To/Mail Code: Albert Lazzarini/51-33
From/Mail Code: Bruce Allen/51-33
Phone/Fax 818-395-3677/818-304-9834
Refer to: LIGO-T970128-02-E
Date: September 10, 1997
Subject: LIGO ADC digitization/quantization accuracy

Quantization Noise in Ligo Interferometers

Bruce Allen
Department of Physics
University of Wisconsin - Milwaukee

Patrick Brady
California Institute of Technology
Pasadena CA 91125

The quantization of the Interferometer Differential Output (IFO) voltage by the Analog-to-Digital Converter (ADC) reduces the Signal-to-Noise Ratio (SNR) obtained by optimal filtering for an astrophysical signal. We calculate the fractional loss of SNR, and show that it is largely determined by the design of the whitening filter used at the IFO. We show that if three simple conditions on the whitened IFO output are satisfied, then the SNR loss is small. If the IFO voltage is perfectly white (in the signal band) then a 6 bit ADC (1) gives SNR loss of less than 0.4% and (2) has enough headroom (dynamic range) so that the IFO output can be up to eight times the root-mean-square voltage $V_{\rm rms}$ before clipping occurs. Our analysis applies to arbitrary astrophysical searches. As a concrete example, we analyze the SNR loss for a binary inspiral search in the the November 1994 configuration of the 40-meter prototype. We also give bounds on the timing jitter of the sample-and-hold clock of the ADC, which ensure that the loss of SNR will be small.

I. INTRODUCTION

In converting analog signals to digital form one must consider effects which arise due to sampling the signal. The Nyquist sampling theorem proves that a bandwidth limited signal can be reconstructed from samples taken at twice the maximum frequency of interest. An additional issue arises because the samples are stored as finite precision numbers. While the requirement that one may reconstruct the continuous input signal drives the choice of sampling rate, potential loss in signal to noise for detection purposes is an important factor in determining the accuracy (number of bits) with which to record the sampled signal.

For the gravitational wave detectors which are currently being built, the following two issues arise:

1. The whitening filters must work well enough to reduce the dynamic range of the IFO voltage so that the there is very small probability that it exceeds the maximum input level of the ADC. In other words, the signal should not clip, or should clip very infrequently.

2. The number of bits should be chosen so that the expected signal to noise from a gravitational wave signal is not significantly reduced by the quantization process.

In section III we present a simple theoretical framework in which to address these issues. Our conclusions may be summarized by the following three conditions, which may be regarded as requirements for the design of the IFO whitening filters:

- **Loss of SNR**
  To ensure that the fractional loss ℓ of SNR is small (for example $\ell < 0.01$) the quantization step size Δ in volts must be less than

$$\Delta^2 < 24 \ell f_N \min_{f \in f_{\rm sig}} |S_v(f)| \, . \tag{1.1}$$

  Here f_N is the Nyquist frequency (half the sample rate) and $S_v(f)$ is the voltage power spectrum (volts/Hz^2) of the whitened IFO. The value of f is that value which minimizes the right-hand-side in the astrophysical signal band, typically 100 Hz – 2000 Hz.

- **Dynamic Range**
  To ensure sufficient headroom (dynamic range) in the ADC process (i.e. to prevent clipping) there must be enough bits b so that

$$(2^{b-1}\Delta)^2 \geq N^2 \int_0^{f_N} S_v(f) df \ . \tag{1.2}$$

  Here N is the safety factor: the ratio of ADC input clipping voltage to the ADC rms input voltage. One would typically like to have $N > 32$ to enable careful inspection of transient glitches in the IFO voltage, i.e. for diagnostic purposes.

- **Dithering**
  The final condition that must be satisfied by the whitening filter is that there is enough power at high frequencies to adequately "dither" the ADC. This requires that

$$\int_0^{f_N} df \, [1 - \cos(2\pi f \tau_{min})] \, S_v(f) \gg \Delta^2/2 \ . \tag{1.3}$$

  Here $\tau_{\rm min}$ is (less than or equal to) the period of the highest frequency waves of astrophysical interest. It may safely be taken to be the sample time. In this case, the integral above gives no weight to the IFO voltage output spectrum at DC, and maximum weight to the spectrum at the Nyquist frequency.

Provided that the IFO whitening filters are designed to satisfy these three conditions, the loss of SNR from the digitization/quantization process will be small.

Note that there is an additional restriction, arising from the accuracy of the clock that drives the ADC sample-and-hold circuit. Timing jitter in this clock gives rise to an effective noise source. Detailed analysis shows that the effects of this timing jitter decrease the SNR by less than 1% if the largest timing jitter is less than about 1/22nd the width of the clock signal. For a 16kHz sample rate, this corresponds to less than 3μsec of timing jiter.

