Quantization optimized with respect to the Haar basis

Shu Nakamura

January 12, 2021

Abstract. We propose a method of data quantization of finite discrete-time signals which optimizes the error estimate of low frequency Haar coefficients. We also discuss the error/noise bounds of this quantization in the Fourier space. Our result shows one can quantize any discrete-time analog signal with high precision at low frequencies. Our method is deterministic, and it employs no statistical arguments, nor any probabilistic assumptions.

1. Introduction

It has been known for a long time that the fixed bit depth quantization can encode higher bit resolution for low frequencies, possibly by allowing higher level of high frequency noise (see, e.g., [2], [3] Section 4.9). However, as far as the author is aware of, the most research results so far are either empirical or based on statistical analysis, which essentially assume the signal is a random noise. Here we propose a procedure, partly inspired by the discrete wavelet theory (see, e.g., [1]), to construct a quantization of signals, which optimize the error with respect to the Haar coefficients. Then we show that this construction also gives good low frequency noise bounds in the Fourier space. This argument is completely deterministic, and the estimates holds for all the signals without any statistical assumptions. We briefly discuss improvements by assuming randomness of the signal in Remark 1.5.

We consider analog data on a finite time sequence as follows. We set the size of the time sequence as $2^N$, with $N \in \mathbb{N}$. We denote the time sequence set by

$$
\Omega = \left\{ t[n] = -\frac{1}{2} + \frac{2n - 1}{2^{N+1}} \mid n = 1, \ldots, 2^N \right\} \subset [-1/2, 1/2].
$$

We set $X = \ell^2(\Omega : \mathbb{R}) = \mathbb{R}^\Omega$ as our analog data space, and we denote

$$
f = f[n] = f(t[n]) \in [-1/2, 1/2], \quad n = 1, \ldots, 2^N
$$

for $f \in X$. For $f, g \in X$, we introduce an inner product $\langle f, g \rangle$ by

$$
\langle f, g \rangle = \langle f, g \rangle_X = \frac{1}{|\Omega|} \sum_{t \in \Omega} f(t) g(t) = \frac{1}{2^N} \sum_{n=1}^{2^N} f[n]^* g[n]
$$

where $|\Omega| = 2^N$. We also denote the quantized data space by $X_Q = \ell^2(\Omega : \mathbb{Z}) = \mathbb{Z}^\Omega \subset X$.

2020 Mathematics Subject Classification. 65G40, 68P01.
We now recall the definition of the Haar basis. We use the notation:

\[ \hat{\Omega} = \{(0, 1) \} \cup \{(k, j) \mid k = 1, \ldots, N, j = 1, \ldots, 2^{k-1}\} \]

is the Haar index sets, and we set the index points by

\[ P[k, j] = -\frac{1}{2} + \frac{2j - 1}{2^k}, \quad (k, j) \in \hat{\Omega}. \]

We set

\[ \text{Haar}[0, 1](t) = 1, \quad \text{Haar}[1, 1](t) = \begin{cases} 1, & \text{if } 0 < t < 1/2, \\ -1, & \text{if } -1/2 < t < 0, \\ 0, & \text{otherwise}, \end{cases} \]

and for \( k = 2, \ldots, N \), we set

\[ \text{(1.1)} \quad \text{Haar}[k, j](t) = 2^{(k-1)/2}\text{Haar}[1, 1](2^{k-1}(t - P[k, j])), \quad j = 1, \ldots, 2^{k-1}. \]

We consider \( \text{Haar}[k, j](t) \) as functions on \( \Omega \), i.e., \( \text{Haar}[\cdot, \cdot] \in \ell^2(\Omega) \). It is well-known that the Haar basis \( \{\text{Haar}[k, j] \mid (k, j) \in \hat{\Omega}\} \) is an orthonormal basis of \( \ell^2(\Omega) \). For \( f \in X \), we denote the coefficients of the expansion in the Haar basis by

\[ Hf[k, j] = \langle \text{Haar}[k, j], f \rangle, \quad (k, j) \in \hat{\Omega}, \]

and we write the *Haar transform* of \( f \) by

\[ Hf = \{Hf[k, j]\}_{k, j} \in \ell^2(\hat{\Omega}). \]

We recall \( H \) is a unitary map from \( X = \ell^2(\Omega) \) to \( \hat{X} = \ell^2(\hat{\Omega}) \), where the inner product on \( \hat{X} \) is defined as usual (without weight):

\[ \langle \phi, \psi \rangle_{\hat{X}} = \sum_{(k,j) \in \hat{\Omega}} \phi[k, j]^* \psi[k, j], \quad \phi, \psi \in \hat{X}. \]

Now we can state our main result.

