The Validity of the Additive Noise Model for Uniform Scalar Quantizers

Daniel Marco, *Member, IEEE,* and David L. Neuhoff, *Fellow, IEEE*

***Abstract*—A uniform scalar quantizer with small step size, large support, and midpoint reconstruction levels is frequently modeled as adding orthogonal noise to the quantizer input. This paper rigorously demonstrates the asymptotic validity of this model when the input probability density function (pdf) is continuous and satisfies several other mild conditions. Specifically, as step size decreases, the correlation between input and quantization error becomes negligible relative to the mean-squared error (MSE). The model is even valid when the input density is discontinuous at the origin, but discontinuities elsewhere can prevent the correlation from being negligible. Though this invalidates the additive model, an asymptotic formula for the correlation is found in terms of the step size and the heights and positions of the discontinuities.**

For a finite support input density, such as uniform, it is shown that the support of the uniform quantizer can be matched to that of the density in ways that make the correlation approach a variety of limits.

The derivations in this paper are based on an analysis of the asymptotic convergence of cell centroids to cell midpoints. This convergence is fast enough that the centroids and midpoints induce the same asymptotic MSE, but not fast enough to induce the same correlations.

***Index Terms*—Asymptotic quantization, cell centroids, cell midpoints, high-resolution quantization, orthogonal error, orthogonal noise, quantization error, quantization noise, uncorrelated quantization error, uncorrelated quantization noise.**

I. Introduction

In his pioneering 1948 paper, Bennett [1] argued that the quantization error of a uniform scalar quantizer with small cells, reproduction levels at the cell midpoints, and large support region can be approximately modeled as being orthogonal to the quantizer input. That is, with X and Y denoting the quantizer input and output, respectively,

$$EX(Y-X) \approx 0\,. \qquad (1)$$

It follows that, as illustrated in Fig. 1(a), the quantizer output Y can be modeled as the sum of X plus orthogonal quantization

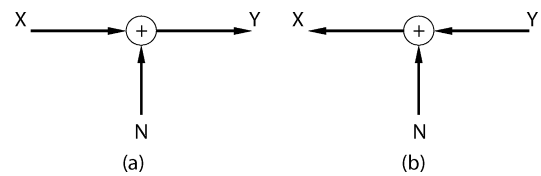

Fig. 1. Additive models of uniform scalar quantization. (a) The levels are midpoints and the quantization error is orthogonal to the input. (b) The levels are centroids and the quantization error is orthogonal to the output.

error $N = Y - X$. This is the *additive noise model*. Since $EY^2 = EX^2 + D + 2EX(Y-X)$, where $D = E(Y-X)^2$ is the mean-squared error (MSE), an equivalent property is

$$EY^2 \approx EX^2 + D \qquad (2)$$

i.e., the output power approximately equals the input power plus the MSE. Though the additive noise model is very widely used (cf. [2, pp. 193ff.], [3, pp. 753ff.], [4]), its validity has never been rigorously demonstrated. The principal goal of this paper is to do this and, in addition, to discover the correlation structure when the additive noise model is not valid.

It is easy to see that the left- and right-hand sides of (1), respectively, (2), tend to the same values as $\Delta \to 0$. This, however, is not sufficient to validate the additive noise model. Instead, we assert that the additive noise model is *asymptotically valid* when and only when

$$EX(Y-X) = o(D)$$

where $o(z)$ denotes a quantity such that $o(z)/z \to 0$ as $z \to 0$. Equivalently, it is asymptotically valid when and only when $EY^2 = EX^2 + D + o(D)$. In other words, the discrepancies in the approximations (1) and (2) must be asymptotically negligible relative to the MSE. Equivalently, using the well-known approximation $D = \frac{\Delta^2}{12} + o(\Delta^2)$, where Δ denotes the width of a quantization cell, the errors must be asymptotically negligible relative to Δ^2.

With this definition in mind, our principal result, Corollary 12, shows that the additive noise model is asymptotically valid when, in addition to satisfying several mild technical conditions, the probability density function (pdf) of X is continuous, except possibly for tending to infinity at the origin or having a finite jump discontinuity at the origin. If, on the other hand, there are finite jump discontinuities not at the origin, then Corollary 11 shows that

$$\frac{EX(Y-X)}{\Delta^2} = \frac{1}{12}\sum_{k=1}^{N} t_k e_k \big(1 - 6\alpha_\Delta(t_k)(1-\alpha_\Delta(t_k))\big) + o\,(1)$$

Manuscript received January 14, 2004; revised January 21, 2005. This work was supported by the National Science foundation under Grant ANI-0112801. The material in this paper was presented in part at the IEEE International Symposium on Information Theory, Yokohama, Japan, July 2003.

D. Marco was with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109 USA. He is now with the Department of Electrical Engineering, California Institute of Technology, Pasadena, CA 91125 USA (e-mail: idaniel@caltech.edu).

D. L. Neuhoff is with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109 USA (email: neuhoff@eecs.umich.edu).

Communicated by M. Effros, Associate Editor for Source Coding.

Digital Object Identifier 10.1109/TIT.2005.846397

where $t_1, \dots, t_N$ are the positions of the jumps in the pdf, $e_1, \dots, e_N$ are their heights, $\alpha_\Delta(t_k)$ is the fractional position of t_k within its quantization cell, and $o(1)$ denotes a quantity that approaches 0 as $\Delta \to 0$. It would be nice if the right-hand side of the above converged to some function of the t_k's and e_k's, with no dependence on the $\alpha_\Delta(t_k)$'s. In such a case, one could easily estimate the correlation, even in the presence of jumps. However, Theorem 13 shows that this is not possible. We conclude that when there are jumps in the pdf, the correlation structure depends intimately on the positions of such jumps within quantization cells.

To avoid overload issues, we focus on uniform quantizers with infinitely many levels, i.e., with infinite support. However, the results have significance for uniform quantizers with finitely many levels. Specifically, since the performance of a uniform quantizer with n levels approaches that of an infinite uniform quantizer as n tends to infinity, the results indicate conditions under which the additive noise model is asymptotically valid when Δ is sufficiently small and n is sufficiently large.

In deriving our results, we find it necessary to explore and exploit relations between quantization cell midpoints and centroids that yield insight into the behavior of uniform quantizers. It is well known that for a given Δ, MSE is minimized when centroids rather than midpoints are used as levels. Not surprisingly, as can be deduced from the results of [5, p. 15], the MSE with centroids is again well approximated by $\Delta^2/12$ when Δ is small. Thus, asymptotically, midpoints and centroids induce the same distortion. On the other hand, midpoints and centroids lead to rather different correlation structures. Specifically, with centroids, it is well known that for any Δ, the quantization error is exactly orthogonal to the quantizer output Y, rather than the quantizer input X, i.e.,

$$EY(Y-X) = 0 \tag{3}$$

or equivalently

$$EY^2 = EX^2 - D \tag{4}$$

i.e., the output power equals the input power minus the MSE. Thus, instead of the usual additive noise model, we have the additive model illustrated in Fig. 1(b). This is somewhat surprising in light of the fact that cell centroids approach cell midpoints as Δ decreases. (This intuitive fact is shown in [6].) Clearly, there is subtle behavior here. In this paper, we strengthen previous results on the convergence of cell centroids to midpoints, and we show that this convergence happens rapidly enough to account for the fact that the MSE with centroids is asymptotically the same as that with midpoints. However, it is not rapid enough to induce the same asymptotic correlation structure. We also note that the proofs of the principal theorems are based on a measure of the difference between the values assumed by EY^2 when centroids are used versus midpoints.

For completeness, we mention that Widrow [7], and Sripad and Snyder [8] found conditions, on Δ and the pdf, involving zeros of the characteristic function, under which the quantizer input and error are exactly orthogonal. Note, however, that these results are not asymptotic and that the conditions are rather restrictive. We also mention that Bennett's paper [1] argued that in addition to being orthogonal to the input, the quantization errors of a uniform scalar quantizer are, approximately, white. A rigorous demonstration of this was given in [9].

The remainder of the paper is organized as follows. Section II introduces infinite uniform scalar quantizers and the framework for considering such with step size Δ decreasing to zero, as well as notation and other essential background material. Section III shows that centroids approach midpoints rapidly enough to account for the fact that the MSE due to centroids is asymptotically the same as that due to midpoints. Section IV discusses the additive noise model and introduces a key functional $r(f)$ measuring the closeness of cell midpoints and centroids, whose value determines the validity of the additive noise model. Section V evaluates $r(f)$ and states the main results regarding the correlation of input and quantization error and the asymptotic validity of the additive noise model. Section VI discusses alternative noise models for uniform quantizers whose support is matched to that of a pdf with finite support. Section VII proves the principal results. Section VIII offers concluding remarks. Finally, the Appendix contains proofs of certain lemmas.

II. Background

An infinite level uniform scalar quantizer is characterized by a *step size* $\Delta > 0$, an *offset* θ, $0 \le \theta < 1$, and a set of *(reconstruction) levels*

$$\dots < y_{-2} < y_{-1} < y_0 < y_1 < y_2 < \cdots.$$

The *thresholds* of such a quantizer are

$$\dots < x_{-2} < x_{-1} < x_0 < x_1 < x_2 < \cdots$$

where $x_i = (i-\theta)\Delta$, and the ith (quantization) cell is $S_i = [x_i, x_{i+1})$. Note that θ is the fractional position of the origin within its quantization cell. Given an input x, the quantizer outputs $q(x) = y_i$ when $x \in S_i$. The quantization error is $q(x) - x$, and when the input is a random variable X with pdf f, the MSE is

$$D = E(q(X)-X)^2 = \int_{-\infty}^{\infty} (q(x)-x)^2 f(x)\,dx.$$

We focus on two choices for the levels: midpoints and centroids. In the former case

$$y_i = x_i + \Delta/2 = (i - \theta + 1/2)\Delta.$$

In the latter

$$y_i = E[X | x_i \le X < x_{i+1}] = \frac{\int_{x_i}^{x_{i+1}} x f(x)\,dx}{\int_{x_i}^{x_{i+1}} f(x)\,dx}$$

where f is the pdf of X.[1] Let $m_{\Delta,\theta}(x)$, $c_{\Delta,\theta}(x)$, and $u_{\Delta,\theta}(x)$ denote the midpoint, centroid, and left threshold, respectively, of the quantization cell in which x lies. These functions are constant on quantization cells. Let $M_{\Delta,\theta}$ and $C_{\Delta,\theta}$ denote the

[1]When a cell $[x_i, x_{i+1})$ has zero probability, the value of $E[X|x_i \le X < x_{i+1}]$ is of no consequence. However, for concreteness, we take it to be the midpoint of the cell.

output random variable $Y = q(X)$ when midpoints and centroids are used, respectively. In addition, let

$$m_{\Delta,u} = u + \Delta/2$$

and

$$c_{\Delta,u} = E[X | u \le X < u + \Delta]$$

denote, respectively, the midpoint and centroid of the interval $[u, u+\Delta)$.[2] For brevity, we usually omit the subscript θ from $m_{\Delta,\theta}(x)$, $c_{\Delta,\theta}(x)$, etc., when they are clear from context. When we wish to emphasize dependence on the pdf, we add a superscript, as in $c^f_{\Delta,\theta}(x)$.

In most of the results in later sections, we consider limiting characteristics of families of uniform quantizers in which the step size Δ goes to zero and the offset θ varies arbitrarily. That is, the offset θ is an arbitrary function of Δ, denoted $\theta(\Delta)$. It can be shown that if $g(\Delta,\theta)$ is a function and c is a constant such that $\lim_{\Delta\to 0} g(\Delta, \theta(\Delta)) = c$ for any function $\theta : \mathbb{R} \to [0,1)$, then the convergence is uniform over all such functions θ.

Throughout this paper, we focus on continuous-input random variables X with finite first and second moments, whose pdfs are either continuous or have finite jump discontinuities, or have points at which f goes to infinity from the left or right. (f is said to have a finite jump discontinuity at t if the following limits exist, and are finite and different: $f(t^-) \triangleq \lim_{x\nearrow t} f(x)$, $\lim_{x\searrow t} f(x) \triangleq f(t^+)$.) Other conditions on f will be specified as needed. It should be noted that for any result in this paper that is concerned with expected values, if f_1 and f_2 are pdfs such that $f_1 = f_2$ almost everywhere (a.e.) and f_2 satisfies the specified conditions for the result, then the result applies to f_1 as well. If a density f has finite support, then an infinite uniform quantizer has, effectively, finitely many levels.

We will occasionally introduce a symbol like f or g to represent a function that is like a pdf, but may lack the property of integrating to one. Accordingly, in all statements of results where f is required to be a pdf, we will explicitly specify such. Where there is no specification, f denotes an arbitrary nonnegative function.

Finally, a function f is said to be *piecewise differentiable* if there exists a countable collection of disjoint open intervals $\{B_i\}$ such that a) f is differentiable on each B_i, b) $\mathbb{R} = (\cup_i B_i) \cup E$, where E is the set of interval endpoints (not including $-\infty$ and ∞), and c) any finite interval contains at most a finite number of B_i's. We let $B \triangleq \cup_i B_i$.

III. Mean-Squared Error (MSE)

As mentioned earlier, when Δ is small, the MSE when using midpoints, denoted $D_{m,\Delta}$, is approximately $\Delta^2/12$. Linder and Zeger [10] showed rigorously that this holds for any pdf. The precise statement is

$$\lim_{\Delta\to 0} \frac{D_{m,\Delta}}{\Delta^2/12} = 1 \tag{5}$$

or, equivalently, $D_{m,\Delta} = \frac{\Delta^2}{12} + o(\Delta^2)$. Although the authors did not claim such, their proof is sufficient to show that (5) holds for any offset function $\theta(\Delta)$.

[2]When $\Pr(u \le X < u+\Delta) = 0$, we let $c_{\Delta,u} = u + \Delta/2$.

It is quite intuitive that M and C become closer as $\Delta \to 0$. The question is how quickly. The following two lemmas, whose proofs are left to the Appendix, provide some answers.

Lemma 1: If f is continuous and positive at x, then for any offset function

$$\lim_{\Delta\to 0} \frac{m_\Delta(x) - c_\Delta(x)}{\Delta} = 0$$

or, equivalently, $c_\Delta(x) = m_\Delta(x) + o(\Delta)$.

Lemma 2: If f is a continuous a.e. pdf, then for any offset function

$$\lim_{\Delta\to 0} \frac{E(M_\Delta - C_\Delta)^2}{\Delta^2} = 0$$

or, equivalently, $E(M_\Delta - C_\Delta)^2 = o(\Delta^2)$.

Remark: Notice that while quantities such as $m_\Delta(x)$, $c_\Delta(x)$, $u_\Delta(x)$, $M_\Delta(x)$, and $C_\Delta(x)$ depend on the offset function, limit expressions, such as in these two lemmas, usually do not. Whenever appropriate, such insensitivity to the offset function will be explicitly stated in future lemmas and theorems. In their proofs, an arbitrary fixed offset function will be assumed. However, it will not appear explicitly therein. Instead, its influence on $m_\Delta(x)$, $c_\Delta(x)$, etc., is implicit.

