3-2022

OPTIMAL QUANTIZATION FOR MIXED DISTRIBUTIONS GENERATED BY TWO UNIFORM DISTRIBUTIONS

Ashley Gomez
The University of Texas Rio Grande Valley

Olga Lopez
The University of Texas Rio Grande Valley

Mrinal Kanti Roychowdhury
The University of Texas Rio Grande Valley

OPTIMAL QUANTIZATION FOR MIXED DISTRIBUTIONS GENERATED BY TWO UNIFORM DISTRIBUTIONS

ASHLEY GOMEZ, OGLA LOPEZ, AND MRINAL KANTI ROYCHOWDHURY

Abstract. In this paper, for mixed distributions generated by two uniform distributions we investigate the optimal sets of n-means and the nth quantization errors for all positive integers n. Some conjectures and open problems are also given.

1. Introduction

The most common form of quantization is rounding-off. Its purpose is to reduce the cardinality of the representation space, in particular, when the input data is real-valued. It has broad application in engineering and technology (see [GG, GN, Z]). For mathematical treatment of quantization one is referred to Graf-Luschgy's book (see [GL1]).

Let $\mathbb{R}^d$ denote the d-dimensional Euclidean space equipped with a metric $\|\cdot\|$ compatible with the Euclidean topology. Let P be a Borel probability measure on $\mathbb{R}^d$ and α be a finite subset of $\mathbb{R}^d$. Then, $\int \min_{a\in\alpha} \|x-a\|^2 dP(x)$ is often referred to as the *cost,* or *distortion error* for α with respect to the probability measure P, and is denoted by $V(P;\alpha)$. Write $\mathcal{D}_n := \{\alpha \subset \mathbb{R}^d : 1 \leq \text{card}(\alpha) \leq n\}$. Then, $\inf\{V(P;\alpha) : \alpha \in \mathcal{D}_n\}$ is called the *nth quantization error* for the probability measure P, and is denoted by $V_n := V_n(P)$. A set α for which the infimum occurs and contains no more than n points is called an *optimal set of n-means.* Since $\int \|x\|^2 dP(x) < \infty$ such a set α always exists (see [AW, GKL, GL1, GL2]). For some recent work in this direction one can see [CR, DR1, DR2, GL3, L1, R1, R2, R3, R4, R5, R6, RR1].

Let us now state the following proposition (see [GG, GL1]):

Proposition 1.1. *Let α be an optimal set of n-means for P, and $a \in \alpha$. Then,*
(i) $P(M(a|\alpha)) > 0$, (ii) $P(\partial M(a|\alpha)) = 0$, (iii) $a = E(X : X \in M(a|\alpha))$, where $M(a|\alpha)$ is the Voronoi region of $a \in \alpha$, i.e., $M(a|\alpha)$ is the set of all elements x in $\mathbb{R}^d$ which are closest to a among all the elements in α.

Proposition 1.1 says that if α is an optimal set and $a \in \alpha$, then a is the *conditional expectation* of the random variable X given that X takes values in the Voronoi region of a. The following theorem is known.

Theorem 1.2. *(see [RR2]) Let P be a uniform distribution on the closed interval $[a,b]$. Then, the optimal set n-means is given by $\alpha_n := \{a + \frac{2i-1}{2n}(b-a) : 1 \leq i \leq n\}$, and the corresponding quantization error is $V_n := V_n(P) = \frac{(a-b)^2}{12n^2}$.*

Mixed distributions are an exciting new area for optimal quantization. For any two Borel probability measures P_1 and P_2, and $p \in (0,1)$, if $P := pP_1 + (1-p)P_2$, then the probability measure P is called the *mixture* or the *mixed distribution* generated by the probability measures (P_1, P_2) associated with the probability vector $(p, 1-p)$. Let P_1 and P_2 be two uniform distributions on the two disconnected line segments $J_1 := [0, \frac{1}{3}]$ and $J_2 := [\frac{2}{3}, 1]$ of equal lengths, and P be a mixed distribution generated by (P_1, P_2) associated with a probability vector $(p, 1-p)$. In this paper, for three different mixed distributions, in Section 2 for $p = \frac{1}{100}$ and in Section 3 for $p = \frac{2}{5}$, and for $p = \frac{1}{1000}$, we determine the optimal sets of n-means and the nth quantization errors for all $n \in \mathbb{N}$. Using the similar technique, given in this paper, one can investigate the

2010 *Mathematics Subject Classification.* 60Exx, 94A34.
Key words and phrases. Mixed distribution, uniform distribution, optimal sets of n-means, quantization error.

optimal sets of n-means and the nth quantization errors for all $n \in \mathbb{N}$ for any mixed distribution P generated by (P_1, P_2) associated with any probability vectors $(p, 1-p)$. In this regard, at the end of Section 3 we give a conjecture Conjecture 3.4, and two open problems Open 3.5 and Open 3.6. Under a conjecture in Section 4, we give a partial answer of the open problem Open 3.6.

2. Quantization for the mixed distribution P when $p = \frac{1}{100}$

Let P_1 and P_2 be two uniform distributions, respectively, on the intervals given by

$$J_1 := [0, \frac{1}{3}], \text{ and } J_2 := [\frac{2}{3}, 1].$$

Let f_1 and f_2 be their respective density functions. Then, $f_1(x) = 3$ if $x \in [0, \frac{1}{3}]$, and zero otherwise; and $f_2(x) = 3$ if $x \in [\frac{2}{3}, 1]$, and zero otherwise. The underlying mixed distribution considered is given by $P := pP_1 + (1-p)P_2$, where $p = \frac{1}{100}$. By $E(X)$ we mean the expectation of a random variable X with distribution P, and $V(X)$ represents the variance of X. By $\alpha_n(\mu)$, we denote an optimal set of n-means with respect to a probability distribution μ, and $V_n(\mu)$ represents the corresponding quantization error for n-means. If μ is the mixed distribution P, sometimes we denote them by α_n instead of $\alpha_n(P)$, and the corresponding quantization error by V_n instead of $V_n(P)$.

