Quantization Considerations for Distortion-Controlled Data Compression

M. Klimesh[1]

Distortion-controlled data compression is lossy signal compression in which the amount of distortion introduced into any small portion of a signal is strictly limited. This article gives results on various practical aspects of quantization and distortion-controlled compression. The primary focus is near-lossless compression, as accomplished by predictive techniques. For compression of noisy signals, the effect of quantization step size q on rate, distortion, and the effective noise characteristics is considered. It is demonstrated that surprisingly large values of q often may be employed without significantly affecting the scientific analysis of the compressed data. For low-noise signals, the use of subtractive dither to reduce or eliminate compression artifacts is analyzed. This analysis includes determination of a class of optimal dither signals based on certain reasonable assumptions. The effect on compression of the step size used in analog-to-digital conversion also is discussed.

I. Introduction

Most of the data gathered by spacecraft consists of digitized samples of signals. These signals may have multiple dimensions; for example, gray-scale image data are obtained from a two-dimensional signal. In the absence of other considerations, scientists would prefer to retrieve the data unaltered. Usually, however, if only lossless compression is used, then a spacecraft is capable of gathering much more data than can be transmitted, due to downlink rate constraints. If lossy compression is used, the amount of data that can be sent is increased, but one must consider the effect of the added distortion on the value of the data. We address this concern by considering compression in which the amount of distortion that may be introduced into any local region of the data is guaranteed to be less than a specified limit. We refer to this type of compression as distortion-controlled compression.

Distortion-controlled compression protects small details of a signal to a well-defined (and selectable) degree, so one can have complete assurance that the scientific value of the signal is adequately preserved. On the other hand, typical lossy compression techniques are well suited for minimizing a quantitative distortion measure such as root-mean-squared error (RMSE) or an equivalent metric such as mean-squared error (MSE) or peak-signal-to-noise ratio (PSNR). Unfortunately, one cannot infer protection of individual details of a signal from an RMSE value, and in practice algorithms have a tendency to cause more distortion in areas of interesting details than in other areas.

It should be noted that, in general, the scientific usefulness of a compressed signal is *not* well measured by a maximum local distortion metric. For a wide variety of compression methods and distortion levels,

[1] Communications Systems and Research Section.

the RMSE provides a reasonable (though far from perfect) metric for comparing the value of reconstructed versions of a signal. The same is not true for metrics that measure local distortion. For example, suppose two identical copies of a signal are compressed, one by a predictive compression algorithm that limits the total distortion, the other by a transform-based algorithm. If the RMSE values of both reconstructed signals are similar, then the overall subjective appraisals will likely also be somewhat similar, although the two reconstructed signals may have very different good and bad aspects. However, the reconstructed signal from the predictive algorithm generally will have a somewhat lower maximum local distortion, and, unless operating at high fidelities, the transform algorithm typically will yield a lower rate.[2]

At small RMSE values, however, distortion-controlled compression methods tend to provide lower rates than do more general methods. Thus, when operating at high fidelities, distortion-controlled compression methods give a maximum local-distortion guarantee essentially for free.

Ideally, it would be possible to use distortion-controlled compression that is competitive in RMSE performance over a wide range of distortion levels. Rate-distortion theory suggests that, to accomplish this at moderate and higher distortion levels with a minimal penalty in RMSE performance, it is important that the local constraint be fairly loose. That is, a conventionally compressed signal would not violate the local constraint very often. Thus, it appears that distortion-controlled compression is inherently less useful (but by no means useless) at moderate and higher distortion levels.

In fact, we do not know of any compression techniques that produce competitive rate-distortion performance in terms of RMSE at high distortion levels along with a reasonable maximum local distortion. If one attempts to artificially construct such a technique by taking a good RMSE method and adding an extra step to reduce the large local residuals, it is found to be difficult to reduce the local distortion significantly without increasing the rate substantially.

Thus, although the definition of distortion-controlled compression encompasses a wide range of degrees of lossiness, we confine ourselves to the cases when the maximum local distortion is restricted to be quite low, resulting in high-fidelity compression.

In this article, a "local region" of the data always will be a single sample. That is, we will consider compression in which the absolute difference between the value of a sample in the original data and the corresponding value in the reconstructed data is less than a specified value. Compression with this constraint has been called "L_∞-constrained compression" [17] and "near-lossless compression."

The primary purpose of this article is to exhibit basic results related to near-lossless compression, thereby laying the groundwork for the design of distortion-controlled compression algorithms. This article also will provide support for compression recommendations; this includes serving as a convenient reference on the trade-offs available. Throughout this article, we give categorizations of types of signal data and discuss the effect of these categorizations on compression.

It also is hoped that this article will prove useful to users of signal compression. They can determine from our results if distortion-controlled compression is likely to be appropriate for their needs.[3] This article also will help increase the understanding of the trade-offs associated with compression, instrument design, quantity and fidelity of signal data returned, and downlink rate.

The remainder of this article is organized as follows. Section II discusses the connections between quantization and analog-to-digital conversion. Section III gives a basic description of predictive compression and gives some quantitative results on the distortions introduced and on the amount of compression

[2] The amount of compression often is given as a rate, in bits/sample. A lower rate indicates more compression. The rate-distortion performance refers to the trade-off between rate and average distortion.

[3] Of course, there is no substitute for consulting compression specialists, but such interaction can be facilitated by an awareness of the issues.

obtained. Section IV provides basic calculations of the distortion from quantization, and Section V compares the amount of compression of near-lossless compression with that of lossless compression. Section VI contains a detailed analysis of near-lossless compression of noisy signals. Section VII contains a discussion of the use of subtractive dither to reduce or eliminate quantization artifacts, including characterization of optimal dither distributions. Section VIII discusses the effect of the fineness of the original analog-to-digital conversion of signals on rate and distortion in near-lossless compression. Finally, in Section IX, we conclude with a summary of the major points and a brief discussion of the possibility of onboard analysis.

II. Quantization and Analog-to-Digital Conversion

An important purpose of this article is to analyze several aspects of quantization, including the effect of quantization step size on rate, distortion, and specific artifacts in the reconstructed samples. Thus, we begin with a general discussion of quantization.

At one or more points in the process of collecting, processing, compressing, and transmitting sampled signal data, quantization must occur. Analog-to-digital conversion is one form of quantization (see Section 1.1 of [3]). Near-lossless compression methods generally include a quantization step that is applied to samples already in digital form. Thus, it is common for a fine quantization step (such as analog-to-digital conversion) to be followed by at least one coarser quantization step.

Transform-based compression methods generally involve a quantization step applied to functions of several samples. A sample is reconstructed from several of these quantized values. Although some of the principles that we discuss apply to this situation, we do not tailor our analysis to it since transform coding is not well-suited to near-lossless compression. Although near-lossless predictive compression (discussed in the next section) is in some ways similar to transform coding, our analysis does apply to near-lossless predictive compression because a sample is reconstructed primarily based on a single quantization.

When a sample is quantized by multiple, progressively coarser quantization steps, the final quantization step is the primary factor in determining the quantization noise statistics and the bit rate needed to encode the quantized data. However, the fineness of an earlier quantization step may be a contributing factor: if it is finer, then overall compression performance may be better. We discuss this effect in Section VIII.

We primarily consider uniform quantization—that is, quantization in which the bins all have equal size. Uniform quantization of noise is known to be good in that, if the quantized values are entropy coded,[4] the resulting rate-distortion performance is close to optimal (but usually not exactly optimal) under several sets of assumptions (for example, Gaussian noise and mean-squared error [16]). We may infer that uniform quantization is also often close to optimal for signal-sample quantization, since an interesting signal generally may be modeled as a random process not unlike noise.

Nonuniform quantization may be appropriate if the noise level varies with the signal level or if the importance of accuracy varies with the signal level. A common reason for using nonuniform compression is to obtain a large dynamic range without requiring an enormous number of quantization levels. Companding (see Section 5.5 of [3]) is one method of achieving this. More generally, any nonuniform quantization can be obtained by applying a reversible nonlinear transform to each sample and then applying uniform quantization. If the RMSE of the transformed data is a useful measure of its value, then the analysis in this article may apply to quantization of the transformed data.

We consider situations in which analog-to-digital conversion or more general quantization is applied to pure noise, to signals with little or no noise, or to noisy signals in which the underlying signal varies slowly compared with the sample-to-sample variations caused by noise.

[4] The entropy (see, for example, [2]) of a distribution is the theoretical lowest rate, in bits/sample, at which random variables with this distribution may be losslessly encoded. Entropy coding (see, for example, [2,3,8]) is the process of encoding the values near this rate. Useful techniques for entropy coding include the well-known techniques of Huffman coding, arithmetic coding [5,15], run-length coding, and Golomb–Rice coding [4,9].

III. Predictive Compression

A natural and well-known method of lossless and near-lossless compression is predictive compression [8], of which differential pulse-coded modulation (DPCM) is a simple form. In lossless predictive compression, each sample is estimated from the previous sample values, and the difference between the estimate and the actual value is encoded. The decoder is able to compute the same estimates as the encoder, since those estimates are based on values it has already reconstructed.

Near-lossless predictive encoding has two main differences. First, the difference between the estimated sample value and the actual value is quantized before it is encoded (this is the lossy step). Second, the sample estimates must be based on the reconstructed sequence that will be observed by the decoder, not on the original samples, so that the decoder can form the same estimates as the encoder. Thus, the encoder must perform a limited form of decoding to form the reconstructed values.

The sample value estimation should be designed using knowledge of the anticipated signal characteristics. For images, the estimator typically incorporates several adjacent pixel values into its estimate. It often is desirable to make the estimation procedure adaptive, so that the estimator will perform well on a wide range of signals.

Let x_i denote the (possibly digitized) sample value at index i. For multidimensional signals, the index can be multidimensional (for example, the value of a pixel of an image would be denoted as $x_{(i,j)}$). The estimate of a sample value is denoted $\hat{x}_i$, and the reconstructed sample value is denoted $\tilde{x}_i$.

A basic form of the quantization procedure consists of quantizing the prediction error, $x_i - \hat{x}_i$, to the nearest multiple of a fixed quantization step size, q. Specifically, the index of the quantization level is given by

$$\eta_i = \left\lfloor \frac{x_i - \hat{x}_i}{q} + \frac{1}{2} \right\rfloor \tag{1}$$

The value of η_i is losslessly encoded in the compressed bit stream. The reconstructed sample is computed as

$$\tilde{x}_i = \hat{x}_i + q\eta_i \tag{2}$$

The values $\hat{x}_i + qj$, where j ranges over the integers, are known as the reconstruction levels of x_i. If there are minimum and maximum possible values of x_i, then Eq. (2) should be modified to include clipping to the allowed range. The effects of this clipping usually are insignificant and generally can be ignored in the analysis.

When the samples have been converted to digital form before the compression, we may assume each x_i is an integer and that the minimum difference between possible values of x_i is 1. (If not, the x_i can be scaled linearly so that this condition holds.) We assume that an analog-to-digital conversion step always yields values with this property.

Typically, a maximum allowable error value δ is supplied to or chosen by the predictive compressor. That is, for each i, we require $|\tilde{x}_i - x_i| \leq \delta$. If the x_i are integers (and δ is an integer), choosing $q = 2\delta + 1$ (or smaller) yields, through Eqs. (1) and (2), a procedure that produces reconstructed values satisfying this constraint.

Again suppose the x_i are integers. Although it is natural to choose q to be an odd integer, q also may be chosen to be an even integer. However, the possible reconstructed values of a sample then will range

from $x_i - q/2 + 1$ to $x_i + q/2$, giving an average of $x_i + 1/2$. Equivalently, the possible values of x_i given $\tilde{x}_i$ range from $\tilde{x}_i - q/2$ to $\tilde{x}_i + q/2 - 1$, so the best guess of x_i is $\tilde{x}_i - 1/2$, which is not an integer. There is nothing inherently wrong with this, since this bias is known and can be corrected (by using $\tilde{x}_i - 1/2$ as the reconstructed value), but it is often inconvenient to have reconstructed samples that have a different set of possible values than the original digitized samples. An advantage of allowing q to be even is that, if q is a power of 2, the quantization may be performed without a division operation.

IV. Basic Distortion Estimation

In this section, we give some well-known and useful properties of the RMSE distortion resulting from uniform quantization. Here we let x_i represent an original (analog) sample value and $\tilde{x}_i$ represent the corresponding quantized value. The RMSE distortion of a collection of $\tilde{x}_i$, denoted by D_{RMSE}, is given by

$$D_{\mathrm{RMSE}} = \sqrt{\frac{1}{n}\sum_i (\tilde{x}_i - x_i)^2}$$

where n is the number of samples.

When the final quantization step size is q, the reconstruction errors $\tilde{x}_i - x_i$ will be in the range[5] $[-q/2, q/2]$. If we assume that the reconstruction errors are uniformly distributed in this range, then D_{RMSE} is given by

$$D_{\mathrm{RMSE}} = \sqrt{\int_{-q/2}^{q/2} \frac{1}{q} x^2 dx} = \frac{q}{\sqrt{12}} \tag{3}$$

Thus, under the uniform error assumption, if the RMSE distortion is required to be at most a given value D_{RMSE}, then it suffices to choose

$$q \le \sqrt{12}\, D_{\mathrm{RMSE}} \tag{4}$$

If the quantization step is a simple analog-to-digital conversion or other quantization procedure in which the reconstruction levels are the same for each sample, then Eq. (3) generally is fairly accurate if the region over which the distortion is computed contains sample values whose range is somewhat larger than the quantization step size. If the quantization step occurs in a predictive compression algorithm, then Eq. (3) generally is slightly pessimistic (more so with larger q and less so when the signal is noisy) because the estimator should produce a distribution on $x_i - \hat{x}_i$ that has its peak near 0, which usually gives a distribution of $\tilde{x}_i - x_i$ that is slightly peaked at 0.

