Chapter 3

Quantization

3.1 Introduction to quantization

The previous chapter discussed coding and decoding for discrete sources. Discrete sources are a subject of interest in their own right (for text, computer files, etc.) and also serve as the inner layer for encoding analog source sequences and waveform sources (see Figure 3.1). This chapter treats coding and decoding for a sequence of analog values. Source coding for analog values is usually called *quantization*. Note that this is also the middle layer for waveform encoding/decoding.

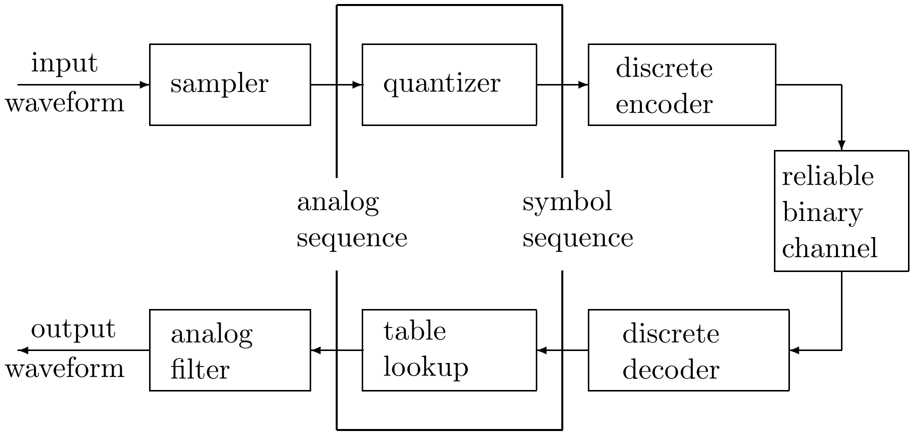


Figure 3.1: Encoding and decoding of discrete sources, analog sequence sources, and waveform sources. Quantization, the topic of this chapter, is the middle layer and should be understood before trying to understand the outer layer, which deals with waveform sources.

The input to the quantizer will be modeled as a sequence $U_1, U_2, \cdots$, of analog random variables (rv's). The motivation for this is much the same as that for modeling the input to a discrete source encoder as a sequence of random symbols. That is, the design of a quantizer should be responsive to the set of possible inputs rather than being designed for only a single sequence of numerical inputs. Also, it is desirable to treat very rare inputs differently from very common inputs, and a probability density is an ideal approach for this. Initially, $U_1, U_2, \ldots$ will be taken as independent identically distributed (iid) analog rv's with some given probability density function (pdf) $f_U(u)$.

A quantizer, by definition, maps the incoming sequence $U_1, U_2, \cdots$, into a sequence of discrete rv's $V_1, V_2, \cdots$, where the objective is that V_m, for each m in the sequence, should represent U_m with as little distortion as possible. Assuming that the discrete encoder/decoder at the inner layer of Figure 3.1 is uniquely decodable, the sequence $V_1, V_2, \cdots$ will appear at the output of the discrete encoder and will be passed through the middle layer (denoted 'table lookup') to represent the input $U_1, U_2, \cdots$. The output side of the quantizer layer is called a 'table lookup' because the alphabet for each discrete random variables V_m is a finite set of real numbers, and these are usually mapped into another set of symbols such as the integers 1 to M for an M symbol alphabet. Thus on the output side a look-up function is required to convert back to the numerical value V_m.

As discussed in Section 2.1, the quantizer output V_m, if restricted to an alphabet of M possible values, cannot represent the analog input U_m perfectly. Increasing M, *i.e.*, quantizing more finely, typically reduces the distortion, but cannot eliminate it.

When an analog rv U is quantized into a discrete rv V, the mean-squared distortion is defined to be $\mathsf{E}[(U-V)^2]$. Mean-squared distortion (often called mean-sqared error) is almost invariably used in this text to measure distortion. When studying the conversion of waveforms into sequences in the next chapter, it will be seen that mean-squared distortion is a particularly convenient measure for converting the distortion for the sequence into the distortion for the waveform.

There are some disadvantages to measuring distortion only in a mean-squared sense. For example, efficient speech coders are based on models of human speech. They make use of the fact that human listeners are more sensitive to some kinds of reconstruction error than others, so as, for example, to permit larger errors when the signal is loud than when it is soft. Speech coding is a specialized topic which we do not have time to explore (see, for example, [10]. However, understanding compression relative to a mean-squared distortion measure will develop many of the underlying principles needed in such more specialized studies.

In what follows, scalar quantization is considered first. Here each analog rv in the sequence is quantized independently of the other rv's. Next vector quantization is considered. Here the analog sequence is first segmented into blocks of n rv's each; then each n-tuple is quantized as a unit.

Our initial approach to both scalar and vector quantization will be to minimize mean-squared distortion subject to a constraint on the size of the quantization alphabet. Later, we consider minimizing mean-squared distortion subject to a constraint on the *entropy* of the quantized output. This is the relevant approach to quantization if the quantized output sequence is to be source-encoded in an efficient manner, *i.e.*, to reduce the number of encoded bits per quantized symbol to little more than the corresponding entropy.

3.2 Scalar quantization

A *scalar quantizer* partitions the set $\mathbb{R}$ of real numbers into M subsets $\mathcal{R}_1, \dots, \mathcal{R}_M$, called *quantization regions*. Assume that each quantization region is an interval; it will soon be seen why this assumption makes sense. Each region $\mathcal{R}_j$ is then represented by a *representation point* $a_j \in \mathbb{R}$. When the source produces a number $u \in \mathcal{R}_j$, that number is quantized into the point a_j. A scalar quantizer can be viewed as a function $\{v(u) : \mathbb{R} \to \mathbb{R}\}$ that maps analog real values u into discrete real values $v(u)$ where $v(u) = a_j$ for $u \in \mathcal{R}_j$.

An analog sequence $u_1, u_2, \dots$ of real-valued symbols is mapped by such a quantizer into the discrete sequence $v(u_1), v(u_2) \dots$. Taking $u_1, u_2 \dots$, as sample values of a random sequence $U_1, U_2, \dots$, the map $v(u)$ generates an rv V_k for each U_k; V_k takes the value a_j if $U_k \in \mathcal{R}_j$. Thus each quantized output V_k is a discrete rv with the alphabet $\{a_1, \dots, a_M\}$. The discrete random sequence $V_1, V_2, \dots$, is encoded into binary digits, transmitted, and then decoded back into the same discrete sequence. For now, assume that transmission is error-free.

We first investigate how to choose the quantization regions $\mathcal{R}_1, \dots, \mathcal{R}_M$, and how to choose the corresponding representation points. Initially assume that the regions are intervals, ordered as in Figure 3.2, with $\mathcal{R}_1 = (-\infty, b_1], \mathcal{R}_2 = (b_1, b_2], \dots, \mathcal{R}_M = (b_{M-1}, \infty)$. Thus an M-level quantizer is specified by $M-1$ interval endpoints, $b_1, \dots, b_{M-1}$, and M representation points, $a_1, \dots, a_M$.

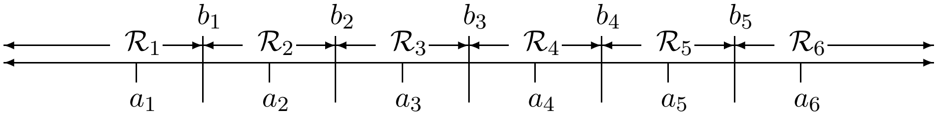


Figure 3.2: Quantization regions and representation points.

For a given value of M, how can the regions and representation points be chosen to minimize mean-squared error? This question is explored in two ways:

- Given a set of representation points $\{a_j\}$, how should the intervals $\{\mathcal{R}_j\}$ be chosen?
- Given a set of intervals $\{\mathcal{R}_j\}$, how should the representation points $\{a_j\}$ be chosen?

3.2.1 Choice of intervals for given representation points

The choice of intervals for given representation points, $\{a_j; 1 \le j \le M\}$ is easy: given any $u \in \mathbb{R}$, the squared error to a_j is $(u - a_j)^2$. This is minimized (over the fixed set of representation points $\{a_j\}$) by representing u by the closest representation point a_j. This means, for example, that if u is between a_j and a_{j+1}, then u is mapped into the closer of the two. Thus the boundary b_j between $\mathcal{R}_j$ and $\mathcal{R}_{j+1}$ must lie halfway between the representation points a_j and $a_{j+1}, 1 \le j \le M-1$. That is, $b_j = \frac{a_j + a_{j+1}}{2}$. This specifies each quantization region, and also shows why each region should be an interval. Note that this minimization of mean-squared distortion does not depend on the probabilistic model for $U_1, U_2, \dots$.

3.2.2 Choice of representation points for given intervals

For the second question, the probabilistic model for $U_1, U_2, \ldots$ is important. For example, if it is known that each U_k is discrete and has only one sample value in each interval, then the representation points would be chosen as those sample value. Suppose now that the rv's $\{U_k\}$ are iid analog rv's with the pdf $f_U(u)$. For a given set of points $\{a_j\}$, $V(U)$ maps each sample value $u \in \mathcal{R}_j$ into a_j. The mean-squared distortion (or mean-squared error MSE) is then

$$\text{MSE} = \mathsf{E}[(U - V(U))^2] = \int_{-\infty}^{\infty} f_U(u)(u - v(u))^2\, du = \sum_{j=1}^{M} \int_{\mathcal{R}_j} f_U(u)\, (u - a_j)^2\, du. \qquad (3.1)$$

In order to minimize (3.1) over the set of a_j, it is simply necessary to choose each a_j to minimize the corresponding integral (remember that the regions are considered fixed here). Let $f_j(u)$ denote the conditional pdf of U given that $\{u \in \mathcal{R}_j\}$; *i.e.*,

$$f_j(u) = \begin{cases} \frac{f_U(u)}{Q_j}, & \text{if} \quad u \in \mathcal{R}_j; \\ 0, & \text{otherwise,} \end{cases} \qquad (3.2)$$

where $Q_j = \Pr\{U \in \mathcal{R}_j\}$. Then, for the interval $\mathcal{R}_j$,

$$\int_{\mathcal{R}_j} f_U(u)\, (u - a_j)^2\, du = Q_j \int_{\mathcal{R}_j} f_j(u)\, (u - a_j)^2\, du. \qquad (3.3)$$

Now (3.3) is minimized by choosing a_j to be the mean of a random variable with the pdf $f_j(u)$. To see this, note that for any rv Y and real number a,

$$\overline{(Y - a)^2} = \overline{Y^2} - 2a\overline{Y} + a^2,$$

which is minimized over a when $a = \overline{Y}$.

