# DSP Homework 09

Xu, Minhuan

October 27, 2022

## Contents

**1 Videos** **1**

**2 Digital Number Representations** **1**
2.1 Exact Meanings of Common Representations . . . . . 1
2.1.1 Fixed-Point Data Types . . . . . 1
2.1.2 Floating-Point Data Types . . . . . 2
2.1.3 Analysis of Double Type Floating-point Numbers . . . . . 2
2.1.4 My Proposal . . . . . 2

**3 Lloyd-Max Quantization Algorithm** **2**

**4 Optimal Quantization Strategy When PDF Has Uniform Distribution** **2**
4.1 Find the Optimal Quantization Level . . . . . 3
4.2 Find the Optimal Quantization Interval . . . . . 3
4.3 Conclude the Optimal Quantization Strategy . . . . . 3

**5 Conclusion** **4**

**Abstract**

# 1 Videos

# 2 Digital Number Representations

## 2.1 Exact Meanings of Common Representations

### 2.1.1 Fixed-Point Data Types

In computer we use bits to store numbers, between these bits, there's a virtual point which divided this number into integer part and fraction part. Fixed-point means this point will not move. This point is usually placed to the tail of binary digits, so that this piece of digits can represent one integer.

Byte, Short Integer and Integer are fixed-point type. Details in Table 1.

| Type | Bytes | Signed Number range | Unsigned Number Range |
|---|---|---|---|
| Byte | 1 | -128 $\sim$ 127 | 0 $\sim$ 255 |
| Short Integer | 2 | -32,768 $\sim$ 32,767 | 0 $\sim$ 65535 |
| Integer | 4 | $-2^{31} \sim 2^{31}-1$ | $0 \sim 2^{32}-1$ |

Table 1: Common Properties of Fixed-Point Data Types

### 2.1.2 Floating-Point Data Types

Floating-point means that division point mentioned above can move. There are 3 parts of digits with floating-point.

First, sign, always 1 bit, represents whether this number is positive or negative.

Second, exponent, several bits, represents an non-negative number, assuming $n$. Engineers want the bits to represent the shift of the fraction number. So, engineers give $n$ an initial shift assuming $-N$, and this $N$ is called Exponent Bias. This means we can just think the exponent bits represents $10^{n-N}$ in binary.

Third, fraction, several bits, represents a number in $[0, 1)$, assuming $1.xxxx\cdots$. However, the 1 before the decimal point is ignored according to the rules of floating-point data type in order to save bits. So, in memory, fraction part is like $.xxxx\cdots$.

So, this floating-point number should be represented as $1.xxxx\cdots \times 10^n$ (all in binary except n). The arrangement of Floating-point Data is as below, see Table 2.

| Sign | Exponent | | | | Fraction | | | | |
|---|---|---|---|---|---|---|---|---|---|
| S | E | E | $\cdots$ | E | F | F | F | $\cdots$ | F |

Table 2: Common Arrangement of Floating-point Data

Float, Double, quadruple are floating-point data type. Details in

| Type | Bytes | Bits for Exponent | Bits for Fraction | Exponent Bias |
|---|---|---|---|---|
| Float | 4 | 8 | 23 | $2^7 - 1$ |
| Double | 8 | 11 | 52 | $2^{10} - 1$ |
| quadruple | 16 | 15 | 112 | $2^{14} - 1$ |

Table 3: Common Properties of Floating-point Data Types

### 2.1.3 Analysis of Double Type Floating-point Numbers

The classification of floating-point data types uses the minimum error the specific data type has. Double precision is the common name of binary64 which means 64 bits to represent numbers. The smallest absolute value (except 0) of double $\Delta$ is represented by bits like:

$$0,\ \overbrace{00\cdots0}^{\text{10 bits of 0}}\ 1,\ \overbrace{00\cdots0}^{\text{51 bits of 0}}\ ,1$$

According to the IEEE 754, the Exponent Bias for Double is 1023. We can find that (D: Decimal; B: Binary)

$$\Delta = 1.\ \overbrace{00\cdots0}^{\text{51 bits of 0}}\ 1B \times 10B^{1D-1023D} = 2D^{-1022D} + 2D^{-1074D} \approx 2.23 \times 10^{-308}$$

So, $\Delta$ also the precision of Double is $2.23 \times 10^{-308}$. So, the usual error of Double floating-point numbers is about $2.23 \times 10^{-308}$.

