interpreted as the average voltage swing of the signal (2σ) divided by the full-scale range of the sampler ($2^N v_0$).

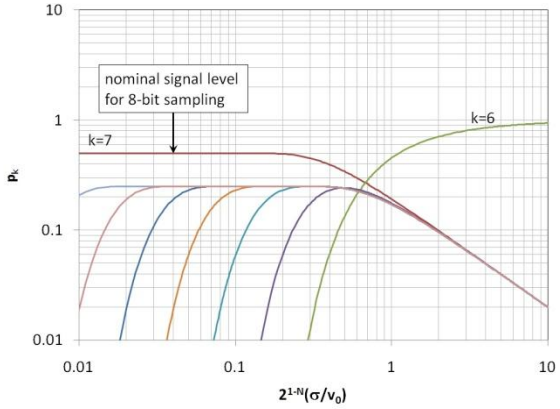

Figure 5. Plot of the scoring probability for two's complement format. N=8 bits. The abscissa may be interpreted as the average voltage swing of the signal (2σ) divided by the full-scale range of the sampler ($2^N v_0$).

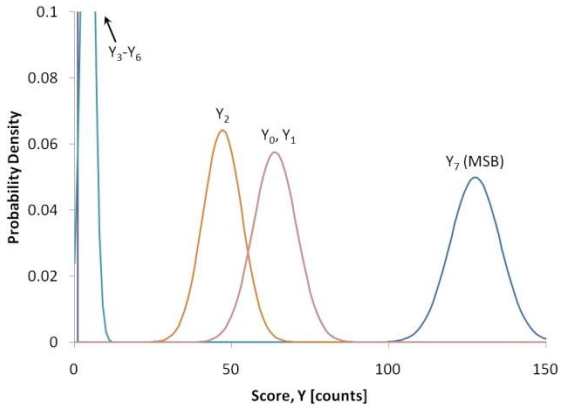

Figure 6. Probability distributions of scores for 8-bit, two's-complement data using the described algorithm with $\sigma=5v_0$ and M=255 words counted.

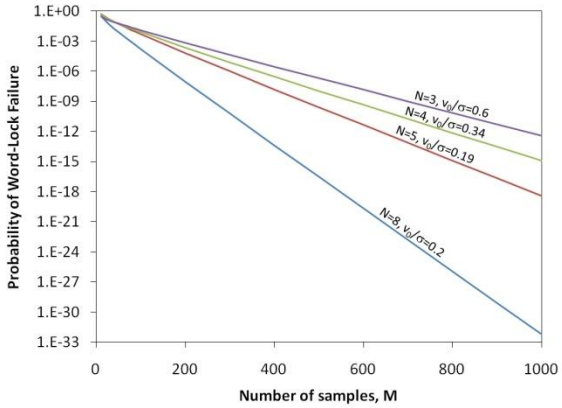

Figure 7. Probability of the algorithm failing (locking onto the incorrect bit) as a function of samples counted, M, for a number of common use cases in two's-complement format.

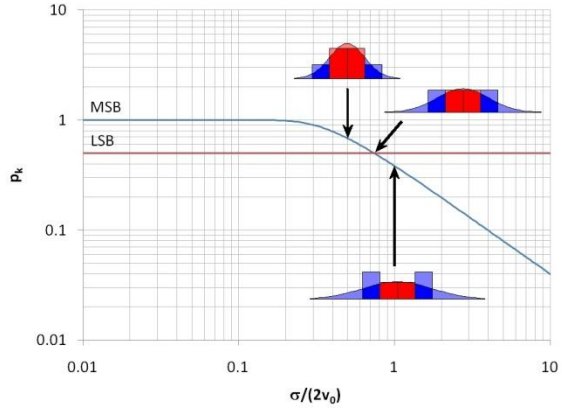

Figure 8. Plot of the scoring probability when N=2. The result is the same for all binary formats. The insets show the Gaussian distribution of the analog signal at different power levels and the corresponding probabilities in the four sampler bins.

