SRC TR 88-106

MARYLAND

TECHNICAL RESEARCH REPORT

On the Structural Synthesis of Tendon-Drive Manipulators Having Pseudo-Triangular Structure Matrix

by

J.-J. Lee and L.-W. Tsai

SYSTEMS RESEARCH CENTER

UNIVERSITY OF MARYLAND

COLLEGE PARK, MARYLAND 20742

ON THE STRUCTURAL SYNTHESIS OF TENDON-DRIVEN MANIPULATORS HAVING PSEUDO-TRIANGULAR STRUCTURE MATRIX

Jyh-Jone Lee
Graduate Research Assistant

Lung-Wen Tsai
Member of ASME
Associate Professor

Mechanical Engineering Department
and
Systems Research Center
University of Maryland
College Park, MD 20742

* This research has been supported in part by the Department of Engergy, Grant No. DE-FG05-88ER13977, and in part by the NSF's Engineering Research Centers Program, NSFD CDR 8803012. Such supports do not constitute an endorsement by the supporting agencies of the views expressed in the article.

ABSTRACT

Tendons have been widely used for power transmission in the field of anthropomorphic manipulating systems. This paper deals with the identification and enumeration of the kinematic structure of tendon-driven robotic mechanisms. The structural isomorphism of tendon-driven manipulators is defined and the structural characteristics of such mechanical systems are described. Applying these structural characteristics, a methodology for the enumeration of tendon-driven robotic mechanisms has been developed. Mechanism structures with up to six degrees of freedom have been enumerated.

1. Introduction

An open-loop kinematic chain is considered to be the simplest structure among various designs of manipulators. However, actuators for driving such sytem must be installed along the joint axes, which in turn increases the inertia of the manipulator. For this reason, a large portion of the power is consumed on driving the actuators themselves other than the load.

An alternative way of reducing the inertia of the system is to install actuators on the ground and transmit power or torques through tendons. A tendon-driven articulated manipulator, such as the UTAH-MIT Hand (Jacobsen et al., 1985) and Stanford/JPL Hand (Salisbury, 1982), has the advantage of remote control and weight reduction. A few tendon-driven mechanical systems can be found in the literature (Leaver and McCarthy, 1987; Mason and Salisbury, 1985; Morecki et al., 1984; Pham and Heginbotham, 1986). To date, most of the studies on such mechanical systems have been emphasized on the mechanics of manipulation and control of some specialized designs. The purpose of this investigation is to establish a method for the identification of structural isomorphism and the enumeration of the kinematic structure of certain type of tendon-driven manipulators.

Theory of synthesis in the area of epicyclic gear trains has been thoroughly studied by Buchsbaum and Freudenstein (1970), Freudenstein (1971), Tsai (1987), and Tsai and Lin (1988). In this work, we extend the theory developed for epicyclic gear trains to the structural synthesis of tendon-driven mechanisms.

2. Principle of Operation

For tendon-driven manipulators, Morecki, et al. (1980) carried out the structural analysis yielding the relationship between joint angles and tendon displacements. Subsequently, Tsai and Lee (1988) showed that the kinematic structure of tendon-driven robotic mechanisms can be represented by a planar schematic from which the structure matrix, relating tendon displacements to the joint angles, can be derived systematically.

For an n-D.O.F. (Degree-of-Freedom) manipulator with m open-ended tendons, the equation of transformation can be written as

$$\underline{S} = A\underline{\Theta} \tag{1}$$

where $\underline{S}$ is an $m \times 1$ column matrix representing linear displacements of tendons, $\underline{\Theta}$ is an $n \times 1$ column matrix for the joint angles and A is an $m \times n$ matrix. If we assume that all pulleys pivoted about one joint axis are of the same radius, then Eq. (1) can be decomposed into the following form:

$$\underline{S} = B\ R\ \underline{\Theta} \tag{2}$$

Here, matrix B whose elements consist of $-1, 0$ and $+1$, is an $m \times n$ matrix which describes the routing of the tendons and, matrix R is an $n \times n$ diagonal matrix whose non-zero elements are the radii of the pulleys.

Resultant torques, $\underline{t}$, about the joint axes in the equivalent open-loop chain, can be related to forces exerted by tendons, $\underline{f}$, through the principle of energy conservation:

$$\dot{\underline{\Theta}}^T \underline{t} = \dot{\underline{S}}^T \underline{f} \tag{3}$$

Substituting the time derivative of Eq. (2) into (3), yields

$$\underline{t} = R^T B^T \underline{f} \tag{4}$$

For a given set of joint torques, Eq. (4) constitutes n linear equations in m unknowns. In order to achieve positive tensions in the tendons, m should be greater than n. Thus, the solution for the forces in tendons consists of a particular solution plus an $(m - n)$

dimensional homogeneous solution. The particular solution can be determined by the generalized inverse transformation of Eq. (4) and the homogeneous solution corresponds to certain sets of tendon tensions that result in no net joint torques. The homogeneous solution can be expressed as a sum of $(m-n)$ basis vectors each of them being multiplied by an arbitrary constant. The $(m-n)$ basis vectors span the null space of B^T. The homogeneous solution is necessary to be non-negative. Thus, by adjusting the constants, positive tension can be maintained in all of the tendons.

3. Structural Characteristics of Tendon-Driven Manipulators

In this paper, we require all the actuators to be installed on the base link. Thus, every tendon regardless of which link it is attached, must be routed through the base joint of the manipulator and connected to its actuator on the base link.

We define the **Degree-of-Incidence** (D.O.I.) of a tendon as the number of joints that the tendon has been routed over. For example, if a tendon has been routed over five consecutive joints, then we say the D.O.I. of the tendon is five.

Let n be the D.O.F. of a robotic system, m the total number of tendons, and m_i the number of tendons with i D.O.I. Then, we have

$$\sum_{i=1}^{n} m_i = m, \tag{5}$$

and

$$m \geq n+1 \tag{6}$$

Note that m_i defines the number of columns with i non-zero elements in the matrix B^T.

Since the robotic system is controllable, a subsystem containing any number of links and their associated joints taken from the far end of the original system should be also controllable. This means the number of tendons contained in the subsystem should also be greater than the number of joints by a minimum of one. If we consider the link farthest away from the base of the system and its associated joint as a subsystem, then we can conclude that there should be at least 2 tendons routed over the last joint. Since these two tendons must be routed over all the joints of the original system, the sum of the D.O.I. for

these two tendons is equal to $2n$. Likewise, if we consider the subsystem consisting of the last two moving links, then it should contain at least 3 tendons, and the minimum number of total D.O.I. is $2n + (n-1)$. Following the same argument, we conclude the lower limit for the sum of D.O.I. for an n D.O.F. robotic system with m tendons is given by:

$$2n + (n-1) + (n-2) + \ldots + 1 + (m-n+1) \leq m_1 + 2m_2 + 3m_3 + \ldots + nm_n \quad (7.a)$$

The upper limit is reached when all the tendons are attached to the last moving link. Hence,

$$m_1 + 2m_2 + 3m_3 + \ldots + nm_n \leq nm \quad (7.b)$$

Combining Eqs. (7.a) and (7.b), we obtain

$$n(n+3)/2 + (m-n+1) \leq m_1 + 2m_2 + \ldots + nm_n \leq nm \quad (7.c)$$

These m_i's are also subjected to the following constraints:

$$\begin{aligned} m_n &\geq 2 \\ m_n + m_{n-1} &\geq 3 \\ &\vdots \\ m_n + m_{n-1} + \ldots + m_1 &\geq n+1 \end{aligned} \quad (8)$$

We observe that the set $(m_n, m_{n-1}, \ldots, m_1)$, when assembled into a structure matrix, takes the following form:

$$B^T = \begin{bmatrix} \# & \# & \# & \cdot & \cdot & 0 & 0 & 0 & \cdot & \cdot & 0 & 0 & 0 \\ \# & \# & \# & \cdot & \cdot & \# & \# & \# & \cdot & \cdot & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \# & \# & \# & \cdot & \cdot & \# & \# & \# & \cdot & \cdot & 0 & 0 & 0 \\ \# & \# & \# & \cdot & \cdot & \# & \# & \# & \cdot & \cdot & \# & \# & \# \end{bmatrix} \quad (9)$$

where the "#" sign represents the existence of a non-zero element.

In Eq. (9), the existence of a non-zero element at the i^{th} row and j^{th} column implies that the j^{th} tendon is incident with the $(n-i+1)^{th}$ joint, and the sign is determined by

the direction of the tendon routing. In this notation, the rows have been arranged in a reversed sequence, i.e. the first row represents the last joint in a manipulator. This will facilitate the structural synthesis to be discussed next.

Based on the above discussion, we summarize the structural characteristics for the tendon-driven manipulators as follows:

C1. There is a minimum of two non-zero elements and one sign change between elements in each row. This guarantees every joint can be manipulated in both directions.

C2. Exchanging any two columns, which is equivalent to the renaming of two tendons, does not affect the function and structure of the system.

C3. The structure matrix can always be arranged in a form such that all the zero elements appear on the upper-right-hand corner of the matrix.

C4. Changing the sign of every element in a row does not affect the generic characteristics of the structure. This is equivalent to a change in the definition of positive direction of a joint axis.

C5. The rank of the structure matrix is equal to the degrees of freedom of the system. Hence, for an n-D.O.F. system with m tendons, at least one determinant of a submatrix formed by deleting $(m-n)$ columns from the structure matrix shall not be equal to zero. Furthermore, if $m = n+1$, then the determinant of a submatrix formed by deleting any column shall not be equal to zero.

C6. There exist $(m-n)$ dimensional homogeneous solution to Eq. (4) such that all the elements in the homogeneous solution are non-negative.

4. Structural Isomorphism

In order to define isomorphic structures, we assign a positive direction of rotation to each joint axis and then sketch the mechanism in a planar schematic introduced by Tsai and Lee (1988). According to Tsai and Lee the element of the structure matrix is determined by the tendon routing and the definition of positive direction of rotation for the joint axes. If the direction of a joint axis is defined in a reversed manner, then the sign for each element in the corresponding row of the structure matrix is altogether reversed. Since the definition of positive direction for a joint axis has no effect on the function of

a mechanism, two tendon-driv manipulators are said to be structurally isomorphic if the structure matrices of the tw systems are identical; or if they become identical after a change of sign for every element in one or more rows, or after rearranging the sequence of certain columns, or a combination of both.

For example, if we define the axes of positive rotation to be pointing out of the paper, the structure matrices for the manipulators shown in Figs.1(a) and 1(b) are given by

$$B_1^T = \begin{bmatrix} -1 & 1 & 0 \\ -1 & -1 & 1 \end{bmatrix}, \quad \text{and} \quad B_2^T = \begin{bmatrix} -1 & 1 & 0 \\ 1 & 1 & -1 \end{bmatrix},$$

respectively. It can be seen that after switching the sign of each element in the second row of B_2^T, B_2^T becomes identical to B_1^T. Physically, if we flip over the base joint axis of the mechanism shown in Fig.1(b), then it becomes identical to that of Fig.1(a). Therefore, the two mechanisms are said to be structurally isomorphic.

Another example is shown in Fig.2. The structure matrices for the mechanisms shown in Figs. 2 (a) and 2 (b) are given by:

$$B_1^T = \begin{bmatrix} -1 & 1 & 0 & 0 \\ -1 & -1 & 1 & 0 \\ -1 & 1 & -1 & 1 \end{bmatrix}, \quad \text{and} \quad B_2^T = \begin{bmatrix} -1 & 1 & 0 & 0 \\ -1 & -1 & 1 & 0 \\ 1 & -1 & -1 & 1 \end{bmatrix},$$

respectively. By switching the sign of each element in the first row of B_2^T, and then exchanging the first and second columns, B_2^T becomes identical to B_1^T. This can also be explained as follows. If we reverse the direction of the last joint axis in Fig. 2(b) and rename F_1 to F_2 and F_2 to F_1, then the two mechanisms become identical. Therefore, the two mechanisms are said to be structurally isomorphic.

5. Structure Synthesis

In what follows, we shall limit ourselves to those tendon-driven robotic systems with the number of tendons greater than the number of D.O.F. by one. For example, if $n = 3$, then $m = 4$. Writting Eqs. (5), (7.c), and (8) for $n = 3$ and $m = 4$, yields

$$m_1 + m_2 + m_3 = 4 \tag{10.a}$$

$$9 \leq m_1 + 2m_2 + 3m_3 \leq 12 \tag{10.b}$$

$$m_3 \geq 2 \tag{10.c}$$

$$m_3 + m_2 \geq 3 \tag{10.d}$$

Solving eq. (10.a) for m_1, m_2 and m_3 subjected to the inequality constraints, eqs. (10.b - 10.d), yields $(m_1, m_2, m_3) = (1,1,2)$, or $(0,2,2)$, or $(1,0,3)$, or $(0,1,3)$, or $(0,0,4)$. Table 1 shows the admissible structure matrices corresponding to the above possible solutions. Note that the structure matrix for those system with $m-n=1$ must have a positive homogeneous solution $\underline{f}$ to Eq. (4). Otherwise, it is not controllable and, hence, not admissible. The number of admissible structure matrices increases as the number of D.O.F. increases. Morecki, et al. (1980) predicted there could exit up to 23040 admissible structures for six-D.O.F. manipulators having seven tendons. In what follows, we shall consider only those systems whose structure matrix takes the pseudo-triangular form, i.e.

$$B^T = \begin{bmatrix} e_{11} & e_{12} & 0 & 0 & \cdots & 0 \\ e_{21} & e_{22} & e_{23} & 0 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\ e_{n-1,1} & e_{n-1,2} & \cdot & \cdot & \cdots & 0 \\ e_{n1} & e_{n2} & \cdot & \cdot & \cdots & e_{n,n+1} \end{bmatrix}, \tag{11}$$

or $(m_1, m_2, m_3, \ldots, m_n) = (1,1,1,\ldots,2)$ where

$$\sum_{i=1}^{n} m_i = m = n+1.$$

Each non-zero element in Eq. (11) can assume the value of $+1$ or -1 provided that the resultant matrix satisfies the structural characteristics C1 - C6 outlined in the previous section. Since there are $n(n+3)/2$ non-zero elements in a robotic system having n-D.O.F. and $(n+1)$ tendons, the number of possible structure matrices is equal to 2 to the $n(n+3)/2\underline{th}$ order. It would be very difficult if not impossible, to identify the admissible structures from such a large number of possible combinations. In what follows, we present a simpler approach.