It is well known that centroids minimize MSE. The following theorem uses the convergence of centroids to midpoints demonstrated in Lemma 2 to show that the MSE induced by centroids, denoted $D_{c,\Delta}$, is asymptotically the same as that induced by midpoints. This result can also be deduced from the results of [5, p. 15] without reference to the closeness of midpoints and centroids and without requiring the pdf to be continuous a.e.

Theorem 3: If f is a continuous a.e. pdf, then for any offset function

$$D_{c,\Delta} = D_{m,\Delta} + o(\Delta^2) = \frac{\Delta^2}{12} + o(\Delta^2). \tag{6}$$

Proof:

$$\begin{aligned} D_{m,\Delta} &= E[(X - C_\Delta) + (C_\Delta - M_\Delta)]^2 \\ &= E(X - C_\Delta)^2 + 2E(X - C_\Delta)(C_\Delta - M_\Delta) + E(C_\Delta - M_\Delta)^2 \\ &= D_{c,\Delta} + E(C_\Delta - M_\Delta)^2 = D_{c,\Delta} + o(\Delta^2) \end{aligned}$$

where the first equality is by definition of $D_{m,\Delta}$, the third is by the orthogonality principle, and the last is due to Lemma 2. The second equality in (6) is from (5). □

IV. Additive Noise Model

Our primary goal is to determine when the additive noise model is asymptotically valid for uniform scalar quantizers with infinitely many levels located at cell midpoints. With M denoting the quantizer output with midpoint levels, we consider the additive noise model to be asymptotically valid when and only when for any offset function

$$EX(M - X) = o(\Delta^2) \tag{7}$$

or equivalently

$$EM^2 = EX^2 + D_m + o(\Delta^2). \tag{8}$$

We focus on the latter condition. To determine when it holds, we write

$$\begin{aligned}EM^2 &= EM^2 + EX^2 - EC^2 - D_c \\ &= EX^2 - D_m + (EM^2 - EC^2) + o(\Delta^2)\end{aligned}$$

where C and D_c denote the output and MSE, respectively, of a quantizer with centroid levels, and where the first equality follows from (4) applied to a quantizer with centroid levels and the second follows from Theorem 3, assuming the pdf is continuous a.e. It is now clear that the relationship between EM^2, EX^2, and D_m depends on the quantity $EM^2 - EC^2$. This motivates us to define a functional r that captures the behavior of this quantity.

Definition 4: Given a pdf f

$$r_\Delta(f) \triangleq \frac{EM^2 - EC^2}{\Delta^2/6} = \int_{-\infty}^{\infty} G_\Delta(x)\,dx \tag{9}$$

where

$$G_\Delta(x) \triangleq \frac{m_\Delta^2(x) - c_\Delta^2(x)}{\Delta^2/6} f(x).$$

When the limit of $r_\Delta(f)$ exists and is the same for all offset functions

$$r(f) \triangleq \lim_{\Delta\to 0} r_\Delta(f). \tag{10}$$

Using the above definition, we obtain the following lemma.

Lemma 5: If f is a continuous a.e. pdf, then for any offset function

$$EM^2 = EX^2 + (2r_\Delta(f) - 1)\frac{\Delta^2}{12} + o(\Delta^2). \tag{11}$$

Proof:

$$\begin{aligned}\frac{EM^2 - EX^2}{\Delta^2/12} &= \frac{EM^2 - EC^2}{\Delta^2/12} - \frac{EX^2 - EC^2}{\Delta^2/12} \\ &= 2r_\Delta(f) - \frac{D_c}{\Delta^2/12} = 2r_\Delta(f) - 1 + \frac{o(\Delta^2)}{\Delta^2}\end{aligned}$$

where the second equality uses (4) and the definition of $r_\Delta(f)$, and the last uses Theorem 3. □

We now consider the ramifications of this lemma.

Corollary 6: If the input pdf f is continuous a.e., then for any offset function

$$EX(M - X) = \frac{\Delta^2}{12}(r_\Delta(f) - 1) + o(\Delta^2) \tag{12}$$

and

$$EM(M - X) = \frac{\Delta^2}{12} r_\Delta(f) + o(\Delta^2)\,. \tag{13}$$

Proof: The first relation is

$$\begin{aligned}EX(M - X) &= \frac{1}{2}\left[EM^2 - EX^2 - D_m\right] \\ &= \frac{1}{2}\left[(2r_\Delta(f) - 1)\frac{\Delta^2}{12} + o(\Delta^2) - D_m\right] \\ &= (r_\Delta(f) - 1)\frac{\Delta^2}{12} + o(\Delta^2)\end{aligned}$$

where the first equality is elementary, the second is from Lemma 5, and the third is from (5).

The second relation is

$$\begin{aligned}EM(M - X) &= EM^2 - EX^2 - EX(M - X) \\ &= (2r_\Delta(f) - 1)\frac{\Delta^2}{12} + o(\Delta^2) \\ &\quad - (r_\Delta(f) - 1)\frac{\Delta^2}{12} + o(\Delta^2) \\ &= r_\Delta(f)\frac{\Delta^2}{12} + o(\Delta^2)\end{aligned}$$

where the first equality is elementary, the second is from Lemma 5 and the first part of this proof, and third is from (5). □

By comparing (11) to (8) and using (5), or equivalently comparing (12) to (7), we obtain the following.

Theorem 7: If midpoints are used and the input pdf is continuous a.e., then the additive noise model is asymptotically valid if and only if $r(f)$ exists and equals one.

V. Evaluating $r(f)$

In this section, we give the main results of the paper, which characterize the behavior of $r(f)$, and consequently, determine the validity or invalidity of the additive noise model. We begin with a definition. Proofs are given in Section VII.

Definition 8: A pdf f is *nice* if each of the following holds.

1. f has finite second moment.
2. $\lim_{x\to 0} x f(x) = 0$.
3. There exists $\varepsilon > 0$ such that
$$\lim_{\substack{x\to-\infty \\ x\in B}} |x|^{2+\varepsilon} f'(x) = 0 \quad \text{and} \quad \lim_{\substack{x\to\infty \\ x\in B}} x^{2+\varepsilon} f'(x) = 0$$
where f' is the derivative of f and B is the set over which f is differentiable.
4. f is continuous, bounded, and piecewise differentiable with bounded derivative, except perhaps at a finite set of *exceptional points* $\{s_1, \ldots, s_n\}$ such that any of the following might hold:
   a) $|f'(x)| \to \infty$ as $x \to s_i$ from left and/or right;
   b) f has a finite jump discontinuity at s_i;
   c) $s_i = 0$ and $f(x) \to \infty$ as $x \to 0$ from left and/or right;

   and if at any s_i, $|f'|$ goes to infinity from left (right), it does so monotonically in some left (right) neighborhood of s_i.

Remark: The class of densities of the form $b|x|^\beta e^{-a|x|^\alpha}$, $\beta > -1$, and $\alpha > 0$, which includes the Gaussian, Laplacian, and gamma densities, and one-sided versions of these, such as the Rayleigh and exponential densities, are nice pdfs.

Theorem 9: If f is a nice pdf with no exceptional points, then $r(f) = 1$.

Theorem 10: If f is a nice pdf, and $T = \{t_1, \ldots, t_N\}$ is the set of exceptional points where f has discontinuities, then for any offset function

$$r_\Delta(f) = s_\Delta(f) + o(1) \tag{14}$$

where

$$s_\Delta(f) \triangleq 1 + \sum_{k=1}^{N} t_k e_k \big[1 - 6\alpha_\Delta(t_k)(1-\alpha_\Delta(t_k))\big]. \qquad (15)$$

$e_k \triangleq f(t_k^+) - f(t_k^-)$ is the height of the discontinuity at t_k, and

$$\alpha_\Delta(t_k) \triangleq \frac{t_k - u_\Delta(t_k)}{\Delta}$$

is the fractional position of t_k within its cell, and where the summand is taken to be zero when $t_k = 0$, even if $f(t_k^+)$ and/or $f(t_k^-)$ are infinite.

From the preceding theorem, Lemma 5, and Corollary 6, we obtain the following corollaries.

Corollary 11: If f is a nice pdf, then for any offset function

$$EX(M-X) = \frac{\Delta^2}{12}(s_\Delta(f) - 1) + o(\Delta^2)$$
$$EM(M-X) = \frac{\Delta^2}{12} s_\Delta(f) + o(\Delta^2)$$
$$EM^2 = EX^2 + \frac{\Delta^2}{12}(2s_\Delta(f) - 1) + o(\Delta^2).$$

Corollary 12: If f is a nice pdf with no discontinuities, except perhaps at 0, then $s_\Delta(f) = 1$ for all Δ, and consequently, $r(f) = 1$ and the additive noise model is asymptotically valid.

On the one hand, when f is a nice pdf with no discontinuities except possibly at the origin, Corollary 12 shows that the additive noise model is asymptotically valid, i.e., for small values of Δ. On the other hand, when there are discontinuities elsewhere, Corollary 11 permits one to determine the validity of the additive noise model for any given Δ by computing $s_\Delta(f)$. It is conceivable that $s_\Delta(f)$ converges to some value depending on the t_k's and e_k's, but not on the offset function $\theta(\Delta)$. In such a case, for small values of Δ, it would be sufficient to know this value, so one would not have to be concerned about the detailed calculation of $s_\Delta(f)$ for the specific values of Δ and $\theta(\Delta)$ being used. Unfortunately, the following theorem shows that this is not possible.

Theorem 13: If f is a nice pdf, and the set of exceptional points $T = \{t_1, \ldots, t_N\}$ where there are discontinuities is not comprised of only the single point 0, then there exists an offset function $\theta(\Delta)$ such that $\lim_{\Delta\to 0} s_\Delta(f)$ does not exist. Thus, $r(f) = \lim_{\Delta \to 0} r_\Delta(f)$ does not exist, and, consequently, the additive noise model is not asymptotically valid.

The effect of jump discontinuities: In light of Corollary 12 and Theorem 13, we observe that jump discontinuities have a determining effect on the correlation between quantizer input and error and the existence of $r(f)$. To see why, consider the case that f has a single finite jump discontinuity at t, and rewrite $r_\Delta(f)$ as

$$r_\Delta(f) = \int_{-\infty}^{u_\Delta(t)} G_\Delta(x)\,dx + \int_{u_\Delta(t)}^{u_\Delta(t)+\Delta} G_\Delta(x)\,dx + \int_{u_\Delta(t)+\Delta}^{\infty} G_\Delta(x)\,dx\,.$$

The methods used in the proof of Theorem 10 can be easily used to show that the left and right terms in the above converge to finite values. Therefore, the existence of $r(f)$ is determined by whether or not the middle term has a limit. We now rewrite the middle term in greater detail as

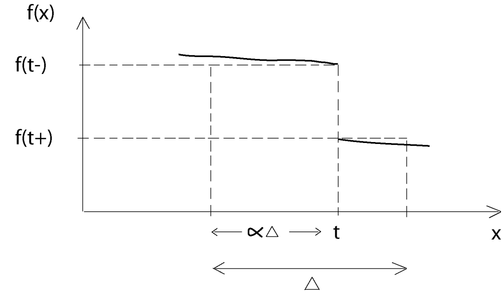

Fig. 2. The pdf f, having a jump discontinuity at $x = t$, can be viewed as being approximately constant on the left and right parts of the cell containing t.

$$\int_{u_\Delta(t)}^{u_\Delta(t)+\Delta} G_\Delta(x)\,dx = 6\Big[m_\Delta(t) + c_\Delta(t)\Big]\Big[\frac{1}{\Delta}\int_{u_\Delta(t)}^{u_\Delta(t)+\Delta} f(x)\,dx\Big]\Big[\frac{m_\Delta(t) - c_\Delta(t)}{\Delta}\Big] \qquad (16)$$

where we used the fact that $m_\Delta(x)$ and $c_\Delta(x)$ are constant on quantization cells.

On the one hand, if f were continuous at t, then the right-hand side of the above would tend to zero. Specifically, the first term in brackets approaches $2t$ and the second approaches $f(t)$. If $f(t) > 0$, then Lemma 1 implies that the third term goes to zero. If $f(t) = 0$, then the second term approaches 0 while the third has magnitude no larger than $1/2$.

On the other hand, when f has a finite jump discontinuity at $t \neq 0$, as illustrated in Fig. 2, $\frac{m_\Delta(t) - c_\Delta(t)}{\Delta}$ no longer goes to zero, necessarily, as $\Delta \to 0$, as we will shortly demonstrate. In fact, it can be made to converge to different values depending on how Δ approaches zero. Furthermore, $\frac{1}{\Delta}\int_{u_\Delta(t)}^{u_\Delta(t)+\Delta} f(x)\,dx$ also does not converge. The important question is whether the product of these two terms converges. We will show that it does not. Therefore, $r(f)$ does not exist.

For example, fix $\theta(\Delta) = 0$ for all Δ and suppose Δ_n is a sequence going to zero as $n \to \infty$ in such a way that t always lies in the center of its cell; i.e., $t = u_{\Delta_n}(t) + \alpha\Delta_n$ for all n, where $\alpha = 1/2$. Assume further that f is constant in neighborhoods to the right and left of t, as will approximately be the case when Δ_n is small. Then the first term in (16) converges to $2t$, the second term converges to $(f(t^-) + f(t^+))/2$, and the third term can be straightforwardly shown to converge to $\frac{1}{4}\frac{f(t^-) - f(t^+)}{f(t^-) + f(t^+)}$. It follows that the right-hand side of (16) converges to $\frac{3}{2}t[f(t^-) - f(t^+)]$. (A careful derivation, not assuming f is constant in neighborhoods, is given later in the derivation of (26).) On the other hand, if $\Delta'_n \to 0$ in such a way that $t = u_{\Delta'_n}(t) + \alpha\Delta'_n$ for all n, with $\alpha \neq 1/2$, then the right-hand side of (16) converges to some other value. This implies that $r(f)$ does not exist.

We comment that although the contribution of the cell containing t is nonvanishing, this fact alone is insufficient to invalidate the additive noise model, since it is conceivable that this substantial nonvanishing contribution might combine with the *sum* of vanishing contributions of all other cells (where f is continuous), which is substantial as well, so as to make $r_\Delta(f)$ converge to 1. However, as mentioned, the fact that the contribution of the cell containing t does not converge, while the sum of contributions of all other cells does converge, ultimately causes $r_\Delta(f)$ not to exist. Finally, one might imagine that if there were several jump discontinuities, then their nonconverging contributions might perhaps cancel each other so that $r_\Delta(f)$ would still converge. This, however, cannot happen, as shown by Theorem 13.