II. THE 40-METER PROTOTYPE

It is illustrative to examine these three conditions, which we will derive in the following sections, for the November 1994 configuration of the Caltech 40-meter prototype interferometer. For this system, the sample rate was approximately 9868 Hz, corresponding to a Nyquist frequency $f_N = 4934$ Hz. Fig. 1 shows a graph of the power spectrum $S_v(f)$ of the IFO during a period of quiet operation, as well as a line showing the spectrum of quantization noise, with amplitude $S_q(f) = \Delta^2/12 f_N = 1.69 \times 10^{-5} \Delta^2/\text{Hz}$.

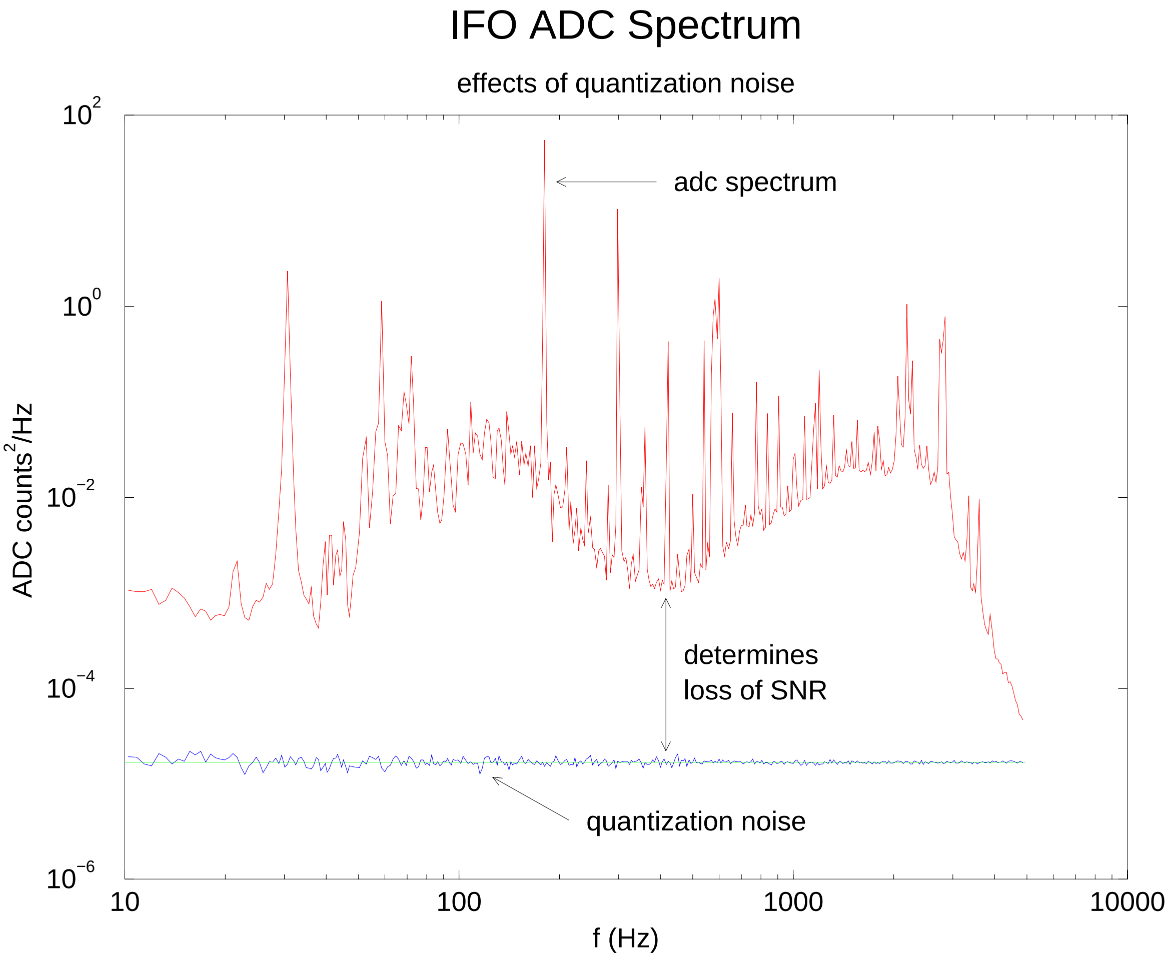

FIG. 1. The IFO power spectrum in (ADC counts)^2/Hz is shown in red. This may be compared to the expected spectrum of quantization error $\Delta^2/12f_N$ shown in green. The expected level of quantization noise was also simulated by replacing the IFO output with uniformly distributed random numbers on the interval $(-\Delta/2, \Delta/2)$, shown in blue. The ratio of the powers at the indicated point determines our upper bound on the fractional loss in signal to noise ratio.

