ON THE QUANTIZERS FOR SMALL TRAINING SEQUENCES

Dong Sik Kim

School of Electronics and Information Engineering, Hankuk University of Foreign Studies
Yongin, Gyonggi-do, 449-791, Korea, dskim@hufs.ac.kr

ABSTRACT

In order to design a good quantizer for the underlying distribution using a training sequence (TS), the traditional approach is seeking for the empirical minimum based on the empirical risk minimization principle. As the size of TS increases, we may obtain a good quantizer for the true distribution. However, if we have a relatively small TS, searching the empirical minimum for the TS causes the overfitting problem, which even worsens the performance of the trained codebook. In this paper, the performance of codebooks trained by small TSs is studied, and it is shown that a piecewise uniform codebook can be better than an empirically minimized codebook is.

1. INTRODUCTION

Quantizer design is to find the reproduction levels and the corresponding quantizer regions, which minimize the average distortion yielded by the underlying distribution, for a fixed number of levels. Even though there is no explicit solution for the quantizer design problem, efficient algorithms are available by Lloyd and Max [1]. The Lloyd algorithm is generalized for vector inputs by Linde *et al.*. This algorithm is called the generalized Lloyd Algorithm (GLA). GLA can use the training sequence (TS) drawn from the underlying distribution if the distribution is not explicitly known. Using TS in designing vector quantizers is common as shown in an image coding scheme recently proposed by Mukherjee and Mitra [6].

When we use TS instead of the explicit distribution function, the traditional algorithms usually minimize the empirical distortion yielded by TS. Note that the empirical distortion yielded by an empirical distribution is different from the average distortion yielded by the true distribution function. However, based on the *empirical risk minimization* principle, which is supported by a work of Pollard [7], we may obtain a good codebook by minimizing the empirical distortion as the TS size increases. In other words, the traditional codebook design algorithms focus on searching an empirically optimal codebook for a given TS. However, we may surmise that simply searching the empirical minimum, for a given TS, is not the best approach in designing a good codebook for the underlying distribution since the TS size is usually finite. Further, we sometimes observe the case that the locally optimal codebook designed by GLA can be better than the globally optimal case [3].

In this paper, the performance of codebooks trained by small TSs is studied, and it is shown that a uniform codebook can be better than an empirically minimized codebook is, especially for small TSs. The performance of a piecewise uniform codebook is then studied to improve the performance of the uniform codebooks for the small TS case.

2. PRELIMINARY

Let F denote the underlying distribution function. The quantizer design problem for F is to find a set that minimizes the *average distortion* defined by

$$D(C) := \int \min_{\mathbf{y}\in C}(x-y)^2 F(dx)$$

over all possible choices of the *codebook* $C \subset \mathbb{R}$, of which size is n. In this paper, we consider a scalar quantizer design problem. We may extend the analysis to the vector quantizer case by constructing a piecewise uniform vector quantizer [5]. Let C^* denote an optimal codebook if $D(C^*) = \inf_C D(C)$. We call $D(C^*)$ the *optimum*, which is assumed $D(C^*) \neq 0$. Suppose that $X, X_1, \cdots, X_m$ are independent, and identically distributed random variables taking values in $\mathbb{R}$ with F. Let $X_1, \cdots, X_m$ denote a finite TS and m denote the TS size. For a codebook C, we define the *empirical distortion* as

$$D_m(C) := \frac{1}{m}\sum_{\ell=1}^{m} \min_{\mathbf{y}\in C}(X_\ell - y)^2,$$

where we suppose that $E\{X^2\} < \infty$. Let C_m^* denote a codebook that minimizes $D_m(\cdot)$. We call C_m^* and $D_m(C_m^*)$ ($= \min_C D_m(C)$) an *empirically optimal codebook* and the *empirical minimum*, respectively.