VI. Uniform Densities and Quantizers With Matched Support

Consider a uniform pdf. It has discontinuities at each end of its support. Thus, according to Theorem 13, the additive noise model is not asymptotically valid. While the discontinuities cause $r(f)$ not to exist, i.e., there are offset functions for which neither $s_\Delta(f)$ nor $r_\Delta(f)$ have limits, the simple fact that the midpoints are centroids (ignoring the cells containing the endpoints of the support of the pdf, for which midpoints might not equal centroids) would already lead one to suspect that $r(f)$ does not equal 1. Instead, in view of (3), one would more likely expect the quantization error to be approximately orthogonal to the output rather than the input, and from Corollaries 6 and 11, one would expect $r_\Delta(f) \approx s_\Delta(f) \approx 0$. Indeed, Theorem 10 shows this will be true if the endpoints of the pdf support are close to the thresholds of the quantizer, in which case the α's for the endpoints will be nearly zero or one, and the $t_k e_k$'s will sum to -1.

In view of the above discussion, it is interesting to consider the limiting characteristics of uniform quantizers in an alternative framework. Specifically, if a pdf has finite support (a,b), consider the sequence of uniform quantizers such that the nth quantizer partitions (a,b) into n cells of width $\Delta_n = (b-a)/n$, with thresholds exactly at a and b.[3] Such quantizers are said to have *matched support*.

Though we do not expect the usual additive noise model to be valid for quantizers with matched support, there can nevertheless be a well-defined asymptotic correlation structure, i.e., asymptotic formulas for the second moment of the output, and the correlations between input, output, and quantization error. These are characterized by modified versions of $r_\Delta(f)$ and $r(f)$, namely

$$\tilde{r}_n(f) \triangleq \int_{-\infty}^{\infty} G_{(b-a)/n}(x)\,dx$$

and when $\tilde{r}_n(f)$ has a limit

$$\tilde{r}(f) \triangleq \lim_{n\to\infty} \tilde{r}_n(f)$$

which are just like $r_\Delta(f)$ and $r(f)$ except we now require that $\Delta_n = (b-a)/n$ and that there be thresholds at a and b. With $\tilde{r}(f)$ replacing $r(f)$, one may easily check that Lemma 5 and Corollary 6 apply for pdfs with finite support and uniform quantizers with matched support. Moreover, slightly modified versions of Theorems 9, 10 and Corollaries 11, 12 hold. The following is an example of what is possible.

[3]Note that the offsets of the quantizers change with n.

Theorem 14: Let f be a nice pdf with finite support (a,b) with no discontinuities except, possibly, jump discontinuities at a, b, and the origin. Then for uniform quantizers with matched support

$$\begin{aligned} EM^2 &= EX^2 + (2\tilde{r}_n(f)-1)D_{m,\Delta} + o(\Delta_n^2) \\ EX(M-X) &= \frac{\Delta_n^2}{12}(\tilde{r}_n(f)-1) + o(\Delta_n^2) \\ EM(M-X) &= \frac{\Delta_n^2}{12}\tilde{r}_n(f) + o(\Delta_n^2) \\ \tilde{r}(f) &= 1 + af(a^+) - bf(b^-)\,. \end{aligned} \tag{17}$$

Proof: The first three relations are derived just as in Lemma 5 and Corollary 6. The last relation follows by deriving a modified version of Theorem 10, and then using the facts that $\alpha_{\Delta_n}(a) = \alpha_{\Delta_n}(b) = 0$ and that the corresponding $t_i e_i$ terms sum to $af(a^+) - bf(b^-)$. □

This theorem shows that for matched quantizers, a variety of different correlations are possible, i.e., a variety of alternative noise models are possible. For a uniform source, $\tilde{r}(f) = 0$ and the theorem predicts the additive noise model illustrated in Fig. 1(b). However, for nonuniform pdfs, (17) indicates that appropriate choices of a, b, $f(a^+)$, and $f(b^-)$ can make $\tilde{r}(f)$ attain any value whatsoever, making possible a broad range of alternative noise models.

When the pdf has jump discontinuities within (a,b) that are not at the origin, $\tilde{r}$ will not exist, for reasons like those that cause r not to exist in Theorem 13. For such cases, one may develop a generalization of Theorem 10. Or one may try a more complicated matching such that all jump discontinuities occur at quantizer thresholds. This, however, is not always possible.

As a final set of options, we mention that one could also consider the family of uniform quantizers whose supports are *matched* to (a,b) in the sense of having cell midpoints at a and b, rather than boundaries at a and b. In this case, for a uniform pdf, Theorem 10 shows that the modified version of $r(f)$, again denoted $\tilde{r}(f)$, would equal $3/2$. More generally, for a uniform pdf on (a,b) and any $\alpha_1, \alpha_2 \in [0,1)$, one could consider the family of uniform quantizers that are matched in the sense that

$$a = u_{\Delta_n}(a) + \alpha_1 \Delta_n$$

and

$$b = u_{\Delta_n}(b) + \alpha_2 \Delta_n.$$

In this case, Theorem 10 implies that

$$\tilde{r}(f) = \frac{6}{b-a}\big(b\alpha_2(1-\alpha_2) - a\alpha_1(1-\alpha_1)\big).$$

Thus, one could obtain a wide range of modified $r(f)$ values. One might even attempt to obtain $\tilde{r}(f) = 1$, so that the additive noise model would be valid.

VII. Proofs

Proof of Theorem 9: Let f be nice with no exceptional points. Let $\theta(\Delta)$ be an arbitrary offset function, let (a,b) be some finite interval, and let us write

$$r_\Delta(f) = \int_{-\infty}^{a} G_\Delta(x)\,dx + \int_{a}^{b} G_\Delta(x)\,dx + \int_{b}^{\infty} G_\Delta(x)\,dx\,. \tag{18}$$

The proof follows by taking the limit of the above, while using Facts 2 and 3 below.

Fact 1: For all $x \in B$, $\lim_{\Delta\to 0} G_\Delta(x)$ exists and equals $-xf'(x)$.

Fact 2:

a)

$$\lim_{\Delta\to 0}\int_a^b G_\Delta(x)\,dx = \int_a^b \lim_{\Delta\to 0} G_\Delta(x)\,dx.$$

b)

$$\int_a^b \lim_{\Delta\to 0} G_\Delta(x)\,dx = af(a) + \int_a^b f(x)\,dx - bf(b).$$

Fact 3:

$$\lim_{\Delta\to 0}\int_{-\infty}^a G_\Delta(x)\,dx = \int_{-\infty}^a f(x)\,dx - af(a)$$

and

$$\lim_{\Delta\to 0}\int_b^\infty G_\Delta(x)\,dx = bf(b) + \int_b^\infty f(x)\,dx.$$

Proof of Fact 1: $\lim_{\Delta\to 0} G_\Delta(x) = -xf'(x)$, *for* $x \in B$: We will use the following lemma, proved in the Appendix, which provides a stronger statement than that of Lemma 1, but requires stronger conditions. A similar result was shown in [6]. However, the conditions set here are less restrictive and the statement of this lemma is more precise.

Lemma 15: If f is positive and differentiable at x, then for any offset function

$$c_\Delta(x) = m_\Delta(x) + \frac{\Delta^2}{12}\frac{f'(x)}{f(x)} + o(\Delta^2)$$

or equivalently

$$\lim_{\Delta\to 0}\frac{m_\Delta(x) - c_\Delta(x)}{\Delta^2} = -\frac{f'(x)}{12 f(x)}.$$

To prove Fact 1, we begin by considering some $x \in B$, and expanding $G_\Delta(x)$

$$\begin{aligned} G_\Delta(x) &= \frac{m_\Delta^2(x) - c_\Delta^2(x)}{\Delta^2/6} f(x) \\ &= 6\left[\frac{m_\Delta(x) - c_\Delta(x)}{\Delta^2}\right][m_\Delta(x) + c_\Delta(x)]\,f(x)\,. \end{aligned}$$

If $f(x) > 0$, then Lemma 15 shows that the first bracketed term converges to $-\frac{f'(x)}{12f(x)}$ as $\Delta \to 0$. The second bracketed term goes to $2x$ as $\Delta \to 0$. Therefore,

$$\lim_{\Delta\to 0} G_\Delta(x) = -xf'(x)\,. \tag{19}$$

If $f(x) = 0$, then $G_\Delta(x) = 0$ for any Δ. Also, $f'(x) = 0$ (otherwise, we could move in the direction that would make f negative). Thus, (19) holds again, which completes the proof of Fact 1.

Proof of Fact 2a:

$$\lim_{\Delta\to 0}\int_a^b G_\Delta(x)\,dx = \int_a^b \lim_{\Delta\to 0} G_\Delta(x)\,dx.$$

We will use the bounded convergence theorem [11, p. 210] to show that the limit and integral can be swapped. Fact 1 showed that the limit of the integrand $G_\Delta(x)$ exists almost everywhere. The bounded convergence theorem also requires that $|G_\Delta(x)|$ be uniformly bounded, which we now show. Since f is nice with no exceptional points, there exists $S < \infty$ such that $|f'| \le S$ wherever f' exists. For any $x \in (a,b)$, Lemma A2 of the Appendix shows that for all sufficiently small Δ

$$|G_\Delta(x)| \le 12(|x| + \Delta)S < 24\max\{|a|,|b|\}S$$

where we used the fact that for all sufficiently small Δ, $(|x| + \Delta) < 2\max\{|a|,|b|\}$. It follows that $|G_\Delta(x)|$ is uniformly bounded for $x \in (a,b)$. Finally, since the integration is over a set of finite measure, Fact 2 follows from the bounded convergence theorem.[4]

Proof of Fact 2b:

$$\int_a^b \lim_{\Delta\to 0} G_\Delta(x)\,dx = af(a) + \int_a^b f(x)\,dx - bf(b).$$

Fact 1 implies

$$\int_a^b \lim_{\Delta\to 0} G_\Delta(x)\,dx = \int_a^b -xf'(x)\,dx.$$

Since f is piecewise differentiable, and B is the union of disjoint open intervals on each of which f' exists, there exists some K such that $B \cap (a,b) = \cup_{i=1}^K (y_i, y_{i+1})$, where $y_1 = a$ and $y_{K+1} = b$. Applying integration by parts to each open interval (y_i, y_{i+1}), we obtain

$$\begin{aligned} \int_a^b \lim_{\Delta\to 0} G_\Delta(x)dx &= \sum_{i=1}^K \int_{y_i}^{y_{i+1}} -xf'(x)\,dx \\ &= \sum_{i=1}^K \left(y_i f(y_i) + \int_{y_i}^{y_{i+1}} f(x)dx - y_{i+1} f(y_{i+1})\right) \\ &= af(a) + \int_a^b f(x)\,dx - bf(b)\,. \end{aligned} \tag{20}$$

Proof of Fact 3:

$$\lim_{\Delta\to 0}\int_{-\infty}^a G_\Delta(x)\,dx = \int_{-\infty}^a f(x)\,dx - af(a)$$

and

$$\lim_{\Delta\to 0}\int_b^\infty G_\Delta(x)\,dx = bf(b) + \int_b^\infty f(x)\,dx.$$

We will show

$$\lim_{\Delta\to 0}\int_b^\infty G_\Delta(x)\,dx = bf(b) + \int_b^\infty f(x)\,dx.$$

The result for the other integral follows in a similar way. We decompose the integral $\lim_{\Delta\to 0}\int_b^\infty G_\Delta(x)\,dx$ into

$$\lim_{\Delta\to 0}\sum_{k=0}^\infty \int_{b_k}^{b_{k+1}} G_\Delta(x)\,dx$$

where $b_k \triangleq b + k$. Our main goal is to show that the limit and sum can be swapped. To do so, we shall make use of the following version of the Weierstrass M-test [11, p. 543].

[4]It can be easily shown that the theorem applies when the integrand is parameterized by some t converging continuously to some t_o, rather than some integer n converging to ∞.

Lemma 16: Let $\Phi_k(\Delta)$, $k \in \mathbb{Z}$ be a sequence of functions such that $\lim_{\Delta\to 0}\Phi_k(\Delta)$ exists, $|\Phi_k(\Delta)| \le M_k$ for $0<\Delta<\delta$, for some $\delta>0$, and $\sum_{k=-\infty}^{\infty} M_k < \infty$. Then $\sum_{k=-\infty}^{\infty}\Phi_k(\Delta)$ exists for $0<\Delta<\delta$, and

$$\lim_{\Delta\to 0}\sum_{k=-\infty}^{\infty}\Phi_k(\Delta)=\sum_{k=-\infty}^{\infty}\lim_{\Delta\to 0}\Phi_k(\Delta)\,.$$

Define $\Phi_k(\Delta)=\int_{b_k}^{b_{k+1}} G_\Delta(x)\,dx$ and write

$$\int_b^\infty G_\Delta(x)\,dx=\sum_{k=0}^{\infty}\int_{b_k}^{b_{k+1}} G_\Delta(x)\,dx=\sum_{k=0}^{\infty}\Phi_k(\Delta)\,. \quad (21)$$

We would like to apply Lemma 16 to the right-hand term of (21). By Fact 2, $\lim_{\Delta\to 0}\Phi_k(\Delta)$ exists. We now find a sequence M_k, whose sum is finite, that dominates the sequence $|\Phi_k(\Delta)|$. We begin by bounding $|\Phi_k(\Delta)|$. Let $S<\infty$ be the uniform bound on the derivative of f. Fix $\delta, 0<\delta<1$ and consider throughout $0<\Delta<\delta$. Recalling that f is piecewise differentiable, let

$$W_k \triangleq B\cap(b_k-\delta, b_{k+1}+\delta)$$

denote the subset of the interval $(b_k-\delta, b_{k+1}+\delta)$ over which f is differentiable. Let $S_k \triangleq \sup_{x\in W_k}|f'(x)|$. Since f is a nice pdf

$$\lim_{\substack{x\to\infty\\ x\in B}} x^{2+\varepsilon} f'(x)=0$$

for some $\varepsilon>0$. Thus, there exists a nonnegative integer N, $N>4-b$ (i.e., $N\ge 0$ and $b_N>4$) such that $x^{2+\varepsilon}|f'(x)|<1$, or, equivalently, $|f'(x)|<\frac{1}{x^{2+\varepsilon}}$ for all $x\in[b_N-1,\infty)\cap B$. Therefore, when $k\ge N$,

$$S_k \le \frac{1}{(b_k-\delta)^{2+\varepsilon}} < \frac{1}{(b_k-1)^{2+\varepsilon}}.$$

Using this we obtain

$$\begin{aligned}|\Phi_k(\Delta)| &\le \int_{b_k}^{b_{k+1}} |G_\Delta(x)|\,dx \overset{(a)}{\le} 12S_k(|b_k|+1+\delta)\\ &\overset{(b)}{<} 12S_k(|b_k|+2)\\ &\overset{(c)}{<} \begin{cases} 12S(|b|+N+2), & 0\le k<N\\ 24\frac{1}{(b_k-1)^{1+\varepsilon}}, & k\ge N\end{cases} \triangleq M_k\end{aligned}$$

where (a) follows from Lemma A2, (b) uses $\delta<1$, and (c) is due to having $|b_k|<|b|+N$ for $0\le k<N$, and $|b_k|+2<2(b_k-1)$ for $k\ge N$.