We start the synthesis with a known n by $(n+1)$ structure, called the generic structure, and increase the degree of freedom one at a time. In view of Eq. (11), we conclude that

each time we increase the degree of freedom by one, both the number of links and the number of joints in the equivalent open-loop chain (Tsai and Lee, 1988) should also be increased by one. Let the new link be connected to the base-link of the generic structure by a turning pair and let the base-link of the generic structure be the first moving link and the added link be the base-link for the new mechanism. Then, all the tendons in the generic structure must now be extended over the newly added joint to allow all actuators be connected to the base. Moreover, an additional tendon is required to connect the first moving link to the new base-link.

From the structure matrix point of view, this procedure implies that the matrix of the generic structure are to be supplemented by an additional column and an additional row. All the elements in the supplemental column, except the last, are to be set to zero. A new element in the supplemental row can assume the value of $+1$ or -1. For convenience, we let the last element of supplemental column (and row) be $+1$. Hence, there are potentially 2^n combinations. However, some of the combinations may be rejected due to the violation of the structural characteristics, C1 - C6, or may be eliminated due to the reason of structural isomorphism. So the number of admissible non-isomorphic structure matrices is usually much less than 2^n. This procedure can be automated by a computer program. We summarized the systematic procedure and the results as follows:

One-D.O.F. system. We start with the most fundamental manipulator with two links and one turning pair. The only possible structure is shown in Fig.3(a) where two cables are routed through different sides of the joint. Figure 3(b) shows the structure matrix of Fig.3(a). We note that the element $(1,2)$ has been chosen to $+1$ and the homogeneous solution is given by $\underline{f} = (1,1)^T$.

Two-D.O.F. system. There are $2^2 = 4$ possible combinations to supplement an additional row of non-zero elements to the structure matrix of one D.O.F. system. Only one form, as shown in Fig.4(a) satisfies the structural characteristics C1 - C6. The corresponding structure matrix is shown in Fig.4(b). The homogeneous solution is given by $\underline{f} = (1,1,2)^T$.

Three-D.O.F. systems. There are $2^3 = 8$ possible combinations to supplement

an additional row to the matrix shown in Fig.4(b). Only three were found to satisfy C1 - C6, however, two of them were structurally isomorphic. Hence, there exist only two nonisomorphic structures as shown in Figs.5(a) and 5(c). Figures 5(b) and 5(d) show their corresponding structure matrices.

Four-D.O.F. systems. There are $2^4 = 16$ possible ways to supplement one additional row to each of the matrix shown in Figs.5(b) and 5(d). Since there are two structurally nonisomorphic structures for the three-D.O.F. systems, totally, $16 \times 2 = 32$ possible structures can be generated for four-D.O.F. systems. Thirteen of them satisfy structural characteristics C1 - C6. But, only eleven are structurally nonisomorphic as shown in Figs.6(a) - 6(k). The corresponding structure matrices are also shown in Figs.6(a) - 6(k).

Five-D.O.F. systems. There are $2^5 = 32$ possible ways to supplement one additional row to each of the matrix shown in Figs.6(a) - 6(k). Since there are eleven structurally nonisomorphic four-D.O.F. systems, a total of $11 \times 32 = 352$ structures can be generated for five-D.O.F. systems. After applying structural characteristics, C1 - C6, and checking for structural isomorphism, we obtained 141 types of drives that are structurally nonisomorphic.

Six-D.O.F. systems. There are $2^6 = 64$ possible ways to supplement one additional row to each of the matrix of five-D.O.F. systems. A total of $64 \times 141 = 9024$ possible structures can be generated. Again, after checking for structural characteristics, C1 - C6, and structural isomorphism, we obtained 3905 structurally nonisomorphic drives.

6. Summary

The structural characteristics of tendon-driven manimulator systems have been investigated. A criterion for the identification of structure isomorphism has been established and a methodology for the enumeration of tendon-driven manipulators having pseudo-triangular structure matrix has been developed. All the admisible structure matrices with up to six-D.O.F. have been investigated. We have found 3905 structurally nonisomorphism drives with six-D.O.F. as opposed to 23040 solutions obtained earlier by Morecki, et al. (1980).

We note that there is a planar schematic corresponding to each structure matrix. However, each planar schematic can be converted into various different spatial mechanisms depending on the twist angle chosen for every pair of adjacent joint axes. This is also true for the construction of planar mechanisms, for which the twist angles can be either zero or one hundred and eighty degrees. Hence, the number of functional mechanisms is much larger than that of structurally nonisomorphic structures.

7. References

Buchsboum, F., and Freudenstein, F., 1970, "Synthesis of Kinematic Structure of Geared Kinematic Chains and Other Mechanisms," J. of Mechanisms, Vol. 5, pp.357-392.

Freudenstein, F., 1971, "An Application of Boolean Algebra to the Motion of Epicyclic Drives," ASME J. of Engineering for Industry, Vol. 93, Series B, pp.176-182.

Jacobsen, S.C., Wood, J.E., Knutti, D.F. and Biggers, K.B., 1985, "The Utah/MIT Dextrous Hand: Work in Progress," The International Journal of Robotics Research, Vol. 3, No. 4, pp.21-50.

Leaver, S.O., and McCarthy, J.M., 1987, "The Design of Three Jointed Two-Degree-of-Freedom Robot Fingers, "ASME Advances in Design Automation, Volume Two: Robotics, Mechanisms, and Machine System, DE-Vol. 10-2, pp.127-134.

Mason, M.T. and Salisbury, J.K. Jr., 1985, "Robot hands and the Mechanics of Manipulation," The MIT Press, Cambridge, Mass.

Morecki, A., et al., 1980, "Synthesis and Control of the Anthropomorphic Two-Handed Manipulator," Proceedings of the 10^{th} Inter. Symposium on Industrial Robots, Milan, Italy.

Morecki, A., Ekiel, J., and Fidelus, K., 1984, "Cybernatic Systems of Limb Movements in Man, Animals and Robots," PWN-Polish Scientific Publishers, Ellis Horwood Limited, Warszawa, Poland.

Pham, D.T., and Heginbotham, W.B., 1986, "Robot Grippers," Springer-Verlag, IFS (Publications) Ltd., UK.

Salisbury, J.K. Jr., 1982, "Kinematic and Force Analysis of Articulated Hands," Ph.D. Dissertation, Stanford University, Stanford, CA.

Tsai, L.W., Sep. 1987, "An Application of the Linkage Characteristic Polynomial to the Topological Synthesis of Epicyclic Gear Trains," ASME Trans., J. of Mechanisms, Transmissions, and Automation in Design, Vol. 109, No. 3, pp.329-336.

Tsai, L.W. and Lee, J.J., 1988, "Kinematic Analysis of Tendon-Driven Robotic Mechanisms Using Graph Theory," Trends and Developments in Mechanisms, Machines and Robotics, DE-Vol. 15-3, pp.330-346, presented at the 1988 ASME design technology conferences, Kissimmee, Florida, Sept. 25-28, 1988, accepted for publication in the Trans. of ASME, J. of Mechanisms, Transmissions, and Automation in Design.

Tsai, L.W. and Lin, C.C., 1988, "The Creation of True Two-Degree-of-Freedom Epicyclic Gear Trains," Trends and Developments in Mechanisms, Machines and Robotics, DE-Vol. 15-1, pp.153-164, presented at the 1988 ASME design technology conferences, Kissimmee, Florida, Sept. 25-28, 1988. Submitted for publication in the Trans. of ASME, J. of Mechanisms, Transmissions, and Automation in Design.

Caption Summary

Fig. 1 Two structurally isomorphic tendon-driven manipulators having two-D.O.F.

Fig. 2 Two structurally isomorphic tendon-driven manipulators having three-D.O.F.

Fig. 3(a) Functional schematic of a one-D.O.F. tendon-driven manipulator.

Fig. 3(b) Structure matrix of Fig. 3(a)

Fig. 4(a) Planar schematic of a two-D.O.F. tendon-driven manipulator.

Fig. 4(b) Structure matrix of Fig. 4(a)

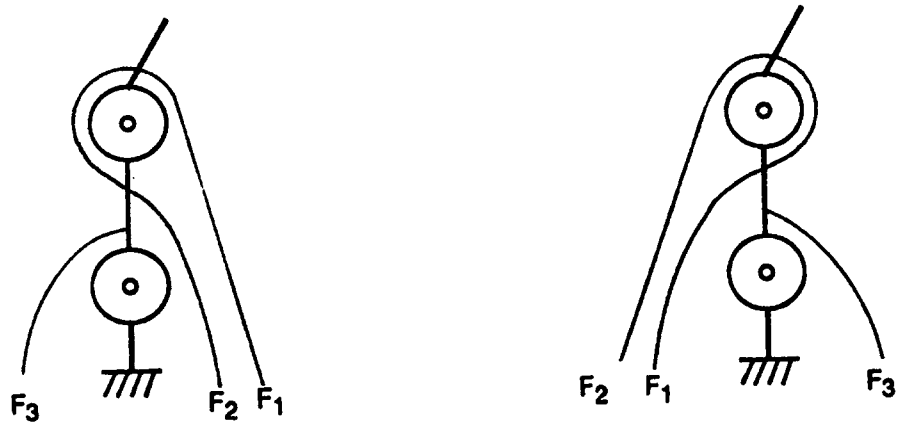
Fig. 5(a) Planar schematic I for three-D.O.F. tendon-driven manipulator.

Fig. 5(b) Structure matrix of Fig. 5 (a).

Fig. 5(c) Planar schematic II for three-D.O.F. tendon-driven manipulator.

Fig. 5(d) Structure matrix of Fig. 5(c).

Fig. 6 Planar schematics and associated structure matrices for four-D.O.F. tendon-driven manipulator.



(a) (b)

Fig. 1

Table 1. Structure matrices for manipulators having three-D.O.F. and four control tendons.

(m_1, m_2, m_3)	(1, 1, 2)	(0, 2, 2)	(1, 0, 3)	(0, 1, 3)	(0, 0, 4)
B^T	# # 0 0 # # # 0 # # # #	# # 0 0 # # # # # # # #	# # # 0 # # # 0 # # # #	# # # 0 # # # # # # # #	# # # # # # # # # # # #
admissible solution	-1 1 0 0 -1 -1 1 0 -1 -1 -1 1	-1 1 0 0 -1 1 1 -1 -1 -1 1 1	no solution	1 1 -1 0 -1 1 1 -1 -1 1 -1 1	1 1 -1 -1 1 -1 -1 1 -1 1 -1 1

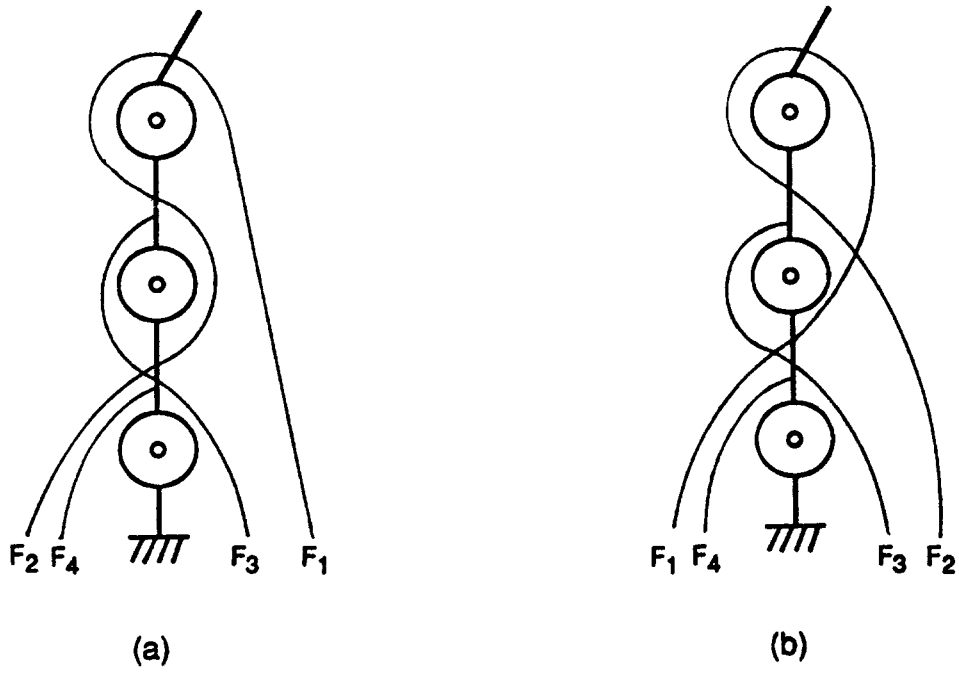
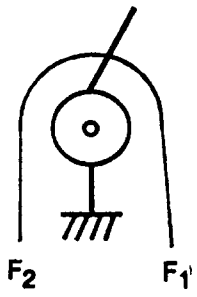