Proposition 2.1. *Let P be the mixed distribution defined by $P = pP_1 + (1-p)P_2$. Then, $E(X) = \frac{1}{6}(5-4p)$, and $V(X) = \frac{1}{108}(-48p^2 + 48p + 1)$.*

Proof. We have

$$E(X) = \int x dP = p \int x d(P_1(x)) + (1-p) \int x d(P_2(x)) = p \int_{J_1} 3x\, dx + (1-p) \int_{J_2} 3x\, dx$$

yielding $E(X) = \frac{1}{6}(5-4p)$, and

$$V(P) = \int (x - E(X))^2 dP = p \int (x - E(X))^2 d(P_1(x)) + (1-p) \int (x - E(X))^2 d(P_2(x)),$$

implying $V(P) = \frac{1}{108}(-48p^2 + 48p + 1)$, and thus, the proposition is yielded. $\square$

Remark 2.2. The optimal set of one-mean is the set $\{\frac{1}{6}(5-4p)\}$, and the corresponding quantization error is the variance $V := V(X)$ of a random variable with distribution $P := pP_1 + (1-p)P_2$. Recall that in our case, $p = \frac{1}{100}$, and then $E(X) = \frac{62}{75}$ and $V(X) = \frac{461}{33750}$.

Proposition 2.3. *The optimal set of two-means is $\{0.731517, 0.910506\}$ with quantization error $V_2 = 0.005682$.*

Proof. Let $\alpha := \{a_1, a_2\}$ be an optimal set of two-means. Since the points in an optimal set are the conditional expectations in their own Voronoi regions, without any loss of generality, we can assume that $0 < a_1 < a_2 < 1$. If $\frac{1}{3} < a_1 < a_2 < \frac{2}{3}$, then the quantization error can be strictly reduced by moving the point a_1 to $\frac{1}{3}$, and a_2 to $\frac{2}{3}$, and so, $\frac{1}{3} < a_1 < a_2 < \frac{2}{3}$ can not happen. Let us now discuss all the possible cases:

Case 1. $0 < a_1 < a_2 \leq \frac{1}{3}$.

Since the boundary of the Voronoi region is $\frac{1}{2}(a_1 + a_2)$, we have the distortion error as

$$\int \min_{a \in \alpha}(x-a)^2 dP = \int_0^{\frac{a_1+a_2}{2}} (x-a_1)^2 dP + \int_{\frac{a_1+a_2}{2}}^{\frac{1}{3}} (x-a_2)^2 dP + \int_{\frac{2}{3}}^{1} (x-a_2)^2 dP$$
$$= \frac{1}{100} \int_0^{\frac{1}{2}(a_1+a_2)} 3\,(x-a_1)^2\, dx + \frac{1}{100} \int_{\frac{1}{2}(a_1+a_2)}^{\frac{1}{3}} 3\,(x-a_2)^2\, dx + \frac{99}{100} \int_{\frac{2}{3}}^{1} 3\,(x-a_2)^2\, dx$$
$$= \frac{81a_1^3 + 81a_2a_1^2 - 81a_2^2a_1 - 81a_2^3 + 10800a_2^2 - 17856a_2 + 7528}{10800},$$

the minimum value of which is $\frac{3119}{12150}$ and it occurs when $a_1 = \frac{1}{9}$, and $a_2 = \frac{1}{3}$.

Case 2. $0 < a_1 < \frac{1}{3} < a_2 < \frac{2}{3}$.

In this case, the boundary $\frac{1}{2}(a_1 + a_2)$ of the Voronoi regions of a_1 and a_2 must satisfy $0 < a_1 < \frac{1}{2}(a_1 + a_2) < \frac{1}{3}$, otherwise the quantization error can be strictly reduced by moving the point a_2 to $\frac{2}{3}$. Hence, the distortion error in this case is given by

$$\begin{aligned}
&\int \min_{a\in\alpha}(x-a)^2 dP \\
&= \frac{1}{100}\int_0^{\frac{1}{2}(a_1+a_2)} 3\,(x-a_1)^2\,dx + \frac{1}{100}\int_{\frac{1}{2}(a_1+a_2)}^{\frac{1}{3}} 3\,(x-a_2)^2\,dx + \frac{99}{100}\int_{\frac{2}{3}}^{1} 3\,(x-a_2)^2\,dx \\
&= \frac{81a_1^3 + 81a_2a_1^2 - 81a_2^2a_1 - 81a_2^3 + 10800a_2^2 - 17856a_2 + 7528}{10800},
\end{aligned}$$

the minimum value of which is $\frac{89}{2430}$ and it occurs when $a_1 = \frac{2}{9}$, and $a_2 = \frac{2}{3}$.

Case 3. $0 < a_1 \leq \frac{1}{3} < \frac{2}{3} \leq a_2$.

In this case, the Voronoi region of a_1 does not contain any point from J_2, if it does, then we must have $\frac{1}{2}(a_1 + a_2) > \frac{2}{3}$ implying $a_2 > \frac{4}{3} - a_1 \geq \frac{4}{3} - \frac{1}{3} = 1$, which is a contradiction as $a_2 < 1$. Similarly, we can show that the Voronoi region of a_2 does not contain any point from J_1. This yields the fact that

$$a_1 = E(X : X \in J_1) = \frac{1}{6}, \text{ and } a_2 = E(X : X \in J_2) = \frac{5}{6},$$

with distortion error

$$\int \min_{a\in\alpha}(x-a)^2 dP = \frac{1}{100}\int_0^{\frac{1}{3}} 3(x-\frac{1}{6})^2\,dx + \frac{99}{100}\int_{\frac{2}{3}}^{1} 3(x-\frac{5}{6})^2\,dx = \frac{1}{108}.$$

Case 4. $\frac{1}{3} < a_1 \leq \frac{2}{3} < a_2$.