Next, we need to show that $\sum_{k=0}^{\infty} M_k<\infty$, which can be seen as follows:

$$\begin{aligned}\sum_{k=0}^{\infty} M_k &= \sum_{k=0}^{N-1} 12S(|b|+N+2)+\sum_{k=N}^{\infty}\frac{24}{(b_k-1)^{1+\varepsilon}}\\ &< 12S(|b|+N+2)N+24\sum_{k=3}^{\infty}\frac{1}{k^{1+\varepsilon}}<\infty\end{aligned}$$

where the first inequality is due to $b_N-1>3$. Thus, taking the limit as $\Delta\to 0$ in (21) we obtain

$$\begin{aligned}\lim_{\Delta\to 0}&\int_b^\infty G_\Delta(x)\,dx\\ &=\lim_{\Delta\to 0}\sum_{k=0}^{\infty}\Phi_k(\Delta)\overset{(a)}{=}\sum_{k=0}^{\infty}\lim_{\Delta\to 0}\Phi_k(\Delta)\\ &=\sum_{k=0}^{\infty}\lim_{\Delta\to 0}\int_{b_k}^{b_{k+1}} G_\Delta(x)\,dx\\ &\overset{(b)}{=}\sum_{k=0}^{\infty}\left(b_k f(b_k)+\int_{b_k}^{b_{k+1}} f(x)\,dx-b_{k+1}f(b_{k+1})\right)\\ &=\lim_{N\to\infty}\sum_{k=0}^{N}\left(b_k f(b_k)+\int_{b_k}^{b_{k+1}} f(x)\,dx-b_{k+1}f(b_{k+1})\right)\\ &=bf(b)+\lim_{N\to\infty}\int_b^{b_{N+1}} f(x)\,dx-\lim_{N\to\infty} b_{N+1}f(b_{N+1})\\ &\overset{(c)}{=}bf(b)+\int_b^\infty f(x)\,dx\end{aligned}$$

where (a) follows from Lemma 16, (b) follows from applying Fact 2 to intervals of the form (b_k, b_{k+1}), and (c) is obtained by applying Lemma A5 of the Appendix, which shows $\lim_{x\to\infty} xf(x)=0$, since f is a pdf with finite mean and $\lim_{x\to\infty, x\in B} xf'(x)=0$. □

Proof of Theorem 10: Let $\{s_1,\dots,s_n\}$ be the exceptional points of f, let $\{v_0, v_1,\dots,v_n\}$ be chosen so that

$$-\infty<v_0<s_1<v_1<s_2<\cdots<v_{n-1}<s_n<v_n<\infty$$

and let us write

$$\int_{-\infty}^{\infty} G_\Delta(x)\,dx=\int_{-\infty}^{v_0} G_\Delta(x)\,dx+\sum_{i=1}^{n}\int_{v_{i-1}}^{v_i} G_\Delta(x)\,dx+\int_{v_n}^{\infty} G_\Delta(x)\,dx\,. \quad (22)$$

It follows from Fact 3 in the proof of Theorem 9, which applies even if f has exceptional points, that the first and third integrals on the right-hand side of (22) converge to $\int_{-\infty}^{v_0} f(x)\,dx - v_0 f(v_0)$ and $v_n f(v_n)+\int_{v_n}^{\infty} f(x)\,dx$, respectively, as $\Delta\to 0$. We further decompose each integral in the sum term above as follows:

$$\int_{v_{i-1}}^{v_i} G_\Delta(x)\,dx=\int_{v_{i-1}}^{u_\Delta(s_i)-2\Delta} G_\Delta(x)\,dx+\int_{u_\Delta(s_i)-2\Delta}^{u_\Delta(s_i)+3\Delta} G_\Delta(x)\,dx+\int_{u_\Delta(s_i)+3\Delta}^{v_i} G_\Delta(x)\,dx\,. \quad (23)$$

Since the treatment of the first and last terms above is similar, we will only consider the first. With the above decomposition in mind, the proof will derive from the following two facts.

Fact 1:

a)

$$\lim_{\Delta\to 0}\int_{v_{i-1}}^{u_\Delta(s_i)-2\Delta} G_\Delta(x)\,dx=\int_{v_{i-1}}^{s_i}\lim_{\Delta\to 0} G_\Delta(x)\,dx$$

b)
$$\int_{v_{i-1}}^{s_i} \lim_{\Delta\to 0} G_\Delta(x)dx = v_{i-1}f(v_{i-1}) + \int_{v_{i-1}}^{s_i} f(x)dx - s_i f(s_i^-).$$

Fact 2:

a)
$$\int_{u_\Delta(s_i)-2\Delta}^{u_\Delta(s_i)+3\Delta} G_\Delta(x)\,dx = o(1)$$
when f is continuous at s_i or when $s_i = 0$.

b)
$$\int_{u_\Delta(s_i)-2\Delta}^{u_\Delta(s_i)+3\Delta} G_\Delta(x)\,dx = 6s_i[f(s_i^-) - f(s_i^+)]\alpha_\Delta(s_i)[1-\alpha_\Delta(s_i)] + o(1)$$
when f has a finite jump discontinuity at s_i and $s_i \neq 0$.

Combining (22), the discussion right after it, (23), and the above two facts, we get the equation at the bottom of the page. Therefore, recalling that $T = \{t_1, \ldots, t_N\}$ is the set of exceptional points where there are discontinuities in f, we may rewrite the equation at the bottom of the page as

$$r_\Delta(f) \triangleq \int_{-\infty}^{\infty} G_\Delta(x)\,dx = 1 + \sum_{k=1}^{N} t_k(f(t_k^+) - f(t_k^-))\left[1 - 6\alpha_\Delta(t_k)(1-\alpha_\Delta(t_k))\right] + o(1)$$

which will conclude the proof of the theorem. We now prove the two facts.

Proof of Fact 1a:
$$\lim_{\Delta\to 0}\int_{v_{i-1}}^{u_\Delta(s_i)-2\Delta} G_\Delta(x)\,dx = \int_{v_{i-1}}^{s_i}\lim_{\Delta\to 0} G_\Delta(x)\,dx$$
To simplify notation, we write v for v_{i-1} and s instead of s_i. We begin with
$$\int_{v}^{u_\Delta(s)-2\Delta} G_\Delta(x)\,dx = \int_v^s G_\Delta(x) I_{(v,u_\Delta(s)-2\Delta)}(x)\,dx$$
where $I_F(x)$ denotes the indicator function of the event F. Observe that for any $x \in (v,s)$
$$\lim_{\Delta\to 0} G_\Delta(x) I_{(v,u_\Delta(s)-2\Delta)}(x) = \lim_{\Delta\to 0} G_\Delta(x).$$
We will use the bounded and dominated convergence theorems to show that when taking the limit of the right-hand side, the integral can be swapped with the limit.

There are four cases to consider, depending on the behavior of f and f' on (v,s):

- Case i): f is continuous and bounded, and f' is bounded wherever it exists;
- Case ii): f is continuous and bounded, and $f' \nearrow \infty$ monotonically as $x \nearrow s$ in some left neighborhood of s;
- Case iii): f is continuous and bounded, and $f' \searrow -\infty$ monotonically as $x \nearrow s$ in some left neighborhood of s;
- Case iv): $s = 0$, and on (v,s), $f(x) \to \infty$ as $x \nearrow s$, and $f'(x) \nearrow \infty$ monotonically as $x \nearrow s$ in some left neighborhood of s.

Case i): (On (v,s), f is continuous and bounded, and f' is bounded wherever it exists.) The proof, which uses the bounded convergence theorem is similar to that of Fact 2a in the proof of Theorem 9. Let $S < \infty$ be such that $|f'(x)| \leq S$ for all $x \in B \cap (v-\delta, s)$ for some $\delta > 0$. Then for $\Delta < \delta$ and $x \in (v, u_\Delta(s) - 2\Delta)$, Lemma A2 shows that
$$|G_\Delta(x)| \leq 12(|x| + \Delta)S < 12(\max\{|v|,|s|\} + \delta)S\ .$$
Therefore, $|G_\Delta(x)|I_{(v,u_\Delta(s)-2\Delta)}(x)$ is uniformly bounded on (v,s). The bounded convergence theorem then gives
$$\lim_{\Delta\to 0}\int_v^s G_\Delta(x) I_{(v,u_\Delta(s)-2\Delta)}(x)dx = \int_v^s \lim_{\Delta\to 0} G_\Delta(x)dx.$$

Case ii): (On (v,s), f is continuous and bounded, and $f' \nearrow \infty$ monotonically as $x \nearrow s$ in some left neighborhood of s.) Let w be chosen so that $v < w < s$, f' exists everywhere and increases monotonically to infinity on (w,s), $f'(w) > 0$, and $|f'(x)| < f'(w)$ for $x \in B \cap (v,w)$. Then
$$\int_v^s G_\Delta(x) I_{(v,u_\Delta(s)-2\Delta)}(x)\,dx = \int_v^w G_\Delta(x) I_{(v,u_\Delta(s)-2\Delta)}(x)\,dx + \int_w^s G_\Delta(x) I_{(v,u_\Delta(s)-2\Delta)}(x)\,dx\ . \quad (24)$$
Since $f'(x)$ is bounded, wherever it exists on (v,w), the same argument as that in Fact 2a in the proof of Theorem 9 yields,
$$\lim_{\Delta\to 0}\int_v^w G_\Delta(x) I_{(v,u_\Delta(s)-2\Delta)}(x)dx = \int_v^s \lim_{\Delta\to 0} G_\Delta(x)dx.$$

To justify swapping the limit and integral in the second term of (24), we use the dominated convergence theorem[5] [11, p. 209], which requires us to find an integrable function $\widetilde{G}(x)$ that dominates $|G_\Delta(x)|I_{(v,u_\Delta(s)-2\Delta)}(x)$ for all $x \in (w,s)$ and all Δ. With $M \triangleq \max\{|w|,|s|\}$, we choose
$$\widetilde{G}(x) = 24Mf'\left(\frac{x+s}{2}\right), \qquad w < x < s.$$
To show that $\widetilde{G}$ dominates $|G_\Delta|I$, we first observe that for $w < x < u_\Delta(s) - 2\Delta$, the positivity and monotonicity of f' on (w,s)

[5]As with the bounded convergence theorem, the integrand's index parameter is allowed to approach 0 continuously.

$$\int_{-\infty}^{\infty} G_\Delta(x)\,dx = \int_{-\infty}^{v_0} f(x)\,dx - v_0 f(v_0) + \sum_{i=1}^{n}\left(v_{i-1}f(v_{i-1}) + \int_{v_{i-1}}^{s_i} f(x)\,dx - s_i f(s_i^-) + s_i f(s_i^+) + \int_{s_i}^{v_i} f(x)\,dx - v_i f(v_i)\right) + \sum_{i=1}^{n}\left(6s_i[f(s_i^-) - f(s_i^+)]\alpha_\Delta(s_i)[1-\alpha_\Delta(s_i)]\right) + v_n f(v_n) + \int_{v_n}^{\infty} f(x)\,dx + o(1)\ .$$

implies $|f'(x)| = f'(x) \leq f'(u_\Delta(x)+\Delta)$. Then Lemma A2 implies that for all sufficiently small Δ and for $w < x < u_\Delta(s) - 2\Delta$

$$\begin{aligned}|G_\Delta(x)| &\leq 12(|x|+\Delta)f'(u_\Delta(x)+\Delta) \leq 24Mf'(u_\Delta(x)+\Delta)\\ &\leq 24Mf'(x+\Delta) \leq 24Mf'\big(\frac{x+s}{2}\big) = \widetilde{G}(x)\end{aligned}$$

where the third inequality uses the monotonicity of f' on (w,s), and the fourth inequality derives from the fact that $x < u_\Delta(s) - 2\Delta$ implies $x < s - 2\Delta$, which in turn implies $\Delta < \frac{s-x}{2}$. It follows that $|G_\Delta(x)| I_{(v,u_\Delta(s)-2\Delta)}(x) \leq \widetilde{G}(x)$ for all $x \in (w,s)$.

We now check that $\widetilde{G}(x)$ is integrable over (w,s)

$$\begin{aligned}\int_w^s \widetilde{G}(x)\,dx &= 24M\int_w^s f'\big(\frac{x+s}{2}\big)dx = 24M\int_{\frac{w+s}{2}}^s 2f'(y)dy\\ &= 48M\big(f(s) - f\big(\frac{w+s}{2}\big)\big) < \infty\end{aligned}$$

where the inequality follows from the fact that f is bounded. Applying the dominated convergence theorem yields

$$\lim_{\Delta\to 0}\int_w^s G_\Delta(x) I_{(v,u_\Delta(s)-2\Delta)}\,dx = \int_w^s \lim_{\Delta\to 0} G_\Delta(x)\,dx$$

which concludes Case ii). Case iii) is proved in the same manner.

Case iv): ($s=0$, and on (v,s), $f(x)\to\infty$ as $x \nearrow s$, and $f'(x) \nearrow \infty$ monotonically as $x \nearrow s$ in some left neighborhood of s.) This case is similar to Case ii), up to a point. Let w be chosen so that $v < w < s = 0$, f' exists everywhere and increases monotonically to ∞ on $(w,0)$, $f'(w) > 0$, and $|f'(x)| < f'(w)$ for $x \in B \cap (v,w)$. Then

$$\begin{aligned}\int_v^0 G_\Delta(x) I_{(v,u_\Delta(0)-2\Delta)}(x)\,dx &= \int_v^w G_\Delta(x) I_{(v,u_\Delta(0)-2\Delta)}(x)\,dx\\ &\quad + \int_w^0 G_\Delta(x) I_{(v,u_\Delta(0)-2\Delta)}(x)\,dx. \qquad (25)\end{aligned}$$

Since $f'(x)$ is bounded, wherever it exists on (v,w), the same argument as that in Fact 2a in the proof of Theorem 9 yields

$$\lim_{\Delta\to 0}\int_v^w G_\Delta(x) I_{(v,u_\Delta(0)-2\Delta)}(x)dx = \int_v^w \lim_{\Delta\to 0} G_\Delta(x)dx.$$

To justify swapping the limit and integral in the second term of (25), we use the dominated convergence theorem. As the function that dominates $|G_\Delta(x)| I_{(v,u_\Delta(0)-2\Delta)}(x)$, we choose

$$\widetilde{G}(x) = 18|x| f'\big(\frac{x}{2}\big), \qquad w < x < 0.$$

To show that this is indeed a dominating function, recall that Lemma A2 shows that for $w < x < u_\Delta(0) - 2\Delta$

$$|G_\Delta(x)| \leq 12(|x| + |\Delta|) f'(u_\Delta(x)+\Delta)$$

where we used the positivity and monotonicity of f' on $(w,0)$. Now if $w < x < u_\Delta(0) - 2\Delta$, then $x < -2\Delta$, or equivalently, $\Delta < -x/2$. Using these and using again the positivity and monotonicity of f' yields for $w < x < u_\Delta(0) - 2\Delta$

$$\begin{aligned}|G_\Delta(x)| &< 12(|x| + |x|/2) f'(u_\Delta(x)+\Delta)\\ &\leq 18|x| f'(x+\Delta) \leq 18|x| f'\big(\frac{x}{2}\big) = \widetilde{G}(x).\end{aligned}$$

This, in turn, implies $|G_\Delta(x)| I_{(v,u_\Delta(0)-2\Delta)}(x) < \widetilde{G}(x)$, for all $x \in (w,0)$.