This provides a set of conditions that the endpoints $\{b_j\}$ and the points $\{a_j\}$ must satisfy to achieve the MSE — namely, each b_j must be the midpoint between a_j and a_{j+1} and each a_j must be the mean of an rv U_j with pdf $f_j(u)$. In other words, a_j must be the conditional mean of U conditional on $U \in \mathcal{R}_j$.

These conditions are necessary to minimize the MSE for a given number M of representation points. They are not sufficient, as shown by an example at the end of this section. Nonetheless, these necessary conditions provide some insight into the minimization of the MSE.

3.2.3 The Lloyd-Max algorithm

The *Lloyd-Max algorithm*[1] is an algorithm for finding the endpoints $\{b_j\}$ and the representation points $\{a_j\}$ to meet the above necessary conditions. The algorithm is almost obvious given the necessary conditions; the contribution of Lloyd and Max was to define the problem and develop the necessary conditions. The algorithm simply alternates between the optimizations of the previous subsections, namely optimizing the endpoints $\{b_j\}$ for a given set of $\{a_j\}$, and then optimizing the points $\{a_j\}$ for the new endpoints.

[1] This algorithm was developed independently by S. P. Lloyd in 1957 and J. Max in 1960. Lloyd's work was done in the Bell Laboratories research department and became widely circulated, although unpublished until 1982 [16]. Max's work [18] was published in 1960.

The Lloyd-Max algorithm is as follows. Assume that the number M of quantizer levels and the pdf $f_U(u)$ are given.

1. Choose an arbitrary initial set of M representation points $a_1 < a_2 < \cdots < a_M$.
2. For each $j; 1 \leq j \leq M{-}1$, set $b_j = \frac{1}{2}(a_{j+1} + a_j)$.
3. For each $j; 1 \leq j \leq M$, set a_j equal to the conditional mean of U given $U \in (b_{j-1}, b_j]$ (where b_0 and b_M are taken to be $-\infty$ and $+\infty$ respectively).
4. Repeat steps (2) and (3) until further improvement in MSE is negligible; then stop.

The MSE decreases (or remains the same) for each execution of step (2) and step (3). Since the MSE is nonnegative, it approaches some limit. Thus if the algorithm terminates when the MSE improvement is less than some given $\varepsilon > 0$, then the algorithm must terminate after a finite number of iterations.

Example 3.2.1. This example shows that the algorithm might reach a local minimum of MSE instead of the global minimum. Consider a quantizer with $M = 2$ representation points, and an rv U whose pdf $f_U(u)$ has three peaks, as shown in Figure 3.3.

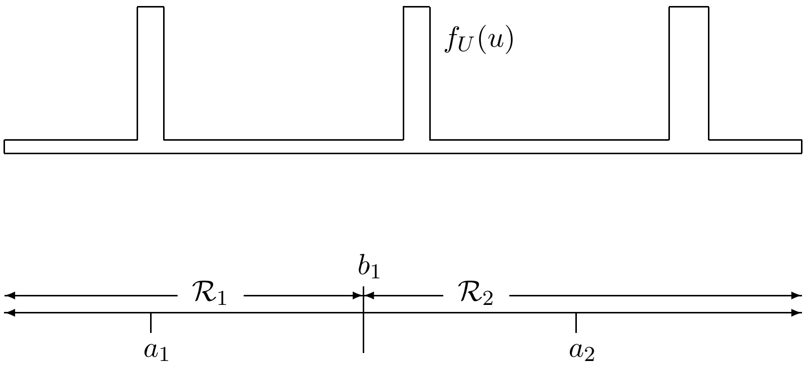


Figure 3.3: Example of regions and representaion points that satisfy Lloyd-Max conditions without minimizing mean-squared distortion.

It can be seen that one region must cover two of the peaks, yielding quite a bit of distortion, while the other will represent the remaining peak, yielding little distortion. In the figure, the two rightmost peaks are both covered by $\mathcal{R}_2$, with the point a_2 between them. Both the points and the regions satisfy the necessary conditions and cannot be locally improved. However, it can be seen in the figure that the rightmost peak is more probable than the other peaks. It follows that the MSE would be lower if $\mathcal{R}_1$ covered the two leftmost peaks.

The Lloyd-Max algorithm is a type of hill-climbing algorithm; starting with an arbitrary set of values, these values are modified until reaching the top of a hill where no more local improvements are possible.[2] A reasonable approach in this sort of situation is to try many randomly chosen starting points, perform the Lloyd-Max algorithm on each and then take the best solution. This is somewhat unsatisfying since there is no general technique for determining when the optimal solution has been found.

[2] It would be better to call this a valley-descending algorithm, both because a minimum is desired and also because binoculars can not be used at the bottom of a valley to find a distant lower valley.

3.3 Vector quantization

As with source coding of discrete sources, we next consider quantizing n source variables at a time. This is called *vector quantization*, since an n-tuple of rv's may be regarded as a vector rv in an n-dimensional vector space. We will concentrate on the case $n = 2$ so that illustrative pictures can be drawn.

One possible approach is to quantize each dimension independently with a scalar (one-dimensional) quantizer. This results in a rectangular grid of quantization regions as shown below. The MSE per dimension is the same as for the scalar quantizer using the same number of bits per dimension. Thus the best 2D vector quantizer has an MSE per dimension at least as small as that of the best scalar quantizer.

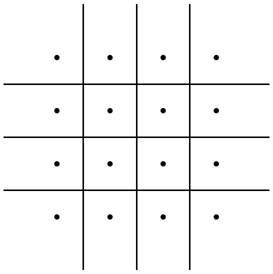


Figure 3.4: 2D rectangular quantizer.

To search for the minimum-MSE 2D vector quantizer with a given number M of representation points, the same approach is used as with scalar quantization.

Let (U, U') be the two rv's being jointly quantized. Suppose a set of M 2D representation points $\{(a_j, a'_j)\}$, $1 \le j \le M$ is chosen. For example, in the figure above, there are 16 representation points, represented by small dots. Given a sample pair (u, u') and given the M representation points, which representation point should be chosen for the given (u, u')? Again, the answer is easy. Since mapping (u, u') into (a_j, a'_j) generates a squared error equal to $(u - a_j)^2 + (u' - a'_j)^2$, the point (a_j, a'_j) which is closest to (u, u') in Euclidean distance should be chosen.

Consequently, the region $\mathcal{R}_j$ must be the set of points (u, u') that are closer to (a_j, a'_j) than to any other representation point. Thus the regions $\{\mathcal{R}_j\}$ are minimum-distance regions; these regions are called the *Voronoi* regions for the given representation points. The boundaries of the Voronoi regions are perpendicular bisectors between neighboring representation points. The minimum-distance regions are thus in general convex polygonal regions, as illustrated in the figure below.

As in the scalar case, the MSE can be minimized for a given set of regions by choosing the representation points to be the conditional means within those regions. Then, given this new set of representation points, the MSE can be further reduced by using the Voronoi regions for the new points. This gives us a 2D version of the Lloyd-Max algorithm, which must converge to a local minimum of the MSE. This can be generalized straightforwardly to any dimension n.

As already seen, the Lloyd-Max algorithm only finds local minima to the MSE for scalar quantizers. For vector quantizers, the problem of local minima becomes even worse. For example, when $U_1, U_2, \cdots$ are iid, it is easy to see that the rectangular quantizer in Figure 3.4 satisfies the Lloyd-Max conditions if the corresponding scalar quantizer does (see Exercise 3.10). It will

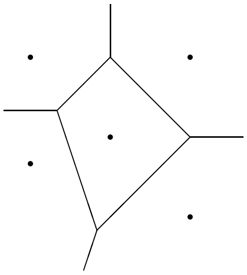


Figure 3.5: Voronoi regions for given set of representation points.

soon be seen, however, that this is not necessarily the minimum MSE.

Vector quantization was a popular research topic for many years. The problem is that quantizing complexity goes up exponentially with n, and the reduction in MSE with increasing n is quite modest, unless the samples are statistically highly dependent.

3.4 Entropy-coded quantization

We must now ask if minimizing the MSE for a given number M of representation points is the right problem. The minimum expected number of bits per symbol, $\overline{L}_{\min}$, required to encode the quantizer output was shown in Chapter 2 to be governed by the entropy $\mathsf{H}[V]$ of the quantizer output, not by the size M of the quantization alphabet. Therefore, anticipating efficient source coding of the quantized outputs, we should really try to minimize the MSE for a given entropy $\mathsf{H}[V]$ rather than a given number of representation points.

This approach is called *entropy-coded quantization* and is almost implicit in the layered approach to source coding represented in Figure 3.1. Discrete source coding close to the entropy bound is similarly often called entropy coding. Thus entropy-coded quantization refers to quantization techniques that are designed to be followed by entropy coding.

The entropy $\mathsf{H}[V]$ of the quantizer output is determined only by the probabilities of the quantization regions. Therefore, given a set of regions, choosing the representation points as conditional means minimizes their distortion without changing the entropy. However, given a set of representation points, the optimal regions are not necessarily Voronoi regions (e.g., in a scalar quantizer, the point separating two adjacent regions is not necessarily equidistant from the two represention points.)

For example, for a scalar quantizer with a constraint $\mathsf{H}[V] \leq \frac{1}{2}$ and a Gaussian pdf for U, a reasonable choice is three regions, the center one having high probability $1 - 2p$ and the outer ones having small, equal probability p, such that $\mathsf{H}[V] = \frac{1}{2}$.

