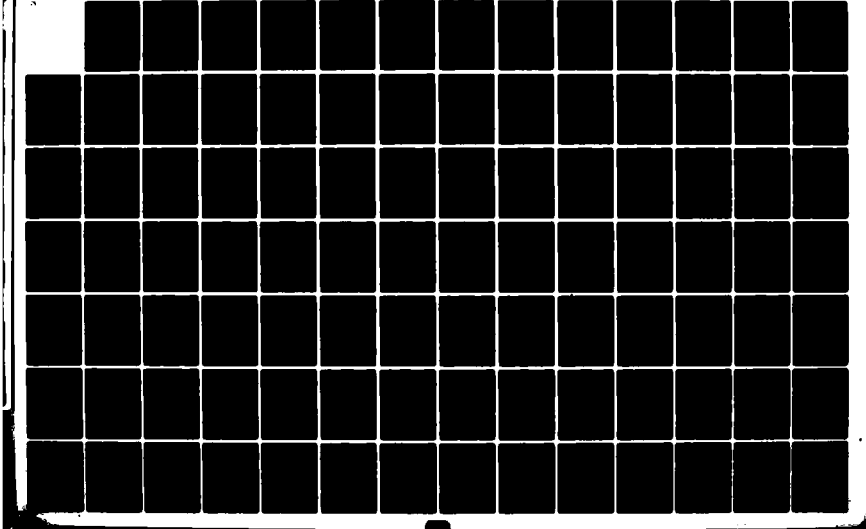
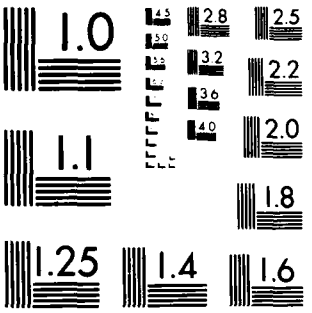
AD-A127 258 ROBUST AND VECTOR QUANTIZATION(U) PRINCETON UNIV NJ 1/2
INFORMATION SCIENCES AND SYSTEMS LAB
P F SWASZEK ET AL. MAR 83 N00014-81-K-0146

UNCLASSIFIED F/G 12/1 NL





Report Number 10

# ROBUST AND VECTOR QUANTIZATION

P.F. Swaszek and J.B. Thomas

INFORMATION SCIENCES AND SYSTEMS LABORATORY

Department of Electrical Engineering and Computer Science
Princeton University
Princeton, New Jersey 08544

MARCH 1983



Prepared for

OFFICE OF NAVAL RESEARCH (Code 411SP)
Statistics and Probability Branch
Arlington, Virginia 22217
under Contract N00014-81-K-0146
SRO(103) Program in Non-Gaussian Signal Processing

S.C. Schwartz, Principal Investigator

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

| REPORT DOCUMENTATION PAGE | | READ INSTRUCTIONS BEFORE COMPLETING FORM |
|---|---|---|
| 1. REPORT NUMBER<br>10 | 2. GOVT ACCESSION NO.<br>AD-A127758 | 3. RECIPIENT'S CATALOG NUMBER |
| 4. TITLE (and Subtitle)<br>"Robust and Vector Quantization | | 5. TYPE OF REPORT & PERIOD COVERED<br>Technical Report<br>Feb.'81-Aug.'82<br>6. PERFORMING ORG. REPORT NUMBER |
| 7. AUTHOR(s)<br>Peter F. Swaszek and John B. Thomas | | 8. CONTRACT OR GRANT NUMBER(s)<br>N00014-81-K-0146 |
| 9. PERFORMING ORGANIZATION NAME AND ADDRESS<br>Information Sciences and Systems Lab.<br>Dept. Electrical Eng. and Computer Sci.<br>Princeton University, Princeton NJ 08544 | | 10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS<br>NR SRO-103 |
| 11. CONTROLLING OFFICE NAME AND ADDRESS<br>Office of Naval Research (Code 411 SP)<br>Department of the Navy<br>Arlington, Virginia 22217 | | 12. REPORT DATE<br>March 1983<br>13. NUMBER OF PAGES<br>116 |
| 14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) | | 15. SECURITY CLASS. (of this report)<br>Unclassified<br>15a. DECLASSIFICATION/DOWNGRADING SCHEDULE |

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

This is part of the Ph.D. Dissertation submitted by P.F. Swaszek to the EECS Department , Princeton University, Princeton, NJ 08544, August 1982.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Robust Quantization
Vector Quantization
Polar Quantizers

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

In this report, we consider the quantization of random sources. The problem of signal quantizer design under an incomplete statistical description of the source is first considered. It is assumed that a histogram of the source on a finite domain is known. The compandor model for a non-uniform quantizer with a large number of output levels is employed. Both minimum mean and minimax error criteria are investigated leading to the design of

(con.t on next pg.)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

(Abstract con't)

piecewise linear compressors. Topics on the partitioning of the histogram are included.

For vector quantization, the design of a spherical coordinates quantizer in k dimensions is discussed. Exact and compandor model solutions are derived as is the factorization of the quantization levels to each quantizer. Numerical examples are presented along with asymptotic results. Also investigated is the optimality of polar quantizers with the subsequent development of optimal circularly symmetric quantizers. Examples of these Dirichlet polar quantizers for the bivariate Gaussian source are included and their performance is compared to optimum error rates. The topic of implementation is considered.

This report closes with a review of the presented material and suggestions for further research.

# ROBUST AND VECTOR QUANTIZATION

by
P.F. Swaszek and John B. Thomas
Department of Electrical Engineering
and Computer Science
Princeton University
Princeton, N.J. 08544



## ABSTRACT

In this report, we consider the quantization of random sources. The problem of signal quantizer design under an incomplete statistical description of the source is first considered. It is assumed that a histogram of the source on a finite domain is known. The compandor model for a non-uniform quantizer with a large number of output levels is employed. Both minimum mean and minimax error criteria are investigated leading to the design of piecewise linear compressors. Topics on the partitioning of the histogram are included.

For vector quantization, the design of a spherical coordinates quantizer in $k$ dimensions is discussed. Exact and compandor model solutions are derived as is the factorization of the quantization levels to each quantizer. Numerical examples are presented along with asymptotic results. Also investigated is the optimality of polar quantizers with the subsequent development of optimal circularly symmetric quantizers. Examples of these Dirichlet polar quantizers for the bivariate Gaussian source are included and their performance is compared to optimum error rates. The topic of implementation is considered.

This report closes with a review of the presented material and suggestions for further research.

TABLE OF CONTENTS

# CHAPTER 1 - INTRODUCTION

The process of quantization is ubiquitous in the areas of communications and signal processing [1]. In its most general sense, quantization is a nonlinear mapping of a continuous time, vector-valued source onto a finite set of values. Many electrical engineering schemes include some form of quantization: digital data communications, digital storage, digital filtering, etc. From these few examples, we see that digital techniques involve quantization, in fact, any analog-to-digital conversion (A/D) requires a simple form of quantizer. A major goal in the design of signal quantizers is accurate representation or reproduction of the source. Quantizers have also been applied to problems in detection and estimation; however, we will consider mainly their use in the direct signal representation sense.

In most cases of interest, the source is not deterministic; hence a statistical measure of performance is required. The performance of the device is measured by some suitable functional of the quantizer itself and of the source's statistical properties. The most common form of measure is that of mean $r$-th error

$$D_r = \int |\mathbf{x} - Q(\mathbf{x})|^r \, p(\mathbf{x}) \, d\mathbf{x}$$

where $Q(\mathbf{x})$ is the output of the quantizer for an input $\mathbf{x}$, the exponent $r$ is some appropriate value usually greater than unity, $p(\mathbf{x})$ is the source density function and the integral is taken over the domain of the source. Other functionals have been suggested in place of $|\mathbf{x} - Q(\mathbf{x})|^r$ and are

often amenable to the techniques considered in this dissertation.

A quantizer is most simply characterized as a set of disjoint regions on the domain of the source whose union completely covers the space. Assigned to each region is an output value. The quantizer's operation consists of deciding which region contains the input value and assigning to $Q(\mathbf{x})$ the associated output value. In one dimension, the input space is the real line, or some subset of it, and the regions are intervals. This zero memory quantizer is the simpliest to analyze; much of the previous research into quantizer design involved solving for the endpoints of these intervals and the associated output points.

A useful model for a zero memory quantizer with unequal step size is the compandor system. This method models the quantizer as a series connection of three elements: a compressor nonlinearity followed by a uniform quantizer followed by an expandor nonlinearity. Any non-uniform quantizer can be modeled in this fashion by appropriately choosing the three components.

Signal quantizers can be separated into three categories: scalar, multidimensional and robust quantizers. As mentioned above, scalar quantizers have received much attention. Their simplicity is embodied in the fact that the regions (intervals) are easily defined. For a vector source, the choice of quantization region is not as obvious; a multitude of shapes and patterns will cover the space. Rate distortion theory, however, suggests that large increases in performance are possible with block coding. The previous research in multidimensional quantizers includes the analysis of asymptotic performance rates of optimal quantizers for vari-

ous sources along with algorithms for their design. Unfortunately, the resulting implementation is usually much more complex than that of scalar quantizers. The design of suboptimal vector quantizers has also been considered. These include the use of uncorrelating filters and polar coordinates representation quantizers.

Robust quantizers are also of interest. By robust we mean that the quantizer performs well over a class of input sources rather than just the one it was designed for. In fact, robust design often means that some minimum level of performance is guaranteed if the input is of a particular class of sources. The source density classes that have already received attention are those with finite domain or with moment constraints.

**OUTLINE**

An outline of the chapters is presented below. Each chapter is self contained so that they may be read in any order.

Chapter 1 contains the above summaries of the problems of data quantization and this dissertation outline. Following these brief notes, a bibliography on data quantization is included. This bibliography is a probe of the available engineering and statistical literature and it provides a reasonable starting point for someone looking into the area for the first time. Brief notes on some of the articles are provided.

The problem of signal quantizer design under an incomplete statistical description of the source is considered in Chapter 2. Previous research in this area is reviewed. For the solution described in this chapter, the statistical information assumed known is that of a histogram of the source on a finite domain. The compandor model for non-uniform quantizers with a large number of output levels is employed. Both minimum mean and minimax error criteria are investigated leading to the design of piecewise linear compressors. Topics on the partitioning of the histogram are included. Quantizers are designed accordingly and compared to other designs.

Chapters 3 and 4 consider the quantization of multidimensional sources. Several investigators [5,6,7,8] have considered polar coordinates quantization of a bivariate, circularly symmetric source. Their schemes quantize the polar coordinates representation of the random variables independently in an attempt to reduce the mean square error below that of an analogous rectangular coordinates quantizer yet retain an implementation simpler than that of the optimal bivariate quantizer. Chapter 3 considers the design of a spherical coordinates quantizer in $k$ dimensions with $k>2$ ($k=2$ matches published results). Exact and compandor model solutions are derived as is the factorization of the quantization levels to each quantizer. Numerical examples are presented along with asymptotic results. Comparisons to the rectangular (one-dimensional) and optimal schemes are included for the multidimensional Gaussian, Pearson Type II and Pearson Type VII spherically symmetric sources.

In the above mentioned literature, it has been shown that for the Gaussian case the polar quantizer outperforms the rectangular quantizer when the number of levels N is large, while for small N, the rectangular form is often better than the polar form. Chapter 4 is an investigation of the optimality of polar quantizers with the subsequent development of optimal circularly symmetric quantizers (labeled Dirichlet polar quantizers). Examples of these Dirichlet polar quantizers for the bivariate Gaussian source are included and their performance is compared to optimum error rates. The topic of implementation is also considered.

Chapter 5 summarizes the results presented and suggests areas for further research.

**REFERENCES**

1. A. Gersho, "Principles of Quantization," *IEEE Trans. Circuits & Systems*, Vol. CAS-25, July 1978, pp.427-437, also *IEEE Comm. Soc. Mag.*, Sept. 1977, pp.16-29.
2. J.A. Bucklew & N.C. Gallagher Jr., "Quantization Schemes for Bivariate Gaussian Random Variables," *IEEE Trans. Inform. Theory*, Vol. IT-25, Sept. 1979, pp.537-543.
3. J.A. Bucklew & N.C. Gallagher Jr., "Two-Dimensional Quantization of Bivariate Circularly Symmetric Densities," *IEEE Trans. Inform. Theory*, Vol. IT-25, Nov. 1979, pp.667-671.
4. W.A. Pearlman, "Polar Quantization of a Complex Gaussian Random Variable," *IEEE Trans. Comm.*, Vol. COM-27, June 1979, pp.892-899.
5. S.G. Wilson, "Magnitude/Phase Quantization of Independent Gaussian Variates," *IEEE Trans. Comm.*, Vol. COM-28, Nov. 1980, pp.1924-1929.

## QUANTIZATION BIBLIOGRAPHY

SCALAR DATA QUANTIZATION


1. E.F. Abaya & G.L. Wise, "Some Remarks on Optimal Quantization," to appear in *Proc. Princeton Conf. Info. Sci. Systems*, 1982. (existence and uniqueness of solution)
2. W.C. Adams Jr. & C.E. Giesler, "Quantization Characteristics for Signals Having Laplacian Amplitude Probability Density Functions," *IEEE Trans. Comm.*, Vol. COM-26, Aug. 1978, pp.1295-1297.
3. V.R. Algazi, "Useful Approximations to Optimal Quantization," *IEEE Trans. Comm. Tech.*, Vol COM-14, June 1966, pp.297-301. (uses calculus of variations to solve for best compandor)
4. E.D. Banta, "On the Autocorrelation Function of Quantized Signal plus Noise," *IEEE Trans. Inform. Theory*, Vol. IT-11, Jan. 1965, pp.114-117.
5. W.R. Bennett, "Spectra of Quantized Signals," *Bell System Tech. Jour.*, Vol. 27, July 1948, pp.446-472.
6. T. Berger, "Optimum Quantizers and Permutation Codes," *IEEE Trans. Inform. Theory*, Vol. IT-18, Nov. 1972, pp.759-765.
7. T. Berger, "Minimum Entropy Quantizers and Permutation Codes," *IEEE Trans. Inform. Theory*, Vol. IT-28, March 1982, pp.149-157.
8. L.E. Brennan & I.S. Reed, "Quantization Noise in Digital Moving Target Indication Systems," *IEEE Trans. Aero. Elect. Systems*, Vol. AES-2, Nov. 1966, pp.655-658. (computes moments of noise with Gauss input)
9. J.D. Bruce, "Optimum Quantization," *MIT Res. Lab. Elect. Tech. Rep. #429*, March 1965. (dynamic programming approach)
10. J.D. Bruce, "On the Optimum Quantization of Stationary Signals," *IEEE Int'l. Conv. Rec.*, 1964, pt. 1, pp.118-124. (dynamic programming approach)
11. J.A. Bucklew & N.C. Gallagher Jr., "A Note on Optimal Quantization," *IEEE Trans. Inform. Theory*, Vol. IT-25, May 1979, pp.365-366. (quantizers preserve mean; MSE equals input minus output variance; uncorrelated and minimum MSE quantizers are different)
12. J.A. Bucklew & N.C. Gallagher Jr., "A Note on the Computation of Optimal Minimum Mean Square Error Quantizers," *IEEE Trans. Comm.* Vol. COM-30, Jan. 1982, pp.298-301. (selection of initial point for iterative solution)
13. J.A. Bucklew & N.C. Gallagher Jr., "Some Properties of Uniform Step Size Quantizers," *IEEE Trans. Inform. Theory*, Vol. IT-26, Sept. 1980, pp.610-613. (MSE equals input variance minus output variance; uniform error rate is not of order $1/N^2$ as is optimum rate; uniform error converges to $\Delta^2/12$)

14. D. Cox, "Note on Grouping," *J. Amer. Stat. Assoc.*, Vol. 52, Dec. 1957, pp.543-547. (similar to Max, but earlier)

15. Cutler, *US Patent Office*, #2605361, July 29, 1952 and #2724740, Nov. 20, 1955.

16. J.A. Demaret & P.P. Bergmans, "Uniform Quantizers: Optimization and Universal Curves," *IEEE Trans. Ind. Elevt. Cont. Instrum.* Vol. IECI-22, Feb. 1975, pp.86-89. (Gaussian inputs)

17. P. Elias, "Bounds on the Performance of Optimum Quantizers," *IEEE Trans. Inform. Theory*, Vol. IT-16, March 1970, pp.172-184. (unusual error functional)

18. T. Fine, "Properties of an Optimum Digital System and Applications," *IEEE Trans. Inform. Theory*, Vol. IT-10, Oct. 1964, pp.287-296. (general system analysis)

19. P.E. Fleischer, "Sufficient Conditions for Achieving Minimum Distortion in a Quantizer," *IEEE Int'l Conv. Rec.* 1964, part 1, pp.104-111

20. N.T. Gaarder & D Slepian, "On Optimal Finite-State Digital Transmission Systems," *IEEE Trans. Inform. Theory*, Vol. IT-28, March 1982, pp.167-186.

21. N.C. Gallagher Jr. & J.A. Bucklew, "Some Recent Developments in Quantization Theory," *Proc. Southeastern Symp. System Theory*, Virginia Beach, May 1980, pp.295-301.

22. V.A. Garmash, "The Quantization of Signals with Non-Uniform Steps," *Telecommunications* 10, Oct. 1957, pp.10-12. (similar to Max, but output is lower breakpoint)

23. M.R. Garey, D.S. Johnson, & H.S. Witsenhausen, "The Complexity of the Generalized Lloyd-Max Problem," *IEEE Trans. Inform. Theory*, Vol. IT-28, March 1982, pp.255-256.

24. A. Gersho, "Principles of Quantization," *IEEE Trans. Circuits & Systems*, Vol. CAS-25, July 1978, pp.427-437, also *IEEE Comm. Soc. Mag.*, Sept. 1977, pp.16-29. (good tutorial)

25. H. Gish & J.N. Pierce, "Asymptotically Efficient Quantizing," *IEEE Trans. Inform. Theory*, Vol. IT-14, Sept. 1968, pp.676-683. (minimum entropy yields uniform levels)

26. T.J. Goblich & J.L. Holsinger, "Analog Source Digitization: a comparison of theory and practice," *IEEE Trans. Inform. Theory*, Vol IT-13, Apr. 1967, pp.323-326. (entropy vs. MSE)

27. L.M. Goodman, "Optimum Sampling and Quantizing Rates," *Proc. IEEE*, Vol. 54, Jan. 1966, pp.90-92.

28. L.M. Goodman & P.R. Drouilhet Jr., "Asymptotically Optimum Pre-emphasis and De-emphasis Networks for Sampling and Quantizing," *Proc. IEEE*, Vol. 54, May 1966, pp.795-796. (Compandor model with linear filter rather than ZNL)

29. R.M. Gray & A.H. Gray Jr., "Asymptotically Optimal Quantizers," *IEEE Trans. Inform. Theory*, Vol. IT-23, Jan. 1977, pp.143-144. (repeat of Gish and Pierce without variational arguments)

30. G.A. Gray & G.W. Zeoli, "Quantization and Saturation Noise Due to Analog-to-Digital Conversion," *IEEE Trans. Aero. Elect. Systems*, Vol AES-7, Jan. 1971, pp.222-223. (Gaussian input to uniform quantizer)

31. A. Habibi, "A Note on the Performance of Memoryless Quantizers," *National Telecomm. Conf. Rec.*, 1975, Vol. 2, no. 38, pp.16-21. (entropy vs MSE)

32. W.J. Hurd, "Correlation Function of a Quantized Sine Wave Plus Gaussian Noise," *IEEE Trans. Inform. Theory*, Vol. IT-13, Jan. 1967, pp.65-68.

33. N. Jayant, Editor, *Waveform Quantization and Coding*, IEEE Press, 1976.

34. H.W. Jones Jr., "Minimum Distortion Quantizers," *NASA Tech. Note TN-D-8384*, March 1977. (tables of quantizers)

35. S.A. Kassam, "Quantization Based on the Mean Absolute Error Criteria," *IEEE Trans. Comm.*, Vol. COM-26, Feb. 1978, pp.267-270.

36. J. Katzenelson, "On Errors Introduced by Combined Sampling and quantization," *IRE Trans. Auto. Control*, Vol. AC-7, April 1962, pp.58-68.

37. J.C. Kieffer, "Exponential Rate of Convergence for Lloyd's Method 1," *IEEE Trans. Inform. Theory*, Vol. IT-28, March 1982, pp.205-210.

38. R.E. Larson, "Optimum Quantization in Dynamic Systems," *IEEE Trans. Auto. Control*, Vol AC-12, April 1967, pp 162-168 and "Reply," Vol. AC-17, April 1972, pp.274-276.

39. Limb, "Design of Dither Waveforms for Quantized Visual Signals," *Bell System Tech. Jour.*, Vol. 48, Sept. 1969, pp 2555-2599.

40. B. Lippel, M. Kurland & A.H. Marsh, "Ordered Dither Patterns for Coarse Quantization of Pictures," *Proc. IEEE*, Vol. 59, March 1971, pp.429-431.

41. S.P. Lloyd, "Least Squares Quantization in PCM," *IEEE Trans. Inform. Theory*, Vol. IT-28, March 1982, pp.129-137.

42. W. Mauersberger, "An Analytic Function Describing the Error Performance of Optimum Quantizers," *IEEE Trans. Inform. Theory*, Vol. IT-27, July 1981, pp.519-521. (generalized Gaussian family)

43. J. Max, "Quantizing for Minimum Distortion," *IRE Trans. Inform. Theory*, Vol. IT-6, March 1960, pp.7-12. (necessary conditions, a design algorithm, asymptotic error rates)

44. J.E. Mazo, "Quantization Noise and Data Transmission," *Bell System Tech. Jour.*, Vol. 47, Oct. 1968, pp.1737-1753. (probability of error in transmission)

45. D.G. Messerschmitt, "Quantizing for Maximum Output Entropy," *IEEE Trans. Inform. Theory*, Vol.IT-17, Sept. 1971, pp.612. (equally probable regions)

46. A.N. Netravali & R. Saigal, "Optimum Quantizer Design Using a Fixed-Point Algorithm," *Bell System Tech. Jour.*, Vol. 55, Nov. 1976, pp.1423-1435. (with entropy constraint)

47. R. Nune & K.R. Rao, "Optimal Quantization of Standard Distribution Functions," *Southeastern Symp. System Theory*, Virginia Beach, 1980, pp.319-325. (mean fourth and sixth error)

48. J.B. O'Neal Jr., "A Bound on Signal-to-Quantizing Noise ratios for Digital Encoding Systems," *Proc. IEEE*, Vol. 55, March 1967, pp.287-292. (Shannon theory for quantizer error, entropy coding)

49. M.D. Paez & T.H. Glisson, "Minimum Mean-Square-Error Quantization in Speech PCM and DPCM Systems," *IEEE Trans. Comm.*, Vol. COM-20, April 1972, pp.225-230. (mu-law vs. optimum for gamma and Laplace densities)

50. P.F. Panter & W. Dite, "Quantization Distortion in Pulse Count Modulation with Non-linear Spacing of Levels," *Proc. IRE*, Vol. 39, Jan. 1951, pp.44-48. (does Pearson II example, optimal and log compressors)

51. W.A. Pearlman & G.H. Senge, "Optimal Quantization of the Rayleigh Probability Density," *IEEE Trans. Comm.*, Vol. COM-27, Jan. 1979, pp.101-112.

52. A.A.G. Requicha, "Expected Values of Functions of Quantized Ravdom Variables," *IEEE Trans. Comm.*, Vol. COM-21, July 1973, pp.850-854.

53. L.G. Roberts, "Picture Coding Using Pseudo Random Noise," *IRE Trans. Inform. Theory*, Vol. IT-8, Feb. 1962, pp.145-154.

54. G.M. Roe, "Quantizing for Minimum Distortion," *IEEE Trans. Inform. Theory*, Vol. IT-10, Oct. 1964, pp.384-385. (companding results)

55. L. Schuchman, "Dither Signals and Their Effects on Quantizing Noise," *IEEE Trans. Comm. Tech.*, Vol. Com-12, Dec 1964, pp.162-165.

56. M.P. Schutzenberger, "On the Quantization of Finite Dimensional Messages," *Inform. & Control*, Vol. 1, May 1958, pp.153-158. (relationship of error and entropy)

57. D. Sharma, "Design of Absolutely Optimal Quantizers for a Wide Class of Distortion Measures," *IEEE Trans. Inform. Theory*, Vol. IT-24, Nov. 1978, pp.693-702, "Comments on" by A.V. Trushkin and "Reply" by D.K. Sharma, Vol. IT-28, May 1982, pp.555. (dynamic programming solution)

58. V.M. Shtein, "On Group Signal Transmission with Frequency Division of Channels by the Pulse Code Modulation Method," *Telecomm.*, 1959, #2, pp.169-184.

59. B. Smith, "Instantaneous Companding of Quantized Signals," *Bell System Tech. Jour.*, Vol. 36, May 1957, pp.653-709.

60. A.B. Sripad & D.L. Snyder, "A Necessary and Sufficient Condition for Quantization Errors to be Uniform and White," *IEEE Trans. Acoust. Speech & Signal Proc.*, Vol. ASSP-25, Oct. 1977, pp.442-448.

61. R.W. Stroh & M.D. Paez, "A Comparison of Optimum and Logarithmic Quantization for Speech PCM and DPCM Systems," *IEEE Trans. Comm.*, Vol. COM-21, June 1973, pp.752-757. (histogram of speech)

62. J.E. Thompson & J.J. Sparkes, "A Pseudo-Random Quantizer for Television Signals," *Proc. IEEE*, Vol. 55, March 1967, pp.353-355.

63. A.V. Trushkin, "Sufficient Conditions for Uniqueness of a Locally Optimal Quantizer for a Class of Convex Error Weighting Functions," *IEEE Trans. Inform. Theory*, Vol. IT-28, March 1982, pp.187-198. (generalized Fleischer)

64. A.I. Velichkin, "Correlation Function and Spectral Density of a Quantized Process," *Telecomm. & Radio Eng.*, Vol. II, #7, 1962, pp.70-77.

65. A.I. Velichkin, "Optimum Characteristics of Quantizers," *Telecomm. & Radio Eng.*, Vol. 18, part 2, Feb. 1963, pp.1-7. (geometric construction of optimal compressor; maximum entropy design)

66. B. Widrow, "A Study of Rough Amplitude Quantization by Means of Nyquist Sampling Theory," *IRE Trans. Circuit Theory*, Vol. CT-3, Dec. 1956, pp.266-276. (statistics of quantized signal)

67. B. Widrow, "Statistical Analysis of Amplitude Quantized Sampled-Data Systems," *Trans. AIEE / App. & Ind.*, Vol. 79 pt. II, Dec. 1960, pp.555-568. 1961.

68. M.J. Wiggins & R.A. Branham, "Reduction in Quantizing Levels for Digital Voice Transmission," *IEEE Int'l. Conf. Rec.*, 1963, part 8, pp.282-288. (modulation system discussion)

69. G. Williams, "Quantizing for Minimum Error with Particular Reference to Speech," *IEE Electronic Letters*, Vol.3, April 1967, pp 134-135 (attempt to analytically calculate Laplace Max quantizer; employed medians instead of centroids)

70. R.C. Wood, "On Optimal Quantization," *IEEE Trans. Inform. Theory*, Vol. IT-5, March 1969, pp.248-252. (companding and entropy results)

71. E.M. Zochevskii, V.M. Nikolaev & V.I. Samoilenko, "Estimating the Effect of the Form of an Analog to Digital Converter Output Characteristic on the Nature of the Quantization Noise," *Automation Remote Control*, Vol. 35, Nov. 1974, pp.1853-1856. (moments and densities of noise)

MULTIDIMENSIONAL RESULTS

72. J.A. Bucklew, "Companding and Random Quantization in Several Dimensions," *IEEE Trans. Inform. Theory*, Vol. IT-27, March 1981, pp.207-211.