**Theorem 1.** Let \( f \in X = \mathbb{R}^\Omega \). Then there exists \( g \in X_Q = \mathbb{Z}^\Omega \) such that

\[ |Hf[0, 1] - Hg[0, 1]| \leq 2^{-N-1}. \]

and

\[ |Hf[k, j] - Hg[k, j]| \leq 2^{-N+(k-1)/2}, \quad (j, k) \in \hat{\Omega}, k \geq 1, \]

Moreover,

\[ \sup_{t \in \Omega} |f(t) - g(t)| \leq 1 - 2^{-N-1} < 1. \]

**Remark 1.1.** By the construction in the proof, we learn \( g \) is almost surely unique, in the sense of the Lebesgue measure, and thus the choice is optimal in the sense of (1.2) and (1.3). If the original data is discrete already, we may add very small noise/randomness to make it almost surely unique. This is actually equivalent to employ the randomized choice, and the results are independent of the choice of \( g \) satisfying these conditions, in any case.

**Remark 1.2.** For \( f \in X \), we write the *simple quantization* by

\[ \tilde{f}[n] = k \in \mathbb{Z} \quad \text{such that } |f[n] - k| \leq \frac{1}{2}. \]
Then, naturally, we have

\[ \sup_{t \in \Omega} |f(t) - \bar{f}(t)| \leq \frac{1}{2}, \]

which is better than (1.4). On the other hand, we only have \(|Hf[0, 1] - H\bar{f}[0, 1]| \leq 1/2\), and

\[ |Hf[k, j] - H\bar{f}[k, j]| \leq \frac{1}{2} \cdot 2^{-(k-1)/2}, \quad (j, k) \in \tilde{\Omega}, k \geq 1, \]

which are worse than (1.3) and (1.2), especially if \(k\) is small.

By virtue of (1.4), we can easily show the following result on finite value quantization. Let \(\ell, m \in \mathbb{Z}\), \(\ell + 2 < m\), and we denote \(I = \{\ell, \ell + 1, \ldots, m\}\), and \(\tilde{I} = [\ell + 1, m - 1]\).

**Corollary 2.** Suppose \(f \in X\) with \(\text{Ran}[f] = \{f(t) \mid t \in \Omega\} \subset \tilde{I}\), then there is \(g \in I^\Omega \subset X_Q\) which satisfies the conditions of Theorem 1, (1.2), and (1.3).

**Remark 1.3.** Corollary 2 implies we can allow the analog signal to have values in an interval of size \(m - \ell - 2\). If the resolution \(m - \ell\) is smaller or equal to 2, then this result is useless; For example, the so-called one-bit quantization case, \(m - \ell = 1\) and hence \(\tilde{I} = \emptyset\).

In such cases, we can still use the method combined with a simpler quantization method, for example the pulse width modulation (PWM) to increase the apparent resolution, and obtain \(g\) which enjoy the properties (1.3) for small \(k\), but little control for \(k\) close to \(N\).

Now we consider the error/noise of the quantization in the Fourier domain. We first fix notation of the discrete Fourier transform. We denote the Fourier variable space by

\(\Xi = \{-2^{N-1} + 1, \ldots, 2^{N-1}\} = (-2^{N-1}, 2^{N-1}] \cap \mathbb{Z}\),

and we write

\[ \text{Fourier}[^{\xi}][f](t) = e^{i\omega_0 t \xi}, \quad t \in [-1/2, 1/2], \quad \xi \in \Xi, \]

where \(\omega_0 = 2\pi\) and \(i = \sqrt{-1}\). We denote the discrete Fourier transform by

\[ \mathcal{F}f[^{\xi}][f] = \langle \text{Fourier}[^{\xi}][f], f \rangle = \frac{1}{2N} \sum_{n=1}^{2N} e^{-i\omega_0 n^{\xi} f[n]} \]

for \(f \in X = \ell^2(\Omega)\). The quantization noise in the Fourier space can be estimated as follows:

