Minimum Analog-Digital Quantization Resolution Requirements for Digital Communications Systems

Gareth B Middleton and Behnaam Aazhang
Rice University
Houston, Texas
Email: {gbmidd, aaz}@rice.edu

Abstract—This paper considers the effects of front-end quantization on the performance of digital communications systems. Instead of modeling the quantizer as a source of additive noise, we use a novel geometric approach to understand the interactions between the quantizer and received constellation points. We first find the minimum number of bins a uniform quantizer must possess in order to recover all symbols from a fading channel without error, then show that although all symbols passed through the minimum precision quantizer are recoverable at high SNR, they may not be equally reliable in realistic SNR regimes. We conclude by showing that, for the case of a uniform quantizer, the QAM modulation scheme is optimal in a symbol-reliability sense.

I. INTRODUCTION

Today’s advanced wireless systems use a variety of technologies to deliver high datarate communications. Multiple-input multiple-output (MIMO) schemes achieve a combination of higher reliability at faster rates [1], and ultrawide band (UWB) methods use a larger frequency band but operate only just above the noise floor [2].

Receivers in these systems are therefore inundated with data: in the UWB case, the high bandwidth necessitates a quantizer running in the gigahertz range, rapidly passing high precision samples to the receiver [3]. While the bandwidth may be smaller in the MIMO case, each antenna produces its own stream of high precision samples [4]. This flow of information from the analog-digital converters (ADCs) causes a data overload in the receiver, requiring large amounts of power and complex architectures in mobile, handheld systems [5].

We study a fundamental way to reduce the complexity of such a receiver: we limit the precision of the front end ADCs. Previous literature has shown that quantization of any kind can cause error floors in system performance [6], [7], [8], but many of these works have assumed that quantization may be modeled as an additive noise term with some density, usually taken to be uniform. In this paper, we present a fundamentally novel approach: we will consider the interaction between the quantizer and the constellation symbols from a geometric perspective.

Our contributions are three-fold. First, we find the minimum precision required for a system to operate without error over a fading channel at high SNR, regardless of modulation scheme. Second, we show that a front-end quantizer can cause inequality in the reliability of the symbols in a constellation, and third, we show that under the special case of a uniform quantizer, quadrature-amplitude modulation is optimal in terms of symbol reliability. We will conclude with simulation results and a discussion of future work.

II. SYSTEM MODEL

To focus on the quantization effect, we consider a SISO system as depicted in Fig. 1. We assume a perfect automatic gain control, such that \( g = \frac{1}{|h|} \), where \( h \) denotes a Rayleigh block-fading channel. The action of the AGC results in symbols being rotated by the channel, but not scaled. To counter this channel phase, we train the receiver with pilot symbols before each block. Inspired by real-world implementations, we
pass the inphase and quadrature channels through independent hard-limiting mid-rise uniform quantizers after downconversion, so that constellation points are presented to the receiver as shown in Fig. 2. The quantizer ranges between $[-1, 1]$, and the width of each quantization bin is denoted by $\bar{q}$. Since the quantizer is mid-rise (i.e., it does not have a “zero bin”), we require that it has an even number of bins. We will assume that the constellation symbols $m$ are indexed, and we will use $i$ and $j$ to reference specific constellation symbols.

III. MAIN RESULT

Here we present our main contribution: a minimum number of bins required to guarantee zero error performance at high SNR, over a fading channel. In such noise regimes, the primary effect of the channel is symbol rotation. As AWGN tends to zero, the channel phase can be corrected with arbitrarily low error using a pilot-symbol channel estimate. In our case, even at high SNR, the quantizer introduces error to channel estimate. The error will be related to the resolution of the quantizer, and will be bounded by a closed form function arising from the geometry of the symbol space. We will relate the estimation error to the resolution of the quantizer, and will be bounded by a closed form function arising from the geometry of the symbol space. We will relate the estimation error to the resolution of the quantizer, which we will then connect to the probability of symbol detection error, using the concept of an extended decision region. Last, we will show that there is some minimum precision above which the channel estimate is accurate enough to allow all symbols to be correctly recovered.

A. Channel Estimation Error

Lemma 1 (Channel Estimation Error): For the system described in Section II, the maximum channel estimation error caused by the quantizer is tightly bounded by

$$\theta_{emax} = \tan^{-1} \left( \frac{\text{Im}[c_i] + \frac{\bar{q}}{2}}{\text{Re}[c_i] - \frac{\bar{q}}{2}} \right) + \frac{\pi}{4}$$

(1)

where $c_i$ is the value of the bin through which a 45 degree line passes.