Fig. 2

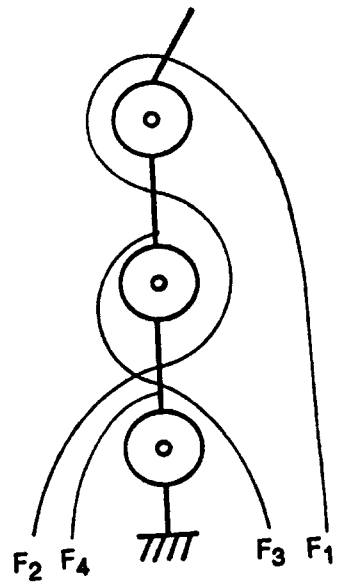


(a)

$$B^T = [-1 \qquad 1]$$

(b)

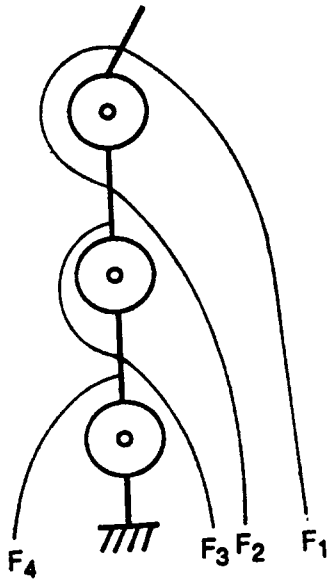
Fig. 3



(a)

$$B^T = \begin{bmatrix} -1 & 1 & 0 & 0 \\ -1 & -1 & 1 & 0 \\ -1 & 1 & -1 & 1 \end{bmatrix}$$

(b)

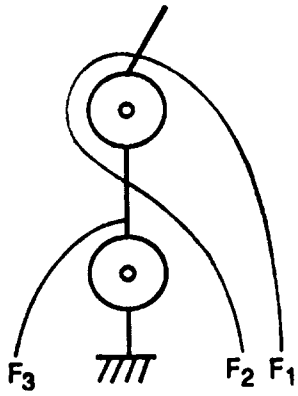


(c)

$$B^T = \begin{bmatrix} -1 & 1 & 0 & 0 \\ -1 & -1 & 1 & 0 \\ -1 & -1 & -1 & 1 \end{bmatrix}$$

(d)

Fig. 5

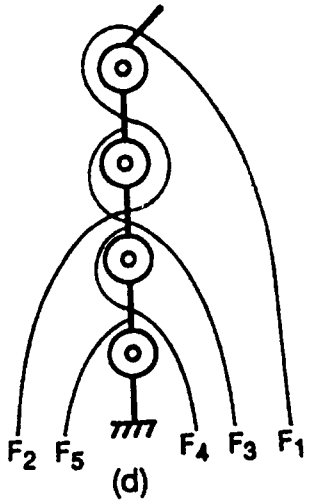


(a)

$$B^T = \begin{bmatrix} -1 & 1 & 0 \\ -1 & -1 & 1 \end{bmatrix}$$

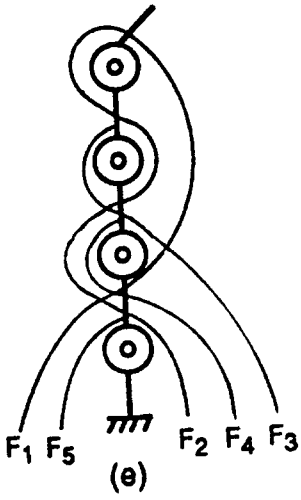
(b)

Fig. 4



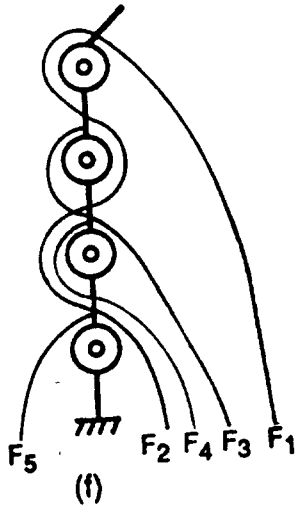
(d)

$$B^T = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 \\ -1 & -1 & 1 & 0 & 0 \\ -1 & 1 & -1 & 1 & 0 \\ -1 & 1 & -1 & -1 & 1 \end{bmatrix}$$



(e)

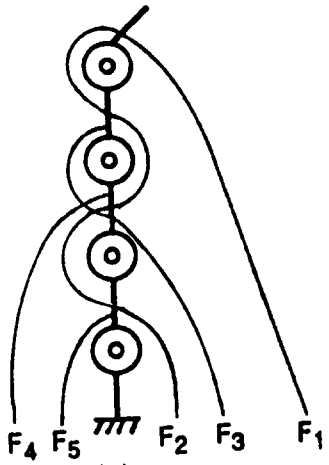
$$B^T = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 \\ -1 & -1 & 1 & 0 & 0 \\ -1 & 1 & -1 & 1 & 0 \\ 1 & -1 & -1 & -1 & 1 \end{bmatrix}$$



(f)

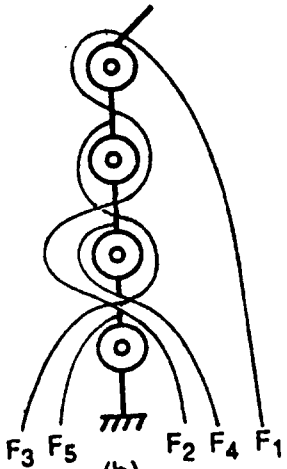
$$B^T = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 \\ -1 & -1 & 1 & 0 & 0 \\ -1 & 1 & -1 & 1 & 0 \\ -1 & -1 & -1 & -1 & 1 \end{bmatrix}$$

Fig 6



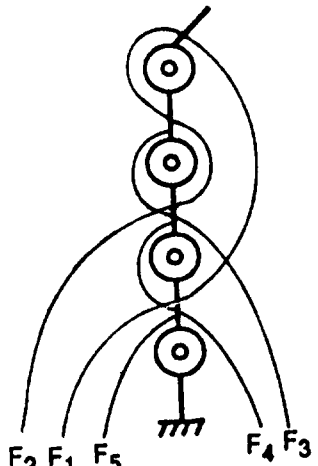
(a)

$$B^T = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 \\ -1 & -1 & 1 & 0 & 0 \\ -1 & 1 & -1 & 1 & 0 \\ -1 & -1 & -1 & 1 & 1 \end{bmatrix}$$



(b)

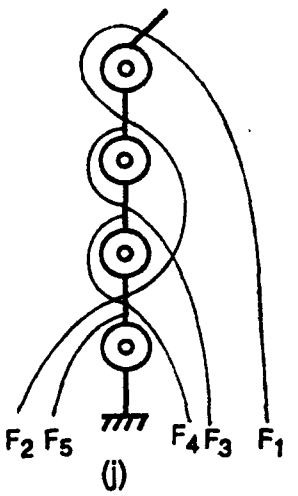
$$B^T = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 \\ -1 & -1 & 1 & 0 & 0 \\ -1 & 1 & -1 & 1 & 0 \\ -1 & -1 & 1 & -1 & 1 \end{bmatrix}$$



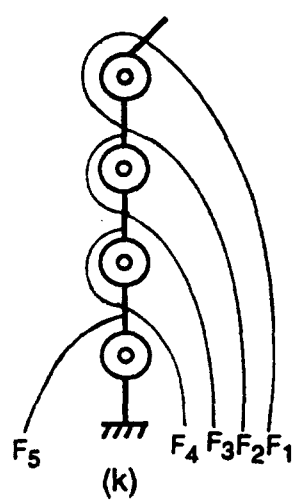
(c)

$$B^T = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 \\ -1 & -1 & 1 & 0 & 0 \\ -1 & 1 & -1 & 1 & 0 \\ 1 & 1 & -1 & -1 & 1 \end{bmatrix}$$

Fig 4

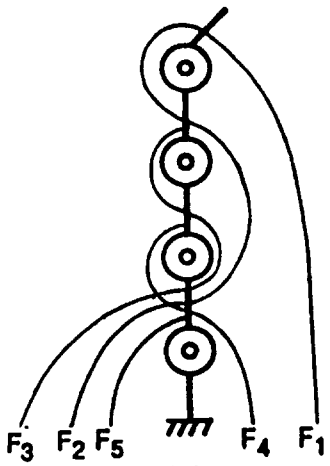


$$B^T = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 \\ -1 & -1 & 1 & 0 & 0 \\ -1 & -1 & -1 & 1 & 0 \\ -1 & 1 & -1 & -1 & 1 \end{bmatrix}$$



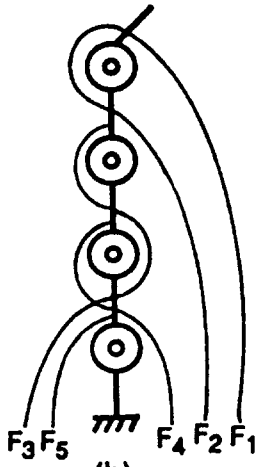
$$B^T = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 \\ -1 & -1 & 1 & 0 & 0 \\ -1 & -1 & -1 & 1 & 0 \\ -1 & -1 & -1 & -1 & 1 \end{bmatrix}$$

Fig. 6 continued



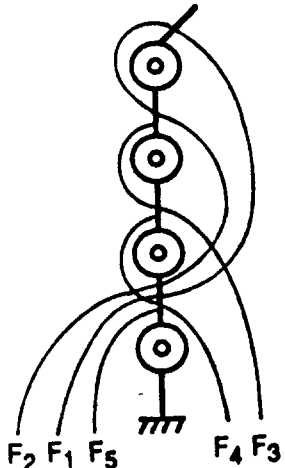
(g)

$$B^T = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 \\ -1 & -1 & 1 & 0 & 0 \\ -1 & -1 & -1 & 1 & 0 \\ -1 & 1 & 1 & -1 & 1 \end{bmatrix}$$



(h)

$$B^T = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 \\ -1 & -1 & 1 & 0 & 0 \\ -1 & -1 & -1 & 1 & 0 \\ -1 & -1 & 1 & -1 & 1 \end{bmatrix}$$



(i)

$$B^T = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 \\ -1 & -1 & 1 & 0 & 0 \\ -1 & -1 & -1 & 1 & 0 \\ 1 & 1 & -1 & -1 & 1 \end{bmatrix}$$

Fig 6 cont'd

5 Calculating the Quantities Needed for the Optimization Algorithm

In order to use a gradient-based numerical optimization algorithm, we need to calculate the average distortion for a specified SSQ and the gradient of the average distortion with respect to the SSQ parameters. In Subsection 5.1, we will describe how to evaluate the average distortion. In Subsection 5.2, we will determine the partial derivative of the average distortion with respect to each of the thresholds of each of the quantizers in the transmitter. In Subsection 5.3, we will determine the partial derivative of the average distortion with respect to each of the next-quantizer map parameters.

5.1 Evaluating the Average Distortion

The average distortion D incurred by an SSQ is given by (4). Thus, to evaluate D for a given SSQ, we need to calculate the π_{ji}'s and D_{ji}'s. We begin with the π_{ji}'s. We will show that the π_{ji}'s can be found by solving a system of $S(T-1)$ linear equations in $S(T-1)$ unknowns.

From the definition of π_{ji}, we have

$$\begin{aligned}\pi_{ji} &= \sum_{m=1}^{S} P[S_t = j, S_{t-1} = m, T_t = i] \\ &= \sum_{m=1}^{S} P[S_t = j | S_{t-1} = m, T_t = i] P[S_{t-1} = m, T_t = i].\end{aligned}$$

But the transmitter state T_t is a function of the prior HMS switch states $S_{t-1}, S_{t-2}, \ldots$, and $(S_t)_{t=1}^{\infty}$ is a Markov chain, so

$$P[S_t = j | S_{t-1} = m, T_t = i] = P[S_t = j | S_{t-1} = m] \triangleq \lambda_{j|m}.$$

Hence

$$\pi_{ji} = \sum_{m=1}^{S} \lambda_{j|m} P[S_{t-1} = m, T_t = i]. \tag{12}$$

We have now used the assumption that the switch is Markov, so the rest of our derivation will not hold for more general composite sources.