We now check the integrability of $\widetilde{G}$

$$\begin{aligned}\int_w^0 \widetilde{G}(x)\,dx &= -18\int_w^0 x f'\big(\frac{x}{2}\big)\,dx = -72\int_{\frac{w}{2}}^0 y f'(y)\,dy\\ &= -72\lim_{x\to 0} x f(x) + 72 w f(w) + \int_w^0 f(x)\,dx\\ &= 0 + 72 w f(w) + \int_w^0 f(x)\,dx < \infty\end{aligned}$$

where the third equality uses integration by parts, and the fourth equality derives from the definition of a nice pdf. Applying the dominated convergence theorem yields

$$\lim_{\Delta\to 0}\int_w^0 G_\Delta(x) I_{(v,u_\Delta(0)-2\Delta)}\,dx = \int_w^0 \lim_{\Delta\to 0} G_\Delta(x)\,dx$$

which concludes Case iv). This completes the proof of Fact 1a.

Proof of Fact 1b:

$$\int_{v_{i-1}}^{s_i} \lim_{\Delta\to 0} G_\Delta(x)dx = v_{i-1} f(v_{i-1}^+) + \int_{v_{i-1}}^{s_i} f(x)dx - s_i f(s_i^-)$$

This follows in a similar way to Fact 2b in the proof of Theorem 9, where v_{i-1} and s_i play the role of a and b, respectively.

Proof of Fact 2a :

$$\int_{u_\Delta(s_i)-2\Delta}^{u_\Delta(s_i)+3\Delta} G_\Delta(x)\,dx = o(1)$$

when f is continuous at s_i or when $s_i = 0$ The considered integral of $G_\Delta(x)$ is over five adjacent quantization cells. We will show that the integral over each of these cells approaches 0 as $\Delta \to 0$. To simplify notation, we write s instead of s_i. There are two cases: $s=0$ and $s \neq 0$.

If $s=0$, then the integral over any one of the five quantization cells has the form shown in the first equation at the bottom of the page, for some $j \in \{-2,-1,0,1,2\}$. The magnitude of the first bracketed term is at most one half, the magnitude of the second bracketed term is easily seen to be no larger than 6, and the third bracketed term goes to zero as $\Delta \to 0$. Therefore,

$$\int_{u_\Delta(0)-j\Delta}^{u_\Delta(0)-j\Delta+\Delta} G_\Delta(x)\,dx = o(1)$$

and since this holds for the integral over each of the five adjacent cells, the result follows in the case $s=0$.

Next, if $s \neq 0$, then f is continuous at s. In this case see the second equation at the bottom of the page. Suppose $f(s)=0$. Then the magnitude of the first bracketed term is at most one half, the second bracketed term tends to $2s$ as $\Delta \to 0$, and due to the continuity of f at s, the third bracketed term tends to $f(s)=0$. Therefore, the product of the three bracketed terms approaches zero. Now suppose $f(s) \neq 0$. By Lemma A1, which is a slightly strengthened version of Lemma 1, the first bracketed

$$\int_{u_\Delta(0)-j\Delta}^{u_\Delta(0)-j\Delta+\Delta} G_\Delta(x)\,dx = 6\left[\frac{m_\Delta(0-j\Delta) - c_\Delta(0-j\Delta)}{\Delta}\right]\left[\frac{m_\Delta(0-j\Delta) + c_\Delta(0-j\Delta)}{\Delta}\right]\left[\int_{u_\Delta(0)-j\Delta}^{u_\Delta(0)-j\Delta+\Delta} f(x)\,dx\right]$$

$$\int_{u_\Delta(s)-j\Delta}^{u_\Delta(s)-j\Delta+\Delta} G_\Delta(x)\,dx = 6\left[\frac{m_\Delta(s-j\Delta) - c_\Delta(s-j\Delta)}{\Delta}\right]\left[m_\Delta(s-j\Delta) + c_\Delta(s-j\Delta)\right]\left[\frac{1}{\Delta}\int_{u_\Delta(s)-j\Delta}^{u_\Delta(s)-j\Delta+\Delta} f(x)\,dx\right].$$

term approaches 0. The second and third terms approach $2s$ and $f(s)$, respectively, as before. Therefore, again the product of the three bracketed terms approaches zero. This completes the proof of Fact 2a.

Proof of Fact 2b:

$$\int_{u_\Delta(s_i)-2\Delta}^{u_\Delta(s_i)+3\Delta} G_\Delta(x)dx = 6s_i[f(s_i^-) - f(s_i^+)]\alpha_\Delta(s_i)[1-\alpha_\Delta(s_i)] + o(1)$$

when f has a finite jump discontinuity at s_i and $s_i \neq 0$ As before, to simplify notation, we write s instead of s_i. We shall also write α_Δ instead of $\alpha_\Delta(s)$. First observe that having a discontinuity at s has no effect on the integral over the four non-middle cells. Thus, the integral over these cells tends to zero as $\Delta \to 0$ as shown in Fact 2a. It remains to consider the integral over the middle cell, which contains the discontinuity at s. Specifically, it needs to be shown that

$$\int_{u_\Delta(s)}^{u_\Delta(s)+\Delta} G_\Delta(x)dx = 6s[f(s^-)-f(s^+)]\alpha_\Delta(1-\alpha_\Delta) + o(1). \tag{26}$$

We decompose the integral above as shown in (27) at the bottom of the page. The three terms in (27) converge as follows:

$$\frac{1}{\Delta}\int_{s-\Delta\alpha_\Delta}^{s+\Delta(1-\alpha_\Delta)} f(x)\,dx = \alpha_\Delta f(s^-) + (1-\alpha_\Delta)f(s^+) + o(1) \tag{28}$$

$$m_\Delta(s) + c_\Delta(s) = 2s + o(1) \tag{29}$$

$$\frac{m_\Delta(s) - c_\Delta(s)}{\Delta/6} = 6\frac{\frac{\alpha_\Delta(1-\alpha_\Delta)}{2}[f(s^-)-f(s^+)]}{\alpha_\Delta f(s^-) + (1-\alpha_\Delta)f(s^+)} + o(1) \tag{30}$$

where (28) is due to the continuity of f on $(s-\Delta\alpha_\Delta, s)$ and $(s, s+\Delta(1-\alpha_\Delta))$. Equation (30) can be obtained by observing that $m_\Delta(s) = s + \Delta(\frac{1}{2} - \alpha_\Delta)$ and by noting that it can be shown that

$$c_\Delta(s) = \gamma_{L,\Delta} c_{L,\Delta}(s) + \gamma_{R,\Delta} c_{R,\Delta}(s)$$

where

$$\gamma_{L,\Delta} = \frac{\int_{u_\Delta(s)}^{s} f(x)\,dx}{\gamma_\Delta}$$
$$\gamma_{R,\Delta} = \frac{\int_{s}^{u_\Delta(s)+\Delta} f(x)\,dx}{\gamma_\Delta}$$
$$\gamma_\Delta = \int_{u_\Delta(s)}^{u_\Delta(s)+\Delta} f(x)\,dx$$

and where $c_{L,\Delta}(s)$ is the centroid of $(u_\Delta(s), s)$ and $c_{R,\Delta}(s)$ is the centroid of $(s, u_\Delta(s)+\Delta)$. Next, by Lemma 15 it follows that $c_{L,\Delta}(s) = m_{L,\Delta}(s) + O(\Delta^2)$ and $c_{R,\Delta}(s) = m_{R,\Delta}(s) + O(\Delta^2)$, where $m_{L,\Delta}(s) = s - \frac{\Delta\alpha_\Delta}{2}$ is the midpoint of $(u_\Delta(s), s)$, $m_{R,\Delta}(s) = s + \frac{\Delta(1-\alpha_\Delta)}{2}$ is the midpoint of $(s, u_\Delta(s)+\Delta)$, and $O(\Delta^2)$ is a quantity such that $O(\Delta^2)/\Delta^2$ remains bounded as $\Delta \to 0$. Thus, we have

$$\begin{aligned}\frac{m_\Delta(s) - c_\Delta(s)}{\Delta} &= \frac{m_\Delta(s)}{\Delta} - \gamma_{L,\Delta}\frac{\left(m_{L,\Delta}(s) + O(\Delta^2)\right)}{\Delta} \\ &\quad - \gamma_{R,\Delta}\frac{\left(m_{R,\Delta}(s) + O(\Delta^2)\right)}{\Delta} \\ &= \gamma_{L,\Delta}\frac{\left(m_\Delta(s) - m_{L,\Delta}(s) + O(\Delta^2)\right)}{\Delta} \\ &\quad + \gamma_{R,\Delta}\frac{\left(m_\Delta - m_{R,\Delta}(s) + O(\Delta^2)\right)}{\Delta}.\end{aligned}$$

Using the continuity of f on the intervals $(u_\Delta(s), s)$ and $(s, u_\Delta(s)+\Delta)$, it is easily seen that

$$\gamma_{L,\Delta} = \frac{\alpha_\Delta f(s^-)}{\alpha_\Delta f(s^-) + (1-\alpha_\Delta)f(s^+)} + o(1)$$

and

$$\gamma_{R,\Delta} = \frac{(1-\alpha_\Delta) f(s^+)}{\alpha_\Delta f(s^-) + (1-\alpha_\Delta)f(s^+)} + o(1)$$

Plugging these into the above, together with some algebraic steps, establishes (30).

Finally, since $\alpha_\Delta f(s^-) + (1-\alpha_\Delta)f(s^+)$, $2s$, and

$$6\frac{\frac{\alpha_\Delta(1-\alpha_\Delta)}{2}[f(s^-) - f(s^+)]}{\alpha_\Delta f(s^-) + (1-\alpha_\Delta)f(s^+)}$$

are bounded, substituting (28)–(30) into (27) yields (26), which completes the proof of Fact 2b and Theorem 10. □

Proof of Theorem 13: We need to show that there exists some offset function $\theta(\Delta)$ for which $\lim_{\Delta\to 0} s_\Delta(f)$ does not exist. We begin by writing

$$\begin{aligned} s_\Delta(f) &= 1 + \sum_{k=1}^N t_k e_k - 6\sum_{k=1}^N t_k e_k \alpha_\Delta(t_k)(1-\alpha_\Delta(t_k)) \\ &= 1 + \sum_{k=1}^N t_k e_k - 6\langle \underline{\lambda}, \underline{\beta}_\Delta \rangle, \end{aligned} \tag{31}$$

where $e_k = f(t_k^+) - f(t_k^-)$ is the jump height at t_k,

$$\alpha_\Delta(t_k) = \frac{t_k - u_\Delta(t_k)}{\Delta}$$

is the offset of t_k within its cell, $\underline{\lambda} = (\lambda_1, \ldots, \lambda_N)$, $\lambda_k = t_k e_k$, $\underline{\beta}_\Delta = (\beta_{1,\Delta}, \ldots, \beta_{N,\Delta})$, $\beta_{k,\Delta} = \alpha_\Delta(t_k)(1-\alpha_\Delta(t_k))$, and $\langle \underline{u}, \underline{v}\rangle = \sum_{k=1}^N u_k v_k$ is the usual inner product. We use underbars throughout to denote vectors.

Since the first two terms of (31) do not depend on Δ, it suffices to find an offset function $\theta(\Delta)$ such that $\lim_{\Delta\to 0}\langle \underline{\lambda}, \underline{\beta}_\Delta\rangle$ does not exist. To this end, we will set $\Delta_\tau = \frac{1}{\tau}$, and let $\tau \to \infty$. We will then show that there exists a fixed quantity τ_δ such that for any $\tau_o > 0$, there exists $\tau_1 \geq \tau_o$ and a value θ_{τ_1} such that with $\theta(\tau) = \theta_{\tau_1}$ for $\tau_1 < \tau < \tau_1 + 2\tau_\delta$, the above inner product

$$\int_{u_\Delta(s)}^{u_\Delta(s)+\Delta} G_\Delta(x)\,dx = \left[\frac{m_\Delta(s) - c_\Delta(s)}{\Delta/6}\right][m_\Delta(s) + c_\Delta(s)]\left[\frac{1}{\Delta}\int_{s-\Delta\alpha_\Delta}^{s+\Delta(1-\alpha_\Delta)} f(x)\,dx\right]. \tag{27}$$

varies by some prespecified positive amount as τ goes from τ_1 to $\tau_1 + 2\tau_\delta$. This will then imply that $\langle \underline{\lambda}, \underline{\beta}(\Delta_\tau)\rangle$ does not converge, which is equivalent to $\lim_{\Delta\to 0} \langle \underline{\lambda}, \underline{\beta}_\Delta\rangle$ not existing.

We notice that if $t_k = 0$, then by definition of $s_\Delta(f)$, the discontinuity at t_k contributes nothing to $s_\Delta(f)$. Thus, without loss of generality, we assume $t_k \neq 0$ for all $k \in \{1, \dots, N\}$. Furthermore, without loss of generality, we assume that the components of $\underline{t} = (t_1, \dots, t_N)$ are ordered by magnitude, i.e.,

$$0 < |t_1| \le |t_2| \le \cdots \le |t_{N-1}| \le t_N$$

where we also assume, without loss of generality, that $t_N > 0$. To simplify matters, we change slightly our notation for the α's and β's. Specifically, let $\alpha_k^\phi(\tau)$ denote the α value at t_k when $\Delta = \Delta_\tau$ and $\theta(\Delta) = \phi$. Similarly, let $\beta_k^\phi(\tau)$ denote the corresponding β value. In this notation

$$\begin{aligned}\alpha_k^\phi(\tau) &= \frac{(t_k + \Delta_\tau\phi) \bmod \Delta_\tau}{\Delta_\tau} = \frac{\Delta_\tau[(t_k/\Delta_\tau + \phi) \bmod 1]}{\Delta_\tau} \\ &= \Big(\frac{t_k}{\Delta_\tau} + \phi\Big) \bmod 1 = (t_k\tau + \phi) \bmod 1\,. \end{aligned} \tag{32}$$

In addition, from now on, we consider the offset to be a function of τ, namely, $\theta(\tau)$, rather than a function of Δ. Our goal will be to find an offset function $\theta(\tau)$ for which we can show that $\lim_{\tau\to\infty} \langle \underline{\lambda}, \underline{\beta}^{\theta(\tau)}(\tau)\rangle$ does not exist, where

$$\underline{\beta}^{\theta(\tau)} = (\beta_1^{\theta(\tau)}(\tau), \dots, \beta_N^{\theta(\tau)}(\tau)).$$

Before going into the details of the proof, we give an intuitive view of the meaning of α in light of (32), followed by an outline of the proof. We identify the unit interval with the unit circle, with 0 located at 12 o'clock. As τ goes to ∞, we view $\alpha_k^{\theta(\tau)}(\tau)$ as rotating around the unit circle—clockwise when $t_k > 0$, and counterclockwise when $t_k < 0$. If $\theta(\tau)$ is constant over an interval of τ's, then we see from (32) that each $\alpha_k^{\theta(\tau)}$ changes linearly with τ unless it passes through 0, in which case the $\bmod 1$ comes into effect—subtracting 1 from α if $t_k > 0$, and adding 1 if $t_k < 0$.