Proof:

Note that the effect of the quantizer is to represent all points within a bin by the value of the center of the bin. To find maximal phase error, we therefore seek to find the point in each bin with phase most different from the phase of the center of the bin, searching over all bins which may quantize the pilot symbol $m$ (i.e., all bins which quantize points of magnitude $|m|$).

We can relate arc-length to phase angle as $s = r\phi$. The maximum angle $\phi$ must therefore subtend the maximum arc length $s$. Therefore we must find the point which, when quantized, has been moved furthest along the arc defined by the magnitude of $m$. This occurs in the quantization bin situated at an angle $\frac{k\pi}{4}$, as shown in Figure 3. This is the effect of the largest linear distance introduced by the quantizer, $\bar{q}\sqrt{2}$, aligning with the maximal arc length on the arc defined by $|m|$. We finish by computing this angle.

Note that at high SNR, the maximal channel estimation error is purely a function of the quantization bin width $\bar{q}$. As $\bar{q} \to 0$, the error term $\theta_{emax} \to 0$ as well, as we would expect.

B. Decision Regions

Our system will estimate a channel phase $\hat{\theta}$ through the use of a pilot symbol transmitted with a known phase, but which will be quantized at the front-end of the receiver. From Lemma 1, we know that the phase estimate will be subject to quantization error, but bounded within the range $[-\theta_{emax}, \theta_{emax}]$. The receiver may now perform detection in either of two ways: rotating the received symbols by $-\hat{\theta}$ and then using the traditional decision regions, or by equivalently rotating the decision regions by $\hat{\theta}$ and detecting the symbols as they were received. We will proceed using the latter scheme, in which the receiver rotates its decision regions $R_i$ by $\hat{\theta}$. We will denote rotated decision regions with $R_{\theta i}$. We now define an extended decision region, denoted $R_{\theta e}^{\theta i}$:
The extended decision region for symbol $i$ is the union of all decision regions for symbol $i$ which have been rotated by any phase in the interval $[-\theta_{\text{emax}}, \theta_{\text{emax}}]$. In practice, given any estimated channel phase $\hat{\theta}$, the receiver will estimate the phase and rotate the decision regions such that $R^\theta_i \in R^{\theta_e}_i$. We have illustrated a set of extended decision regions in Figure 4, for an 8-PSK modulation scheme. When the phase estimate is known to fall within a bounded range, it is certain that the decision regions will be rotated by at most $\theta_{\text{emax}}$ in either a clockwise or counter-clockwise direction. Adjacent extended decision regions will overlap by an amount $2\theta_{\text{emax}}$, indicated in the figure.

We use extended decision regions to form a criterion for accurate detection at high SNR:

**Theorem 1 (Minimum Bins for Zero Error):** For the system described in Section II, zero error performance at infinitely high SNR may be guaranteed provided that

\[
Q[m_i] \notin R^{\theta_e}_i \quad \forall \ i \neq j.
\]

where $i,j$ index all symbols in the chosen constellation.

This is to say that zero-error performance is guaranteed if, after quantization, no symbol $i$ falls into the extended decision region of any another symbol $j$. In proving this result, we first by show why the condition is required, and then that it may be met.

**Proof:**

To show that this condition is required, we will establish a contradiction. Assume the system is operating with zero error, and further that $Q[m_i] \in R^{\theta_e}_j$. Now, suppose the channel phase estimate $\hat{\theta} \in [-\theta_{\text{emax}}, \theta_{\text{emax}}]$ is such that $Q[m_i] \in R^\theta_j \subset R^{\theta_e}_j$. The detector will recover symbol $j$ when symbol $i$ was transmitted, and the system has an error. This contradicts our original zero-error assumption, meaning that the Theorem 1 must be satisfied for zero-error performance. Now suppose the condition is met. If $Q[m_i] \notin R^{\theta_e}_j$, then $Q[m_i] \notin R^\theta_j$, which means that symbol $i$ will never be detected as some other symbol $j$, meaning that the system operates without error.

As $\bar{q} \to 0$, the quantizer becomes more precise. This implies that the minimum distance between any point and its quantized representation tends to zero, which further implies that the phase estimation error tends to zero. Thus

\[
\bar{q} \to 0 \quad \Rightarrow \quad \theta_{\text{emax}} \to 0
\]

\[
\Rightarrow \quad Q[m_i] \to m_i
\]

\[
\Rightarrow \quad R^{\theta_e}_i \to R^\theta_i
\]

Thus there exists some $\bar{q}_{\text{max}}$ such that our condition is met, since the decision regions converge to those which would exist without quantization, as the symbols converge to those which would be received without quantization.