Next, we express $P[S_{t-1} = m, T_t = i]$ in a formula that includes the transmitter state and the index of the observed cell at the previous time instant:

$$\begin{aligned} P[S_{t-1} = m, T_t = i] &= \sum_{k=1}^{T} \sum_{l=1}^{C} P[S_{t-1} = m, T_t = i, T_{t-1} = k, C_{t-1} = l] \\ &= \sum_{k=1}^{T} \sum_{l=1}^{C} \{ P[T_t = i | S_{t-1} = m, T_{t-1} = k, C_{t-1} = l] \\ &\qquad \times P[C_t = l | S_t = m, T_t = k] \\ &\qquad \times P[S_t = m, T_t = k] \}. \end{aligned}$$

Now the quantizer that is to be used at any particular time instant is a function of the quantizer that was used at the previous time instant and the observed cell. Conditioned on this information, the quantizer decision is independent of the state of the switch at the previous time instant. Thus,

$$P[T_t = i | S_{t-1} = m, T_{t-1} = k, C_{t-1} = l] = P[T_t = i | T_{t-1} = k, C_{t-1} = l].$$

Using eq. (6) and defining

$$\gamma_{l|mk} \triangleq P[C_t = l | S_t = m, T_t = k] \tag{13}$$

$$= \int_{\xi_{l-1}^k}^{\xi_l^k} g_m(x)\, dx, \tag{14}$$

we obtain

$$P[S_{t-1} = m, T_t = i] = \sum_{k=1}^{T} \sum_{l=1}^{C} \tau_{i|kl} \gamma_{l|mk} \pi_{mk}. \tag{15}$$

Combining (15) and (12) and rearranging terms,

$$\pi_{ji} = \sum_{m=1}^{S} \sum_{k=1}^{T} \left(\lambda_{j|m} \sum_{l=1}^{C} \tau_{i|kl} \gamma_{l|mk} \right) \pi_{mk}. \tag{16}$$

Define

$$A_{ji|mk} \triangleq \lambda_{j|m} \sum_{l=1}^{C} \tau_{i|kl} \gamma_{l|mk}. \tag{17}$$

$A_{ji|mk}$ is the conditional probability of the source/transmitter system state moving to $(S_{t+1} = j, T_{t+1} = i)$ given that $(S_t = m, T_t = k)$. Equation (17) allows us to rewrite equation (16) as

$$\pi_{ji} = \sum_{m=1}^{S} \sum_{k=1}^{T} A_{ji|mk} \pi_{mk}. \tag{18}$$

There is one such equation for each $j \in J_S$ and $i \in J_T$, so the π_{ji}'s satisfy ST linear equations in ST unknowns. However, the solution for this system of equations is not unique; for example, the all-zero vector solves (18). Of course, we are only interested in that particular solution for which the π_{ji}'s define a probability distribution. By using this constraint, we can derive a new system of equations with fewer unknowns that does have a unique solution. Observe that

$$P[S_t = j] \triangleq \rho_j = \sum_{i=1}^{T} \pi_{ji}. \tag{19}$$

Hence, we can treat π_{jT} as a dependent variable[4] and replace it in equation (18) by

$$\pi_{jT} = \rho_j - \sum_{k=1}^{T-1} \pi_{jk}, \quad j = 1, 2, \ldots, S. \tag{20}$$

Thus,

$$\begin{aligned} \pi_{ji} &= \sum_{m=1}^{S} \sum_{k=1}^{T} A_{ji|mk} \pi_{mk} \\ &= \sum_{m=1}^{S} \sum_{k=1}^{T-1} (A_{ji|mk} - A_{ji|mT}) \pi_{mk} + \sum_{m=1}^{S} A_{ji|mT} \rho_m. \end{aligned} \tag{21}$$

This last equation must hold for each $i \in J_{T-1}$ and $j \in J_S$, yielding $S(T-1)$ equations in $S(T-1)$ unknowns.

We now describe one way of putting this system of equations into the standard form

$$\Phi z = b, \tag{22}$$

[4]Of course, we could have made π_{jk} with some other $k \in \{1, 2, \ldots, T-1\}$ the dependent variable; theoretically, it makes no difference which one we select. Notationally, however, it is easier to use π_{jT}.

so that the π_{ji}'s can be found using any standard subroutine for solving simultaneous linear equations.

Let δ_{ab}, where a and b are integers, be the Kronecker delta function:

$$\delta_{ab} = \begin{cases} 1, & \text{if } a = b; \\ 0, & \text{otherwise.} \end{cases} \tag{23}$$

Equation (21) can then be rewritten as

$$\sum_{m=1}^{S} \sum_{k=1}^{T-1} (A_{ji|mk} - A_{ji|mT} - \delta_{jm}\delta_{ik})\pi_{mk} = -\sum_{m=1}^{S} A_{ji|mT}\rho_m, \tag{24}$$

with one such equation for each $j \in J_S$ and $i \in J_{T-1}$.

We will use r to index the rows of Φ and b, and c to index the columns of Φ. For each $j \in J_S$ and $i \in J_{T-1}$, let $r = (j-1)(T-1) + i$, and for each $m \in J_S$ and $k \in J_{T-1}$, let $c = (m-1)(T-1) + k$. This indexing means that the π_{ji}'s are being placed into the vector z in the following order:

$$z = [\pi_{1,1}\ \pi_{1,2}\ \cdots\ \pi_{1,T-1}\ \pi_{2,1}\ \pi_{2,2}\ \cdots\ \pi_{2,T-1}\ \cdots\ \pi_{S,1}\ \pi_{S,2}\ \cdots\ \pi_{S,T-1}]^T,$$

where the superscript T means *transpose* (z is a column vector). Φ and b are now completely determined. The element in the r-th row of the vector b of (22) is given by

$$b_r = -\sum_{m=1}^{S} A_{ji|mT}\rho_m, \tag{25}$$

and the element in the r-th row and the c-th column of the matrix Φ of (22) is given by

$$\Phi_{rc} = A_{ji|mk} - A_{ji|mT} - \delta_{jm}\delta_{ik}. \tag{26}$$

Any subroutine for solving linear systems can now be used to get the vector z, from which the π_{ji}'s are obtained by observing that the element in the r-th row of z corresponds to π_{ji} for $r = (j-1)(T-1) + i$, $j \in J_S$ and $i \in J_{T-1}$. The π_{jT}'s are then readily obtained by using equation (20).

We now calculate the average distortion incurred by applying the i-th quantizer to the j-th subsource, D_{ji}. In general, the formula for D_{ji} is given by (3). Obviously, to evaluate D_{ji} we need to select the reproduction levels η_l^i. We now show how to do this optimally.

From equation (4), we get

$$\frac{\partial D}{\partial \eta_l^k} = \sum_{j=1}^{S}\sum_{i=1}^{T} \frac{\partial \pi_{ji}}{\partial \eta_l^k} D_{ji} \; + \; \sum_{j=1}^{S}\sum_{i=1}^{T} \pi_{ji} \frac{\partial D_{ji}}{\partial \eta_l^k}. \tag{27}$$

Note that the reproduction levels for any particular quantizer have no influence over the next-quantizer decision rule. Therefore, $\partial \pi_{ji}/\partial \eta_l^k$ is equal to zero for all values of j and i, so the first double summation vanishes. Furthermore, the reproduction levels for the i-th quantizer do not affect the average distortion incurred by the k-th quantizer when $k \neq i$, so the second double summation reduces to a single summation. Thus,

$$\frac{\partial D}{\partial \eta_l^k} = \sum_{j=1}^{S} \pi_{jk} \frac{\partial D_{jk}}{\partial \eta_l^k} \tag{28}$$

with

$$\frac{\partial D_{jk}}{\partial \eta_l^k} = \int_{\xi_{l-1}^k}^{\xi_l^k} g_j(x) \frac{\partial d_j(x, \eta_l^k)}{\partial \eta_l^k} \, dx. \tag{29}$$

To proceed, we must be more specific about the distortion measure. We will use a weighted-squared-error criterion, with possibly a different weight for each subsource:[5]

$$d_j(x, y) = w_j (x - y)^2, \quad w_j > 0, \quad j \in J_S. \tag{30}$$

Then equation (28) becomes

$$\frac{\partial D}{\partial \eta_l^k} = 2 \sum_{j=1}^{S} \pi_{jk} w_j \int_{\xi_{l-1}^k}^{\xi_l^k} (\eta_l^k - x) g_j(x) \, dx. \tag{31}$$

Let

$$\mu_{l|jk} = \int_{\xi_{l-1}^k}^{\xi_l^k} x g_j(x) \, dx. \tag{32}$$

[5] For example, $w_1 = w_2 = \ldots = w_S$ means minimising the average squared error, while $w_j = 1/\sigma_j^2$, $j \in J_S$, with σ_j^2 being the variance of the j-th subsource, means minimising the average noise-to-signal ratio.

Table 1: Summary of the steps required for computing the weighted mean squared error D for a given SSQ.

1. Compute the probability of each of the C cells for each of the T quantizers using (14).
2. Compute the $A_{ji|mk}$'s using (17).
3. Solve the linear system of equations $\Phi z = b$ for the vector z, where b and Φ are given by (25) and (26), respectively.
4. Recover the π_{ji}'s from z via the formula $\pi_{ji} = z_{(j-1)(T-1)+i}$, for $j \in J_S$ and $i \in J_{T-1}$.
5. Recover the π_{jT}'s using (20)
6. Compute the centroid for each of the C cells of each of the T quantizers using (32).
7. Compute the optimal reproduction levels for the quantizers using (33).
8. Compute the distortion incurred by applying the i-th quantizer to the j-th source using (3).
9. Compute the average distortion using (4).

Setting (31) equal to zero and simplifying yields

$$\eta_l^k = \sum_{j=1}^{S} \pi_{jk} w_j \mu_{l|jk} \Big/ \sum_{j=1}^{S} \pi_{jk} w_j \gamma_{l|jk} \,. \tag{33}$$

Once the optimal reproduction levels are found, the D_{ji}'s can be calculated. Having calculated the π_{ji}'s and the D_{ji}'s, the average distortion incurred by an SSQ can be evaluated using equation (4). The steps required to compute the average distortion D achieved by a given SSQ are listed in Table 1.

5.2 Evaluating the Partial Derivative of the Average Distortion with Respect to an Arbitrary Quantizer Threshold

From (4), we get

$$\frac{\partial D}{\partial \xi_l^q} = \sum_{j=1}^{S}\sum_{i=1}^{T} D_{ji}\frac{\partial \pi_{ji}}{\partial \xi_l^q} + \sum_{j=1}^{S}\sum_{i=1}^{T} \pi_{ji}\frac{\partial D_{ji}}{\partial \xi_l^q}. \tag{34}$$

We now evaluate the first of the two double summations. By (19),

$$\sum_{i=1}^{T} \frac{\partial \pi_{ji}}{\partial \xi_l^q} = 0,$$

so

$$\sum_{j=1}^{S}\sum_{i=1}^{T} D_{ji}\frac{\partial \pi_{ji}}{\partial \xi_l^q} = \sum_{j=1}^{S}\sum_{i=1}^{T-1} (D_{ji} - D_{jT})\frac{\partial \pi_{ji}}{\partial \xi_l^q}. \tag{35}$$

The D_{ji}'s were computed in the last section; the remaining problem is to calculate

$$\frac{\partial \pi_{ji}}{\partial \xi_l^q}, \quad j \in J_S,\ i \in J_{T-1}.$$

Recall that the π_{ji}'s, with $j \in J_S$ and $i \in J_{T-1}$, were found by solving a linear system of equations denoted by $\Phi z = b$, where the formulas for the elements of Φ and b were given by equations (26) and (25), respectively. It follows that

$$\Phi \frac{\partial z}{\partial \xi_l^q} = \theta^{ql} \tag{36}$$

with

$$\theta^{ql} = \frac{\partial b}{\partial \xi_l^q} - \frac{\partial \Phi}{\partial \xi_l^q} z. \tag{37}$$

We now explain how to compute $\partial b/\partial \xi_l^q$ and $\partial \Phi/\partial \xi_l^q$. Afterwards, we will be able to give a succinct formula for the elements of the vector θ^{ql}.

For $i, k \in J_{T-1}$ and $j, m \in J_S$, define $r = (j-1)(T-1)+i$ and $c = (m-1)(T-1)+k$. Then the element in the r-th row and the c-th column of $\partial \Phi/\partial \xi_l^q$ is given by (see eq. (26))

$$\frac{\partial \Phi_{rc}}{\partial \xi_l^q} = \frac{\partial A_{ji|mk}}{\partial \xi_l^q} - \frac{\partial A_{ji|mT}}{\partial \xi_l^q}. \tag{38}$$

By equation (17),

$$\frac{\partial A_{ji|mk}}{\partial \xi_l^q} = \lambda_{j|m} \sum_{n=1}^{C} \tau_{i|kn} \frac{\partial \gamma_{n|mk}}{\partial \xi_l^q}. \tag{39}$$

From equation (14),

$$\frac{\partial \gamma_{n|mk}}{\partial \xi_l^q} = \begin{cases} g_m(\xi_l^k), & \text{if } k = q \text{ and } n = l; \\ -g_m(\xi_l^k), & \text{if } k = q \text{ and } n = l+1; \\ 0, & \text{otherwise.} \end{cases}$$

Thus,

$$\frac{\partial A_{ji|mk}}{\partial \xi_l^q} = \begin{cases} \lambda_{j|m}(\tau_{i|kl} - \tau_{i|k,l+1}) g_m(\xi_l^k), & \text{if } k = q; \\ 0, & \text{otherwise.} \end{cases} \tag{40}$$

Using the Kronecker delta function of (23), we can simplify this to

$$\frac{\partial A_{ji|mk}}{\partial \xi_l^q} = \delta_{kq} \lambda_{j|m}(\tau_{i|kl} - \tau_{i|k,l+1}) g_m(\xi_l^k). \tag{41}$$

Since $\partial A_{ji|mk}/\partial \xi_l^q = 0$ when $q \neq k$, and since k is only taken to run from 1 through $T-1$ when forming the matrix Φ, equation (38) becomes

$$\begin{aligned} \frac{\partial \Phi_{rc}}{\partial \xi_l^q} &= \frac{\partial A_{ji|mk}}{\partial \xi_l^q} \\ &= \delta_{kq} \lambda_{j|m}(\tau_{i|kl} - \tau_{i|k,l+1}) g_m(\xi_l^k) \end{aligned} \tag{42}$$

for $q \in J_{T-1}$. For $q = T$, equation (38) becomes

$$\begin{aligned} \frac{\partial \Phi_{rc}}{\partial \xi_l^T} &= -\frac{\partial A_{ji|mT}}{\partial \xi_l^T} \\ &= -\lambda_{j|m}(\tau_{i|Tl} - \tau_{i|T,l+1}) g_m(\xi_l^T). \end{aligned} \tag{43}$$

The r-th element of the vector $(\partial \Phi/\partial \xi_l^q)z$, denoted by

$$[\frac{\partial \Phi}{\partial \xi_l^q} z]_r,$$

is found by taking the inner product of the r-th row of $\partial \Phi/\partial \xi_l^q$ and the vector z. Because the elements in the c-th column of $\partial \Phi/\partial \xi_l^q$ and the c-th row of z correspond to the m-th

subsource and the k-th quantizer via $c = (m-1)(T-1)+k$, where $m \in J_S$ and $k \in J_{T-1}$, the r-th element of $(\partial\Phi/\partial\xi_l^q)z$ is given by (recall that $z_c = \pi_{mk}$ for this ordering)

$$\begin{aligned}
[\frac{\partial\Phi}{\partial\xi_l^q}z]_r &= \sum_{c=1}^{S(T-1)} \frac{\partial\Phi_{rc}}{\partial\xi_l^q} z_c \\
&= \sum_{m=1}^{S}\sum_{k=1}^{T-1} \delta_{kq}\lambda_{j|m}(\tau_{i|kl} - \tau_{i|k,l+1})g_m(\xi_l^k)\pi_{mk} \\
&= (\tau_{i|ql} - \tau_{i|q,l+1}) \sum_{m=1}^{S} \pi_{mq}\lambda_{j|m}g_m(\xi_l^q) \qquad (44)
\end{aligned}$$

for $q < T$. For $q = T$, the formula is different:

$$\begin{aligned}
[\frac{\partial\Phi}{\partial\xi_l^T}z]_r &= \sum_{c=1}^{S(T-1)} \frac{\partial\Phi_{rc}}{\partial\xi_l^T} z_c \\
&= \sum_{m=1}^{S}\sum_{k=1}^{T-1} -\lambda_{j|m}(\tau_{i|Tl} - \tau_{i|T,l+1})g_m(\xi_l^T)\pi_{mk} \\
&= -(\tau_{i|Tl} - \tau_{i|T,l+1}) \sum_{m=1}^{S} (\rho_m - \pi_{mT})\lambda_{j|m}g_m(\xi_l^T), \qquad (45)
\end{aligned}$$

using equation (20).