It is easy to see from (32) that if the offset were held constant, then no α_k would converge and consequently no β_k and no product $\lambda_k\beta_k$ would converge. The difficulty lies in showing that the inner product, which is the sum of products $\lambda_k\beta_k$, does not converge either. Essentially, one must show that nonconvergent terms in the sum cannot somehow negate each other's nonconvergence. We do this by choosing an offset function that is piecewise constant rather than constant. Specifically, we show that for any $\tau_o > 0$ there exists $\tau_1 \ge \tau_o$, an interval $[\tau_1, \tau_1 + 2\tau_\delta)$, and a constant offset in this interval that cause the following favorable property to hold. All α_k's, except α_N, do not pass through zero and, consequently, change linearly with τ over this interval. α_N, on the other hand, passes through zero in the middle of the interval (but nowhere else). Thus, it changes linearly over the first half of the interval and has a discontinuity as τ passes to the second half of the interval and the $\bmod 1$ comes into effect.

Using this property, the inner product $\langle \underline{\lambda}, \underline{\beta}^{\theta(\tau)}(\tau)\rangle$ turns out to be a parabolic function of τ in the first half of the interval, i.e., $A\tau^2 + B\tau + C$. If A is not zero or B is bounded away from zero for large values of τ_o, then it is easily shown that the inner product must change by some nonzero amount that can be specified in advance. Otherwise, we use the discontinuity in α_N at the halfway point of the interval to lower-bound the amount of change.

To keep notation short, we will assume throughout the proof that $k \in \{1, \dots, N\}$, unless otherwise specified. We set $\mu = \min_{k\in\{1,\dots,N-1\}}(t_N - t_k)$ and $\delta = \frac{\mu}{32Nt_N}$, which remain fixed for the rest of the proof. Let W denote the set of all $\tau \ge 0$ such that

$$\left|\alpha_N^0(\tau) - \alpha_k^0(\tau)\right|_L \ge 4\delta, \qquad \text{for all } k \neq N \tag{33}$$

where $|a-b|_L$ denotes Lee distance, i.e.,

$$|a-b|_L = \min\{(a-b) \bmod 1, 1 - ((a-b) \bmod 1)\}.$$

The following lemma asserts that W is unbounded.

Lemma 17: For any $\tau_o > 0$ there exists $\tau \ge \tau_o$ such that $\tau \in W$.

Proof: The proof is constructive. If $\tau_o \in W$, there is nothing to show. From now on, assume $\tau_o \notin W$. Consider first $\tau_1 = \tau_o + \frac{8\delta}{\mu}$. For every k such that $|\alpha_N^0(\tau) - \alpha_k^0(\tau)|_L < 4\delta$, we have

$$\begin{aligned}\left|\alpha_N^0(\tau_1) - \alpha_k^0(\tau_1)\right|_L &= \Big|\alpha_N^0(\tau_o) + \frac{8\delta}{\mu}t_N - \alpha_k^0(\tau_o) - \frac{8\delta}{\mu}t_k\Big|_L \\ &= \Big|\alpha_N^0(\tau_o) - \alpha_k^0(\tau_o) + \frac{8\delta}{\mu}(t_N - t_k)\Big|_L. \end{aligned} \tag{34}$$

Combining (34) and the facts that $\frac{8\delta}{\mu}(t_N - t_k) \ge 8\delta$ and $|\alpha_N^0(\tau) - \alpha_k^0(\tau)|_L < 4\delta$, it follows straightforwardly that $|\alpha_N^0(\tau_1) - \alpha_k^0(\tau_1)|_L > 4\delta$.

We have shown that $|\alpha_N^0(\tau_1) - \alpha_k^0(\tau_1)|_L > 4\delta$ for those k's considered above. However, now that τ has been increased from τ_o to τ_1, it is possible that other α_k's no longer satisfy (33). We can "fix" these by increasing τ, yet again, to $\tau_2 = \tau_1 + \frac{8\delta}{\mu}$. Since t_N has largest magnitude, the distance between α_N and previously fixed α's will only increase, and so a fixed α need not be fixed again. Thus, repeating this process at most $N-1$ times will guarantee that all α_k's are fixed, i.e., (33) is satisfied for all $k \neq N$, provided we make one additional check.

From a geometrical point of view, the process of fixing some α_k involves sufficient advancement of α_N in the clockwise direction, thus letting α_N gain sufficient distance from α_k. We observe, however, that the assertion that "repeating this process at most $N-1$ times will guarantee that all α_k's are fixed" is correct if the distance between a previously fixed α_k and α_N cannot become small again due to having α_N get close to α_k from the "other" direction as τ is increased. This, however, cannot happen, since in $N-1$ steps, τ increases from τ_o to

$$\begin{aligned}\tau_{N-1} &= \tau_o + (N-1)\frac{8\delta}{\mu} \\ &= \tau_o + (N-1)\frac{8}{\mu}\frac{\mu}{32Nt_N} < \tau_o + \frac{1}{4t_N}.\end{aligned}$$

Consequently, α_N advances less than $\frac{1}{4t_N}t_N = \frac{1}{4}$. Since t_N has largest magnitude, all other α_k's advance less than $\frac{1}{4}$. Therefore, $\alpha_N^0(\tau_{N-1})$ is at least $\frac{1}{2}$ away in clockwise direction from any $\alpha_k^0(\tau_{N-1})$ that was fixed using the above process. □

Now, we use Lemma 17 to show that for any τ_o there exists $\tau_1 \ge \tau_o$, an interval $[\tau_1, \tau_1 + 2\tau_\delta)$, and a choice of constant offset in this interval with the favorable property discussed in the proof outline described earlier.

Let $\tau_o > 0$ be given. It follows from Lemma 17 that there exist $\tau_1 \geq \tau_o$ for which $|\alpha_N^0(\tau_1) - \alpha_k^0(\tau_1)|_L \geq 4\delta$ for all $k \neq N$. Fix $\theta_{\tau_1} = \big(1 - \delta - \alpha_N^0(\tau_1)\big) \bmod 1$, and let $\theta(\tau) = \theta_{\tau_1}$ for $\tau_1 \leq \tau < \tau + 2\tau_\delta$, where $\tau_\delta \triangleq \frac{\delta}{t_N} = \frac{\mu}{32Nt_N^2}$. This choice of τ_1 and θ_{τ_1} makes

$$\begin{aligned}\alpha_N^{\theta_{\tau_1}}(\tau_1) &= (t_N\tau_1 + \theta_{\tau_1}) \bmod 1\\ &= \big(\alpha_N^0(\tau_1) + 1 - \delta - \alpha_N^0(\tau_1)\big) \bmod 1 = 1 - \delta.\end{aligned}$$

Moreover, since $t_N\tau_\delta = \delta$, we deduce from (32) and the above that $\alpha_N^{\theta_{\tau_1}}(\tau)$ increases linearly from its value $1-\delta$ at $\tau = \tau_1$, and passes through zero precisely at $\tau = \tau_1 + \tau_\delta$. And since $\delta = \min_{k\in\{1,\ldots,N-1\}} t_N - t_k/(32Nt_N) \leq 1/(16N)$, it will not pass through zero anywhere else in the interval $[\tau_1, \tau_1 + 2\tau_\delta)$. It follows that

$$\alpha_N^{\theta_{\tau_1}}(\tau_1 + s) = \begin{cases}\alpha_N^{\theta_{\tau_1}}(\tau_1) + st_N, & 0 \leq s < \tau_\delta\\ \alpha_N^{\theta_{\tau_1}}(\tau_1) + st_N - 1, & \tau_\delta \leq s < 2\tau_\delta 2\tau_\delta.\end{cases} \tag{35}$$

Next, for any $k \neq N$, the facts that $\alpha_N^{\theta_{\tau_1}}(\tau_1) = 1 - \delta$,

$$|\alpha_N^{\theta_{\tau_1}}(\tau_1) - \alpha_k^{\theta_{\tau_1}}(\tau_1)|_L = |\alpha_N^0(\tau_1) - \alpha_k^0(\tau_1)|_L \geq 4\delta,$$

and $|t_N| \geq |t_k|$ imply that $\alpha_k^{\theta_{\tau_1}}(\tau)$ cannot pass through zero in the interval $[\tau_1, \tau_1 + 2\tau_\delta)$. Therefore,

$$\alpha_k^{\theta_{\tau_1}}(\tau_1 + s) = \alpha_k^{\theta_{\tau_1}}(\tau_1) + st_k, \qquad 0 \leq s < 2\tau_\delta \text{ and } k \neq N. \tag{36}$$

Having established (35) and (36) we are now ready to express $\langle \underline{\lambda}, \underline{\beta}^{\theta(\tau)}(\tau)\rangle$ as a parabolic function of τ, when $\tau \in [\tau_1, \tau_1 + \tau_\delta)$. To do so, we observe that for all $s \in [0, \tau_\delta)$ and for all k

$$\begin{aligned}&\beta_k^{\theta_{\tau_1}}(\tau_1 + s)\\ &\quad= (\alpha_k^{\theta_{\tau_1}}(\tau_1) + st_k)(1 - \alpha_k^{\theta_{\tau_1}}(\tau_1) - st_k)\\ &\quad= \alpha_k^{\theta_{\tau_1}}(\tau_1)\big(1 - \alpha_k^{\theta_{\tau_1}}(\tau_1)\big) + t_k\big(1 - 2\alpha_k^{\theta_{\tau_1}}(\tau_1)\big)s - t_k^2 s^2\\ &\quad= \beta_k^{\theta_{\tau_1}}(\tau_1) + t_k\big(1 - 2\alpha_k^{\theta_{\tau_1}}(\tau_1)\big)s - t_k^2 s^2\,.\end{aligned} \tag{37}$$

Using (37), we evaluate $\langle \underline{\lambda}, \underline{\beta}^{\theta_{\tau_1}}(\tau_1 + s)\rangle$ for $s \in [0, \tau_\delta)$ as follows:

$$\begin{aligned}\langle \underline{\lambda}, \underline{\beta}^{\theta_{\tau_1}}(\tau_1 + s)\rangle &= \sum_{k=1}^N \lambda_k \beta_k^{\theta_{\tau_1}}(\tau_1) + \left[\sum_{k=1}^N \lambda_k t_k\big(1 - 2\alpha_k^{\theta_{\tau_1}}(\tau_1)\big)\right]s\\ &\quad + \left[-\sum_{k=1}^N \lambda_k t_k^2\right]s^2\\ &= As^2 + B_{\tau_1}s + C_{\tau_1}\end{aligned} \tag{38}$$

where

$$A \triangleq -\sum_{k=1}^N \lambda_k t_k^2$$

$$B_{\tau_1} \triangleq \sum_{k=1}^N \lambda_k t_k\big(1 - 2\alpha_k^{\theta_{\tau_1}}(\tau_1)\big)$$

and

$$C_{\tau_1} \triangleq \sum_{k=1}^N \lambda_k \beta_k^{\theta_{\tau_1}}(\tau_1).$$

We proceed by showing how to lower-bound the amount of change in the inner product. To keep notation short, we set $\mathcal{Y}_\tau(s) \triangleq \langle \underline{\lambda}, \underline{\beta}^{\theta_\tau}(\tau + s)\rangle$. If $A \neq 0$, then (38) shows that $\mathcal{Y}_{\tau_1}(s)$ is a parabolic function. Therefore,

$$\left|\mathcal{Y}_{\tau_1}\left(\frac{\tau_\delta}{4}\right) - \mathcal{Y}_{\tau_1}(0)\right| \geq |A|\left(\frac{\tau_\delta}{4}\right)^2$$

and /or

$$\left|\mathcal{Y}_{\tau_1}\left(\frac{\tau_\delta}{2}\right) - \mathcal{Y}_{\tau_1}\left(\frac{\tau_\delta}{4}\right)\right| \geq |A|\left(\frac{\tau_\delta}{4}\right)^2 \tag{39}$$

which derives from the fact that if $y(x) = ax^2 + bx + c$ and $a \neq 0$, then for any $t \in \mathbb{R}$,

$$|y(t) - y(0)| \geq |a|t^2 \quad \text{and/or} \quad |y(2t) - y(t)| \geq |a|t^2.$$

This fact can be seen as follows. $y(0) = c$, $y(t) = at^2 + bt + c$ and $y(2t) = 4at^2 + 2bt + c$. Thus, $|y(t) - y(0)| = |at^2 + bt|$ and if $|at^2 + bt| \geq |a|t^2$, then the fact is shown. Otherwise

$$\begin{aligned}|y(2t) - y(t)| &= |3at^2 + bt|\\ &= |(at^2 + bt) + 2at^2| > |2at^2 - at^2| = |a|t^2\end{aligned}$$

which shows the fact.

Thus, for the case $A \neq 0$, we have established a lower bound to the change of the inner product $\langle \underline{\lambda}, \underline{\beta}^{\theta(\tau)}(\tau)\rangle$ as τ ranges over the interval $[\tau_o, \infty)$.

