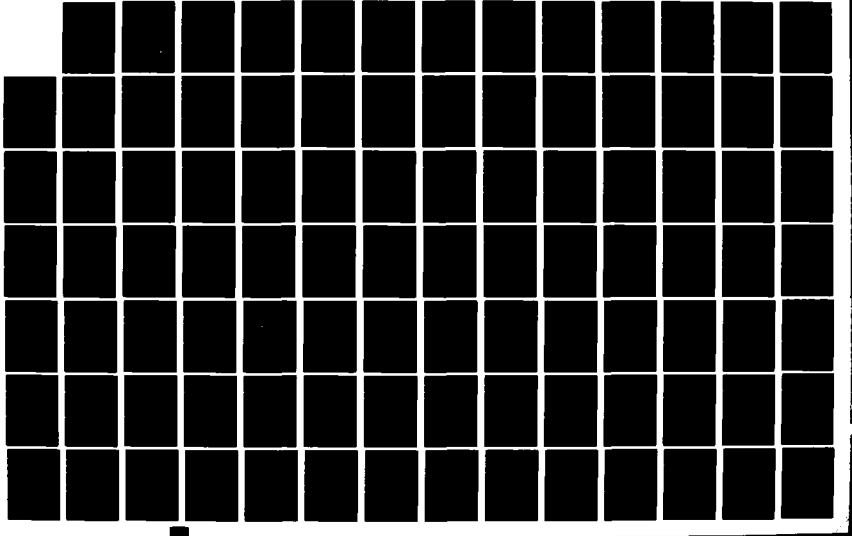
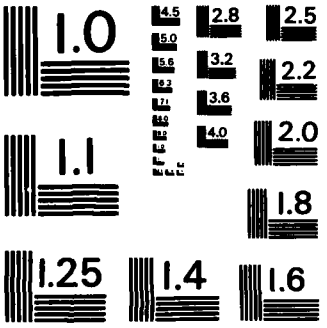
AD-A121 294 THE ANALYSIS OF DESIGN OF ROBUST NONLINEAR ESTIMATORS AND ROBUST SIGNAL C.. (U) PURDUE UNIV LAFAYETTE IN SCHOOL OF ELECTRICAL ENGINEERING N C GALLAGHER 1/2

UNCLASSIFIED 16 SEP 82 AFOSR-TR-82-0933 AFOSR-78-3605 F/G 12/1 NL





UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

(4)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER AFOSR-TR- 82-0933	2. GOVT ACCESSION NO. AD-A121294	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) THE ANALYSIS OF DESIGN OF ROBUST NONLINEAR ESTIMATORS AND ROBUST SIGNAL CODING SCHEMES		5. TYPE OF REPORT & PERIOD COVERED FINAL, 15 JUN 78-14 JUN 82 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Neal C. Gallagher, Jr.		8. CONTRACT OR GRANT NUMBER(s) AFOSR-78-3605
9. PERFORMING ORGANIZATION NAME AND ADDRESS School of Electrical Engineering Purdue University West Lafayette IN 47907		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS PE61102F; 2304/A6
11. CONTROLLING OFFICE NAME AND ADDRESS Directorate of Mathematical & Information Sciences Air Force Office of Scientific Research Bolling AFB DC 20332		12. REPORT DATE 16 SEP 82 13. NUMBER OF PAGES 155
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 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

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)
Source coding; estimation; quantization; median filtering.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)
Two topics of engineering interest have been treated in this research. One is block or vector quantization, which deals with the digital representation of multi-dimensional signals. Two Ph.D. dissertations and one patent application, in addition to numerous technical articles have resulted from this work. The other area of study has been nonlinear signal estimating which has lead to a study of median filtering. This work on median filtering has resulted in two Ph.D. dissertations in addition to a number of technical publications.

DD FORM 1 JAN 73 1473

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

AFOSR-TR- 82-0933

Final Report
Air Force Office of Scientific Research
Grant No.
AFOSR-78-3605

The Analysis of Design of Robust Nonlinear Estimators and Robust Signal Coding Schemes

by

Neal C. Gallagher, Jr.
School of Electrical Engineering
Purdue University
W. Lafayette, IN 47907

I. General

The four year duration of this grant has resulted in twenty-five technical publications in the general areas of signal estimation and source coding. A copy of each publication is found in the Appendix. In addition to these publications, one patent application has been filed dealing with a novel method of multi-dimensional quantization.