73. J.A. Bucklew, "Upper Bounds to the Asymptotic Performance of Block Quantizers," *IEEE Trans. Inform. Theory*, Vol. IT-27, Sept. 1981, pp.577-581.

74. J.A. Bucklew & N.C. Gallagher Jr., "Quantization Schemes for Bivariate Gaussian Random Variables," *IEEE Trans. Inform. Theory*, Vol. IT-25, Sept. 1979, pp.537-543.

75. J.A. Bucklew & N.C. Gallagher Jr., "Two-Dimensional Quantization of Bivariate Circularly Symmetric Densities," *IEEE Trans. Inform. Theory*, Vol. IT-25, Nov. 1979, pp.667-671.

76. J.A. Bucklew & N.C. Gallagher Jr., "Some Results in Multidimensional Quantization Theory," *Proc. Princeton Conf. Info. Sci. Systems*, March 1980, pp. 172-176.

77. J.A. Bucklew & G.L. Wise, "A Note on Multidimensional Asymptotic Quantization Theory," *Proc. Allerton Conf. Comm., Cont. Comp.*, 1979, pp.454-463.

78. J.A. Bucklew & G.L. Wise, "Multidimensional Asymptotic Quantization Theory with r-th Power Distortion Measures," *IEEE Trans. Inform. Theory*, Vol. IT-28, March 1982, pp.239-247.

79. D.T.S. Chen, "On Two or more Dimensional Optimum Quantizers," *IEEE Int'l. Conf. ASSP*, 1977, pp.640-643.

80. J.H. Conway & N.J.A. Sloane, "Voronoi Regions of Lattices, Second Moments of Polytopes, and Quantization," *IEEE Trans. Inform. Theory*, Vol. IT-28, March 1982, pp.211-226.

81. J.H. Conway & N.J.A. Sloane, "Fast Quantizing and Decoding Algorithms for Lattice Quantizers and Codes," *IEEE Trans. Inform. Theory*, Vol. IT-28, March 1982, pp.227-232.

82. W.J. Dallas, "Magnitude-coupled Phase Quantization," *Appl. Optics*, Vol. 13, Oct. 1974, pp.2274-2279. (initial paper on Diri-polar results)

83. P. Elias, "Bounds and Asymptotes for the Performance of Multivariate Quantizers," *Ann. Math. Stat.*, Vol. 41, 1970, pp.1249-1259.

84. L. Fejes Toth, "Sur la Representation d'une population infinie par un nombre fine d'elements," *Acta Mathematica*, Magyar Tudomanyos Akademia Budapest, Vol. 10, 1959, pp.299-304.

85. N.C. Gallagher Jr., "Discrete Spectral Phase Coding," *IEEE Trans. Inform. Theory*, Vol. IT-22, Sept. 1976, pp.622-624. (follows Dallas)

86. N.C. Gallagher Jr., "Quantizing Schemes for the Discrete Fourier Transform of a Random Time Series," *IEEE Trans. Inform. Theory*, Vol. IT-24, March 1978, pp.156-163. (polar quantizer, did N=400 example)

87. A. Gersho, "Asymptotically Optimal Block Quantization," *IEEE Trans. Inform. Theory*, Vol. IT-25, July 1979, pp.373-380. (excellent intro and extension of optimal multidimensional quantizing)

88. A. Gersho, "On the Structure of Vector Quantizers," *IEEE Trans. Inform. Theory*, Vol. IT-28, March 1982, pp.157-166.

89. R.M. Gray & E.D. Karnin, "Multiple Local Optima in Vector Quantizers," *IEEE Trans. Inform. Theory*, Vol. IT-28, March 1982, pp.256-261.

90. R.M. Gray, J.C. Kieffer & Y. Linde, "Locally Optimal Block Quantizer Design," *Inform. & Control*, Vol. 45, May 1980, pp.178-198.

91. J.J.Y. Huang & P.M. Schultheiss, "Block Quantization of Correlated Gaussian Random Variables," *IEEE Trans. Comm. Sys.*, Vol. CS-11, Sept. 1963, pp.289-296.

92. Y. Linde, A. Buzo & R.M. Gray, "An Algorithm for Vector Quantizer Design," *IEEE Trans. Comm*, Vol. COM-28, Jan. 1980, pp.84-95. (iterative solution with training data)

93. J. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," *5th Berkeley Symp. Math, Stat. & Prob.*, 1967, Vol. 1, pp.281-297.

94. J. Menez, F. Boeri & D.J. Esteban, "Optimum Quantizer Algorithm for Real Time Block Quantizing," *Proc. Int'l Conf. ASSP*, 1979, pp.980-984.

95. D.J. Newman, "The Hexagon Theorem," *IEEE Trans. Inform. Theory*, Vol. IT-28, March 1982, pp.137-139.

96. W.A. Pearlman, "Quantization Error Bounds for Computer Generated Holograms," *Stanford Info. Systems Lab.*, Tech. Rep. 6503-1, Aug. 1974.

97. W.A. Pearlman, "Polar Quantization of a Complex Gaussian Random Variable," *IEEE Trans. Comm.*, Vol. COM-27, June 1979, pp.892-899.

98. W.A. Pearlman, "Optimum Fixed Level Quantization of the DFT of Achromatic Images," *Proc. Allerton Conf. Comm., Cont. Comp.*, 1979, pp.313-319.

99. D. Pollard, "Quantization and the Method of k-Means," *IEEE Trans. Inform. Theory*, Vol. IT-28, March 1982, pp.199-205. (statistical clustering)

100. K.D. Rines & N.C. Gallagher Jr., "The Design of Two-Dimensional Quantizers using Prequantization," *Proc. Allerton Conf. Comm., Cont. Comp.*, 1979, pp.446-453.

101. K.D. Rines & N.C. Gallagher Jr., "Quantization in Spectral Phase Coding," *Proc. Hopkins Conf. Info. Sci. Systems*, 1979, pp.131-135.

102. K.D. Rines & N.C. Gallagher Jr., "The Design of Two-Dimensional Quantizers using Prequantization," *IEEE Trans. Inform. Theory*, Vol. IT-28, March 1982, pp.232-239.

103. K.D. Rines, N.C. Gallagher Jr. & J.A. Bucklew, "Nonuniform Multidimensional Quantizers," to appear in the *Proc. Princeton Conf. Info. Sci. Systems*, 1982. (uniform quantizers on regions of space)

104. A. Segall, "Bit Allocation and Encoding for Vector Sources," *IEEE Trans. Inform. Theory*, Vol. IT-22, March 1976, pp.162-169. (extends Huang and Schultheiss to other systems: Max, entropy coding, etc.)

105. P.F. Swaszek & J.B. Thomas "k-Dimensional Polar Quantizers for Gaussian Sources," *Proc. Allerton Conf. Comm., Cont. & Comp.*, Sept-Oct. 1981, pp.89-97.

106. P.F. Swaszek & J.B. Thomas "Optimal Circularly Symmetric Quantizers," to appear in *The Journal of the Franklin Institute*,

107. M. Tasto & P.A. Wintz, "Note on the Error Signal of Block Quantizers," *IEEE Trans. Comm.*, Vol. COM-21, March 1973, pp.216-219

108. S.G. Wilson, "Magnitude/Phase Quantization of Independent Gaussian Variates," *IEEE Trans. Comm.*, Vol. COM-28, Nov. 1980, pp.1924-1929. (varies number of phase divisions per level)

109. Y. Yamada, S. Tazaki & R.M. Gray, "Asymptotic Performance of Block Quantizers with Difference Distortion Measures," *IEEE Trans. Inform. Theory*, Vol. IT-26, Jan. 1980, pp.6-14. (extends Gish and Pierce to multidimensional case)

110. P. Zador, *Development and Evaluation of Procedures for Quantizing Multivariate Distributions*, Stanford Univ. Dissert., Dept of Stat., Dec. 1963.

111. P.L. Zador, "Asymptotic Quantization Error of Continuous Signals and the Quantization Dimension," *IEEE Trans. Inform. Theory*, Vol. IT-28, March 1982, pp.139-149.

ROBUST and MISMATCH RESULTS

112. W.G. Bath & V.D. VandeLinde "Robust Quantizers for Signals with Known Moments and Modes," *22nd Midwest Symp. Circuits & Systems*, Philadelphia, June 1979.

113. W.G. Bath & V.D. VandeLinde, "Robust Quantizers Designed using the Companding Approximation," *IEEE Conf. Decision & Control*, Ft Lauderdale, Dec. 1979, pp.483-487.

114. W.G. Bath & V.D. VandeLinde, "Robust Memoryless Quantization for Minimum Signal Distortion," *IEEE Trans. Inform. Theory*, Vol. IT-28, March 1982, pp.296-306.

115. R.M. Gray & L.D. Davisson, "Quantizer Mismatch," *IEEE Trans. Comm.*, Vol. COM-23, April 1975, pp.439-443. (distance measure bound)

116. D. Kazakos, "Robust Quantization," to appear in the *Proc. Princeton Conf. Info. Sci. Systems*, 1982. (minimax solution like Bath and VandeLinde)

117. D. Kazakos, "On the Design of Robust Quantizers," to appear in the *Proc. NTC*, Dec. 1981, New Orleans.

118. W. Mauersberger, "Experimental Results on the Performance of Mismatched Quantizers," *IEEE Trans. Inform. Theory*, Vol. IT-25, July 1979, pp.381-386.

119. J.M. Morris & V.D. VandeLinde, "Robust Quantization of Stationary Signals," *The Johns Hopkins Univ. Tech. Rep. 73-18, Dec. 1973.* (minimax quantizer for finite support density is uniform - analysis approach)

120. J.M. Morris & V.D. VandeLinde, "Robust Quantization of Discrete-Time Signals with Independent Samples," *IEEE Trans. Comm.*, Vol. COM-22, Dec. 1974, pp.1897-1902.

121. P. Papantoni-Kazakos, "A Robust and Efficient Quantization Scheme," *Proc. Princeton Conf. Info. Sci. Systems*, March 1980, pp.177-187.

122. P.F. Swaszek & J.B. Thomas, "Quantization in Unsure Statistical Environments," *Proc. Johns Hopkins Conf. Info. Sci. & Systems*, March 1981, pp.339-344.

DETECTION and ESTIMATION

123. G.C. Bagley, "Digital Processing of Signal Phase Angle," *IEEE Trans. Aero. Elect. Systems*, Vol. AES-9, Nov. 1973, pp.953-954. (sine wave in Gaussian noise)

124. L. Bluestein, "Asymptotically Optimum Quantizers and Optimum Analog to Digital Converters for Continuous Signals," *IEEE Trans. Inform. Theory*, Vol. IT-10, July 1964, pp.242-246.

125. H. Broman, "Quantization of Signals in Additive Noise: Efficiency and Linearity of the Equal Step-size Quantizer," *IEEE Trans. Acoust., Speech & Signal Proc.*, Vol. ASSP-25, Dec. 1977, pp.572-574. (employs uniform quantizer to estimate signal in noise; considers efficacy like values for quantizers)

126. M.M. Buchner, "A System Approach to Quantization and Transmission Errors," *Bell System Tech. Jour.* Vol. 48, May 1969, pp.1219-1247.

127. B.G. Clark, "The Effect of Digitization Errors on Detection of Weak Signals in Noise," *Proc. IEEE*, Vol. 61, Nov. 1973, pp.1654-1655.

128. T. Fine, "Optimum Mean Square Quantization of a Noisy Input," *IEEE Trans. Inform. Theory*, Vol. IT-11, Apr. 1965, pp.293-294.

129. A.R. Gedance, "Estimation of the Mean of a Quantized Signal," *Proc. IEEE*, Aug. 1972, pp.1007-1008.

130. S.A. Kassam, "Optimum Quantization for Signal Detection," *IEEE Trans. Comm.*, Vol. COM-25, May 1977, pp.479-484.

131. F. Kuhlmann, J.A. Bucklew & G. Wise, "Nonuniform Quantization and Transmission of Generalized Gaussain Signals over Noisy Channels," *Proc. Johns Hopkins Conf. Info. Sci. Systems*, March 1981, pp.345-350.

132. A.J. Kurtenbach & P.A. Wintz, "Quantizing for Noisy Channels," *IEEE Trans. Comm. Tech.*, Vol. COM-17, Apr. 1969, pp.291-302.

133. J.M. Morris, "On the Performance of Quantizers for Noise-Corrupted Signal Sources," *Nat. Telecomm. Conf.*, 1975, Vol. 2, #38, pp.22-25.

134. J.M. Morris, "The Performance of Quantizers for a Class of Noise-Corrupted Signal Sources," *IEEE Trans. Comm.*, Vol. COM-24, Feb. 1976, pp.184-189.

135. E.G. Peters, J.S. Boland, L.J. Pinson & W.W. Malcolm, "Quantization Effects on Signal Matching Functions," *IEEE Trans. Inform. Theory*, Vol. IT-24, May 1978, pp.395-398. (uses Sequential Similarity Detection Algorithm)

136. H.V. Poor & D. Alexandrou, "A General Relationship Between Two Quantizer Design Criteria," *IEEE Trans. Inform. Theory*, Vol. IT-26, March 1980, pp.210-212. (the detection quantizer is shown to be the best MSE fit to the optimum nonlinearity for various detection problems when the criterion is efficacy)

137. H.V. Poor & D. Alexandrou, "The Analysis and Design of Data Quantization Schemes for Stochastic-Signal Detection Systems," *IEEE Trans. Comm.* Vol. COM-28, July 1980, pp.983-991.

138. H.V. Poor & J.B. Thomas, "Memoryless Quantizer-Detectors for Constant Signals in m-Dependent Noise," *IEEE Trans. Inform. Theory*, Vol. IT-26, July 1980, pp.423-432.

139. H.V. Poor & J.B. Thomas, "Application of Ali-Silvey Distance Measures in the Design of Generalized Quantizers for Binary Decision Systems," *IEEE Trans. Comm.*, Vol. COM-25, Sept. 1977, pp.893-900.

140. H.V. Poor & J.B. Thomas, "Optimum Quantization for Local Decisions Based on Independent Samples," *The Journal of The Franklin Institute*, Vol. 303, June 1977, pp.549-561.

141. H.V. Poor & J.B. Thomas, "Optimum Data Quantization for a General Signal Detection Problem," *Proc. Asilomar Conf. Cir., Sys. & Comp.*, Nov 1977.

142. H.V. Poor & J.B. Thomas, "Maximum Distance Quantization for Detection," *Proc. Allerton Conf. Circuits and System Theory*, Sept-Oct. 1976, pp.925-934.

143. H.V. Poor & J.B Thomas, "Asymptotically Robust Quantization for Detection," *IEEE Trans. Inform. Theory*, Vol. IT-24, March 1978, pp.222-229. (Tukey-Huber local detection problem)

144. R.J. Richardson, "Quantization of Noisy Channels," *IEEE Trans. Aero. Elect. Systems*, Vol. AES-2, May 1966, pp.362-364.

145. P.K. Varshney, "Combined Quantization-Detection on Uncertain Signals," *IEEE Trans. Inform. Theory*, Vol. IT-27, March 1981, pp.262-265. (N signal models, use detector to pick one of N quantizers)

146. M. Vinokur, "Amplitude Quantization: A New, More General Approach," *Proc. IEEE*, Vol. 57, Feb. 1969, pp.246-247. (error density for signal plus noise input)

# CHAPTER 2 - HISTOGRAM QUANTIZERS

## INTRODUCTION

A signal quantizer is a device which projects a possibly infinitely-valued, k-dimensional space onto a finite set of points. Specification of an N-level quantizer consists of partitioning the k-space into N disjoint regions $S_i$, i=1,...N, and allocating to each region an output point, $y_i$. In the one dimensional case, the input space is the real line or some subset of it and the N regions are intervals. Hence, specifying the N+1 interval endpoints and the N output points uniquely determines the quantizer.

In general, the quantizer output will not equal the input signal, the difference being the quantization error. The design procedure should reduce the effect of this error by minimizing some suitable measure of the distortion induced by the error. One common criterion is mean $r$-th error

$$D_r = \int_{-\infty}^{\infty} |x - Q(x)|^r p(x)\, dx$$

where $p(x)$ is the source probability density function (pdf), Q(x) is the quantizer output for the input x and the integral is taken over the domain of the source. This error can be rewritten as

$$D_r = \sum_{i=1}^{N} \int_{x_i}^{x_{i+1}} |x - y_i|^r p(x)\, dx \qquad (1)$$

where the $x_i$ are the quantizer breakpoints $\{S_i = [x_i, x_{i+1})\}$ and the $y_i$ are the associated output points. Max [1] considered this distortion measure for r=2 and p(x) the Gaussian density and found necessary conditions for

the $x_i$ and $y_i$ to minimize $D_2$. One condition, which holds for most mean error criteria, is that the breakpoints should be Dirichlet partitions of the output points

$$x_i = \frac{1}{2}(y_{i-1}+y_i) \;\; ; i=2,3,\ldots N \;, x_1=-\infty, x_{N+1}=\infty$$

This condition states that the optimal quantizer should map each input point to the nearest output point. All schemes considered herein will have this property which reduces the specification of an N-level quantizer to the allocation of the N output points.

For a large number of levels, several authors [2,3,4] have modeled a non-uniform quantizer as a three part system: a compressor nonlinearity $g$, a uniform quantizer $Q_U$ and an expandor nonlinearity $g^{-1}$ (see Fig. 1). The compressor function $g$ maps the domain of x onto [-1,1] and the quantizer $Q_U$ projects [-1,1] onto N equally spaced output points. Selection of the compressor function $g$ determines the system performance. For the mean $r$-th error criterion, the asymptotic error ($N \to \infty$) for a compressor $g$ with signal pdf $p(x)$ is

$$D_r \approx \frac{1}{(r+1)N^r} \int_{-\infty}^{\infty} \frac{p(x)}{|g'(x)|^r} dx \qquad (2)$$

The calculus of variations can be used to find the best compressor function for the particular pdf:

$$g_r(x) = -1 + \frac{2\int_{-\infty}^{x} p(y)^{1/(r+1)}\, dy}{\int_{-\infty}^{\infty} p(y)^{1/(r+1)}\, dy} \qquad (3)$$

For this compressor, the associated asymptotic mean $r$-th distortion is

$$D_r \approx \frac{1}{2^r N^r (r+1)} \left[\int_{-\infty}^{\infty} p(x)^{1/(r+1)}\, dx\right]^{r+1} \qquad (4)$$

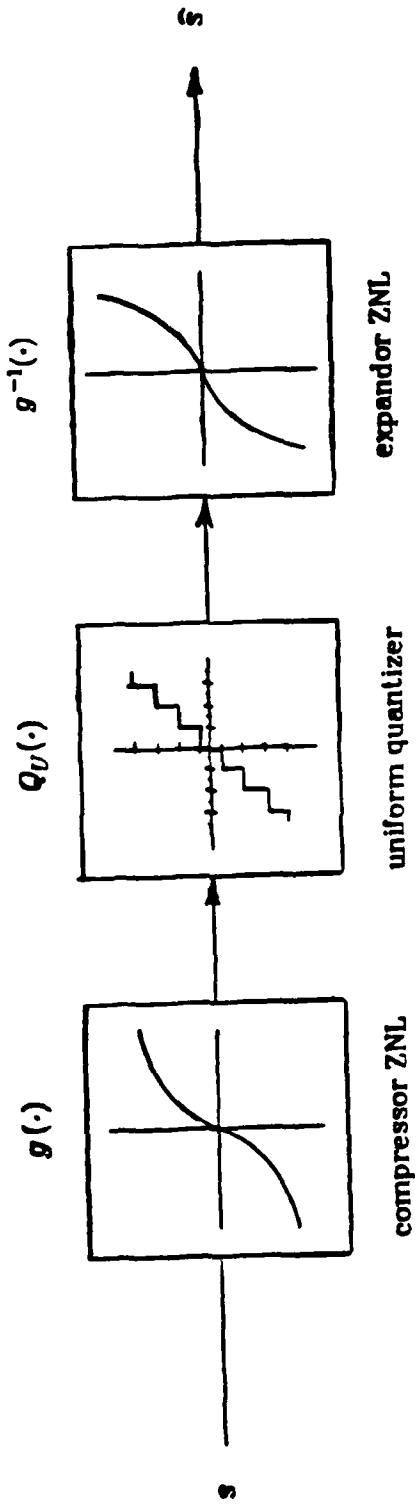



Fig. 1 - Compandor system model.

## ROBUST QUANTIZERS

Minimization of this mean $r$th distortion measure requires exact knowledge of the signal pdf. Most of the available literature on quantizer design assumes this pdf is known. When the source and quantizer are not matched, severe degradation can occur. For example, Fig. 2 presents the MSE ($r=2$) of several quantizers with $N=16$ for a Gaussian input. The quantizers considered are the optimal Gaussian, the optimal uniform Gaussian and the $\mu$-law (see Examples section) quantizers for a unit power source. The input signal is allowed to range in power from $-30$ to 10 $dB$. Both the Gaussian quantizers show large variations in SNR. The $\mu$-law quantizer, although relatively insensitive to variations in source power, has substantially poorer performance when the source and quantizer are nearly matched.

The performance of quantizers when the source and quantizer are not matched was considered in greater detail by Mauersberger [5]. He evaluated mean square error rates for variance and density shape mismatches of generalized Gaussian density quantizers. Suggestions for design under this known density functional form with unknown parameters were presented by him. Robust quantizers, defined as those that perform well over a range of inputs, are desirable

For the situation in which the only available statistical information is that the source has a finite domain (the interval [-c,c]), Morris and VanDeLinde [6] solved the minimax problem

$$\min_{q \in Q} \max_{p \in P} D(p,q)$$

where $Q$ is the set of N-level quantizers, $P$ is the class of density

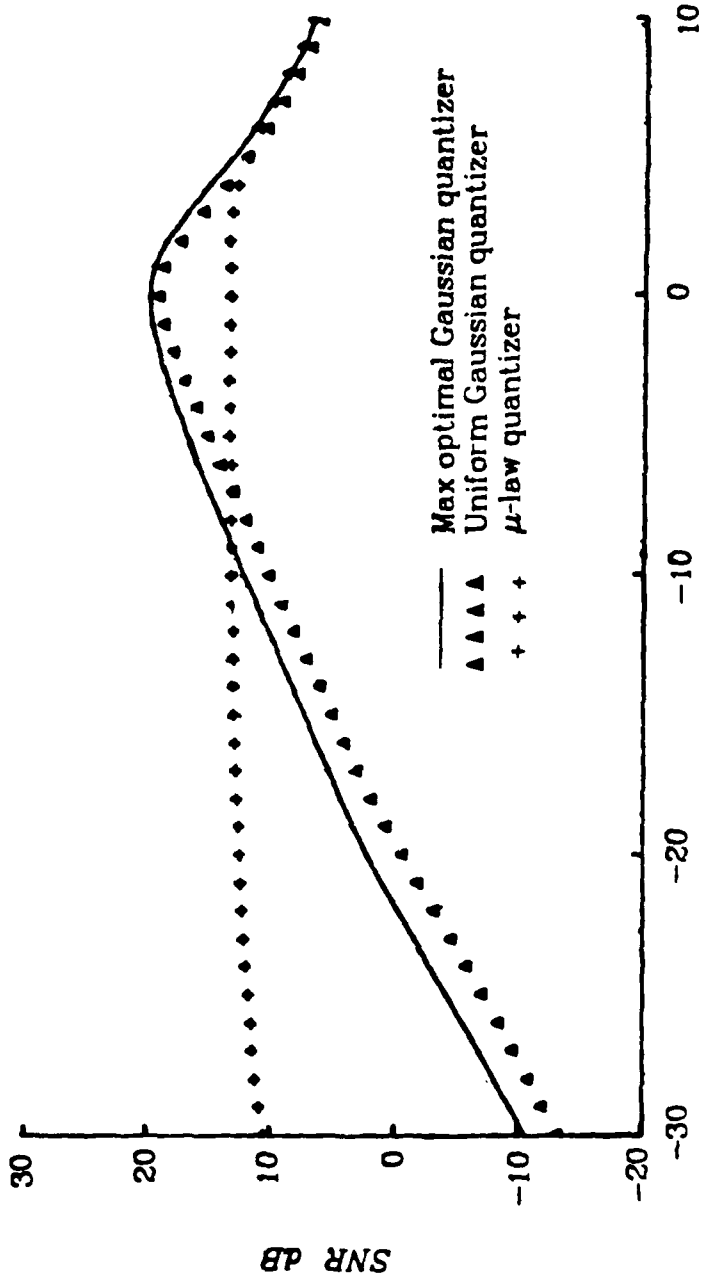



Fig. 2 - MSE curves for a Gaussian source into several $N$=16 quantizers.

functions on [-c,c] and $D(p,q)$ is the distortion measure for density $p$ and quantizer $q$. They investigated mean error distortion measures and characterized the worst case density and the resulting minimax quantizer. The worst case pdf consists of atoms located at the quantizer breakpoints (the points of maximum error) and the minimax quantizer is a uniform (equal step size) quantizer on [-c,c]. Bath and VandeLinde [7] later investigated the minimax quantizer when the source is unimodal and conforms to an integral (moment) constraint. In this case, the worst case pdf is piecewise uniform and a numerical solution procedure is described.

The specific problem considered herein is the design of quantizers when the available statistical information consists of a source histogram. An M-region histogram is characterized by the division of the real line into M disjoint regions $[h_i, h_{i+1})$, i=1,...M, and associating with each region the probability $p_i$ that the source takes a value in that interval. If $p(x)$ is the underlying source density, then

$$p_i = \int_{h_i}^{h_{i+1}} p(x)\,dx \quad ; -\infty \le h_1 \le \cdots h_M \le \infty$$

It will be assumed that $p(x)$ has finite support of $[-L,L]$. Finite support is necessary for the histogram quantizer design and relaxation of this condition will be mentioned later. Without loss of generality, it will be assumed that the underlying density and available histogram are symmetric about zero. The optimal compressor specification then reduces to a function mapping $[0,L]$ onto $[0,1]$:

$$g_r(x) = \frac{\int_0^x p(y)^{1/(r+1)}\,dy}{\int_0^L p(y)^{1/(r+1)}\,dy} \quad ; x \in [0,L] \tag{5}$$

The compressor on $[-L,0]$ is defined odd symmetrically. The histogram on $[0,L]$ has breakpoints (M even)

$$0 = h_{\frac{M}{2}+1} < h_{\frac{M}{2}+2} < \cdots < h_{M+1} = L$$

with the associated probabilities $p_i$, $i = \frac{M}{2}+1,\ldots M$ such that

$$\sum_{i=\frac{M}{2}+1}^{M} p_i = \frac{1}{2}$$

Extensions to the non-symmetric case are easily made.