In Eq. (3), it matters little whether finer quantization (such as analog-to-digital conversion) was performed before the final, coarse quantization (but see Section VIII). It is common, however, to measure the distortion between the digitized samples and the final quantized samples, in part because the digitized samples are more often available than the original analog samples. We denote the RMSE distortion measured in this way by D'_{RMSE}. If the x_i are integers and δ is the maximum error value and $q = 2\delta + 1$, the possible reconstruction errors $\tilde{x}_i - x_i$ are

[5] Note that if the quantization procedure given by Eqs. (1) and (2) is used, the range will be the half-open interval $(-q/2, q/2]$.

$$\{-\delta, -\delta+1, \cdots, \delta-1, \delta\}$$

If it is assumed that the distribution of errors is uniform over this set, then

$$D'_{\mathrm{RMSE}} = \sqrt{\frac{\delta^2+\delta}{3}} \tag{5}$$

Solving Eq. (5) for δ shows that a distortion of at most D'_{RMSE} would be achieved if δ were chosen as

$$\delta = \left\lfloor \frac{-1+\sqrt{1+12(D'_{\mathrm{RMSE}})^2}}{2} \right\rfloor \tag{6}$$

It usually is reasonable to model the quantization noise introduced by the analog-to-digital conversion as uncorrelated with the additional noise introduced by the later quantization step. With D_{MSE} and D'_{MSE} defined analogously to D_{RMSE} and D'_{RMSE}, we have $D_{\mathrm{MSE}} = D'_{\mathrm{MSE}} + 1/12$ or, equivalently, $D_{\mathrm{RMSE}} = \sqrt{(D'_{\mathrm{RMSE}})^2 + 1/12}$. This is consistent with Eqs. (3) and (5).

V. Lossless and Near-Lossless Rate Comparison

Near-lossless compression has a large advantage over lossless compression in terms of the rate, in bits/sample, needed to encode a digitized signal.

Suppose the (integer) sample sequence $\{x_i\}$ can be modeled as a realization of a random process, $\{X_i\}$. Let r_0^* be the best possible expected rate, in bits/sample, needed to losslessly compress the sequence. (Note that r_0^* is actually the entropy of the entire sequence divided by the number of samples; however, we wish to emphasize practical coding here.) Now suppose that a near-lossless algorithm is used to compress the sequence such that the maximum sample error is δ, for some integer $\delta \geq 0$. Let r_δ be the expected rate achieved and let $\{\tilde{X}_i\}$ be the resulting reconstructed sequence (a random sequence that is a function of $\{X_i\}$). Each difference $\lfloor \tilde{X}_i \rfloor - X_i$ can take on at most $2\delta+1$ values. (For simplicity, we have assumed that the reconstructed sequence takes on integer values, although this assumption is not essential.) Thus, the sequence $\{\lfloor \tilde{X}_i \rfloor - X_i\}$ can be encoded with an expected rate r_e (with units of bits/sample) satisfying

$$r_e \leq \log_2(2\delta+1) \tag{7}$$

The sequence $\{X_i\}$ can be obtained from the sequences $\{\tilde{X}_i\}$ and $\{\lfloor \tilde{X}_i \rfloor - X_i\}$, so it follows that

$$r_0^* \leq r_\delta + r_e \tag{8}$$

Combining Eqs. (7) and (8) yields

$$r_\delta \geq r_0^* - \log_2(2\delta+1) \tag{9}$$

In other words, no near-lossless encoding algorithm that guarantees a maximum sample error of δ can achieve an expected rate that is lower than $\log_2(2\delta+1)$ less than the best expected rate achievable with a lossless encoder.

In practice, of course, it usually is impossible to determine r_0^* exactly. However, it is as difficult to approach the theoretical compression limits for near-lossless compression as it is for lossless compression (except in some uninteresting concocted cases), so Eq. (9) generally will be satisfied when r_0^* is replaced by the best lossless rate achieved. When δ is small, the rates achieved by practical near-lossless compression algorithms tend to be close to this bound. Incidentally, Eq. (9) may be generalized; if a compression algorithm achieves rate r while generating a distribution of reconstruction errors with entropy H_e, then Eq. (9) may be replaced by

$$r \geq r_0^* - H_e$$

Table 1 shows the potential and experimental rate advantage of using near-lossless compression, as opposed to lossless compression, for a variety of choices of δ. The maximum savings are a practical upper bound. The actual savings values are an indication of what is achievable; they were obtained for a simple near-lossless compression algorithm applied to 8-bit images. The top row of the table ($\delta = 0$) corresponds to lossless compression. The "munar," "thor," and "olaf" images are planetary images, and the "boat" image is an image with an outdoors setting. Clearly, there can be a large rate advantage over lossless compression, even if a small value of δ is used.

Table 1. Decrease in rate resulting from near-lossless compression with maximum sample error δ, as compared with lossless compression.

δ	Maximum savings, bits/sample	"Munar"		"Thor"		"Olaf"		"Boat"	
		Actual savings, bits/pixel	Resulting rate, bits/pixel	Actual savings, bits/pixel	Resulting rate, bits/pixel	Actual savings, bits/pixel	Resulting rate, bits/pixel	Actual savings, bits/pixel	Resulting rate, bits/pixel
0	0	0	5.05	0	4.97	0	4.70	0	4.65
1	1.58	1.54	3.51	1.50	3.47	1.53	3.17	1.51	3.14
2	2.32	2.23	2.82	2.16	2.81	2.19	2.51	2.15	2.50
3	2.81	2.68	2.37	2.59	2.38	2.62	2.08	2.54	2.11
4	3.17	3.00	2.05	2.90	2.07	2.88	1.82	2.81	1.84
5	3.46	3.21	1.84	3.11	1.86	3.04	1.66	3.00	1.65
6	3.70	3.35	1.70	3.25	1.72	3.15	1.55	3.13	1.52
7	3.91	3.46	1.59	3.35	1.62	3.23	1.47	3.22	1.43
8	4.09	3.53	1.52	3.43	1.54	3.29	1.41	3.28	1.37
9	4.25	3.59	1.46	3.49	1.48	3.33	1.37	3.32	1.33
10	4.39	3.64	1.41	3.54	1.43	3.37	1.33	3.36	1.29

VI. Compression of Noisy Signals

In this section, we assume the signal can be conceptually viewed as an underlying signal that is corrupted by additive noise. We refer to the noise as instrument noise, but it may be background noise in the physical signal. For meaningful estimation of total distortion (which includes instrument noise and distortion from compression), our primary requirement on the noise is that the scientists are not interested in the individual noise values. The noise statistics may be of interest, however. For example, an image of the surface of Mars may contain small pixel-to-pixel variations that look much like noise, but, if these variations actually exist on the surface, then they should be considered part of the underlying signal since scientists are presumably interested in their individual values. (For some purposes, statistics may be sufficient, so in such cases the variations could be treated as noise.) However, if there is also a

noticeable amount of noise from sources such as thermal noise in the sensor electronics or radiation hits on the sensor, then the distortion results in this section may apply. For rate calculations, the source of signal variations is obviously unimportant.

When near-lossless predictive compression is used to compress a signal containing a significant amount of memoryless noise, the predictor should attempt to estimate the underlying signal, since it usually cannot hope to predict the individual noise values. This estimation often consists of performing some sort of running average of the previously quantized samples. In such cases, the predictor's estimate of a sample may be fairly accurate when the underlying signal is changing slowly.

A. Rate Estimation

Compression of a noisy signal is similar to compression of noise, the performance of which can be estimated fairly accurately. Suppose the X_i sequence is formed from an underlying signal with zero-mean additive noise. We derive some useful bounds on the compression achievable. We are interested in the value of $(1/n)H(\{\tilde{X}_i\}_{i=1}^n)$, the entropy per sample of the random sequence of quantized instrument readings. For noisy signals, it often is possible to achieve compression that is very close to this value (perhaps within 0.1 to 0.3 bits/sample). Each $\tilde{X}_i$ is a quantized version of the instrument reading X_i. We denote a general quantization function by Q and a uniform quantization function with step size q (and a reconstruction level at 0) by Q_q; that is, $Q_q(x) = q\lfloor x/q + 1/2 \rfloor$. Each X_i is formed as the sum of the underlying signal Y_i and the instrument noise N_i.

For readability, we have relegated most calculations of this section to Appendix A.

Most generally, the average entropy per sample may be bounded by

$$\frac{1}{n} H\left(\left\{\tilde{X}_i\right\}_{i=1}^n\right) \geq \inf_{\{y_i\}_{i=1}^n} \frac{1}{n} H\left(\{Q(Y_i + N_i)\}_{i=1}^n \,|\{Y_i\}_{i=1}^n = \{y_i\}_{i=1}^n\right) \tag{10}$$

where the infimum is over all possible signal sequences. If the sequence $\{Y_i\}_{i=1}^n$ is independent from the sequence $\{N_i\}_{i=1}^n$, then

$$\frac{1}{n} H\left(\left\{\tilde{X}_i\right\}_{i=1}^n\right) \geq \inf_{\{y_i\}_{i=1}^n} \frac{1}{n} H\left(\{Q(y_i + N_i)\}_{i=1}^n\right) \tag{11}$$

If, in addition, the N_i are independent and identically distributed, then Eq. (11) becomes

$$\frac{1}{n} H\left(\left\{\tilde{X}_i\right\}_{i=1}^n\right) \geq \inf_y H\left(Q(y + N)\right) \tag{12}$$

where N is a random variable with the distribution of the N_i.

It is observed experimentally that Eqs. (11) and (12) are useful as approximations (and not simply lower bounds) to the rate obtained with near-lossless compression. They are most accurate when the underlying signal value changes slowly from sample to sample compared with the signal variations caused by the noise. It is not necessary for the overall variations in the underlying signal to be small compared with the noise. The right-hand sides of Eqs. (11) and (12) are good approximations for the respective left-hand sides when the estimator in the compressor is able to closely predict the values of the underlying signal. When this occurs, the entropy of the quantized residuals will be only slightly larger than the entropy of the quantized instrument noise. Only moderate accuracy in estimating Y_i is needed to accomplish this, since the entropy of a random variable grows as the logarithm of its standard deviation. If the

RMSE between the estimate and the underlying signal is D_{est} and the standard deviation of the noise is σ_N, then a rough estimate of the extra rate (in bits/sample) needed to encode the quantized signal plus noise (versus the noise only) is $\log_2 \sqrt{1 + D_{\text{est}}^2/\sigma_N^2}$. Note, however, that an algorithm for compressing the signal sequence $\{X_i\}_{i=1}^n$ may need to be more complicated than an algorithm for compressing the noise sequence $\{N_i\}_{i=1}^n$, even if there is little difference in the compression achieved; this is because estimation of the underlying signal is not needed for compression of $\{N_i\}_{i=1}^n$.

Naturally, if the behavior of the total signal (underlying signal plus noise) can be modeled accurately, then one can compute the entropy directly based on that model. However, it is often much more difficult to model the total signal (often, a purpose of the instrument is to provide data to develop such a model). In addition, the above results provide an intuitive feel for the effect of noise on the resulting rate.

We next turn our attention to the entropy of the quantized noise. We consider only the case when the quantizer is Q_q; that is, uniform with quantization step size q. Let $h(N)$ denote the differential entropy of the continuous random variable N, given as

$$h(N) = -\int_{-\infty}^{\infty} p_N(x) \log_2 p_N(x) dx \tag{13}$$

where p_N is the probability density function for N. If there is a reconstruction level at a, then the probabilities of the possible values of $Q_q(a + N)$ are given by

$$p_i = \int_{a+(i-1/2)q}^{a+(i+1/2)q} p_N(x) dx$$

for $i \in \mathbf{Z}$. With this notation, the entropy of $Q_q(a + N)$ is

$$H\left(Q_q(a + N)\right) = -\sum_{i \in \mathbf{Z}} p_i \log_2 p_i \tag{14}$$

With a little work (see Appendix A), we obtain

$$H\left(Q_q(a + N)\right) \geq -\log_2 q + h(N) \tag{15}$$

The right-hand side of Eq. (15) is independent of a, so it follows that

$$\inf_a H\left(Q_q(a + N)\right) \geq -\log_2 q + h(N)$$

In fact (see Theorem 9.3.1 of [2]), it is also true that

$$\lim_{q \downarrow 0} \left(H\left(Q_q(a + N)\right) + \log_2 q\right) = h(N)$$

and, for relatively smooth distributions, the approximation

$$H\left(Q_q(a + N)\right) \approx -\log_2 q + h(N) \tag{16}$$

becomes accurate enough for practical purposes about when q becomes smaller than the standard deviation of N. Overall,

$$\frac{1}{n} H\left(\left\{\tilde{X}_i\right\}_{i=1}^{n}\right) \geq -\log_2 q + h(N)$$

when the underlying signal is independent of the instrument noise and the latter is independent and identically distributed, with approximate equality holding in many practical situations.