In this case, the boundary $\frac{1}{2}(a_1 + a_2)$ of the Voronoi regions of a_1 and a_2 must satisfy $\frac{2}{3} < \frac{1}{2}(a_1 + a_2) < a_2 < 1$, otherwise the quantization error can be strictly reduced by moving the point a_1 to $\frac{1}{3}$. Hence, the distortion error in this case is given by

$$\begin{aligned}
&\int \min_{a\in\alpha}(x-a)^2 dP \\
&= \frac{1}{100}\int_0^{\frac{1}{3}} 3\,(x-a_1)^2\,dx + \frac{99}{100}\int_{\frac{2}{3}}^{\frac{1}{2}(a_1+a_2)} 3\,(x-a_1)^2\,dx + \frac{99}{100}\int_{\frac{1}{2}(a_1+a_2)}^{1} 3\,(x-a_2)^2\,dx \\
&= \frac{8019a_1^3 + 27\,(297a_2 - 788)\,a_1^2 - 9\,(891a_2^2 - 1580)\,a_1 - 8019a_2^3 + 32076a_2^2 - 32076a_2 + 7528}{10800},
\end{aligned}$$

the minimum value of which is $\frac{1}{150}$ and it occurs when $a_1 = \frac{2}{3}$, and $a_2 = \frac{8}{9}$.

Case 5. $\frac{2}{3} < a_1 < a_2 < 1$.

In this case, the distortion error is given by

$$\begin{aligned}
&\int \min_{a\in\alpha}(x-a)^2 dP \\
&= \frac{1}{100}\int_0^{\frac{1}{3}} 3\,(x-a_1)^2\,dx + \frac{99}{100}\int_{\frac{2}{3}}^{\frac{1}{2}(a_1+a_2)} 3\,(x-a_1)^2\,dx + \frac{99}{100}\int_{\frac{1}{2}(a_1+a_2)}^{1} 3\,(x-a_2)^2\,dx \\
&= \frac{8019a_1^3 + 27\,(297a_2 - 788)\,a_1^2 - 9\,(891a_2^2 - 1580)\,a_1 - 8019a_2^3 + 32076a_2^2 - 32076a_2 + 7528}{10800},
\end{aligned}$$

the minimum value of which is 0.005682 and it occurs when $a_1 = 0.731517$, and $a_2 = 0.910506$.

Comparing the distortion errors obtained in all the above possible cases, we see that the distortion error in Case 5 is smallest. Thus, the optimal set of two-means is $\{0.731517, 0.910506\}$ with quantization error $V_2 = 0.005682$, which is the proposition. $\square$

Proposition 2.4. *Optimal set of three-means is* $\{\frac{1}{6}, \frac{3}{4}, \frac{11}{12}\}$ *with quantization error* $V_3 = \frac{103}{43200}$.

Proof. Let $\alpha = \{a_1, a_2, a_3\}$ be an optimal set of three-means. Proposition 1.1 implies that if α contains a point from the open interval $(\frac{1}{3}, \frac{2}{3})$, it cannot contain more than one point from the open interval $(\frac{1}{3}, \frac{2}{3})$. First, we assume that α contains one point from J_1, and two points from J_2. Then, $0 < a_1 \leq \frac{1}{3} < \frac{2}{3} \leq a_2 < a_3 < 1$ yielding the fact that the Voronoi region of a_1 does not contain any point from J_1, and the Voronoi region of a_2, and so of a_3 cannot not contain any point from J_1. This yields $a_1 = \frac{1}{6}$, $a_2 = \frac{3}{4}$, and $a_3 = \frac{11}{12}$ with distortion error

$$\int \min_{a\in\alpha}(x-a)^2 dP = \frac{1}{100}\int_0^{\frac{1}{3}} 3(x-\frac{1}{6})^2\,dx + \frac{99}{100}\int_{\frac{2}{3}}^{\frac{5}{6}} 3(x-\frac{3}{4})^2\,dx + \frac{99}{100}\int_{\frac{5}{6}}^{1} 3(x-\frac{11}{12})^2\,dx$$

yielding

$$\int \min_{a\in\alpha}(x-a)^2 dP = \frac{103}{43200}.$$

Since V_3 is the quantization error for three-means, we have $V_3 \leq \frac{103}{43200} = 0.00238426$. If $a_3 < \frac{2}{3}$, then

$$V_3 \geq \frac{99}{100}\int_{\frac{2}{3}}^{1} 3(x-\frac{2}{3})^2\,dx = \frac{11}{300} > V_3,$$

which is a contradiction. Hence, we can assume that $\frac{2}{3} < a_3$. Suppose that $a_2 < \frac{2}{3}$. Then,

$$V_3 \geq \int_{J_3} \min_{a\in\alpha}(x-a)^2 dP = \int_{J_3} \min_{a\in\{a_2,a_3\}}(x-a)^2 dP \geq \int_{J_3} \min_{a\in\{\frac{2}{3},a_3\}}(x-a)^2 dP$$

implying

$$\begin{aligned} V_3 &\geq \frac{99}{100}\int_{\frac{2}{3}}^{\frac{1}{2}(a_3+\frac{2}{3})} 3(x-\frac{2}{3})^2\,dx + \frac{99}{100}\int_{\frac{1}{2}\left(a_3+\frac{2}{3}\right)}^{1} 3(x-a_3)^2\,dx \\ &= -\frac{11(81a_3^3 - 270a_3^2 + 288a_3 - 100)}{1200}, \end{aligned}$$

the minimum value of which is $\frac{11}{2700}$ and it occurs when $a_3 = \frac{8}{9}$. Notice that $\frac{11}{2700} = 0.00407407 > V_3$, which leads to a contradiction. Hence, we can assume that $\frac{2}{3} \leq a_2 < a_3 < 1$. Notice that for $\frac{2}{3} \leq a_2 < a_3 < 1$, the Voronoi region of a_2 does not contain any point from J_1. Suppose that $\frac{2}{3} \leq a_1$. Then,

$$V_3 > \frac{1}{100}\int_0^{\frac{1}{3}} 3(x-\frac{2}{3})^2\,dx = \frac{7}{2700} = 0.00259259 > V_3,$$

which leads to a contradiction. So, we can assume that $a_1 < \frac{2}{3}$. Suppose that $\frac{1}{3} < a_1 < \frac{2}{3}$. Then, we must have $\frac{2}{3} < \frac{1}{2}(a_1+a_2) < a_2 < a_3 < 1$ yielding the distortion error as