## QUANTIZER DESIGN FROM A SOURCE HISTOGRAM

Given an M-region histogram with regions $[h_i,h_{i+1})$ and probabilities $p_i$, define the region widths

$$\Delta_i = h_{i+1} - h_i \quad ; \quad i = 1,\ldots M$$

A simple approach to the design of the quantizer would be to assume that the density is piecewise constant of value $p_i/\Delta_i$ over the region $[h_i,h_{i+1})$. With this assumption, the optimal compressor characteristic on $[0,L]$ from Eq.(5) is

$$g_r(x) = s_j x + b_j \quad ; \quad x \in [h_j,h_{j+1}) \ , \ j = \frac{M}{2}+1,\ldots M \tag{6}$$

where $s_j$ and $b_j$ are defined by

$$s_j = \frac{\left[\frac{p_j}{\Delta_j}\right]^{1/(r+1)}}{\sum_{i=\frac{M}{2}+1}^{M} (p_i \Delta_i^r)^{1/(r+1)}} \tag{7}$$

$$b_j = \frac{\sum_{i=\frac{M}{2}+1}^{j-1} (p_i \Delta_i^r)^{1/(r+1)} - \left[\frac{p_j}{\Delta_j}\right]^{1/(r+1)} h_j}{\sum_{i=\frac{M}{2}+1}^{M} (p_i \Delta_i^r)^{1/(r+1)}} \tag{8}$$

This compressor is piecewise linear and asymptotically has mean $r$-th error

$$D_r \approx \frac{1}{2^r N^r (r+1)} \left[ \sum_{i=1}^{M} (p_i \Delta_i^r)^{1/(r+1)} \right]^{r+1} \tag{9}$$

A somewhat more conservative approach would be to consider a minimax-type problem. Direct minimization of the maximum error leads to a uniform quantizer on $[-L,L]$. Instead, consider a single histogram region $[h_j, h_{j+1})$. Generalizing Morris and VandeLinde's result, the quantizer on this region should be uniform. On this region, $g(x)$ is linear and the overall compressor is again piecewise linear

$$g_m(x) = \alpha_j x + \beta_j \quad ; \; x \in [h_j, h_{j+1})$$

Continuity of the compressor function requires that

$$\beta_j = \sum_{i=\frac{M}{2}+1}^{j-1} \alpha_i \Delta_i - \alpha_j h_j$$

The uniform quantizer $Q_U$ has $N$ equispaced outputs on $[-1,1]$. Region $j$ of the histogram, $x \in [h_j, h_{j+1})$, with width $\Delta_j$ and compressor slope $\alpha_j$, maps onto an interval of width $\Delta_j \alpha_j$ in $[-1,1]$. For large $N$, the number of outputs covered by $[h_j, h_{j+1})$ is

$$N_j = N \times \frac{\alpha_j \Delta_j}{2}$$

The maximum error in Region $j$ (since the spacing of levels in $[h_j, h_{j+1})$ is uniform) is

$$d_j = \frac{\Delta_j}{2N_j} = \frac{1}{N\alpha_j}$$

The constraint on the $\alpha_j$'s is

$$\sum_{i=\frac{M}{2}+1}^{M} \alpha_i \Delta_i = 1 \quad ; \; \alpha_i \geq 0$$

For a general error functional $e(\cdot)$ (monotonically increasing in $|\cdot|$), the maximum error on Region $j$ is $e(1/N\alpha_j)$. This error would be due to a point mass located within $[h_j, h_{j+1})$ with error $1/N\alpha_j$. Since error occurs in each region where $p_i > 0$, taking a mean maximum error measure yields

$$D_M = \sum_{i=1}^{M} p_i \, e(1/N\alpha_i)$$

Minimizing this sum with respect to the constraint gives the condition

$$\alpha_j^2 \Delta_j \propto p_j e'(1/N\alpha_j)$$

For the error measure $e(\cdot) = |\cdot|^r$, this condition simplifies to

$$\alpha_j = s_j$$

for $s_j$ as defined by Eq.(7). Using this value of $\alpha_j$ yields

$$\beta_j = b_j$$

for $b_j$ from Eq.(8). The resulting mean $r$-th maximum error is

$$D_M = \frac{1}{N^r}\left[\sum_{i=1}^{M} (p_i \Delta_i^r)^{1/(r+1)}\right]^{r+1} = 2^r (r+1) D_r \qquad (10)$$

Both the piecewise constant density approach and the mean maximum error method produce the same solution when the error functional is $r$-th power. Also, the errors are proportional.

## HISTOGRAM SELECTION

If the histogram data is not prespecified, the designer may have control over the allocation of the histogram regions. Both of the previously considered error measures in Eqs.(9) and (10) resulted in increasing functions of the sum

$$S = \sum_{i=1}^{M} (p_i \Delta_i^r)^{1/(r+1)} \qquad (11)$$

Bounding this term will bound the error. Chebychev type probability inequalities will be used to provide upper bounds on the histogram region probabilities, the $p_j$'s, and the above sum will be minimized over the region widths, the $\Delta_j$'s.

Consider the four region histogram for a symmetric density on $[-L,L]$ with unit variance. Denote the two regions on $[0,L]$ by $[0,a]$ and $[a,L]$. The Chebychev inequality [8] ($\sigma^2$ is the source power)

$$\text{Prob}(x \geq k) \leq \begin{cases} \frac{1}{2} & ; 0 \leq k \leq \sigma \\ \sigma^2 / 2k^2 & ; \sigma \leq k \end{cases}$$

bounds the $p_i$

$$p_3 = \text{Prob}(0 \leq x \leq a) \leq \text{Prob}(0 \leq x) = \frac{1}{2}$$

$$p_4 = \text{Prob}(a \leq x \leq L) \leq \frac{\sigma^2}{2a^2} = \frac{1}{2a^2}$$

The sum in Eq.(11) is then bounded by (since $p_1=p_4$, $\Delta_1=\Delta_4$, $p_2=p_3$ and $\Delta_2=\Delta_3$)

$$S \leq 2\left[\frac{a^2}{2}\right]^{1/(r+1)} + 2\left[\frac{(L-a)^r}{2a^2}\right]^{1/(r+1)}$$

and can be minimized for $a \varepsilon [0,L]$. Optimal region placement is a function of the actual underlying distribution and hence does not result; however, suboptimal allocations do occur and an understanding of region placement develops. Larger values of $M$ are analyzed in a similar manner.

As more information about the underlying density becomes known, tighter bounds on the region probabilities may be found. For example, for symmetric, unimodal densities, the Gauss inequality [8]

$$\text{Prob}(x \geq k) \leq \begin{cases} \frac{1}{2}(1 - k/\sigma\sqrt{3}) & ; 0 \leq k \leq 2\sigma/\sqrt{3} \\ 2\sigma^2/9k^2 & ; 2\sigma/\sqrt{3} \leq k \end{cases}$$

may be employed to solve for the histogram boundaries.

Fig. 3 plots the resulting region placements for $M$=4, 6 and 8 region histograms using the Chebychev inequality. The graphs indicate the locations of the histogram breakpoints for the case $r$=2. Notice that as $\sigma/L$ increases, the solution for $M$=8 degenerates to the $M$=6 selection. Similarly, the $M$=6 lines collapse to the 4 region case. The set of permissible solutions ($-L \le h_1 < h_2 \cdots < h_M \le L$) is convex and the degeneration signifies that the minimum is achieved on the set's boundary (one or more of the $\Delta_i$ going to zero). For larger M, the ratio of $\sigma/L$ must be small to obtain non-zero $\Delta_i$. Fig.4 depicts similar results employing the Gauss inequality (again r=2).

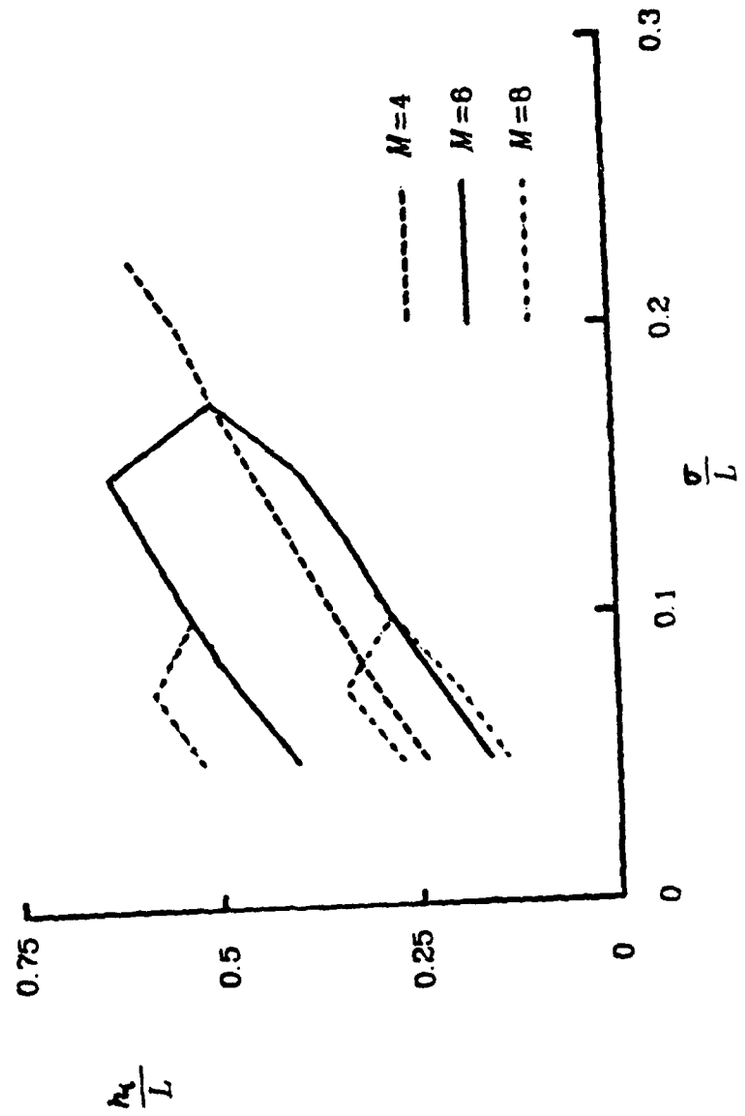



Fig. 3 - Chebychev inequality region placement.

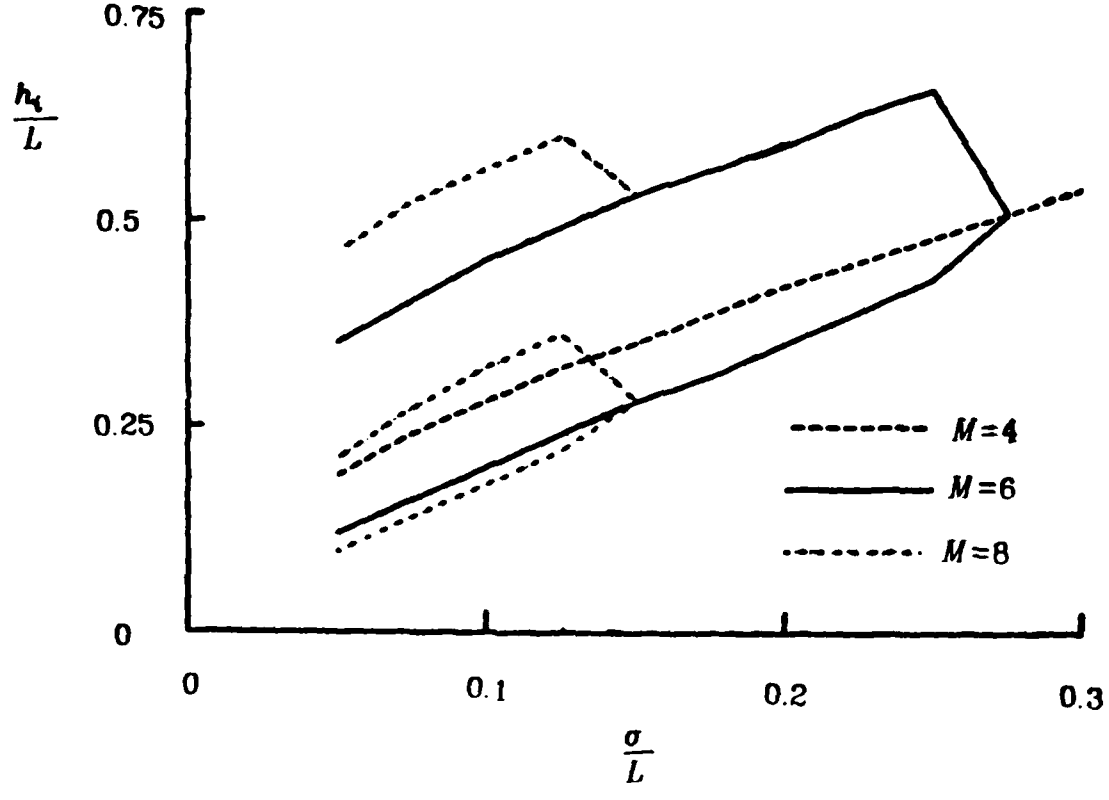



Fig. 4 - Gauss inequality region placement.

## NUMERICAL COMPARISONS

The following examples compare the piecewise linear compressors to the optimal and $\mu$-law compressors under the mean square error criterion. For the example pdf's, the optimal compressor is found from Eq.(3) with the asymptotic error given in Eq.(4). The $\mu$-law quantizer has compressor function on $[0,L]$ of

$$g_\mu(x) = \frac{\ln(1+\mu x / L)}{\ln(1+\mu)}$$

with $\mu$=255. This compressor is an approximation on $[0,L]$ to the function

$$g^*(x) = L + c \log(x / L)$$

where c is a constant. The performance of $g^*(\cdot)$ is found from Eq.(2) to be

$$D_r \approx \frac{\sigma^2}{c^2(r+1)N^r}$$

which is truly robust, being totally independent of $p(x)$. Unfortunately, $g^*(\cdot)$ is undefined for $x$=0; hence, the $\mu$-law approximation is employed. Substituting into Eq.(2) with $r$=2 yields the $\mu$-law compressor's asymptotic mean square error

$$D_2 \approx \frac{L^2 \ln^2(1+\mu)}{3\mu^2 N^2}\left[1 + \frac{2\mu}{L}E\{|x|\} + \frac{\mu^2\sigma^2}{L^2}\right]$$

The comparison of performance for small N (N=16) quantizers is also tabulated. The outputs are found by an inverse mapping through the compressor functions of eight equally spaced points in [0,1]. Dirichlet partitions define the quantizer breakpoints (the $x_i$) and the mean square error is found from Eq.(1).

GAUSSIAN SOURCE: The unit normal source is the canonical choice for the comparison of quantization schemes. The pdf for $x \in [-L,L]$ is

$$p(x) = K\, e^{-x^2/2}$$

with $K$ chosen for unit mass on $[-L,L]$. The optimal compressor function on [0,L] is

$$g_{opt}(x) = \frac{\int_0^x e^{-y^2/6} dy}{\int_0^L e^{-y^2/6} dy}$$

The following results are for $L=5$ ($5\sigma$ loading).

Piecewise linear compressors for 4,6 and 8 region histograms are compared (corresponding to 2,3 and 4 regions on [0,5]). The Gauss inequality bound with $\sigma/L=0.2$ yields suboptimal region placement for M equal to 4 and 6. For M=8, equispacing and a modified Gauss placement are tabulated. Figs. 5 and 6 display these compressor functions. Table I lists the histogram region endpoints and the associated asymptotic error rates. The M=2 quantizer is the uniform quantizer on [-5,5].

For small N (N=16), Table II lists the positive output values for Max's optimal, $\mu$-law and piecewise linear compressors. The piecewise linear examples are the M=4 (Gauss bound) and M=8 (equispaced) versions. Values of mean square error and the associated Signal-to-Noise Ratio are tabulated where

$$SNR = 10 \log_{10} \frac{\sigma^2}{MSE} dB$$

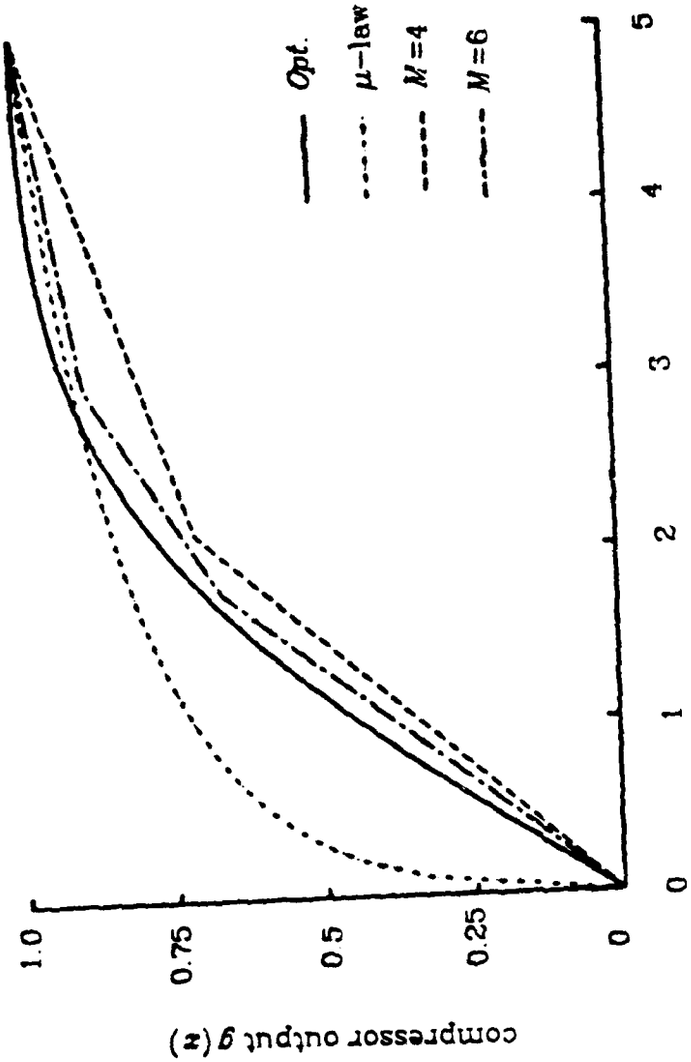



Fig. 5 - Gaussian source compressors.

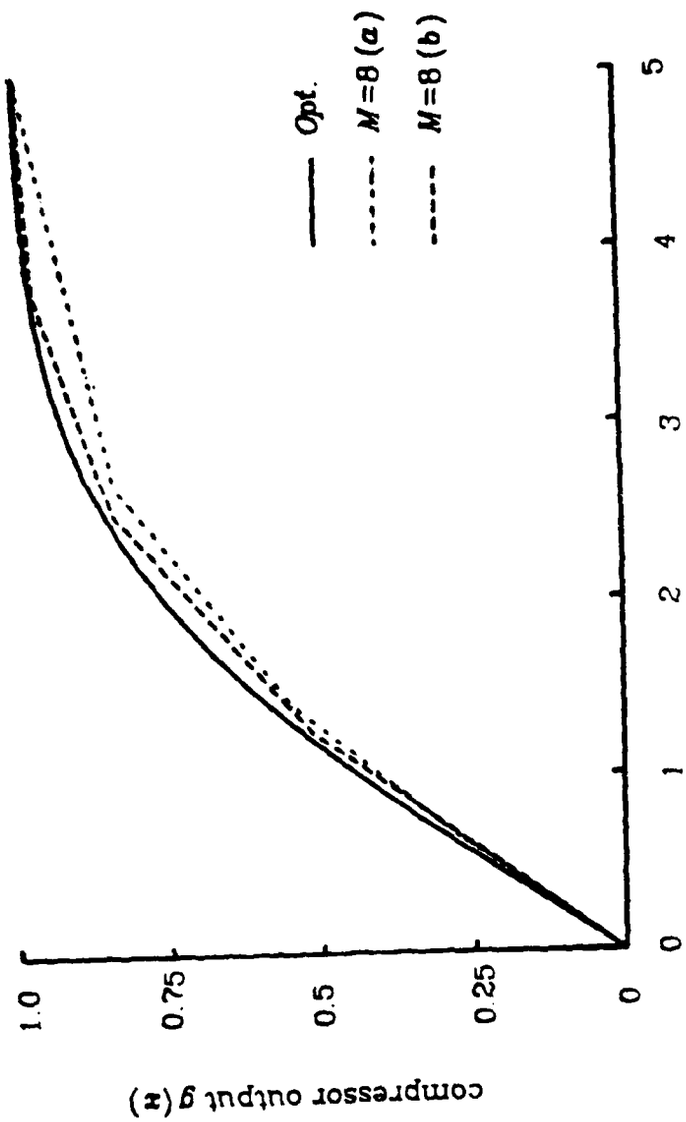



Fig. 6 - Gaussian source compressors

| M | Histogram Divisions | | | | $N^2D$ |
|---|---|---|---|---|---|
| 2 | 5. | | | | 8.33 |
| 4 | 2.1 | 5. | | | 4.00 |
| 6 | 1.75 | 2.95 | 5. | | 3.22 |
| 8(a) | 0.7 | 1.4 | 2.65 | 5. | 3.31 |
| 8(b) | 1.25 | 2.5 | 3.75 | 5. | 3.01 |
| $\mu$-law compressor | | | | | 10.6 |
| optimal compressor | | | | | 2.69 |

Table I - Gaussian source asymptotic error rates

| | Max Opt | $\mu$-law | 4 region | 8 region |
|---|---|---|---|---|
| $y_1$ | 0.1284 | 0.008122 | 0.1855 | 0.1517 |
| $y_2$ | 0.3881 | 0.03585 | 0.5565 | 0.4551 |
| $y_3$ | 0.6568 | 0.09131 | 0.9276 | 0.7586 |
| $y_4$ | 0.9424 | 0.2022 | 1.299 | 1.062 |
| $y_5$ | 1.256 | 0.4241 | 1.670 | 1.433 |
| $y_6$ | 1.618 | 0.8677 | 2.041 | 1.913 |
| $y_7$ | 2.069 | 1.755 | 3.141 | 2.393 |
| $y_8$ | 2.733 | 3.530 | 4.380 | 3.446 |
| MSE | 0.009513 | 0.04098 | 0.01452 | 0.01090 |
| SNR (dB) | 20.2 | 13.9 | 18.4 | 19.6 |

Table II - $N$=16 Gaussian source quantizers.

LAPLACIAN SOURCE. When modeling signal sources, the Laplace source on [-L,L] is sometimes considered

$$p(x) = K\, e^{-\alpha |x|}$$

The optimal compressor on [0,L] is

$$g_{opt}(x) = \frac{1-e^{-\alpha x/3}}{1-e^{-\alpha L/3}}$$

For this pdf, taking L=8 and $\alpha=\sqrt{2}$ yields unit variance. Again, the Gauss inequality bound yields suboptimal region selection for $\sigma/L=0.125$. Figs. 7 and 8 depict the optimal, $\mu$-law and piecewise linear compressors for M=4,6 and 8. Table III lists the histogram breakpoints and the asymptotic error rates

Adams and Geisler [9] tabulated optimal output values for the Laplace source when N=16 and r=2. Table IV lists these values along with the outputs for the $\mu$-law device and the M=4 and 8 histogram quantizers. As in the previous example, MSE and SNR are tabulated for each scheme.

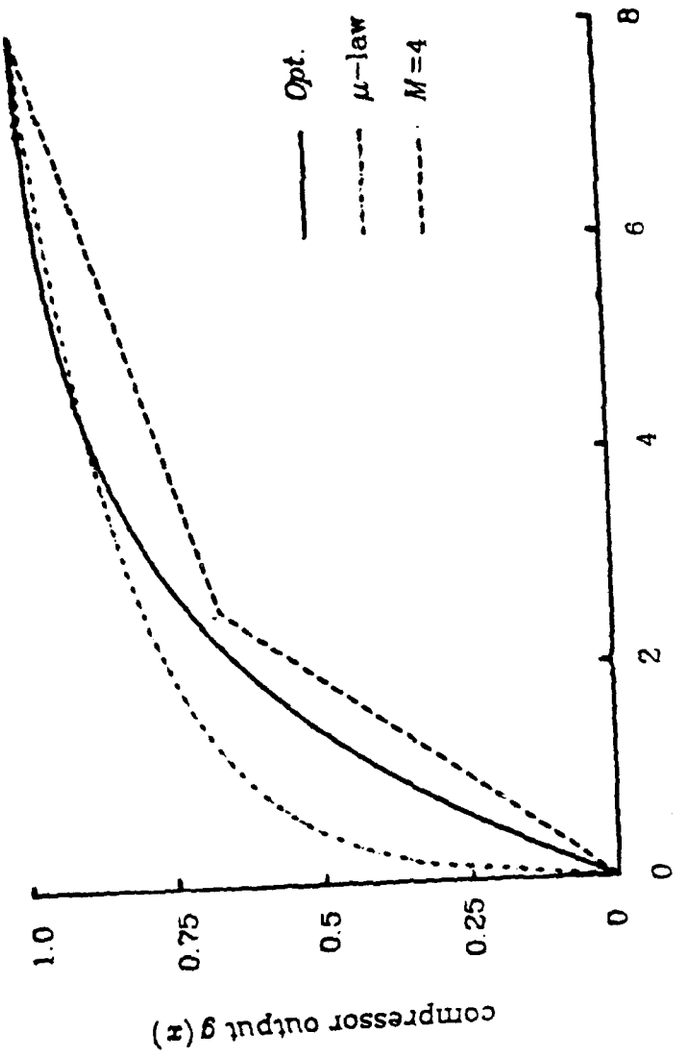



Fig. 7 - Laplacian source compressors.

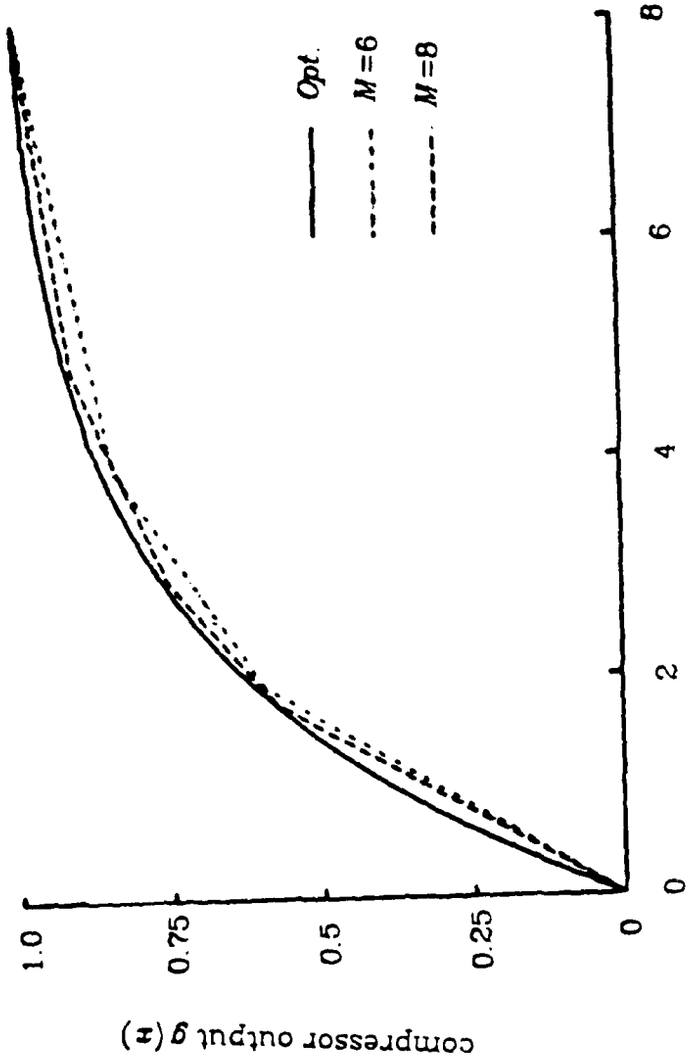



Fig. 8 - Laplacian source compressors.

| M | Histogram Divisions | | | | $N^2D$ |
|---|---|---|---|---|---|
| 2 | 8. | | | | 21.3 |
| 4 | 2.56 | 8. | | | 7.16 |
| 6 | 1.92 | 3.92 | 8. | | 5.47 |
| 8 | 1.76 | 2.88 | 4.8 | 8. | 5.00 |
| $\mu$-law compressor | | | | | 10.7 |
| optimal compressor | | | | | 4.20 |

Table III - Laplacian source asymptotic error rates

| | A&G Opt. | $\mu$-law | 4 region | 8 region |
|---|---|---|---|---|
| $y_1$ | 0.1240 | 0.01299 | 0.2398 | 0.1915 |
| $y_2$ | 0.4048 | 0.05736 | 0.7195 | 0.5745 |
| $y_3$ | 0.7287 | 0.1461 | 1.199 | 0.9575 |
| $y_4$ | 1.111 | 0.3236 | 1.679 | 1.341 |
| $y_5$ | 1.578 | 0.6785 | 2.158 | 1.724 |
| $y_6$ | 2.177 | 1.388 | 2.892 | 2.477 |
| $y_7$ | 3.017 | 2.808 | 4.935 | 3.628 |
| $y_8$ | 4.431 | 5.648 | 6.978 | 5.810 |
| MSE | 0.0161 | 0.0408 | 0.0255 | 0.0192 |
| SNR (dB) | 17.9 | 13.9 | 15.9 | 17.2 |

Table IV - $N=16$ Laplacian source quantizers

## CONCLUSIONS

Robust quantizer design depends entirely upon the amount of information assumed about the class of permissible input distributions. As previously mentioned, the uniform quantizer is the minimax quantizer when only finite support of the density is known and the $\mu$-law quantizer performs well for most input statistics. From the examples, the proposed method of piecewise linear compressor quantizer design is seen to present a viable alternative to uniform quantization or optimal quantization of the "known" pdf. A few other points need to be considered.