Even for scalar quantizers, minimizing MSE subject to an entropy constraint is a rather messy problem. Considerable insight into the problem can be obtained by looking at the case where the target entropy is large— *i.e.*, when a large number of points can be used to achieve small MSE. Fortunately this is the case of greatest practical interest.

Example 3.4.1. For the following simple example, consider the minimum-MSE quantizer using a constraint on the number of representation points M compared to that using a constraint on the entropy $\mathsf{H}[V]$.

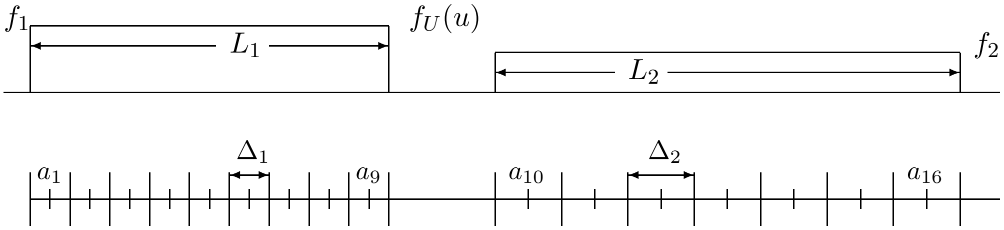


Figure 3.6: Comparison of constraint on M to constraint on $\mathsf{H}[U]$.

The example shows a piecewise constant pdf $f_U(u)$ that takes on only two positive values, say $f_U(u) = f_1$ over an interval of size L_1, and $f_U(u) = f_2$ over a second interval of size L_2. Assume that $f_U(u) = 0$ elsewhere. Because of the wide separation between the two intervals, they can be quantized separately without providing any representation point in the region between the intervals. Let M_1 and M_2 be the number of representation points in each interval. In the figure, $M_1 = 9$ and $M_2 = 7$. Let $\Delta_1 = L_1/M_1$ and $\Delta_2 = L_2/M_2$ be the lengths of the quantization regions in the two ranges (by symmetry, each quantization region in a given interval should have the same length). The representation points are at the center of each quantization interval. The MSE, conditional on being in a quantization region of length Δ_i, is the MSE of a uniform distribution over an interval of length Δ_i, which is easily computed to be $\Delta_i^2/12$. The probability of being in a given quantization region of size Δ_i is $f_i\Delta_i$, so the overall MSE is given by

$$\text{MSE} = M_1\frac{\Delta_1^2}{12}f_1\Delta_1 + M_2\frac{\Delta_2^2}{12}f_2\Delta_2 = \frac{1}{12}\Delta_1^2 f_1 L_1 + \frac{1}{12}\Delta_2^2 f_2 L_2. \tag{3.4}$$

This can be minimized over Δ_1 and Δ_2 subject to the constraint that $M = M_1 + M_2 = L_1/\Delta_1 + L_2/\Delta_2$. Ignoring the constraint that M_1 and M_2 are integers (which makes sense for M large), Exercise 3.4 shows that the minimum MSE occurs when Δ_i is chosen inversely proportional to the cube root of f_i. In other words,

$$\frac{\Delta_1}{\Delta_2} = \left(\frac{f_2}{f_1}\right)^{1/3}. \tag{3.5}$$

This says that the size of a quantization region decreases with increasing probability density. This is reasonable, putting the greatest effort where there is the most probability. What is perhaps surprising is that this effect is so small, proportional only to a cube root.

Perhaps even more surprisingly, if the MSE is minimized subject to a constraint on entropy for this example, then Exercise 3.4 shows that, in the limit of high rate, the quantization intervals all have the same length! A scalar quantizer in which all intervals have the same length is called a *uniform scalar quantizer*. The following sections will show that uniform scalar quantizers have remarkable properties for high-rate quantization.

3.5 High-rate entropy-coded quantization

This section focuses on high-rate quantizers where the quantization regions can be made sufficiently small so that the probability density is approximately constant within each region. It will be shown that under these conditions the combination of a uniform scalar quantizer followed by discrete entropy coding is nearly optimum (in terms of mean-squared distortion) within the class of scalar quantizers. This means that a uniform quantizer can be used as a universal quantizer with very little loss of optimality. The probability distribution of the rv's to be quantized can be explointed at the level of discrete source coding. Note however that this essential optimality of uniform quantizers relies heavily on the assumption that mean-squared distortion is an appropriate distortion measure. With voice coding, for example, a given distortion at low signal levels is for more harmful than the same distortion at high signal levels.

In the following sections, it is assumed that the source output is a sequence $U_1, U_2, \ldots$, of iid real analog-valued rv's, each with a probability density $f_U(u)$. It is further assumed that the probability density function (pdf) $f_U(u)$ is smooth enough and the quantization fine enough that $f_U(u)$ is almost constant over each quantization region.

The analogue of the entropy $\mathsf{H}[X]$ of a discrete rv is the differential entropy $\mathsf{h}[U]$ of an analog rv. After defining $\mathsf{h}[U]$,the properties of $\mathsf{H}[U]$ and $\mathsf{h}[U]$ will be compared.

The performance of a uniform scalar quantizer followed by entropy coding will then be analyzed. It will be seen that there is a tradeoff between the rate of the quantizer and the mean-squared error (MSE) between source and quantized output. It is also shown that the uniform quantizer is essentially optimum among scalar quantizers at high rate.

The performance of uniform vector quantizers followed by entropy coding will then be analyzed and similar tradeoffs will be found. A major result is that vector quantizers can achieve a gain over scalar quantizers (*i.e.*, a reduction of MSE for given quantizer rate), but that the reduction in MSE is at most a factor of $\pi e/6 = 1.42$.

The changes in MSE for different quantization methods, and similarly, changes in power levels on channels, are invariably calculated by communication engineers in decibels (dB). The number of decibels corresponding to a reduction of α in the mean squared error is defined to be $10 \log_{10} \alpha$. The use of a logarithmic measure allows the various components of mean squared error or power gain to be added rather than multiplied.

The use of decibels rather than some other logarithmic measure such as natural logs or logs to the base 2 is partly motivated by the ease of doing rough mental calculations. A factor of 2 is $10 \log_{10} 2 = 3.010\cdots$ dB, approximated as 3 dB. Thus $4 = 2^2$ is 6 dB and 8 is 9 dB. Since 10 is 10 dB, we also see that 5 is 10/2 or 7 dB. We can just as easily see that 20 is 13 dB and so forth. The limiting factor of 1.42 in MSE above is then a reduction of 1.53 dB.

As in the discrete case, generalizations to analog sources with memory are possible, but not discussed here.

3.6 Differential entropy

The differential entropy $\mathsf{h}[U]$ of an analog random variable (rv) U is analogous to the entropy $\mathsf{H}[X]$ of a discrete random symbol X. It has many similarities, but also some important differences.

Definition The *differential entropy* of an analog real rv U with pdf $f_U(u)$ is

$$\mathsf{h}[U] = \int_{-\infty}^{\infty} -f_U(u) \log f_U(u)\; du.$$

The integral may be restricted to the region where $f_U(u) > 0$, since $0 \log 0$ is interpreted as 0. Assume that $f_U(u)$ is smooth and that the integral exists with a finite value. Exercise 3.7 gives an example where $\mathsf{h}(U)$ is infinite.

As before, the logarithms are base 2 and the units of $\mathsf{h}[U]$ are bits per source symbol.

Like $\mathsf{H}[X]$, the differential entropy $\mathsf{h}[U]$ is the expected value of the rv $-\log f_U(U)$. The log of the joint density of several independent rv's is the sum of the logs of the individual pdf's, and this can be used to derive an AEP similar to the discrete case.

Unlike $\mathsf{H}[X]$, the differential entropy $\mathsf{h}[U]$ can be negative and depends on the scaling of the outcomes. This can be seen from the following two examples.

Example 3.6.1 (Uniform distributions). Let $f_U(u)$ be a uniform distribution over an interval $[a, a+\Delta]$ of length Δ; *i.e.*, $f_U(u) = 1/\Delta$ for $u \in [a, a+\Delta]$, and $f_U(u) = 0$ elsewhere. Then $-\log f_U(u) = \log \Delta$ where $f_U(u) > 0$ and

$$\mathsf{h}[U] = \mathsf{E}[-\log f_U(U)] = \log \Delta.$$

Example 3.6.2 (Gaussian distribution). Let $f_U(u)$ be a Gaussian distribution with mean m and variance σ^2; *i.e.*,

$$f_U(u) = \sqrt{\frac{1}{2\pi\sigma^2}} \exp\left\{-\frac{(u-m)^2}{2\sigma^2}\right\}.$$

Then $-\log f_U(u) = \frac{1}{2}\log 2\pi\sigma^2 + (\log e)(u-m)^2/(2\sigma^2)$. Since $\mathsf{E}[(U-m)^2] = \sigma^2$,

$$\mathsf{h}[U] = \mathsf{E}[-\log f_U(U)] = \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2}\log e = \frac{1}{2}\log(2\pi e\sigma^2).$$

It can be seen from these expressions that by making Δ or σ^2 arbitrarily small, the differential entropy can be made arbitrarily negative, while by making Δ or σ^2 arbitrarily large, the differential entropy can be made arbitrarily positive.

If the rv U is rescaled to αU for some scale factor $\alpha > 0$, then the differential entropy is increased by $\log \alpha$, both in these examples and in general. In other words, $\mathsf{h}[U]$ is not invariant to scaling. Note, however, that differential entropy is invariant to translation of the pdf, *i.e.*, an rv and its fluctuation around the mean have the same differential entropy.

One of the important properties of entropy is that it does not depend on the labeling of the elements of the alphabet, *i.e.*, it is invariant to invertible transformations. Differential entropy is very different in this respect, and, as just illustrated, it is modified by even such a trivial transformation as a change of scale. The reason for this is that the probability density is a probability per unit length, and therefore depends on the measure of length. In fact, as seen more clearly later, this fits in very well with the fact that source coding for analog sources also depends on an error term per unit length.