Suppose next that $A = 0$. Then (38) reduces to

$$\mathcal{Y}_{\tau_1}(s) = B_{\tau_1}s + C_{\tau_1}, \qquad \text{when } s \in [0, \tau_\delta). \tag{40}$$

If

$$\lim_{\substack{\tau\to\infty\\ \tau\in W}} B_\tau \neq 0$$

(in particular the limit need not exist), then for any $\tau_o > 0$, there exists $\tau_1 \in W$ such that $\tau_1 \geq \tau_o$ and

$$|B_{\tau_1}| > u \triangleq \frac{1}{2}\limsup_{\substack{\tau\to\infty\\ \tau\in W}} |B_\tau| > 0.$$

Consequently

$$\left|\mathcal{Y}_{\tau_1}\left(\frac{\tau_\delta}{2}\right) - \mathcal{Y}_{\tau_1}(0)\right| = |B_{\tau_1}|\frac{\tau_\delta}{2} > u\frac{\tau_\delta}{2} = u\frac{\delta}{2t_N}\,. \tag{41}$$

This establishes a lower bound to the change of the inner product $\langle \underline{\lambda}, \underline{\beta}^{\theta(\tau)}(\tau)\rangle$ for the case $A = 0$ and

$$\lim_{\substack{\tau\to\infty\\ \tau\in W}} B_\tau \neq 0.$$

It remains to consider the case that $A = 0$ and

$$\lim_{\substack{\tau\to\infty\\ \tau\in W}} B_\tau = 0.$$

For any $\tau_o > 0$, there exists $\tau_1 \in W$ such that $\tau_1 \geq \tau_o$ and

$$|B_{\tau_1}| < \frac{|\lambda_N|\delta}{\tau_\delta}.$$

Using (35) we obtain

$$\begin{aligned}\beta_N^{\theta_{\tau_1}}(\tau_1+s) &= (\alpha_N^{\theta_{\tau_1}}(\tau_1)+st_N-1)(2-\alpha_N^{\theta_{\tau_1}}(\tau_1)-st_N)\\ &= \beta_N^{\theta_{\tau_1}}(\tau_1)+t_N\big(1-2\alpha_N^{\theta_{\tau_1}}(\tau_1)\big)s-t_N^2s^2\\ &\quad -2\big[1-\alpha_N^{\theta_{\tau_1}}(\tau_1)-st_N\big], \qquad \text{when } \tau_\delta \leq s < 2\tau_\delta. \end{aligned} \tag{42}$$

Using (42) and observing that the expression in (37) holds for $s \in [\tau_\delta, 2\tau_\delta)$ for all $1 \leq k \leq N-1$, it follows via a derivation similar to that of (40), that for all $s \in [\tau_\delta, 2\tau_\delta)$

$$\mathcal{Y}_{\tau_1}(s) = \big(B_{\tau_1}s + C_{\tau_1}\big) + \big(-2\lambda_N[1-\alpha_N^{\theta_{\tau_1}}(\tau_1)] + 2\lambda_N t_N s\big).$$

Consequently

$$\left|\mathcal{Y}_{\tau_1}\left(\frac{7\tau_\delta}{4}\right) - \mathcal{Y}_{\tau_1}\left(\frac{5\tau_\delta}{4}\right)\right| = \left|B_{\tau_1}\frac{\tau_\delta}{2} + \lambda_N t_N \tau_\delta\right| > \frac{|\lambda_N|\delta}{2} \tag{43}$$

where the inequality follows from having $|B_{\tau_1}| < \frac{|\lambda_N|\delta}{\tau_\delta}$. This establishes a lower bound to the change of the inner product $\langle \underline{\lambda}, \underline{\beta}^{\theta(\tau)}(\tau)\rangle$ for the case $A = 0$ and

$$\lim_{\substack{\tau\to\infty\\ \tau\in W}} B_\tau = 0.$$

Combining (39), (41), and (43) we have shown that for any $\tau_o > 0$ there exist $\tau_b > \tau_a \geq \tau_o$ such that

$$\big|\langle \underline{\lambda}, \underline{\beta}(\tau_b)\rangle - \langle \underline{\lambda}, \underline{\beta}(\tau_a)\rangle\big| > |A|\Big(\frac{\delta}{4t_N}\Big)^2 + \frac{\delta}{2|t_N|}V_1 + \frac{|\lambda_N|\delta}{2}V_2$$

where V_1 is a quantity that equals 0 if $A \neq 0$, or if $A = 0$ and

$$\lim_{\substack{\tau\to\infty\\ \tau\in W}} B_\tau = 0.$$

Otherwise, $V_1 = u$. And V_2 is a quantity that equals 0 if $A \neq 0$, or if $A = 0$ and

$$\lim_{\substack{\tau\to\infty\\ \tau\in W}} B_\tau \neq 0.$$

Otherwise, $V_2 = 1$. This shows that $\langle \underline{\lambda}, \underline{\beta}^{\theta(\tau)}(\tau)\rangle$ does not converge and concludes the proof of the theorem. □

Remark: There is a simple way to prove Theorem 13 for almost all vectors $\underline{t} \in \mathbb{R}^N$. Specifically, for the cases that $\underline{t}$ is rationally independent (i.e., for $\underline{t}$'s such that for all nonzero $\underline{h} \in \mathbb{Z}^N$, $\langle \underline{t}, \underline{h}\rangle \notin \mathbb{Z}$), which is almost all of $\mathbb{R}^N$. The following is a brief sketch. Fix $\theta(\Delta) = 0$ and set $\Delta_n = \frac{1}{n}$ for $n \in \mathbb{Z}^+$. From (32), we have that $\alpha_{k,n} = t_k n \bmod 1$. Since $\underline{t}$ is rationally independent, it follows via a theorem of Kronecker [12] (cf. [13, p. 158], which also cites [14]) that the sequence $\{\alpha_{1,n}, \ldots, \alpha_{N,n}\}_{n=1}^\infty$ is dense in $[0,1)^N$ and so the sequence $\{\beta_{1,n}, \ldots, \beta_{N,n}\}_{n=1}^\infty$ is dense in $[0, \frac{1}{4}]^N$. Therefore, there exists a subsequence n_l of n, such that β_{k,n_l} is arbitrarily small for all $k \in \{1, \ldots, N-1\}$ and β_{N,n_l} is dense in $[0, \frac{1}{4}]$. Finally,

$$\langle \underline{\lambda}, \underline{\beta}_{n_l}\rangle = \sum_{k=1}^N \lambda_k \beta_{k,n_l} \approx \lambda_N \beta_{N,n_l}.$$

Since β_{N,n_l} is dense in $[0, \frac{1}{4}]$, it follows that $\langle \underline{\lambda}, \underline{\beta}_{n_l}\rangle$ does not converge (in fact, it has an uncountable number of limit points) and consequently, $\langle \underline{\lambda}, \underline{\beta}_n\rangle$ does not converge.

VIII. Conclusion

Corollary 12 rigorously establishes that the widely used additive noise model for uniform scalar quantization is, as one would hope, valid in an asymptotic sense whenever the input pdf is continuous and also satisfies certain other benign conditions. Specifically, the correlation between input and quantization error is asymptotically negligible relative to the MSE, or equivalently, to the square of the quantizer level spacing Δ. The model is even valid when there is a discontinuity at the origin. On the other hand, Theorem 13 shows that discontinuities elsewhere can cause the correlation between the input and quantization error to no longer be negligible relative to the MSE. In such cases, the additive noise model is not asymptotically valid. Nevertheless, Theorem 10 permits one to estimate the correlation when Δ is small, in terms of the heights of the discontinuities and their fractional positions within quantization cells.

The derivation of these results is based on an analysis of the asymptotic convergence of cell centroids to cell midpoints, as expressed in the functional r. This convergence is shown to be fast enough to account for the fact that the distortion induced by midpoints is asymptotically the same as that induced by centroids. But it is not fast enough to cause the correlation induced by midpoints to be similar to that induced by centroids.

For a pdf with finite support, such as a uniform pdf, we have also shown that it is possible to design the uniform quantizer to be matched to the support in such a way that the correlation has an asymptotic limit. Depending on the pdf and the manner of matching, a wide variety of correlations may be possible.

Finally, it is interesting to consider that any discontinuous pdf can be well approximated by a continuous pdf. For example, suppose a pdf with jump discontinuities is approximated by a continuous pdf that replaces each jump with a ramp of width δ. When $\Delta \ll \delta$, $r(f) \approx 1$ and the additive noise model is valid. On the other hand, when $\Delta \geq \delta$, the value of $r(f)$ can be quite far from 1, and consequently, the correlation need not be small relative to the MSE.

Appendix

Lemma 1: We will prove the following slightly stronger version of Lemma 1.

Lemma A1: If f is continuous and positive at x, then for any offset function and any integer j

$$\lim_{\Delta\to 0} \frac{m_\Delta(x - j\Delta) - c_\Delta(x - j\Delta)}{\Delta} = 0\,.$$

Proof: Suppose f is continuous and positive at x. Recall that $u_\Delta(x)$ denotes the left boundary of the cell containing x. It follows from the definitions of $m_\Delta(x)$ and $c_\Delta(x)$ that

$$\frac{m_\Delta(x-j\Delta)-c_\Delta(x-j\Delta)}{\Delta} = \frac{\frac{1}{\Delta^2}\int_{u_\Delta(x-j\Delta)}^{u_\Delta(x-j\Delta)+\Delta}(u_\Delta(x-j\Delta)+\frac{\Delta}{2}-t)f(t)\,dt}{\frac{1}{\Delta}\int_{u_\Delta(x-j\Delta)}^{u_\Delta(x-j\Delta)+\Delta} f(t)\,dt}. \quad \text{(A1)}$$

Since f is continuous at x, the denominator of the above converges to $f(x)$. Now consider the numerator

$$\begin{aligned}\frac{1}{\Delta^2}&\int_{u_\Delta(x-j\Delta)}^{u_\Delta(x-j\Delta)+\Delta}(u_\Delta(x-j\Delta)+\frac{\Delta}{2}-t)f(t)\,dt\\ &=\frac{1}{\Delta^2}\int_0^{\frac{\Delta}{2}}(\frac{\Delta}{2}-y)f(x+(u_\Delta(x-j\Delta)-x+y))\,dy\\ &\quad+\frac{1}{\Delta^2}\int_{\frac{\Delta}{2}}^{\Delta}(\frac{\Delta}{2}-y)f(x+(u_\Delta(x-j\Delta)-x+y))\,dy.\end{aligned}$$

Let $\varepsilon>0$ be given. Since f is continuous at x, there exists $\delta>0$ such that $|f(x+t)-f(x)|<\varepsilon$ for all $t\in(-\delta,\delta)$. Therefore, when $\Delta<\frac{\delta}{j+2}$ and $0<y<\Delta$, we have $|u_\Delta(x-j\Delta)-x+y|<(j+2)\Delta<\delta$ which in turn implies $|f(x+(u_\Delta(x-j\Delta)-x+y))-f(x)|<\varepsilon$. Using this in the right-hand side of the above, we have that for all sufficiently small Δ

$$\frac{1}{\Delta^2}\int_0^{\frac{\Delta}{2}}(\frac{\Delta}{2}-y)f(x+(u_\Delta(x-j\Delta)-x+y))\,dy < \frac{f(x)+\varepsilon}{8}$$

and

$$\frac{1}{\Delta^2}\int_{\frac{\Delta}{2}}^{\Delta}(\frac{\Delta}{2}-y)f(x+(u_\Delta(x-j\Delta)-x+y))\,dy < \frac{\varepsilon-f(x)}{8}.$$

Thus,

$$\frac{1}{\Delta^2}\int_0^{\Delta}(\frac{\Delta}{2}-y)f(u_\Delta(x-j\Delta)+y)\,dy<\frac{\varepsilon}{4}$$

for all sufficiently small Δ. In much the same way, it can be shown that

$$\frac{1}{\Delta^2}\int_0^{\Delta}(\frac{\Delta}{2}-y)f(u_\Delta(x-j\Delta)+y)\,dy>-\frac{\varepsilon}{4}.$$

Since ε is arbitrary, we obtain that

$$\lim_{\Delta\to0}\frac{1}{\Delta^2}\int_0^{\Delta}(\frac{\Delta}{2}-y)f(u_\Delta(x-j\Delta)+y)\,dy=0$$

i.e., the numerator of (A1) converges to zero. Since the denominator converges to $f(x)\neq 0$, we conclude that

$$\lim_{\Delta\to0}\frac{m_\Delta(x-j\Delta)-c_\Delta(x-j\Delta)}{\Delta}=0$$

which completes the proof of the lemma. □

Proof of Lemma 2: Let f be a continuous a.e. pdf. Let us define

$$W_\Delta(x)\triangleq\frac{\left(c_\Delta(x)-m_\Delta(x)\right)^2}{\Delta^2}.$$

We may then write

$$\lim_{\Delta\to0}\frac{E(C_\Delta-M_\Delta)^2}{\Delta^2}=\lim_{\Delta\to0}\int_{-\infty}^{\infty}W_\Delta(x)f(x)\,dx\,. \quad \text{(A2)}$$

To show that the limit above is zero, we will swap the limit and the integral using the bounded convergence theorem. We may view the integration as being with respect to the measure $\mu(E)\triangleq\int_E f(x)\,dx$ [11, p. 214], and then the integration is over a set of finite measure. Furthermore, since $|c_\Delta(x)-m_\Delta(x)|\le\frac{\Delta}{2}$ for all x and Δ, it follows that $0\le W_\Delta(x)\le\frac{1}{4}$ for all x and Δ. Hence, $W_\Delta(x)$ is uniformly bounded for all x and Δ. Therefore, using the bounded convergence theorem

$$\begin{aligned}\lim_{\Delta\to0}\int_{-\infty}^{\infty}W_\Delta(x)f(x)\,dx&=\int_{-\infty}^{\infty}\lim_{\Delta\to0}W_\Delta(x)f(x)\,dx\\ &=\int_S\lim_{\Delta\to0}W_\Delta(x)f(x)\,dx=0\end{aligned}$$

where S denotes the set over which f is continuous and positive, and where the last equality follows from Lemma 1. □

Proof of Lemma 15: Suppose f is positive and differentiable at x. As in the proof of Lemma A1, with $j=0$

$$\frac{m_\Delta(x)-c_\Delta(x)}{\Delta^2}=\frac{\frac{1}{\Delta^3}\int_0^{\Delta}(\frac{\Delta}{2}-y)f(u_\Delta(x)+y)\,dy}{\frac{1}{\Delta}\int_0^{\Delta}f(u_\Delta(x)+y)\,dy} \quad \text{(A3)}$$

where the limit, as $\Delta\to0$ of the denominator of (A3), equals $f(x)$, since f is continuous at x.