We now turn our attention to the vector $\partial b/\partial\xi_l^q$. For $j \in J_S$ and $i \in J_{T-1}$, let $r = (j-1)(T-1)+i$. Then the r-th element of the vector b is given by equation (25). Hence

$$\frac{\partial b_r}{\partial\xi_l^q} = \begin{cases} 0, & q \neq T; \\ -\sum_{m=1}^{S} \rho_m(\partial A_{ji|mT}/\partial\xi_l^T), & \text{otherwise.} \end{cases} \qquad (46)$$

Using equation (41), we get

$$\frac{\partial b}{\partial\xi_l^T} = (\tau_{i|T,l+1} - \tau_{i|Tl}) \sum_{m=1}^{S} \rho_m\lambda_{j|m}g_m(\xi_l^T). \qquad (47)$$

We can now compute the entries of the vector θ^{ql} appearing on the right-hand side of (36). Since $\partial b/\partial\xi_l^q$ equals zero when $q < T$, we get

$$\theta^{ql} = \begin{cases} -\frac{\partial\Phi}{\partial\xi_l^q}z, & \text{when } q < T; \\ \frac{\partial b}{\partial\xi_l^T} - \frac{\partial\Phi}{\partial\xi_l^T}z, & \text{when } q = T. \end{cases} \qquad (48)$$

Therefore, for $q < T$,

$$\begin{aligned}
\theta_r^{ql} &= -[\frac{\partial \Phi}{\partial \xi_l^q} z]_r \\
&= -(\tau_{i|ql} - \tau_{i|q,l+1}) \sum_{m=1}^{S} \pi_{mq} \lambda_{j|m} g_m(\xi_l^q) \\
&= (\tau_{i|q,l+1} - \tau_{i|ql}) \sum_{m=1}^{S} \pi_{mq} \lambda_{j|m} g_m(\xi_l^q) \qquad (49)
\end{aligned}$$

for $r = (j-1)(T-1) + i$. For $q = T$, subtracting (45) from (47) and simplifying yields

$$\theta_r^{Tl} = (\tau_{i|T,l+1} - \tau_{i|Tl}) \sum_{m=1}^{S} \pi_{mT} \lambda_{j|m} g_m(\xi_l^T), \qquad (50)$$

which has the same form as (49). Hence, we can combine these two formulas into one:

$$\theta_r^{ql} = (\tau_{i|q,l+1} - \tau_{i|ql}) \sum_{m=1}^{S} \pi_{mq} \lambda_{j|m} g_m(\xi_l^q), \quad q \in J_T,\ l \in J_{C-1}. \qquad (51)$$

Once this system of equations is solved for $\partial z / \partial \xi_l^q$, we can compute (35). This completes the evaluation of the first double summation on the right-hand side of (34). We now go after the second double summation on the right-hand side of (34).

At first glance, it might appear that a change in the thresholds of the q-th quantizer will not affect the average distortion incurred by the i-th quantizer when $q \neq i$, which would mean $\partial D_{ji} / \partial \xi_l^q = 0$ for $i \neq q$. However, this is not true in general because a change in any of the thresholds affects *all* of the π_{ji}'s, which in turn affects the optimal reproduction levels for every quantizer (see equation (33)), and, of course, this affects the average distortion incurred by each of the quantizers. Thus, $\partial D_{ji} / \partial \xi_l^q \neq 0$ in general, even if $i \neq q$.

Considering the reproduction levels as functions of the thresholds rather than as independent variables and applying *Leibniz's Formula* [17, p. 245], we get:

$$\begin{aligned}
\frac{\partial D_{ji}}{\partial \xi_l^q} &= \frac{\partial}{\partial \xi_l^q} \left[\sum_{n=1}^{C} \int_{\xi_{n-1}^i}^{\xi_n^i} d_j(x, \eta_n^i) g_j(x)\, dx \right] \\
&= \sum_{n=1}^{C} \int_{\xi_{n-1}^i}^{\xi_n^i} g_j(x) \frac{\partial d_j(x, \eta_n^i)}{\partial \xi_l^q}\, dx \ + \ \delta_{iq} g_j(\xi_l^q) \left[d_j(\xi_l^q, \eta_l^q) - d_j(\xi_l^q, \eta_{l+1}^q) \right],
\end{aligned}$$

where δ_{iq} is the Kronecker Delta.

Recall that we are assuming that

$$d_j(x,y) = w_j(x-y)^2, \quad w_j > 0, \quad j = 1, \ldots, M,$$

so

$$\begin{aligned}
\frac{\partial D_{ji}}{\partial \xi_l^q} &= 2w_j \sum_{n=1}^{C} \frac{\partial \eta_n^i}{\partial \xi_l^q} \int_{\xi_{n-1}^i}^{\xi_n^i} g_j(x)(\eta_n^i - x)\, dx \\
&\quad + \delta_{iq} w_j g_j(\xi_l^q)(\eta_{l+1}^q - \eta_l^q)(2\xi_l^q - \eta_{l+1}^q - \eta_l^q) \\
&= 2w_j \sum_{n=1}^{C} \frac{\partial \eta_n^i}{\partial \xi_l^q} [\eta_n^i \gamma_{n|ji} - \mu_{n|ji}] \\
&\quad + \delta_{iq} w_j g_j(\xi_l^q)(\eta_{l+1}^q - \eta_l^q)(2\xi_l^q - \eta_{l+1}^q - \eta_l^q). \qquad (52)
\end{aligned}$$

Observe that

$$\begin{aligned}
\sum_{j=1}^{S} \sum_{\substack{i=1 \\ i \neq q}}^{T} \pi_{ji} \frac{\partial D_{ji}}{\partial \xi_l^q} &= \sum_{j=1}^{S} \sum_{\substack{i=1 \\ i \neq q}}^{T} \pi_{ji} \left[2w_j \sum_{n=1}^{C} \frac{\partial \eta_n^i}{\partial \xi_l^q} [\eta_n^i \gamma_{n|ji} - \mu_{n|ji}] \right] \\
&= 2 \sum_{\substack{i=1 \\ i \neq q}}^{T} \sum_{n=1}^{C} \frac{\partial \eta_n^i}{\partial \xi_l^q} \left(\eta_n^i \sum_{j=1}^{S} \pi_{ji} w_j \gamma_{n|ji} - \sum_{j=1}^{S} \pi_{ji} w_j \mu_{n|ji} \right).
\end{aligned}$$

But the quantity inside the parentheses is zero as long as we choose the reproduction levels to be optimal for the given thresholds and next-quantizer map via equation (33). Thus,

$$\sum_{j=1}^{S} \sum_{i=1}^{T} \pi_{ji} \frac{\partial D_{ji}}{\partial \xi_l^q} = \sum_{j=1}^{S} \pi_{jq} \frac{\partial D_{jq}}{\partial \xi_l^q}.$$

But we can simplify this expression further. Observe that

$$\begin{aligned}
\sum_{j=1}^{S} \pi_{jq} \frac{\partial D_{jq}}{\partial \xi_l^q} &= \sum_{j=1}^{S} \pi_{jq} \left[w_j g_j(\xi_l^q)(\eta_{l+1}^q - \eta_l^q)(2\xi_l^q - \eta_{l+1}^q - \eta_l^q) \right. \\
&\qquad \left. + 2w_j \sum_{n=1}^{C} \frac{\partial \eta_n^q}{\partial \xi_l^q} [\eta_n^q \gamma_{n|jq} - \mu_{n|jq}] \right] \\
&= (\eta_{l+1}^q - \eta_l^q)(2\xi_l^q - \eta_{l+1}^q - \eta_l^q) \left[\sum_{j=1}^{S} \pi_{jq} w_j g_j(\xi_l^q) \right]
\end{aligned}$$

$$+ 2 \sum_{n=1}^{C} \frac{\partial \eta_n^q}{\partial \xi_l^q} \left(\eta_n^q \sum_{j=1}^{S} \pi_{jq} w_j \gamma_{n|jq} - \sum_{j=1}^{S} \pi_{jq} w_j \mu_{n|jq} \right).$$

But the quantity inside the parentheses is zero, again by equation (33). Thus,

$$\sum_{j=1}^{S} \sum_{i=1}^{T} \pi_{ji} \frac{\partial D_{ji}}{\partial \xi_l^q} = (\eta_{l+1}^q - \eta_l^q)(2\xi_l^q - \eta_{l+1}^q - \eta_l^q) \left[\sum_{j=1}^{S} \pi_{jq} w_j g_j(\xi_l^q) \right]. \tag{53}$$

Remark. This last formula looks similar to a formula encountered in the derivation of the Lloyd-Max design equations. The difference is the presence of the function

$$\sum_{j=1}^{S} \pi_{jq} w_j g_j(\xi_l^q)$$

which is *not*, in general, a probability density function. □

Plugging (35) and (53) into (34), we have the complete formula for the partial derivative of the average distortion with respect to the l-th threshold of the q-th quantizer when the distortion measure is weighted squared error:

$$\begin{aligned} \frac{\partial D}{\partial \xi_l^q} &= \sum_{j=1}^{S} \sum_{i=1}^{T-1} (D_{ji} - D_{jT}) \frac{\partial \pi_{ji}}{\partial \xi_l^q} \\ &\quad + (\eta_{l+1}^q - \eta_l^q)(2\xi_l^q - \eta_{l+1}^q - \eta_l^q) \left[\sum_{j=1}^{S} \pi_{jq} w_j g_j(\xi_l^q) \right] \end{aligned} \tag{54}$$

for all $q \in J_T$ and $l \in J_{C-1}$.

The steps required to compute $\partial D / \partial \xi_l^q$ for a given SSQ are listed in Table 2.

5.3 Evaluating the Partial Derivative of the Average Distortion with Respect to an Arbitrary Next-Quantizer Map Parameter

From (4), we get

$$\frac{\partial D}{\partial \tau_{q|pl}} = \sum_{j=1}^{S} \sum_{i=1}^{T} D_{ji} \frac{\partial \pi_{ji}}{\partial \tau_{q|pl}} + \sum_{j=1}^{S} \sum_{i=1}^{T} \pi_{ji} \frac{\partial D_{ji}}{\partial \tau_{q|pl}}. \tag{55}$$

Table 2: Summary of the steps required for computing the partial derivative of the average distortion with respect to an arbitrary quantizer threshold.

To compute the derivative w.r.t. ξ_l^q:

1. Form θ^{ql} using (51).
2. Solve the system $\Phi y = \theta^{ql}$ for y, where Φ is given by (26).
3. Recover $\partial\pi_{ji}/\partial\xi_l^q$ from y via the formula $\partial\pi_{ji}/\partial\xi_l^q = y_{(j-1)(T-1)+i}$.
4. Obtain $\partial D/\partial\xi_l^q$ from (54).

We now prove that the second double summation on the right-hand side of this equation is identically zero. To begin,

$$\begin{aligned}\frac{\partial D_{ji}}{\partial\tau_{q|pl}} &= \frac{\partial}{\partial\tau_{q|pl}}\left[\sum_{n=1}^{C}\int_{\xi_{n-1}^i}^{\xi_n^i} d_j(x,\eta_n^i)g_j(x)\,dx\right] \\ &= \sum_{n=1}^{C}\int_{\xi_{n-1}^i}^{\xi_n^i} g_j(x)2w_j(\eta_n^i - x)\frac{\partial\eta_n^i}{\partial\tau_{q|pl}}\,dx,\end{aligned}$$

since $d_j(x,y) = w_j(x-y)^2$. Using equations (32) and (14),

$$\frac{\partial D_{ji}}{\partial\tau_{q|pl}} = 2w_j\sum_{n=1}^{C}\frac{\partial\eta_n^i}{\partial\tau_{q|pl}}[\eta_n^i\gamma_{n|ij} - \mu_{n|ij}].$$

Therefore,

$$\begin{aligned}\sum_{j=1}^{S}\sum_{i=1}^{T}\pi_{ji}\frac{\partial D_{ji}}{\partial\tau_{q|pl}} &= \sum_{j=1}^{S}\sum_{i=1}^{T}\pi_{ji}2w_j\sum_{n=1}^{C}\frac{\partial\eta_n^i}{\partial\tau_{q|pl}}[\eta_n^i\gamma_{n|ji} - \mu_{n|ji}] \\ &= 2\sum_{i=1}^{T}\sum_{n=1}^{C}\frac{\partial\eta_n^i}{\partial\tau_{q|pl}}\left(\eta_n^i\sum_{j=1}^{S}\pi_{ji}w_j\gamma_{n|ij} - \sum_{j=1}^{S}\pi_{ji}w_j\mu_{n|ij}\right).\end{aligned}$$

But the quantity inside the parentheses is zero because the reproduction levels are chosen to satisfy equation (33). Thus, the second double summation in equation (55) is identically zero, so (55) reduces to

$$\frac{\partial D}{\partial\tau_{q|pl}} = \sum_{j=1}^{S}\sum_{i=1}^{T}D_{ji}\frac{\partial\pi_{ji}}{\partial\tau_{q|pl}}. \tag{56}$$

Equation (56) can be further simplified. Since

$$\sum_{i=1}^{T} \pi_{ji} = \rho_j,$$

we have

$$\frac{\partial \pi_{jT}}{\partial \tau_{q|pl}} = -\sum_{i=1}^{T-1} \frac{\partial \pi_{ji}}{\partial \tau_{q|pl}},$$

which enables us to rewrite the double sum as

$$\frac{\partial D}{\partial \tau_{q|pl}} = \sum_{j=1}^{S} \sum_{i=1}^{T-1} (D_{ji} - D_{jT}) \frac{\partial \pi_{ji}}{\partial \tau_{q|pl}}. \tag{57}$$

To evaluate this double sum, we can now proceed as we did in the previous subsection. But now the the system of equations to be solved for the partial derivatives of the π_{ji}'s has the form $\Phi y = \theta^{qpl}$, where Φ is the matrix whose elements are given by (26), and θ^{qpl} is a vector whose elements depend on q, p and l.