The differential entropy of a distribution can be computed using Eq. (13); also, Table 16.1 of [2] contains differential entropies of many distributions. For convenience, we mention that, for the Gaussian distribution

$$p_N(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

the differential entropy is

$$h(N) = \frac{1}{2} \log_2\left(2\pi e \sigma^2\right)$$

and, for the Laplacian distribution

$$p_N(x) = \frac{\alpha}{2} e^{-\alpha|x-\theta|} \tag{17}$$

(which has variance $2/\alpha^2$), the differential entropy is

$$h(N) = \log_2\left(\frac{2e}{\alpha}\right)$$

As an example, we have computed bounds on $H(Q_q(a+N))$ for the Laplacian distribution. The result is shown in Fig. 1. Analytic expressions for the entropies involved are given in Appendix B.

In practice, it is not especially relevant to examine the range of entropies resulting from variations in a predictor offset, a. The offset a generally is not constant, so the overall discrete distribution (on the η_i) to be compressed actually will be a convex combination of the distributions resulting from all different possible values of a. Even if a is constant, the entropy coder may not be able to encode the resulting discrete distribution to near its entropy, because, if a is nonzero, then the peak of the distribution will be offset slightly from 0, which will make some entropy coders inefficient. This offset can be accounted for by the entropy coder if a is known, but, if a is known, the prediction could simply be adjusted so that a becomes 0.

Our calculations are useful in that they give an indication of whether Eq. (16) is likely to be accurate. In addition, these results can give insight into the effects of quantizing samples before near-lossless compression. (These effects are discussed in Section VIII.)

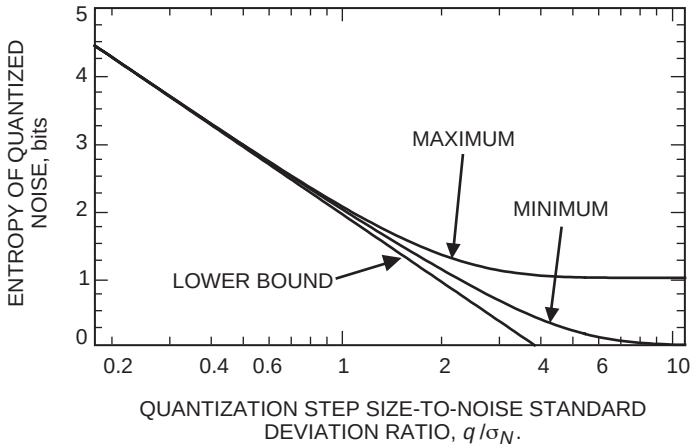

Fig. 1. Bounds on the entropy of quantized Laplacian noise. The upper two curves are $\sup_a H(Q_q(a+N))$ and $\inf_a H(Q_q(a+N))$, while the lower curve is the lower bound $h(N) - \log_2 q$.

B. Distortion and Bias

As in Section VI.A, we assume that the samples obtained from the instrument are random variables of the form $X = Y + N$, where Y is the underlying signal and N is the instrument noise. We assume that X may be approximated by a continuous (undigitized) random variable. The final reconstructed value of X is denoted $\tilde{X}$, obtained by uniform quantization with step size q. As long as the amount of noise is significant, it does not matter for these calculations if the final quantization step is during near-lossless compression or analog-to-digital conversion. We assume that N is memoryless noise with mean 0. (If N has a known nonzero mean, this value can be subtracted before further processing.)

A well-known approximation for the noise $\tilde{X} - X$ in the final quantization step is to assume it is uniformly distributed on $[-q/2, q/2]$ and uncorrelated with N. With this assumption, we have

$$E\left[\tilde{X} - Y\right] = 0 \tag{18}$$

$$E\left[\left(\tilde{X} - Y\right)^2\right] = E\left[N^2\right] + \frac{q^2}{12} \tag{19}$$

Note that $q^2/12$ is the MSE equivalent to the approximation Eq. (3) for the RMSE distortion between $\tilde{X}$ and X. The approximations of Eqs. (18) and (19) are together equivalent to the well-known corrections

$$\left.\begin{aligned} E\left[\tilde{X}\right] &= E[X] \\ E\left[\tilde{X}^2\right] &= E\left[X^2\right] + \frac{q^2}{12} \end{aligned}\right\} \tag{20}$$

for the quantization of pure noise [6].

When applying near-lossless compression to a noisy signal, it usually is logical to choose the quantization step size based on the resulting total (quantization-plus-instrument) noise. In this context, Eq. (19) may be rewritten

$$\sigma_{\text{total}} = \sqrt{\sigma_N^2 + D_{\text{RMSE}}^2} \tag{21}$$

where σ_N is the instrument-noise standard deviation, D_{RMSE} is the distortion, Eq. (3), from quantization, and σ_{total} is the standard deviation of the resulting total noise.

As an example, suppose that σ_N is estimated to be about 10.0, and it is determined that a value of σ_{total} up to 5 percent larger than σ_N can be tolerated (where the increase results from compression). From Eq. (21), we find that we need $D_{\text{RMSE}} \leq 3.2$. Applying Eq. (4) yields $q \leq 11.1$. [Alternatively, this value could have been obtained directly from Eq. (19).] If the samples have been digitized to integers before compression, we may apply near-lossless compression with maximum sample error $\delta = 5$. As the results of Sections V and VI.A indicate, this could result in a savings of more than 3 bits/sample as compared with lossless compression and result in a rate lower than 2 bits/sample.

We now turn our attention to the distribution of the total noise and in particular to the accuracy of Eqs. (18) and (19). Let $\hat{X}$ be a random variable indicating the reference level for the quantization (if the quantization occurs in a predictive compression algorithm, then $\hat{X}$ is the estimate of X as usual). Let $p_{\hat{X}-Y}(x)$ denote the density function, which we assume exists, of the random quantity $\hat{X} - Y$ and $p_{\tilde{X}-Y}(x)$ the density function of $\tilde{X} - Y$. The latter is presumably of most interest to scientists concerned with the effects of compression and noise on their data.

Using the standard quantization scheme, we may calculate

$$p_{\tilde{X}-Y}(x) = \Pr\left(N \in \left[x - \frac{q}{2}, x + \frac{q}{2}\right)\right) \sum_{i \in \mathbf{Z}} p_{\hat{X}-Y}(x + iq)$$

An example of the density function produced from this calculation in shown in Fig. 2. Although the total noise distribution is somewhat disturbing in appearance, it does have a zero mean, and its standard deviation is in line with Eqs. (21) and (3), so under many circumstances this would be acceptable to scientists.

In a near-lossless compression algorithm, there inevitably will be portions of the data where the estimator $\hat{X}$ will be biased. This is likely to happen in regions where the underlying signal is especially interesting; for example, in an image, the estimator may have trouble accurately estimating the values near one side of a boundary between two regions of different brightness. An example of the effect of this bias on the total noise distribution is illustrated in Fig. 3. In this case, the total noise distribution is more disturbing since it is not symmetric. However, the mean of this distribution is still extremely close to zero, so it still may be acceptable.

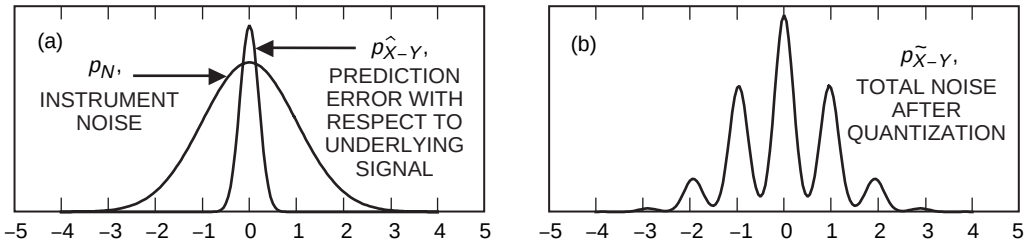

Fig. 2. Example noise distributions: (a) instrument-noise density function (Gaussian with mean 0 and standard deviation 1) and a possible prediction error $\hat{X} - Y$ density function (Gaussian with mean 0 and standard deviation 0.2) and (b) the resulting total noise density function when the quantization step size q is 1. Each density is plotted on a different vertical scale.

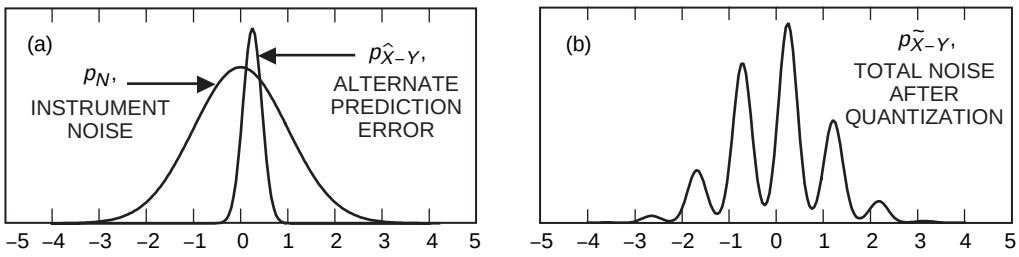

Fig. 3. Example noise distributions when the estimator is biased: (a) instrument-noise density function (Gaussian with mean 0 and standard deviation 1) and prediction-error density function (Gaussian with mean 0.25 and standard deviation 0.2) and (b) the resulting total noise density function when the quantization step size q is 1. Each density is plotted on a different vertical scale.

The bias in the total noise resulting from a given estimator value x is given by

$$\sum_{i \in \mathbf{Z}} (x + iq) \Pr\left(N \in \left[x + \left(i - \frac{1}{2}\right) q, x + \left(i + \frac{1}{2}\right) q\right)\right) \tag{22}$$

Figure 4 shows an example of this quantity as a function of x. Notice that this function always will be periodic with period q, since changing the estimate by any multiple of q does not affect how the signal will be quantized (although it will usually affect the cost in bits).

When N is Gaussian (with zero mean) and q/σ_N is less than about 4, the worst-case total noise bias occurs when the predicted value is very close to $\pm q/4$. We can thus obtain an accurate estimate of the worst-case bias by evaluating Eq. (22) at $x = q/4$. The result is shown in Fig. 5. Perhaps surprisingly, the bias is not significant when q/σ_N is less than about 2. The maximum bias decays to 0 exponentially in σ_N^2/q^2 [6].

When N has a distribution other than Gaussian, we expect the results to be similar.

The problem of determining the worst-case bias resulting from quantizing noisy samples is essentially equivalent to the problem of determining the effect of quantization on the measured mean of samples of noise. The latter problem is addressed in detail in [6] using complex Fourier series. Our Fig. 5 is equivalent to Fig. 4 in [6] (although they are presented quite differently). Also in [6] are results for uniform distributions and for sine waves.

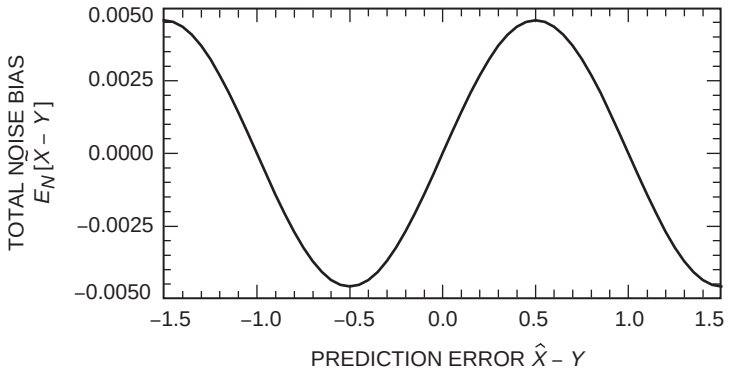

Fig. 4. Total noise bias as a function of the underlying signal prediction error value $\widehat{X} - Y$ when N is Gaussian with mean 0 and standard deviation 1, and the quantization step size is $q = 2$.

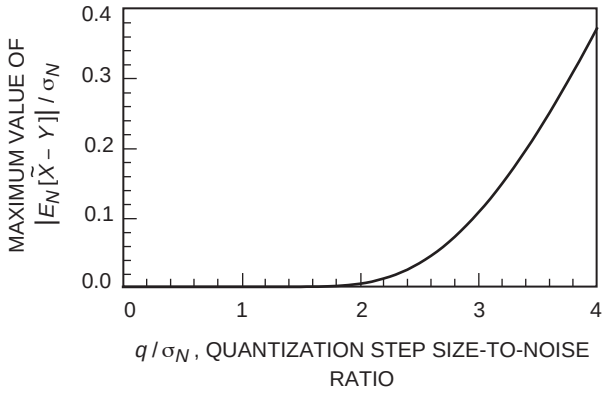

Fig. 5. Worst-case total noise bias as a function of q/σ_N, when N is Gaussian with mean 0.

The accuracy of the correction to the second moment in Eq. (20) also is considered in [6]. Those results are closely related to the accuracy of Eq. (19). If it is desired to determine the variance of a signal with great accuracy, then this may be important (and perhaps [6] should be consulted), but in our application the underlying signal is usually the primary concern, so any inaccuracies to Eq. (19) are corrections to a correction and generally will be insignificant.

VII. Artifacts and Dithering

When the instrument noise level is small compared with the final quantization step size, predictive near-lossless compression can result in disturbing artificial features (artifacts) in the reconstructed signal. Of course, the magnitude of the individual sample errors will be strictly limited, but several correlated errors can produce artifacts that are evident on close inspection or that may interfere with scientific analysis of the signal. These artifacts may include

(1) A biased average value of some regions of the signal

(2) Contouring, or step-like signal-value profiles, in slowly changing portions of the signal

(3) Erasure of faint features that would be detectable in the original signal because they occupy a large area

A technique for reducing or eliminating these artifacts is subtractive dither [10,12]. In the simplest form, which may be used with uniform quantization, a pseudorandom value, uniformly distributed on $[-q/2, q/2]$, is added to the sample before quantization occurs. The same value then is subtracted from the quantized sample during reconstruction. Thus, in a predictive compression algorithm, the reconstructed value is $\tilde{x} = Q_q(x - \hat{x} + D) + \hat{x} - D$, where D is the dither pseudorandom variable. It is well-known that the resulting quantization noise, $\tilde{x} - x$, is uniform on $[-q/2, q/2]$ and independent of x and $\hat{x}$.