1- Initially, the input was assumed to have finite support of $[-L,L]$. Without this constraint, some of the histogram region widths, the $\Delta_i$, would be infinite and the resulting piecewise linear compressor would have sections of zero slope. For non-finite support, select $L$ such that the probability of overload ($x$ outside $[-L,L]$) is small and map the overload regions to the nearest output, $y_1$ or $y_N$.

2- The use of probability inequalities in histogram region selection provided degenerate solutions for $M>8$ unless $\sigma/L$ approached zero. The results for $M=8$ from Table I indicate that equal subdivisions of $[-L,L]$ is a reasonable procedure for larger $M$. For $\Delta_i$ of order $1/M$, the histogram converges uniformly to the underlying density [10] and the piecewise linear compressor converges to the optimum compressor.

3- The $M-1$ cusps of the piecewise linear compressor may seem undesirable. Initially, the quantizer breakpoints were defined as Dirichlet partitions of the output points. On the sides of a cusp, the two

linear segments will in general have different output point spacing and the resulting partition will not fall exactly on the cusp. Hence, a slight rounding of the cusp occurs.

4- The simplicity of the compressor curve calculation [Eqs.(7) and (8)] suggests that this scheme may be employed adaptively. Occasional histogram measurement would keep the quantizer tuned to a non-stationary input.

## REFERENCES


1. J Max, "Quantizing for Minimum Distortion," *IRE Trans. Inform. Theory*, Vol. IT-6, March 1960, pp.7-12.
2. W.R. Bennett, "Spectra of Quantized Signals," *Bell System Tech. Jour.*, Vol. 27, July 1948, pp.446-472.
3. V.R. Algazi, "Useful Approximations to Optimal Quantization," *IEEE Trans. Comm. Tech.*, Vol. COM-14, June 1966, pp.297-301.
4. A. Gersho, "Principles of Quantization," *IEEE Trans. Circuits & Systems*, Vol. CAS-25, July 1978, pp.427-437, also *IEEE Comm. Soc. Mag.*, Sept. 1977, pp.16-29.
5. W. Mauersberger, "Experimental Results on the Performance of Mismatched Quantizers," *IEEE Trans. Inform. Theory*, Vol. IT-25, July 1979, pp.381-386.
6. J.M. Morris & V.D. VandeLinde, "Robust Quantization of Discrete-Time Signals with Independent Samples," *IEEE Trans. Comm.*, Vol. COM-22, Dec. 1974, pp.1897-1902.
7. W.G. Bath & V.D. VandeLinde, "Robust Memoryless Quantization for Minimum Signal Distortion," *IEEE Trans. Inform. Theory*, Vol. IT-28, March 1982, pp.296-306.
8. J.K. Patel, C.H. Kapadia & D.B Owen, *Handbook of Statistical Distributions*, New York; Marcel Dekker, 1976.
9. W.C. Adams Jr. & C.E. Giesler, "Quantization Characteristics for Signals Having Laplacian Amplitude Probability Density Functions," *IEEE Trans. Comm.*, Vol. COM-26, Aug. 1978, pp.1295-1297.
10. W. Wertz, *Statistical Density Estimation: A Survey*, Vandehoech & Ruprecht, 1978.

# CHAPTER 3 - MULTIDIMENSIONAL SPHERICAL COORDINATES QUANTIZATION

## INTRODUCTION

A data quantizer is a mapping of a vector-valued source onto a finite number of points. A general multidimensional characterization of an $N$-level quantizer consists of a partition of the input space into $N$ disjoint regions and the assignment of a particular output value to each region. Implementation requires deciding which of the regions the input is an element of, often a time consuming task. In one dimension, a scalar or zero-memory quantizer has regions which are intervals on the real line; hence its implementation is simple. Both uniform and non-uniform interval width scalar quantizers have been designed for a variety of fidelity criteria and source statistics. The canonical example is Max's unit power, Gaussian probability density function quantizer [1] for the performance criterion Mean Square Error (MSE). The MSE criterion has universal appeal in its tractability and the intuitive notions of noise power and signal-to-noise ratio. Rate distortion theory, however, suggests that multidimensional quantizers may be more efficient.

Research in multidimensional quantization began with the works of Huang and Schultheiss [2] and Zador [3]. Huang and Schultheiss considered the quantization of a correlated Gaussian source. Their system transformed the input vector by a linear device to a set of uncorrelated (hence independent) coordinates and quantized each new coordinate separately. This scheme, although suboptimal, retained a simple implementation and reduced the MSE below that of separate quantizers for the

correlated rectangular coordinates. The solution included the factorization of the levels to each quantizer since the product of the number of levels in each quantizer must equal the total number of levels $N$. Zador considered asymptotic error rates for optimal multidimensional quantization in $k$ dimensions. He derived bounds on the minimally attainable distortion but did not present the actual quantizer design. Later, Gersho [4] and Conway and Sloane [5] discussed optimal quantizer designs of particular dimensionalities.

A major area of interest in suboptimal multidimensional quantizer design involves the use of polar coordinates ($k=2$). After effecting a change from rectangular to polar coordinates, the resulting magnitude and phase are quantized using separate scalar, Max-type quantizers. Both Pearlman [6] and Bucklew and Gallagher [7] considered the quantizing of an independent, bivariate Gaussian random variable in this manner. DFT coefficients, holographic data or a pair of inputs from an independent and identically distributed Gaussian source can be considered as the output of a bivariate Gaussian source. Published results show that the polar form almost always outperforms the rectangular format, a pair of Max quantizers, for Gaussian variates. Bucklew and Gallagher [8] later extended the polar form to any circularly symmetric density of which the bivariate Gaussian is one example. Noting that rectangular quantizers outperform the polar quantizers when $N$ is small, Wilson [9] defined unrestricted polar quantizers which have lower MSE values. Swaszek and Thomas [10] investigated the optimality of the polar schemes and introduced a scheme which resembles the optimal quantizer for the bivariate

Gaussian source, a tessellation of distorted hexagons, but has a simpler implementation.

This chapter describes an extension of the polar quantizers mentioned above to the quantization of a $k$-dimensional spherically symmetric random source. The next section considers the data vector $\mathbf{x}$ of length $k$ with rectangular coordinate elements $x_j$, $j=1,2,...k$. The source statistics are contained in its $k$-dimensional density function $f(\mathbf{x})$. Using a transformation to $k$-dimensional spherical coordinates, the resulting magnitude and $k-1$ angles will be separately quantized. In the third section, the magnitude and angles quantizers are derived when MSE is the performance criterion. Asymptotic results and allocation of the number of levels to each separate quantizer are the topics addressed in the fourth section. Finally, we present several examples.

## SPHERICAL COORDINATES QUANTIZERS

Spherically symmetric sources are characterized by contours of constant height that are hyperspheres in the $k$ dimensional space. These spherically symmetric densities can be generated by replacing the independent variable of a zero mean, unit power univariate density, say $p(x)$, with the square root of a quadratic form [11]. The resulting density has the form:

$$f(\mathbf{x}) = \gamma\, p\left(\sqrt{\mathbf{x}^T\mathbf{x}}\right) \;;\; \mathbf{x} \in \mathbf{R}^k$$

where $\gamma$ is a scaling constant so that the density has unit mass. The resulting multivariate density, however, does not in general have as its marginals the univariate $p(x)$

There is a method of generating spherically symmetric densities with a specific marginal. Assume the marginal has characteristic function $\Phi_1(u)$, and let the m-dimensional characteristic function be

$$\Phi_m(\mathbf{u})=\Phi_1\left[\sqrt{\mathbf{u}^T\mathbf{u}}\right]$$

Taking the inverse transform yields the m-dimensional density $f(\mathbf{x})$. Care must be taken to ensure that $f(\mathbf{x})$ integrates to one and is positive for all $\mathbf{x}$.

In one-dimensional or rectangular quantization, each coordinate $x_j$ is quantized independently by a Max-type, scalar quantizer $Q_j$. The resulting quantization regions, being the cross product of $k$ intervals, are $k$-dimensional rectangular parallelepipeds. Each coordinate has a MSE term $E_1(N_j)$ which is the error associated with a scalar quantizer for the marginal density with $N_j$ levels ($N_1 \times N_2 \cdots \times N_k = N$). The errors, being independent and orthogonal, sum to the total error. A symmetry argument shows that this rectangular quantization error is minimized when each $N_j = N^{1/k}$. The total error is then $k \times E_1(N^{1/k})$.

Another set of coordinates with which to describe an input $\mathbf{x}$ of $\mathbf{R}^k$ is the magnitude $r$ and $k-1$ angles $\varphi_j$'s of the $k$-dimensional spherical coordinates system [12]. The following transformations produce these coordinates:

$$r = \left[\sum_{j=1}^{k} x_j^2\right]^{1/2} \quad ; \quad \varphi_j = \tan^{-1}\left[\frac{x_{j+1}}{\left[\sum_{i=1}^{j} x_i^2\right]^{1/2}}\right] \quad , j = 1,2,\ldots k-1$$

The reverse transformations are

$$x_1 = r \cos\varphi_{k-1} \cos\varphi_{k-2} \cdots \cos\varphi_2 \cos\varphi_1$$

$$x_2 = r \cos\varphi_{k-1} \cos\varphi_{k-2} \cdots \cos\varphi_2 \sin\varphi_1$$

$$\vdots$$

$$x_j = r \cos\varphi_{k-1} \cos\varphi_{k-2} \cdots \cos\varphi_j \sin\varphi_{j-1}$$

$$\vdots$$

$$x_k = r \sin\varphi_{k-1}$$

The $k=2$ case has already received much theoretical attention and is a special case of this analysis.

A change of variables produces the source density in spherical coordinates

$$f(\mathbf{x}) \rightarrow f(r,\underline{\varphi}) = f_k(r) \prod_{j=1}^{k-1} f_j(\varphi_j)$$

where

$$f_k(r) = \frac{2\pi^{k/2}}{\Gamma(k/2)} r^{k-1} \gamma p(r) \quad ; \quad r \in [0,\infty)$$

is the magnitude density with $p(\cdot)$ as defined above. The $k-1$ angle densities are

$$f_1(\varphi_1) = \frac{1}{2\pi} \quad ; \quad \varphi_1 \in [0,2\pi)$$

and

$$f_j(\varphi_j) = \frac{\Gamma[(j+1)/2]}{\Gamma(1/2)\Gamma(j/2)} \cos^{j-1}\varphi_j \quad ; \quad \varphi_j \in [-\pi/2,\pi/2] \ , \ j = 2,3,\ldots k-1$$

The resulting spherical coordinates are statistically independent.

In the spherical coordinates representation, employing separate scalar quantizers $(Q_r, Q_1, Q_2, \ldots Q_{k-1})$ defines the typical quantization region as the intersection of a non-zero width spherical shell centered at zero with a pyramid of apex zero (see Fig. 1 for $k=2$ and 3 examples). The spherical coordinates MSE expression is not as simple as that of the

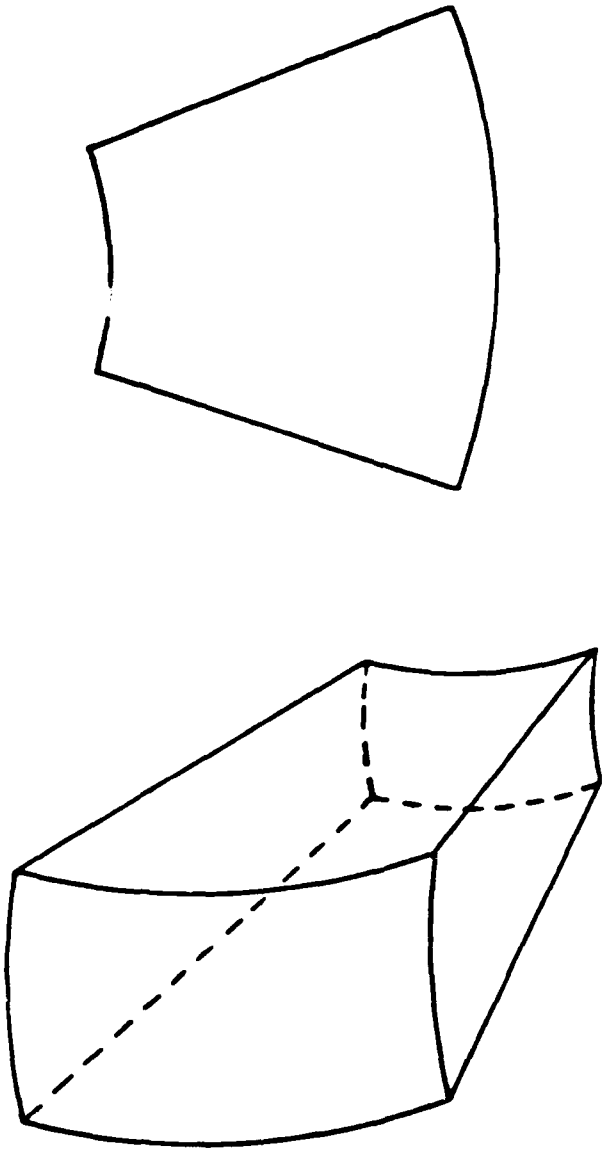


Fig. 1 - Examples of $k=2$ and $k=3$ spherical quantizer region shapes

rectangular quantizer and is derived below. Another problem of interest is the factorization of the number of levels to each quantizer ($\Lambda = N_r \times N_1 \times N_2 \cdots \times N_{k-1}$). In the bivariate case ($k=2$), the ratio of $N_\varphi$ to $N_r$ which minimizes the MSE has already been found [8].

For a quantizer with input $\mathbf{x}$ and output $\hat{\mathbf{x}}$ the MSE $D$ is

$$D = \int_{\mathbf{R}^k} |\mathbf{x}-\hat{\mathbf{x}}|^2 f(\mathbf{x})\, d\mathbf{x}$$

Transforming to spherical coordinates, letting hats indicate quantized values and employing the notation $c_i = \cos\varphi_i$ and $s_i = \sin\varphi_i$, the error transforms to

$$D = \int_{\mathbf{P}^k} [r^2 + \hat{r}^2 - 2r\hat{r}\,\beta(\underline{\varphi},\underline{\hat{\varphi}})]\, p(r,\underline{\varphi})\, dr\, d\underline{\varphi} \qquad (1)$$

where $p(r,\underline{\varphi})$ is the spherical coordinates density, $\mathbf{P}^k$ is the $k$-dimensional spherical coordinates space and

$$\beta(\underline{\varphi},\underline{\hat{\varphi}}) = s_{k-1}\hat{s}_{k-1} + c_{k-1}\hat{c}_{k-1}[s_{k-2}\hat{s}_{k-2} + c_{k-2}\hat{c}_{k-2}[s_{k-3}\hat{s}_{k-3} + \cdots [s_1\hat{s}_1 + c_1\hat{c}_1]\cdots]]$$

The above integral expression for $D$ can be simplified. The first two terms in the brackets are independent of $\underline{\varphi}$ and since $p(r,\underline{\varphi})$ factors, $\underline{\varphi}$ can be integrated out of these terms. The last term in the brackets is the product of an integral over $r$ and one over $\underline{\varphi}$. Eq.(1) becomes

$$D = \int_0^\infty r^2 f_k(r)\,dr + \int_0^\infty \hat{r}^2 f_k(r)\,dr - 2M_{k-1}\int_0^\infty r\hat{r} f_k(r)\,dr \qquad (2)$$

The term

$$M_{k-1} = \int_0^{2\pi}\int_{-\pi/2}^{\pi/2}\cdots\int_{-\pi/2}^{\pi/2} \beta(\underline{\varphi},\underline{\hat{\varphi}}) \prod_{i=1}^{k-1} f_i(\varphi_i)\, d\underline{\varphi}$$

is independent of $r$ and the magnitude quantizer; thus, Eq (2) can be minimized over the magnitude quantizer's parameters. The three

integrals present in that expression are all positive since the magnitude $r$ is always positive, hence, $D$ can also be minimized independently of $r$ by maximizing the value of $M_{k-1}$.

Working sequentially through the $\varphi_j$'s, the $M_{k-1}$ term can be written as

$$M_{k-1} = \int_{-\pi/2}^{\pi/2} \cdots \int_{-\pi/2}^{\pi/2} \left[ s_{k-1}\hat{s}_{k-1} + c_{k-1}\hat{c}_{k-1}[s_{k-2}\hat{s}_{k-2} + \cdots [s_{j+1}\hat{s}_{j+1}] \cdots ] \right] F_{j+1}(\underline{\varphi}) d\underline{\varphi}$$
$$+ M_j \int_{-\pi/2}^{\pi/2} \cdots \int_{-\pi/2}^{\pi/2} c_{k-1}\hat{c}_{k-1} \cdots c_{j+1}\hat{c}_{j+1} F_{j+1}(\underline{\varphi}) d\underline{\varphi} \qquad (3)$$

where $M_j$ is defined sequentially by ($M_0 = 1$)

$$M_j = \frac{\Gamma[(j+1)/2]}{\Gamma(1/2)\Gamma(j/2)} \int_{-\pi/2}^{\pi/2} (\sin\varphi_j \sin\hat{\varphi}_j + M_{j-1}\cos\varphi_j \cos\hat{\varphi}_j) \cos^{j-1}\varphi_j \, d\varphi_j \qquad (4)$$

and

$$F_{j+1}(\underline{\varphi}) d\underline{\varphi} = \prod_{i=j+1}^{k-1} f_i(\varphi_i) d\varphi_i$$

The integrals in Eq.(3) are independent of $\varphi_j$. The integrand of the second, being non-negative on $\mathop{\mathbf{X}}_{i=j+1}^{k-1} [-\pi/2, \pi/2]$, insures that the second integral is positive, hence, maximizing each $M_j(\varphi_j)$ term over the $\varphi_j$ quantizer sequentially maximizes $M_{k-1}$.

## QUANTIZER OPTIMIZATION

The quantizers designed in this paper are all scalar processors. Their specification requires the computation of the output values and the end-points of the quantization intervals. For a quantizer $Q_s(\cdot)$ with $N_s$ levels operating upon an input $s$, adopt the notation $\hat{s}_i$ as the $i$th output value and $[s_i, s_{i+1})$ as the $i$th interval, $i = 1, 2, \ldots N_s$, with $s_i$ as the $i$th breakpoint.

$i=1,2,\ldots N_s+1$. When the number of quantization levels is large, the compandor model of Bennett [13] for a nonuniform quantizer will be employed (see Fig. 2). Under this model, the quantizer $Q_s(\cdot)$ is a three-part system: an invertible, differentiable compressor nonlinearity $g(\cdot)$ mapping the range of the input to [0,1], a uniform quantizer $Q_U(\cdot)$ with $N_s$ levels on [0,1] and an expandor $h(\cdot)=g^{-1}(\cdot)$ mapping [0,1] back to the range of the input signal. For this model, specification of the compressor $g(\cdot)$ and the number of levels $N_s$ completely determines the quantizer.

The minimization of $D$ in Eq.(2) [maximization of each $M_j$ in Eq.(4)] can be accomplished in two ways depending upon the value of $N_s$. When $N_s$ is small, partial derivatives with respect to the quantizer's parameters will yield necessary conditions for the extremum similar to those found by Max. Positive [negative] definiteness of the matrix of second partial derivatives evaluated at the stationary point demonstrates the sufficiency of the necessary conditions. This may also be shown using a second derivative test similar to Fleischer's analysis [14]. These necessary conditions may be employed iteratively, as also suggested by Fleischer, to solve numerically for the optimal quantizer's parameters.

When $N_s$ is large, the output of a compandor system for an input $s$ can be approximated by

$$\hat{s} \approx s + \varepsilon h'[g(s)] = s + \frac{\varepsilon}{g'(s)}$$

where $\varepsilon$ is an independent noise source uniformly distributed on $[-\Delta/2,\Delta/2]$ ($\Delta = 1/N_s$ = the step size of the quantizer $Q_U$). After substituting for $\hat{s}$, the calculus of variations may be employed to yield the best compressor function. In both cases (maximizing $M_j$ or minimizing $D$),

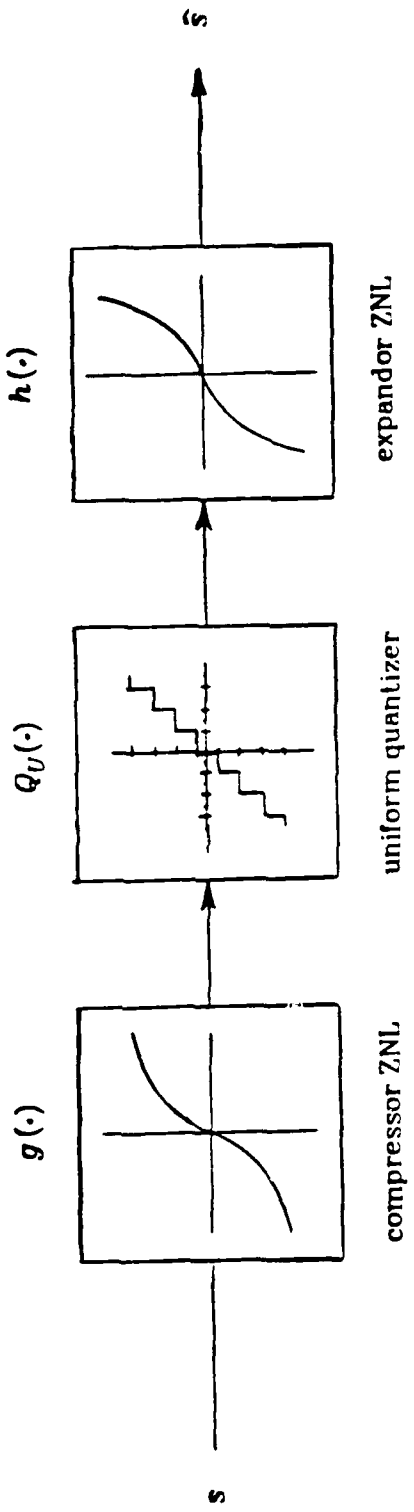



Fig. 2 - Compandor system model.

the functional is of the form

$$\int h[s,\hat{s}]\,ds$$

Employing the compandor approximation, this becomes

$$\int h[s,g'(s)]\,ds$$

The Euler-Lagrange differential equation [15]

$$\frac{\partial h}{\partial g} - \frac{d}{ds}\left[\frac{\partial h}{\partial g'}\right] = 0$$

applied to this problem yields the solution

$$\frac{\partial h}{\partial g'} = \text{constant}$$

This expression is solved for $g'$ which is then integrated to find the optimal compressor nonlinearity. The sign of the second variation exhibits the sufficiency of the compressor function solution.

MAGNITUDE QUANTIZER

For small $N_r$, taking partial derivatives of $D$ in Eq.(2) with respect to the magnitude quantizer's parameters yield:

$$r_i = \frac{1}{2M_{k-1}}(\hat{r}_{i-1} + \hat{r}_i) \quad ; \quad i = 2,3,\ldots N_r, \quad r_1 = 0, \; r_{N_r+1} = \infty \tag{5}$$

$$\hat{r}_i = \frac{M_{k-1}\int_{r_i}^{r_{i+1}} r f_k(r)\,dr}{\int_{r_i}^{r_{i+1}} f_k(r)\,dr} \quad ; \quad i = 1,2,\ldots N_r \tag{6}$$

These expressions, except for the $M_{k-1}$ terms, are equivalent to the equations defining the minimum MSE quantizer derived by Max. His optimal, $N_r$-level quantizer is defined by

$$t_i = \frac{1}{2}(\hat{t}_{i-1} + \hat{t}_i) \quad ; \quad i = 2,3,\ldots N_r, \quad t_1 = 0, \; t_{N_r+1} = \infty$$

$$\hat{t}_i = \frac{\int_{t_i}^{t_{i+1}} t f_k(t)dt}{\int_{t_i}^{t_{i+1}} f_k(t)dt} \quad ; \quad i = 1,2,\ldots N_r$$

From the Max quantizer, define a new quantizer with the same break-points and the outputs scaled by $M_{k-1}$

$$r_i = t_i \quad ; \quad \hat{r}_i = M_{k-1}\hat{t}_i$$

This new quantizer can be shown to satisfy the necessary conditions imposed on the magnitude quantizer by Eqs.(5) and (6). It is shown in Appendix A that $0 \leq M_{k-1} \leq 1$ and is usually approximately unity so that the scaling does not remove the output points from their respective regions. Employing the notation $E_k$ as the unscaled quantizer's MSE, the spherical distortion can be written as

$$D = M_{k-1}^2 \; E_k(N_r) + k \; (1 - M_{k-1}^2)$$

Since the optimal magnitude quantizer is a scaled version of the Max quantizer for the magnitude density, for large $N_r$ we employ the minimum MSE compressor for the magnitude density

$$g_r(r) = K_r \int_0^r f_k^{1/3}(t)\, dt \tag{7}$$

where $K_r$ is a constant such that $g_r$ maps to $[0,1]$. The actual compandor system has its expandor scaled by $M_{k-1}$. The same compressor function results if we directly apply the calculus of variations to the minimization of $D$. Examples of the magnitude compressor function for a multivariate Gaussian source appear in Fig. 3 (the magnitude random variable has a chi density).

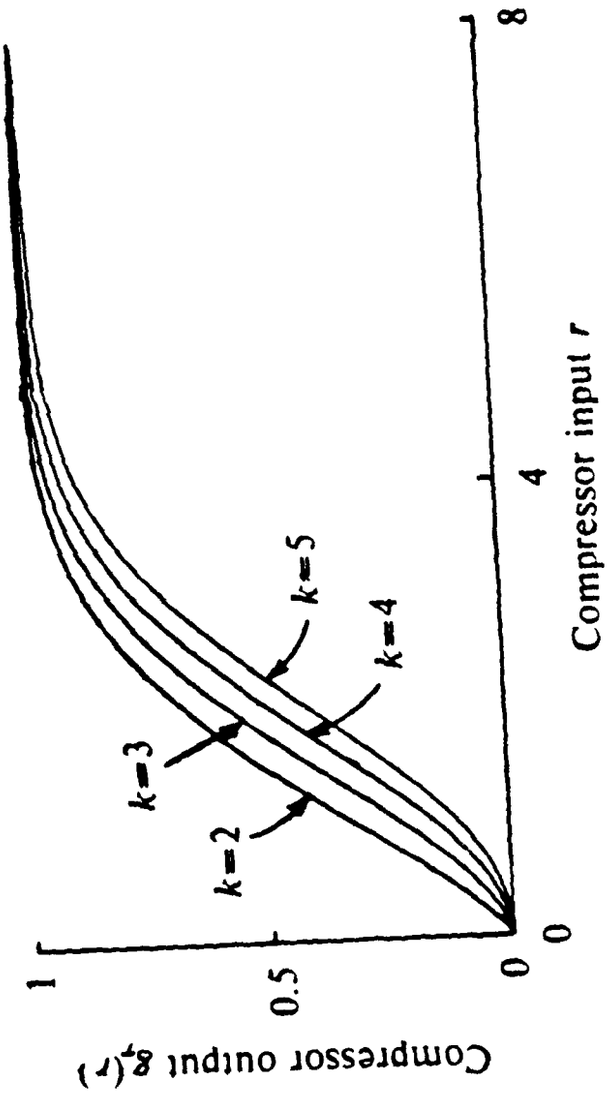



Fig. 3 - Magnitude compressors for the chi density.