During this (typical) period of interferometer operation, the RMS output value was about 23Δ (i.e., ± 23 ADC output counts). The ADC itself had $b = 12$ bits and an output range from -2048Δ to $+2047\Delta$. Examining each of the above three conditions in turn, we find:

- **Loss of SNR**
  An upper bound on the fractional loss ℓ in signal to noise ratio (SNR) for a gravitational wave signal is provided by

  $$\ell \le \max_{f_{\rm sig}} \left| \frac{S_q(f)}{2S_v(f)} \right| = \max_{f_{\rm sig}} \left| \frac{\Delta^2}{24 f_N S_v(f)} \right| \ , \tag{2.1}$$

  In Fig. 1, the minimum value of the ratio ℓ in the signal band from 120 Hz to 2000 Hz is 9×10^{-3}. Hence, for this particular stretch of data, no more than 0.9% of the SNR is lost due to quantization error.

- **Dynamic Range**
  The safety factor N is simply the ratio of the rms output voltage to the peak (clipping level) input of the ADC, which is $\pm 2^{b-1}\Delta$. For this 40-meter data, one finds that:

  $$N = \frac{2^{b-1}\Delta}{V_{\rm rms}} = \frac{2^{b-1}\Delta}{\left[\int_0^{f_N} df\ S_v(f)\right]^{1/2}} = 89. \tag{2.2}$$

Thus, the IFO can exceed 89 times its rms value without overloading or clipping. In practice, clipping is *very* infrequent.

- **Dithering**
As we will show later, the dithering condition is set by requiring that the output of the ADC changes by more than a single count over the timescale of interest. It is easy to show that for the spectrum we have shown, that

$$\int_0^{f_N} df\,[1-\cos(\pi f/f_N)]\,S_v(f) \approx 100\Delta^2/2\;, \tag{2.3}$$

so this condition is easily satisfied, even when the timescale of interest is the sample time, $\tau_{\min} = \Delta t = 1/2f_N$.

Later, we will return to the November 1994 configuration of the 40-meter prototype, and show some additional details concerning SNR loss from digitization error.

III. QUANTIZATION PROCESS

In this section we derive some the results which have just been outlined. In particular, we consider the properties of the IFO voltage and the effects of the analog-to-digital conversion on the recorded signal. During normal operation the partially whitened* IFO voltage $v(t)$ will be a stationary random process. This voltage is passed through an ADC which determines an output voltage $\overline{v} = Q(v)$ for the given input voltage v. The function $Q(v)$ is represented in Fig. 2 and maps the real numbers into signed integer multiples of Δ by rounding; Δ has units of volts. The number of output levels available is 2^b where b is the number of bits used by the ADC, thus the dynamic range is $[-2^{b-1}\Delta, (2^{b-1}-1)\Delta]$. The error introduced by the quantization process is given by

$$W(v) = Q(v) - v \tag{3.1}$$

and lies in the range $(-\Delta/2, \Delta/2]$ as shown in Fig 2.

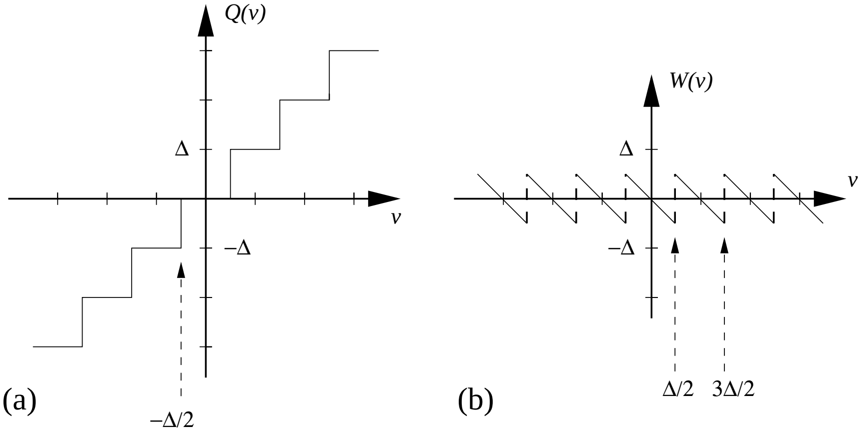

FIG. 2. (a) The quantization function $Q(v)$ which maps the analog signal into integer multiples of Δ, the quantization level, by rounding. (b) The error function which is defined as $W(v) = Q(v) - v$.

*The raw voltage is electronically filtered to reduce the dynamic range. This process removes the dominant second-order correlations from IFO, however the filtered voltage is not completely white.