**Theorem 3.** Let \(f \in X\) and \(g \in X_Q\) as in Theorem 1. Then

\[ |\mathcal{F}f[0] - \mathcal{F}g[0]| \leq 2^{-N-1}, \]

and for \(\xi \in \Xi, \xi \neq 0\),

\[ |\mathcal{F}f[^{\xi}][f] - \mathcal{F}g[^{\xi}][f]| \leq \sum_{k=1}^{N} 2^{-2N+2(k-1)} \frac{1 - \cos(2^{-k} \cdot 2\pi \xi)}{|\sin(2^{-N} \pi \xi)|}. \]

In particular,

\[ |\mathcal{F}f[^{\xi}][f] - \mathcal{F}g[^{\xi}][f]| \leq \frac{N \pi^2}{2^{N+1} \pi |\xi|}, \quad \xi \in \Xi, \xi \neq 0. \]

**Remark 1.4.** Since we have an apriori bound \(|\mathcal{F}f[^{\xi}][f] - \mathcal{F}g[^{\xi}][f]| \leq 1\), (1.7) is useful only when \(|\xi| \ll 2^N\). If \(|\xi| \ll 2^N\), this gives us a very low noise floor, i.e., high resolution in the low frequencies. If we use the simple quantization, we only have

\[ |\mathcal{F}f[^{\xi}][f] - \mathcal{F}g[^{\xi}][f]| \leq \frac{1}{2}, \quad \xi \in \Xi, \]
and it is not possible to improve it in general.

Remark 1.5. Our results are completely deterministic, and these inequalities holds for any signal \( f \in X \). If we suppose the signal is very random, we can use statistical argument to show that the quantization noise can be much smaller.

We recall the argument for the simple quantization case. If \( \{ f[n] - f[n] \mid n = 1, \ldots, 2^N \} \) are identically distributed independent random variables, then for each \( \xi \in \Xi \), we have

\[
\mathbb{E}[|F f[\xi] - F f[\xi]|^2] = O(2^{-N}),
\]

where \( \mathbb{E}[:]= \) is the expectation, and hence we expect \( |F f[\xi] - F f[\xi]| = O(2^{-N/2}) \). This random assumption is unrealistic in applications, and usually high frequency random noise (dithering) is added to achieve the approximate independence, but with higher level of high frequency noise.

In our setting, if \( \{ Hf[j, k] - Hg[k, j] \mid (k, j) \in \hat{\Omega} \} \) are independent random variables, we can show

\[
\mathbb{E}[(F f[\xi] - F g[\xi])^2] \leq C 2^{-2N} |\xi|^2,
\]

and hence we expect \( |F f[\xi] - F g[\xi]| = O(2^{-N} |\xi|) \), which improves (1.7). We can also show

\[
\mathbb{E}[(F f[\xi] - F g[\xi])^2] \leq C 2^N |\xi|^{-2},
\]

and hence we expect \( |F f[\xi] - F g[\xi]| = O(2^{N/2} / |\xi|) \), which is useful for \( |\xi| \gg 2^{N/2} \). The correlations between Haar coefficients \( \{ H(f - g)[k, j] \mid (k, j) \in \Omega \} \) might usually be smaller than that of \( \{ f[n] - f[n] \mid n = 1, \ldots, 2^N \} \), but it is not obvious, and we might need to add some noise to achieve some approximate independence.