$$\begin{aligned} &\int \min_{a\in\alpha}(x-a)^2 dP \\ &= \frac{1}{100}\int_0^{\frac{1}{3}} 3\,(x-a_1)^2\,dx + \frac{99}{100}\int_{\frac{2}{3}}^{\frac{1}{2}(a_1+a_2)} 3\,(x-a_1)^2\,dx + \frac{99}{100}\int_{\frac{1}{2}(a_1+a_2)}^{\frac{1}{2}(a_2+a_3)} 3\,(x-a_2)^2\,dx \\ &\qquad + \frac{99}{100}\int_{\frac{1}{2}(a_2+a_3)}^{1} 3\,(x-a_3)^2\,dx \\ &= \frac{1}{10800}\Big(8019a_1^3 + 27(297a_2 - 788)a_1^2 - 9(891a_2^2 - 1580)a_1 - 8019a_3^3 - 8019(a_2-4)a_3^2 \\ &\qquad + 8019(a_2^2-4)a_3 + 7528\Big) \end{aligned}$$

the minimum value of which is $\frac{137}{33750}$, and it occurs when $a_1 = \frac{2}{3}$, $a_2 = \frac{4}{5}$, and $a_3 = \frac{14}{15}$. Notice that $\frac{137}{33750} = 0.00405926 > V_3$, and thus a contradiction arises. Hence, we can assume that $a_1 \leq \frac{1}{3}$.

Thus, as described before, we deduce that $a_1 = \frac{1}{6}$, $a_2 = \frac{3}{4}$, and $a_3 = \frac{11}{12}$, and the quantization error for three-means is $V_3 = \frac{103}{43200}$. Thus, the proof of the proposition is complete. $\square$

Lemma 2.5. *For $n \geq 3$ let α_n be an optimal set of n-means for P. Then, $\alpha_n \cap J_1 \neq \emptyset$ and $\alpha_n \cap J_2 \neq \emptyset$.*

Proof. By Proposition 2.4, the lemma is true for $n = 3$. Let us now prove the lemma for $n \geq 4$. The distortion error due to the set $\beta := \{\frac{1}{6}, \frac{2}{3} + \frac{1}{18}, \frac{2}{3} + \frac{3}{18}, \frac{2}{3} + \frac{5}{18}\}$ is given by

$$\int \min_{a\in\beta}(x-a)^2 dP = \int_{J_1}(x-\frac{1}{6})^2 dP + \int_{J_2} \min_{a\in(\beta\setminus\{\frac{1}{6}\})}(x-a)^2 dP = \frac{1}{900}.$$

Since V_n is the quantization error for n-means, where $n \geq 4$, we have $V_n \leq \frac{1}{900} = 0.00111111$. For $n \geq 4$, let $\alpha_n := \{a_1, a_2, \cdots, a_n\}$ be an optimal set of n-means for P such that $0 < a_1 < a_2 < \cdots < a_n < 1$. If $a_n < \frac{2}{3}$, then

$$V_n > \int_{J_2}(x-\frac{2}{3})^2 dP = \frac{11}{300} = 0.0366667 > V_n,$$

which leads to a contradiction. Hence, we can assume that $a_n \geq \frac{2}{3}$, i.e., $\alpha_n \cap J_2 \neq \emptyset$. We now show that $\alpha_n \cap J_1 \neq \emptyset$. If $\frac{1}{2} \leq a_1$, then

$$V_n > \int_{J_1}(x-\frac{1}{2})^2 dP = \frac{13}{10800} = 0.0012037 > V_n,$$

which yields a contradiction. Hence, we can assume that $a_1 < \frac{1}{2}$. Assume that $\frac{1}{3} \leq a_1 < \frac{1}{2}$. Then, by Proposition 1.1, we must have $\frac{1}{2}(a_1 + a_2) > \frac{2}{3}$ yielding $a_2 > \frac{4}{3} - a_1 \geq \frac{4}{3} - \frac{1}{2} = \frac{5}{6}$, and so,

$$V_n > \int_{J_1}(x-\frac{1}{3})^2 dP + \int_{[\frac{2}{3},\frac{5}{6}]}(x-\frac{1}{2})^2 dP = \frac{2681}{21600} = 0.12412 > V_n,$$

which leads to a contradiction. Hence, we can assume that $a_1 < \frac{1}{3}$, i.e., $\alpha_n \cap J_1 \neq \emptyset$. Thus, the proof of the lemma is complete. $\square$

Lemma 2.6. *An optimal set of four-means does not contain any point from the open interval $(\frac{1}{3}, \frac{2}{3})$.*

Proof. Let $\alpha := \{a_1, a_2, a_3, a_4\}$, where $0 < a_1 < a_2 < a_3 < a_4 < 1$, be an optimal set of four-means. As mentioned in the proof of Lemma 2.5, we have $V_4 \leq \frac{1}{900} = 0.00111111$. Suppose that $a_2 \leq \frac{2}{3}$. Then,

$$V_4 > \frac{99}{100}\int_{\frac{2}{3}}^{\frac{1}{2}\left(a_3+\frac{2}{3}\right)} 3(x-\frac{2}{3})^2\,dx + \frac{99}{100}\int_{\frac{1}{2}\left(a_3+\frac{2}{3}\right)}^{\frac{1}{2}(a_3+a_4)} 3\left(x-a_3\right)^2 dx + \frac{99}{100}\int_{\frac{1}{2}(a_3+a_4)}^{1} 3\left(x-a_4\right)^2 dx$$
$$= \frac{11\left(-81a_4^3+324a_4^2-324a_4+27a_3^2\left(3a_4-2\right)-9a_3\left(9a_4^2-4\right)+100\right)}{1200}$$

the minimum value of which is $\frac{11}{7500}$, and it occurs when $a_3 = \frac{4}{5}$ and , $a_4 = \frac{14}{15}$ implying

$$V_4 > \frac{11}{7500} = 0.00146667 > V_4,$$

which leads to a contradiction. Thus, we can assume that $\frac{2}{3} < a_2$, and so $\frac{2}{3} < a_2 < a_3 < a_4$. Again, by Lemma 2.5, we see that $a_1 < \frac{1}{3}$. Hence, an optimal set of four-means does not contain any point from the open interval $(\frac{1}{3}, \frac{2}{3})$, which is the lemma. $\square$

Remark 2.7. Proceeding in the similar way as Lemma 2.6, we can show that the optimal set of five-means does not contain any point from the open interval $(\frac{1}{3}, \frac{2}{3})$.