Since we have no knowledge of the precise input voltage which produced a given ADC output, it is tempting to think of the quantization process as the addition of noise to the IFO voltage in such a way that

$$\overline{v}(t) = Q[v(t)] = v(t) + W[v(t)] = v(t) + n_q(t) \ , \tag{3.2}$$

where $n_q(t) = W[v(t)]$. To make progress we must make several assumptions about this noise process which will be justified *a posteriori* in the case of the 40m prototype interferometer:

1. $n_q(t)$ is a stationary, white random process.
2. $n_q(t)$ is uncorrelated with the voltage $v(t)$.
3. $n_q(t)$ is uniformly distributed from $(-\Delta/2, \Delta/2]$.

Now, if the sampling rate is $2f_N$, where f_N is the Nyquist frequency measured in Hz, the one-sided power spectral density of the noise $n_q(t)$ can be determined from

$$\int_0^{f_N} S_q(f) df = \langle n_q(t) n_q(t) \rangle \ , \tag{3.3}$$

where $\langle \ldots \rangle$ indicates the ensemble average. Assumption 1 implies that $S_q(f)$ is constant, while the right-hand side is just the variance of the uniform random variable associated with $n_q(t)$. This is

$$\langle W^2(v) \rangle = \frac{1}{\Delta} \int_{-\Delta/2}^{\Delta/2} v^2 dv = \frac{1}{\Delta} \left[\frac{v^3}{3} \right]_{-\Delta/2}^{\Delta/2} = \frac{\Delta^2}{12}. \tag{3.4}$$

Hence, the one-sided power spectral density associated with the quantization noise is

$$S_q(f) = \frac{\Delta^2}{12 f_N} \ . \tag{3.5}$$

Note that the dimension of the power spectrum is Volts^2/Hz.

One can also derive a simple relation which the IFO must satisfy in order that the above assumptions should be valid. Since we are interested in detecting bandwidth limited signals using the interferometer, there is a shortest timescale of interest. This timescale is related to the maximum frequency $f_{\max}$ of the expected astrophysical signals by $\tau_{\min} = 1/(2f_{\max})$. If, on average, the IFO voltage changes by more than one quantization level on timescales greater than or equal to $\tau_{\min}$, then assumptions 1 and 2 are justified. The mathematical expression of this condition is

$$\langle (v(t) - v(t + \tau_{\min})^2 \rangle \gg \Delta^2 \ . \tag{3.6}$$

By assumption $v(t)$ is a stationary random process, therefore we can expand the left hand side and express it in terms of a simple integral involving the voltage power spectrum. The final condition is that

$$\int_0^{f_N} df \, [1 - \cos(2\pi f \tau_{min})] \, S_v(f) \gg \Delta^2/2 \ . \tag{3.7}$$

Notice that if the (whitened) IFO voltage was perfectly white, i.e. $S_v(f) = S_0$ a constant, this is equivalent to saying that the RMS voltage $V_{\rm rms}^2 = f_N S_0$ should be much greater than the square of the quantization level Δ^2.

Provided that the dithering condition Eq. (3.7) is satisfied, then the assumptions above are valid. In this case, it follows from Eq. (3.2) that the power spectra of the signal and the quantization noise add in quadrature, so that

$$S_{\overline{v}}(f) = S_v(f) + S_q(f) = S_v(f) + \frac{\Delta^2}{12 f_N}. \tag{3.8}$$

This relation makes it easy to analyze the effects the ADC quantization process; the digitization of the signal may be modeled as an additional independent noise source in the IFO.

An additional source of effective noise arises from errors in the timing signals that are used to trigger or clock the ADC. This noise source can also be described as an additional (in quadrature) error in recording the signal value. Denote the signal value by the Fourier transform:

$$v(t) = \int df\, \tilde{v}(f) \exp(2\pi i f t). \tag{3.9}$$

Errors in the clock that triggers the ADC have the effect of sampling this signal at the wrong times. Let us suppose that the timing error is denoted by $\Delta\tau$. In this case the error in the signal value is

$$\Delta v = v'(t)\Delta\tau = \int df (2\pi i f \Delta\tau)\tilde{v}(f) \exp(2\pi i f t). \tag{3.10}$$

We assume that the timing errors are not correlated with the signal, and that they are described by a characteristic value $\Delta\tau$. The spectrum of timing error noise S_{timing} is then easily related to the signal spectrum $S_v(f)$ by:

$$S_{\text{timing}}(f) = (2\pi f \Delta\tau)^2 S_v(f). \tag{3.11}$$

With these assumptions, we can write

$$S_{\overline{v}}(f) = \left[1 + 4\pi^2 (f\Delta\tau)^2\right] S_v(f) + S_q(f) = S_v(f)\left[1 + 4\pi^2 (f\Delta\tau)^2\right] + \frac{\Delta^2}{12 f_N}. \tag{3.12}$$

Thus the effects of signal quantization and timing noise jitter can both be described as additional sources of noise.