2. Proof of Theorem I

We first recall the expansion of data in the Haar basis: For \( f \in \ell^2(\Omega) \), we have

\[
f(t) = \sum_{(k,j) \in \hat{\Omega}} Hf[k,j] \text{Haar}[k,j](t), \quad t \in \Omega.
\]

We write down \( Hf[k,j] \) more explicitly. We denote the Haar domains by

\[
D[k,j] = \left( -\frac{1}{2} + \frac{j-1}{2^k}, -\frac{1}{2} \right) \cup \Omega,
\]

for \( k = 0, 1, \ldots, N, j = 1, \ldots, 2^k \). We note \( D[N,j] = \{ t[j] \}, j = 1, \ldots, 2^N \). We note

\[
D[k-1,j] = \left( -\frac{1}{2} + \frac{j-1}{2^{k-1}}, -\frac{1}{2} \right) \cap \Omega = \left( -\frac{1}{2} + \frac{2j-2}{2^k}, -\frac{1}{2} + \frac{2j}{2^k} \right) \cap \Omega
\]

(2.1) for \( k = 1, \ldots, N, j = 1, \ldots, 2^{k-1} \). We also note \( \text{Haar}[k,j] = \pm 2^{(k-1)/2} \) on \( D[k,2j] \) and \( D[k,2j-1] \), respectively, and \( \text{Haar}[k,j] = 0 \) otherwise. Thus we have

\[
Hf[k,j] = 2^{-N+(k-1)/2} \left( \sum_{t \in D[k,2j]} f(t) - \sum_{t \in D[k,2j-1]} f(t) \right)
\]

(2.2) for \( (k,j) \in \hat{\Omega}, k \geq 1 \). We also recall \( Hf[0,1] = 2^{-N} \sum_{t \in \Omega} f(t) \).

Inspired by these computations, we define the total value of \( f \in X \) in \( D[k,j] \) by

\[
Vf[k,j] = \sum_{t \in D[k,j]} f(t), \quad k = 0, 1, \ldots, N, \ j = 1, \ldots, 2^k.
\]
Then we have, by (2.2) and (2.1), respectively,

\begin{equation}
Hf[k, j] = 2^{-N+(k-1)/2}(Vf[k, 2j] - Vf[k, 2j - 1]),
\end{equation}

for \(k \geq 1\), and

\begin{equation}
Vf[k - 1, j] = Vf[k, 2j] + Vf[k, 2j - 1].
\end{equation}

In the following, for a given \(f \in X\), we will decide \(Vg[k, j]\) inductively on \(k\), and show that there exists \(g \in X_Q\) such that it satisfies these conditions, i.e., we construct \(G[k, j]\) so that \(G[k, j] = Vg[k, j]\). At first, we choose \(G[0, 1] \in Z\) so that

\begin{equation}
|Hf[0, 1] - 2^{-N}G[0, 1]| \leq 2^{-N-1}.
\end{equation}

Such \(G[0, 1]\) is unique except for a measure zero values of \(Hf[0, 1]\).

Then, inspired by (2.3) and (2.4), we find \(G[1, 1], G[1, 2] \in Z\) so that

\begin{equation}
G[1, 1] + G[1, 2] = G[0, 1]
\end{equation}

and

\begin{equation}
|Hf[1, 1] - 2^{-N}(G[1, 2] - G[1, 1])| \leq 2^{-N}.
\end{equation}

Such \(G[1, 1]\) and \(G[1, 2] \in Z\) always exist and almost surely unique (up to measure zero sets), since the set of possible \(G[1, 2] - G[1, 1]\) values with the condition (2.6) is \(G[0, 1] - 2Z\). Repeating this procedure, we can construct \(G[k, j] \in Z\), \(k = 2, \ldots, N, j = 1, \ldots, 2^k\), so that

\begin{equation}
G[k - 1, j] = G[k, 2j] + G[k, 2j - 1] \quad \text{for } j = 1, \ldots, 2^{k-1},
\end{equation}

and

\begin{equation}
|Hf[k, j] - 2^{-N+(k-1)/2}(G[k, 2j] - G[k, 2j - 1])| \leq 2^{-N+(k-1)/2}
\end{equation}

for \(j = 1, \ldots, 2^{k-1}\). Now we set

\begin{equation}
g[j] = G[N, j],
\end{equation}

for \(j = 1, \ldots, 2^N\). Then it is easy to show from the above construction, in particular by (2.4) and (2.7), that

\begin{equation}
Vg[k, j] = G[k, j], \quad k \geq 0, j = 1, \ldots, 2^k.
\end{equation}

Recalling (2.3) and (2.8), we learn

\begin{equation}
|Hf[k, j] - Hg[k, j]| \leq 2^{-N+(k-1)/2}, \quad (k, j) \in \hat{\Omega}, k \geq 1.
\end{equation}

The estimate (1.2) follows immediately from (2.5).