When this dithering technique is used, the resulting quantization noise will satisfy Eq. (3) exactly. This usually will represent an increase in the RMSE as compared to when dithering is not used, since, as discussed in Section IV, Eq. (3) is generally pessimistic. The resulting rate (in bits/sample) generally will be higher also, since the values being quantized will have a larger variance. See [12] for some useful techniques for reducing these increases.

The reconstructed signal will have none of the artifacts mentioned above, since those artifacts occur due to correlations between the quantization noise and the signal. However, the entire signal may appear somewhat grainy due to the uniform noise on all samples.

It is possible to compromise between the dithering described above and no dither. This can be accomplished by using a dither signal that tends to take on values of smaller magnitude than would a signal uniformly distributed on $[-q/2, q/2]$. This dither signal still will increase the rate and distortion, but by a smaller amount, and it will remove some correlation between the reconstruction errors and the signal.

We denote a general dither distribution by P_D (so that $P_D(S) = \Pr(D \in S)$). Suppose $A(P_D)$ is a metric for the degree to which the reconstructed signal will contain artifacts, and $C(P_D)$ is a metric for the cost of the dither signal as manifested by the increase in rate and distortion. Then we would like to determine the dither signal distributions that achieve the optimal trade-off between minimizing $A(P_D)$ and minimizing $C(P_D)$.

Two reasonable choices for $C(P_D)$ are the variance of D and the second moment of D. (Note that these are the same if the dither distribution has mean 0.) Experimentally, the variance of D is a good indicator of the increase in distortion from dithering. It is logical that the variance of D also gives an indication of the increase in rate, since the rate generally is roughly equal to a constant plus the logarithm of the variance of the residual distribution. Figure 6(a) of [12] suggests that a similar relation will hold if the values of D are supplied to the entropy coder and decoder.

Reasonable choices of $A(P_D)$ are more complicated. When dithering is not used, the error $\tilde{x} - x$ in a reconstructed sample is dependent on the estimate $\hat{x}$ of the sample by $\tilde{x} - x = \hat{x} + Q_q(x - \hat{x}) - x$. When subtractive dither is used, this error is random, and the dependence between the reconstruction error and the estimate takes the form of a possible bias in the reconstructed value. Specifically, $E[\tilde{x} - x] = E[Q_q(x - \hat{x} + D) - (x - \hat{x} + D)]$. Treating the signal and the estimate as random variables makes this quantity a random variable; the (random) bias is then

$$E_D\left[Q_q\left(X - \hat{X} + D\right) - \left(X - \hat{X} + D\right)\right]$$

To deal more easily with this expression, we introduce quantities that behave better than X and $\hat{X}$. Let $R = X - \hat{X} - Q_q(X - \hat{X})$. Note that R is the difference between a sample and its estimate, translated by a multiple of q to be in the range $[-q/2, q/2]$. If no dither is used, R is the error in the reconstructed sample, and, when a dither signal is used, R still should be distributed in the same way as the no-dither reconstruction error (that is, ideally uniformly distributed over $[-q/2, q/2]$ but in practice typically slightly peaked at 0, as remarked in Section IV). Let $\tilde{R} = Q_q(R + D) - D$, so that $\tilde{R}$ is the quantized value of subtractively dithered R. Note that if no dither is used, then $\tilde{R} = 0$.

With these definitions, we have

$$\begin{aligned}
E_D\left[\tilde{R} - R\right] &= E_D\left[Q_q(R + D) - D - \left(X - \hat{X}\right) + Q_q\left(X - \hat{X}\right)\right] \\
&= E_D\left[Q_q\left(X - \hat{X} - Q_q\left(X - \hat{X}\right) + D\right) - \left(X - \hat{X} + D\right) + Q_q\left(X - \hat{X}\right)\right] \\
&= E_D\left[Q_q\left(X - \hat{X} + D\right) - \left(X - \hat{X} + D\right)\right]
\end{aligned}$$

so $E_D\left[\tilde{R} - R\right]$ expresses the sample bias in terms of R and D. We let $A(P_D)$ be the mean-squared value of this bias (averaged over R); that is,

$$A(P_D) = E_R\left[\left(E_D\left[\tilde{R} - R\right]\right)^2\right]$$

or, equivalently,

$$A(P_D) = \int_{[-q/2,q/2]} \left(r - E\left[\tilde{R}|R = r\right]\right)^2 dP_R(r) \tag{23}$$

We primarily consider the case when R is distributed uniformly on $[-q/2, q/2]$ (or at least when we weight the bias uniformly over this range of R) so that

$$A(P_D) = \int_{-q/2}^{q/2} \left(r - E\left[\tilde{R}|R = r\right]\right)^2 dr \tag{24}$$

This metric is discussed in [10] and [13]. Intuitively, $A(P_D)$ indicates the degree to which the expected mean of a sample can be biased by the quantization reference point.

Suppose $A(P_D)$ is given by Eq. (24) and $C(P_D)$ is the variance or the second moment of P_D. When no dither is used, $C = 0$ and $A = q^2/12$. When a dither that is uniform on $[-q/2, q/2]$ is used, $C = q^2/12$ and $A = 0$. Note that 0 is the minimum value of both $A(P_D)$ and $C(P_D)$. Our definitions of A and C are intended for comparison among dither distributions when q is constant. There is no obvious significance of a specific value of A or C, so we can only determine the range of best compromises between minimization of A and minimization of C. In a particular application, experimentation and subjective judgment will be needed to determine which of these "best compromises" to use.

In Appendix C, we present several results concerning the nature of optimal dither distributions. In particular, we show (Theorem C-5) that, when $A(P_D)$ is given by Eq. (24) and $C(P_D)$ is either the variance or the second moment of P_D, then the optimal trade-off between $C(P_D)$ and $A(P_D)$ occurs for dither distributions that are uniform on $[-k/2, k/2]$ for $k \in [0, q]$ (where $k = 0$ corresponds to no dither).

The discrete case also is considered in Appendix C. In that case, the samples are integers, $q = 2\delta + 1$, where δ is the maximum absolute error allowed, and D must take on integer values. When $A(P_D)$ is the discrete analogue of Eq. (24) and $C(P_D)$ is the second moment of P_D, then the optimal distributions are those that are uniform on $\{-k, \cdots, k\}$, where $k \in \{0, \cdots, \delta\}$, or are a convex combination of two such distributions with consecutive values of k (Theorem C-9).

As an example of the use of these results, an image we refer to as "munar" was compressed with a simple predictive algorithm with maximum sample error $\delta = 2$. A dither signal uniformly distributed on $\{-k, \cdots, k\}$ was used, where $k = 0$, 1, or 2. Note that $k = 0$ corresponds to no dither and $k = 2$ corresponds to standard subtractive dither. The results are given in Table 2. Note how the cost of dithering increases as the dither signal amplitude increases. For comparison, the lossless compression rate obtained with the same algorithm was 5.049 bits/pixel. Observe that the error distribution becomes more uniform as the dither signal amplitude increases. Figure 6 contains a portion of the original "munar" image, and Fig. 7 shows an enlarged and contrast-enhanced detail area of the original and reconstructed images, from which it can be seen how the dither signal amplitude affects the appearance of the reconstructed image. We make no claim as to which version is best. When displayed normally, the original and all three reconstructed images are virtually indistinguishable.

Table 2. Result of compressing "munar" with maximum pixel error $\delta = 2$ and a dither distribution uniform on $\{-k, \cdots, k\}$.

k	Rate, bits/pixel	RMSE	Error distribution, percentage of pixels				
			-2	-1	0	1	2
0	2.819	1.370	18.6	21.2	22.1	20.1	18.0
1	2.842	1.405	19.8	20.2	20.5	20.1	19.5
2	2.873	1.414	19.9	20.0	19.9	20.1	20.0

Fig. 6. A portion of the "munar" image used in our dither example. The detail area shown in Fig. 7 is indicated.

When the instrument-noise level is not small as compared with the final quantization step size, the instrument noise itself reduces the occurrence of harmful artifacts in the reconstructed data. In this case, dithering could be used to prevent total noise distributions like those in Figs. 2(b) and 3(b) from occurring, but, as indicated in Section VI.B, such distributions may not be very harmful. Thus, in this case, dithering may not be worthwhile. However, the rate and distortion costs of dithering are small when the signal is noisy.

VIII. Analog-to-Digital Conversion and Near-Lossless Compression

In addition to the quantization step size q used for near-lossless compression, the quantization step size used in the analog-to-digital conversion, which we call q_d, also affects the rate-distortion performance of the compression. This effect is secondary but important.

Decreasing q_d never results in worse rate-distortion performance. Thus, one could simply make q_d as small as possible (very fine quantization) and ignore compression considerations. However, there frequently are practical considerations that make it more convenient to use a larger q_d: increasing the number of bits of precision of analog-to-digital conversion may be difficult or expensive; it may be convenient to use, say, 16 bits to store an integer instead of 20; and the compression software may be quicker if it can work with smaller integers.

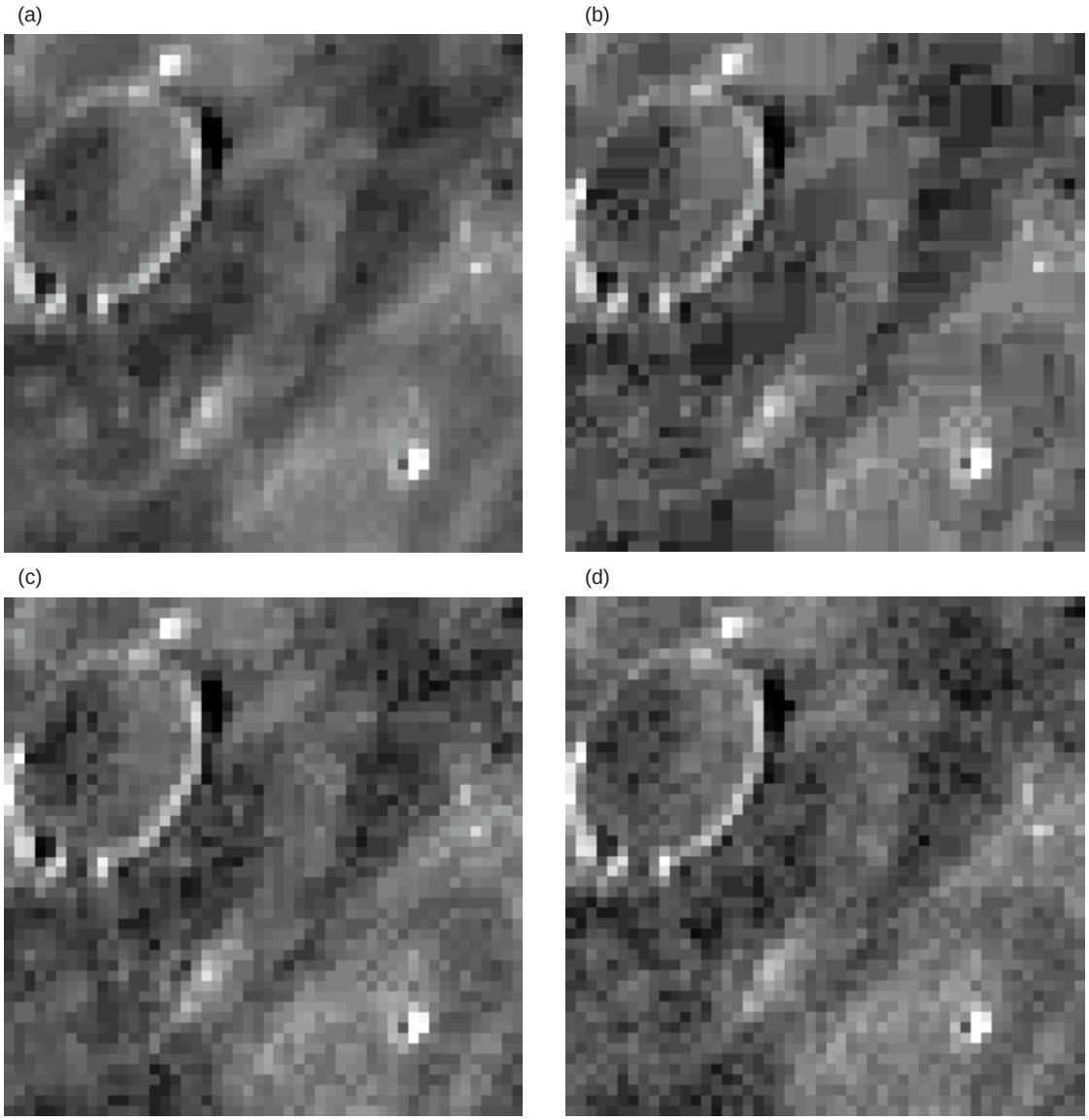

Fig. 7. Dither example with maximum pixel error δ = 2. Each image is magnified and contrast enhanced: (a) detail of original, (b) compressed and decompressed with *k* = 0 (no dither), (c) *k* = 1, and (d) *k* = 2 (standard dither). As *k* increases from 0 to 2, the appearance of streaks and artificial regions of constant intensity decreases, but an overall grainy look becomes more evident.