We consider then the numerator of (A3). We begin by expressing $f(x+z)$ using the derivative

$$f(x+z)=f(x)+zf'(x)+z\delta_x(z) \quad \text{(A4)}$$

where $\delta_x(z)$ is a quantity that goes to zero as $z\to0$. (Note that z may be either positive or negative.) Using (A4) to evaluate the numerator of (A3), we obtain

$$\begin{aligned}\frac{1}{\Delta^3}&\int_0^{\Delta}(\frac{\Delta}{2}-y)f(u_\Delta(x)+y)\,dy\\ &=\frac{1}{\Delta^3}\int_0^{\Delta}(\frac{\Delta}{2}-y)f(x+(u_\Delta(x)-x+y))\,dy\\ &=\frac{1}{\Delta^3}\int_0^{\Delta}(\frac{\Delta}{2}-y)f(x)\,dy\\ &\quad+\frac{1}{\Delta^3}\int_0^{\Delta}(\frac{\Delta}{2}-y)(u_\Delta(x)-x+y)f'(x)\,dy\\ &\quad+\frac{1}{\Delta^3}\int_0^{\Delta}(\frac{\Delta}{2}-y)(u_\Delta(x)-x+y)[\delta_x(u_\Delta(x)-x+y)]\,dy\\ &=0-\frac{f'(x)}{12}\\ &\quad+\frac{1}{\Delta^3}\int_0^{\Delta}(\frac{\Delta}{2}-y)(u_\Delta(x)-x+y)[\delta_x(u_\Delta(x)-x+y)]dy\end{aligned}$$

$$= -\frac{f'(x)}{12} + \frac{1}{\Delta^3}\int_0^{\Delta}(\frac{\Delta}{2} - y)(u_\Delta(x) - x)[\delta_x(u_\Delta(x) - x + y)]dy$$
$$+ \frac{1}{\Delta^3}\int_0^{\Delta}(\frac{\Delta}{2} - y)y[\delta_x(u_\Delta(x) - x + y)]\,dy. \quad \text{(A5)}$$

It remains to show that the last two integrals in (A5) converge to zero as $\Delta \to 0$. Let $\varepsilon > 0$ be given. Since $(u_\Delta(x) - x + y) \in (-\Delta, \Delta]$ when $y \in [0, \Delta]$, and since $\delta_x(z) \to 0$ as $z \to 0$, it follows that for all sufficiently small Δ, $|\delta_x(u_\Delta(x) - x + y)| < \varepsilon$ for all $y \in [0, \Delta]$. Since $-\Delta < (u_\Delta(x) - x) \le 0$, it is not hard to see that

$$\frac{1}{\Delta^3}\left|\int_0^{\Delta}(\frac{\Delta}{2} - y)(u_\Delta(x) - x)[\delta_x(u_\Delta(x) - x + y)]\,dy\right| < \frac{\varepsilon}{4}$$

and that

$$\frac{1}{\Delta^3}\left|\int_0^{\Delta}(\frac{\Delta}{2} - y)y[\delta_x(u_\Delta(x) - x + y)]\,dy\right| < \frac{\varepsilon}{8}.$$

Since ε is arbitrary, it follows that

$$\lim_{\Delta\to 0}\frac{1}{\Delta^3}\int_0^{\Delta}(\frac{\Delta}{2} - y)f(u_\Delta(x) + y)\,dy = -\frac{f'(x)}{12}.$$

By combining this with the fact that the limit of the denominator of (A3) equals $f(x)$, we obtain that

$$\lim_{\Delta\to 0}\frac{m_\Delta(x) - c_\Delta(x)}{\Delta^2} = -\frac{f'(x)}{12f(x)}. \qquad \square$$

Lemma A2: Let f, x, $\Delta > 0$, and $S \ge 0$ be such that f is a continuous and piecewise differentiable function on $(x - \Delta, x + \Delta)$ and $|f'(y)| \le S$ for all $y \in (x - \Delta, x + \Delta)$ where f' exists. Then for any offset function

$$|G_\Delta(x)| \le 12(|x| + \Delta)S$$

where $G_\Delta(x)$ is as given in Definition 4.

Proof: First note that if $S = 0$, then the lemma holds trivially, since $G_\Delta(x) = 0$ (even if $f = 0$ over the interval $(x - \Delta, x + \Delta)$, since we recall that by convention $c_\Delta(x) = m_\Delta(x)$ in such a case). Suppose then that $S > 0$. We begin by writing

$$|G_\Delta(x)| = \frac{|m_\Delta^2(x) - c_\Delta^2(x)|}{\Delta^2/6}f(x)$$
$$= 6|m_\Delta(x) + c_\Delta(x)|\frac{|m_\Delta(x) - c_\Delta(x)|}{\Delta^2}f(x)$$
$$\le 12(|x| + \Delta)\frac{|m_\Delta(x) - c_\Delta(x)|}{\Delta^2}f(x).$$

It remains to show $\frac{|m_\Delta(x) - c_\Delta(x)|}{\Delta^2}f(x) \le S$. We do this assuming $c_\Delta(x) \le m_\Delta(x)$. The proof is essentially the same when the reverse inequality holds. There are two cases to consider.

1. $f(u_\Delta(x)) < S\Delta$: Since f is piecewise differentiable and $|f'(y)| \le S$ for almost all $y \in (u_\Delta(x), u_\Delta(x) + \Delta)$, we have $f(x) \le f(u_\Delta(x)) + S\Delta < 2S\Delta$, and consequently
$$\frac{|m_\Delta(x) - c_\Delta(x)|}{\Delta^2}f(x) < \frac{\Delta/2}{\Delta^2}2S\Delta = S.$$
2. $f(u_\Delta(x)) \ge S\Delta$: Define
$$g(y) \triangleq f(u_\Delta(x)) - S(y - u_\Delta(x))$$
for $y \in (u_\Delta(x), u_\Delta(x) + \Delta)$. From Lemma A3, which appears next, it follows that $c_\Delta^g(x) \le c_\Delta^f(x)$. Thus,
$$\frac{|m_\Delta(x) - c_\Delta^f(x)|}{\Delta^2}f(x) \le \frac{|m_\Delta(x) - c_\Delta^g(x)|}{\Delta^2}f(x).$$
$c_\Delta^g(x)$ can be simplified to
$$c_\Delta^g(x) = u_\Delta(x) + \frac{\Delta}{2} - \frac{\frac{S\Delta^2}{12}}{f(u_\Delta(x)) - \frac{S\Delta}{2}}.$$
It follows that
$$\frac{|m_\Delta(x) - c_\Delta^g(x)|}{\Delta^2}f(x) = \frac{S}{12}\frac{1}{f(u_\Delta(x)) - \frac{S\Delta}{2}}f(x)$$
$$\le \frac{S}{12}\frac{f(u_\Delta(x)) + S\Delta}{f(u_\Delta(x)) - \frac{S\Delta}{2}}$$
$$= \frac{S}{12}\frac{k + 1}{k - 1/2} \le \frac{S}{12}4 < S$$
where $k \triangleq \frac{f(u_\Delta(x))}{S\Delta} \ge 1$ implies $(k + 1)/(k - 1/2) \le 4$. This completes the proof that $\frac{|m_\Delta(x) - c_\Delta(x)|}{\Delta^2}f(x) \le S$ and, consequently, the proof of the lemma. $\square$

Lemma A3: Let f be a continuous and piecewise differentiable function on $(u, u + \Delta)$. Let also $f(u) \ge S\Delta$ and $|f'(x)| \le S$ for all $x \in W \cap (u, u + \Delta)$ for some $S > 0$ and $\Delta > 0$, where W is the set over which f is differentiable. Let $g(x) \triangleq f(u) - (x - u)S$. Then $c_{\Delta,u}^g \le c_{\Delta,u}^f$.

Proof: First note that $g(u) = f(u)$ and g decreases as sharply as possible among functions that satisfy the derivative constraint. If $g = f$ on $(u, u + \Delta)$, the lemma holds trivially. Suppose then that $g \ne f$ on a subset of $(u, u + \Delta)$ with positive measure. Let $h \triangleq f - g$, or, equivalently, $f = g + h$. Observe that for $x \in W \cap (u, u + \Delta)$, the fact that $|f'(x)| \le S$ implies $h'(x) \ge 0$ and $h(x) \ge 0$. From the definition of centroid we may write

$$c_u^f = c_u^g\frac{\int_u^{u+\Delta}g(x)\,dx}{\int_u^{u+\Delta}f(x)\,dx} + c_u^h\frac{\int_u^{u+\Delta}h(x)\,dx}{\int_u^{u+\Delta}f(x)\,dx}.$$

Thus, c_u^f is a weighted average of c_u^g and c_u^h. Since g is a strictly decreasing function and h is an increasing function (though not necessarily strictly increasing), it is easy to see that $c_u^g < u + \frac{\Delta}{2} < c_u^h$. It follows that $c_u^f > c_u^g$ since c_u^f is the average of c_u^g and something larger. $\square$

Lemma A4: Let f be a nonnegative, continuous and piecewise differentiable function such that

$$\lim_{\substack{x\to-\infty \\ x\in W}} f'(x) = 0 \quad \text{and} \quad \lim_{\substack{x\to\infty \\ x\in W}} f'(x) = 0$$

where W denotes the set over which f is differentiable. Let also $\int_{-\infty}^{\infty} f(x)\,dx < \infty$. Then

$$\lim_{x\to-\infty} f(x) = 0 \quad \text{and} \quad \lim_{x\to\infty} f(x) = 0.$$

Proof: We will show that $\lim_{x\to\infty} f(x) = 0$. The other case follows in a similar way. Let

$$M \triangleq \int_{-\infty}^{\infty} f(x)\,dx < \infty.$$

Let $\varepsilon > 0$ be given. Set $m = \frac{\varepsilon^2}{8M}$. There exists $T_1 > 0$ such that $|f'(x)| \le m$ for all $x \in (T_1, \infty) \cap W$. Since $\int_{-\infty}^{\infty} f(x)\,dx < \infty$, it follows that there exists $T_2 > T_1$ such that $f(T_2) \le \frac{\varepsilon}{2}$. Suppose there exists $T_3 > T_2$ such that $f(T_3) \ge \varepsilon$. Then

$$f(x) \ge \begin{cases} (x - T_3)m + \varepsilon, & T_3 - \frac{\varepsilon}{2m} \le x \le T_3 \\ 0, & T_2 \le x < T_3 - \frac{\varepsilon}{2m}. \end{cases}$$

Note that $T_3 - T_2 \ge \frac{\varepsilon}{2m}$. It now follows that

$$\int_{T_2}^{T_3} f(x)\,dx \ge \int_{T_3 - \frac{\varepsilon}{2m}}^{T_3} ((x - T_3)m + \varepsilon)\,dx = \int_0^{\frac{\varepsilon}{2m}} (\varepsilon - my)\,dy = \frac{3\varepsilon^2}{8m} = 3M$$

where the last equality follows from recalling that $m = \frac{\varepsilon^2}{8M}$. The above contradicts the fact that $\int_{-\infty}^{\infty} f(x)\,dx = M$. Therefore, $f(x) < \varepsilon$ for all $x > T_2$. Since ε is arbitrary, it follows that $\lim_{x\to\infty} f(x) = 0$. □

Lemma A5: Let f be a nonnegative, continuous, and piecewise differentiable function such that

$$\lim_{\substack{x\to-\infty \\ x\in W}} xf'(x) = 0 \quad \text{and} \quad \lim_{\substack{x\to\infty \\ x\in W}} xf'(x) = 0$$

where W denotes the set over which f is differentiable. Let also

$$\int_{-\infty}^{\infty} f(x)\,dx < \infty \quad \text{and} \quad \int_{-\infty}^{\infty} |x| f(x)\,dx < \infty.$$

Then

$$\lim_{x\to-\infty} xf(x) = 0 \quad \text{and} \quad \lim_{x\to\infty} xf(x) = 0.$$

Proof:

$$\lim_{\substack{x\to-\infty \\ x\in W}} xf'(x) = 0 \quad \text{and} \quad \lim_{\substack{x\to\infty \\ x\in W}} xf'(x) = 0$$

imply that

$$\lim_{\substack{x\to-\infty \\ x\in W}} f'(x) = 0 \quad \text{and} \quad \lim_{\substack{x\to\infty \\ x\in W}} f'(x) = 0.$$

Thus, applying Lemma A4 to the function f, we obtain that

$$\lim_{x\to-\infty} f(x) = 0 \quad \text{and} \quad \lim_{x\to\infty} f(x) = 0.$$

Next, let

$$g(x) \triangleq \begin{cases} xf(x), & x \ge 0 \\ -xf(x), & x < 0 \end{cases}$$

and obtain

$$g'(x) = \begin{cases} f(x) + xf'(x), & x \in [0,\infty) \cap W \\ -f(x) - xf'(x), & x \in (-\infty, 0) \cap W \end{cases}$$

Since

$$\lim_{x\to-\infty} f(x) = 0, \qquad \lim_{x\to\infty} f(x) = 0$$
$$\lim_{\substack{x\to-\infty \\ x\in W}} xf'(x) = 0, \quad \text{and} \quad \lim_{\substack{x\to\infty \\ x\in W}} xf'(x) = 0$$

it follows that

$$\lim_{\substack{x\to-\infty \\ x\in W}} g'(x) = 0 \quad \text{and} \quad \lim_{\substack{x\to\infty \\ x\in W}} g'(x) = 0.$$

Finally, since g is also nonnegative, continuous, piecewise differentiable, and integrable, we may apply Lemma A4 to it, and obtain that $\lim_{x\to-\infty} g(x) = 0$ and $\lim_{x\to\infty} g(x) = 0$, which is equivalent to

$$\lim_{x\to-\infty} xf(x) = 0 \quad \text{and} \quad \lim_{x\to\infty} xf(x) = 0.$$ □

Acknowledgment

The authors would like to thank the anonymous reviewers for their valuable comments.

References

[1] W. R. Bennett, “Spectra of quantized signals,” *Bell Syst. Tech. J.*, vol. 27, pp. 446–472, Jul. 1948.
[2] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*, 2nd ed. Upper Saddle River, NJ: Prentice-Hall, 1999.
[3] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing*, 3rd ed. Upper Saddle River, NJ: Prentice-Hall, 1996.
[4] A. Gersho, “Principles of quantization,” *IEEE Trans. Circuits Syst.*, vol. CAS-25, no. 7, pp. 427–436, Jul. 1978.
[5] T. Linder, “On Asymptotic Quantization Theory,” Ph.D. dissertation, Hungarian Academy of Sciences, Budapest, 1992.
[6] R. C. Wood, “On optimum quantization,” *IEEE Trans. Inf. Theory*, vol. IT-15, no. 2, pp. 248–252, Mar. 1969.
[7] B. Widrow, “Statistical analysis of amplitude-quantized sampled-data systems,” *Trans. AIEE*, pp. 555–568, Jan. 1961.
[8] A. B. Sripad and D. L. Snyder, “A necessary and sufficient condition for quantization errors to be uniform and white,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-25, no. 5, pp. 442–448, Oct. 1977.
[9] H. Viswanathan and R. Zamir, “On the whiteness of high-resolution quantization errors,” *IEEE Trans. Inf. Theory*, vol. 47, no. 5, pp. 2029–2038, Jul. 2001.
[10] T. Linder and K. Zeger, “Asymptotic entropy-constrained performance of tessellating and universal randomized lattice quantization,” *IEEE Trans. Inf. Theory*, vol. 40, no. 2, pp. 575–579, Mar. 1994.
[11] P. Billingsley, *Probability and Measure*, 3rd ed. New York: Wiley, 1995.
[12] L. Kronecker, “Naherungsweise ganzzahilge auflosung linearer gleichungen,” *Preuss. Akad. Wiss.*, pp. 1179–1193, 1271–1299, 1884. Reprint: K. Hensel, *L. Kronecker Werke*, vol. III(1), pp. 47-109. Leipzig, Germany: Teubner, 1899.
[13] K. Peterson, *Ergodic Theory*. New York: Cambridge Univ. Press, 1983.
[14] P. L. Tchebychef, “Sur une question arithmetique,” *Denkschr. Akad. Wiss. St. Petersburg*, no. 4, 1866. Reprint: A. Markoff and N. Sonin, *Oeuvres de P. L. Tchebychef*, vol. I, pp. 637-684. New York: Chelsea, 1962.