Now it remains to show the uniform error bound (1.4). Since for each \(k\), \(H_{\text{haar}}[k, j]\) has disjoint support for \(j = 1, \ldots, 2^{N-1}\), we have

\[
\left| \sum_{j=1}^{2^{k-1}} (Hf[k, j] - Hg[k, j])H_{\text{haar}}[k, j](t) \right| \leq 2^{-N+(k-1)/2} \cdot 2^{(k-1)/2} = 2^{-N+k-1}
\]
for \( k \geq 1 \) and \( t \in \Omega \). By summing up these in \( k \), we learn
\[
|f(t) - g(t)| = \left| \sum_{(k,j) \in \Omega} (H_f[k,j] - H_g[k,j])\text{Haar}[k,j](t) \right|
\leq 2^{-N-1} + \sum_{k=1}^{N} 2^{-N+k-1} = 2^{-N-1} + 2^{-N}(2^N - 1)
= 1 - 2^{-N-1},
\]
which completes the proof. \( \square \)

### 3. Proof of Theorem 3

We first compute the Fourier coefficients of the Haar functions. We denote
\[
\text{FH}[\xi, k,j] = F(\text{Haar}[k,j])|\xi| = \langle \text{Fourier}[\xi, \text{Haar}[k,j] \rangle
\]
for \( \xi \in \Xi; (k,j) \in \Omega \). It is obvious that \( \text{FH}[0,0,1] = 1; \)

\[
\text{FH}[0,k,j] = 0 \quad \text{for} \quad (j,k) \in \Omega, k \geq 1; \quad \text{FH}[\xi, 0,1] = 0 \quad \text{for} \quad \xi \neq 0.
\]
In fact, we have \( \text{FH}[0,1] = Ff[0] \). In the following, we only consider the case \((k,j) \in \Omega, k \geq 1\). Recalling the definition of the Haar basis (1.1), we have
\[
\text{FH}[\xi, k,j] = \frac{2^{(k-1)}2^{N}}{2N} \sum_{n=1}^{2^N} e^{-i\omega_0t[n]\xi} \text{Haar}[1,1](2^{k-1}(t[n] - P[k,j]))
\]
\[
= \frac{2^{(k-1)}2^{N}}{2N} \sum_{n=1}^{2^N} e^{-i\omega_0t[n]+P[k,j]\xi} \text{Haar}[1,1](2^{k-1}t[n])
= \frac{2^{(k-1)}2^{N}}{2N} e^{-i\omega_0P[k,j]\xi} \left( \sum_{n=1}^{2^{N-k}} e^{-i\omega_02^{-N}(n-\frac{1}{2})} \right) - \sum_{n=1}^{2^{N-k}} e^{i\omega_02^{-N}(n-\frac{1}{2})} \right).
\]

We now compute
\[
\sum_{n=1}^{2^{N-k}} e^{-i\omega_02^{-N}(n-\frac{1}{2})} = \frac{e^{-i\omega_N2^{N-k}2^{-N}(n-\frac{1}{2})} - e^{-i\omega_N2^{-N}(n-\frac{1}{2})}}{2i\sin(\omega_N2^{-N}/2)),}
\]
where \( \omega_N = 2^{-N}\omega_0 \). Similarly, we have
\[
\sum_{n=1}^{2^{N-k}} e^{i\omega_02^{-N}(n-\frac{1}{2})} = -\frac{1 - e^{i\omega_N2^{N-k}}}{2i\sin(\omega_N2^{-N}/2)),}
\]
and hence
\[
\sum_{n=1}^{2^{N-k}} e^{-i\omega_02^{-N}(n-\frac{1}{2})} - \sum_{n=1}^{2^{N-k}} e^{i\omega_02^{-N}(n-\frac{1}{2})} = \frac{1 - \cos(\omega_N2^{-N-k}\xi)}{i\sin(\omega_N2^{-N}\xi/2))} = \frac{1 - \cos(\omega_02^{-k}\xi)}{i\sin(\omega_02^{-N}\xi/2))}.
\]