Definition The *differential entropy* of an n-tuple of rv's $\boldsymbol{U}^n = (U_1, \cdots, U_n)$ with joint pdf $f_{\boldsymbol{U}^n}(\boldsymbol{u}^n)$ is

$$\mathsf{h}[\boldsymbol{U}^n] = \mathsf{E}[-\log f_{\boldsymbol{U}^n}(\boldsymbol{U}^n)].$$

Like entropy, differential entropy has the property that if U and V are independent rv's, then the entropy of the joint variable UV with pdf $f_{UV}(u,v) = f_U(u)f_V(v)$ is $\mathsf{h}[UV] = \mathsf{h}[U] + \mathsf{h}[V]$.

Again, this follows from the fact that the log of the joint probability density of independent rv's is additive, *i.e.*, $-\log f_{UV}(u,v) = -\log f_U(u) - \log f_V(v)$.

Thus the differential entropy of a vector rv $\boldsymbol{U}^n$, corresponding to a string of n iid rv's $U_1, U_2, \dots, U_n$, each with the density $f_U(u)$, is $\mathsf{h}[\boldsymbol{U}^n] = n\mathsf{h}[U]$.

3.7 Performance of uniform high-rate scalar quantizers

This section analyzes the performance of uniform scalar quantizers in the limit of high rate. Appendix A continues the analysis for the nonuniform case and shows that uniform quantizers are effectively optimal in the high-rate limit.

For a uniform scalar quantizer, every quantization interval $\mathcal{R}_j$ has the same length $|\mathcal{R}_j| = \Delta$. In other words, $\mathbb{R}$ (or the portion of $\mathbb{R}$ over which $f_U(u) > 0$), is partitioned into equal intervals, each of length Δ.

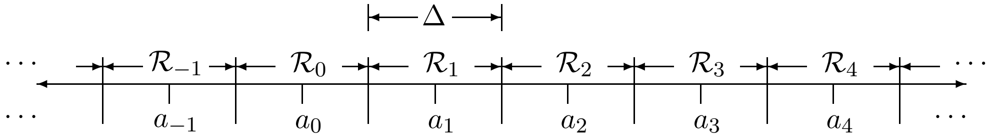


Figure 3.7: Uniform scalar quantizer.

Assume there are enough quantization regions to cover the region where $f_U(u) > 0$. For the Gaussian distribution, for example, this requires an infinite number of representation points, $-\infty < j < \infty$. Thus, in this example the quantized discrete rv V has a countably infinite alphabet. Obviously, practical quantizers limit the number of points to a finite region $\mathcal{R}$ such that $\int_{\mathcal{R}} f_U(u)\, du \approx 1$.

Assume that Δ is small enough that the pdf $f_U(u)$ is approximately constant over any one quantization interval. More precisely, define $\overline{f}(u)$ (see Figure 3.8) as the average value of $f_U(u)$ over the quantization interval containing u,

$$\overline{f}(u) = \frac{\int_{\mathcal{R}_j} f_U(u)du}{\Delta} \qquad \text{for} \quad u \in \mathcal{R}_j. \tag{3.6}$$

From (3.6) it is seen that $\Delta\overline{f}(u) = \Pr(\mathcal{R}_j)$ for all integer j and all $u \in \mathcal{R}_j$.

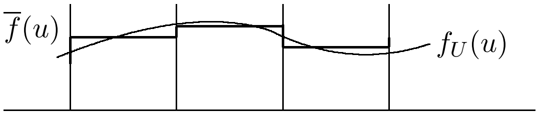


Figure 3.8: Average density over each $\mathcal{R}_j$.

The *high-rate assumption* is that $f_U(u) \approx \overline{f}(u)$ for all $u \in \mathbb{R}$. This means that $f_U(u) \approx \Pr(\mathcal{R}_j)/\Delta$ for $u \in \mathcal{R}_j$. It also means that the conditional pdf $f_{U|\mathcal{R}_j}(u)$ of U conditional on $u \in \mathcal{R}_j$ is approximated by

$$f_{U|\mathcal{R}_j}(u) \approx \begin{cases} 1/\Delta, & u \in \mathcal{R}_j; \\ 0, & u \notin \mathcal{R}_j. \end{cases}$$

Consequently the conditional mean a_j is approximately in the center of the interval $\mathcal{R}_j$, and the mean-squared error is approximately given by

$$\text{MSE} \approx \int_{-\Delta/2}^{\Delta/2} \frac{1}{\Delta} u^2 du = \frac{\Delta^2}{12} \tag{3.7}$$

for each quantization interval $\mathcal{R}_j$. Consequently this is also the overall MSE.

Next consider the entropy of the quantizer output V. The probability p_j that $V = a_j$ is given by both

$$p_j = \int_{\mathcal{R}_j} f_U(u)\, du \qquad \text{and, for all } u \in \mathcal{R}_j, \qquad p_j = \overline{f}(u)\Delta. \tag{3.8}$$

Therefore the entropy of the discrete rv V is

$$\begin{aligned} \mathsf{H}[V] &= \sum_j -p_j \log p_j = \sum_j \int_{\mathcal{R}_j} -f_U(u) \log[\overline{f}(u)\Delta]\, du \\ &= \int_{-\infty}^{\infty} -f_U(u) \log[\overline{f}(u)\Delta]\, du \qquad (3.9) \\ &= \int_{-\infty}^{\infty} -f_U(u) \log[\overline{f}(u)]\, du - \log\Delta, \qquad (3.10) \end{aligned}$$

where the sum of disjoint integrals were combined into a single integral.

Finally, using the high-rate approximation[3] $f_U(u) \approx \overline{f}(u)$, this becomes

$$\begin{aligned} \mathsf{H}[V] &\approx \int_{-\infty}^{\infty} -f_U(u) \log[f_U(u)\Delta]\, du \\ &= \mathsf{h}[U] - \log\Delta. \qquad (3.11) \end{aligned}$$

Since the sequence $U_1, U_2, \ldots$ of inputs to the quantizer is memoryless (iid), the quantizer output sequence $V_1, V_2, \ldots$ is an iid sequence of discrete random symbols representing quantization points— *i.e.*, a discrete memoryless source. A uniquely-decodable source code can therefore be used to encode this output sequence into a bit sequence at an average rate of $\overline{L} \approx \mathsf{H}[V] \approx \mathsf{h}[U] - \log\Delta$ bits/symbol. At the receiver, the mean-squared quantization error in reconstructing the original sequence is approximately MSE $\approx \Delta^2/12$.

The important conclusions from this analysis are illustrated in Figure 3.9 and are summarized as follows:

- Under the high-rate assumption, the rate $\overline{L}$ for a uniform quantizer followed by discrete entropy coding depends only on the differential entropy $\mathsf{h}[U]$ of the source and the spacing Δ of the quantizer. It does not depend on any other feature of the source pdf $f_U(u)$, nor on any other feature of the quantizer, such as the number M of points, so long as the quantizer intervals cover $f_U(u)$ sufficiently completely and finely.

[3] Exercise 3.6 provides some insight into the nature of the approximation here. In particular, the difference between $\mathsf{h}[U] - \log\Delta$ and $\mathsf{H}[V]$ is $\int f_U(u) \log[\overline{f}(u)/f_U(u)]\, du$. This quantity is always nonpositive and goes to zero with Δ as Δ^2. Similarly, the approximation error on MSE goes to 0 as Δ^4.

- The rate $\overline{L} \approx \mathsf{H}[V]$ and the MSE are parametrically related by Δ, *i.e.*,

$$\overline{L} \approx \mathsf{h}(U) - \log \Delta; \qquad \text{MSE} \approx \frac{\Delta^2}{12}. \tag{3.12}$$

Note that each reduction in Δ by a factor of 2 will reduce the MSE by a factor of 4 and increase the required transmission rate $\overline{L} \approx \mathsf{H}[V]$ by 1 bit/symbol. Communication engineers express this by saying that each additional bit per symbol decreases the mean-squared distortion[4] by 6 dB. Figure 3.9 sketches MSE as a function of $\overline{L}$.

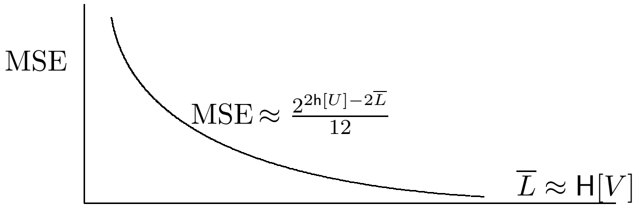


Figure 3.9: MSE as a function of $\overline{L}$ for a scalar quantizer with the high-rate approximation. Note that changing the source entropy $\mathsf{h}(U)$ simply shifts the figure right or left. Note also that log MSE is linear, with a slope of -2, as a function of $\overline{L}$.

Conventional b-bit analog-to-digital (A/D) converters are uniform scalar 2^b-level quantizers that cover a certain range $\mathcal{R}$ with a quantizer spacing $\Delta = 2^{-b}|\mathcal{R}|$. The input samples must be scaled so that the probability that $u \notin \mathcal{R}$ (the "overflow probability") is small. For a fixed scaling of the input, the tradeoff is again that increasing b by 1 bit reduces the MSE by a factor of 4.

Conventional A/D converters are not usually directly followed by entropy coding. The more conventional approach is to use A/D conversion to produce a very high rate digital signal that can be further processed by digital signal processing (DSP). This digital signal is then later compressed using algorithms specialized to the particular application (voice, images, etc.). In other words, the clean layers of Figure 3.1 oversimplify what is done in practice. On the other hand, it is often best to view compression in terms of the Figure 3.1 layers, and then use DSP as a way of implementing the resulting algorithms.