![Fig. 5. An erroneous channel estimate can combine with the rounding effects of the quantizer to cause an error; here, the received symbol should have been detected as $m_{10}$, but is instead detected as $m_1$, indicating a lack of precision.](image-url)

### TABLE I

<table>
<thead>
<tr>
<th>M</th>
<th>PSK</th>
<th>QAM</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>10</td>
<td>4</td>
<td>16</td>
<td>18</td>
<td>5</td>
<td>12</td>
</tr>
<tr>
<td>32</td>
<td>36</td>
<td>6</td>
<td>64</td>
<td>72</td>
<td>7</td>
<td>28</td>
</tr>
<tr>
<td>256</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### C. Intuition

The result states that no symbol, after quantization, should fall within an overlapping decision region. This can be thought of as requiring that no symbol be subject to ambiguous detection. In this way, we have captured both the channel estimation error effect (through the extended decision regions) and also the “rounding” nature of the quantizer, by requiring that $Q[m_i]$ be defined to fall outside the extended decision regions. Figure 5 depicts a case in which minimum resolution does not exist; an erroneous phase estimate combines with the rounding effects of the quantizer to cause the transmitted symbol to be received in error.

### D. Application

Applying the result to common PSK and QAM systems of varying orders, we compute Table I. For any number of bins less than shown in the table, not all $Q[m_i]$ are guaranteed to fall in unambiguous decision regions, hence errors are possible, even at high SNR.

Notice that in the PSK case, the number of required bins scales roughly with the number of points in the constellation.

The linear scaling is largely due to the fact that, for a PSK...
constellation, the information is encoded in the phase of the signal, mandating an accurate phase recovery and requiring more bins of precision. In the case of QAM, note now that the bins required scales with the square-root of \(M\), reflecting the fact that in a QAM system, most information is encoded in the magnitude of the symbol, rather than in its phase alone.

**IV. Extended Results**

The previous result for minimum quantization resolution arose from a geometrical analysis of the system. Here we continue our study of the quantization impact on symbols from a geometric perspective, presenting first a discussion of how the uniform quantizer favors some symbols over others, and then showing that QAM is most immune to this effect.

**Lemma 2 (Effects on Symbol Reliability):** For the system described in Section II and employing a modulation scheme different from Quadrature-Amplitude (QAM), constellation symbols \(m_i\) suffer differing probabilities of error.

*Proof:*

As shown in Figure 2, the quantizer slices the I-Q plane. Decision regions not shaped as squares are therefore pixelated, with the result that some decision regions \(R_i\) are larger than others, \(R_j\). As such, when computing symbol error probabilities, the result of the integration of the AWGN bell over the larger decision regions \(R_i\) is greater than that over the smaller regions, meaning the symbols \(m_i\) are more reliable than those corresponding to smaller decision regions \(m_j\).

This lemma shows that even if sufficient resolution exists in order to recover each symbol, the quantizer causes differences in the reliability of constellation symbols at reasonable SNR. An example is shown in Figure 6 for 8-PSK. This result makes a case for tailoring the constellation design to the characteristics of the receiver quantization element. Below, we show that QAM is best suited to receivers with uniform quantizers.

**Lemma 3 (QAM Performance Invariant):** For a system \(y = Q[x+n]\) where \(x\) is drawn from an equiprobable M-ary square QAM constellation defined over \([-c, c]\) and \(Q[]\) is a uniform quantizer with at least \(b = \sqrt{M}\) levels of precision over \([-c, c]\), with \(b\) a power of 2, the system will experience the same symbol error rate as the system \(y = x+n\) when operating over an AWGN channel.

*Proof:*

Our method of proving this result will be to show that the quantizer boundaries and the decision region boundaries are the same. Therefore quantizer does not affect the mapping from the continuous received signal to the symbol decision.

The decision regions for a continuous QAM system are symmetric between the I and Q channels, so we will consider only the I channel. Since we have \(\sqrt{M}\) constellation points distributed evenly across 2c, the points will be positioned as \(m_i = -c + (2i+1)\frac{d}{2}\), where \(d = \frac{2c}{\sqrt{M}}\) and \(i \in \{0, 1, \ldots, M-1\}\).