For the purposes of this discussion, we consider q_d to be the step size used in the last quantization performed with a fixed set of reconstruction levels. (The set of reconstruction levels varies with the sample estimate, $\hat{x}$, in lossy predictive compression.) Thus, we include the possibility of reducing the precision of the data by dropping some of the least-significant bits after the analog-to-digital conversion. This may be useful if one desires to vary q_d (perhaps with the noise level of the data). On occasion the crude strategy of dropping a number of bits (depending on the noise level) and sending the resulting quantized data compressed losslessly ($q_d = q$) can be convenient and reasonably effective, although better results usually are obtained with $q_d < q$. (A rough comparison of lossy compression to dropping bits and using lossless compression is given in [14]; however, near-lossless compression is not considered there.)

For a given q, a possible detrimental effect of a larger q_d is that the estimate of a sample (in the near-lossless compression) will be more limited in what values it can take on. This is best illustrated

with an example. Suppose the signal is noisy but the underlying signal value is accurately estimated (perhaps by averaging over several previous quantized samples) to have the value 11. Suppose that $q = 4$. If $q_d = 1$, then the estimate can be exactly 11, but if q_d is 2, then the estimate may be constrained to be either 10 or 12. Thus, the entropy of the quantized residuals may be larger. The maximum extent of this rate increase may be estimated by examining the effect of prediction bias on the entropy of the quantized residuals. When the signal is noisy, we may use the results of Section VI.A for this purpose. The difference between the worst-case and best-case curves in Fig. 1 is an indication of the approximate maximum cost of a prediction bias when the noise has a Laplacian distribution, as a function of q/σ_N. However, the bias needed to give this maximum cost is $q/2$, so it can occur only when $q_d = q$ (lossless compression). Note that the cost could occur with respect to arbitrarily small q_d or to q_d as large as $q/2$. A slightly higher distortion also may result.

The other detrimental effects of coarse digitization are more difficult to quantify analytically. These effects are a cost in both rate and distortion due to the fact that the predictor is less accurate because the previous samples are encoded less accurately. The rate is increased because of the less accurate estimate, and the distortion is increased because the distribution of distortion values is less peaked (it is closer to being uniform on $[-q/2, q/2]$).

We offer general guidelines for the fineness needed in the analog-to-digital conversion. No significant detrimental effects occur if $q_d \leq \max\{q/4, \sigma_N/4\}$. Thus, if the minimum possible amount of instrument noise is known, it is safe to digitize to no coarser than 1/4 of the standard deviation of this noise. Alternatively, if it is known that the compressor will not use a quantizer step size smaller than q, then it is safe to choose $q_d \leq q/4$. (The latter is perhaps less likely in general since usually some amount of data is sent losslessly, but it may apply to a subset of the data.)

IX. Conclusion

This article has covered several aspects of quantization that will be of interest to those using data compression as well as to those implementing it. The more important practical observations concern the rate achievable with near-lossless compression, the largest acceptable quantization step size, and the use of dither. Near-lossless compression has a large rate improvement over lossless compression, even if the maximum allowed sample error δ is small, as illustrated in Table 1. When the signal to be compressed is noisy, the effect of quantization on the scientific value of the signal often can be determined rather precisely. Figure 5 shows that surprisingly large quantization step sizes may be used without introducing a significant bias into the reconstructed samples. Subtractive dither is a technique for reducing or eliminating artifacts in the reconstructive image. We have found the best dither signal distributions to use to produce a range of degrees of dithering. Examples are shown in Fig. 7.

Our results suggest some possibilities for onboard analysis of signal data. We have been concerned with the ability to accurately measure certain properties of a signal, especially the mean of a group of samples. It is natural to ask whether one may simply measure the desired statistics onboard the spacecraft and transmit those statistics, along with some general descriptive information of the signal to give context to the statistics. Clearly, such a strategy has the potential to greatly reduce the volume of transmitted data without sacrificing analysis accuracy.

Such a strategy often is worth consideration and likely will prove useful in a variety of situations. However, there are complications in onboard analysis that will ensure that there is a continuing demand for distortion-controlled compression. The primary disadvantage of onboard analysis is that the spacecraft may not be able to determine which regions of the data should be characterized by statistics. The regions of interest may be determined using data from another instrument or another (perhaps future) mission. It may be very difficult to determine regions of interest automatically, since they depend on recognition of some feature in the data such as, in the case when the signal is an image, a crater, fault, or shadow.

The regions may be irregularly shaped, and there may be too many possible regions of interest to account for all possibilities. Scientists often will not be able to determine beforehand what sort of interesting features may be present. Finally, it may be necessary to remove spurious samples (occurring, e.g., from radiation) from the region of interest before determining statistics.

Future work on distortion-controlled compression could include compression in which the local regions of constrained distortion are larger than a single sample. Although such a generalization probably would not yield significantly better algorithms for high-fidelity compression, there is evidence that such compression could be competitive over a wider range of fidelities than near-lossless compression.

Acknowledgments

The author would like to thank Aaron Kiely and Sam Dolinar for many helpful discussions.

References

[1] P. Billingsley, *Convergence of Probability Measures*, New York: John Wiley & Sons, Inc., 1968.

[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, New York: Wiley-Interscience, 1991.

[3] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Boston: Kluwer Academic Publishers, 1992.

[4] S. W. Golomb, "Run-Length Encodings," *IEEE Transactions on Information Theory*, vol. IT-12, no. 3, pp. 399–401, July 1966.

[5] A. B. Kiely, "Bit-Wise Arithmetic Coding for Data Compression," *The Telecommunications and Data Acquisition Progress Report 42-117, January–March 1994*, Jet Propulsion Laboratory, Pasadena, California, pp. 145–160, May 15, 1994. http://tmo.jpl.nasa.gov/tmo/progress_report/42-117/117n.pdf

[6] I. Kollár, "Bias of Mean Value and Mean Square Value Measurements Based on Quantized Data," *IEEE Transactions on Instrumentation and Measurement*, vol. 43, no. 5, pp. 733–739, October 1994.

[7] H. W. Kuhn and A. W. Tucker, "Nonlinear Programming," *Proc. 2nd Berkeley Symposium on Mathametical Statistics and Probability*, University of California Press, Berkeley, pp. 481–492, 1951.

[8] M. Rabbani and P. Jones, *Digital Image Compression Techniques*, SPIE Publications, 1991.

[9] R. F. Rice, *Some Practical Universal Noiseless Coding Techniques*, JPL 79-22, Jet Propulsion Laboratory, Pasadena, California, March 1979.

[10] L. G. Roberts, "Picture Coding Using Pseudo-Random Noise," *IRE Transactions on Information Theory*, vol. 8, pp. 145–154, February 1962.

[11] W. Rudin, *Principles of Mathematical Analysis*, third edition, New York: McGraw-Hill, Inc., 1976.

[12] D. W. E. Schobben, R. A. Beuker, and W. Oomen, "Dither and Data Compression," *IEEE Transactions on Signal Processing*, vol. 45, no. 8, pp. 2097–2101, August 1997.

[13] L. Schuchman, "Dither Signals and Their Effect on Quantization Noise," *IEEE Transactions on Communication Technology*, vol. COM-12, pp. 162–165, December 1964.

[14] J. C. Tilton and M. Manohar, "Radiometric Resolution Enhancement by Lossy Compression as Compared to Truncation Followed by Lossless Compression," *Proc. Science Information Management and Data Compression Workshop*, pp. 27–39, 1994.

[15] I. H. Witten, R. M. Neal, and J. G. Cleary, "Arithmetic Coding for Data Compression," *Communications of the ACM*, vol. 30, no. 6, pp. 520–540, June 1987.

[16] R. C. Wood, "On Optimum Quantization," *IEEE Transactions on Information Theory*, vol. IT-15, no. 2, pp. 248–252, March 1969.

[17] X. Wu, W. K. Choi, and P. Bao, "L_∞-Constrained High-Fidelity Image Compression via Adaptive Context Modeling," *Proc. of the 1997 Data Compression Conference (DCC '97)*, pp. 91–100, 1997.

Appendix A
Rate-Estimation Calculations for Noisy Signals

This Appendix contains calculations for results in Section VI.A.

Let A and B be two arbitrary random variables. Suppose B takes on values in a set S with probability one. Then

$$H(A) \geq H(A|B) \geq \inf_{b \in S} H(A|B=b)$$

Using similar reasoning,

$$\begin{aligned}
\frac{1}{n} H\left(\left\{\tilde{X}_i\right\}_{i=1}^{n}\right) &= \frac{1}{n} H\left(\{Q(Y_i+N_i)\}_{i=1}^{n}\right) \\
&\geq \frac{1}{n} H\left(\{Q(Y_i+N_i)\}_{i=1}^{n} \,|\, \{Y_i\}_{i=1}^{n}\right) \\
&\geq \inf_{\{y_i\}_{i=1}^{n}} \frac{1}{n} H\left(\{Q(Y_i+N_i)\}_{i=1}^{n} \,|\, \{Y_i\}_{i=1}^{n} = \{y_i\}_{i=1}^{n}\right)
\end{aligned}$$

where the infimum is over all[6] possible signal sequences. This is Eq. (10). If the sequence $\{Y_i\}_{i=1}^{n}$ is independent from the sequence $\{N_i\}_{i=1}^{n}$, then

$$H\left(\{Q\left(Y_i+N_i\right)\}_{i=1}^{n} \,|\, \{Y_i\}_{i=1}^{n} = \{y_i\}_{i=1}^{n}\right) = H\left(\{Q\left(y_i+N_i\right)\}_{i=1}^{n}\right)$$

so Eq. (11) follows.

We now derive Eq. (15). Define f on $[0,\infty)$ by $f(x) = -x \log_2 x$. Starting from Eq. (14), we have

$$\begin{aligned}
H\left(Q_q(a+N)\right) &= -\sum_{i \in \mathbf{Z}} p_i \log_2 p_i \\
&= -\sum_{i \in \mathbf{Z}} p_i \left(\log_2 q + \log_2 \frac{p_i}{q}\right) \\
&= -\log_2 q - q \sum_{i \in \mathbf{Z}} \frac{p_i}{q} \log_2 \frac{p_i}{q} \\
&= -\log_2 q + q \sum_{i \in \mathbf{Z}} f\left(\frac{p_i}{q}\right)
\end{aligned}$$

[6] Actually, a set with probability one is sufficient.

Note that f is convex $\cap$ on its domain. This implies that

$$\begin{aligned} qf\left(\frac{p_i}{q}\right) &= qf\left(\int_{a+(i-1/2)q}^{a+(i+1/2)q} p_N(x)\frac{1}{q}dx\right) \\ &\geq q\int_{a+(i-1/2)q}^{a+(i+1/2)q} f(p_N(x))\frac{1}{q}dx \\ &= \int_{a+(i-1/2)q}^{a+(i+1/2)q} f(p_N(x))dx \end{aligned}$$

where we have used Jensen's inequality. Thus,

$$\begin{aligned} H(Q_q(a+N)) \geq & -\log_2 q + \sum_{i\in\mathbf{Z}}\int_{a+(i-1/2)q}^{a+(i+1/2)q} f(p_N(x))dx \\ = & -\log_2 q + \int_{-\infty}^{\infty} f(p_N(x))dx \\ = & -\log_2 q - \int_{-\infty}^{\infty} p_N(x)\log_2 p_N(x)dx \\ = & -\log_2 q + h(N) \end{aligned}$$

of which the final result is Eq. (15).

Appendix B
Entropy of the Quantized Laplacian Distribution

In this Appendix, we compute the entropy of a Laplacian random variable that has been uniformly quantized, when the quantization levels are offset from 0 by a. These results are referred to in Section VI.A, where they are shown as an indication of the accuracy of entropy bounds.

Suppose N is a Laplacian random variable, with parameter α as in Eq. (17) but with zero mean ($\theta = 0$). We wish to find $H(Q_q(a+N))$. For convenience, let $\beta = a/q$. From the symmetries involved, we need only consider $\beta \in [0, 1/2]$; thus, we assume β is in this range. Also let $c = \alpha q$. It is straightforward (but tedious) to compute

$$\begin{aligned} H(Q_q(a+N)) = {} & \frac{1}{2}e^{-c(\beta+1/2)}\left(-\ln\left(\frac{1}{2}-\frac{1}{2}e^{-c}\right)+c\left(\beta-\frac{1}{2}\right)+\frac{c}{1-e^{-c}}\right) \\ & + \frac{1}{2}e^{-c(-\beta+1/2)}\left(-\ln\left(\frac{1}{2}-\frac{1}{2}e^{-c}\right)+c\left(-\beta-\frac{1}{2}\right)+\frac{c}{1-e^{-c}}\right) \\ & + \left(\frac{1}{2}e^{-c(-\beta+1/2)}+\frac{1}{2}e^{-c(\beta+1/2)}-1\right)\ln\left(1-\frac{1}{2}e^{-c(-\beta+1/2)}-\frac{1}{2}e^{-c(\beta+1/2)}\right) \end{aligned}$$

in nats. Note that dividing this result by $\ln 2$ yields the result in bits. For a fixed value of c, the extreme values for $\beta \in [0, 1/2]$ apparently always occur at the end points of that interval. Specifically, the minimum always occurs at $\beta = 0$ with

$$H(Q_q(N)) = e^{-c/2}\left(-\ln\left(\frac{1}{2}-\frac{1}{2}e^{-c}\right)+\frac{c+ce^{c}}{2e^{c}-2}\right)+\left(e^{-c/2}-1\right)\ln\left(1-e^{-c/2}\right) \tag{B-1}$$

in nats, and the maximum always occurs at $\beta = 1/2$, where

$$H\left(Q_q\left(\frac{q}{2}+N\right)\right) = -\ln\left(\frac{1}{2}-\frac{1}{2}e^{-c}\right)+\frac{c}{e^{c}-1} \tag{B-2}$$

in nats. Note that these results can be given in terms of q/σ_N by letting $c = \sqrt{2}q/\sigma_N$. The approximation Eq. (16) for this noise distribution, which is also the lower bound of Eq. (15), becomes

$$H\left(Q_q\left(a+N\right)\right) \approx \log_2\left(\sqrt{2}e\right)+\log_2\left(\frac{\sigma_N}{q}\right) \approx 1.9427+\log_2\left(\frac{\sigma_N}{q}\right) \tag{B-3}$$

Figure 1 of the main text compares the functions Eqs. (B-1) and (B-2), converted to bits, and Eq. (B-3), which is already in bits.