The relation $\mathsf{H}[V] \approx \mathsf{h}[u] - \log \Delta$ provides an elegant interpretation of differential entropy. It is obvious that there must be some kind of tradeoff between MSE and the entropy of the representation, and the differential entropy specifies this tradeoff in a very simple way for high rate uniform scalar quantizers. $\mathsf{H}[V]$ is the entropy of a finely quantized version of U, and the additional term $\log \Delta$ relates to the "uncertainty" within an individual quantized interval. It shows explicitly how the scale used to measure U affects $\mathsf{h}[U]$.

Appendix A considers nonuniform scalar quantizers under the high rate assumption and shows that nothing is gained in the high-rate limit by the use of nonuniformity.

[4] A quantity x expressed in dB is given by $10 \log_{10} x$. This very useful and common logarithmic measure is discussed in detail in Chapter 6.

3.8 High-rate two-dimensional quantizers

The performance of uniform two-dimensional (2D) quantizers are now analyzed in the limit of high rate. Appendix B considers the nonuniform case and shows that uniform quantizers are again effectively optimal in the high-rate limit.

A 2D quantizer operates on 2 source samples $\boldsymbol{u} = (u_1, u_2)$ at a time; *i.e.*, the source alphabet is $\boldsymbol{U} = \mathbb{R}^2$. Assuming iid source symbols, the joint pdf is then $f_{\boldsymbol{U}}(\boldsymbol{u}) = f_U(u_1)f_U(u_2)$, and the joint differential entropy is $\mathsf{h}[\boldsymbol{U}] = 2\mathsf{h}[U]$.

Like a uniform scalar quantizer, a uniform 2D quantizer is based on a fundamental quantization region $\mathcal{R}$ ("quantization cell") whose translates tile[5] the 2D plane. In the one-dimensional case, there is really only one sensible choice for $\mathcal{R}$, namely an interval of length Δ, but in higher dimensions there are many possible choices. For two dimensions, the most important choices are squares and hexagons, but in higher dimensions, many more choices are available.

Notice that if a region $\mathcal{R}$ tiles $\mathbb{R}^2$, then any scaled version $\alpha\mathcal{R}$ of $\mathcal{R}$ will also tile $\mathbb{R}^2$, and so will any rotation or translation of $\mathcal{R}$.

Consider the performance of a uniform 2D quantizer with a basic cell $\mathcal{R}$ which is centered at the origin $\boldsymbol{0}$. The set of cells, which are assumed to tile the region, are denoted by[6] $\{\mathcal{R}_j;\, j \in \mathbb{Z}^+\}$ where $\mathcal{R}_j = \boldsymbol{a}_j + \mathcal{R}$ and $\boldsymbol{a}_j$ is the center of the cell $\mathcal{R}_j$. Let $A(\mathcal{R}) = \int_{\mathcal{R}} d\boldsymbol{u}$ be the area of the basic cell. The average pdf in a cell $\mathcal{R}_j$ is given by $\Pr(\mathcal{R}_j)/A(\mathcal{R}_j)$. As before, define $\overline{f}(\boldsymbol{u})$ to be the average pdf over the region $\mathcal{R}_j$ containing $\boldsymbol{u}$. The high-rate assumption is again made, *i.e.*, assume that the region $\mathcal{R}$ is small enough that $f_{\boldsymbol{U}}(\boldsymbol{u}) \approx \overline{f}(\boldsymbol{u})$ for all $\boldsymbol{u}$.

The assumption $f_{\boldsymbol{U}}(\boldsymbol{u}) \approx \overline{f}(\boldsymbol{u})$ implies that the conditional pdf, conditional on $\boldsymbol{u} \in \mathcal{R}_j$ is approximated by

$$f_{\boldsymbol{U}|\mathcal{R}_j}(\boldsymbol{u}) \approx \begin{cases} 1/A(\mathcal{R}), & \boldsymbol{u} \in \mathcal{R}_j; \\ 0, & \boldsymbol{u} \notin \mathcal{R}_j. \end{cases} \tag{3.13}$$

The conditional mean is approximately equal to the center $\boldsymbol{a}_j$ of the region $\mathcal{R}_j$. The mean-squared error per dimension for the basic quantization cell $\mathcal{R}$ centered on 0 is then approximately equal to

$$\text{MSE} \approx \frac{1}{2}\int_{\mathcal{R}} \|\boldsymbol{u}\|^2 \frac{1}{A(\mathcal{R})}\; d\boldsymbol{u}. \tag{3.14}$$

The right side of (3.14) is the MSE for the quantization area $\mathcal{R}$ using a pdf equal to a constant; it will be denoted MSE_c. The quantity $\|\boldsymbol{u}\|$ is the length of the vector u_1, u_2, so that $\|\boldsymbol{u}\|^2 = u_1^2 + u_2^2$. Thus MSE_c can be rewritten as

$$\text{MSE} \approx \text{MSE}_c = \frac{1}{2}\int_{\mathcal{R}} (u_1^2 + u_2^2)\, \frac{1}{A(\mathcal{R})}\; du_1 du_2. \tag{3.15}$$

MSE_c is measured in units of squared length, just like $A(\mathcal{R})$. Thus the ratio $G(\mathcal{R}) = \text{MSE}_c/A(\mathcal{R})$ is a dimensionless quantity called the normalized second moment. With a little effort, it can

[5] A region of the 2D plane is said to *tile* the plane if the region, plus translates and rotations of the region, fill the plane without overlap. For example the square and the hexagon tile the plane. Also, rectangles tile the plane, and equilateral triangles with rotations tile the plane.

[6] $\mathbb{Z}^+$ denotes the set of positive integers, so $\{\mathcal{R}_j;\, j \in \mathbb{Z}^+\}$ denotes the set of regions in the tiling, numbered in some arbitrary way of no particular interest here.

be seen that $G(\mathcal{R})$ is invariant to scaling, translation and rotation. $G(\mathcal{R})$ does depend on the shape of the region $\mathcal{R}$, and, as seen below, it is $G(\mathcal{R})$ that determines how well a given shape performs as a quantization region. By expressing

$$\mathrm{MSE}_c = G(\mathcal{R})A(\mathcal{R}),$$

it is seen that the MSE is the product of a shape term and an area term, and these can be chosen independently.

As examples, $G(\mathcal{R})$ is given below for some common shapes.

- Square: For a square Δ on a side, $A(\mathcal{R}) = \Delta^2$. Breaking (3.15) into two terms, we see that each is identical to the scalar case and $\mathrm{MSE}_c = \Delta^2/12$. Thus $G(\text{Square}) = 1/12$.
- Hexagon: View the hexagon as the union of 6 equilateral triangles Δ on a side. Then $A(\mathcal{R}) = 3\sqrt{3}\Delta^2/2$ and $\mathrm{MSE}_c = 5\Delta^2/24$. Thus $G(\text{hexagon}) = 5/(36\sqrt{3})$.
- Circle: For a circle of radius r, $A(\mathcal{R}) = \pi r^2$ and $\mathrm{MSE}_c = r^2/4$ so $G(\text{circle}) = 1/(4\pi)$.

The circle is not an allowable quantization region, since it does not tile the plane. On the other hand, for a given area, this is the shape that minimizes MSE_c. To see this, note that for any other shape, differential areas further from the origin can be moved closer to the origin with a reduction in MSE_c. That is, the circle is the 2D shape that minimizes $G(\mathcal{R})$. This also suggests why $G(\text{Hexagon}) < G(\text{Square})$, since the hexagon is more concentrated around the origin than the square.

Using the high rate approximation for any given tiling, each quantization cell $\mathcal{R}_j$ has the same shape and area and has a conditional pdf which is approximately uniform. Thus MSE_c approximates the MSE for each quantization region and thus approximates the overall MSE.

Next consider the entropy of the quantizer output. The probability that $\boldsymbol{U}$ falls in the region $\mathcal{R}_j$ is

$$p_j = \int_{\mathcal{R}_j} f_{\boldsymbol{U}}(\boldsymbol{u})\, d\boldsymbol{u} \qquad \text{and, for all } \boldsymbol{u} \in \mathcal{R}_j, \quad p_j = \overline{f}(\boldsymbol{u})A(\mathcal{R}).$$

The output of the quantizer is the discrete random symbol V with the pmf p_j for each symbol j. As before, the entropy of $\boldsymbol{V}$ is given by

$$\begin{aligned}
\mathsf{H}[\boldsymbol{V}] &= -\sum_j p_j \log p_j \\
&= -\sum_j \int_{\mathcal{R}_j} f_{\boldsymbol{U}}(\boldsymbol{u}) \log[\overline{f}(\boldsymbol{u})A(\mathcal{R})]\, d\boldsymbol{u} \\
&= -\int f_{\boldsymbol{U}}(\boldsymbol{u})\ [\log \overline{f}(\boldsymbol{u}) + \log A(\mathcal{R})]\, d\boldsymbol{u} \\
&\approx -\int f_{\boldsymbol{U}}(\boldsymbol{u})\ [\log f_{\boldsymbol{U}}(\boldsymbol{u})]\, d\boldsymbol{u} + \log A(\mathcal{R})] \\
&= 2\mathsf{h}[U] - \log A(\mathcal{R}),
\end{aligned}$$

where the high rate approximation $f_{\boldsymbol{U}}(\boldsymbol{u}) \approx \bar{f}(\boldsymbol{u})$ was used. Note that, since $\boldsymbol{U} = U_1U_2$ for iid variables U_1 and U_2, the differential entropy of $\boldsymbol{U}$ is $2\mathsf{h}[U]$.

Again, an efficient uniquely-decodable source code can be used to encode the quantizer output sequence into a bit sequence at an average rate per source symbol of

$$\overline{L} \approx \frac{\mathsf{H}[\boldsymbol{V}]}{2} \approx \mathsf{h}[U] - \frac{1}{2}\log A(\mathcal{R}) \quad \text{bits/symbol.} \tag{3.16}$$

At the receiver, the mean-squared quantization error in reconstructing the original sequence will be approximately equal to the MSE given in (3.14).