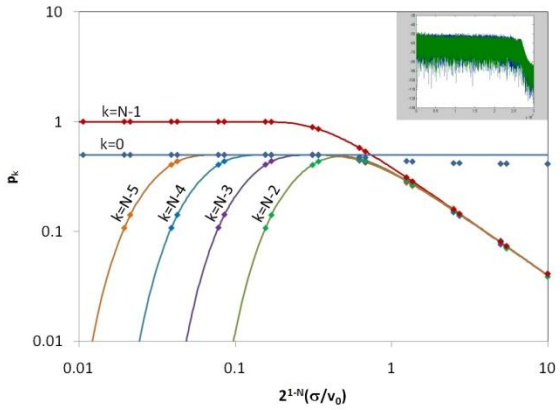

Figure 9. Theoretical (lines) and measured (markers) scoring probability for straight-binary data with no CW components. The spectrum of the data stream used for the measurement is shown in the upper-right corner. Multiple signal levels were simulated numerically by clipping the waveform and truncating the bits.

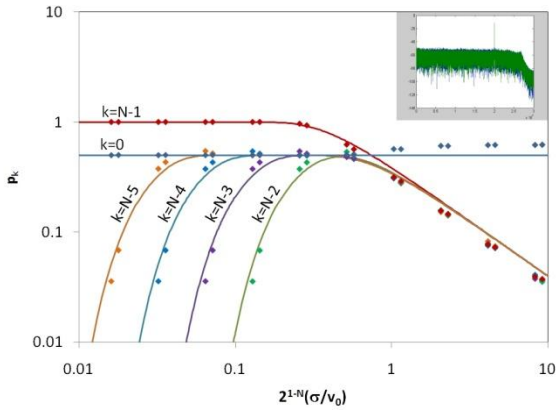

Figure 10. Theoretical (lines) and measured (markers) scoring probability for straight binary data with a high-band CW component. The spectrum of the data stream used for the measurement is shown in the upper-right corner. Multiple signal levels were simulated numerically by clipping the waveform and truncating the bits.

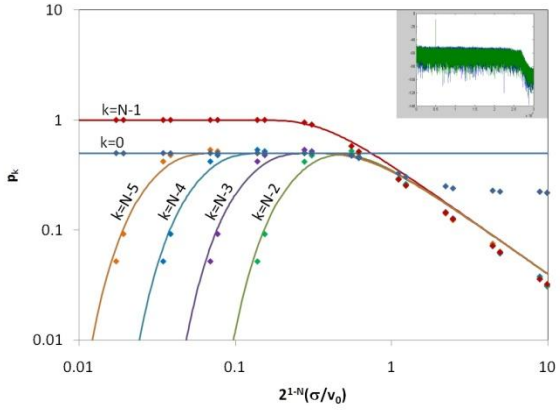

Figure 11.Theoretical (lines) and measured (markers) scoring probability for straight binary data with a low-band CW component. The spectrum of the data stream used for the measurement is shown in the upper-right corner. Multiple signal levels were simulated numerically by clipping the waveform and truncating the bits.

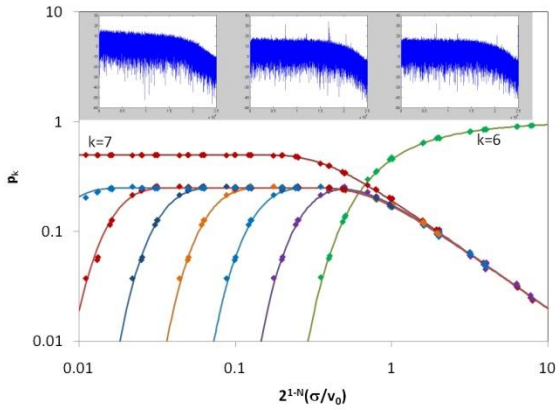

Figure 12. Theoretical (lines) and measured (markers) scoring probability for two's-complement data. Data points were taken from multiple spectra, including some with low-band, mid-band, and high-band CW tones injected, as well as from the L-Band front-end on the Green Bank Telescope in which significant levels of RFI were present. Some of the spectra are shown across the top of the plot. Multiple signal levels were simulated numerically by clipping the waveform and truncating the bits.