As before, let $i, k \in J_{T-1}$, $j, m \in J_S$, and

$$r = (j-1)(T-1) + i \quad \text{and} \quad c = (m-1)(T-1) + k.$$

Then the element in the r-th row and the c-th column of $\partial \Phi / \partial \tau_{q|pl}$ is given by (see equation (26))

$$\frac{\partial \Phi_{rc}}{\partial \tau_{q|pl}} = \frac{\partial A_{ji|mk}}{\partial \tau_{q|pl}} - \frac{\partial A_{ji|mT}}{\partial \tau_{q|pl}}. \tag{58}$$

By equation (17), we have

$$\begin{aligned} \frac{\partial A_{ji|mk}}{\partial \tau_{q|pl}} &= \frac{\partial}{\partial \tau_{q|pl}} (\lambda_{j|m} \sum_{n=1}^{C} \tau_{i|kn} \gamma_{n|mk}) \\ &= \delta_{iq} \delta_{kp} \lambda_{j|m} \gamma_{l|mk}. \end{aligned} \tag{59}$$

Therefore, since k is only taken to run from 1 through $T-1$ when forming Φ, we have

$$\frac{\partial \Phi_{rc}}{\partial \tau_{q|pl}} = \begin{cases} \frac{\partial A_{ji|mk}}{\partial \tau_{q|pl}}, & \text{if } p < T; \\ -\frac{\partial A_{ji|mT}}{\partial \tau_{q|Tl}}, & \text{if } p = T. \end{cases} \tag{60}$$

Thus,

$$\frac{\partial \Phi_{rc}}{\partial \tau_{q|pl}} = \begin{cases} \delta_{iq}\delta_{kp}\lambda_{j|m}\gamma_{l|mk} & \text{if } p < T; \\ -\delta_{iq}\lambda_{j|m}\gamma_{l|mT} & \text{if } p = T. \end{cases} \tag{61}$$

The r-th element of the vector $(\partial\Phi/\partial\tau_{q|pl})z$, denoted by

$$[\frac{\partial \Phi}{\partial \tau_{q|pl}} z]_r,$$

is found by taking the inner product of the r-th row of $\partial\Phi/\partial\tau_{q|pl}$ and the vector z. Because the elements in the c-th column of $\partial\Phi/\partial\tau_{q|pl}$ and the c-th row of z correspond to the m-th subsource and the k-th quantizer via $c = (m-1)(T-1)+k$, where $m \in J_S$ and $k \in J_{T-1}$, the r-th element of $(\partial\Phi/\partial\tau_{q|pl})z$ is given by (recall that $z_c = \pi_{mk}$ for this ordering)

$$\begin{aligned} [\frac{\partial \Phi}{\partial \tau_{q|pl}} z]_r &= \sum_{c=1}^{S(T-1)} \frac{\partial \Phi_{rc}}{\partial \tau_{q|pl}} z_c \\ &= \sum_{m=1}^{S}\sum_{k=1}^{T-1} \delta_{iq}\delta_{kp}\lambda_{j|m}\gamma_{l|mk}\pi_{mk} \\ &= \delta_{iq}\sum_{m=1}^{S} \lambda_{j|m}\gamma_{l|mp}\pi_{mp} \end{aligned} \tag{62}$$

for $p < T$. On the other hand, for $p = T$ we get

$$\begin{aligned} [\frac{\partial \Phi}{\partial \tau_{q|Tl}} z]_r &= \sum_{c=1}^{S(T-1)} \frac{\partial \Phi_{rc}}{\partial \tau_{q|Tl}} z_c \\ &= \sum_{m=1}^{S}\sum_{k=1}^{T-1} -\delta_{iq}\lambda_{j|m}\gamma_{l|mT}\pi_{mk} \\ &= -\delta_{iq}\sum_{m=1}^{S} \lambda_{j|m}\gamma_{l|mT}(\rho_m - \pi_{mT}). \end{aligned} \tag{63}$$

Now

$$\begin{aligned} \frac{\partial b_r}{\partial \tau_{q|pl}} &= \frac{\partial}{\partial \tau_{q|pl}}[-\sum_{m=1}^{S} A_{ji|mT}\rho_m] \\ &= \begin{cases} 0, & \text{if } p < T; \\ -\delta_{iq}\sum_{m=1}^{S}(\partial A_{ji|mT}/\partial \tau_{q|Tl})\rho_m, & \text{if } p = T. \end{cases} \end{aligned}$$

Thus,

$$\frac{\partial b_r}{\partial \tau_{q|Tl}} = -\delta_{iq} \sum_{m=1}^{S} \rho_m \lambda_{j|m} \gamma_{l|mT}.$$

We can now compute the entries of the vector θ^{qpl}. Since $\partial b / \partial \tau_{q|pl}$ is the all-zero vector when $p < T$, we get

$$\theta^{qpl} = \begin{cases} -\frac{\partial \Phi}{\partial \tau_{q|pl}} z, & \text{when } q < T; \\ \frac{\partial b}{\partial \tau_{q|Tl}} - \frac{\partial \Phi}{\partial \tau_{q|Tl}} z, & \text{when } p = T. \end{cases} \tag{64}$$

Hence,

$$\theta_r^{qpl} = -\delta_{iq} \sum_{m=1}^{S} \lambda_{j|m} \gamma_{l|mp} \pi_{mp} \tag{65}$$

if $p < T$ and

$$\theta_r^{qTl} = -\delta_{iq} \sum_{m=1}^{S} \rho_m \lambda_{j|m} \gamma_{l|mT} + \delta_{iq} \sum_{m=1}^{S} (\rho_m - \pi_{mT}) \lambda_{j|m} \gamma_{l|mT} \tag{66}$$

$$= -\delta_{iq} \sum_{m=1}^{S} \lambda_{j|m} \gamma_{l|mT} \pi_{mT} \tag{67}$$

if $p = T$. Hence, for all $q \in J_{T-1}$, $p \in J_T$, and $l \in J_C$, the formula for the r-th element of the right-hand side vector is given by

$$\theta_r^{qpl} = -\delta_{iq} \sum_{m=1}^{S} \lambda_{j|m} \gamma_{l|mp} \pi_{mp}. \tag{68}$$

Once the system of equations $\Phi(\partial z / \partial \tau_{q|pl}) = \theta^{qpl}$ is solved, we can evaluate (57). This completes our discussion of how to evaluate the partial derivative of the average distortion with respect to an arbitrary next-quantizer map parameter. The steps required to compute $\partial D / \partial \tau_{q|pl}$ for a given SSQ are listed in Table 3.

6 Performance Bound

In this section we obtain a lower bound for the average distortion that can be achieved by an SSQ.

Table 3: Summary of the steps required for computing the partial derivative of the average distortion with respect to an arbitrary next-quantizer-map parameter.

To compute the derivative w.r.t. $\tau_{q|pl}$:

1. Form θ^{qpl} using (68).
2. Solve the system $\Phi y = \theta^{qpl}$ for y, where Φ is given by (26).
3. Recover $\partial\pi_{ji}/\partial\tau_{q|pl}$ from y via the formula $\partial\pi_{ji}/\partial\tau_{q|pl} = y_{(j-1)(T-1)+i}$.
4. Obtain $\partial D/\partial\tau_{q|pl}$ from (57).

Consider the situation where the transmitter knows the state of the switch at the *current* time instant before the next-quantizer decision is made. For this case, we need a bank of at most S quantizers because there are only at most S distinct probability distributions for the next observation. For notational convenience, we will proceed as if we need S distinct quantizers. Thus, the quantizers can be put into one-to-one correspondence with the switch states.

Recall that D_{ji} denotes the average distortion incurred by applying the i-th quantizer to the j-th subsource and D_i denotes the average distortion incurred by applying the i-th quantizer. Since the i-th quantizer is now used only when the switch at the previous time instant was pointing at the i-th subsource, we have

$$D_i = \sum_{m=1}^{M} \lambda_{m|i} D_{im}.$$

The average distortion D for the SSQ is now given by

$$D = \sum_{i=1}^{M} \rho_i D_i,$$

where $\rho_i = P[S_t = i]$. To minimize D, we need to minimize D_i for each $i \in J_S$. Each quantizer can be designed independently since we do not have to worry about the next-quantizer decision rule.

For the weighted squared error distortion of (30), we have

$$D_i = \sum_{m=1}^{M} \lambda_{m|i} \sum_{l=1}^{L} \int_{\xi_{l-1}^i}^{\xi_l^i} w_m(x - \eta_l^i)^2 g_m(x)\, dx.$$

It is straightforward to show that

$$\frac{\partial D_i}{\partial \xi_l^i} = (\eta_{l+1}^i - \eta_l^i)(2\xi_l^i - \eta_{l+1}^i - \eta_l^i) \sum_{m=1}^{M} \lambda_{m|i} w_m g_m(\xi_l^i),$$

and, using equations (32) and (14), that

$$\frac{\partial D_i}{\partial \eta_l^i} = 2 \sum_{m=1}^{M} \lambda_{m|i} w_m [\eta_l^i \gamma_{l|mi} - \mu_{l|mi}].$$

Thus, given a set of quantizer reproduction levels, the optimal decision thresholds are given by

$$\xi_l^i = \frac{\eta_{l+1}^i + \eta_l^i}{2}.$$

Given a set of decision thresholds, the optimal reproduction levels are found via

$$\eta_l^i = \frac{\sum_{m=1}^{M} \lambda_{m|i} w_m \mu_{l|mi}}{\sum_{m=1}^{M} \lambda_{m|i} w_m \gamma_{l|mi}}.$$

Thus, a modified Lloyd algorithm [18] can be used to design each of the M quantizers. Clearly, the average distortion incurred by this system provides a lower bound for the average distortion of all other switched quantizer systems which have quantizers with L levels, regardless of the number of quantizers used. This includes the Jayant Adaptive Quantizer.[6]

Remark. Note that the quantizers are optimized for weighted mixtures of the subsource distributions, and not for the individual subsource distributions themselves. For example, even if all of the subsources were Gaussian, the mixture function would not necessarily be Gaussian (in fact, will usually not even be a density), so a Gaussian quantization characteristic would be suboptimal for that mixture. This is an important observation,

[6] Actually, this number is a lower bound on *all* finite-state quantization systems with tracking receivers.

because it means that adaptive schemes which use a fixed quantization characteristic while adjusting only the scale, like the popular Jayant Adaptive Quantizer, are inherently suboptimal. □

If we view the switched quantizer as attempting to estimate the current switch position before choosing the next quantizer, a tighter lower bound would incorporate the lowest possible decision error rate. This error rate will not in general be zero (as assumed above) since the switch is not observable. We have not yet found this tighter bound.

7 Numerical Results

We conclude this chapter with some numerical results obtained using the algorithm described in the preceding sections. We first briefly describe how we solved the optimization problem, and then present and discuss numerical results for several examples.

We have shown how to calculate the average distortion and the gradient of the average distortion so we can use a gradient-based descent algorithm for solving the design problem. It is desirable to use an algorithm which makes use of second-order information as well so that the algorithm is efficient. We used the IMSL subroutine DNCONG [19], which implements a sequential quadratic programming technique [20] for solving a nonlinear programming problem with nonlinear constraints. This routine does not require an analytic evaluation of the Hessian matrix (i.e., the matrix of second-order partial derivatives), but uses the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method for approximating the Hessian matrix [20]. This subroutine is easy to use, and proved to be quite effective in solving our problem.

Remark. The examples that follow demonstrate the kind of studies that can be conducted using our algorithm. It is obvious, however, that there are many parameters that could be varied in a study of switched scalar quantizers for Hidden Markov sources; namely, the number of quantizers, the number of quantization levels, the number of

subsources, the Markov chain statistics, and the parameters of the subsource probability distributions. Thus, these examples do not exhaust the possible situations of interest. □

Example 1

A major strength of our design approach is that there can be more quantizers than subsources. We now give an example of how additional quantizers can improve performance.