Proposition 2.8. *For $n \geq 3$ let α_n be an optimal set of n-means for P. Then, α_n does not contain any point from the open interval $(\frac{1}{3}, \frac{2}{3})$. Moreover, the Voronoi region of any point in $\alpha_n \cap J_1$ does not contain any point from J_2, and the Voronoi region of any point in $\alpha_n \cap J_2$ does not contain any point from J_1.*

Proof. By Proposition 2.4, Lemma 2.6 and Remark 2.7, the proposition is true for $n = 3, 4, 5$. We now prove the proposition for $n \geq 6$. Let $\alpha_n := \{a_1, a_2, \cdots, a_n\}$, where $n \geq 6$, be an optimal set of n-means. Without any loss of generality, we can assume that $0 < a_1 < a_2 < \cdots < a_n < 1$. Let us now consider the set of six points $\beta := \{\frac{1}{6}, \frac{7}{10}, \frac{23}{30}, \frac{5}{6}, \frac{9}{10}, \frac{29}{30}\}$. By routine calculation, the distortion error due to the set β is given by

$$\int \min_{a\in\beta}(x-a)^2 dP = \frac{1}{100}\int_{J_1}(x-\frac{1}{6})^2 dP_1 + \frac{99}{100}\int_{J_2}\min_{b\in(\beta\setminus\{\frac{1}{6}\})}(x-b)^2 dP_2 = \frac{31}{67500},$$

and so, $V_6 \leq \frac{31}{67500} = 0.000459259$. Since V_n is the quantization error for six-means with $n \geq 6$, we have $V_n \leq V_6 \leq 0.000459259$. By Lemma 2.5, we know that $a_1 < \frac{1}{3}$ and $a_n > \frac{2}{3}$. Let k be the largest positive integer such that $a_k \leq \frac{1}{3}$. For the sake of contradiction, assume that α_n contains a point from the open interval $(\frac{1}{3}, \frac{2}{3})$. Then, by Proposition 1.1, we must have $a_{k+1} \in (\frac{1}{3}, \frac{2}{3})$, and $\frac{2}{3} \leq a_{k+2}$. The following two cases can arise:

Case 1. $\frac{1}{3} < a_{k+1} \leq \frac{1}{2}$.

Then, the Voronoi region of a_{k+1} must contain points from J_2, i.e., $\frac{1}{2}(a_{k+1}+a_{k+2}) \geq \frac{2}{3}$ implying $a_{k+2} \geq \frac{4}{3} - a_{k+1} \geq \frac{4}{3} - \frac{1}{2} = \frac{5}{6}$, otherwise the quantization error can be strictly reduced by moving the point a_{k+1} to $\frac{1}{3}$. Then,

$$V_n \geq \int_{[\frac{2}{3},\frac{5}{6}]}(x-\frac{5}{6})^2 dP = \frac{11}{2400} = 0.00458333 > V_n,$$

which is a contradiction.

Case 2. $\frac{1}{2} \leq a_{k+1} < \frac{2}{3}$.

Then, we must have $\frac{1}{2}(a_k + a_{k+1}) \leq \frac{1}{3}$ implying $a_k \leq \frac{2}{3} - a_{k+1} \leq \frac{2}{3} - \frac{1}{2} = \frac{1}{6}$, and so

$$V_n \geq \int_{[\frac{1}{6},\frac{1}{3}]}(x-\frac{1}{6})^2 dP = \frac{1}{21600} = 0.0000462963 > V_n,$$

which leads to a contradiction.

By Case 1 and Case 2, we deduce that α_n does not contain any point from the open interval $(\frac{1}{3}, \frac{2}{3})$. Thus, $\frac{2}{3} \leq a_{k+1}$. To complete the proof, assume that the Voronoi region of a_k contains points from J_2. Then, $\frac{1}{2}(a_k + a_{k+1}) > \frac{2}{3}$ implying $a_{k+1} > \frac{4}{3} - a_k \geq \frac{4}{3} - \frac{1}{3} = 1$, which is a contradiction. Similarly, we can show that if the Voronoi region of a_{k+1} contains points from J_1, then a contradiction arises. Thus, the proof of the proposition is complete. □

We are now ready to prove the following theorem.

Theorem 2.9. *For $n \geq 3$ let α_n be an optimal set of n-means for P. Let $card(\alpha_n \cap J_1) = k$. Then, α_n contains k elements from J_1, and $(n-k)$ elements from J_2, i.e., $\alpha_n(P) = \alpha_k(P_1) \cup \alpha_{n-k}(P_2)$ with quantization error*

$$V_n(P) = \frac{1}{324}\Big(\frac{1}{k^2} + \frac{2}{(n-k)^2}\Big).$$

Proof. By Proposition 2.8, we have $\alpha_n \cap J_1 \neq \emptyset$ and $\alpha_n \cap J_2 \neq \emptyset$. Thus, there exist two positive integers n_1 and n_2 such that $\text{card}(\alpha_n \cap J_1) = n_1$, and $\text{card}(\alpha_n \cap J_2) = n_2$. Again, by Proposition 2.8, α_n does not contain any point from the open interval $(\frac{1}{3}, \frac{2}{3})$, and so we have $n = n_1 + n_2$. Hence, by taking $n_1 = k$, we see that α_n contains k elements from J_1, and $(n-k)$ elements from J_2. Again, by Proposition 2.8, we know that the Voronoi region of any point in $\alpha_n \cap J_1$ does not contain any point from J_2, and the Voronoi region of any point from $\alpha_n \cap J_2$

does not contain any point from J_1. This implies the fact that $\alpha_n(P) = \alpha_k(P_1) \cup \alpha_{n-k}(P_2)$, and the corresponding quantization error is given by

$$V_n(P) = \frac{1}{100} V_k(P_1) + \frac{99}{100} V_{n-k}(P_2) = \frac{1}{10800}\Big(\frac{1}{k^2} + \frac{99}{(n-k)^2}\Big).$$