Appendix C
Optimal Dithering

In this Appendix, we demonstrate that the dither distributions described in Section VII are optimal as claimed. Along the way we present general results that could simplify the process of determining optimal distributions if the basic assumptions are modified.

I. Continuous Case

We first consider the case in which R and D may take on a continuum of values. We scale the numbers involved so that $q = 1$.

A. General Results

Although our choices of $A(P_D)$ and $C(P_D)$ under which we find optimal dither distributions are reasonable, it is possible that a more detailed analysis of a particular situation may yield more refined functions $A(P_D)$ and $C(P_D)$. We have not analyzed any specific alternate formulation in detail, but we present some results that may be helpful in determining optimal dither distributions for alternate formulations. Theorems C-1, C-3, and C-4 give cases in which it is sufficient to consider distributions that are concentrated on certain intervals or that are symmetric. Theorem C-2 gives conditions guaranteeing the existence of optimal dither distributions.

In this section, we frequently use the notation P_D to denote the probability measure corresponding to the random variable D. We extend this notation to expressions such as $P_{f(D)}$, P_{D+c}, and P_{-D}, where, for example, $P_{f(D)}$ is the probability measure corresponding to the random quantity $f(D)$. When we say that P_D is concentrated on a set S, we mean $P_D(\mathbf{R} - S) = 0$.

We will make use of the topology of weak convergence, for which a good reference is Appendix III of [1]. This is a topology on the set of probability measures on $\mathbf{R}$, under which a sequence $\{P_i\}$ converges to P if and only if for all continuous and bounded functions $f : \mathbf{R} \to \mathbf{R}$,

$$\lim_{i\to\infty} \int f dP_i = \int f dP$$

This topology is equivalent to the weak-$*$ ("weak-star") topology and is important because it is small enough ("coarse" enough) to make many commonly occurring sets compact while at the same time large enough ("fine" enough) to allow needed continuity properties.

For these results, we assume $A(P_D)$ is of the form of Eq. (23), but we no longer assume P_R is uniform on $[-1/2, 1/2]$. We also do not assume that $C(P_D)$ is the variance of D or the second moment of D. However, we do place several restrictions on C. We assume that the cost of no dither is 0, and that $C(P_D) \geq 0$ for each P_D. The cost function also must be one of the following two types:

Type A: For any (measurable) function $f : \mathbf{R} \to \mathbf{R}$ satisfying $0 \leq |f(x)| \leq |x|$ for all real x, and for any P_D, we require $C(P_{f(D)}) \leq C(P_D)$.

Type B: The cost function satisfies the following:

(1) The cost is invariant to translations of D; that is, for any x, we have $C(P_{D+x}) = C(P_D)$.

(2) For any (measurable) function $f : \mathbf{R} \to \mathbf{R}$ satisfying $0 \leq |f(x)| \leq |x|$ for all real x, and for any P_D with mean 0, we require that $C(P_{f(D)}) \leq C(P_D)$.

Intuitively, a Type A cost function must assign higher cost to distributions that have larger amplitudes. A Type A cost function must be symmetric about 0 in the sense that $C(P_D) = C(P_{-D})$, since with $f(x) = -x$ we have $C(P_D) \leq C(P_{-D}) \leq C(P_D)$. The second moment of D is a Type A cost function.

A Type B cost function must be symmetric about μ when restricted to distributions with mean μ. The variance of D is a Type B cost function.

Theorem C-1. *For any $P_{D'}$, there exists a $P_{D''}$ with $C(P_{D''}) \leq C(P_{D'})$ and $A(P_{D''}) = A(P_{D'})$, and*

(i) *if C is a Type A cost function, then $P_{D''}$ can be chosen to be concentrated on $[-1/2, 1/2]$;*

(ii) *if C is a Type B cost function, then $P_{D''}$ can be chosen to be concentrated on an interval of the form $[c - 1/2, c + 1/2]$, where $c \in [-1/2, 1/2]$;*

(iii) *if C is a Type B cost function that is continuous under the topology of weak convergence, then $P_{D''}$ can be chosen so that its mean, μ, is in $[-1/2, 1/2]$ and $P_{D''}$ is concentrated on $[\mu - 1/2, \mu + 1/2]$.*

Note that Theorem C-1 implies that, in all cases that obey our general restrictions on the cost function, it suffices to consider dither distributions that are concentrated on $[-1, 1]$.

Proof. Rewriting Eq. (23) for $q = 1$ gives

$$A(P_D) = \int_{[-1/2,1/2]} \left(r - E\left[\tilde{R}|R = r\right]\right)^2 dP_R(r) \tag{C-1}$$

Since $\tilde{R}$ is defined as $\tilde{R} = Q_1(R + D) - D$, we have

$$E\left[\tilde{R}|R = r\right] = E\left[Q_1(r + D) - D\right] \tag{C-2}$$

It is clear from Eqs. (C-1) and (C-2) that $A(P_D)$ can be determined from $P_{(D \bmod 1)}$. Intuitively, shifting any part of the distribution P_D by any integer does not affect $A(P_D)$. In particular, if we let $f(x) = x - \lfloor x + 1/2 \rfloor$, then $A(P_{f(D)}) = A(P_D)$. Note that f satisfies $0 \leq |f(x)| \leq |x|$ and $f(x) \in [-1/2, 1/2]$ for all real x.

In Case (i), we let $P_{D''} = P_{f(D')}$. This $P_{D''}$ is concentrated on $[-1/2, 1/2]$ since $f(x) \in [-1/2, 1/2]$, and the Type A property implies $C(P_{D''}) \leq C(P_{D'})$.

In Case (ii), we first form P_{D_1} by translating $P_{D'}$ by an integer ($D_1 = D' + i$ for some i) so that the mean of D_1 is in $[-1/2, 1/2]$. Let c be this mean. Clearly, D_1 and D' are equivalent modulo 1, so $A(P_{D_1}) = A(P_{D'})$, and by Type B Property (1), $C(P_{D_1}) = C(P_{D'})$. Now let $D'' = c + f(D_1 - c)$. Then D'' and D_1 are equivalent modulo 1 so $A(P_{D'}) = A(P_{D''}) = A(P_D)$. We also have

$$C\left(P_{D''}\right) = C\left(P_{c+f(D_1-c)}\right) = C\left(P_{f(D_1-c)}\right) \leq C\left(P_{D_1-c}\right) = C\left(P_{D_1}\right) = C\left(P_{D'}\right)$$

where we have used both Type B properties. Note that $P_{f(D_1-c)}$ is concentrated on $[-1/2, 1/2]$, so $P_{D''}$ is concentrated on $[c - 1/2, c + 1/2]$. Thus, Case (ii) is established. Note that c might not be the mean of D''.

In Case (iii), a formal proof can be obtained by parametrizing all distributions that are equivalent to $P_{D'}$ modulo 1 and concentrated on a closed interval of width 1 with midpoint in $[-1/2, 1/2]$. At least

one of these distributions must have cost less than or equal to $C(P_{D'})$. It can be shown that, among all of the parametrized distributions, there exists one, call it P_D^*, that achieves the minimum cost. This distribution usually will be concentrated on the interval of width 1 about its mean. If not, then we can form a sequence of parameters that converge to one for which the distribution does satisfy this condition. We omit the details. ❒

Theorem C-2. *If the (Type A or B) cost function $C(P_D)$ is continuous under the topology of weak convergence and P_R can be described by a density function p_R (that is, $P_R(S) = 0$ whenever the Lebesgue measure of S is 0), then for each $\alpha \geq 0$ there exists a P_D^* that minimizes $A(P_D)$ subject to $C(P_D) \leq \alpha$.*

Proof. By Theorem C-1, we may restrict our attention to dither distributions that are concentrated on $[-1, 1]$. Let T be the set of all such distributions. It follows from Prohorov's Theorem [1] that T is compact under the topology of weak convergence. Let $S = \{P_D \in T : C(P_D) \leq \alpha\}$. Our continuity assumption on C implies that S is closed under the topology of weak convergence. Since $S \subset T$, it follows that S is compact under the topology of weak convergence.

Next we show that $A(P_D)$ is continuous under the topology of weak convergence. Suppose $\{P_{D_n}\}_{n=1}^{\infty}$ is a sequence of probability measures that converges to P_D under the topology of weak convergence. It suffices to show that $\lim_{n\to\infty} A(P_{D_n}) = A(P_D)$.

Using the density function p_R, we can write

$$A\left(P_D\right) = \int_{-1/2}^{1/2} \left(r + E\left[D - Q_1(r + D)\right]\right)^2 p_R(r)dr$$

Thus,

$$A(P_D) = \int_{-1/2}^{1/2} \left(r + E[D] - \sum_{i\in\mathbf{Z}} iP_D\left(\left[i - r - \frac{1}{2}, i - r + \frac{1}{2}\right)\right)\right)^2 p_R(r)dr$$

$$= \int_{-1/2}^{1/2} \left(\int_{\mathbf{R}} \left(x + r - \left\lfloor x + r + \frac{1}{2} \right\rfloor \right) dP_D(x)\right)^2 p_R(r)dr$$

Let $f_r(x) = x + r - \lfloor x + r + 1/2 \rfloor$ and let $g_r(x) = x + r - \lceil x + r - 1/2 \rceil$. We then may write

$$A(P_D) = \int_{-1/2}^{1/2} \left(\int_{\mathbf{R}} f_r(x) dP_D(x)\right)^2 p_R(r)dr \tag{C-3}$$

We will eventually apply Lebesgue's Dominated Convergence Theorem (see [11], for example) to this formulation of A to conclude that $\lim_{n\to\infty} A(P_{D_n}) = A(P_D)$.

Observe that $g_r(x) \geq f_r(x)$ and if $x \notin \mathbf{Z} - r - 1/2$, then $g_r(x) = f_r(x)$. Each f_r is lower semicontinuous and each g_r is upper semicontinuous; it thus follows [1] that

$$\liminf_{n\to\infty} \int_{\mathbf{R}} f_r(x) dP_{D_n}(x) \geq \int_{\mathbf{R}} f_r(x) dP_D(x) \tag{C-4}$$

and

$$\limsup_{n\to\infty} \int_{\mathbf{R}} g_r(x) dP_{D_n}(x) \leq \int_{\mathbf{R}} g_r(x) dP_D(x) \tag{C-5}$$

Also note that, since $g_r(x) \geq f_r(x)$, we have

$$\limsup_{n\to\infty} \int_{\mathbf{R}} g_r(x) dP_{D_n}(x) \geq \limsup_{n\to\infty} \int_{\mathbf{R}} f_r(x) dP_{D_n}(x) \tag{C-6}$$

If $P_D(\mathbf{Z} - r - 1/2) = 0$, then

$$\int_{\mathbf{R}} f_r(x) dP_D(x) = \int_{\mathbf{R}} g_r(x) dP_D(x) \tag{C-7}$$

and combining Eqs. (C-4) through (C-7) yields

$$\liminf_{n\to\infty} \int_{\mathbf{R}} f_r(x) dP_{D_n}(x) \geq \int_{\mathbf{R}} f_r(x) dP_D(x) \geq \limsup_{n\to\infty} \int_{\mathbf{R}} f_r(x) dP_{D_n}(x)$$

Thus, if $P_D(\mathbf{Z} - r - 1/2) = 0$, then

$$\lim_{n\to\infty} \int_{\mathbf{R}} f_r(x) dP_{D_n}(x) = \int_{\mathbf{R}} f_r(x) dP_D(x) \tag{C-8}$$

Note that the set $\{r : P_D(\mathbf{Z} - r - 1/2) > 0\}$ must have Lebesgue measure zero.

Finally, we bound the integrand (of the outer integral) of Eq. (C-3). Since $|f_r(x)| \leq 1/2$, we have, for each n,

$$\left| \int_{\mathbf{R}} f_r(x) dP_{D_n}(x) \right| \leq \frac{1}{2}$$

and, thus,

$$\left(\int_{\mathbf{R}} f_r(x) dP_{D_n}(x) \right)^2 p_R(r) \leq \frac{1}{4} p_R(r)$$

The Dominated Convergence Theorem may now be applied to formulation Eq. (C-3) of A to conclude that

$$\lim_{n\to\infty} A(P_{D_n}) = A(P_D)$$

establishing the continuity of A under the topology of weak convergence.