MAXIMIZING $M_j$ OVER THE $\varphi_j$ QUANTIZER; $j=1,2,\ldots k-1$

When $N_j$, the number of levels in the $j$-th angle quantizer is small, partial derivatives of $M_j$ from Eq.(4) with respect to the angle quantizer $Q_j$'s parameters $\vartheta_i$ and $\hat{\vartheta}_i$ yield the following necessary conditions ($M_0=1$):

$$\vartheta_i = \tan^{-1}\left[\frac{M_{j-1}(\cos\hat{\vartheta}_i - \cos\hat{\vartheta}_{i-1})}{(\sin\hat{\vartheta}_{i-1} - \sin\hat{\vartheta}_i)}\right] \quad ; \quad i = 2,3,\ldots N_j \tag{8}$$

$$\hat{\vartheta}_i = \tan^{-1}\left[\frac{\int_{\vartheta_i}^{\vartheta_{i+1}} \cos^{j-1}\vartheta \sin\vartheta \, d\vartheta}{M_{j-1}\int_{\vartheta_i}^{\vartheta_{i+1}} \cos^{j}\vartheta \, d\vartheta}\right] \quad ; \quad i = 1,2,\ldots N_j \tag{9}$$

where $\vartheta_1$ and $\vartheta_{N_j+1}$ are the endpoints of the interval of definition of $\varphi_j$. For the first angle quantizer, these expressions yield a uniform quantizer:

$$\vartheta_i = 2\pi(i-1)/N_1 \quad ; \quad i = 1,2,\ldots N_1+1$$
$$\hat{\vartheta}_i = \pi(2i-1)/N_1 \quad ; \quad i = 1,2,\ldots N_1$$

and the resulting value of $M_1$ is

$$M_1 = \frac{\sin(\pi/N_1)}{(\pi/N_1)}$$

The other angle quantizers $Q_j(\,)$, $j\geq 2$, are nonuniform.

For large $N_j$ we cannot immediately decide to employ the minimum MSE compressor for the angle densities since we are trying to maximize $M_j$ in Eq.(4) for each $j$, not minimize the mean square error between $\varphi_j$ and $\hat{\varphi}_j$. Assuming that $M_{j-1}\approx 1$ in Eq.(4) for $M_j$ (since $N_{j-1}$ is also large, $M_{j-1}$ will be close to unity), the term in parenthesis simplifies to $\cos(\varphi_j-\hat{\varphi}_j)$. This term is expanded in a Taylor series about zero, since

for large $N_j$ the region widths will be small, to give

$$\cos(\varphi_j - \hat{\varphi}_j) \approx 1 - \frac{(\varphi_j - \hat{\varphi}_j)^2}{2}$$

Now applying the compandor approximation and the calculus of variations yields the compressor function for the $j$-th angle:

$$g_j(\varphi_j) = K_j \int_{-\pi/2}^{\varphi_j} \cos^{(j-1)/3}\vartheta \, d\vartheta \qquad (10)$$

where again $K_j$ is a constant so that $g_j$ maps to [0,1].

This resulting compressor is seen to be the minimum MSE compressor for the angle density and is proportional to an incomplete beta function [16]. For the $\varphi_1$ quantizer, the lower limit in Eq.(10) is zero, $g_1(\varphi_1)$ is linear and the quantizer is uniform. Fig. 4 presents the compressor functions for the second, third and fourth angle quantizers. To see if these approximations are reasonable for small $N_j$, we computed the parameters of the second angle quantizer. Fig. 5 depicts the graph of the compressor solution and the actual values satisfying the necessary conditions for $N_2$=16. We note that the values are very close.

The overall result is that to minimize the MSE of a $k$-dimensional spherical coordinates quantizer, a factorization of the total number of levels $N$ must be selected: $N = N_r \times N_1 \cdots \times N_{k-1}$. For small $N$, the $N_r$-level magnitude density quantizer is found by Eqs.(5) and (6), the $\varphi_1$ quantizer is uniform with $N_1$ levels and the $M_j$, $j=2,\ldots,k-1$, are maximized sequentially, each maximization in turn specifying the $\varphi_j$ quantizer $Q_j$ by Eqs.(8) and (9). When $N_j$ is large, the factorization is still made and the compressor functions are computed from Eqs.(7) and (10).

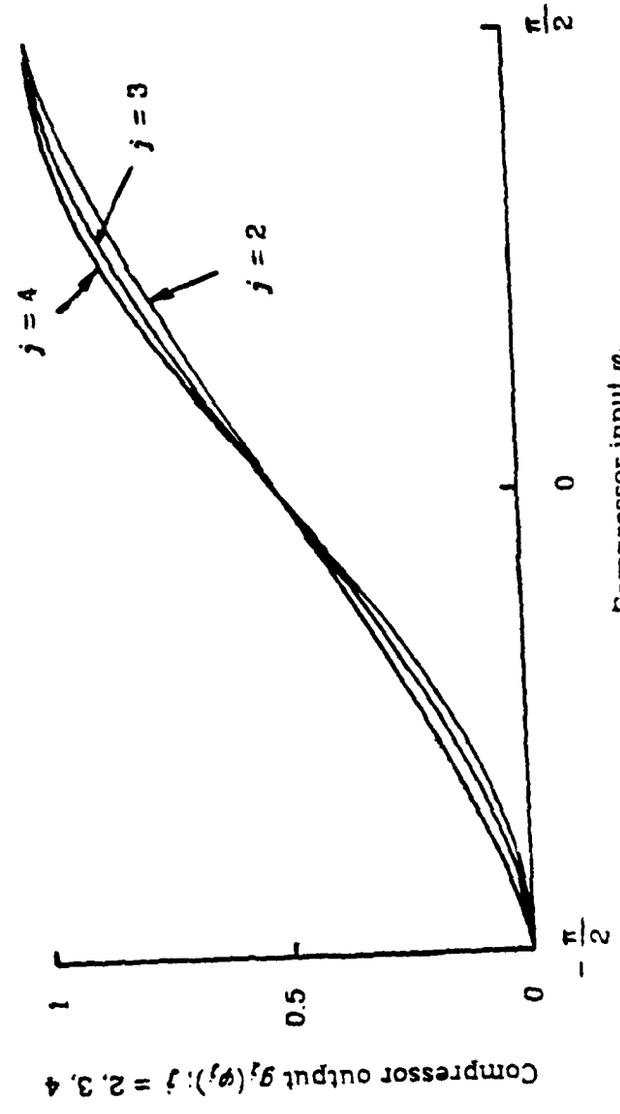



Fig. 4 - Compressors for the second, third and fourth angles.

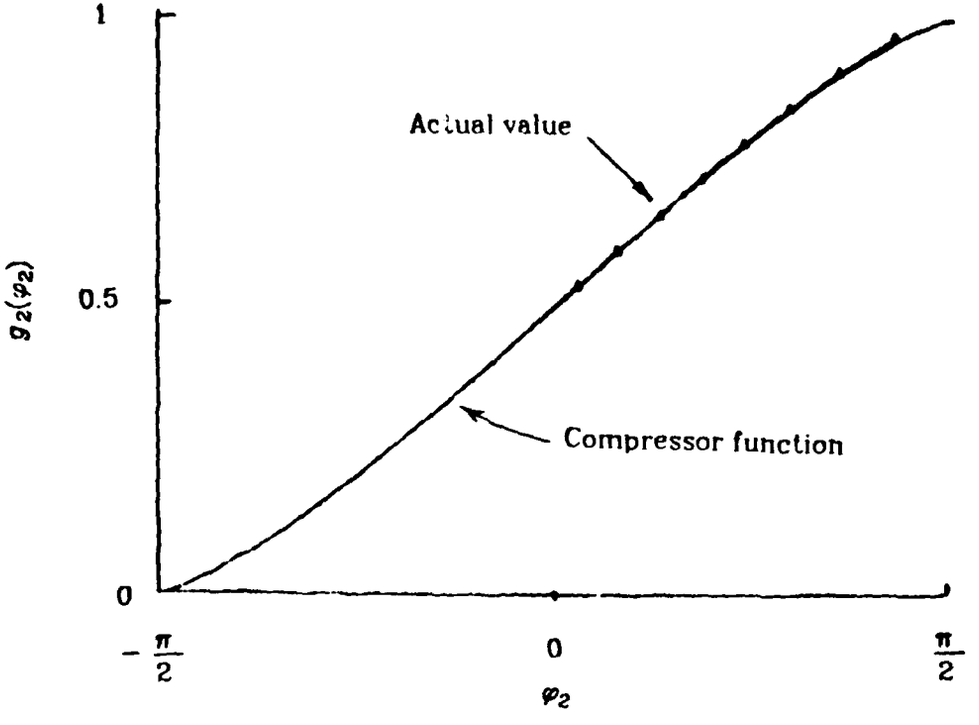



Fig. 5 - Comparison of compressor values and actual outputs for a $\varphi_2$ quantizer ($N_2$=16)

## ASYMPTOTIC RESULTS

This section considers asymptotic MSE rates and the solution to the problem of factoring the number of levels $N$ to each of the spherical coordinates quantizers. We assume that the number of levels in each quantizer is large so that the compandor approximation is appropriate. The levels are factored to each quantizer by

$$N = N_r \times N_1 \times N_2 \cdots \times N_{k-1} = N_r \times N_{\underline{\ell}}$$

where $N_{\underline{\ell}}$ is denoted as the product of the number of levels in all of the angle quantizers. Previously, we developed the expression

$$MSE = M_{k-1}^2(N_{\underline{\ell}}) \times E_k(N_r) \quad + \quad k\left[1 - M_{k-1}^2(N_{\underline{\ell}})\right] \tag{11}$$

hence, we require expressions for the magnitude error $E_k$ as a function of $N_r$ and for $M_{k-1}$ as a function of $N_{\underline{\ell}}$. Previous asymptotic results [17] yield the $E_k$ term

$$E_k(N_r) \approx \frac{1}{12\,N_r^2}\left\{\int_0^{\infty} f^{1/3}(r)\,dr\right\}^3 = \frac{E_r}{N_r^2}$$

where $f_k(r)$ is the magnitude density.

It is shown in Appendix C that $M_{k-1}$ is of the form

$$M_{k-1} \approx 1 - \frac{C_{k-1}}{N_{\underline{\ell}}^{2/(k-1)}}$$

where the $C_j$ are defined sequentially by

$$C_j = \frac{j^2\,C_{j-1}}{j^2-1}\,T_j$$

with $C_1 = \pi^2/6$ and

$$T_j = \left\{\frac{\pi\,\Gamma[(j+1)/2]\,\Gamma^3[(j+2)/6]}{24\,\Gamma(j/2)\,\Gamma^3[(j+5)/6]}\,\frac{(j^2-1)}{j\,C_{j-1}}\right\}^{1/j}$$

This result also yields the solution to the factorization of the number of levels in the angle quantizers:

$$N_j \approx \frac{T_j^{(j-1)/2}}{\prod_{i=j+1}^{k-1} T_i^{1/2}} N_{\underline{\ell}}^{1/(k-1)}$$

A second derivative test shows that this factorization of $N_{\underline{\ell}}$ maximizes $M_{k-1}$ for any spherically symmetric density.

From the value of $C_{k-1}$, Eq.(11) is minimized by

$$N_r \approx \left\{ \frac{E_r\ (k-1)}{2k\ C_{k-1}} \right\}^{(k-1)/2k} N^{1/k}$$

Remembering that $N_{\underline{\ell}} = N\ /\ N_r$, the minimum spherical MSE is

$$MSE \approx k \left\{ \frac{2k\ C_{k-1}}{k-1} \right\}^{(k-1)/k} E_r^{1/k}\ N^{-2/k} \qquad (12)$$

**EXAMPLES**

The compared quantization schemes are the rectangular coordinates quantizer ($k$ Max-type quantizers), the above described spherical coordinates quantizer and the optimal $k$-dimensional quantizer discussed by Zador. In order to compare error rates of schemes for different numbers of dimensions, we divide the MSE by $k$ yielding a MSE rate per dimension. Since all of the presented schemes have error rates proportional to $N^{-2/k}$, only the coefficient of the rate will be compared. Rectangular quantizers yield orthogonal errors making the coefficient a constant, independent of dimension. The spherical coordinates quantizer's MSE rate is found by evaluating Eq (12) with the appropriate $E_r$ term. The minimally achievable MSE is presented by the upper and lower bounds derived by Zador. This value is

$$MSE_{\min} \approx \frac{C(k,2)}{N^{2/k}} \left\{ \int_{\mathbf{R}^k} p^{k/(k+2)}(\mathbf{x})\, d\mathbf{x} \right\}^{(k+2)/k}$$

where $C(k,2)$ is a constant depending upon the optimal $k$-dimensional uniform quantizer. Zador provided bounds on this constant. More recently, Conway and Sloane found tighter upper limits on $C(k,2)$ for $k$ between 3 and 10. Note, however that the optimal scheme requires the implementation of the optimal $k$-dimensional uniform quantizer, usually a vector input device, while the rectangular and spherical schemes require only scalar processors.

The first source we consider is the independent Gaussian source with probability density function

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{k/2}} e^{-\frac{\mathbf{x}^T\mathbf{x}}{2}}$$

This source has standard Gaussian marginals; hence, the results will be comparable to others in the literature. For the three-dimensional case, the factorization is $N=N_r \times N_1 \times N_2$. Tables of $E_3(N_r)$ and $M_2(N_1,N_2)$ were generated. The best combinations for the $k=3$ spherical quantizers are listed in Table I along with their Signal-to-Noise Ratio (SNR) rates. The per dimension SNR is

$$SNR = 10 \log_{10} \frac{k}{MSE}\ dB$$

For comparison, the one-dimensional [1] and two-dimensional [6] results are also tabulated. The values of small $N$ considered correspond to the integers 8 through 18 cubed which allow easy comparison to the published one and two-dimensional results. For large $N$, several representative values were selected ($N=50^3$ and $100^3$). The values in parentheses are the actual factorizations employed and in all cases $N_r \times N_1 \times N_2 \leq N$.

| $N^{1/3}$ | N | Rect. SNR | Polar SNR | Three-dim SNR | Best dim |
|---|---|---|---|---|---|
| 8 | 512 | 14.62 | 14.58 | 14.52 (5x14x7) | 1 |
| 9 | 729 | 15.55 | 15.59 | 15.58 (5x16x9) | 2 |
| 10 | 1000 | 16.40 | 16.31 | 16.41 (6x18x9) | 3 |
| 11 | 1331 | 17.16 | 17.24 | 17.27 (7x19x10) | 3 |
| 12 | 1728 | 17.87 | 17.90 | 17.93 (7x22x11) | 3 |
| 13 | 2197 | 18.52 | 18.63 | 18.58 (9x22x11) | 2 |
| 14 | 2744 | 19.13 | 19.22 | 19.22 (8x26x13) | 3 |
| 15 | 3375 | 19.69 | 19.88 | 19.87 (9x25x15) | 2 |
| 16 | 4096 | 20.22 | 20.35 | 20.39 (10x27x15) | 3 |
| 17 | 4913 | 20.72 | 20.80 | 20.85 (10x30x16) | 3 |
| 18 | 5832 | 21.20 | 21.34 | 21.37 (11x33x16) | 3 |
| 50 | 125000 | 29.63 | 30.02 | 30.06 (32x83x47) | 3 |
| 100 | 1000000 | 35.65 | 36.05 | 36.08 (64x168x93) | 3 |

Table I - Comparison of SNR for $k = 1,2$ and 3 polar quantizers

This occurs since cubes of integers do not often have suitable factorizations or some of the results for smaller N are better. For this case, the best allocation of levels to each coordinate quantizer is

$$N_r \approx 0.654\, N^{1/3} \quad , \quad N_1 \approx 1.63\, N^{1/3} \quad , \quad N_2 \approx 0.937\, N^{1/3}$$

As $N_r \to \infty$, the minimum MSE quantizer for the chi density has error:

$$E_r = \frac{1}{12}\left|\int_0^\infty f_k^{1/3}(r)\, dr\right|^3 = \frac{3^{k/2}\,\Gamma^3[(k+2)/6]}{8\Gamma(k/2)}$$

When $k=3$, this reduces to $E_r = 1.054$ and the resulting asymptotic Gaussian source error rate is $7.401 \times N^{-2/3}$. This asymptotic result compares favorably to the above numerical results.

For other values of $k$, the spherical MSE rate is found through Eq.(12) with $E_r$ as above. Rectangular coordinates quantizers have a total error rate of

$$MSE_{rect} \approx \frac{k\pi\sqrt{3}}{2} N^{-2/k}$$

The optimal rate for Gaussian sources is

$$MSE_{min} \approx C(k,2)\, 2\pi \left[\frac{k+2}{k}\right]^{1+k/2} N^{-2/k}$$

Fig. 6 and Table II provide a comparison of the coefficient of the error rate for rectangular, spherical and optimal quantizers versus the number of dimensions. Notice that $k=3$ yields the best of the spherical error rates and that this value is only slightly below that of the polar quantizers.

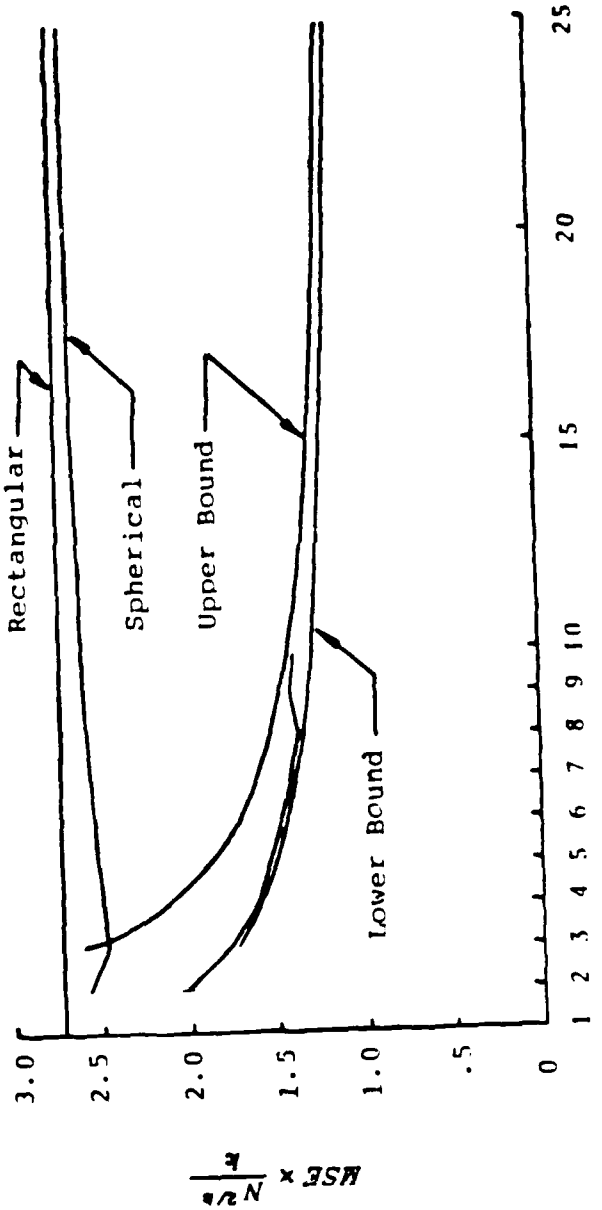



Fig. 6 - Asymptotic error rates for a Gaussian source.

| dim | Rect. MSE | Sph. MSE | Lower MSE | Optimum MSE | Upper MSE |
|---|---|---|---|---|---|
| 2 | 2.72 | 2.5665 | | 2.015 | |
| 3 | 2.72 | 2.467 | 1.734 | 1.769 | 2.609 |
| 4 | 2.72 | 2.48675 | 1.591 | 1.624 | 2.115 |
| 5 | 2.72 | 2.51 | 1.5 | 1.542 | 1.863 |
| 6 | 2.72 | 2.53 | 1.436 | 1.475 | 1.70833 |
| 7 | 2.72 | 2.54714 | 1.388 | 1.425 | 1.60571 |
| 8 | 2.72 | 2.56125 | 1.35125 | 1.375 | 1.53125 |
| 9 | 2.72 | 2.57444 | 1.32111 | 1.42111 | 1.47333 |
| 10 | 2.72 | 2.586 | 1.296 | 1.401 | 1.428 |
| 11 | 2.72 | 2.59545 | 1.27636 | | 1.39182 |
| 12 | 2.72 | 2.60333 | 1.25833 | | 1.36167 |
| 13 | 2.72 | 2.61077 | 1.24308 | | 1.33615 |
| 14 | 2.72 | 2.61714 | 1.23 | | 1.315 |
| 15 | 2.72 | 2.62267 | 1.218 | | 1.296 |
| 16 | 2.72 | 2.62812 | 1.20687 | | 1.27875 |
| 17 | 2.72 | 2.63294 | 1.19824 | | 1.26412 |
| 18 | 2.72 | 2.63722 | 1.18889 | | 1.25111 |
| 19 | 2.72 | 2.64105 | 1.18211 | | 1.24 |
| 20 | 2.72 | 2.6445 | 1.175 | | 1.229 |
| 21 | 2.72 | 2.64762 | 1.1681 | | 1.22 |
| 22 | 2.72 | 2.65045 | 1.16182 | | 1.21091 |
| 23 | 2.72 | 2.65304 | 1.15696 | | 1.20304 |
| 24 | 2.72 | 2.65583 | 1.15208 | | 1.19583 |
| 25 | 2.72 | 2.658 | 1.1472 | | 1.1892 |

Table II - Asymptotic results for a Gaussian source

Another spherically symmetric source is the Pearson Type II source with pdf

$$f(\mathbf{x}) = \frac{\Gamma(\nu+1)\,[\,2(\nu+1) - \mathbf{x}^T\mathbf{x}\,]^{\nu-k/2}}{\pi^{k/2}\,2^{\nu}\,(\nu+1)^{\nu}\,\Gamma(\nu+1-k/2)} \quad ; \quad \mathbf{x}^T\mathbf{x} \le 2(\nu+1)$$

This source has finite range and a Pearson II marginal density with parameter $\nu$ ($\nu>0$). Results for this source are not presented because the two dimensional case performed best in all of the examples attempted (polar results can be found in [8]).

Another source with infinite range is the Pearson Type VII source with Pearson VII marginals ($\nu>1$):

$$f(\mathbf{x}) = \frac{2^{\nu}\,(\nu-1)^{\nu}\,\Gamma(\nu+k/2)}{\pi^{k/2}\,\Gamma(\nu)\,[\,2(\nu-1) + \mathbf{x}^T\mathbf{x}\,]^{\nu+k/2}}$$

For this density, the rectangular error (for a bank of Max quantizers) is

$$MSE_{rect} \approx \frac{k(\nu-1)\,B^3[1/2,(\nu-1)/2]}{6\,B[1/2,\nu]}\,N^{-2/k}$$

where $B(\,,\,)$ is the Beta function [16]. For this source, the magnitude error term needed for the spherical coordinates quantizer MSE can be found to be

$$E_r = \frac{(\nu-1)\,B^3[(k+2)/2,(\nu-1)/2]}{24\,B[k/2,\nu]}$$

The spherical MSE rate is found from Eq.(12) and this term. The minimal MSE rate can be computed to be

$$MSE_{\min} \approx C(k,2)\,2\pi(\nu-1)\,\frac{\Gamma(\nu+k/2)}{\Gamma(\nu)}\left[\frac{\Gamma\left[\frac{k(\nu-1)}{k+2}\right]}{\Gamma\left[\frac{k(k+2\nu)}{2(k+2)}\right]}\right]^{(k+2)/k} N^{-2/k}$$

These Pearson sources have restrictions on the value of the parameter $\nu$ in order to assure unit power marginals [18]. Plots and tables of the

coefficient of the MSE rate versus dimension for the Pearson VII source with various values of the parameter $\nu$ are presented in Figs. 7 through 9 and Tables III through V. The greatest performance gains by the spherical quantizers were for those Pearson VII sources which are furthest from the Gaussian ($\nu$ approaches unity). These sources are more peaked at the origin and heavier tailed. Figure 10 compares the marginal densities of the considered sources to illustrate this point. Note that as $\nu \to \infty$, the Pearson VII source approaches the Gaussian source.