We now briefly overview the bias formed by the trained codebook, and the convergence rate as a function of the *training ratio* β [1], which is defined by the ratio of the TS size to the codebook size, i.e., $\beta := m/n$. Note that the training ratio is an important parameter in designing codebooks [3]. It is known that the bias between $E\{D_m(C_m^*)\}$ and the optimum $D(C^*)$ has a form of $\beta^{-\alpha}$, which will be decreased as the TS size m increases, in the following form:

$$E\{D(C_m^*)\} - D(C^*) = c\beta^{-\alpha},$$

where $\alpha \geq 0$ and $c > 0$, and the sequence $(c)_m$ is bounded. Hence, increasing the training ratio for a given TS is very important in reducing the bias between the designed codebook and the true optimal codebook. In other words, the bias is quite large if the training ratio is relatively small. Hence, if we have a small TS, then simply designing the codebook based on seeking to find the empirically optimal codebook could be an inefficient approach due to the large bias.

3. CODEBOOKS FOR SMALL TS

As shown in some experimental results of [3], where $n = 2$, the codebooks designed by GLA could be better than empirically optimal codebooks for both the Gaussian and Laplacian densities. In [3, Fig. 2], it is clear that the designed codebooks by GLA are locally optimal for the given TS. However by experiment, we may notice that the performance of the designed codebook is better than the empirically optimal codebooks. Further, in order to solve the local minimum problem of GLA, if we use the *simulated annealing* method, then the method could even deteriorate the performance of the designed codebook for the scalar quantization case. Hence, simply reducing the empirical distortion is not the best approach in codebook designing.

Imposing a constraint upon a structure of the codebook obvious prevents designing an optimal codebook. However, this constraint can generalize the trained codebook for a small TS. The structure of the piecewise uniform codebook may also be good for small TSs. In order to analyze the performance of a uniform and piecewise uniform codebooks for a small TS, we will observe two important notions, which are concerned with *changing the training ratio*.

The first notion is developed from *increasing the training ratio* β for a fixed bit rate as follows. It should be noted that the training ratio can be different depending on the structure of the quantizer. In other words, $\beta = m/n$ can be different depending on the quantizer structure since n can be different for a fixed bit rate. Structurally constrained quantizers can have better (or larger) training ratios than the full-search quantizer case, for a given bit rate [3]. For example, if we consider a multistage vector quantizer [1, p. 451] and use the same size of two codebooks for the two stages, then the training ratio increases to $m/\sqrt{n}$ compared to the normal training ratio m/n [3, Eq. (6)]. Hence, structurally constrained quantizers can have better training ratios than the conventional full-search quantizer case, which implies that the structurally constrained quantizer can yield even better training performance than the full-search quantizer case.

The second notion is devised from the fact that a *fixed codebook* can be better than the trained codebook. If we have too small TS, then searching the empirically optimal codebook is overfitting the codebook to an particular empirical distribution, which is very different from the true distribution. Designing codebooks for such small TS could worsen the codebook performance. Hence, we may reach a fact that, if we have too small TS, then we should simply use a fixed codebook, such as the uniform codebook instead of clustering the TS.

In the next section, the performance of the piecewise uniform codebook is numerically investigated based on the notions: *increasing training ratio* and *fixed codebooks*.

4. PIECEWISE UNIFORM CODEBOOKS

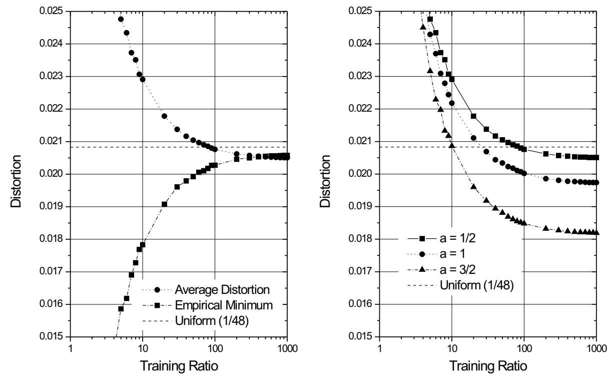
In order to analyze the performance of the piecewise uniform codebook, we first compare the performances of the empirically optimal and uniform codebooks for a simple density relying on the notion about fixed codebooks. We then extend the approach to a general case by increasing the training ratio, which is another notion in the previous section.