The theorem now follows from the fact that a continuous function achieves its minimum on a compact set. ❒

Theorem C-3. *Suppose P_R is symmetric about 0 and can be described by a density function p_R. Suppose also that C is a Type A cost function and C is convex $\cup$. Then, for any $P_{D'}$, there exists a $P_{D''}$ that is symmetric about 0, is concentrated on $[-1/2, 1/2]$, and for which $A(P_{D''}) \leq A(P_{D'})$ and $C(P_{D''}) \leq C(P_{D'})$.*

Note that Theorem C-3 implies that, when its hypothesis holds, it suffices to consider dither distributions that are symmetric about 0 (and thus have mean 0) and are concentrated on $[-1/2, 1/2]$.

Proof. The symmetry condition on P_R implies $p_R(r) = p_R(-r)$. We first show that $A(P_{-D'}) = A(P_{D'})$. Using the functions f_r and g_r defined in the proof of Theorem C-2 and starting with Eq. (C-3), we have

$$\begin{aligned}
A(P_{-D'}) &= \int_{-1/2}^{1/2} \left(\int_{\mathbf{R}} f_r(x) dP_{-D'}(x) \right)^2 p_R(r) dr \\
&= \int_{-1/2}^{1/2} \left(\int_{\mathbf{R}} \left(x + r - \left\lfloor x + r + \frac{1}{2} \right\rfloor \right) dP_{-D'}(x) \right)^2 p_R(r) dr \\
&= \int_{-1/2}^{1/2} \left(\int_{\mathbf{R}} \left(-y - s - \left\lfloor -y - s + \frac{1}{2} \right\rfloor \right) dP_{D'}(y) \right)^2 p_R(-s) ds \\
&= \int_{-1/2}^{1/2} \left(\int_{\mathbf{R}} \left(y + s + \left\lfloor -y - s + \frac{1}{2} \right\rfloor \right) dP_{D'}(y) \right)^2 p_R(s) ds \\
&= \int_{-1/2}^{1/2} \left(\int_{\mathbf{R}} \left(y + s - \left\lceil y + s - \frac{1}{2} \right\rceil \right) dP_{D'}(y) \right)^2 p_R(s) ds \\
&= \int_{-1/2}^{1/2} \left(\int_{\mathbf{R}} g_s(y) dP_{D'}(y) \right)^2 p_R(s) ds
\end{aligned}$$

where for the third equality we have changed variables ($y = -x$ and $s = -r$) and replaced $dP_{-D'}(-y)$ by $dP_{D'}(y)$. If $P_{D'}(\mathbf{Z} - s - 1/2) = 0$, then $\int_{\mathbf{R}} g_s(y) dP_{D'}(y) = \int_{\mathbf{R}} f_s(y) dP_{D'}(y)$. But the set $\{s : P_{D'}(\mathbf{Z} - s - 1/2) > 0\}$ must have Lebesgue measure zero, so

$$\int_{-1/2}^{1/2} \left(\int_{\mathbf{R}} g_s(y) dP_{D'}(y) \right)^2 p_R(s) ds = \int_{-1/2}^{1/2} \left(\int_{\mathbf{R}} f_s(y) dP_{D'}(y) \right)^2 p_R(s) ds = A(P_{D'})$$

and, thus, $A(P_{-D'}) = A(P_{D'})$ as claimed.

Now let $P_{D_1} = (P_{D'} + P_{-D'})/2$. Note that P_{D_1} is symmetric about 0. From Eq. (C-3), it can be seen that $A(P_D)$ is a convex $\cup$ function of P_D, and, by hypothesis, $C(P_D)$ is a convex $\cup$ function of P_D. Therefore, $A(P_{D_1}) \leq A(P_{D'})$ and $C(P_{D_1}) \leq C(P_{D'})$.

Now let f be given by

$$f(x) = \begin{cases} x - \left\lceil x - \frac{1}{2} \right\rceil & \text{if } x \geq 0 \\ x - \left\lfloor x + \frac{1}{2} \right\rfloor & \text{if } x \leq 0 \end{cases} \tag{C-9}$$

Note that f is symmetric in that $f(x) = -f(-x)$. Also, $f(x) - x \in \mathbf{Z}$ and $f(x) \in [-1/2, 1/2]$. Let $P_{D''} = P_{f(D_1)}$. Then $A(P_{D''}) = A(P_{D_1}) \leq A(P_{D'})$ and, by the Type A property, $C(P_{D''}) \leq C(P_{D_1}) \leq C(P_{D'})$. Note that $P_{D''}$ is symmetric and concentrated on $[-1/2, 1/2]$, so the proof is complete. ❒

Theorem C-4. *Suppose P_R is the uniform distribution on $[-1/2, 1/2]$. Suppose also that $C(P_D)$ is a Type B cost function and that $C(P_D)$ is convex $\cup$ when restricted to P_D with mean 0. Then, for any $P_{D'}$, there exists a $P_{D''}$ that is symmetric about 0, is concentrated on $[-1/2, 1/2]$, and for which $A(P_{D''}) \leq A(P_{D'})$ and $C(P_{D''}) \leq C(P_{D'})$.*

Proof. Under the hypothesis, Eq. (C-3) becomes

$$A(P_D) = \int_{-1/2}^{1/2} \left(\int_{\mathbf{R}} f_r(x) dP_D(x) \right)^2 dr$$

From the definition of f_r, for any real c we may write

$$\int_{\mathbf{R}} f_r(x) dP_{D+c}(x) = \int_{\mathbf{R}} f_{r-c}(x) dP_D(x)$$

But for a fixed x, the value of $f_r(x)$ depends only on r mod 1, and as r ranges over $[-1/2, 1/2]$, r mod 1 ranges over the same range as $(r - c)$ mod 1. Thus,

$$\begin{aligned} A(P_{D+c}) &= \int_{-1/2}^{1/2} \left(\int_{\mathbf{R}} f_r(x) dP_{D+c}(x) \right)^2 dr \\ &= \int_{-1/2}^{1/2} \left(\int_{\mathbf{R}} f_{r-c}(x) dP_D(x) \right)^2 dr \\ &= \int_{-1/2}^{1/2} \left(\int_{\mathbf{R}} f_r(x) dP_D(x) \right)^2 dr \\ &= A(P_D) \end{aligned}$$

Thus, $A(P_D) = A(P_{D+c})$, so $A(P_D)$ is invariant to translations of D. Let D_1 be D' translated to have mean 0. We have shown that $A(P_{D_1}) = A(P_{D'})$ and, by Type B Property (1), we have $C(P_{D_1}) = C(P_{D'})$.

As in the proof of Theorem C-3, we have

$$A(P_{-D_1}) = \int_{-1/2}^{1/2} \left(\int_{\mathbf{R}} f_r(x) dP_{-D_1}(x) \right)^2 dr$$

$$= \int_{-1/2}^{1/2} \left(\int_{\mathbf{R}} g_r(x) dP_{D_1}(x) \right)^2 dr$$

$$= \int_{-1/2}^{1/2} \left(\int_{\mathbf{R}} f_r(x) dP_{D_1}(x) \right)^2 dr$$

$$= A(P_{D_1})$$

Now let $P_{D_2} = (P_{D_1} + P_{-D_1})/2$. Then P_{D_2} is symmetric about 0. Since A is convex $\cup$, $A(P_{D_2}) \leq A(P_{D_1})$. By hypothesis, C is also convex $\cup$ in a region that includes P_{D_1}, P_{-D_1}, and P_{D_2}, and, by Type B Property (2), we have $C(P_{D_1}) = C(P_{-D_1})$, so $C(P_{D_2}) \leq C(P_{D_1})$.

Using f as given by Eq. (C-9), let $P_{D''} = P_{f(D_2)}$. Then $A(P_{D''}) = A(P_{D_2}) \leq A(P_{D'})$ and, by Type B Property (2), $C(P_{D''}) \leq C(P_{D_2}) \leq C(P_{D'})$. Note that $P_{D''}$ is symmetric about 0 and concentrated on $[-1/2, 1/2]$, so the proof is complete. ❒

B. Solution for the Basic Continuous Case

We now prove the following result.

Theorem C-5. *Suppose $A(P_D)$ is given by Eq. (24) and $C(P_D)$ is either the variance or the second moment of P_D. Suppose $k \in [0, q]$ and let P_D^* be the uniform distribution on $[-k/2, k/2]$. Then P_D^* minimizes $A(P_D)$ subject to $C(P_D) = k^2/12$.*

Proof. The theorem is obvious when $k = 0$, so we assume that $k \in (0, q]$. It is easily calculated that $A(P_D^*) = (1-k)^2/12$ and $C(P_D^*) = k^2/12$. Without loss of generality, we assume that $q = 1$.

When $C(P_D)$ is the second moment of P_D, by Theorem C-3 it suffices to find the optimum P_D among those which are concentrated on $[-1/2, 1/2]$ and are symmetric about 0. When $C(P_D)$ is the variance of P_D, Theorem C-4 implies the same result. Since these distributions have mean 0, both cost functions become equivalent. We may, therefore, confine our attention to P_D that are concentrated on $[-1/2, 1/2]$ and are symmetric about 0, and to the cost function $C(P_D) = 2\int_{[0,1/2]} x^2 dP_D(x)$, which exploits the symmetry of the P_D.

Using Eqs. (C-1) and (C-2) and the symmetry of the P_D, we have

$$A(P_D) = 2\int_0^{1/2} \left(r - E\left[Q_1(r+D) - D\right]\right)^2 dr$$

$$= 2\int_0^{1/2} \left(r - E\left[Q_1(r+D)\right]\right)^2 dr$$

When $0 < r < 1/2$ and $-1/2 \leq D \leq 1/2$, observe that

$$Q_1(r+D) = \begin{cases} 1 & \text{if } r + D \geq \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$

so that

$$E\left[Q, (r+D)\right] = \Pr\left(D \geq \frac{1}{2} - r\right) = P_D\left(\left[\frac{1}{2} - r, \frac{1}{2}\right]\right)$$

and, thus,

$$A(P_D) = 2\int_0^{1/2} \left(r - P_D\left(\left[\frac{1}{2} - r, \frac{1}{2}\right]\right)\right)^2 dr$$

As already mentioned, A is a convex $\cup$ function of P_D. It also is apparent that the distributions we need to consider with $C(P_D) = k^2/12$ form a convex set. Thus, P_D^* must minimize A if

$$\frac{\partial}{\partial \epsilon} A\left(\left(1-\epsilon\right) P_D^* + \epsilon P_D\right)\Big|_{\epsilon=0} \geq 0 \tag{C-10}$$

for all P_D with $C(P_D) = k^2/12$.

We have

$$\begin{aligned}
\frac{\partial}{\partial \epsilon} A\left((1-\epsilon)P_D^* + \epsilon P_D\right) &= \frac{\partial}{\partial \epsilon} 2\int_0^{1/2} \left(r - (1-\epsilon)P_D^*\left(\left[\frac{1}{2} - r, \frac{1}{2}\right]\right) - \epsilon P_D\left(\left[\frac{1}{2} - r, \frac{1}{2}\right]\right)\right)^2 dr \\
&= 4\int_0^{1/2} \left(r - (1-\epsilon)P_D^*\left(\left[\frac{1}{2} - r, \frac{1}{2}\right]\right) - \epsilon P_D\left(\left[\frac{1}{2} - r, \frac{1}{2}\right]\right)\right) \\
&\quad \times \left(P_D^*\left(\left[\frac{1}{2} - r, \frac{1}{2}\right]\right) - P_D\left(\left[\frac{1}{2} - r, \frac{1}{2}\right]\right)\right) dr
\end{aligned}$$

and so

$$\frac{\partial}{\partial\epsilon}A\left((1-\epsilon)P_D^* + \epsilon P_D\right)\Big|_{\epsilon=0}$$

$$= 4\int_0^{1/2}\left(r - P_D^*\left(\left[\frac{1}{2}-r,\frac{1}{2}\right]\right)\right)\left(P_D^*\left(\left[\frac{1}{2}-r,\frac{1}{2}\right]\right) - P_D\left(\left[\frac{1}{2}-r,\frac{1}{2}\right]\right)\right)dr$$

$$= -4\int_0^{1/2-k/2} rP_D\left(\left[\frac{1}{2}-r,\frac{1}{2}\right]\right)dr$$

$$+ 4\int_{1/2-k/2}^{1/2}\left(r - \left(\frac{1}{2}-\frac{1}{2k}+\frac{r}{k}\right)\right)\left(\frac{1}{2}-\frac{1}{2k}+\frac{r}{k} - P_D\left(\left[\frac{1}{2}-r,\frac{1}{2}\right]\right)\right)dr$$

$$= -4\int_0^{1/2-k/2} rP_D\left(\left[\frac{1}{2}-r,\frac{1}{2}\right]\right)dr + 4\int_{1/2-k/2}^{1/2}\left(r - \left(\frac{1}{2}-\frac{1}{2k}+\frac{r}{k}\right)\right)\left(\frac{1}{2}-\frac{1}{2k}+\frac{r}{k}\right)dr$$

$$- 4\int_{1/2-k/2}^{1/2}\left(r - \left(\frac{1}{2}-\frac{1}{2k}+\frac{r}{k}\right)\right)P_D\left(\left[\frac{1}{2}-r,\frac{1}{2}\right]\right)dr$$

$$= \frac{k}{12} - \frac{k^2}{12} - 4\int_0^{1/2} rP_D\left(\left[\frac{1}{2}-r,\frac{1}{2}\right]\right)dr$$

$$+ 4\int_{1/2-k/2}^{1/2}\left(\frac{1}{2}-\frac{1}{2k}+\frac{r}{k}\right)\left(P_D\left(\left[\frac{1}{2}-r,\frac{k}{2}\right]\right) + P_D\left(\left(\frac{k}{2},\frac{1}{2}\right]\right)\right)dr$$

$$= \frac{k}{12} - \frac{k^2}{12} + \frac{k}{2}P_D\left(\left(\frac{k}{2},\frac{1}{2}\right]\right) - 4\int_0^{1/2} rP_D\left(\left[\frac{1}{2}-r,\frac{1}{2}\right]\right)dr$$

$$+ 4\int_{1/2-k/2}^{1/2}\left(\frac{1}{2}-\frac{1}{2k}+\frac{r}{k}\right)P_D\left(\left[\frac{1}{2}-r,\frac{k}{2}\right]\right)dr$$