Thus, the proof of the theorem is complete. $\square$

Remark 2.10. Let k be the positive integer as stated in Theorem 2.9. By Theorem 1.2, $\alpha_k(P_1)$ and $\alpha_{n-k}(P_2)$ are known. Thus, once k is known, we can easily determine the optimal sets of n-means and the nth quantization errors for all $n \in \mathbb{N}$ with $n \geq 3$. For $n \geq 3$, consider the real valued function

$$F(n, x) = \frac{1}{10800}\Big(\frac{1}{x^2} + \frac{99}{(n-x)^2}\Big)$$

defined in the domain $1 \leq x \leq n-1$. Notice that $F(n, x)$ is concave upword, and so $F(n, x)$ attains its minimum at a unique x in the interval $[1, n-1]$. Thus, we can say that for a given positive integer $n \geq 3$, there exists a unique positive integer k for which $F(n, k)$ is minimum if x ranges over the positive integers in the interval $[1, n-1]$.

Remark 2.11. For $n \geq 3$ let us write

$$V(j, n-j) := \frac{1}{100} V_j(P_1) + \frac{99}{100} V_{n-j}(P_2),$$

where $1 \leq j \leq n-1$. For a given n let $k := k(n)$ be the positive integer as stated in Theorem 2.9. Then, we have $V_n = V(k, n-k)$. Notice that

$$V_n = V(k, n-k) = \min\{V(j, n-j) : 1 \leq j \leq n-1\}.$$

Moreover, if we order the elements of the set $\{V(j, n-j) : 1 \leq j \leq n-1\}$ in a sequence as

$$\{V(1, n-1), V(2, n-2), \cdots, V(n-1, 1)\}$$

then $V(k, n-k)$ is the kth term in the sequence. Using this fact, for a given n we can easily determine the value of the positive integer k as follows: Define the function

$$f : \mathbb{N} \to \mathbb{N} \text{ such that } f(n) = k, \tag{1}$$

where k is the unique positive integer such that

$$V_n := V(k, n-k) = \min\{V(j, n-j) : j \in \mathbb{N}, \ 1 \leq j \leq n-1\}.$$

For a given positive integer n, once $k := k(n)$ is known, using Theorem 1.2 and Theorem 2.9, we can determine the optimal set of n-means and the corresponding quantization error.

In the following example, we calculate the values of k for different values of n. For such calculations we have used Mathematica.

Example 2.12. Recall the function f defined by (1). Then, we see that

$$\{f(n)\}_{n=3}^{\infty} = \{1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 6, 7, 7, 7, 7, 7, 7, 8, 8, 8, 8, 8, 9, 9, 9, 9, 9, 9, 10, 10, 10, 10, 10, 10, 11, \cdots .\}.$$

In fact,

$$\{f(n)\}_{n=4985}^{5011} = \{886, 886, 886, 887, 887, 887, 887, 887, 887, 888, 888, 888, 888, 888, 889, 889, 889, 889, 889, 889, 890, 890, 890, 890, 890, 890, 891\}.$$

3. Quantization for the mixed distribution P when $p = \frac{2}{5}$, and $p = \frac{1}{1000}$

Let P_1 and P_2 be two uniform distributions, respectively, on the intervals given by

$$J_1 := [0, \frac{1}{3}], \text{ and } J_2 := [\frac{2}{3}, 1].$$

Let $P := pP_1 + (1-p)P_2$ be the mixed distribution generated by (P_1, P_2) associated with the probability vector $(p, 1-p)$, where $0 < p < 1$. For $n \in \mathbb{N}$ let α_n be an optimal set of n-means for P. Using the similar techniques as given in the previous section, we can show that if $p = \frac{2}{5}$, then

$$\alpha_1 = \{\frac{17}{30}\} \text{ with } V_1 = \frac{313}{2700}, \ \alpha_2 = \{\frac{1}{6}, \frac{5}{6}\} \text{ with } V_2 = \frac{1}{108}, \ \alpha_3 = \{\frac{1}{6}, \frac{3}{4}, \frac{11}{12}\} \text{ with } V_3 = \frac{11}{2160},$$
$$\alpha_4 = \{\frac{1}{12}, \frac{1}{4}, \frac{3}{4}, \frac{11}{12}\} \text{ with } V_4 = \frac{1}{432}, \text{ and so on.}$$

On the other hand, if $p = \frac{1}{1000}$, then we see that

$$\begin{aligned}
&\alpha_1 = \{\frac{1249}{1500}\} \text{ with } V_1 = 0.00970326,\\
&\alpha_2 = \{0.74824116, 0.91608039\} \text{ with } V_2 = 0.0026610135,\\
&\alpha_3 = \{0.719398, 0.831639, 0.94388\} \text{ with } V_3 = 0.00134412,\\
&\alpha_4 = \{0.704407, 0.788862, 0.873317, 0.957772\} \text{ with } V_4 = 0.00087869,\\
&\alpha_5 = \{\frac{1}{6}, \frac{17}{24}, \frac{19}{24}, \frac{7}{8}, \frac{23}{24}\} \text{ with } V_5 = 0.000587384 \text{ and so on.}
\end{aligned}$$

Remark 3.1. By the results in Section 2 and Section 3, we see that for $p = \frac{1}{100}$ and $p = \frac{1}{1000}$ the optimal sets of two-means do not contain any point from J_1, but for $p = \frac{2}{5}$ it contains a point from J_1. Moreover, we see that for $p = \frac{1}{100}$ and $p = \frac{2}{5}$ the optimal sets of three-means contain points from J_1, but for $p = \frac{1}{1000}$, the optimal set of three-means, and four-means do not contain any point from J_1. Using the similar technique as given in Section 2, it can be shown that Lemma 2.5 and Proposition 2.8, and Theorem 2.9 are also true here for $p = \frac{2}{5}$ and $p = \frac{1}{1000}$. The main difference is that for $p = \frac{1}{100}$ they are true for all $n \geq 3$, but for $p = \frac{2}{5}$, they are true for all $n \geq 2$, on the other hand, for $p = \frac{1}{1000}$ they are true for all $n \geq 5$.