IV. LOSS IN SIGNAL TO NOISE

In the previous section, we characterized the effects of ADC quantization as an additional source of IFO noise. If the dithering condition is satisfied, this noise appears as an independent white noise source, which adds in quadrature to the other noise sources in the IFO. We will now estimate the loss in expected signal to noise due to quantization noise which occurs when using matched filtering to detect a gravitational wave signal. The strain at the detector is determined by

$$\tilde{h}(f) = R(f)\tilde{V}(f) \tag{4.1}$$

where $R(f)$ is the complex transfer function, having units of strain per Volt, and a tilde indicates the Fourier transform. Denote the Fourier transform of the postulated signal by $\tilde{h}_T(f)$, and the one-sided power spectrum of the IFO strain h by $S_h(f) = |R(f)|^2 S_v(f)$. The expected signal to noise ratio from optimal filtering is then given by

$$\left(\frac{S}{N}\right)^2 = 4 \int_0^{f_N} \frac{|\tilde{h}_T(f)|^2}{S_h(f)} df\ . \tag{4.2}$$

We wish to compare the signal to noise that would be obtained in the absence of quantization effects

$$\left(\frac{S}{N}\right)^2_{\text{opt}} = 4 \int_0^{f_N} \frac{|\tilde{h}_T(f)|^2}{|R(f)|^2 S_v(f)} df \tag{4.3}$$

to that obtained using the quantized ADC output. Provided that the dithering condition Eq. (3.7) is satisfied, the noise due to quantization will be independent of the IFO noise, so the power spectra of the two processes add in quadrature. Thus the SNR obtained with the quantized ADC output is

$$\begin{aligned}\left(\frac{S}{N}\right)^2_{\text{quant}} &= 4 \int_0^{f_N} \frac{|\tilde{h}_T(f)|^2}{|R(f)|^2 S_{\overline{v}}(f)} df \\ &= 4 \int_0^{f_N} \frac{|\tilde{h}_T(f)|^2}{|R(f)|^2 [S_v(f) + S_q(f)]} df \\ &\approx 4 \int_0^{f_N} \frac{|\tilde{h}_T(f)|^2}{|R(f)|^2 S_v(f)} \left[1 - \frac{S_q(f)}{S_v(f)}\right] df\ ,\end{aligned} \tag{4.4}$$

where $S_q(f)$ is given by Eq. (3.5). Note: we have assumed that there is no timing clock jitter, so $\Delta\tau = 0$ (we will return to this point later.) Now the fractional loss in signal to noise ratio ℓ is given by

$$\frac{(S/N)_{\text{quant}}}{(S/N)_{\text{opt}}} \approx 1 - \ell\ , \tag{4.5}$$

where

$$2\ell = \frac{\int_0^{f_N} \frac{|\tilde{h}_T(f)|^2}{|R(f)|^2 S_v(f)} \frac{S_q(f)}{S_v(f)} df}{\int_0^{f_N} \frac{|\tilde{h}_T(f)|^2}{|R(f)|^2 S_v(f)} df} \leq \max_{f_{\rm sig}} \left| \frac{S_q(f)}{S_v(f)} \right| . \tag{4.6}$$

The notation $\max_{f_{\rm sig}}$ indicates the maximum over the bandwidth of the signal. Using Eq. (3.5) we arrive at the compact expression

$$\ell \leq \max_{f_{\rm sig}} \left| \frac{\Delta^2}{24 f_N S_v(f)} \right| \tag{4.7}$$

This result shows that the fractional loss is signal to noise ratio is bounded above by the largest relative magnitude of the quantization noise spectrum to the partially whitened IFO voltage spectrum in the signal band. In particular, for narrow band signals ℓ *is* the fractional loss in signal to noise. It makes intuitive sense that the most significant effect should occur where the whitened noise floor most closely approaches the quantization noise floor.