In addition to the numerous publications noted above, this project has resulted in the graduation of four Ph.D. students who have been supported in whole or in part through this grant. Two of these dissertations, one by Jim Bucklew and one by Kerry Rines, deal with the analysis and design of block quantizers. The remaining two dissertations by Gonzalo Arce and Tom Nodes treat properties of median filters. A fifth dissertation by Tom McCannon is still being researched. This research concerns the design of nonlinear estimators and predictions. Here we will present a brief description of the technical results; however, the detailed discussion is contained in the attached reprints.

The work on multidimensional quantizers began with a search to find better ways of quantizing multidimensional vectors. We started with a study of vectors with Gaussian distributions and then generalized to circularly symmetric distributions. We developed new derivations for bounds on quantizer performance. Finally, we developed a very simple procedure by which to implement the known optimum quantizer structures. This procedure has been the subject of a patent application.

Our work in nonlinear estimation began with a study of estimation schemes which used an extended form of the projection theorem in their design. We combined polynomial operations with linear operations in the estimator design.

Our work led to an investigation of the properties of the median filters. Our initial interest in the median filter began because of the fact that these median methods really seem to work in many situations where linear estimators are nearly useless. The problem with median filters (and therefore our opportunity) has been the atmost complete lack of theory on their properties and for their design. We have viewed this as a chance to make a significant contribution in this relatively new field of median methods. We believe we have made several major contributions to the analysis of median filters as illustrated by two Ph.D. dissertations and a number of invited technical presentations on the topic of median filters. Copies of these dissertations will be mailed as separate technical reports.

APPENDIX

REPRINTS OF TECHNICAL PAPERS

1. A NOVEL APPROACH FOR DESIGNING NONLINEAR DISCRETE TIME FILTERS: PART I
2. A NOVEL APPROACH FOR DESIGNING NONLINEAR DISCRETE TIME FILTERS: PART II
3. QUANTIZATION OF BIVARIATE CIRCULARLY SYMMETRIC DENSITIES
4. QUANTIZATION IN SPECTRAL PHASE CODING
5. A NOTE ON OPTIMAL QUANTIZATION
6. SOME PROPERTIES OF UNIFORM STEP SIZE QUANTIZERS*
7. ON THE DETERMINATION OF REGRESSION FUNCTIONS
8. QUANTIZATION SCHEMES FOR BIVARIATE GUASSIAN RANDOM VARIABLES
9. TWO-DIMENSIONAL QUANTIZATION OF BIVARIATE CIRCULARLY SYMMETRIC DENSITIES
10. SOME RESULTS IN MULTIDIMENSIONAL QUANTIZATION THEORY*
11. SOME RECENT DEVELOPMENTS IN QUANTIZATION THEORY*
12. PASSBAND AND STOPBAND PROPERTIES OF MEDIAN FILTERS*
13. ROOT-SIGNAL SET ANALYSIS FOR *MEDIAN FILTERS*
14. SOME PROPERTIES OF UNIFORM STEP SIZE QUANTIZERS
15. SOME MODIFICATIONS TO THE MEDIAN FILTER PROCESS AND THEIR PROPERTIES*
16. THE DESIGN OF MULTIDIMENSIONAL QUANTIZERS USING PREQUANTIZATION
17. A NOVEL APPROACH FOR THE COMPUTATION OF ORTHONORMAL POLYNOMIAL EXPANSIONS
18. SOME RESULTS ON THE MEDIAN FILTERING OF SIGNALS AND ADDITIVE WHITE NOISE*
19. ON A CLASS OF RANDOM PROCESSES EXHIBITING OPTIMAL NONLINEAR ONE-STEP PREDICTORS
20. A THEORETICAL ANALYSIS OF THE PROPERTIES OF MEDIAN FILTERS
21. PROPERTIES OF MINIMUM MEAN SQUARED ERROR BLOCK QUANTIZERS
22. NONUNIFORM MULTIDIMENSIONAL QUANTIZATION
23. A NOTE ON THE COMPUTATION OF OPTIMAL MINIMUM MEAN-SQUARE ERROR QUANTIZERS
24. ON THE DESIGN OF NONLINEAR DISCRETE-TIME PREDICTORS
25. THE DESIGN OF TWO-DIMENSIONAL QUANTIZERS USING PREQUANTIZATION

A NOVEL APPROACH FOR DESIGNING NONLINEAR DISCRETE TIME FILTERS: PART I

D. MINOO-HAMEDANI and G.L. WISE
Department of Electrical Engineering
University of Texas at Austin
Austin, Texas 78712

and

N.C. GALLAGHER and T.E. McCANNON
School of Electrical Engineering
Purdue University
West Lafayette, Indiana 49707

ABSTRACT

The problem of minimum mean squared error prediction of a discrete time random process using a nonlinear filter consisting of a zero memory nonlinearity followed by a linear filter is studied. Classes of random processes for which the best predictor is realizable using a nonlinear filter of the above form are discussed. For those random processes for which the best predictor is not realizable using the above nonlinear filter, an iterative procedure is presented for finding a suboptimal nonlinear filter.