The function $f : \mathbb{N} \to \mathbb{N}$, defined in (1), is also true here under the condition that $V(j, n-j)$ in this section is defined as follows:

$$V(j, n-j) := \left\{ \begin{array}{ll} \frac{2}{5} V_j(P_1) + \frac{3}{5} V_{n-j}(P_2) & \text{if } p = \frac{2}{5}, \\ \frac{1}{1000} V_j(P_1) + \frac{999}{1000} V_{n-j}(P_2) & \text{if } p = \frac{1}{1000} \end{array} \right.$$

where $1 \leq j \leq n-1$. Now, we give the following two examples which are analogous to Example 2.12 given in the previous section.

Example 3.2. For $p = \frac{2}{5}$, we have

$\{f(n)\}_{n=2}^{\infty} = \{1, 1, 2, 2, 3, 3, 4, 4, 5, 5, 6, 6, 7, 7, 7, 8, 8, 9, 9, 10, 10, 11, 11, 12, 12, 13, 13, 14, 14, \cdots.\}$.

In fact,

$\{f(n)\}_{n=4985}^{5011} = \{2324, 2325, 2325, 2326, 2326, 2327, 2327, 2328, 2328, 2329, 2329, 2329, 2330, 2330, 2331, 2331, 2332, 2332, 2333, 2333, 2334, 2334, 2335, 2335, 2336, 2336, 2336\}$.

Example 3.3. For $p = \frac{1}{1000}$, we have

$\{f(n)\}_{n=5}^{\infty} = \{1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6 \cdots\}$.

In fact,

$$\{f(n)\}_{n=4985}^{5011} = \{453, 453, 454, 454, 454, 454, 454, 454, 454, 454, 454, 454, 454, 455, 455, 455, 455, 455, 455, 455, 455, 455, 455, 455, 456, 456, 456\}.$$

Let us now give the following conjecture and the open problems.

Conjecture 3.4. Let P_1 and P_2 be two uniform distributions defined on any two closed intervals $[a, b]$ and $[c, d]$, where $a < b < c < d$. Let $P := pP_1 + (1-p)P_2$ be a mixed distribution generated by (P_1, P_2) associated with any probability vector $(p, 1-p)$, where $0 < p < 1$. Then, we conjecture that for each probability vector $(p, 1-p)$ there exists a positive integer N such that for all $n \geq N$, the optimal sets α_n contain points from both the intervals $[a, b]$ and $[c, d]$, and do not contain any point from the open interval (b, c). This yields the fact that if $\text{card}(\alpha_n \cap [a, b]) = k := k(n)$, then α_n contains k elements from $[a, b]$, and $(n-k)$ elements from $[c, d]$, i.e., $\alpha_n(P) = \alpha_k(P_1) \cup \alpha_{n-k}(P_2)$ for all $n \geq N$ with quantization error

$$V_n(P) = \frac{1}{108}\Big(\frac{p}{k^2} + \frac{1-p}{(n-k)^2}\Big).$$

Open 3.5. Let $P := pP_1 + (1-p)P_2$ be the mixed distributions generated by the two uniform distributions P_1 and P_2 defined on the closed intervals $[0, \frac{1}{3}]$ and $[\frac{1}{3}, \frac{2}{3}]$ associated with the probability vectors $(p, 1-p)$. It is still not known whether there is any probability vector $(p, 1-p)$, or what is the range of p, for which an optimal set α_2 of two-means for the mixed distributions P will contain a point from the open interval $(\frac{1}{3}, \frac{2}{3})$, and an optimal set α_3 of three-means will contain points from $[0, \frac{1}{3}]$ and $[\frac{2}{3}, 1]$, and also from the open interval $(\frac{1}{3}, \frac{2}{3})$.

If the answer of the above open problem comes in the negative, then it leads to investigate the following open problem.

Open 3.6. It is still not known whether there is a set of values of a, b, c, d, and p, where $a < b < c < d$ and $0 < p < 1$, such that if $P := pP_1 + (1-p)P_2$ is the mixed distribution generated by the two uniform distributions P_1 and P_2 defined on the closed intervals $[a, b]$ and $[c, d]$ associated with the probability vector $(p, 1-p)$, then an optimal set α_2 of two-means for the mixed distributions P will contain a point from the open interval (b, c), and an optimal set α_3 of three-means will contain points from $[a, b]$ and $[c, d]$, and also from the open interval (b, c).

Conjecture 3.7. We conjecture that the answer of the open problem Open 3.5 will be negative.

4. Observation

Under Conjecture 3.7, in this section, we try to give a partial answer of the open problem 'Open 3.6'. Let us choose the two closed intervals as follows:

$$[a, b] = [0, \frac{7}{15}] \text{ and } [c, d] = [\frac{8}{15}, 1].$$

Let P_1 and P_2 be the uniform distributions defined on the closed intervals $[a, b]$ and $[c, d]$. Then, the density functions f_1 and f_2 of the uniform distributions P_1 and P_2 are, respectively, given by $f_1(x) = \frac{15}{7}$ if $x \in [0, \frac{7}{15}]$, and zero otherwise; and $f_2(x) = \frac{15}{7}$ if $x \in [\frac{8}{15}, 1]$, and zero otherwise.

We now give the following two propositions.