We have the following important conclusions for a uniform 2D quantizer under the high-rate approximation:

- Under the high-rate assumption, the rate $\overline{L}$ depends only on the differential entropy $\mathsf{h}[U]$ of the source and the area $A(\mathcal{R})$ of the basic quantization cell $\mathcal{R}$. It does not depend on any other feature of the source pdf $f_U(u)$, and does not depend on the shape of the quantizer region, *i.e.*, it does not depend on the normalized second moment $G(\mathcal{R})$.

- There is a tradeoff between the rate $\overline{L}$ and MSE that is governed by the area $A(\mathcal{R})$. From (3.16), an increase of 1 bit/symbol in rate corresponds to a decrease in $A(\mathcal{R})$ by a factor of 4. From (3.14), this decreases the MSE by a factor of 4, *i.e.*, by 6 dB.

- The ratio $G(\text{Square})/G(\text{Hexagon})$ is equal to $3\sqrt{3}/5 = 1.0392$. This is called the *quantizing gain* of the hexagon over the square. For a given $A(\mathcal{R})$ (and thus a given $\overline{L}$), the MSE for a hexagonal quantizer is smaller than that for a square quantizer (and thus also for a scalar quantizer) by a factor of 1.0392 (0.17 dB). This is a disappointingly small gain given the added complexity of 2D and hexagonal regions and suggests that uniform scalar quantizers are good choices at high rates.

3.9 Summary of quantization

Quantization is important both for digitizing a sequence of analog signals and as the middle layer in digitizing analog waveform sources. Uniform scalar quantization is the simplest and often most practical approach to quantization. Before reaching this conclusion, two approaches to optimal scalar quantizers were taken. The first attempted to minimize the expected distortion subject to a fixed number M of quantization regions, and the second attempted to minimize the expected distortion subject to a fixed entropy of the quantized output. Each approach was followed by the extension to vector quantization.

In both approaches, and for both scalar and vector quantization, the emphasis was on minimizing mean square distortion or error (MSE), as opposed to some other distortion measure. As will be seen later, MSE is the natural distortion measure in going from waveforms to sequences of analog values. For specific sources, such as speech, however, MSE is not appropriate. For an introduction to quantization, however, focusing on MSE seems appropriate in building intuition; again, our approach is building understanding through the use of simple models.

The first approach, minimizing MSE with a fixed number of regions, leads to the Lloyd-Max algorithm, which finds a local minimum of MSE. Unfortunately, the local minimum is not necessarily a global minimum, as seen by several examples. For vector quantization, the problem of local (but not global) minima arising from the Lloyd-Max algorithm appears to be the typical case.

The second approach, minimizing MSE with a constraint on the output entropy is also a difficult problem analytically. This is the appropriate approach in a two layer solution where the quantizer is followed by discrete encoding. On the other hand, the first approach is more appropriate when vector quantization is to be used but cannot be followed by fixed-to-variable-length discrete source coding.

High-rate scalar quantization, where the quantization regions can be made sufficiently small so that the probability density in almost constant over each region, leads to a much simpler result when followed by entropy coding. In the limit of high rate, a uniform scalar quantizer minimizes MSE for a given entropy constraint. Moreover, the tradeoff between Minimum MSE and output entropy is the simple univeral curve of Figure 3.9. The source is completely characterized by its differential entropy in this tradeoff. The approximations in this result are analyzed in Exercise 3.6. Two-dimensional vector quantization under the high-rate approximation with entropy coding leads to a similar result. Using a square quantization region to tile the plane, the tradeoff between MSE per symbol and entropy per symbol is the same as with scalar quantization. Using a hexagonal quantization region to tile the plane reduces the MSE by a factor of 1.0392, which seems hardly worth the trouble. It is possible that non-uniform two-dimensional quantizers might achieve a smaller MSE than a hexagonal tiling, but this gain is still limited by the circular shaping gain, which is $\pi/3 = 1.0472$ (0.2 dB). Using non-uniform quantization regions at high rate leads to a lowerbound on MSE which is lower than that for the scalar uniform quantizer by a factor of 1.0472, which, even if achievable, is scarcely worth the trouble.

The use of high-dimensional quantizers can achieve slightly higher gains over the uniform scalar quantizer, but the gain is still limited by a fundamental information-theoretic result to $\pi e/6 = 1.423$ (1.53 dB).

3A Appendix A: Nonuniform scalar quantizers

This appendix shows that the approximate MSE for uniform high-rate scalar quantizers in Section 3.7 provides an approximate lower bound on the MSE for any nonuniform scalar quantizer, again using the high-rate approximation that the pdf of U is constant within each quantization region. This shows that in the high-rate region, there is little reason to further consider nonuniform scalar quantizers.

Consider an arbitrary scalar quantizer for an rv U with a pdf $f_U(u)$. Let Δ_j be the width of the jth quantization interval, *i.e.*, $\Delta_j = |\mathcal{R}_j|$. As before, let $\overline{f}(u)$ be the average pdf within each quantization interval, *i.e.*,

$$\overline{f}(u) = \frac{\int_{\mathcal{R}_j} f_U(u)\, du}{\Delta_j} \qquad \text{for} \quad u \in \mathcal{R}_j.$$

The high-rate approximation is that $f_U(u)$ is approximately constant over each quantization region. Equivalently, $f_U(u) \approx \overline{f}(u)$ for all u. Thus, if region $\mathcal{R}_j$ has width Δ_j, the conditional mean a_j of U over $\mathcal{R}_j$ is approximately the midpoint of the region, and the conditional mean-squared error, MSE_j, given $U{\in}\mathcal{R}_j$, is approximately $\Delta_j^2/12$.

Let V be the quantizer output, *i.e.*, the discrete rv such that $V = a_j$ whenever $U \in \mathcal{R}_j$. The probability p_j that $V{=}a_j$ is $p_j = \int_{\mathcal{R}_j} f_U(u)\, du$

The unconditional mean-squared error, *i.e.*. $\mathsf{E}[(U-V)^2]$ is then given by

$$\text{MSE} \approx \sum_j p_j \frac{\Delta_j^2}{12} = \sum_j \int_{\mathcal{R}_j} f_U(u)\, \frac{\Delta_j^2}{12}\, du. \tag{3.17}$$

This can be simplified by defining $\Delta(u) = \Delta_j$ for $u \in \mathcal{R}_j$. Since each u is in $\mathcal{R}_j$ for some j, this defines $\Delta(u)$ for all $u \in \mathbb{R}$. Substituting this in (3.17),

$$\begin{aligned} \text{MSE} &\approx \sum_j \int_{\mathcal{R}_j} f_U(u)\, \frac{\Delta(u)^2}{12}\, du & (3.18)\\ &= \int_{-\infty}^{\infty} f_U(u)\, \frac{\Delta(u)^2}{12}\, du\ . & (3.19) \end{aligned}$$

Next consider the entropy of V. As in (3.8), the following relations are used for p_j

$$p_j = \int_{\mathcal{R}_j} f_U(u)\ du \quad \text{and, for all } u \in \mathcal{R}_j, \quad p_j = \overline{f}(u)\Delta(u).$$

$$\begin{aligned} \mathsf{H}[V] &= \sum_j -p_j \log p_j \\ &= \sum_j \int_{\mathcal{R}_j} -f_U(u) \log[\,\overline{f}(u)\Delta(u)]\ du & (3.20)\\ &= \int_{-\infty}^{\infty} -f_U(u) \log[\overline{f}(u)\Delta(u)]\ du, & (3.21) \end{aligned}$$

where the multiple integrals over disjoint regions have been combined into a single integral. The high-rate approximation $f_U(u) \approx \overline{f}(u)$ is next substituted into (3.21).

$$\begin{aligned} \mathsf{H}[V] &\approx \int_{-\infty}^{\infty} -f_U(u) \log[f_U(u)\Delta(u)]\ du \\ &= \mathsf{h}[U] - \int_{-\infty}^{\infty} f_U(u) \log \Delta(u)\ du. & (3.22) \end{aligned}$$

Note the similarity of this to (3.11).

The next step is to minimize the mean-squared error subject to a constraint on the entropy $\mathsf{H}[V]$. This is done approximately by minimizing the approximation to MSE in (3.22) subject to the approximation to $\mathsf{H}[V]$ in (3.19). Exercise 3.6 provides some insight into the accuracy of these approximations and their effect on this minimization.

Consider using a Lagrange multiplier to perform the minimization. Since MSE decreases as $\mathsf{H}[V]$ increases, consider minimizing $\text{MSE} + \lambda\mathsf{H}[V]$. As λ increases, MSE will increase and $\mathsf{H}[V]$ decrease in the minimizing solution.

In principle, the minimization should be constrained by the fact that $\Delta(u)$ is constrained to represent the interval sizes for a realizable set of quantization regions. The minimum of $\text{MSE} + \lambda\mathsf{H}[V]$ will be lower bounded by ignoring this constraint. The very nice thing that happens is that this unconstrained lower bound occurs where $\Delta(u)$ is constant. This corresponds to a uniform quantizer, which is clearly realizable. In other words, subject to the high-rate approximation, the lower bound on MSE over all scalar quantizers is equal to the MSE for the uniform scalar quantizer. To see this, use (3.19) and (3.22),

$$\begin{aligned} \mathrm{MSE} + \lambda \mathsf{H}[V] &\approx \int_{-\infty}^{\infty} f_U(u)\, \frac{\Delta(u)^2}{12}\; du \;+ \lambda \mathsf{h}[U] - \lambda \int_{-\infty}^{\infty} f_U(u) \log \Delta(u)\; du \\ &= \lambda \mathsf{h}[U] + \int_{-\infty}^{\infty} f_U(u) \left\{ \frac{\Delta(u)^2}{12} - \lambda \log \Delta(u) \right\}\; du. \end{aligned} \qquad (3.23)$$

This is minimized over all choices of $\Delta(u) > 0$ by simply minimizing the expression inside the braces for each real value of u. That is, for each u, differentiate the quantity inside the braces with respect to $\Delta(u)$, getting $\Delta(u)/6 - \lambda(\log e)/\Delta(u)$. Setting the derivative equal to 0, it is seen that $\Delta(u) = \sqrt{\lambda(\log e)/6}$. By taking the second derivative, it can be seen that this solution actually minimizes the integrand for each u. The only important thing here is that the minimizing $\Delta(u)$ is independent of u. This means that the approximation of MSE is minimized, subject to a constraint on the approximation of $\mathsf{H}[V]$, by the use of a uniform quantizer.