The goal is to ascertain the operating characteristics for the whitening filters and the ADC in order to ensure that the fractional loss in signal-to-noise is below some pre-determined value $\ell_{\rm max}$. Let us first consider the ADC. Eq. (4.7) determines that the quantization levels in the ADC satisfy

$$\Delta^2 < 24 \ell_{\rm max} f_N \min_{f \in f_{\rm sig}} |S_v(f)| \; . \tag{4.8}$$

Now, it remains to establish how many bits are needed in order to achieve this value. To avoid overloading the ADC and clipping the IFO, the total dynamic range should be some factor N times the RMS voltage of the IFO, thus we must have enough bits so that

$$(2^{b-1}\Delta)^2 \geq N^2 V_{\rm rms}^2 = N^2 \int_0^{f_N} S_v(f) df \; . \tag{4.9}$$

We will refer to N as the *safety factor.* Combining Eqs. (4.8) and (4.9) gives

$$2^{2(b-1)} \geq \max_{f_{\rm sig}} \left| \frac{N^2 V_{\rm rms}^2}{24 \ell_{\rm max} f_N S_v(f)} \right| \; . \tag{4.10}$$

It is useful to consider a simple example at this point. Suppose that the (whitened) IFO voltage was perfectly white in the signal band, so that $V_{\rm rms}^2 = f_N S_v(f)$, and that we require a maximum fractional loss in SNR $\ell_{\rm max} = 2^{-8}$, then with a safety factor $N = 8$ one should use at least $b = 6$ bits in the ADC. This shows that remarkably little loss in SNR occurs due to quantization noise provided the assumptions made in section III hold (see below for justification).

Returning now to the effects of error in the sample and hold clock for the ADC, similar analysis shows that an upper bound on the fractional loss of SNR due to the timing jitter is then given by

$$\ell \leq \max_{f_{\rm sig}} \left| \frac{S_{\rm timing}(f)}{2 S_v(f)} \right| = 2\pi^2 (f_{\rm max-sig} \Delta\tau)^2. \tag{4.11}$$

For the most conservative possible design, we allow this max frequency to be the Nyquist frequency $f_{\rm max-sig} = f_N = f_{\rm sample}/2 = 1/2\Delta t$, i.e. half of the sample rate. In order that the SNR loss due to timing errors be less than 1%, we need to ensure that

$$\frac{1}{100} < \frac{\pi^2}{2} \left(\frac{\Delta\tau}{\Delta t} \right)^2 \Rightarrow \delta\tau \lesssim \Delta t/22, \tag{4.12}$$

so the errors in clock timing should be less than 1/22nd the actual sample time interval. For a 16 kHz sample rate, this would correspond to no more than a 3 μsec timing error in the ADC sample and hold clock circuit.

V. THE 40M PROTOTYPE (CONTINUED)

The most subtle result here concerning the dithering condition Eq. (3.7). It is interesting to compare the results of the theoretical model introduced above to those obtained by directly analyzing the data from the 40m interferometer. The ADC used during the data collection run in 1994 had 12 bits.