Suppose that we design a two-level codebook for a given TS, which is drawn from the underlying distribution function F. The conventional approach is finding an empirically optimal codebook for the TS. Here, we may employ GLA to find an empirically optimal codebook with low searching complexity. However, there exist better codebooks than the empirically optimal codebook. Let us consider the uniform codebook as a fixed codebook, and we will check whether this uniform codebook could be the codebook that is better than the empirically optimal codebook for a certain distribution function. In other words, we will show a case where the uniform codebook could be better than the trained codebook.

We consider a density $u_{a,b}$, which is described by a constant a ($|a| \leq 2b^2$) for a given positive b as follows:

$$u_{a,b}(x) := \begin{cases} ax + b - a/2b, & \text{for } x \in [0, 1/b] \\ 0, & \text{otherwise.} \end{cases}$$

$u_{0,b}$ implies a uniform density defined on $[0, 1/b]$. Let $C^{\text{U}} := \{1/4b, 3/4b\}$ denote a uniform codebook. C^{U} is the optimal codebook for $u_{0,b}$ when $n = 2$ [7]. Note that the inverse of $|a|$ implies *uniformity* of the density $u_{a,b}$. Suppose that F has the density $u_{a,b}$. The average distortion of C^{U} is then given by $D(C^{\text{U}}) = 1/48b^2$, which is independent of the constant a. If $|a| \neq 0$, then there exits an optimal codebook C^* such that $D(C^*) < D(C^{\text{U}})$. Hence, it is obvious that the uniform codebook C^{U} is not optimal for $u_{a,b}$ for $|a| \neq 0$.



(a) (b)

Fig. 1. Empirically optimal and uniform codebooks for $u_{a,b}$ and $D(C^{\mathrm{U}}) = 1/48$ with respect to the training ratio β, where $b = 1$ and $n = 2$. (a) Empirical minimum and average distortion curves for $a = 1/2$. (b) Average distortion curves of empirically optimal codebooks for $a = 1/2$, 1, and $3/2$.

Further, for a given TS, which is drawn from $u_{a,b}$, we have $D_m(C^{\mathrm{U}}) \geq D_m(C^*_m)$, and we have a strict relationship:

$$E\{D_m(C^{\mathrm{U}})\} > E\{D_m(C^*_m)\}$$

since $E\{D_m(C^{\mathrm{U}})\} = D(C^{\mathrm{U}}) \geq D(C^*)$ and $D(C^*) > E\{D_m(C^*_m)\}$ from [4, Theorem 1]. In other words, C^{U} could be different from the empirically optimal codebook C^*_m. Note that the elements of C^*_m are random variables defined on the underlying distribution. The expectation of $D(C^*_m)$ has a bias as a function of the training ratio β from the optimum $D(C^*)$. Here, the bias decreases as β increases. On the other hand, the uniform codebook C^{U} also has a bias $D(C^{\mathrm{U}}) - D(C^*)$, which is fixed. Hence, $D(C^{\mathrm{U}})$ could be lower than $E\{D(C^*_m)\}$ for relatively small training ratios.

In Fig. 1, numerical results for the density $u_{a,b}$ are illustrated for different values of a for a fixed $b = 1$ with respect to the training ratio[1]. Here, GLA is employed to design an empirically optimal codebook for the given TS[2]. We usually obtain the average and empirical distortion curves of the form as in Fig. 1(a) in training codebooks [4]. Even though GLA does not guarantee the global optimum (empirically optimal codebook), we suppose that the empirical distortion of the trained codebook C_m is equal to or very close to the empirical minimum $D_m(C^*_m)$, i.e., $D_m(C_m) \approx D_m(C^*_m)$. As we may notice in Fig. 1(b), the fixed uniform codebook C^{U} can be better than the trained codebooks, and as the slope a increases, the maximum training ratio, where C^{U} is better than C_m, is decreased. For the case of $a = 1.0$ in Fig. 1(b), C^{U} is better than the codebooks C_m, which are trained at $\beta < 20$. However, for the case of $a = 3/2$, C^{U} is better than C_m for $\beta < 10$ since the uniformity described by the inverse of $|a|$ is decreased. From the numerical observation in Fig. 1, we may notice that the uniform codebook can be better than the trained codebook for the given TS for relatively small training ratios. In other words, for a certain class of distributions, simply using the uniform codebook will provide a better performance if we have a small TS.