The next question is the meaning of minimizing an approximation to something subject to a constraint which itself is an approximation. From Exercise 3.6, it is seen that both the approximation to MSE and that to $\mathsf{H}[V]$ are good approximations for small Δ, *i.e.*, for high-rate. For any given high-rate nonuniform quantizer then, consider plotting MSE and $\mathsf{H}[V]$ on Figure 3.9. The corresponding approximate values of MSE and $\mathsf{H}[V]$ are then close to the plotted value (with some small difference both in the ordinate and abscissa). These approximate values, however, lie above the approximate values plotted in Figure 3.9 for the scalar quantizer. Thus, in this sense, the performance curve of MSE versus $\mathsf{H}[V]$ for the approximation to the scalar quantizer either lies below or close to the points for any nonuniform quantizer.

In summary, it has been shown that for large $\mathsf{H}[V]$ (*i.e.*, high-rate quantization), a uniform scalar quantizer approximately minimizes MSE subject to the entropy constraint. There is little reason to use nonuniform scalar quantizers (except perhaps at low rate). Furthermore the MSE performance at high-rate can be easily approximated and depends only on $\mathsf{h}[U]$ and the constraint on $\mathsf{H}[V]$.

3B Appendix B: Nonuniform 2D quantizers

For completeness, the performance of nonuniform 2D quantizers is now analyzed; the analysis is very similar to that of nonuniform scalar quantizers. Consider an arbitrary set of quantization intervals $\{\mathcal{R}_j\}$. Let $A(\mathcal{R}_j)$ and MSE_j be the area and mean-squared error per dimension respectively of $\mathcal{R}_j$, *i.e.*,

$$A(\mathcal{R}_j) = \int_{\mathcal{R}_j} d\boldsymbol{u}\ ; \qquad \mathrm{MSE}_j = \frac{1}{2} \int_{\mathcal{R}_j} \frac{\|\boldsymbol{u} - \boldsymbol{a}_j\|^2}{A(\mathcal{R}_j)}\; d\boldsymbol{u},$$

where $\boldsymbol{a}_j$ is the mean of $\mathcal{R}_j$. For each region $\mathcal{R}_j$ and each $\boldsymbol{u} \in \mathcal{R}_j$, let $\overline{f}(\boldsymbol{u}) = \Pr(\mathcal{R}_j)/A(\mathcal{R}_j)$ be the average pdf in $\mathcal{R}_j$. Then

$$p_j = \int_{\mathcal{R}_j} f_{\boldsymbol{U}}(\boldsymbol{u})\; d\boldsymbol{u} = \overline{f}(\boldsymbol{u}) A(\mathcal{R}_j).$$

The unconditioned mean-squared error is then

$$\mathrm{MSE} = \sum_j p_j\ \mathrm{MSE}_j.$$

Let $A(\boldsymbol{u}) = A(\mathcal{R}_j)$ and $\text{MSE}(\boldsymbol{u}) = \text{MSE}_j$ for $\boldsymbol{u} \in A_j$. Then,

$$\text{MSE} = \int f_{\boldsymbol{U}}(\boldsymbol{u})\ \text{MSE}(\boldsymbol{u})\ d\boldsymbol{u}. \tag{3.24}$$

Similarly,

$$\begin{aligned}
\mathsf{H}[V] &= \sum_j -p_j \log p_j \\
&= \int -f_{\boldsymbol{U}}(\boldsymbol{u}) \log[\overline{f}(\boldsymbol{u})A(\boldsymbol{u})]\ d\boldsymbol{u} \\
&\approx \int -f_{\boldsymbol{U}}(\boldsymbol{u}) \log[f_{\boldsymbol{U}}(\boldsymbol{u})A(\boldsymbol{u})]\ d\boldsymbol{u} &\quad (3.25) \\
&= 2\mathsf{h}[U] - \int f_{\boldsymbol{U}}(\boldsymbol{u}) \log[A(\boldsymbol{u})]\ d\boldsymbol{u}. &\quad (3.26)
\end{aligned}$$

A Lagrange multiplier can again be used to solve for the optimum quantization regions under the high-rate approximation. In particular, from (3.24) and (3.26),

$$\text{MSE} + \lambda\mathsf{H}[V] \approx \lambda 2\mathsf{h}[U] + \int_{\mathbb{R}^2} f_{\boldsymbol{U}}(\boldsymbol{u})\ \{\text{MSE}(\boldsymbol{u}) - \lambda \log A(\boldsymbol{u})\}\ du. \tag{3.27}$$

Since each quantization area can be different, the quantization regions need not have geometric shapes whose translates tile the plane. As pointed out earlier, however, the shape that minimizes MSE_c for a given quantization area is a circle. Therefore the MSE can be lower bounded in the Lagrange multiplier by using this shape. Replacing $\text{MSE}(\boldsymbol{u})$ by $A(\boldsymbol{u})/(4\pi)$ in (3.27),

$$\text{MSE} + \lambda\mathsf{H}[V] \approx 2\lambda\mathsf{h}[U] + \int_{\mathbb{R}^2} f_{\boldsymbol{U}}(\boldsymbol{u}) \left\{ \frac{A(\boldsymbol{u})}{4\pi} - \lambda \log A(\boldsymbol{u}) \right\}\ du. \tag{3.28}$$

Optimizing for each $\boldsymbol{u}$ separately, $A(\boldsymbol{u}) = 4\pi\lambda \log e$. The optimum is achieved where the same size circle is used for each point $\boldsymbol{u}$ (independent of the probability density). This is unrealizable, but still provides a lower bound on the MSE for any given $\mathsf{H}[V]$ in the high-rate region. The reduction in MSE over the square region is $\pi/3 = 1.0472$ (0.2 dB). It appears that the uniform quantizer with hexagonal shape is optimal, but this figure of $\pi/3$ provides a simple bound to the possible gain with 2D quantizers. Either way, the improvement by going to two dimensions is small.

The same sort of analysis can be carried out for n dimensional quantizers. In place of using a circle as a lower bound, one now uses an n dimensional sphere. As n increases, the resulting lower bound to MSE approaches a gain of $\pi e/6 = 1.4233$ (1.53 dB) over the scalar quantizer. It is known from a fundamental result in information theory that this gain can be approached arbitrarily closely as $n \to \infty$.

3.E Exercises

3.1. Let U be an analog rv (rv) uniformly distributed between -1 and 1.

(a) Find the three-bit ($M = 8$) quantizer that minimizes the mean-squared error.

(b) Argue that your quantizer satisfies the necessary conditions for optimality.

(c) Show that the quantizer is unique in the sense that no other 3-bit quantizer satisfies the necessary conditions for optimality.

3.2. Consider a discrete-time, analog source *with memory*, *i.e.*, $U_1, U_2, \dots$ are dependent rv's. Assume that each U_k is uniformly distributed between 0 and 1 but that $U_{2n} = U_{2n-1}$ for each $n \geq 1$. Assume that $\{U_{2n}\}_{n=1}^{\infty}$ are independent.

(a) Find the one-bit ($M = 2$) scalar quantizer that minimizes the mean-squared error.

(b) Find the mean-squared error for the quantizer that you have found in (a).

(c) Find the one-bit-per-symbol ($M = 4$) two-dimensional vector quantizer that minimizes the MSE.

(d) Plot the two-dimensional regions and representation points for both your scalar quantizer in part (a) and your vector quantizer in part (c).

3.3. Consider a binary scalar quantizer that partitions the reals $\mathbb{R}$ into two subsets, $(-\infty, b]$ and (b, ∞) and then represents $(-\infty, b]$ by $a_1 \in \mathbb{R}$ and (b, ∞) by $a_2 \in \mathbb{R}$. This quantizer is used on each letter U_n of a sequence $\cdots, U_{-1}, U_0, U_1, \cdots$ of iid random variables, each having the probability density $f(u)$. Assume throughout this exercise that $f(u)$ is symmetric, *i.e.*, that $f(u) = f(-u)$ for all $u \geq 0$.

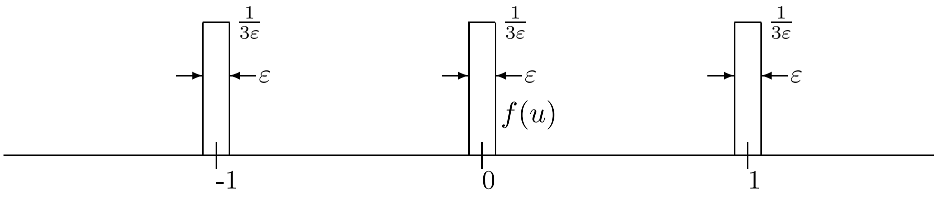
(a) Given the representation levels a_1 and $a_2 > a_1$, how should b be chosen to minimize the mean square distortion in the quantization? Assume that $f(u) > 0$ for $a_1 \leq u \leq a_2$ and explain why this assumption is relevant.

(b) Given $b \geq 0$, find the values of a_1 and a_2 that minimize the mean square distortion. Give both answers in terms of the two functions $Q(x) = \int_x^\infty f(u)\, du$ and $y(x) = \int_x^\infty u f(u)\, du$.

(c) Show that for $b = 0$, the minimizing values of a_1 and a_2 satisfy $a_1 = -a_2$.

(d) Show that the choice of b, a_1, and a_2 in part (c) satisfies the Lloyd-Max conditions for minimum mean square distortion.