Assume that the HMS has two memoryless zero-mean Gaussian subsources, the first having variance $\sigma_1^2 = 1$ and the second having variance $\sigma_2^2 = 100$. Assume that $\lambda_{2|1} = \lambda_{1|2} = 0.01$. The design algorithm was used to obtain optimal SSQ's with different numbers of quantizers. We used $C = 4$ cells per quantizer, for a coding rate of two bits per sample. The distortion measure was taken to be the average weighted squared error, equation (30), with $w_i = 1/\sigma_i^2$, $i = 1, 2$. This measures the average noise-to-signal ratio. The design was carried out for $T = 2, 3, 4, 5$ and 6 quantizers. Since the performance of the system with six quantizers was only slightly better than the performance of the system with five quantizers, we stopped at $T = 6$.

The average noise-to-signal ratios achieved for the different SSQ's are given in Table 4.[7] Here, $D^o_{SSQ}(T, C)$ denotes the minimum average distortion achievable for an SSQ with T quantizers, each having C cells. The case $T = 1$ corresponds to the optimum fixed quantizer system. Denote the lower bound of Section 6 on the average distortion achievable by any SSQ using any number of C-cell quantizers by $D^*_{SSQ}(C)$. For this example, $D^*_{SSQ}(4) = 0.1332$.

The quantizer thresholds and reproduction levels obtained for the different values of T are listed in Table 5. The quantizers have been indexed in such a way that the lower numbered quantizers are more appropriate for the subsource with smaller variance

[7] Of course, our algorithm only computes a local minimum, not a global minimum, so it is possible that there is a better SSQ for any of the different T's considered here. We believe this to be unlikely for this example, however, given how close we are to the theoretical limit for SSQ performance.

Table 4: A comparison of the average noise-to-signal ratios achieved by SSQ's with different numbers of quantizers. The quantizers all have $C = 4$ cells. The source is an HMS with $S = 2$ memoryless Gaussian subsources. The HMS has parameters $\sigma_1^2 = 1$, $\sigma_2^2 = 100$, $\lambda_{2|1} = \lambda_{1|2} = 0.01$. The column labeled "% Improvement" shows the reduction in the noise-to-signal ratio achieved over the SSQ with one less quantizer. The final column compares the performance of the designed SSQ to the lower bound for all SSQ's using 4-cell quantizers. For this case, the lower bound on the achievable distortion is $D^*_{SSQ}(4) = 0.1332$.

T	$D^o_{SSQ}(T,4)$	% Improvement	$D^o_{SSQ}(T,4)/D^*_{SSQ}(4)$
1	.2663	—	2.00
2	.2071	22.2%	1.55
3	.1852	10.6%	1.39
4	.1687	8.9%	1.27
5	.1612	4.4%	1.21
6	.1590	1.4%	1.19

and the higher numbered quantizers are more appropriate for the subsource with larger variance. Note that each SSQ includes quantizers which are asymmetric about zero, even though both subsources have pdfs which are symmetric about zero and the optimum scalar quantizers for the individual subsources are known to be symmetric [8].

The joint probabilities of using the i-th quantizer to encode an output of the j-th subsource, the π_{ji}'s, are listed in Table 6. The probability of using the i-th quantizer, denoted π_i, is listed in the fifth column of this table. For example, for the case $T = 4$, the first quantizer will be used about 41.3% of the time in a sufficiently long block of samples, the second quantizer will be used about 7.5% of the time, etc.

Table 6 also displays the conditional probability that subsource j is active given that we are using quantizer i, denoted by $\rho_{j|i}$. For example, for the case $T = 5$, consider those time instants at which quantizer 2 is used. For 89.4% of these time instants, subsource 1 will be active, while subsource 2 will be active the other 10.6% of the time. The observed conditional distributions are consistent with our observation that the lower

Table 5: A comparison of the thresholds and reproduction levels for the quantizers in each of the SSQ systems designed using our algorithm, for $T = 2, 3, 4, 5$ and 6. The quantizers all have $C = 4$ cells. We also include the optimal fixed quantizer parameters. The source is an HMS with $S = 2$ memoryless Gaussian subsources, with $\sigma_1^2 = 1$, $\sigma_2^2 = 100$, and $\lambda_{2|1} = \lambda_{1|2} = 0.01$.

Case	Quantizer Index i	Thresholds			Reproduction Levels			
		ξ_1^i	ξ_2^i	ξ_3^i	η_1^i	η_2^i	η_3^i	η_4^i
$T = 1$	1	-6.770	0.000	6.770	-12.730	-0.810	0.810	12.730
$T = 2$	1	-0.636	0.600	4.203	-1.255	-0.016	1.215	10.606
	2	-6.067	0.212	6.957	-12.200	-0.730	1.041	12.872
$T = 3$	1	-0.863	0.205	1.389	-1.427	-0.299	0.710	1.868
	2	-3.441	0.041	3.552	-9.644	-0.772	0.825	9.925
	3	-6.180	0.298	7.069	-12.285	-0.719	1.180	12.958
$T = 4$	1	-0.924	0.099	1.169	-1.471	-0.378	0.576	1.671
	2	-1.537	0.000	1.537	-2.093	-0.632	0.632	2.093
	3	-3.520	0.219	5.865	-10.218	-0.670	0.973	12.050
	4	-5.852	3.072	11.867	-12.040	-0.165	6.968	16.766
$T = 5$	1	-0.993	0.028	1.060	-1.523	-0.442	0.498	1.578
	2	-1.258	0.000	1.258	-1.772	-0.551	0.551	1.772
	3	-3.573	0.000	3.573	-10.041	-0.799	0.799	10.041
	4	-5.113	0.279	6.500	-11.494	-0.674	1.085	12.526
	5	-6.038	3.166	11.965	-12.179	-0.235	7.075	16.847
$T = 6$	1	-1.006	0.000	1.006	-1.533	-0.462	0.462	1.533
	2	-1.101	0.000	1.101	-1.618	-0.497	0.497	1.618
	3	-1.434	0.000	1.434	-1.975	-0.604	0.604	1.975
	4	-3.480	0.246	5.409	-10.062	-0.651	0.976	11.711
	5	-5.552	0.305	6.645	-11.817	-0.687	1.150	12.635
	6	-6.071	3.188	11.975	-12.203	-0.257	7.094	16.855

Table 6: A comparison of the $\pi_{j|i}$'s and the $\rho_{j|i}$'s for each of the SSQ systems designed using our algorithm, for $T = 2, 3, 4, 5$ and 6. The quantizers all have $C = 4$ cells. The source is an HMS with $S = 2$ memoryless Gaussian subsources, with $\sigma_1^2 = 1$, $\sigma_2^2 = 100$, and $\lambda_{2|1} = \lambda_{1|2} = 0.01$.

Case	Quantizer Index i	π_{1i}	π_{2i}	π_i	$\rho_{1\|i}$	$\rho_{2\|i}$
$T = 2$	1	.43724	.04648	.48372	.9039	.0961
	2	.06276	.45352	.51628	.1216	.8784
$T = 3$	1	.41837	.02616	.44453	.9412	.0588
	2	.04477	.03265	.07742	.5783	.4217
	3	.03686	.44120	.47806	.0771	.9229
$T = 4$	1	.40054	.01279	.41333	.9691	.0309
	2	.05996	.01508	.07504	.7990	.2010
	3	.03034	.09046	.12080	.2512	.7488
	4	.00917	.38167	.39084	.0235	.9765
$T = 5$	1	.35422	.00719	.36141	.9801	.0199
	2	.10051	.01194	.11245	.8938	.1062
	3	.02540	.02256	.04796	.5296	.4704
	4	.01403	.10891	.12294	.1141	.8859
	5	.00583	.34941	.35524	.0164	.9836
$T = 6$	1	.29006	.00374	.29380	.9873	.0127
	2	.12909	.00590	.13499	.9563	.0437
	3	.04614	.01053	.05667	.8142	.1858
	4	.01933	.03282	.05215	.3707	.6293
	5	.01035	.11235	.12270	.0844	.9156
	6	.00503	.33467	.33970	.0148	.9852

numbered quantizers are more appropriate for the first subsource while the high numbered quantizers are more appropriate for the second subsource. Note that quantizer 2 of the three-quantizer SSQ and quantizer 3 of the five-quantizer SSQ are used when we are almost equally unsure about which subsource is active.

The average noise-to-signal ratios incurred by applying the i-th quantizer to the j-th subsource are given in Table 7.

The quantizer transition diagrams for each of the SSQ's are shown in Figure 1. The numbers inside the circles are the quantizer indices. An arrow is drawn from the circle representing quantizer i to the circle representing quantizer j if there is an i-th quantizer cell for which there is a positive probability of going next to quantizer j. The numbers next to the arrows indicate the cell number which can cause the transition, with the probability of that transition given in parentheses if the decision has a random component. The absence of a number in parentheses means that observing the given cell always causes the specified transition. The quantizer cells are numbered from left to right, so cell number 1 is the left-most cell on the real number line, while cell number 4 is the right-most. Consider now the situation for $T = 5$. If we observe cells 2 or 3 of quantizer 1, we will always use quantizer 1 again. If we observe cell 4 of quantizer 1, we will always use quantizer 2 at the next time instant. On the other hand, if we observe cell 1 of quantizer 1, we will pick a uniformly distributed random number between zero and one. If this number is no greater than 0.563, we will again use quantizer 1. If this number is greater than 0.563, we will use quantizer 2 at the next time instant. From the figure, we see that for $T = 5$ there are three cells which have stochastic next-quantizer decision rules: cell 1 of quantizer 1, and cell 2 of both quantizers 4 and 5. The second and third quantizer do not have any cells which involve a stochastic next-quantizer decision. It is interesting to note that the second and third quantizers are purely transitional quantizers for this case, i.e., we will never use these quantizers twice in a row. All of the other quantizers have self-loops.

Table 7: A comparison of the average noise-to-signal ratios D_{ji}'s for the individual quantizers for each of the SSQ systems designed using our algorithm, for $T = 2, 3, 4, 5$ and 6. The quantizers all have $C = 4$ cells. The source is an HMS with $S = 2$ memoryless Gaussian subsources, with $\sigma_1^2 = 1$, $\sigma_2^2 = 100$, and $\lambda_{2|1} = \lambda_{1|2} = 0.01$.

Case	Quantizer Index i	D_{1i}	D_{2i}
$T = 1$	1	.3635	.1691
$T = 2$	1	.1907	.5061
	2	.3804	.1683
$T = 3$	1	.1263	.7636
	2	.3770	.2211
	3	.4027	.1665
$T = 4$	1	.1196	.7732
	2	.1702	.7077
	3	.3817	.1946
	4	1.0280	.1348
$T = 5$	1	.1180	.7757
	2	.1323	.7474
	3	.3738	.2175
	4	.3863	.1738
	5	1.0549	.1338
$T = 6$	1	.1176	.7781
	2	.1204	.7671
	3	.1544	.7220
	4	.3850	.1980
	5	.3962	.1699
	6	1.0658	.1335

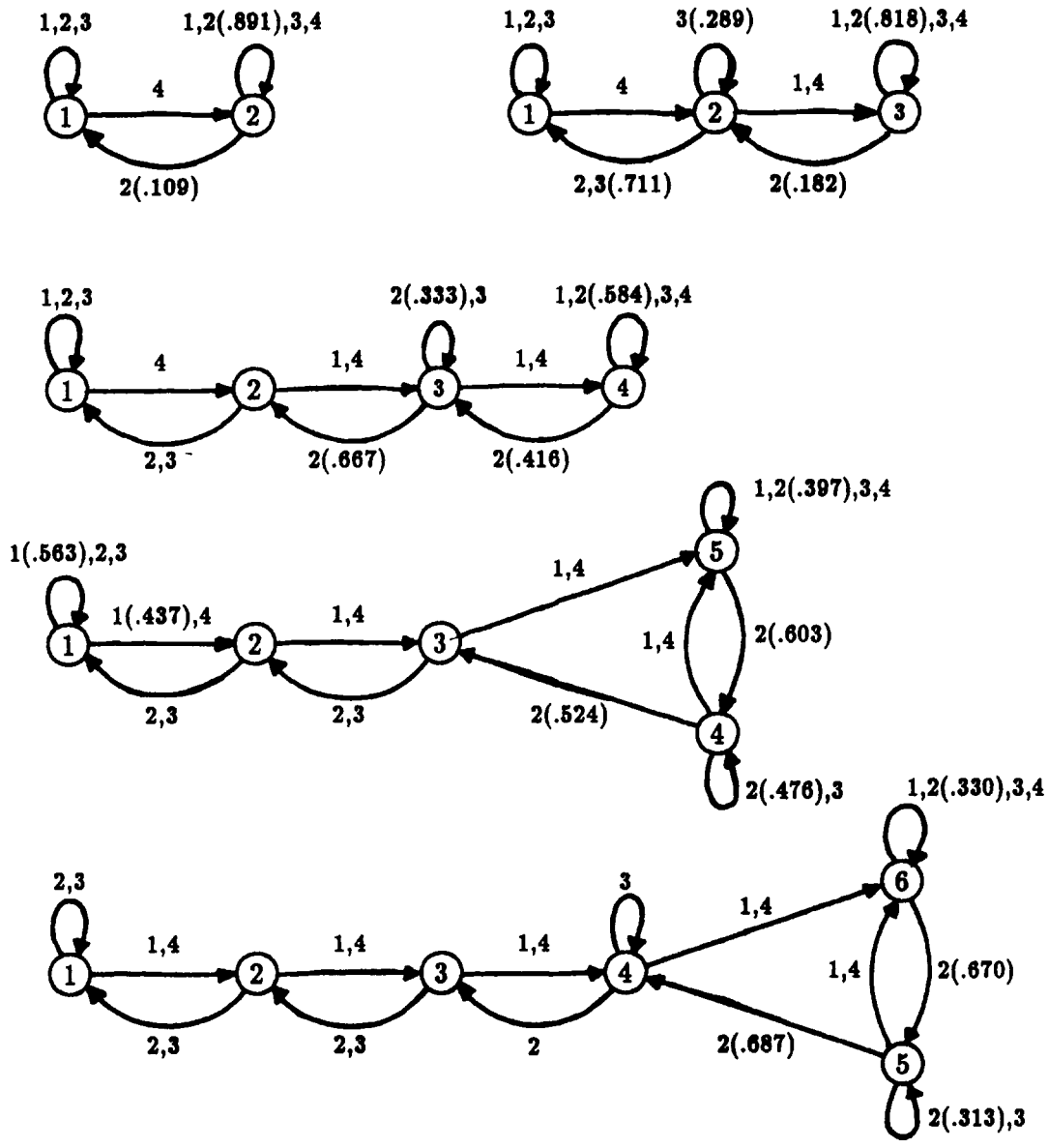


Figure 1: Next-quantizer transition diagrams for $T = 2, 3, 4, 5$ and 6. The quantizers all have $C = 4$ cells. The HMS has $S = 2$ memoryless Gaussian subsources, $\sigma_1^2 = 1$, $\sigma_2^2 = 100$, and $\lambda_{2|1} = \lambda_{1|2} = 0.01$. The numbers inside the circles index the quantizers. The numbers next to the arrows indicate which levels will cause the encoder to go from the quantizer at the tail of the arrow to the quantizer at the head of the arrow, with the number in parentheses indicating the probability of that particular transition when the transition is stochastic.