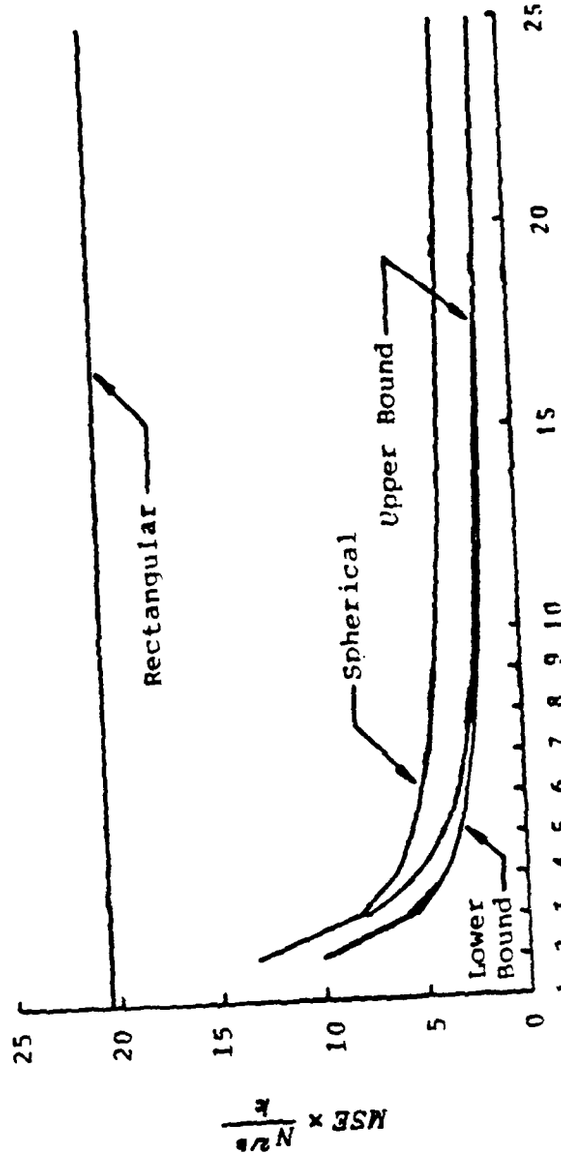



Fig. 7 - Asymptotic error rates for a Pearson VII source, $\nu = 1.25$.

| dim | Rect. MSE | Sph. MSE | Lower MSE | Optimum MSE | Upper MSE |
|---|---|---|---|---|---|
| 2 | 20.432 | 13.259 | | 10.078 | |
| 3 | 20.432 | 8.205 | 5.3786 | 5.4857 | 8.0925 |
| 4 | 20.432 | 6.4407 | 3.8657 | 3.9468 | 5.1389 |
| 5 | 20.432 | 5.5567 | 3.1357 | 3.2246 | 3.8951 |
| 6 | 20.432 | 5.0265 | 2.7089 | 2.7838 | 3.2253 |
| 7 | 20.432 | 4.6723 | 2.4293 | 2.4941 | 2.8102 |
| 8 | 20.432 | 4.4184 | 2.2318 | 2.2713 | 2.5287 |
| 9 | 20.432 | 4.227 | 2.0848 | 2.2423 | 2.3253 |
| 10 | 20.432 | 4.0772 | 1.9709 | 2.1305 | 2.1716 |
| 11 | 20.432 | 3.9565 | 1.88 | | 2.0512 |
| 12 | 20.432 | 3.8572 | 1.8058 | | 1.9544 |
| 13 | 20.432 | 3.7738 | 1.7438 | | 1.8748 |
| 14 | 20.432 | 3.7027 | 1.6914 | | 1.8082 |
| 15 | 20.432 | 3.6414 | 1.6463 | | 1.7516 |
| 16 | 20.432 | 3.5878 | 1.6072 | | 1.7028 |
| 17 | 20.432 | 3.5407 | 1.5729 | | 1.6604 |
| 18 | 20.432 | 3.4988 | 1.5426 | | 1.6231 |
| 19 | 20.432 | 3.4614 | 1.5155 | | 1.59 |
| 20 | 20.432 | 3.4276 | 1.4912 | | 1.5606 |
| 21 | 20.432 | 3.3971 | 1.4693 | | 1.5341 |
| 22 | 20.432 | 3.3693 | 1.4494 | | 1.5102 |
| 23 | 20.432 | 3.3439 | 1.4313 | | 1.4884 |
| 24 | 20.432 | 3.3205 | 1.4147 | | 1.4686 |
| 25 | 20.432 | 3.299 | 1.3994 | | 1.4505 |

Table III - Asymptotic results for a Pearson VII source, $\nu = 1.25$

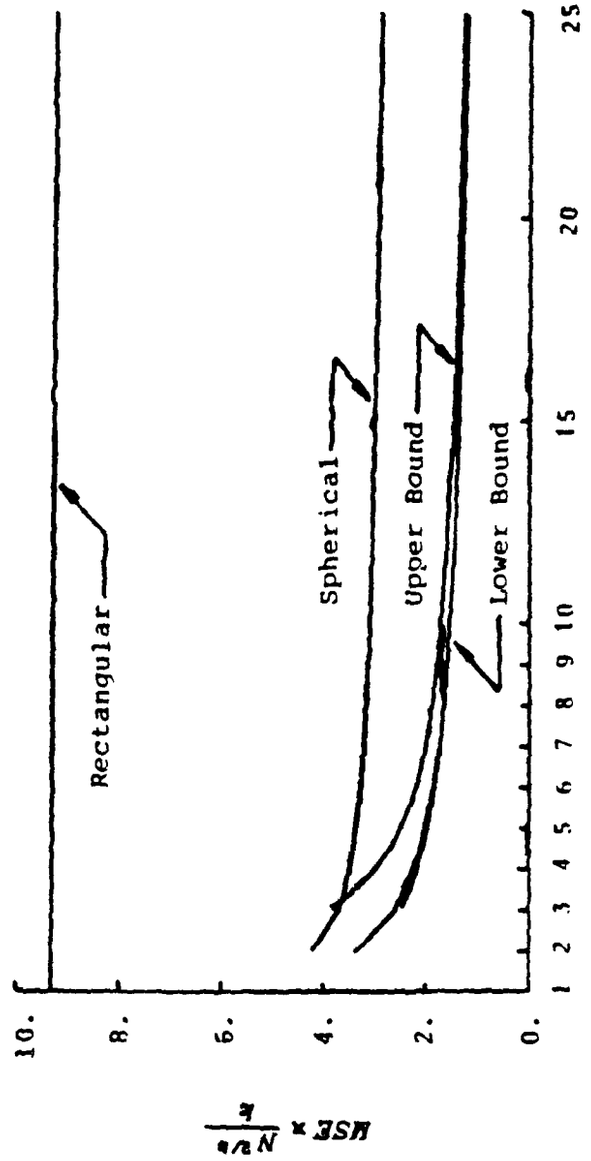



Fig 8 - Asymptotic error rates for a Pearson VII source, $\nu$=2.5.

| dim | Rect. MSE | Sph. MSE | Lower MSE | Optimum MSE | Upper MSE |
|---|---|---|---|---|---|
| 2 | 9.3044 | 4.2193 | | 3.3594 | |
| 3 | 9.3044 | 3.7081 | 2.5478 | 2.5986 | 3.8334 |
| 4 | 9.3044 | 3.4921 | 2.1875 | 2.2334 | 2.9079 |
| 5 | 9.3044 | 3.3719 | 1.977 | 2.0330 | 2.4557 |
| 6 | 9.3044 | 3.2941 | 1.8372 | 1.8880 | 2.1874 |
| 7 | 9.3044 | 3.2385 | 1.7367 | 1.7831 | 2.0091 |
| 8 | 9.3044 | 3.1962 | 1.6606 | 1.6899 | 1.8814 |
| 9 | 9.3044 | 3.1626 | 1.6005 | 1.7215 | 1.7852 |
| 10 | 9.3044 | 3.1349 | 1.5518 | 1.6775 | 1.7098 |
| 11 | 9.3044 | 3.1116 | 1.5114 | | 1.649 |
| 12 | 9.3044 | 3.0915 | 1.4772 | | 1.5989 |
| 13 | 9.3044 | 3.074 | 1.4479 | | 1.5567 |
| 14 | 9.3044 | 3.0585 | 1.4224 | | 1.5207 |
| 15 | 9.3044 | 3.0446 | 1.4001 | | 1.4895 |
| 16 | 9.3044 | 3.0321 | 1.3802 | | 1.4623 |
| 17 | 9.3044 | 3.0208 | 1.3625 | | 1.4383 |
| 18 | 9.3044 | 3.0105 | 1.3466 | | 1.4169 |
| 19 | 9.3044 | 3.001 | 1.3322 | | 1.3977 |
| 20 | 9.3044 | 2.9923 | 1.3191 | | 1.3804 |
| 21 | 9.3044 | 2.9842 | 1.3071 | | 1.3647 |
| 22 | 9.3044 | 2.9767 | 1.296 | | 1.3504 |
| 23 | 9.3044 | 2.9697 | 1.2859 | | 1.3372 |
| 24 | 9.3044 | 2.9631 | 1.2765 | | 1.3252 |
| 25 | 9.3044 | 2.9569 | 1.2678 | | 1.3141 |

Table IV - Asymptotic results for a Pearson VII source, $\nu = 2.5$

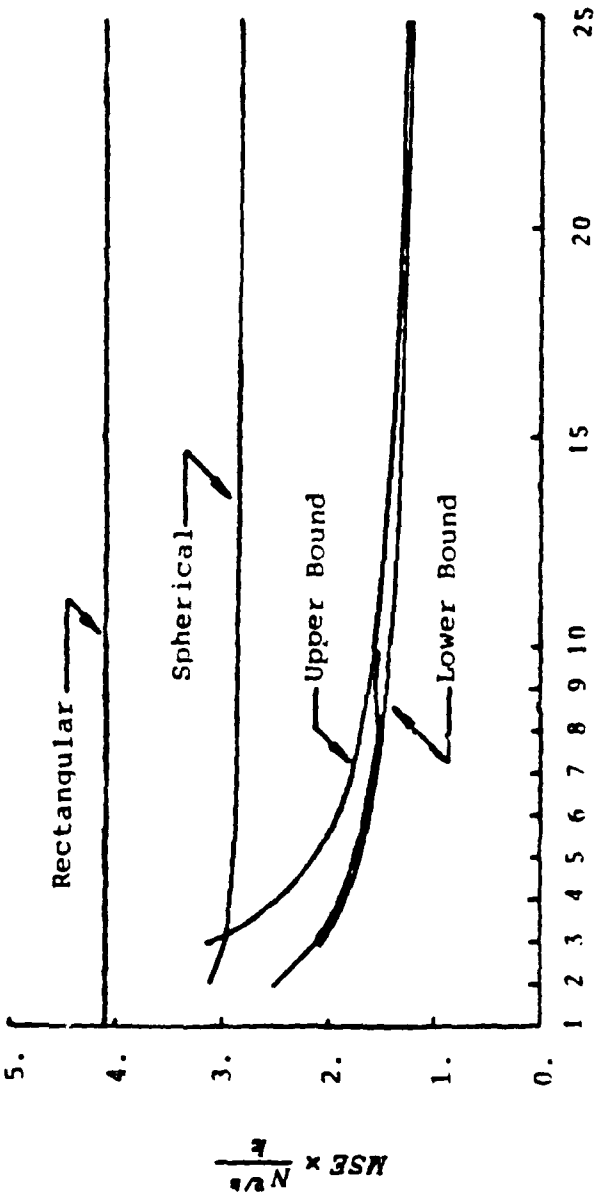



Fig. 9 - Asymptotic error rates for a Pearson VII source, $\nu$=5.

| dim | Rect. MSE | Sph. MSE | Lower MSE | Optimum MSE | Upper MSE |
|---|---|---|---|---|---|
| 2 | 4.1038 | 3.1136 | | 2.5196 | |
| 3 | 4.1038 | 2.9606 | 2.0677 | 2.1089 | 3.111 |
| 4 | 4.1038 | 2.9099 | 1.8502 | 1.889 | 2.4595 |
| 5 | 4.1038 | 2.8884 | 1.7164 | 1.7651 | 2.1321 |
| 6 | 4.1038 | 2.8776 | 1.6246 | 1.6695 | 1.9343 |
| 7 | 4.1038 | 2.8714 | 1.5569 | 1.5985 | 1.8011 |
| 8 | 4.1038 | 2.8672 | 1.5047 | 1.5314 | 1.7049 |
| 9 | 4.1038 | 2.864 | 1.463 | 1.5735 | 1.6318 |
| 10 | 4.1038 | 2.8614 | 1.4287 | 1.5444 | 1.5741 |
| 11 | 4.1038 | 2.859 | 1.3999 | | 1.5274 |
| 12 | 4.1038 | 2.8568 | 1.3754 | | 1.4886 |
| 13 | 4.1038 | 2.8548 | 1.3542 | | 1.4559 |
| 14 | 4.1038 | 2.8528 | 1.3356 | | 1.4279 |
| 15 | 4.1038 | 2.8509 | 1.3192 | | 1.4035 |
| 16 | 4.1038 | 2.849 | 1.3046 | | 1.3821 |
| 17 | 4.1038 | 2.8471 | 1.2914 | | 1.3632 |
| 18 | 4.1038 | 2.8453 | 1.2796 | | 1.3463 |
| 19 | 4.1038 | 2.8436 | 1.2688 | | 1.3311 |
| 20 | 4.1038 | 2.8419 | 1.2589 | | 1.3174 |
| 21 | 4.1038 | 2.8402 | 1.2498 | | 1.3049 |
| 22 | 4.1038 | 2.8386 | 1.2415 | | 1.2935 |
| 23 | 4.1038 | 2.837 | 1.2338 | | 1.283 |
| 24 | 4.1038 | 2.8354 | 1.2266 | | 1.2734 |
| 25 | 4.1038 | 2.8339 | 1.2199 | | 1.2645 |

Table V - Asymptotic results for a Pearson VII source, $\nu$=5.

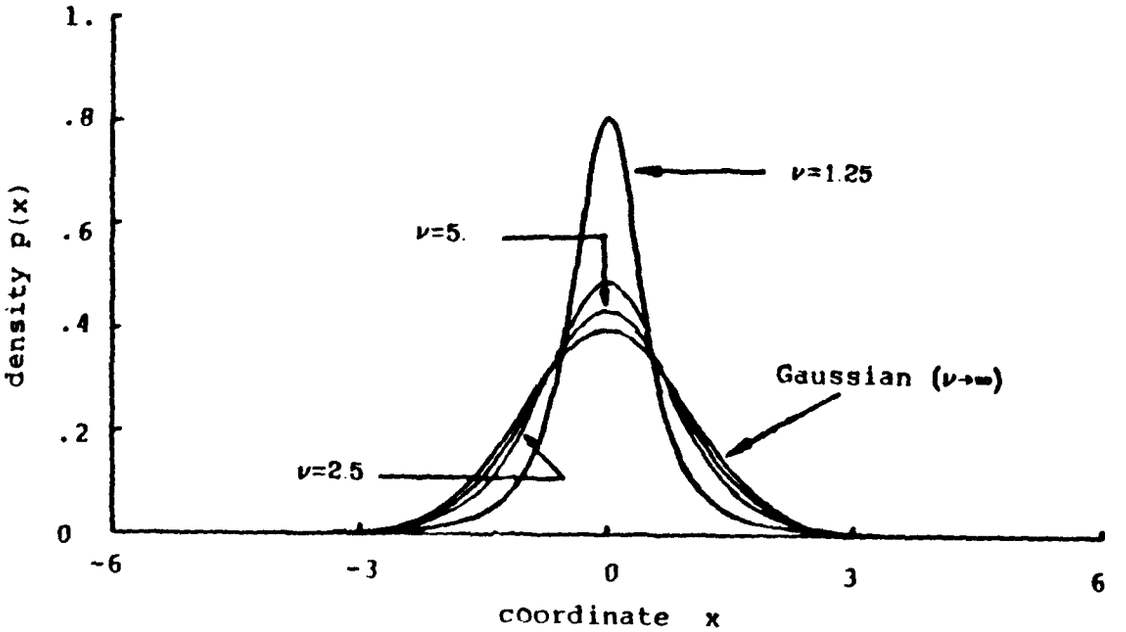



Fig. 10 - Marginal densities of the Gaussian and Pearson VII sources.

## CONCLUSIONS

This chapter presents the generalization of polar quantizers to greater than two dimensions for all spherically symmetric densities. In comparison, the spherical scheme is applicable to any number of dimensions, $k$, and has a scalar processor implementation while the optimal quantizers are available only when the $k$-dimensional uniform quantizer can be implemented. The derived performance expressions may be used to decide if spherical schemes are of value in the particular application.

The MSE rates for $k \geq 4$ presented were for large values of $N$ only. Research in one-dimensional compandor approximations suggest that these error rates are also valid for reasonable data rates (i.e. greater that 5 bits per dimension) and this was observed in the examples of the three dimensional Gaussian quantizers computed for $N \varepsilon [8^3, 18^3]$. It was also seen for the Gaussian source, as was noted in the published polar results, that rectangular quantizers perform better than spherical quantizers for the small values of $N$ while as $N$ increases, the asymptotic rates became valid and the three-dimensional quantizers performed best. Of course, the optimal rate is always lower.

The results presented for the multidimensional spherical quantizers show that spherical coordinates encoding of spherically symmetric sources is often more efficient in a MSE sense than one-dimensional rectangular coordinates quantizing. For the Pearson VII source with $\nu = 1.25$, polar quantizing ($k=2$) has a gain of 1.9 dB in Signal to Noise Ratio (SNR) over rectangular quantization while spherical quantization in six dimensions showed an increase of 6 dB. The optimal rate for $k=6$ is approxi-

mately 9 dB over the rectangular rate. An intuitive explanation for the spherical coordinate superiority over rectangular schemes is that they preserve the spherical symmetry inherent in the considered multidimensional densities.

A straightforward and perhaps important extension of this work is the design of $k$-dimensional spherical quantizers with uniform step-size. The solution set is then only the factorization of $N$ and the magnitude quantizer step-size ($k$ parameters). This extension is important because of the simplicity of its implementation.

## APPENDIX A - BOUNDING $M_{k-1}$

The term $M_j$, defined in Eq.(4), is bounded as follows:

1- Given that $M_{j-1}\varepsilon[0,1]$, then $M_j \leq 1$.

Eq.(4) expresses $M_j$ as the expectation over $\varphi_j \varepsilon[-\pi/2, \pi/2]$ of the function

$$\Psi(\varphi_j) = \sin\varphi_j \sin\hat{\varphi}_j + M_{j-1}\cos\varphi_j \cos\hat{\varphi}_j$$

where $\hat{\varphi}_j$ is the quantized value of $\varphi_j$. Hence, upper bounding $\Psi(\varphi_j)$ by unity also bounds $M_j$ by unity. The maximum of $\Psi(\varphi_j)$ is $\Psi_{max}=\sin\hat{\varphi}_j / \sin\varphi_j$ which is attained at the point $\tan\hat{\varphi}_j = M_{j-1}\tan\varphi_j$. Since $M_{j-1}\varepsilon[0,1]$:

$$\tan\hat{\varphi}_j \leq \tan\varphi_j \rightarrow |\hat{\varphi}_j| \leq |\varphi_j| \rightarrow |\sin\hat{\varphi}_j| \leq |\sin\varphi_j| \rightarrow \Psi_{max} \leq 1 \rightarrow M_j \leq 1$$

2- Given that $M_{j-1}\varepsilon[0,1]$ and the $\varphi_j$ quantizer satisfies Eqs.(8) and (9), then $M_j \geq 0$.

The form of $M_j$ in Eq.(4) involves cosine integrals and sine-cosine integrals. Rearranging Eq.(9) allows substitution and expansion of the sine-cosine integrals over $\varphi$ in Eq.(4). The result is

$$M_j = \frac{\Gamma[(j+1)/2]}{\Gamma(1/2)\Gamma(j/2)} M_{j-1} \sum_{i=1}^{N_j} \int_{\vartheta_i}^{\vartheta_{i+1}} (\sin^2\hat{\vartheta}_i \sec\hat{\vartheta}_i + \cos\hat{\vartheta}_i \cos\varphi_j)\cos^{j-1}\varphi_j \, d\varphi_j$$

Since all of the terms present are non-negative over the range of $\varphi_j$, $M_j \geq 0$.

3- $0 \leq M_j \leq 1$, $j = 2,3,\ldots k-1$.

Since $M_1=\sin(\pi/N_1)\div(\pi/N_1)$ and $N_1 \geq 1$, then $M_1\varepsilon[0,1]$. This fact, combined with 1 and 2 inductively, shows that $0 \leq M_j \leq 1$ $j=2,3,\ldots k-1$.

## APPENDIX B - SUFFICIENCY OF CONDITIONS

In the text of this chapter, necessary conditions on the output points (the $\hat{s}_i$) and the breakpoints (the $s_i$) of the magnitude and angle quantizers are given in Eqs.(5) and (6) and Eqs.(8) and (9) respectively. The purpose of this appendix is to demonstrate the sufficiency of these conditions. The included analysis closely follows that of Fleischer [14] and will draw several results from his paper.

For the magnitude quantizer, we desire to minimize $D$ in Eq.(2) and for the $j$-th angle quantizer, we seek to minimize $-M_j$ from Eq.(4) (equivalent to maximizing $M_j$). The following will be a parallel developement of sufficient conditions for either type (magnitude or angle) quantizer. Sufficiency is shown by determining that the matrix of second partial derivatives, evaluated at the stationary point, is positive definite.

From Eqs.(2) and (4), the functionals are both functions of $2N_s$ variables. To reduce this number, assume that the $\hat{s}_i$ are preassigned and optimize over the $s_i$. Taking derivatives yields

$$\frac{\partial D}{\partial r_i} = f_k(r_i)\,(\hat{r}_{i-1} - \hat{r}_i)\,(\hat{r}_{i-1} + \hat{r}_i - 2M_{k-1}r_i)$$

and

$$\frac{\partial(-M_j)}{\partial \vartheta_i} = K_j \cos^{j-1}\vartheta_i \left[\sin\vartheta_i\,(\sin\hat{\vartheta}_i - \sin\hat{\vartheta}_{i-1}) + M_{j-1}\cos\vartheta_i\,(\cos\hat{\vartheta}_i - \cos\hat{\vartheta}_{i-1})\right]$$

where $K_j$ is the constant term from Eq.(4). Equating these expressions to zero yields the necessary conditions in Eqs.(5) and (8) respectively. In both cases, the matrix of second partial derivatives with respect to the $s_i$, evaluated at the stationary point determined be Eqs.(5) and (8), is a

diagonal matrix with elements

$$\frac{\partial^2 D}{(\partial r_i)^2} = 2M_{k-1} f_k(r_i) ( \hat{r}_i - \hat{r}_{i-1} )$$

and

$$\frac{\partial^2(-M_j)}{(\partial \vartheta_i)^2} = K_j \cos^{j-2}\vartheta_i ( \sin\hat{\vartheta}_i - \sin\hat{\vartheta}_{i-1} )$$

Since $M_{k-1}$ is positive and the $s_i$ are increasing in $i$, the above matrices are both easily seen to be positive definite. The result is that for fixed $\hat{s}_i$, Eqs (5) and (8) are necessary and sufficient to minimize $D$ and $-M_j$ respectively.

Now assume that for the quantizers, the breakpoints are assigned as above. The functional is now dependent only on $N_s$ variables, the $\hat{s}_i$. Taking derivatives and using the conditions of Eqs (5) and (8) yield

$$\frac{\partial D}{\partial \hat{r}_i} = 2\int_{r_i}^{r_{i+1}} ( \hat{r}_i - M_{k-1}r ) f_k(r)\, dr$$

and

$$\frac{\partial(-M_j)}{\partial \hat{\vartheta}_i} = \int_{\vartheta_i}^{\vartheta_{i+1}} ( M_{j-1}\cos\vartheta \sin\hat{\vartheta}_i - \sin\vartheta \cos\hat{\vartheta}_i ) K_j \cos^{j-1}\vartheta\, d\vartheta$$

Again, equating to zero yields the necessary conditions in Eqs (6) and (9). After careful algebra and utilization of the conditions in Eqs (5), (6), (8) and (9), the second partial derivative matrix with respect to the $\hat{s}_i$ can be shown to be of the form

$$\begin{bmatrix} 2a_1-b_1 & -b_1 & 0 & 0 & 0 \\ -b_1 & 2a_2-b_1-b_2 & -b_2 & 0 & \\ 0 & -b_2 & 2a_3-b_2-b_3 & -b_3 & \\ 0 & 0 & -b_3 & & \\ & & & & -b_{N_s-1} \\ 0 & & & -b_{N_s-1} & 2a_{N_s}-b_{N_s-1} \end{bmatrix}$$

where for the $D$ minimization

$$a_i = \int_{r_i}^{r_{i+1}} f_k(r)\,dr \quad \text{and} \quad b_i = \frac{f_k(r)}{2M}(\hat{r}_i - \hat{r}_{i-1})$$

and for the $-M_j$ case

$$a_i = \frac{K_j}{2}\int_{\vartheta_i}^{\vartheta_{i+1}} (\sin\vartheta \sin\hat{\vartheta}_i + M_{j-1}\cos\vartheta\cos\hat{\vartheta}_i)\cos^{j-1}\vartheta\,d\vartheta$$

and

$$b_i = K_j M_{j-1}^2 \cos^{j+2}\vartheta_i \frac{(1 - \sin\hat{\vartheta}_i \sin\hat{\vartheta}_{i-1} - \cos\hat{\vartheta}_i \cos\hat{\vartheta}_{i-1})^2}{(\sin\hat{\vartheta}_i - \sin\hat{\vartheta}_{i-1})^3}$$

Note that both sets of $b_i$ are positive.

For a matrix $\mathbf{M}$ of the above form, a quadratic form can be expanded as

$$\mathbf{x}^T\mathbf{M}\mathbf{x} = \sum_{i=1}^{N_s} x_i^2 (2a_i - 2b_i - 2b_{i-1}) + \sum_{i=1}^{N_s-1} b_i (x_{i+1} - x_i)^2$$

where $b_0=0$. Since the $b_i$ are all positive, a sufficient condition for the quadratic form to be positive and the matrix to be positive definite is that

$$\sigma_i = 2a_i - 2b_i - 2b_{i-1} \geq 0$$

For the minimization of $D$, directly following Fleischer's arguments from this point yields the same sufficient condition

$$\frac{\partial^2}{\partial r^2}\log f_k(r) < 0$$

This condition holds for all of the densities considered in the examples. As to the maximization of $M_j$, no argument similar to Fleischer's is apparent due to the complexity of the $a_i$ and $b_i$; hence, a numerical evaluation of the $\sigma_i$ for the resulting quantizer is appropriate.

## APPENDIX C - ASYMPTOTIC DERIVATIONS

In our previous discussion, we saw that the $M_j$ terms were defined sequentially by

$$M_1 = \frac{\sin(\pi/N_1)}{(\pi/N_1)}$$

and for $2 \le j \le k-1$ by

$$M_j = \frac{\Gamma\left[\frac{j+1}{2}\right]}{\Gamma\left[\frac{1}{2}\right]\Gamma\left[\frac{j}{2}\right]} \int_{-\pi/2}^{\pi/2} \left( \sin\varphi \sin\hat{\varphi} + M_{j-1}\cos\varphi\cos\hat{\varphi} \right) \cos^{j-1}\varphi \, d\varphi \qquad \text{(A1)}$$

When the number of levels is each quantizer, $N_j$, is large, the compandor model approach is appropriate. Using this model, the quantized output is approximately equal to the input plus a random error

$$\hat{\varphi} \approx \varphi + \varepsilon \frac{\Gamma(1/2)\,\Gamma[(j+2)/6]}{\Gamma[(j+5)/6]\cos^{(j-1)/3}\varphi}$$

where $\varepsilon$ is uniformly distributed on $[-1/2N_j, 1/2N_j]$. Substituting in for $\hat{\varphi}$ in Eq.(A1), applying the trigonometric identities

$$\sin(\alpha+\beta) = \sin\alpha\cos\beta + \sin\beta\cos\alpha$$
$$\cos(\alpha+\beta) = \sin\alpha\sin\beta + \cos\alpha\cos\beta$$

using the small angle approximations

$$\cos\gamma \approx 1 - \frac{\gamma^2}{2} \; ; \; \sin\gamma \approx \gamma - \frac{\gamma^3}{6}$$

integrating over $\varepsilon$ and simplifying yields $(2 \le j \le k-1)$:

$$M_j \approx \frac{1 + j\,M_{j-1}}{j+1} - \frac{\pi\,\Gamma[(j+1)/2]\,\Gamma^3[(j+2)/6]}{24\,(j+5)\,\Gamma[j/2]\,\Gamma^3[(j+5)/6]} \frac{3 + (j+2)\,M_{j-1}}{N_j^2} \qquad \text{(A2)}$$

Assume that $M_j$ is of the form

$$M_j \approx 1 - \frac{C_j}{\prod_{i=1}^{j} N_i^{2/j}}$$

This holds for $j=1$ since

$$M_1 = \frac{\sin(\pi/N_1)}{(\pi/N_1)} \approx 1 - \frac{\pi^2}{6}\frac{1}{N_1^2} \quad \rightarrow \quad C_1 = \frac{\pi^2}{6}$$

We intend to show that this expression holds inductively. Substituting for $M_{j-1}$ in Eq.(A2) yields

$$M_j \approx 1 - \frac{j\ C_{j-1}}{j+1}\prod_{i=1}^{j-1} N_i^{-2/(j-1)} - (j+5)\ \frac{\pi\,\Gamma[(j+1)/2]\,\Gamma^3[(j+2)/6]}{24\,\Gamma[j/2]\,\Gamma^3[(j+5)/6]}\,N_j^{-2}$$

$$+\ \frac{\pi\,\Gamma[(j+1)/2]\,\Gamma^3[(j+2)/6]}{24\,\Gamma[j/2]\,\Gamma^3[(j+5)/6]}(j+2)\ C_{j-1}\ N_j^{-2}\prod_{i=1}^{j} N_i^{-2/(j-1)}$$

Ignoring the last term (since it is of higher power of $N^{-1}$) and maximizing over the value of $N_j$ shows that the above assumption is correct and yields:

$$C_j = \frac{j^2\ C_{j-1}}{j^2-1}\ T_j$$

where

$$T_j = \left\{\frac{\pi\,\Gamma[(j+1)/2]\,\Gamma^3[(j+2)/6]}{24\,\Gamma[j/2]\,\Gamma^3[(j+5)/6]}\ \frac{(j^2-1)}{j\ C_{j-1}}\right\}^{1/j}$$

This result also yields the solution to the factorization of the number of levels in the angle quantizers:

$$N_j \approx \frac{T_j^{(j-1)/2}}{\prod_{i=j+1}^{k-1} T_i^{1/2}}\ N_{\ell}^{1/(k-1)}$$

Table VI presents some useful precomputed results for the factorization of $N_{\ell}$. The above expression for $N_j$ requires the computation of $k-1$ values of $T_j$ and is different for each $k$. Instead, we present values of $F_j$, a sequential factorization value. It is defined as the proportion of the unused angle levels which should be assigned to the $j$-th angle quantizer

| $j$ | $F_j$ | $j$ | $F_j$ |
|---|---|---|---|
| 1 | 1.000 | 13 | 0.987 |
| 2 | 0.7573 | 14 | 0.9886 |
| 3 | 0.8584 | 15 | 0.9899 |
| 4 | 0.9074 | 16 | 0.9909 |
| 5 | 0.9348 | 17 | 0.9918 |
| 6 | 0.9515 | 18 | 0.9926 |
| 7 | 0.9625 | 19 | 0.9933 |
| 8 | 0.97 | 20 | 0.9938 |
| 9 | 0.9755 | 21 | 0.9943 |
| 10 | 0.9796 | 22 | 0.9947 |
| 11 | 0.9827 | 23 | 0.9951 |
| 12 | 0.9851 | 24 | 0.9955 |

Table VI - Sequential factorization values for a spherical source.