Proposition 4.1. *Let $p = \frac{51}{500}$, and let $P := pP_1 + (1-p)P_2$ be the mixed distribution generated by the two uniform distributions P_1 and P_2 defined on the closed intervals $[a, b]$ and $[c, d]$. Then, an optimal set α_2 of two-means for the mixed distribution P contains a point from the open interval (b, c).*

Proof. Proceeding in the similar way as Proposition 2.3, we see that an optimal set of two-means for the mixed distribution $P := pP_1 + (1-p)P_2$, where $p = \frac{51}{500}$, is given by $\alpha_2 := \{0.488570, 0.829523\}$ with quantization error $V_2 = 0.0179722$. Since $b < 0.488570 < c < 0.829523 < d$, the assertion of the proposition follows. $\square$

Proposition 4.2. *Let $p = \frac{225}{500}$, and let $P := pP_1 + (1-p)P_2$ be the mixed distribution generated by the two uniform distributions P_1 and P_2 defined on the closed intervals $[a, b]$ and $[c, d]$. Then, an optimal set α_3 of three-means for the mixed distribution P contains points from $[a, b]$ and $[c, d]$, and also from the open interval (b, c).*

Proof. Proceeding in the similar way as Proposition 2.4, we see that an optimal set of three-means for the mixed distribution $P := pP_1 + (1-p)P_2$, where $p = \frac{225}{500}$, is given by $\alpha_3 := \{0.174089, 0.522267, 0.840756\}$ with quantization error $V_3 = 0.00985931$. Since $a < 0.174089 < b < 0.522267 < c < 0.840756 < d$, the assertion of the proposition follows. $\square$

Remark 4.3. Notice that the two mixed distributions $P := pP_1 + (1-p)P_2$ considered in Proposition 4.1 and Proposition 4.2 are different. It is worthwhile to investigate whether the two mixed distributions can be same, in other words, whether there is a mixed distribution $P := pP_1 + (1-p)P_2$, where P_1 and P_2 are two uniform distributions on two different closed intervals $[a, b]$ and $[c, d]$ associated with a probability vector $(p, 1-p)$ with $a < b < c < d$ and $0 < p < 1$, for which the open problem 'Open 3.6' is true.

References

[AW] E.F. Abaya and G.L. Wise, *Some remarks on the existence of optimal quantizers*, Statistics & Probability Letters, Volume 2, Issue 6, December 1984, Pages 349-351.

[BW] J.A. Bucklew and G.L. Wise, *Multidimensional asymptotic quantization theory with rth power distortion measures*, IEEE Transactions on Information Theory, 1982, Vol. 28 Issue 2, 239-247.

[CR] D. Comez and M.K. Roychowdhury, *Quantization for uniform distributions on stretched Sierpinski triangles*, Monatshefte für Mathematik, Volume 190, Issue 1, 79-100 (2019).

[DR1] C.P. Dettmann and M.K. Roychowdhury, *Quantization for uniform distributions on equilateral triangles*, Real Analysis Exchange, Vol. 42(1), 2017, pp. 149-166.

[DR2] C.P. Dettmann and M.K. Roychowdhury, *An algorithm to compute CVTs for finitely generated Cantor distributions*, to appear, Southeast Asian Bulletin of Mathematics.

[GG] A. Gersho and R.M. Gray, *Vector quantization and signal compression*, Kluwer Academy publishers: Boston, 1992.

[GKL] R.M. Gray, J.C. Kieffer and Y. Linde, *Locally optimal block quantizer design*, Information and Control, 45 (1980), pp. 178-198.

[GL1] S. Graf and H. Luschgy, *Foundations of quantization for probability distributions*, Lecture Notes in Mathematics 1730, Springer, Berlin, 2000.

[GL2] A. György and T. Linder, *On the structure of optimal entropy-constrained scalar quantizers*, IEEE transactions on information theory, vol. 48, no. 2, February 2002.

[GL3] S. Graf and H. Luschgy, *The Quantization of the Cantor Distribution*, Math. Nachr., 183, 113-133 (1997).

[GN] R. Gray and D. Neuhoff, *Quantization*, IEEE Trans. Inform. Theory, 44 (1998), pp. 2325-2383.

[L] L.J. Lindsay, *Quantization dimension for probability distributions*, PhD dissertation, 2001, University of North Texas, Texas, USA.

[L1] L. Roychowdhury, *Optimal quantization for nonuniform Cantor distributions*, Journal of Interdisciplinary Mathematics, Vol 22 (2019), pp. 1325-1348.

[R1] M.K. Roychowdhury, *Quantization and centroidal Voronoi tessellations for probability measures on dyadic Cantor sets*, Journal of Fractal Geometry, 4 (2017), 127-146.

[R2] M.K. Roychowdhury, *Optimal quantizers for some absolutely continuous probability measures*, Real Analysis Exchange, Vol. 43(1), 2017, pp. 105-136.

[R3] M.K. Roychowdhury, *Optimal quantization for the Cantor distribution generated by infinite similitudes*, Israel Journal of Mathematics 231 (2019), 437-466.

[R4] M.K. Roychowdhury, *Least upper bound of the exact formula for optimal quantization of some uniform Cantor distributions*, Discrete and Continuous Dynamical Systems- Series A, Volume 38, Number 9, September 2018, pp. 4555-4570.

[R5] M.K. Roychowdhury, *Center of mass and the optimal quantizers for some continuous and discrete uniform distributions*, Journal of Interdisciplinary Mathematics, Vol. 22 (2019), No. 4, pp. 451-471.

[R6] M.K. Roychowdhury, *Optimal quantization for mixed distributions*, Real Analysis Exchange, Vol. 46(2), 2021, pp. 451-484.

[RR1] J. Rosenblatt and M.K. Roychowdhury, *Optimal quantization for piecewise uniform distributions*, Uniform Distribution Theory 13 (2018), no. 2, 23-55.

[RR2] J. Rosenblatt and M.K. Roychowdhury, *Uniform distributions on curves and quantization*, arXiv:1809.08364 [math.PR].

[Z] R. Zam, *Lattice Coding for Signals and Networks: A Structured Coding Approach to Quantization, Modulation, and Multiuser Information Theory*, Cambridge University Press, 2014.

School of Mathematical and Statistical Sciences, University of Texas Rio Grande Valley, 1201 West University Drive, Edinburg, TX 78539-2999, USA.
Email address: ashley.gomez06@utrgv.edu

School of Mathematical and Statistical Sciences, University of Texas Rio Grande Valley, 1201 West University Drive, Edinburg, TX 78539-2999, USA.
Email address: ogla.lopez01@utrgv.edu

School of Mathematical and Statistical Sciences, University of Texas Rio Grande Valley, 1201 West University Drive, Edinburg, TX 78539-2999, USA.
Email address: mrinal.roychowdhury@utrgv.edu