Thus we obtain
\[
\text{FH}[\xi, k,j] = \frac{2^{(k-1)/2}}{2N} e^{-i\omega_0P[k,j]\xi} \left( \frac{1 - \cos(\omega_02^{-k}\xi)}{i\sin(\omega_02^{-N}\xi/2))}.
\]
and

\[
|FH[\xi, k, j]| = \frac{2^{(k-1)/2}}{2^N} \frac{1 - \cos(\omega_0 2^{-k} \xi)}{|\sin(\omega_0 2^{-N} \xi/2)|} = \frac{2^{(k-1)/2}}{2^N} \cdot \frac{1 - \cos(2\pi \cdot 2^{-k} \xi)}{|\sin(2^{-N} \pi \xi)|},
\]

Now we fix \( f \in X \) and its quantization \( g \) as in Theorem \( \text{(1)} \) and we estimate the bound for \( F(f - g)[\xi] \), i.e., the noise bound in the Fourier domain. We recall

\[
F(f - g)[\xi] = \sum_{(j,k) \in \hat{\Omega}} FH[\xi, k, j] H(f - g)[k, j],
\]

and by Theorem \( \text{(1)} \) we know

\[
|H(f - g)[k, j]| \leq 2^{N+(k-1)/2}, \quad (k, j) \in \hat{\Omega}, k \geq 1.
\]

Combining these formulas with (3.1), we have

\[
|F(f - g)[\xi]| \leq \sum_{k=1}^N \sum_{j=1}^{2^{k-1}} |FH[\xi, k, j]| |H(f - g)[k, j]|
\]

\[
\leq \sum_{k=1}^N \sum_{j=1}^{2^{k-1}} 2^{-2N+k-1} \frac{1 - \cos(2\pi \cdot 2^{-k} \xi)}{|\sin(2^{-N} \pi \xi)|}
\]

\[
= \sum_{k=1}^N 2^{-2N+2(k-1)} \frac{1 - \cos(2\pi \cdot 2^{-k} \xi)}{|\sin(2^{-N} \pi \xi)|},
\]

which proves (1.6). We now recall the elementary inequalities:

\[
|\sin \theta| \geq \frac{2}{\pi} |\theta| \quad \text{for} \quad |\theta| \leq \frac{\pi}{2}; \quad 1 - \cos \theta \leq \frac{1}{2} \theta^2 \quad \text{for} \quad \theta \in \mathbb{R},
\]

and we have

\[
1 - \cos(2\pi \cdot 2^{-k} \xi) \leq \frac{1}{2} (2\pi \cdot 2^{-k} \xi)^2 \leq \frac{\pi^2}{2} |2^{-N} \xi|^2 = 2^{-N-2k} \pi^2 |\xi|.
\]

Combining this with (1.6), we have

\[
|F(f - g)[\xi]| \leq \sum_{k=1}^N 2^{-2N+2(k-1)} \cdot 2^N \cdot 2^{-2k} \pi^2 |\xi| = \frac{N \pi^2}{2N+2} |\xi|,
\]

which proves (1.7). The estimate (1.5) follows from (1.3), since \( Ff[0] = Hf[0, 1] \) and \( Fg[0] = Hg[0, 1] \).

\[\square\]

References

[1] Jensen, A., la Cour-Harbo, A.: Ripples in Mathematics. Springer Verlag, 2001.
[2] Lipshitz, S. P., Wannamaker, R. A., Vanderkooy, J.: Quantization and dither: a theoretical survey. J. Audio Eng. Soc. Vol. 40, No. 5, 1992, 355–375.
[3] Oppenheim, A. V., Schafer, R. W., Buck, J. R.: Discrete-Time Signal Processing (Second Edition), Prentice Hall, 1999.

Department of Mathematics, Faculty of Sciences, Gakushuin University, 1-5-1, Mejiro, Toshima, Tokyo, Japan 171-8588

Email address: shu.nakamura@gakushuin.ac.jp