The sequential process, beginning with quantizer $Q_{k-1}$, is as follows:

$$N_{k-1} = F_{k-1}\, N_{\varphi}^{1/(k-1)}$$

$$N_{k-2} = F_{k-2}\left\{\frac{N_{\varphi}}{N_{k-1}}\right\}^{1/(k-2)}$$

$$N_{k-3} = F_{k-3}\left\{\frac{N_{\varphi}}{N_{k-1}\, N_{k-2}}\right\}^{1/(k-3)}$$

$$\vdots$$

$$N_2 = F_2\left\{\frac{N_{\varphi}}{\prod_{i=3}^{k-1} N_i}\right\}^{1/2}$$

$$N_1 = F_1\left\{\frac{N_{\varphi}}{\prod_{i=2}^{k-1} N_i}\right\}$$

## REFERENCES


1. J. Max, "Quantizing for Minimum Distortion," *IRE Trans. Inform. Theory*, Vol. IT-6, March 1960, pp.7-12.
2. J.J.Y. Huang & P.M. Schultheiss, "Block Quantization of Correlated Gaussian Random Variables," *IEEE Trans. Comm. Sys.*, Vol. CS-11, Sept. 1963, pp.289-296.
3. P. Zador, *Development and Evaluation of Procedures for Quantizing Multivariate Distributions*, Stanford Univ. Dissert., Stat. Dept., Dec. 1963.
4. A. Gersho, "Asymptotically Optimal Block Quantizers," *IEEE Trans. Inform. Theory*, Vol. IT-25, July 1979, pp.373-380.
5. J.H. Conway & N.J.A. Sloane, "Voronoi Regions of Lattices, Second Moments of Polytopes and Quantization," *IEEE Trans. Inform. Theory*, Vol. IT-28, March 1982, pp.211-226.
6. W.A. Pearlman, "Polar Quantization of a Complex Gaussian Random Variable," *IEEE Trans. Comm.*, Vol. COM-27, June 1979, pp.892-899.
7. J.A. Bucklew and N.C. Gallagher Jr., "Quantization Schemes for Bivariate Gaussian Random Variables," *IEEE Trans. Inform. Theory*, Vol. IT-25, Sept. 1979, pp.537-543.
8. J.A. Bucklew and N.C. Gallagher Jr., "Two-Dimensional Quantization of Bivariate Circularly Symmetric Densities," *IEEE Trans. Inform. Theory*, Vol. IT-25, Nov. 1979, pp.667-671.
9. S.G. Wilson, "Magnitude/Phase Quantization of Independent Gaussian Variates," *IEEE Trans. Comm.*, Vol. COM-28, Nov. 1980, pp.1924-1929.
10. P.F. Swaszek & J.B. Thomas, "Optimal Circularly Symmetric Quantizers," to appear in *The Journal of the Franklin Institute* or Chapter 4 of this dissertation.
11. R.D. Lord, "The Use of the Hankel Transform in Statistics," *Biometrika*, Vol. 41, June 1954, pp.44-55.
12. M. Kendall, *The Geometry of n Dimensions*, Griffin Co. Ltd., London, 1962.
13. W.R. Bennett, "Spectra of Quantized Signals," *Bell System Tech. Jour.*, July 1948, pp.446-472.
14. P.E. Fleischer, "Sufficient Conditions for Achieving Minimum Distortion in a Quantizer," *IEEE Int'l. Conv. Rec.*, 1964, part 1, pp 104-111.
15. R. Weinstock, *Calculus of Variations*, McGraw-Hill, N.Y., 1952.
16. H. Abramowitz and I.A. Stegun, *Handbook of Mathematical Functions*, National Bureau of Standards Applied Mathematics Series #55, Dec. 1972.
17. V.R. Algazi, "Useful Approximations to Optimal Quantization," *IEEE Trans. Comm. Tech.*, Vol. COM-14, June 1966, pp.297-301.
18. D.K. McGraw & J.F. Wagner, "Elliptically Symmetric Distributions," *IEEE Trans. Inform. Theory*, Vol. IT-14, Jan. 1968, pp 110-120

# CHAPTER 4 - OPTIMAL CIRCULARLY SYMMETRIC QUANTIZERS

## INTRODUCTION

The canonical example of an zero-memory or one-dimensional quantizer is Max's Gaussian probability density function quantizer [1] for the performance criterion Mean Square Error (MSE). The MSE criterion has universal appeal in its tractability and its intuitive relationship to noise power, hence signal-to-noise ratio (SNR). Rate distortion theory, however, suggests that multidimensional or block quantizers may be more efficient. Research interest in multidimensional quantization began with the work of Huang and Schultheiss [2] who considered the problem of quantizing a correlated Gaussian source efficiently. Their solution was to uncorrelate the source by an appropriate linear filter, thereby changing the set of coordinates, and to quantize the resulting independent Gaussian random variables with separate Max-type quantizers.

Zador [3] examined the more general problem of quantizing a multidimensional source under the assumption of a large number of levels. He employed Bennett's compandor model and derived error rates depending upon the compressor function and uniform quantizer used. His expressions showed that the problem of optimal quantization could be divided into two separate problems: finding the best compressor function on the multidimensional input space and implementing the optimal multidimensional uniform quantizer on the unit hypercube. In two dimensions, the optimal uniform quantizer is a honeycomb-like tessellation of hexagons. When mapped by the inverse of the compandor function, the

quantizer becomes a pattern of distorted hexagons on the plane [4].

Another major area of interest in multidimensional quantizer design rests in the use of polar coordinates for the independent, bivariate case. Specifically, rather than separately quantizing the abscissa and ordinate as in Fig. 1, a change of variables to polar coordinates is effected. The resulting magnitude and phase are quantized separately by real-time one-dimensional quantizers. Of particular interest is the quantizing of independent, bivariate Gaussian random variables with density

$$p(x,y) = \frac{1}{2\pi} e^{-(x^2+y^2)/2}$$

For example DFT coefficients, holographic data or pairs of inputs from an iid Gaussian source can be considered as the output of a bivariate Gaussian source.

Independent, unit-power Gaussian variates in rectangular coordinates transform to independent magnitude and phase on the polar coordinates plane by the transformations

$$r = \sqrt{x^2+y^2} \; ; \; \varphi = \tan^{-1} \frac{y}{x}$$

The resulting source density expressed as a function of the polar coordinates is

$$p(r,\varphi) = \frac{1}{2\pi} r \, e^{-r^2/2}$$

The magnitude $r$ is Rayleigh distributed on $[0,\infty)$ and the phase $\varphi$ is uniformly distributed on $[0,2\pi)$. Minimizing MSE results in a uniform quantizer for the phase angle and a scaled Max-type Rayleigh quantizer for the magnitude. It has been shown [5,6] that polar coordinates quantizers for a bivariate Gaussian source almost always have smaller MSE than

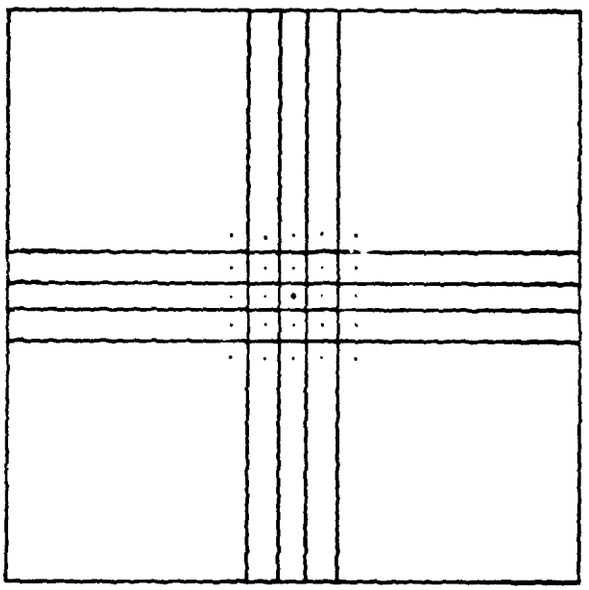


Fig. 1. Typical bivariate rectangular coordinates quantizer.

rectangular quantizers.

This chapter examines the optimality of the polar quantizers developed by Pearlman [5] and Bucklew and Gallagher [6] for the MSE criterion. It is well known [4] that two conditions are necessary for a *local* minimum of MSE [7,8]: centroidal output points and Dirichlet partition boundaries. Polar quantizers do not conform to these conditions. Permutations which do conform (labeled Dirichlet polar quantizers) will be developed and compared to other available two-dimensional quantization schemes. Wilson's technique [9] will be mentioned and considered as an input for the Dirichlet form. Although this chapter will pursue in depth only the bivariate Gaussian case, the extensions to higher dimensions [10] and other circularly symmetric densities [11] will be outlined.

## OPTIMAL TWO-DIMENSIONAL QUANTIZERS

Define the minimum MSE, N-level quantizer $Q_N$ on the plane by $\{S_i, \hat{\mathbf{x}}_i; i=1,2,\ldots,N\}$ where the $S_i$ are disjoint regions such that their union covers the plane and the $\hat{\mathbf{x}}_i$ are the output points associated by the quantizer to these regions. The quantizer's operation for the input vector $\mathbf{x}$ is

$$Q_N(\mathbf{x}) = \hat{\mathbf{x}}_i \quad ; \quad \mathbf{x} \in S_i$$

For an input $\mathbf{x}$ with bivariate pdf $p(\mathbf{x})$, the MSE $D$ is

$$D = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |\mathbf{x} - Q_N(\mathbf{x})|^2 p(\mathbf{x}) d\mathbf{x}$$

Minimizing $D$ over the choice of $S_i$ and $\hat{\mathbf{x}}_i$, the following are necessary conditions:

$$\hat{\mathbf{x}}_i = \frac{\int\int_{S_i} \mathbf{x}\, p(\mathbf{x}) d\mathbf{x}}{\int\int_{S_i} p(\mathbf{x}) d\mathbf{x}} \tag{1}$$

which states that the output $\hat{\mathbf{x}}_i$ is the centroid of the region $S_i$ with density $p(\mathbf{x})$ and

$$S_i = \bigcap_{j=1, j \neq i}^{N} \{\mathbf{x} : |\mathbf{x} - \hat{\mathbf{x}}_i| < |\mathbf{x} - \hat{\mathbf{x}}_j|\} \tag{2}$$

which states that $S_i$ is formed by taking the intersection of nearest neighbor or Dirichlet partitions of $\hat{\mathbf{x}}_i$ and the other output points. The points of equality in Eq.(2) are the region boundaries which are assigned to either region and contribute equivalent error either way. A Dirichlet partition is the perpendicular bisecter of the line segment connecting a pair of output points. From Eq.(2), it can be shown that the resulting $S_i$ are all convex, simply connected regions. This partitioning holds for most mean error measures while centroidal outputs holds only for MSE. The resulting MSE for this optimal quantizer is

$$D = \sigma_{\mathbf{x}}^2 - \sum_{i=1}^{N} |\hat{\mathbf{x}}_i|^2 \int\int_{S_i} p(\mathbf{x}) d\mathbf{x}$$

where $\sigma_{\mathbf{x}}^2$ is the signal power.

For the uniform density on the unit square, the optimal region pattern for fine quantization, ignoring edge effects, is known to be a tessellation of regular hexagons. For other densities, Eqs.(1) and (2) may be used iteratively to converge to a local minimum of MSE. Note that if the regions are fixed, Eq.(1) is necessary and sufficient to minimize D. When the output points are fixed, Eq.(2) is necessary and sufficient to minimize D. The iterative design method, as previously proposed for one-dimensional problem solutions [12], is to select a set of outputs $\{\hat{\mathbf{x}}_i\}$ and to employ Eq.(2) to select the $S_i$ optimally. This set of outputs and regions has a distortion measure $D_1$. Usually, Eq.(1) is not satisfied, the

$\{\hat{\mathbf{x}}_i\}$ not being optimal for the generated regions, so redefining the outputs by Eq.(1) will decrease D to a value smaller than $D_1$, say $D_2$. Similarly, now Eq.(2) is probably not satisfied, so redefining the regions will again decrease the error. This iterative scheme converges to a local minimum of D due to the fact that D is reduced by each application of Eq.(1) or (2) and that D is lower bounded by zero by being the integral of a positive quantity.

## DIRICHLET POLAR QUANTIZERS

Wilson [9] classified two types of polar quantizers: Strictly Polar (SPQ) and Unrestricted Polar (UPQ) Quantizers. For the SPQ's, the total number of outputs N is factored into $N_r \times N_\varphi$, the number of magnitude and phase levels respectively. The UPQ's, a larger class of quantizers, require only that the number of outputs sum up to N ($P_1+P_2+\ldots P_M=N$), hence different radii levels can have different numbers of phase levels. For small N, UPQ's have been shown to substantially reduce the MSE. All polar quantization regions are partial annuli, delimited by rays of constant angle and arcs of constant radius as in Fig. 2. Unfortunately, these quantizers do not satisfy Eqs (1) and (2). In particular, the magnitude boundaries are not Dirichlet partitions. From Eq (2), each $S_i$ is a convex polygon which partial annuli are not.

The iterative technique, as explained above using Eqs.(1) and (2), may be employed to the strictly polar quantizers to reduce MSE and converge to a local minimum. After selecting a factorization of $N=N_r \times N_\varphi$, applying Eq.(2) yields a pattern as in Fig. 3. The inherent symmetry of this pattern allows the analysis to focus on one slice of angle $2\pi/N_\varphi$. The

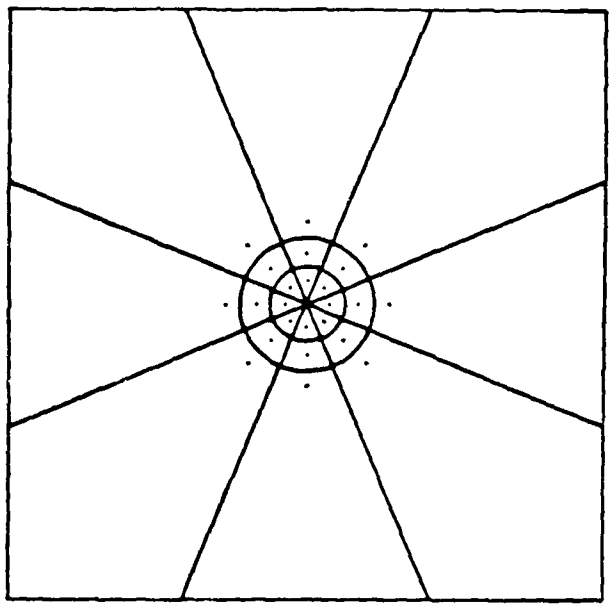


Fig. 2 - Polar quantizer (SPQ).

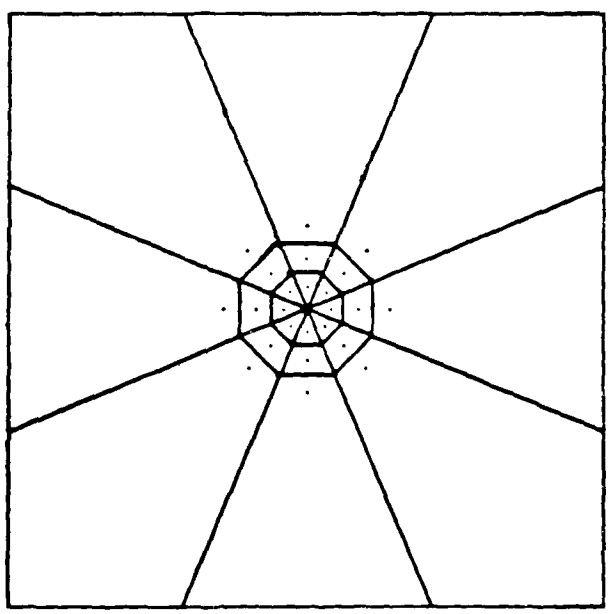
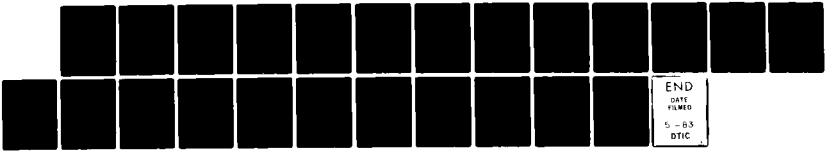
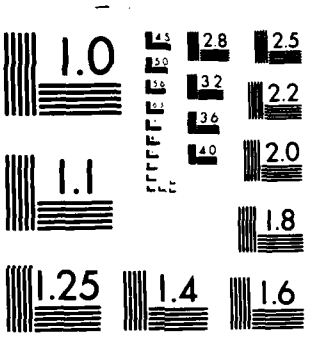


Fig. 3 - Dirichlet Polar Quantizer (DPQ) pattern.

AD-A127 258 ROBUST AND VECTOR QUANTIZATION(U) PRINCETON UNIV NJ INFORMATION SCIENCES AND SYSTEMS LAB P F SWASZEK ET AL. MAR 83 N00014-81-K-0146 2/2

UNCLASSIFIED F/G 12/1 NL





iterative use of Eqs (1) and (2) will not change the phase boundaries; only the magnitude boundaries will move. Similarly, the output points will vary along the ray bisecting the phase boundaries. Hence, a one-dimensional iteration will yield these Dirichlet Polar Quantizers (DPQ's) from the SPQ's. Dallas [13] has applied a similar region shape to the reduction of the Fourier domain phase quantization noise for computer generated holograms.

From Fig. 3, it is seen that the DPQ's can be implemented as follows. First, quantize the phase to one of $N_\varphi$ levels with a uniform quantizer on $[0,2\pi)$. The second coordinate used to specify the output is its distance $s$ along the quantized phase ray

$$s = r\cos(\varphi - \hat{\varphi})$$

for $\hat{\varphi}$ the quantized version of $\varphi$. The univariate probability distribution function of this distance coordinate can be found to be

$$f(s) = \frac{2N_\varphi}{\sqrt{2\pi}} e^{-s^2/2}\left[\Phi(s\tan\pi/N_\varphi) - .5\right]$$

for $\Phi$ the error function integral

$$\Phi(y) = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,dx$$

Letting $N_\varphi \to \infty$ (with l'Hopital's rule), $f(s)$ approaches the Rayleigh density.

Standard Max-type expressions may be used to define the minimum MSE, $N_r$-level quantizer for $s$

$$s_i = \frac{\hat{s}_{i-1}+\hat{s}_i}{2}\ ;\ \hat{s}_i = \frac{\int_{s_i}^{s_{i+1}} s\,f(s)\,ds}{\int_{s_i}^{s_{i+1}} f(s)\,ds}$$

where the $\hat{s}_i$ are the quantizer outputs and the $s_i$ are the region endpoints. Uniqueness of this quantizer is shown by applying Fleischer's condition [12] to the distance density

$$\frac{\partial^2}{\partial s^2} \log f(s) < 0$$

Wilson's solutions of the UPQ's for N=1,...,32 may also be considered with the iterative technique. His N = 1, 2, 3 and 4 cases are already optimum. The N = 5, 6, 7 and 8 solutions are easily extended. Unfortunately, for N > 8, the boundaries are no longer easy to compute and the resulting analysis is not included here. He only considered small N since the number of factorizations grows quickly with N and because the small N region is of importance since it is here that rectangular formats outperform polar forms.

## DIRICHLET ROTATED POLAR QUANTIZERS

The previous section showed that Eqs.(1) and (2) can be applied to a set of outputs to iterate toward a local minimum of MSE. The resulting quantizer will vary depending upon the initial output point pattern. A rectangular starting grid produces a rectangular quantizer, a pair of Max Gaussian quantizers, since the partitions will always move perpendicular to themselves. A polar initial pattern produces the Dirichlet Polar Quantizer already introduced.

Consider the polar quantizer (SPQ) where the magnitude and phase are independently quantized (Fig. 2). A rotation of every other magnitude ring, as in Fig. 4, does not change the associated MSE. This new pattern, when applied as a starting point for the iterative method with

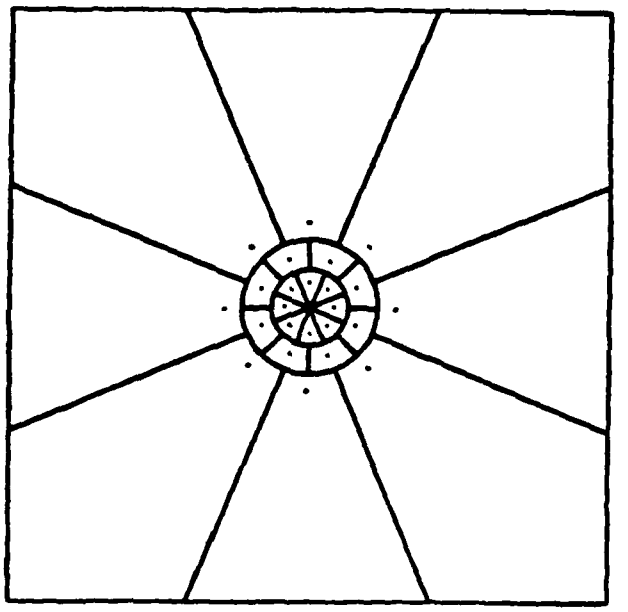


Fig. 4 - Rotated polar pattern.

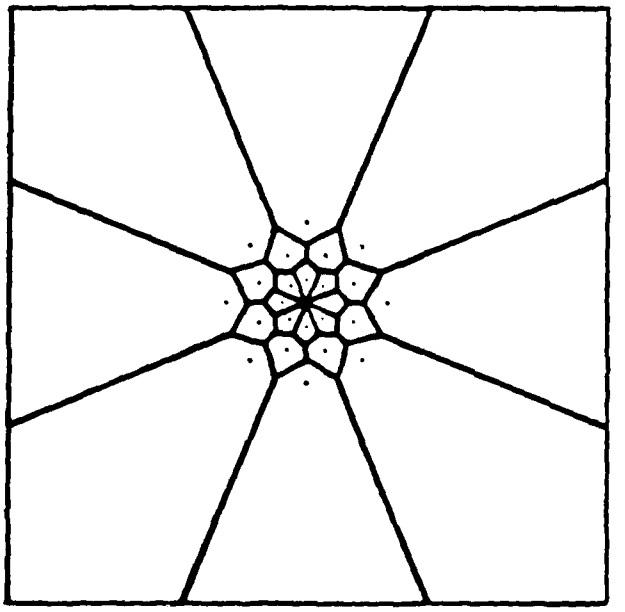


Fig. 5 - Dirichlet Rotated Polar Quantizer (DRPQ) pattern.

Eqs.(1) and (2), will yield a quantization pattern as in Fig. 5, quite different from the Dirichlet Polar Quantizer. This Dirichlet Rotated Polar Quantizer (DRPQ), although more difficult to implement than the DPQ, has lower MSE (for a possible implementation, see Appendix A).

Other rotations and permutations on the plane could be used to solve for better quantizers. However, most other patterns make Eq. (1) difficult to compute. A further extension of this rotated form is to allow a central region with $N_\varphi$ sides and output value zero similar to Wilson's 5 through 8 patterns. The MSE savings could be dramatic, but are not considered here.

**EXAMPLES**

The iterative technique is defined by Eqs.(1) and (2). The numerical calculations of the region probabilities and moments for the bivariate Gaussian density are described in Appendix B. For the examples, the following factorizations of $N$ were employed:

Rectangular: $N_x \approx N_y$
Polar: $N_\varphi \approx 2.6\, N_r$
DPQ: $N_\varphi \approx 2.6\, N_s$
DRPQ: $N_\varphi \approx N_r$

Symmetry arguments show that the rectangular MSE is minimized if the levels are equally divided among the coordinates. Previously published results suggest the factorization for the polar scheme. As the number of levels gets large, the DPQ's and the polar quantizers are equivalent, hence the asymptotic factorizations of $N$ are the same.

For the tabulated results, all factorizations for the DPQ's were com-

pared and the best result occurred concurrent with the polar factorization. For the DRPQ's, all combinations were attempted for $N \leq 144$. It was seen that equal division of the levels produced optimal results. For $N > 144$, only equal factorizations were attempted; hence, the actual error rates may be lower than the tabulated results for those values of $N$. The comparison of MSE rates is in Table I with a plot of the results in Fig. 6. Error values for polar and rectangular quantizers are included for comparison. The number in parenthesis is the actual number of levels if different from the first column. This difference appears due to the necessity to factor N into appropriate integers. Figs. 7 through 13 depict the DRPQ patterns for some of the values listed in Table I.

| N | Polar | Dirichlet Polar | Dirichlet Rotated Polar | Rectangular |
|---|---|---|---|---|
| 16 | .2396 | .2391 | .2224 | .2350 |
| 25 | .1710 (24) | .1702 (24) | .1462 | .1599 |
| 36 | .1176 | .1174 | .1052 | .1159 |
| 49 | .08889 (48) | .08882 (48) | .07899 | .08800 |
| 64 | .06973 | .06967 | .06134 | .06908 |
| 100 | .04392 (102) | .04387 (102) | .04003 | .04586 |
| 144 | .03244 (140) | .03241 (140) | .02816 | .03268 |
| 225 | .02056 | .02055 | ≤.01822 | .02146 |
| 324 | .01468 (320) | .01467 (320) | ≤.01280 | .01519 |
| 529 | .008904(532) | .008899(532) | ≤.008046 | .009482 |
| 900 | .005314 | .005308 | ≤.004684 | .005668 |

Table I - Bivariate Gaussian density quantizer's MSE values.

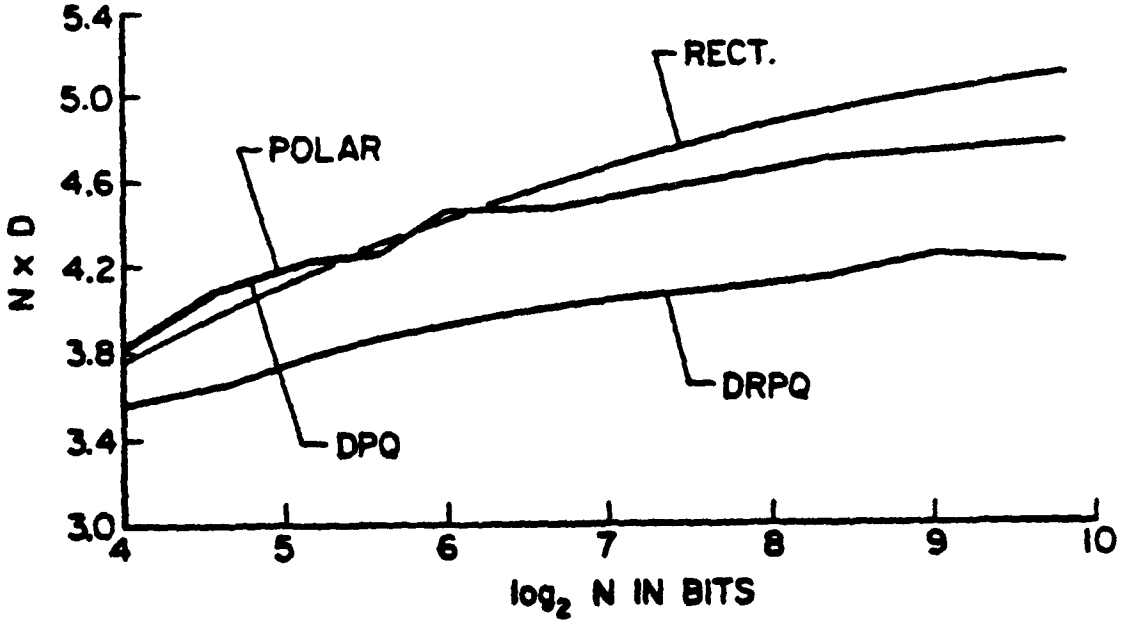



Fig. 6 - Comparison of MSE rates for four bivariate quantizers (Rectangular, SPQ, DPQ and DRPQ).