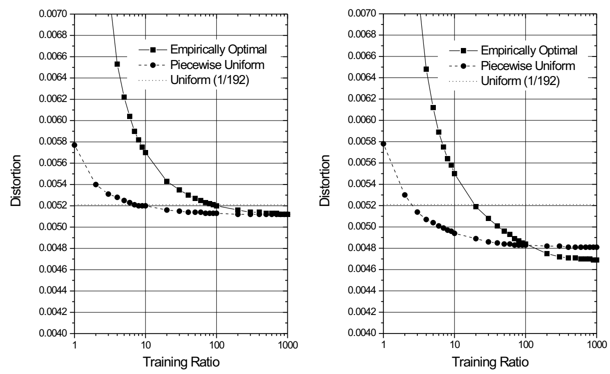
For a general distribution with larger codebooks, however, simply using the uniform codebook does not guarantee the better performance. Hence, we need to modify the uniform codebook to the piecewise uniform codebook for general distributions. Instead of using a uniform codebook, we may use a codebook, which has flexibility introduced by the given TS. The purpose of introducing the flexibility is to adapt the codebook to the underlying distribution. We now consider a simple approach, which can reflect the true distribution in designing codebooks. First, we divide the density $u_{a,b}$ into two intervals by clustering the TS. Hence, the divide between two intervals is given by the midpoint of the cluster centers. We assign two levels to each interval, and then design two different uniform codebooks for the two intervals, respectively. Note that the resultant codebook is a piecewise uniform codebook. A numerical comparison on this piecewise uniform codebook is illustrated in Fig. 2. As we may notice in Fig. 2, the piecewise uniform codebook shows a better performance than the traditionally trained codebook for a fixed training ratio. Further, for an appropriate range of training ratios, the piecewise uniform codebook is better than the uniform codebook, e.g., $\beta > 10$ in Fig. 2(a), where $a = 1/2$. For the $a = 3/2$ case in Fig. 2(b), we can clearly observe that the empirically optimal codebook is the best for $\beta \geq 100$. However, for $2 < \beta < 100$, the piecewise uniform codebook shows the best performance, and for $\beta \leq 2$, the uniform codebook is the best.

5. NUMERICAL RESULTS

We now summarize the piecewise uniform codebook, which is used in the paper for a performance analysis. We consider bounded $L\ (= n/n')$ intervals, where L and n' are positive integers. we divide the input into L intervals using GLA by clustering the given TS into L clusters. We then obtain L quantizer intervals, which are the L intervals in the proposed algorithm. Note that the training ratio when designing L clusters is increased to $\beta = n' \cdot m/n$, which is n' times larger than the n-level clustering case ($\beta = m/n$). Hence, the obtained L quantizer intervals are closer to the desired intervals for the real density f than the n level case, in terms

[1] In the figures, the distortion curves are numerically calculated by performing the expectations over 1,000 sample distortions.

[2] For such large TSs, obtaining the empirically optimal codebooks is very difficult due to the computational complexity. Hence, we employed a clustering algorithm, such as GLA, in order to reduce the complexity.



(a) (b)

Fig. 2. Uniform and piecewise uniform codebooks for $u_{a,b}$ with respect to the training ratio β, where $b = 1$ and $n = 4$. (a) $a = 1/2$. (b) $a = 3/2$.

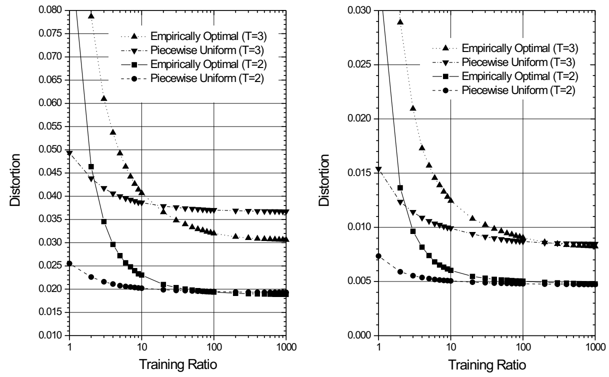
of the bias of the average distortion. In each quantizer interval, we next uniformly add n' levels to make an n'-level uniform codebook. We consider a truncated Gaussian density g'_{G} defined as

$$g'_{\mathrm{G}}(x) := \begin{cases} g_{\mathrm{G}}(x)/\tau_{\mathrm{G}}, & \text{for } |x| \le T \\ 0, & \text{otherwise.} \end{cases} \qquad (1)$$