### 2.1.4 My Proposal

# 3 Lloyd-Max Quantization Algorithm

# 4 Optimal Quantization Strategy When PDF Has Uniform Distribution

We have the quantization error of J which can be described as below:

$$J = \sum_{i=1}^{M} \int_{b_{i-1}}^{b_i} [Qi(x) - x]^2 \, p(x) \, \mathrm{d}x \tag{1}$$

In the special case of uniform distribution, we Have

$$J = \sum_{i=1}^{M} \int_{b_{i-1}}^{b_i} [q_i - x]^2 \, \mathrm{d}x \tag{2}$$

### 4.1 Find the Optimal Quantization Level

To find the best $q_i$, take the partial derivative of $q_i$

$$\begin{aligned}\frac{\partial J}{\partial q_i} &= \sum_{i=1}^{M} \int_{b_{i-1}}^{b_i} \frac{\partial}{\partial q_i}(q_i^2 - 2q_i x + x^2)\,\mathrm{d}x \\ &= \sum_{i=1}^{M} \int_{b_{i-1}}^{b_i} (2q_i - 2x)\,\mathrm{d}x\end{aligned}$$

We want $\frac{\partial J}{\partial q_i} = 0$. Therefore

$$\begin{aligned}q_i \int_{b_{i-1}}^{b_i} \mathrm{d}x &= \int_{b_{i-1}}^{b_i} x\,\mathrm{d}x \\ q_i(b_i - b_{i-1}) &= \frac{1}{2}\,(b_i^2 - b_{i-1}^2)\end{aligned}$$

Therefore

$$q_i = \frac{b_i + b_{i-1}}{2} \tag{3}$$

### 4.2 Find the Optimal Quantization Interval

And, the same as the $q_i$, take the partial of $b_i$, and make $\frac{\partial J}{\partial b_i} = 0$

$$\begin{aligned}\frac{\partial J}{\partial b_i} &= \frac{\partial}{\partial b_i} \sum_{i=1}^{M} \int_{b_{i-1}}^{b_i} (q_i - x)^2\,\mathrm{d}x \\ &= \frac{\partial}{\partial b_i}[\int_{b_{i-1}}^{b_i} (q_i - x)^2\,\mathrm{d}x + \int_{b_i}^{b_{i+1}} (q_{i+1} - x)^2\,\mathrm{d}x] \\ &= 0\end{aligned}$$

We learned (4) in out freshman year that

$$\begin{aligned}\frac{d}{dx}\int_a^x f(t)\,\mathrm{d}t &= f(x) \\ \frac{d}{dx}\int_x^a f(t)\,\mathrm{d}t &= -f(x)\end{aligned} \tag{4}$$

Rewrite the (4.2)

$$\begin{aligned}(q_i - b_i)^2 &= (q_{i+1} - b_i)^2 \\ b_i - q_i &= q_{i+1} - b_i\end{aligned}$$

We can have the other result

$$b_i = \frac{q_{i+1} + q_i}{2} \tag{5}$$

### 4.3 Conclude the Optimal Quantization Strategy

If we combine (3) with (5). First, we can know that

$$\begin{aligned}q_i &= \frac{b_i + b_{i-1}}{2} \\ q_{i+1} &= \frac{b_{i+1} + b_i}{2}\end{aligned}$$

Put them in (5)

$$\begin{aligned}b_i &= \frac{q_i + q_{i+1}}{2} \\ &= \frac{1}{2}\,(b_i + \frac{b_{i-1} + b_{i+1}}{2})\end{aligned}$$

Therefore

$$b_i = \frac{b_{i-1} + b_{i+1}}{2}$$

We can easily know that $b_i$ is an arithmetic sequence, and because (3), $q_i$ is an arithmetic sequence too.

In conclusion, if $p(x)$ is in the special case of uniform distribution, the range of $[0, 1]$ should be equally divided into M parts, and $q_i$ should be the mean of $b_i$ and $b_{i+1}$.

# 5 Conclusion