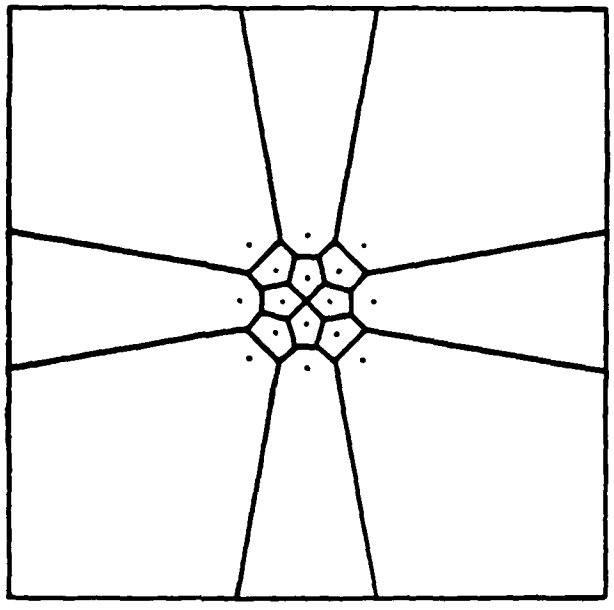


Fig. 7 - N=16 DRPQ pattern.

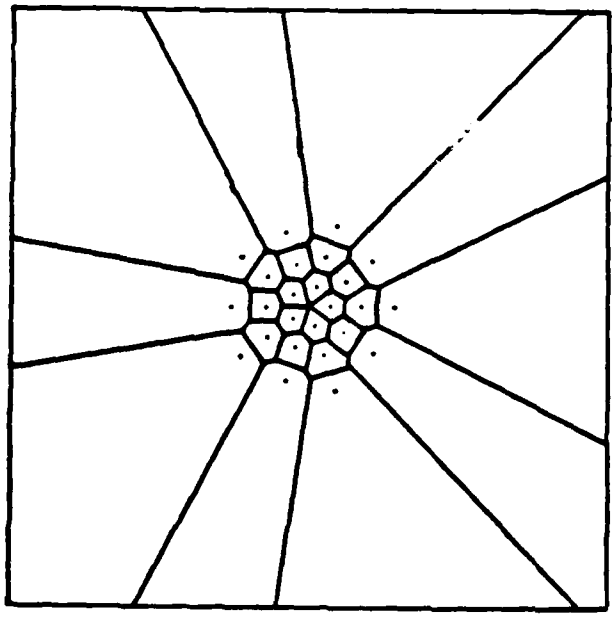


Fig. 8 - N=25 DRPQ pattern.

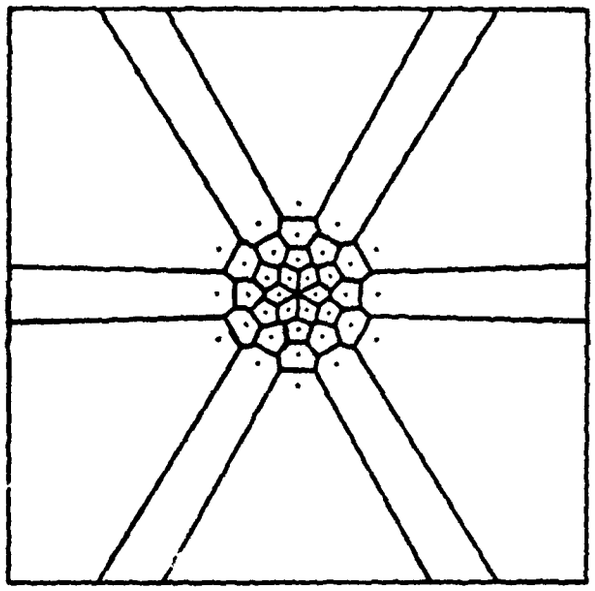


Fig. 9 - N=36 DRPQ pattern.

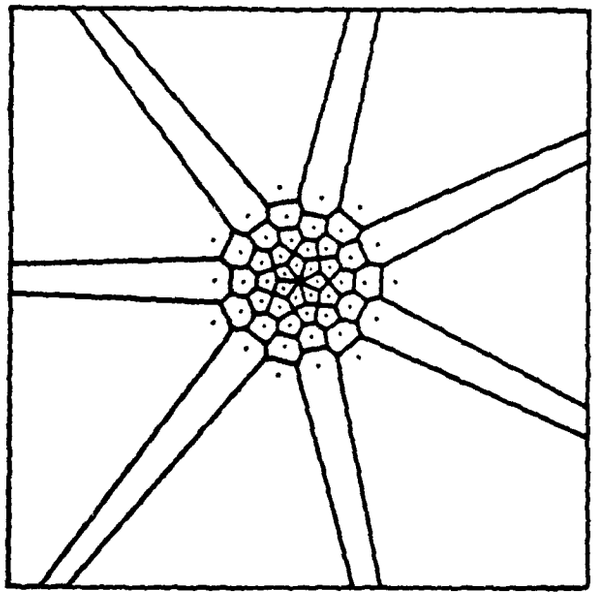


Fig. 10 - N=49 DRPQ pattern.

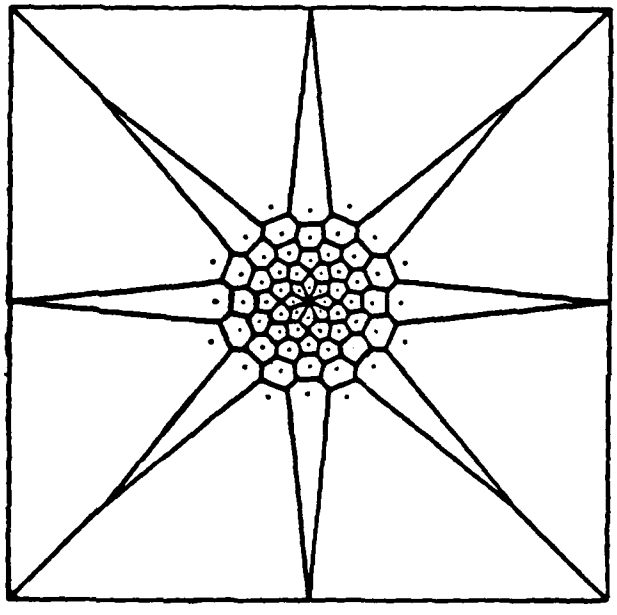


Fig. 11 - N=64 DRPQ pattern.

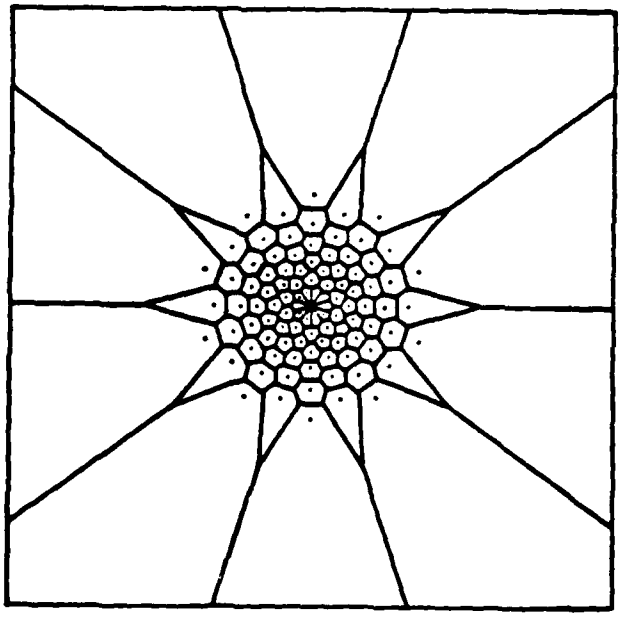


Fig. 12 - N=100 DRPQ pattern.

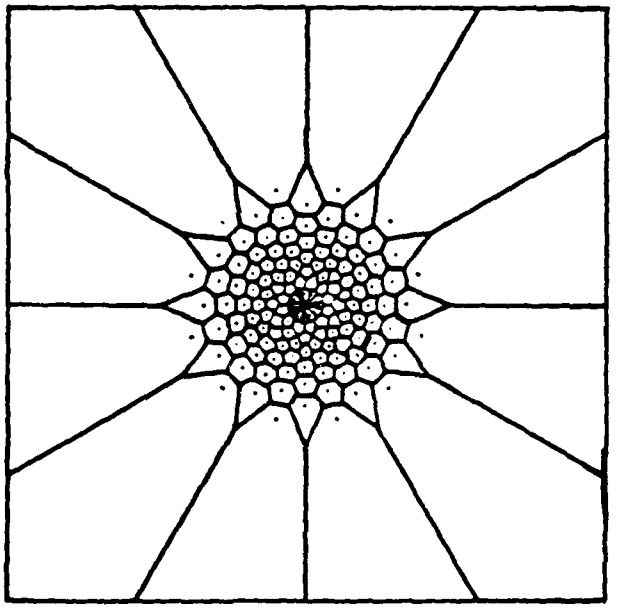


Fig. 13 - N=144 DRPQ pattern.

## CONCLUSIONS

For two reasons, the presented figures are dominated by patterns for DRPQ's. The DPQ patterns are all of the same form as Fig. 3 and the DPQ's MSE is only slightly below that of the SPQ, being equal when $N \to \infty$. The DRPQ patterns are included to demonstrate the hexagonality of the quantization regions, the way in which the hexagon sizes are distributed and because the DRPQ's substantially reduce MSE. At N=100, the gain in SNR is .6 dB over rectangular and .4 dB over polar quantizers.

From Fig. 6, the asymptotic MSE rates can be considered. The SPQ and DPQ (equivalent as $N \to \infty$) have rate $N \times D = 4.95$ for a bivariate Gaussian source. The corresponding optimal rate is 4.03 and the rectangular rate is 5.44. From the graph, the DRPQ rate falls between the polar and optimum. Although the DRPQ is not optimal, it does always perform better than both polar and rectangular schemes. Results for DRPQ's allowing an $N_\varphi$-sided polygonal region at the origin may be even better.

Up to this point, this chapter considered only the bivariate Gaussian case. The trapezoids of the Dirichlet Polar Quantizer and the polygons of the DRPQ become polytopes in higher dimensions. Other circularly symmetric and non-circularly symmetric densities may also be used. The difficulty in both cases is obtaining accurate probability and moment integrals.

Polar quantizers as described in the literature minimize MSE subject to independent coordinates. Loosening the coordinate selection slightly (to angle $\varphi$ and distance $s$) yields DPQ's, again minimizing MSE for their constraint class Further loosening of the coordinate class yields the

DRPQ's with substantially reduced MSE, but increased complexity of implementation. The intuition to be gained from this work is as follows: all of the mentioned schemes (rectangular, polar, DPQ and DRPQ) minimize MSE subject to their implementation constraint. Rectangular formats retain centroids and Dirichlet partitions (necessary conditions), but lose the symmetry of the problem; polar forms preserve the problem symmetry but lose the necessary conditions; the DPQ and DRPQ schemes have both.

## APPENDIX A - DRPQ IMPLEMENTATION

This appendix presents a possible real-time implementation for the Dirichlet Rotated Polar Quantizers (DRPQ's). For an $N$-level quantizer, the levels factorization is $N = N_r \times N_\varphi$. The scheme is as follows:

1 - Convert the input $\mathbf{x}$ to polar coordinates, $r$ and $\varphi$.

2 - Process the phase angle $\varphi$ with a $2N_\varphi$-level uniform quantizer on $[0,2\pi)$. The outputs $\hat{\varphi}$ are of the form $\pi(2k-1)/2N_\varphi$ for $k \in [1,2,\ldots 2N_\varphi]$.

3 - Process the magnitude r with a $N_r$-level, lower value quantizer. This quantizer's output $\hat{r}$ is the magnitude value closest and less than the actual distance.

4 - For a lower magnitude of level j and a phase of level k, compare

$$\hat{r}_j\, e^{j\hat{\varphi}_{kj}} - r\, e^{j\varphi} \quad \text{and} \quad \hat{r}_{j+1}\, e^{j\hat{\varphi}_{kj+1}} - r\, e^{j\varphi}$$

to find the closest output where

$$\hat{\varphi}_{kj} = \frac{\pi}{N_\varphi}\frac{2k-1}{2} + \frac{\pi}{2N_\varphi} \qquad \begin{array}{l} + \text{ if } |j-k| \text{ is even} \\ - \text{ if } |j-k| \text{ is odd} \end{array}$$

This scheme requires no compressor functions as does the optimal scheme and is real-time, digital implementational. Also, this implementation extends trivially to the zero-output extension of the DRPQ previously described.

## APPENDIX B - BIVARIATE GAUSSIAN INTEGRALS

The numerical calculations necessary to solve iteratively Eqs.(1) and (2) for the DPQ's and DRPQ's involve integrations on the bivariate Gaussian plane of polygonal regions. Polygons on the plane can be partitioned into the sum and difference of triangles which have a vertex at the origin and a side along a coordinate axis. Without loss of generality, since the regions are symmetric about a ray of constant angle, we assume the regions to be symmetric about the positive x-axis and use this axis as the side of all the triangles (*see* Fig. 14).

For the calculation of the areas (probabilities), these triangles are again partitioned into the difference of two right triangles with a vertex at the origin as in Fig. 15. A right triangle is rotated about the origin to be equivalent to one with vertices (0,0), (0,$h$) and ($h$,$k$) as in Fig. 16. The area is then

$$V(h,k) = \frac{1}{2\pi} \int_0^h e^{-x^2/2} \int_0^{kx/h} e^{-y^2/2}\, dy\, dx$$

This expression, although not directly integrable, can be expanded into a series summation [14,15]

$$V(h,k) = (2\pi)^{-1} \left[ \lambda(1-e^{-m}) - \frac{1}{3}\lambda^3(1-e^{-m}-me^{-m}) \right.$$
$$\left. + \frac{1}{5}\lambda^5\left(1-e^{-m}-me^{-m}-\frac{m^2}{2!}e^{-m}\right) - \cdots \right.$$

with $\lambda = k/h$ and $m = \frac{1}{2}h^2$. Truncation of this series after 20 terms yielded the approximations employed in the presented results.

The symmetry of the DPQ and DRPQ regions reduces the moment calculation to that along the ray of symmetry (the $x$-axis). For the original

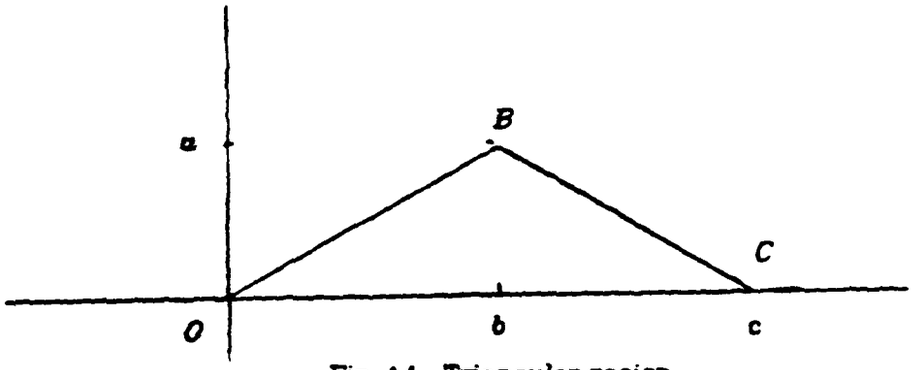



Fig. 14 - Triangular region.

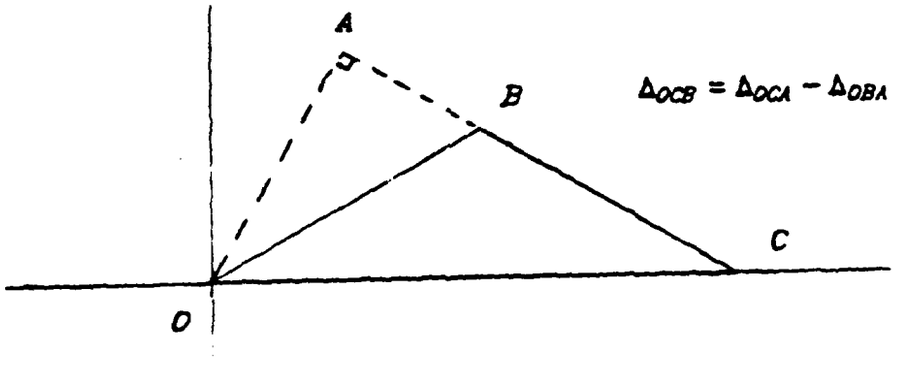



Fig. 15 - Difference of right triangles.

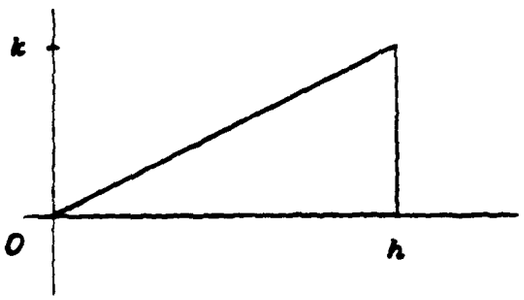



Fig. 16 - Triangle with sides of length $h$ and $k$.

triangle with vertices at (0,0), (c ,0) and (b ,a ) the $x$ moment is

$$M = \int\int_{\Delta_{OAB}} x \, \frac{1}{2\pi} e^{-\frac{x^2+y^2}{2}} \, dx \, dy$$

Inserting the correct limits in the integration, integrating over $x$, completing the square in $y$ in the exponent and integrating over $y$ yields

$$M = \frac{\Phi(ag) - .5}{\sqrt{2\pi}g} - \exp\left\{\frac{-ac}{2(c-b)h^2}\right\} \times \frac{\Phi\left[ah + \frac{a}{(c-b)h}\right] - \Phi\left[\frac{a}{(c-b)h}\right]}{\sqrt{2\pi}h}$$

where

$$g = \sqrt{1+\frac{a^2}{b^2}} \quad \text{and} \quad h = \sqrt{1+\frac{a^2}{(c-b)^2}}$$

**REFERENCES**


1. J. Max, "Quantizing for Minimum Distortion," *IRE Trans. Inform. Theory* Vol. IT-6, March 1960, pp.7-12.
2. J.J.Y. Huang and P.M. Schultheiss, "Block Quantization of Correlated Gaussian Random Variables," *IEEE Trans. Comm. Sys.* Vol. CS-11, Sept. 1963, pp. 289-296.
3. P. Zador, *Development and Evaluation of Procedures for Quantizing Multivariate Distributions*, Stanford Univ. Dissert., Dept. of Stat., Dec. 1963.
4. A. Gersho, "Asymptotically Optimal Block Quantization," *IEEE Trans. Inform. Theory*, Vol. IT-25, July 1979, pp.373-380.
5. W.A. Pearlman, "Polar Quantization of a Complex Gaussian Random Variable," *IEEE Trans. Comm.*, Vol. COM-27, June 1979, pp.892-899.
6. J.A. Bucklew and N.C. Gallagher Jr., "Quantization Schemes for Bivariate Gaussian Random Variables," *IEEE Trans. Inform. Theory*, Vol. IT-25, Sept. 1979, pp.537-543.
7. R.M. Gray & E.D. Karnin, "Multiple Local Optima in Vector Quantization," *IEEE Trans. Inform. Theory*, Vol. IT-28, March 1982, pp.256-261.
8. R.M. Gray, J.C. Kieffer & Y. Linde, "Locally Optimal Block Quantizer Design," *Inform. & Control*, Vol. 45, May 1980, pp.178-198.
9. S. Wilson, "Magnitude/Phase Quantization of Independent Gaussian Variates," *IEEE Trans. Comm.*, Vol. COM-28, Nov. 1980, pp.1924-1929.
10. P.F. Swaszek and J.B. Thomas, "k-Dimensional Polar Quantizers for Gaussian Sources," *Proc. Allerton Conf. Comm. Cont. & Comp.*, Sept. 1981, pp.89-97 or Chapter 3 of this dissertation.
11. J.A. Bucklew and N.C. Gallagher Jr., "Two-Dimensional Quantization of Bivariate Circularly Symmetric Densities," *IEEE Trans. Inform. Theory*, Vol. IT-25, Nov. 1979, pp.667-671.
12. P.E. Fleischer, "Sufficient Conditions for Achieving Minimum Distortion in a Quantizer," *IEEE Int'l. Conv. Rec.*, 1964, part 1,pp.104-111.
13. W.J. Dallas, "Magnitude-Coupled Phase Quantization," *Applied Optics*, Vol. 43, Oct. 1974, pp.2274-2279.
14. N.L. Johnson and S. Kotz *Continuous Multivariate Distributions* Wiley and Sons, Inc., N.Y., 1972.
15. C. Nicholson, "The Probability Integral for Two Variables," *Biometrika*, Vol. 33, April 1943, pp.59-72.

# CHAPTER 5 - CONCLUSIONS

## REVIEW AND FURTHER RESEARCH

In this chapter, the main achievements of each of the preceding chapters will be discussed. Shortcomings and possible avenues of future research will be mentioned when available.

In the second chapter, zero-memory quantizers are designed when the available source statistical description is a histogram. The design of the histogram measurement is also discussed. This scheme is practical in that only information which is easy to obtain is needed to completely design the quantizer and the piecewise linear compressor is easy to implement. Recently, this idea has been extended to block quantizers [1]. Further research in robust quantizer design might include (i) applying other techniques besides Chebychev-like probability inequalities to the problem of allocating the histogram regions, (ii) discussing the design inaccuracy due to the empirical region probability measurements and (iii) in block quantization, using as the histogram cell a cross product of intervals and exploring this form of dependence structure.

The third chapter considered the extension of polar quantizers to greater than two dimensions. The general result for $k$-dimensions and any spherically symmetric source were presented. It was noted in the chapter that the Gaussian source did not exhibit an appreciable gain in performance on allowing the number of dimensions to increase. Among the problems which remain to be answered are the following: (i) can other coordinate systems be applied as easily and (ii) how often do spherically

symmetric sources occur naturally? The design was founded firmly upon the fact that the source was spherically symmetric. This implies a certain dependence structure on the multivariate density. Rectangular quantizers, although not performing as well in the tabulated examples, are robust in the sense that their error rate is independent of the multivariate source structure since only the marginal density matters. This fact suggests that when the multivariate structure is questionable, rectangular quantizers should be employed.

The fourth chapter extends the results of the third chapter by considering the optimality of spherical coordinates quantizers. Several other authors [2,3] have noted the lack of optimality of block quantizers. The chapter stresses the facts that the DRPQ's have hexagonal regions, which are conjectured to be optimal, and nearly optimal performance. Future research in this area could be (i) extending the examples to other sources beyond the bivariate Gaussian and (ii) allowing a central region in the pattern whose output value is zero.

## REFERENCES

1. K.D. Rines, N.C. Gallagher Jr. & J.A. Bucklew, "Nonuniform Multidimensional Quantizers," to appear in the *Proc. Princeton Conf. Info. Sci. Systems*, 1982.
2. R.M. Gray, J.C. Kieffer & Y. Linde, "Locally Optimal Block Quantizer Design," *Inform. & Control*, Vol. 45, May 1980, pp.178-198.
3. R.M. Gray & E.D. Karnin, "Multiple Local Optima in Vector Quantizers," *IEEE Trans. Inform. Theory*, Vol. IT-28, March 1982, pp.256-261.

OFFICE OF NAVAL RESEARCH
STATISTICS AND PROBABILITY PROGRAM

BASIC DISTRIBUTION LIST
FOR
UNCLASSIFIED TECHNICAL REPORTS

FEBRUARY 1982

| | Copies |
|---|---|
| Statistics and Probability Program (Code 411(SP))<br>Office of Naval Research<br>Arlington, VA 22217 | 3 |
| Defense Technical Information Center<br>Cameron Station<br>Alexandria, VA 22314 | 12 |
| Commanding Officer<br>Office of Naval Research<br>Eastern/Central Regional Office<br>Attn: Director for Science<br>Barnes Building<br>495 Summer Street<br>Boston, MA 02210 | 1 |
| Commanding Officer<br>Office of Naval Research<br>Western Regional Office<br>Attn: Dr. Richard Lau<br>1030 East Green Street<br>Pasadena, CA 91101 | 1 |
| U. S. ONR Liaison Office - Far East<br>Attn: Scientific Director<br>APO San Francisco 96503 | 1 |
| Applied Mathematics Laboratory<br>David Taylor Naval Ship Research and Development Center<br>Attn: Mr. G. H. Gleissner<br>Bethesda, Maryland 20084 | 1 |
| Commandant of the Marine Coprs (Code AX)<br>Attn: Dr. A. L. Slafkosky<br>Scientific Advisor<br>Washington, DC 20380 | 1 |
| Navy Library<br>National Space Technology Laboratory<br>Attn: Navy Librarian<br>Bay St. Louis, MS 39522 | 1 |
| U. S. Army Research Office<br>P.O. Box 12211<br>Attn: Dr. J. Chandra<br>Research Triangle Park, NC 27706 | 1 |
| Director<br>National Security Agency<br>Attn: R51, Dr. Maar<br>Fort Meade, MD 20755 | 1 |
| ATAA-SL, Library<br>U.S. Army TRADOC Systems Analysis Activity<br>Department of the Army<br>White Sands Missile Range, NM 88002 | 1 |
| ARI Field Unit-USAREUR<br>Attn: Library<br>c/o ODCSPER<br>HQ USAEREUR & 7th Army<br>APO New York 09403 | 1 |
| Library, Code 1424<br>Naval Postgraduate School<br>Monterey, CA 93940 | 1 |
| Technical Information Division<br>Naval Research Laboratory<br>Washington, DC 20375 | 1 |
| OASD (I&L), Pentagon<br>Attn: Mr. Charles S. Smith<br>Washington, DC 20301 | 1 |

Copies

Director
AMSAA
Attn: DRXSY-MP, H. Cohen
Aberdeen Proving Ground, MD 21005 1

Dr. Gerhard Heiche
Naval Air Systems Command (NAIR 03)
Jefferson Plaza No. 1
Arlington, VA 20360 1

Dr. Barbara Bailar
Associate Director, Statistical Standards
Bureau of Census
Washington, DC 20233 1

Leon Slavin
Naval Sea Systems Command (NSEA 05H)
Crystal Mall #4, Rm. 129
Washington, DC 20036 1

B. E. Clark
RR #2, Box 647-B
Graham, NC 27253 1

Naval Underwater Systems Center
Attn: Dr. Derrill J. Bordelon
Code 601
Newport, Rhode Island 02840 1

Naval Coastal Systems Center
Code 741
Attn: Mr. C. M. Bennett
Panama City, FL 32401 1

Naval Electronic Systems Command (NELEX 612)
Attn: John Schuster
National Center No. 1
Arlington, VA 20360 1

Defense Logistics Studies Information Exchange
Army Logistics Management Center
Attn: Mr. J. Dowling
Fort Lee, VA 23801 1

Copies

Reliability Analysis Center (RAC)
RADC/RBRAC
Attn: I. L. Krulac
Data Coordinator/
Government Programs
Griffiss AFB, New York 13441 1

Technical Library
Naval Ordnance Station
Indian Head, MD 20640 1

Library
Naval Ocean Systems Center
San Diego, CA 92152 1

Technical Library
Bureau of Naval Personnel
Department of the Navy
Washington, DC 20370 1

Mr. Dan Leonard
Code 8105
Naval Ocean Systems Center
San Diego, CA 92152 1

Dr. Alan F. Petty
Code 7930
Naval Research Laboratory
Washington, DC 20375 1

Dr. M. J. Fischer
Defense Communications Agency
Defense Communications Engineering Center
1860 Wiehle Avenue
Reston, VA 22090 1

Mr. Jim Gates
Code 9211
Fleet Material Support Office
U. S. Navy Supply Center
Mechanicsburg, PA 17055 1

Mr. Ted Tupper
Code M-311C
Military Sealift Command
Department of the Navy
Washington, DC 20390 1

| | Copies | | Copies |
|---|---|---|---|
| Mr. F. R. Del Priori<br>Code 224<br>Operational Test and Evaluation<br>Force (OPTEVFOR)<br>Norfolk, VA 23511 | 1 | | |

END

DATE
FILMED

5 -83

DTIC