Here, T is a positive constant, $\tau_{\mathrm{G}} := \int_{|x|\le T} g_{\mathrm{G}}(x)dx$, and g_{G} is a Gaussian density given by $g_{\mathrm{G}}(x) := e^{-x^2/2}/\sqrt{2\pi}$. Fig. 3 shows the performance of the piecewise uniform codebooks, for the $n = 8$ and 16 cases. Here, the size of each uniform codebook is $n' = 2$, and for the densities of g'_{G}, $T = 2$ and $T = 3$. We first consider the $T = 2$ case, where $\tau_{\mathrm{G}} \approx 0.954$. For the $n = 8$ case in Fig. 3(a), the piecewise uniform codebook is better than the empirical minimum case when $\beta \le 60$. For the $n = 16$ case in Fig. 3(b), the piecewise uniform codebook also better than the empirical minimum case when $\beta \le 1000$. We next consider the $T = 3$ case, where $\tau_{\mathrm{G}} \approx 0.997$. (Note that the left and right intervals have smaller probabilities than the $T = 2$ case.) We may notice that the maximum training ratio, where the piecewise uniform codebook is better than the empirically optimal codebook, is reduced, e.g., to 10 as shown for the case of $n = 8$ in Fig. 3(a). This tendency is also observed for the $n = 16$ case in Fig. 3(b). In a manner similar to the relative performance of uniform quantizers [2, p. 127], as the codebook size increases, the piecewise uniform codebook produces more gains, over a fixed uniform codebook. Hence, the relative performance of the piecewise uniform codebook of the $n = 16$ case (Fig. 3(b)) is better than the $n = 8$ case (Fig. 3(a)).

6. CONCLUSION

In this paper, the performance of the piecewise uniform codebook is analyzed for a small TS. It is shown that the piecewise uniform codebook is better than the empirically minimized codebook case if the TS size is relatively small. In training codebooks, if we have relatively small TS, then we should consider a non optimal quantizers, such as the quantizer based on the piecewise uniform codebook, instead of the quantizers based on the empirically optimal codebook.



(a) (b)

Fig. 3. Average distortions of piecewise uniform and empirically optimal codebooks for the truncated Gaussian densities g'_{G} in (1), where $T = 2$ ($\tau_{\mathrm{G}} \approx 0.954$) and $T = 3$ ($\tau_{\mathrm{G}} \approx 0.997$). (a) $n = 8$, $n' = 2$, and $L = 4$. (b) $n = 16$, $n' = 2$, and $L = 8$.

7. REFERENCES

[1] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression.* Boston: Kluwer Academic Publishers, 1992.

[2] N.S. Jayant and P. Noll, *Digital Coding of Waveforms*, New Jersey: Prentice-Hall, 1984.

[3] D. S. Kim, "Training ratio and comparison of trained vector quantizers," *IEEE Trans. Signal Processing*, vol. 51, no. 6, pp. 1632–1641, June 2003.

[4] D. S. Kim and M. R. Bell, "Upper bounds on empirically optimal quantizers," *IEEE Trans. Inform. Theory*, vol. IT-49, no. 4, pp. 1037–1046, April 2003.

[5] F. Kuhlmann and J. A. Bucklew, "Piecewise uniform vector quantizers," *IEEE Trans. Inform. Theory*, vol. IT-34, no. 5, pp. 1259–1263, Sep. 1988.

[6] D. Mukherjee and S. K. Mitra, "Vector SPIHT for embedded wavelet video and image coding," *IEEE Trans. Circuits and Systems for Video Technolodge*, vol. 13, no. 3, pp. 231–246, March 2003.

[7] D. Pollard, "Quantization and the method of k-means," *IEEE Trans. Inform. Theory,* vol. IT-28, no. 2, pp. 199–205, March 1982.