We require \(\sqrt{M}\) decision regions, which imply \(\sqrt{M} + 1\) decision boundaries, with the outermost boundaries set as \(\pm\infty\). Since the signaling is equiprobable, all decision regions are the same width, and so the boundaries exist at \(b_i = -c + id\), for \(i \in \{0, 1, \ldots, M-2\}\).

Now let us look at the quantizer specification. The quantizer possesses \(2^L\) bins, spread uniformly over the range \([-c, c]\). Thus its boundaries will be at \(b_i^q = -c + iq\), where \(q = \frac{2c}{2^L}\) and \(i \in \{0, 1, \ldots, 2^L + 1\}\). The decision region used by the demodulator in the quantized case becomes the set of bins which lie in the original \(R_i\).

Let us suppose we have exactly \(\sqrt{M}\) quantizer levels, exactly equal to the number of decision regions. In this case, \(d = \frac{2c}{2^L} = \sqrt{M}\), so that the quantizer boundaries and the decision region boundaries are identical, that is

\[
b_i = b_i^q = -c + id = -c + iq. \quad (7)
\]

More specifically, for any point \(r\), if \(r \in R_i\) then \(Q[r] = c_i \in R_i\). Similarly, if \(x + n \in R_i\), then \(Q[x + n] = c_i \in R_i\). Therefore \(Q_i = R_i\) if \(b = \sqrt{M}\).

Now suppose we add a bit of precision, such that we have \(2\sqrt{M}\) bins distributed over \([-c, c]\). The bin width \(q' = \frac{2c}{2\sqrt{M}}\). As before, the quantizer boundaries are defined as \(b_i^q = -c + iq'\) for \(i \in \{0, 1, \ldots, 2^L + 1\}\). The decision region boundaries remain as they were, since we have the same modulation order.

Consider the quantization boundaries indexed by \(k = 2i\), where \(i \in \{0, 1, \ldots, 2^L-1\}\). The locations of these boundaries are

\[
b_i^{q'} = -c + id = -c + 2i\frac{2c}{2\sqrt{M}} = -c + i\frac{2c}{\sqrt{M}} = b_i \quad (8)
\]

Thus, every other quantization bin shares a boundary with a decision region. Intuitively this is expected, since in the case
where we had just as many bins as decision regions, they overlapped exactly. Here we have doubled the number of bins, halving their size but preserving every other boundary location. Thus now each decision region is composed of 2 bins, but the effective boundary remains the same. The same happens if we add another bit of precision: now we have 4 bins in each decision region, and so on.

The result is that if \( r \in R_i \), then \( Q[r] \in R_i \) also. Thus to understand the effects of noise in quantized systems, we need to integrate that part of the noise density corresponding to all other decision regions except the one of interest. Since we’ve shown that quantization does not affect the mapping from the continuum to the decision region, it has no effect on the performance of the system, and QAM is optimal.

V. Simulation Results
Here we present simulation results to confirm the previous contributions. We first consider a 16-QAM system, quantized with both 10 and 12 bins. Theorem 1 shows that 10 bins are too few for reliable performance (see Table I), while 12 should be sufficient. The results are shown in Figure 7, and as expected, quantizing with 10 bins causes an error floor to appear as SNR increases. If 12 bins are used, the error floor disappears, yielding the traditional waterfall curve.

We next verify Lemma 2, which suggests that some symbols are more reliable than others. Here we consider an 8-PSK system with 18 bins of precision, enough to guarantee perfect performance at high SNR, shown in Figure 6. We track the error for symbols 0 and 1, shown in Figure 8. As predicted, symbol 0 experiences more error, since it corresponds to a smaller decision region in the quantized \( I - Q \) plane for 8-PSK modulation. The inequality among the symbols worsens as SNR increases, approaching 1.5 dB as the SNR approaches 20 dB.

VI. Conclusions and Future Work
We have considered the problem of recovering constellation symbols after time domain signals have been quantized. We find the minimum resolution required for the receiver to recover all symbols accurately at high SNR, and then go on to show that the quantizer introduces inequality in symbol reliability, but that the square QAM scheme is optimal in a symbol reliability sense.

Our method of considering the quantizer from a geometrical perspective has yielded an important result: that the quantization error floor disappears for systems with some sufficient number of bins. However, we have not described the interaction of the quantizer on system performance in more realistic SNR regimes. Future work will take into account the performance of a system with an imperfect AGC, as well as an analysis of quantization effects on multicarrier systems.

REFERENCES