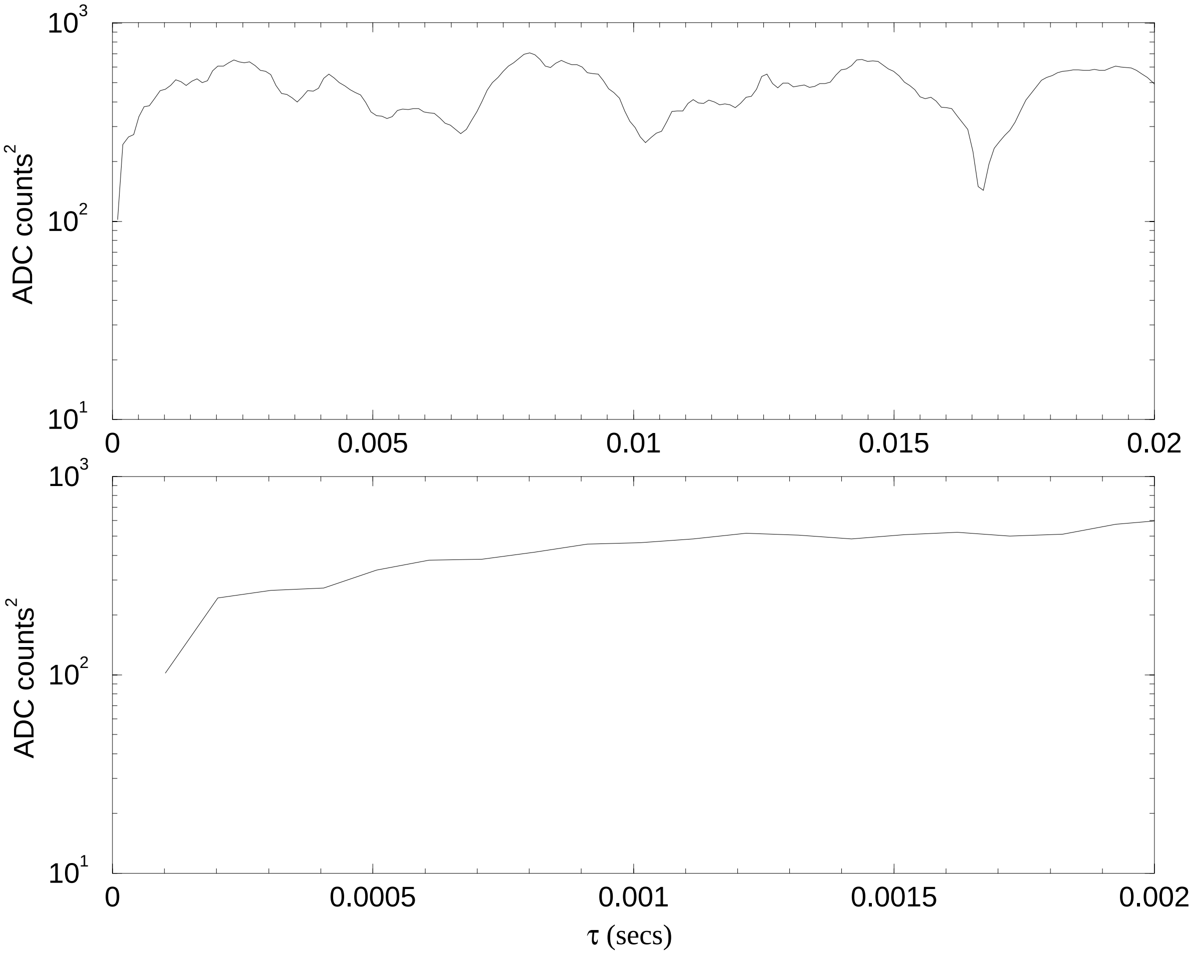

FIG. 3. The function $f(\tau) = \left\langle (v(t) - v(t+\tau)^2 \right\rangle$ in units of (ADC counts)^2 for a quiet stretch of data from the prototype 40m interferometer. The lower graph is zoomed in by a factor of 10 on the upper graph. It is clear that the condition in Eq. (3.6), with $\Delta = 1.0$ is well satisfied even at for $\tau = (2f_N)^{-1} \approx 10^{-4}$ sec .

We have verified that the condition given in Eq. (3.6) [or equivalently in Eq. (3.7)] is indeed satisfied for the data taken during the 1994 observation run with the 40m prototype; a graphical representation of this result is given in Figure 3.

The bound on fractional loss in signal to noise ℓ and the safety factor N are presented in Table I. Moreover, the actual loss in signal to noise for compact coalescing binary chirp signals is also presented for comparison. The results demonstrate how well one does with only 12 bits of ADC provided the whitening filters are reasonably good.

Data file	Safety factor N	$\ell_{\max}$	ℓ (120Hz $\leq f_{\rm sig} \leq$ 2000Hz)
19nov94.1	96.9 ± 2.4	7.2e-02	1.1e-02
19nov94.3	75.0 ± 17.4	2.1e-02	4.4e-03

TABLE I. The safety factor and fractional loss in signal to noise for the 40m prototype interferometer. based on the assumptions in section III Each quantity is computed from the first locked segment of data in the listed files. The error estimate for N corresponds to $(< N^2 > - < N >^2)^{1/2}$ over the entire locked segment. The losses in signal to noise are typical values.

VI. SIMULATIONS

Finally, the validity of the above assumptions has been checked in the following simple numerical experiment. Correlated (colored) Gaussian noise was generated using double precision accuracy. This noise was then quantized using the quantization function shown in Fig. 2, and the error $W(t)$ was recorded. The result of binning up 131072 samples is shown in Fig. 4. A χ^2-test on the resulting distribution determined it to be uniform with probability 0.92. Moreover, the quantization noise is white, and $\langle W(t)W(t+\tau)\rangle$ is approximately stationary.