Now observe that

$$\begin{aligned}-4\int_0^{1/2} rP_D\left(\left[\frac{1}{2}-r,\frac{1}{2}\right]\right)dr &= -4\int_0^{1/2} r\int_{[1/2-r,1/2]} dP_D(x)dr \\ &= -4\int_{[0,1/2]}\int_{1/2-x}^{1/2} r\,dr\,dP_D(x) \\ &= \int_{[0,1/2]}(2x^2-2x)dP_D(x)\end{aligned}$$

and a similar calculation gives

$$4\int_{1/2-k/2}^{1/2}\left(\frac{1}{2}-\frac{1}{2k}+\frac{r}{k}\right)P_D\left(\left[\frac{1}{2}-r,\frac{k}{2}\right]\right)dr=\int_{[0,k/2]}\left(2x-\frac{2x^2}{k}\right)dP_D(x)$$

Also note that

$$\frac{k}{2}P_D\left(\left(\frac{k}{2},\frac{1}{2}\right]\right)=\int_{(k/2,1/2]}\frac{k}{2}dP_D(x)$$

and, since $C(P_D)=2\int_{[0,1/2]}x^2dP_D(x)=k^2/12$, we have

$$\frac{k}{12}-\frac{k^2}{12}=\int_{[0,1/2]}\left(\frac{2x^2}{k}-2x^2\right)dP_D(x)$$

Substituting these results into our earlier expression gives

$$\begin{aligned}
\frac{\partial}{\partial\epsilon}A\left((1-\epsilon)P_D^*+\epsilon P_D\right)\Big|_{\epsilon=0} &= \int_{[0,1/2]}\left(\frac{2x^2}{k}-2x^2\right)dP_D(x)+\int_{(k/2,1/2]}\frac{k}{2}dP_D(x)\\
&\quad+\int_{[0,1/2]}\left(2x^2-2x\right)dP_D(x)+\int_{[0,k/2]}\left(2x-\frac{2x^2}{k}\right)dP_D(x)\\
&= \int_{[0,k/2]}\left(\left(\frac{2x^2}{k}-2x^2\right)+(2x^2-2x)+\left(2x-\frac{2x^2}{k}\right)\right)dP_D(x)\\
&\quad+\int_{(k/2,1/2]}\left(\left(\frac{2x^2}{k}-2x^2\right)+\frac{k}{2}+(2x^2-2x)\right)dP_D(x)\\
&= 0+\int_{(k/2,1/2]}\left(\frac{2x^2}{k}-2x+\frac{k}{2}\right)dP_D(x)\\
&= \int_{(k/2,1/2]}\frac{2}{k}\left(x-\frac{k}{2}\right)^2dP_D(x)\\
&\geq 0
\end{aligned}$$

so Eq. (C-10) holds as promised; thus, P_D^* is optimal and the proof is complete. ❒

II. Discrete Case

For the discrete case, we assume that R and D take on integer values. We consider only the case in which q is odd, with $q=2\delta+1$. We use P_D to denote the (discrete) dither probability distribution, with $P_D(i)=\Pr(D=i)$. Most of the notation is the same as in the continuous case, and we rely on context to distinguish the two. In the basic discrete case, $A(P_D)$ is defined as

$$A(P_D) = \frac{1}{2\delta+1} \sum_{r=-\delta}^{\delta} \left(r - E\left[Q_q(r+D) - D\right]\right)^2 \tag{C-11}$$

and the cost $C(P_D)$ is the second moment of D.

A. General Results

Again we remove some of our specific assumptions on $A(P_D)$ and $C(P_D)$. We consider $A(P_D)$ of the form

$$A(P_D) = \frac{1}{2\delta+1} \sum_{r=-\delta}^{\delta} \left(r - E\left[Q_q(r+D) - D\right]\right)^2 P_R(r)$$

where P_R is the (discrete) distribution on R. We require that the cost of no dither be 0 and that $C(P_D) \geq 0$ for each P_D. The cost function also must be one of the following two types:

Type A: For any function $f : \mathbf{Z} \to \mathbf{Z}$ satisfying $0 \leq |f(i)| \leq |i|$ for all integers i, and for any P_D, we require $C(P_{f(D)}) \leq C(P_D)$.

Type B: The cost function satisfies the following:

(1) The cost is invariant to translations of D; that is, for any integer i, we have $C(P_{D+i}) = C(P_D)$.

(2) For any function $f : \mathbf{R} \to \mathbf{R}$ satisfying $0 \leq |f(x)| \leq |x|$ and $f(x) - x \in \mathbf{Z}$ for all real x, and for any P_D, we require that $C(P_{f(D-\mu)+\mu}) \leq C(P_D)$, where μ is the mean of D.

Intuitively, a Type A cost function must assign higher cost to distributions that have larger amplitudes. As is the continuous case, a Type A cost function must be symmetric about 0 in the sense that $C(P_D) = C(P_{-D})$, since with $f(i) = -i$, we have $C(P_D) \leq C(P_{-D}) \leq C(P_D)$. The second moment of D is a Type A cost function.

A Type B cost function must be symmetric about 0 when restricted to distributions with mean 0. The variance of D is a Type B cost function.

Theorem C-6. *For any $P_{D'}$, there exists a $P_{D''}$ with $C(P_{D''}) \leq C(P_{D'})$ and $A(P_{D''}) \leq A(P_{D'})$, and*

(i) *if C is a Type A cost function, then $P_{D''}$ can be chosen to be concentrated on $\{-\delta, \cdots, \delta\}$;*

(ii) *if C is a Type B cost function, then $P_{D''}$ can be chosen to be concentrated on a set of the form $\{c-\delta, \cdots, c+\delta\}$, where $c \in \{-\delta, \ldots, \delta\}$.*

Note that Theorem C-6 implies that, in all cases that obey our general restriction on the cost function, it suffices to consider dither distributions that are concentrated on $\{-(q-1), \cdots, q-1\}$. The proof of Theorem C-6 is analogous to the proof of (i) and (ii) of Theorem C-1. We omit the details.

Theorem C-7. *If $C(P_D)$ is continuous when restricted to P_D that are concentrated on $\{-(q-1), \cdots, q-1\}$, then for each $\alpha \geq 0$ there exists a P_D that minimizes $A(P_D)$ subject to $C(P_D) \leq \alpha$.*

Theorem C-7 is the discrete analogue of Theorem C-2. The former is much easier to prove, however, since A is readily seen to be continuous and the set

$$\{P_D : C(P_D) \le \alpha \text{ and } P_D \text{ is concentrated on } \{-(q-1), \cdots, q-1\}\}$$

is compact. Again we omit the details.

Theorem C-8. *Suppose P_R is symmetric about 0 and C is a Type A cost function that is convex $\cup$. Then, for any $P_{D'}$, there exists a $P_{D''}$ that is symmetric about 0, is concentrated on $\{-\delta, \cdots, \delta\}$, and for which $A(P_{D''}) \le A(P_{D'})$ and $C(P_{D''}) \le C(P_{D'})$.*

Theorem C-8 is the discrete analogue of Theorem C-3 and is straightforward to prove (we omit the details).

Note that there is no simple discrete analogue of Theorem C-4. Our results for the basic discrete case apply when the cost function is the second moment of P_D and do not apply when the cost function is the variance of P_D.

B. Solution for the Basic Discrete Case

We now consider our basic discrete case. Note that the result below does not apply to the case when $C(P_D)$ is the variance of P_D, since with this cost function there are dither distributions with nonzero mean that are better than any with mean 0.

Theorem C-9. *Suppose $A(P_D)$ is given by Eq. (C-11) and $C(P_D)$ is the second moment of P_D. Suppose P_D^* is the discrete distribution given by*

$$P_D^*(i) = \begin{cases} \alpha & \text{if } |i| < k \\ \dfrac{1+\alpha}{2} - k\alpha & \text{if } |i| = k \\ 0 & \text{if } |i| > k \end{cases}$$

where $k \in \{1, \cdots, \delta\}$ and $\alpha \ge (1+\alpha)/2 - k\alpha$. Then P_D^* minimizes $A(P_D)$ subject to $C(P_D) = C(P_D^*)$.

Note that $\alpha \ge (1+\alpha)/2 - k\alpha$ implies

$$\alpha \ge \frac{1}{2k+1} \tag{C-12}$$

and that P_D^* is a convex combination of the uniform distributions on $\{-k, \cdots, k\}$ and $\{-(k-1), \cdots, k-1\}$.

Proof. By Theorem C-8, it suffices to find the optimum P_D among those that are concentrated on $\{-\delta, \cdots, \delta\}$ and are symmetric about 0. With these conditions, we may rewrite Eq. (C-11) as

$$A(P_D) = \frac{2}{2\delta+1} \sum_{r=1}^{\delta} \left(r - E\left[Q_q(r+D)\right]\right)^2$$

and the cost becomes

$$C(P_D) = 2\sum_{i=1}^{\delta} i^2 P_D(i) \tag{C-13}$$

When $r \in \{1, \cdots, \delta\}$,

$$Q_q(r+D) = \begin{cases} 2\delta+1 & \text{if } r+D > \delta \\ 0 & \text{otherwise} \end{cases}$$

Thus,

$$A(P_D) = \frac{2}{2\delta+1} \sum_{r=1}^{\delta} \left(r - (2\delta+1)\Pr(D > \delta - r)\right)^2$$

$$= (4\delta+2) \sum_{r=1}^{\delta} \left(\frac{r}{2\delta+1} - \sum_{j=\delta-r+1}^{\delta} P_D(j) \right)^2 \tag{C-14}$$

As in the continuous case, $A(P_D)$ is a convex $\cup$ function of P_D. By the Kuhn–Tucker conditions [7], our choice of P_D^* minimizes $A(P_D)$ subject to $C = C(P_D^*)$ if there exists a λ such that

$$\left.\frac{\partial A}{\partial P_D(i)}\right|_{P_D=P_D^*} + \lambda \left.\frac{\partial C}{\partial P_D(i)}\right|_{P_D=P_D^*} = 0 \qquad (1 \le i \le k) \tag{C-15}$$

and

$$\left.\frac{\partial A}{\partial P_D(i)}\right|_{P_D=P_D^*} + \lambda \left.\frac{\partial C}{\partial P_D(i)}\right|_{P_D=P_D^*} \ge 0 \qquad (k < i \le \delta) \tag{C-16}$$

Interestingly, it turns out to be unnecessary to include the condition $\sum_i P_D(i) = 1$.

Starting from Eq. (C-14), we have

$$\frac{\partial A}{\partial P_D(i)} = (8\delta+4) \sum_{r=\delta-i+1}^{\delta} \left(\sum_{j=\delta-r+1}^{\delta} P_D(j) - \frac{r}{2\delta+1} \right)$$

and so

$$\left.\frac{\partial A}{\partial P_D(i)}\right|_{P_D=P_D^*} = -4 \sum_{r=\delta-i+1}^{\delta} r + \begin{cases} (8\delta+4) \displaystyle\sum_{r=\delta-i+1}^{\delta} \left((r-\delta+k-1)\alpha + \frac{1+\alpha}{2} - k\alpha \right) & \text{if } i \le k \\ (8\delta+4) \displaystyle\sum_{r=\delta-k+1}^{\delta} \left((r-\delta+k-1)\alpha + \frac{1+\alpha}{2} - k\alpha \right) & \text{if } i > k \end{cases}$$

These sums are easily evaluated, yielding

$$\left.\frac{\partial A}{\partial P_D(i)}\right|_{P_D=P_D^*} = (2 - 2\alpha - 4\alpha\delta)i^2 \qquad (1 \le i \le k)$$

and

$$\left.\frac{\partial A}{\partial P_D(i)}\right|_{P_D=P_D^*} = (2-2\alpha-4\alpha\delta)i^2 + 2(1+2\delta)(i-k)(\alpha i+\alpha k-1) \qquad (k<i\le\delta) \tag{C-17}$$

Since Eq. (C-17) applies when $i \ge k+1$, we have $i-k \ge 1$ and $\alpha i+\alpha k-1 \ge \alpha(2k+1)-1 \ge 0$ [using Eq. (C-12)]; thus,

$$\left.\frac{\partial A}{\partial P_D(i)}\right|_{P_D=P_D^*} \ge (2-2\alpha-4\alpha\delta)i^2 \qquad (k<i\le\delta)$$

From Eq. (C-13), we have

$$\frac{\partial C}{\partial P_D(i)} = 2i^2$$

so Eqs. (C-15) and (C-16) are satisfied when $\lambda = -1+\alpha+2\alpha\delta$, and we have proven that our P_D^* is optimal. Note that these optimal P_D^* cover the useful range of costs, from 0 to $(\delta^2+\delta)/3$. ❒