Finally, we compare how much is gained by using a stochastic next-quantizer map instead of a deterministic next-quantizer map. If the gain is not significant, we prefer the deterministic next-quantizer map because it is simpler to implement. The average noise-to-signal ratios are shown in Table 8. The deterministic SSQ was designed according to the remark made at the end of Section 4, except that the initial next-quantizer map was obtained by rounding off those random components of the optimal next-quantizer map to either 0 or 1 in order to generate an initial next-quantizer map that was deterministic. Values were rounded to either 0 or 1 depending on which number was closer, except that care was taken to ensure that the next-quantizer map was one for which the transmitter states could all communicate, i.e., in setting the values of the initial next-quantizer map to 1 or 0, we picked the closest value except when this choice would cause one of the transmitter states to be an absorbing state. The design algorithm was then run using the equality constraints of (11) to ensure that the final system would have a deterministic next-quantizer map. For example, for $T = 3$, we rounded $\tau_{2|2,3}$ to zero and $\tau_{1|2,3}$ to one, but $\tau_{3|3,2}$ had to be set to zero even though it is closer to one because setting it equal to one would mean that we could never escape from quantizer 3 once we started using it. Note that the performance gain of the stochastic decision rule over the deterministic decision rule decreases as T increases, until it is very small for $T = 5$ and negligible for $T = 6$.

To summarize, in this example we have observed the following things. First, some of the quantizers in the SSQ can be asymmetric even though the quantizers that are optimal for the individual subsource densities are symmetric. Also, there can be a mix of symmetric and asymmetric quantizers. This is an important observation, because it implies that switched quantization systems like the Jayant Adaptive Quantizer that preserve such things as quantizer symmetry cannot be optimal. Second, the optimum next-quantizer map can have stochastic components. Thus, the approach that was used in order to fa-

Table 8: A comparison of the average noise-to-signal ratios achieved by the different SSQ's when stochastic next-quantizer maps were allowed and when next-quantizer maps were constrained to be deterministic. The column labeled "% Improvement" quantifies the reduction in the noise-to-signal ratio achieved by allowing a stochastic next-quantizer map. The HMS has $S = 2$ memoryless Gaussian subsources, with $\sigma_1^2 = 1$, $\sigma_2^2 = 100$, and $\lambda_{2|1} = \lambda_{1|2} = .01$.

N	Random NQM	Deterministic NQM	% Improvement
2	.2071	.2370	12.6%
3	.1852	.2036	9.0%
4	.1687	.1753	3.8%
5	.1612	.1641	1.8%
6	.1590	.1601	0.7%

cilitate a numerical solution turns out to be necessary for getting the optimal system. Third, the stochastic decision rule showed significant improvement over the deterministic rule when T was small. The performance gain of the stochastic rule decreased, however, as T increased. Finally, we were able to get fairly close to the theoretical limit for SSQ's using a small number of quantizers.

Example 2

We now compare the *segmental signal-to-noise ratio* (SNRSEG) performance of our SSQ and a Jayant Adaptive Quantizer (JAQ). The advantages of using SNRSEG to compare coding systems for speech are discussed in [8] and [21]. The operation of the JAQ is described in [8, Chapter 4]. The segmental SNR, denoted SNRSEG, is the average of the short-term SNR computed over successive segments of the signal:

$$\text{SNRSEG} = \frac{1}{N_S} \sum_{s=1}^{N_S} 10 \log_{10} \frac{\sum_{i=1}^{n} x_{(s-1)n+i}^2}{\sum_{i=1}^{n} e_{(s-1)n+i}^2}, \tag{69}$$

where n is the segment length, s indexes the segments, i indexes the signal and error samples within a segment, $x_{(s-1)n+i}$ is the i-th signal sample in the s-th segment, and

$e_{(s-1)n+i}$ is the i-th reproduction error sample in the s-th segment. N_S is the number of segments in the waveform being analyzed.

In order to achieve a high SNRSEG, a coding system has to have small relative errors for all modes of the HMS. Because of this, we will use SSQ's designed according to the weighted squared error criterion in the comparison. Weighted squared error and SNRSEG are not equivalent, but a system designed to minimize one of these measures should do a reasonably good job with respect to the other, and weighted squared error is easier to work with.

It remains to describe how we selected the parameters of the JAQ. There are basically two things to choose: the basic quantizer transfer characteristic and the stepsize multipliers. Preliminary computer simulations showed that using a Gaussian quantization characteristic is better than using a uniform quantization characteristic when the subsources are Gaussian. Hence, our basic quantizer transfer characteristic will be the Lloyd-Max quantizer for the unit-variance Gaussian source. The parameters of this quantizer can be found in Table 4.3 of [8] or Table I of [22]. The next problem is to pick the values for the stepsize multipliers. Our approach is rather involved because of our desire to be fair to the JAQ, so the details are not discussed here; see [15]. Basically, the multipliers were assumed to have the structure

$$M_i = \gamma^{k_i}, \quad i = 1, 2, 3, 4. \tag{70}$$

Several sets of k_i's were found that approximately satisfied the so-called *stability equation* [23]

$$.3836\, k_1 + .3226\, k_2 + .2133\, k_3 + .0805\, k_4 = 0. \tag{71}$$

For each set of k_i's, the best γ was found using an exhaustive search within a subinterval of the positive real numbers, by simulating for each γ the corresponding JAQ on a long training sequence. The best multipliers for each of the test cases, i.e., that set of multipliers

Table 9: Optimal JAQ Multipliers, determined by simulation for $L = 8$ and a frame length of 50 samples.

σ_2^2	λ_{12}	λ_{21}	M_1	M_2	M_3	M_4
25	.01	.01	.9228	1	1.0410	1.3248
25	.01	.04	.8817	1	1.0650	1.5540
25	.04	.01	.9507	.9833	1.0343	1.2450
25	.04	.04	.9228	.9606	1.0410	1.5558
100	.01	.01	.8883	1	1.0610	1.5136
100	.01	.04	.8558	1	1.0810	1.7250
100	.04	.01	.9518	.9756	1.0250	1.3121
100	.04	.04	.9019	.9497	1.0530	1.7649
400	.01	.01	.8495	1	1.0850	1.7701
400	.01	.04	.8190	1	1.1050	2.0116
400	.04	.01	.9518	.9756	1.0250	1.3121
400	.04	.04	.8686	.9320	1.0730	2.1707

resulting in the largest value of SNRSEG for the training sequence using a 50-sample segment length, are shown in Table 9. These were the multipliers used in the comparison. The results of the comparison are summarized in Table 10. The frame length n for which the segmental SNR was calculated was fifty samples.

Table 10 also includes the SNRSEG performance of the optimal fixed quantizer designed using the same weighted squared error criterion as the SSQ. These results are listed in the column labeled SNRSEG_{FQ}.

Observe that the optimal SSQ beats the JAQ in all but one case, the one for which $\lambda_{12} = .01, \lambda_{21} = .04$, and $\sigma_2^2 = 25$. Also, for a fixed pair of transition probabilities, as the difference between the variances of the two subsources increases, so does the gap between the SSQ and the JAQ, with one exception. The SSQ always outperformed the fixed quantizer, and did so by a significant amount when the comparison was based on weighted squared error [15], but not always by a significant amount when using SNRSEG. The JAQ usually outperformed the fixed quantizer, but failed to do so as the difference

Table 10: SNRSEG (in dB) performance comparison between the SSQ, JAQ, and optimal fixed quantizer, for 10,000 frames of simulated HMS samples using 50 samples/frame

σ_2^2	λ_{12}	λ_{21}	SNRSEG_{SSQ}	SNRSEG_{JAQ}	Gap	SNRSEG_{FQ}
25	.01	.01	13.01	12.96	.05	11.31
25	.01	.04	12.12	12.18	-.06	11.42
25	.04	.01	13.39	13.17	.22	13.05
25	.04	.04	12.50	11.52	.98	11.48
100	.01	.01	12.74	12.21	.53	11.20
100	.01	.04	11.64	11.22	.42	10.61
100	.04	.01	13.26	12.68	.58	12.04
100	.04	.04	11.99	10.29	1.70	11.86
400	.01	.01	12.65	11.64	1.01	10.90
400	.01	.04	11.48	10.39	1.09	10.27
400	.04	.01	12.78	12.29	.49	11.52
400	.04	.04	11.99	9.39	2.60	11.44

between the variances increased and the switch became more active. Hence, we conclude that when the statistics of a Hidden Markov Source are known, a well designed SSQ will outperform a fixed quantizer system or a JAQ, often by a significant amount.

8 Summary

We have shown how to design switched scalar quantizers for Hidden Markov sources. Our approach provides a solution to a problem that has plagued designers of switched quantizer coding systems for some time: the design of the optimal next-quantizer decision rule. The solution was to parameterize the next-quantizer decision rule, and then use a gradient-based descent algorithm for solving the resulting nonlinear optimization problem. In parameterizing the next-quantizer map, we allowed for stochastic next-quantizer switching rules. In many cases, it turned out that the optimal next-quantizer map did in fact require stochastic components. We briefly discussed how such a system could be implemented in

practice.

We have demonstrated the usefulness of the design algorithm by discussing several design examples. We have shown how the optimal SSQ system is capable of outperforming the most popular AQ system, the Jayant Adaptive Quantizer, when the source is a Hidden Markov Source with known statistics.

There are several problems that remain open. First, it would be useful to extend the design approach to more general finite-state scalar quantization systems having more transmitter states than quantizers. While our simulations suggest that such systems might not gain very much when the subsources are memoryless, we conjecture that for subsources with memory the gain will be sufficient to warrant the additional complexity. Second, the design problem could be extended to include entropy-constrained quantizers [24]. Of course, for the entropy-constrained case, a new bound on performance will have to be derived since our bound is applicable only for switched quantizer systems with a specified number of quantization levels. Third, robustness of this system to channel noise remains to be studied. The results presented in Section 7 assumed that the receiver had access to the exact channel symbols selected by the transmitter, a situation unlikely to occur in practice. Fourth, there is the very broad question of how to design switched scalar quantizers in the face of uncertainty in the modeling of the subsources and the Markov chain. Afterall, the parameters for the Hidden Markov Source model will usually be derived from a finite data record. Finally, it will be most interesting to extend our results to the design of finite-state vector quantizers [25,26,27] for the HMS.

Acknowledgement

The authors would like to thank Dr. Robert Simpson of the Aerospace Technology Center for his careful reading of the first drafts of this paper. They would also like to thank their colleagues at the University of Maryland, Dr. André Tits and Mr. John Garnett, for their

helpful suggestions during the course of this research.

[15] D. M. Goblirsch, *Switched Quantization Systems for Hidden Markov Sources.* PhD thesis, Univ. of Maryland, College Park, in preparation.

[16] R. E. Ziemer and R. L. Peterson, *Digital Communications and Spread Spectrum Systems.* Macmillan Publishing Co., 1985.

[17] R. G. Bartle, *The Elements of Real Analysis.* John Wiley & Sons, second ed., 1976.

[18] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Trans. on Information Theory*, vol. IT-28, pp. 129–137, March 1982.

[19] *User's Manual, MATH/LIBRARY.* IMSL, Houston, TX, 1987.

[20] L. E. Scales, *Introduction to Non-linear Optimization.* Springer-Verlag New York, 1985.

[21] P. Noll, "Adaptive quantizing in speech coding systems," in *Proceedings of the 1974 Zürich Seminar on Digital Communications*, pp. B3(1)–B3(6), March 1974.

[22] J. Max, "Quantizing for minimum distortion," *IEEE Trans. on Information Theory*, vol. IT-6, pp. 7–12, Jan. 1960.

[23] D. J. Goodman and A. Gersho, "Theory of an adaptive quantizer," *IEEE Trans. on Communications*, vol. COM-22, pp. 1037–1045, August 1974.

[24] N. Farvardin and J. W. Modestino, "Optimum quantizer performance for a class of non-Gaussian memoryless sources," *IEEE Trans. on Information Theory*, vol. IT-30, pp. 485–497, May 1984.

[25] R. M. Gray, "Vector quantization," *IEEE ASSP Magazine*, vol. 1, pp. 4–29, April 1984.

[26] M. O. Dunham and R. M. Gray, "An algorithm for the design of labeled-transition finite-state vector quantizers," *IEEE Trans. on Communications*, vol. COM-33, pp. 83–89, Jan. 1985.

[27] J. Foster, R. M. Gray, and M. O. Dunham, "Finite-state vector quantization for waveform coding," *IEEE Trans. on Information Theory*, vol. IT-31, pp. 348–359, May 1985.