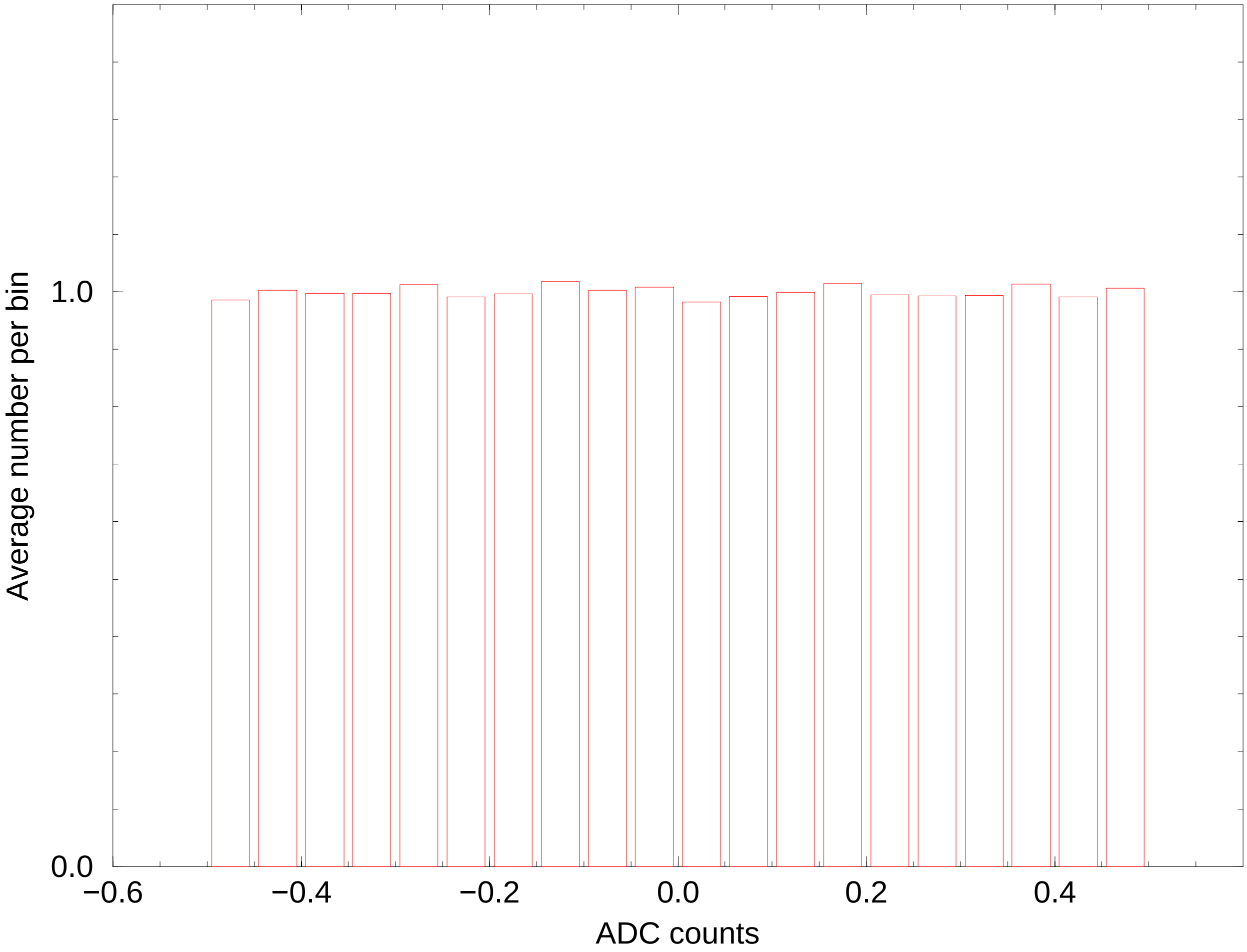

FIG. 4. Simulated Gaussian noise, with the power spectral density of the 40m prototype was quantized as describe above with $\Delta = 1$. The quantization noise from 131072 samples was binned into 20 bins in the range $(-0.5, 0.5]$, and averaged (divided by 131072/20). The results shows that this noise is indeed uniformly distributed throughout the interval.

VII. CONCLUSION

The theoretical framework presented above is quite general. Provided the dithering condition in Eq. (1.3) is satisfied, and the dynamic range condition Eq. (1.2) is satisfied with a reasonable safety factor, then Eq. (1.1) gives a bound on the loss of Signal-to-Noise ratio due to the digitization process. These three conditions should be satisfied by the design of the whitening filter. Provided that this filter is well-designed, and the IFO has a fairly white spectrum, even a small number of bits (say ten) is sufficient to ensure very small loss of SNR. It is the quality of this whitening filter that ultimately determines the loss in signal to noise. Thus, the whitening filters should be designed to keep the broadband noise as flat as possible.

For chirp signals from compact coalescing binaries we have shown that the actual loss in signal to noise is generally an order of magnitude smaller for data taken with the 40m prototype interferometer than the bound given in Eq. (1.1).

For other *broadband* signals the upper bound still applies; however the loss in signal to noise will be smaller than this bound in general.

For narrow-band sources the fractional loss in signal to noise due to quantization noise is equal to the bound in Eq. (4.7). Computing ℓ for data taken with the 40m prototype in 1994 indicates that there are frequencies (between 120 Hz and 2000 Hz) where $\ell = 0.07$. Given the difficulty in detecting weak periodic signals, it is desirable to reduce this number as much as possible in the final design of the interferometric detector