(e) Consider the particular symmetric density below



Find all sets of triples, $\{b, a_1, a_2\}$ that satisfy the Lloyd-Max conditions and evaluate the MSE for each. You are welcome in your calculation to replace each region of non-zero probability density above with an impulse *i.e.*, $f(u) = \frac{1}{3}[\delta(-1) + \delta(0) + \delta(1)]$, but you should use the figure above to resolve the ambiguity about regions that occurs when b is -1, 0, or +1.

(f) Give the MSE for each of your solutions above (in the limit of $\varepsilon \to 0$). Which of your solutions minimizes the MSE?

3.4. In Section 3.4, we partly analyzed a minimum-MSE quantizer for a pdf in which $f_U(u) = f_1$ over an interval of size L_1, $f_U(u) = f_2$ over an interval of size L_2 and $f_U(u) = 0$ elsewhere. Let M be the total number of representation points to be used, with M_1 in the first interval and $M_2 = M - M_1$ in the second. Assume (from symmetry) that the quantization intervals are of equal size $\Delta_1 = L_1/M_1$ in interval 1 and of equal size $\Delta_2 = L_2/M_2$ in interval 2. Assume that M is very large, so that we can approximately minimize the MSE over M_1, M_2 without an integer constraint on M_1, M_2 (that is, assume that M_1, M_2 can be arbitrary real numbers).

(a) Show that the MSE is minimized if $\Delta_1 f_1^{1/3} = \Delta_2 f_2^{1/3}$, *i.e.*, the quantization interval sizes are inversely proportional to the cube root of the density. [Hint: Use a Lagrange multiplier to perform the minimization. That is, to minimize a function $\mathrm{MSE}(\Delta_1, \Delta_2)$ subject to a constraint $M = f(\Delta_1, \Delta_2)$, first minimize $\mathrm{MSE}(\Delta_1, \Delta_2) + \lambda f(\Delta_1, \Delta_2)$ without the constraint, and, second, choose λ so that the solution meets the constraint.]

(b) Show that the minimum MSE under the above assumption is given by

$$\mathrm{MSE} = \frac{\left(L_1 f_1^{1/3} + L_2 f_2^{1/3}\right)^3}{12M^2}.$$

(c) Assume that the Lloyd-Max algorithm is started with $0 < M_1 < M$ representation points in the first interval and $M_2 = M - M_1$ points in the second interval. Explain where the Lloyd-Max algorithm converges for this starting point. Assume from here on that the distance between the two intervals is very large.

(d) Redo part (c) under the assumption that the Lloyd-Max algorithm is started with $0 < M_1 \le M - 2$ representation points in the first interval, one point between the two intervals, and the remaining points in the second interval.

(e) Express the exact minimum MSE as a minimum over $M - 1$ possibilities, with one term for each choice of $0 < M_1 < M$ (assume there are no representation points between the two intervals).

(f) Now consider an arbitrary choice of Δ_1 and Δ_2 (with no constraint on M). Show that the entropy of the set of quantization points is

$$H(V) = -f_1 L_1 \log(f_1\Delta_1) - f_2 L_2 \log(f_2 \Delta_2).$$

(g) Show that if we minimize the MSE subject to a constraint on this entropy (ignoring the integer constraint on quantization levels), then $\Delta_1 = \Delta_2$.

3.5. Assume that a continuous valued rv Z has a probability density that is 0 except over the interval $[-A, +A]$. Show that the differential entropy $h(Z)$ is upper bounded by $1 + \log_2 A$.

(b) Show that $h(Z) = 1 + \log_2 A$ if and only if Z is uniformly distributed between $-A$ and $+A$.

3.6. Let $f_U(u) = 1/2 + u$ for $0 < u \le 1$ and $f_U(u) = 0$ elsewhere.

(a) For $\Delta < 1$, consider a quantization region $\mathcal{R} = (x, x + \Delta]$ for $0 < x \le 1 - \Delta$. Find the conditional mean of U conditional on $U \in \mathcal{R}$.

(b) Find the conditional mean-squared error (MSE) of U conditional on $U \in \mathcal{R}$. Show that, as Δ goes to 0, the difference between the MSE and the approximation $\Delta^2/12$ goes to 0 as Δ^4.

(c) For any given Δ such that $1/\Delta = M$, M a positive integer, let $\{\mathcal{R}_j = ((j-1)\Delta, j\Delta]\}$ be the set of regions for a uniform scalar quantizer with M quantization intervals. Show that the difference between $\mathsf{h}[U] - \log \Delta$ and $\mathsf{H}[V]$ as given (3.10) is

$$\mathsf{h}[U] - \log \Delta - \mathsf{H}[V] = \int_0^1 f_U(u) \log[\overline{f}(u)/f_U(u)]\ du.$$

(d) Show that the difference in (3.6) is nonnegative. Hint: use the inequality $\ln x \leq x - 1$. Note that your argument does not depend on the particular choice of $f_U(u)$.

(e) Show that the difference $\mathsf{h}[U] - \log \Delta - \mathsf{H}[V]$ goes to 0 as Δ^2 as $\Delta \to 0$. Hint: Use the approximation $\ln x \approx (x-1) - (x-1)^2/2$, which is the second-order Taylor series expansion of $\ln x$ around $x = 1$.

The major error in the high-rate approximation for small Δ and smooth $f_U(u)$ is due to the slope of $f_U(u)$. Your results here show that this linear term is insignificant for both the approximation of MSE and for the approximation of $\mathsf{H}[V]$. More work is required to validate the approximation in regions where $f_U(u)$ goes to 0.

3.7. (Example where $\mathsf{h}(U)$ is infinite.) Let $f_U(u)$ be given by

$$f_U(u) = \begin{cases} \frac{1}{u(\ln u)^2} & \text{for} \quad u \geq e \\ 0 & \text{for} \quad u < e, \end{cases}$$

(a) Show that $f_U(u)$ is non-negative and integrates to 1.

(b) Show that $\mathsf{h}(U)$ is infinite.

(c) Show that a uniform scalar quantizer for this source with any separation Δ $(0 < \Delta < \infty)$ has infinite entropy. Hint: Use the approach in Exercise 3.6, parts (c, d.)

3.8. (Divergence and the extremal property of Gaussian entropy) The divergence between two probability densities $f(x)$ and $g(x)$ is defined by

$$D(f\|g) = \int_{-\infty}^{\infty} f(x) \ln \frac{f(x)}{g(x)}\, dx$$

(a) Show that $D(f\|g) \geq 0$. Hint: use the inequality $\ln y \leq y - 1$ for $y \geq 0$ on $-D(f\|g)$. You may assume that $g(x) > 0$ where $f(x) > 0$.

(b) Let $\int_{-\infty}^{\infty} x^2 f(x)\, dx = \sigma^2$ and let $g(x) = \phi(x)$ where $\phi(x)$ is the density of the rv $\mathcal{N}(0, \sigma^2)$. Express $D(f\|\phi(x))$ in terms of the differential entropy (in nats) of a rv with density $f(x)$.

(c) Use (a) and (b) to show that the Gaussian rv $\mathcal{N}(0, \sigma^2)$ has the largest differential entropy of any rv with variance σ^2 and that that differential entropy is $\frac{1}{2}\ln(2\pi e \sigma^2)$.

3.9. Consider a discrete source U with a finite alphabet of N real numbers, $r_1 < r_2 < \cdots < r_N$ with the pmf $p_1 > 0, \ldots, p_N > 0$. The set $\{r_1, \ldots, r_N\}$ is to be quantized into a smaller set of $M < N$ representation points $a_1 < a_2 < \cdots < a_M$.

(a) Let $\mathcal{R}_1, \mathcal{R}_2, \dots, \mathcal{R}_M$ be a given set of quantization intervals with $\mathcal{R}_1 = (-\infty, b_1], \mathcal{R}_2 = (b_1, b_2], \dots, \mathcal{R}_M = (b_{M-1}, \infty)$. Assume that at least one source value r_i is in $\mathcal{R}_j$ for each j, $1 \leq j \leq M$ and give a necessary condition on the representation points $\{a_j\}$ to achieve minimum MSE.

(b) For a given set of representation points $a_1, \dots, a_M$ assume that no symbol r_i lies exactly halfway between two neighboring a_i, *i.e.*, that $r_i \neq \frac{a_j+a_{j+1}}{2}$ for all i, j. For each r_i, find the interval $\mathcal{R}_j$ (and more specifically the representation point a_j) that r_i must be mapped into to minimize MSE. Note that it is not necessary to place the boundary b_j between $\mathcal{R}_j$ and $\mathcal{R}_{j+1}$ at $b_j = (a_j + a_{j+1})/2$ since there is no probability in the immediate vicinity of $(a_j + a_{j+1})/2$.

(c) For the given representation points, $a_1, \dots, a_M$, now assume that $r_i = \frac{a_j+a_{j+1}}{2}$ for some source symbol r_i and some j. Show that the MSE is the same whether r_i is mapped into a_j or into a_{j+1}.

(d) For the assumption in part c), show that the set $\{a_j\}$ cannot possibly achieve minimum MSE. Hint: Look at the optimal choice of a_j and a_{j+1} for each of the two cases of part c).

3.10. Assume an iid discrete-time analog source $U_1, U_2, \cdots$ and consider a scalar quantizer that satisfies the Lloyd-Max conditions. Show that the rectangular 2-dimensional quantizer based on this scalar quantizer also satisfies the Lloyd-Max conditions.

3.11. (a) Consider a square two dimensional quantization region $\mathcal{R}$ defined by $-\frac{\Delta}{2} \leq u_1 \leq \frac{\Delta}{2}$ and $-\frac{\Delta}{2} \leq u_2 \leq \frac{\Delta}{2}$. Find MSE_c as defined in (3.15) and show that it's proportional to Δ^2.

(b) Repeat part (a) with Δ replaced by $a\Delta$. Show that $\mathrm{MSE}_c/A(\mathcal{R})$ (where $A(\mathcal{R})$ is now the area of the scaled region) is unchanged.

(c) Explain why this invariance to scaling of $\mathrm{MSE}_c/A(\mathcal{R})$ is valid for any two dimensional region.