I. INTRODUCTION

In this paper we consider a second order random process $\{X_n, n=1,2,\ldots\}$, and we are interested in predicting the random variable X_{N+1} from an observation of $X_1,\ldots,X_N$. Our estimate is denoted by $\hat{X}_{N+1}$, and we wish to choose it so as to minimize the mean squared error.

It is well known [1, pp.77-78] that the optimal estimate of X_{N+1} in terms of $X_1,\ldots,X_N$ is given by the conditional expectation

$$\hat{X}_{N+1} = E\{X_{N+1} \mid X_N,\ldots,X_1\} \quad .$$

In general, this is a Borel measurable function of $X_1,\ldots,X_N$, and in many cases an exact expression for this quantity is difficult to obtain. Often we do not have the necessary statistical information to evaluate such a quantity. Linear estimation has been widely studied [2], and it is well known that the best linear estimate of X_{N+1} given the observations $X_1,\ldots,X_N$ is obtained by applying the Projection Theorem [1, pp.150-155]. It is clear that in this case the only statistical information required is the second moment characteristics of the random process.

In this paper we restrict our estimate $\hat{X}_{N+1}$ to be of a form that is expressible as the output of a system consisting of a zero memory nonlinearity (ZNL) followed by a linear filter. The ZNL is characterized by a Borel measurable function $g(\cdot)$ such that $g(X_1),\ldots,g(X_{N+1})$ are second order random variables. If the weighting sequence of the linear filter is given by $h_0,\ldots,h_{N-1}$, then the estimate is given by

$$\hat{X}_{N+1} = \sum_{n=1}^{N} g(X_n) h_{N-n} \quad . \tag{1}$$

Presented at the Sixteenth Annual Allerton Conference on Communication, Control, and Computing, October 4-6, 1978; to be published in the Proceedings of the Conference.

We wish to determine a function $g(\cdot)$ and a set of coefficients $h_0,\ldots,h_{N-1}$ in such a way that the resulting mean squared error is minimized. With this form of an estimate, we are guaranteed that the performance can be at least as good as that of the optimal linear filter.

In Section II we consider some cases where the optimal estimate has the form of Eq.(1). In the general case the optimal predictor will not have the form of Eq.(1) and thus a predictor of this form will be suboptimal. This situation is discussed in Section III where an iterative scheme is presented for determining suboptimal predictors. In Section IV examples are given to illustrate the method.

II. OPTIMAL PREDICTION

In this section we consider some cases where the optimal filter has the form of Eq.(1). Whenever the optimal filter is linear, then it obviously has the form of Eq.(1) with $g(x)=x$. The class of spherically invariant random processes [3] admits linear solutions, with the most well-known examples being the Gaussian processes.

It is clear that the performance of the filter given by Eq.(1) can always be made at least as good as that of the optimal linear filter. In some cases the filter given by Eq.(1) can be optimal while the optimal linear filter is useless. For example, let $X_n=P_n(U)$ where U is a random variable uniformly distributed over $[-1,1]$ and $P_n(\cdot)$ is the n-th Legendre polynomial. In this case, the sequence $\{X_n,\ n=1,2,\ldots\}$ is a sequence of uncorrelated zero mean random variables and the optimal linear filter yields an estimate which is zero. However, for $g(x)=P_{N+1}(x)$ and

$$h_n = \begin{cases} 1, & n=0 \\ 0, & n\neq 0 \end{cases},$$

the filter of Eq.(1) gives the estimate $\hat{X}_{N+1}=X_{N+1}$. Numerous examples similar to this can easily be constructed.

When the process is a (first order) Markov process it is well known [1, pp.81-83] that $E\{X_{N+1}|X_N,\ldots,X_1\} = E\{X_{N+1}|X_N\}$, with probability one (wp1). Thus a system of the form of Eq.(1) with a ZNL given by $g(x)= E\{X_{N+1}|X_N = x\}$ and a weighting sequence given by

$$h_n = \begin{cases} 1, & n=0 \\ 0, & n\neq 0 \end{cases}$$

will yield the optimal estimate of X_{N+1}.

Markov processes serve as the model of many physical phenomena that arise in practice. Often they are obtained as the solution of first order stochastic difference equations of the form

$$X_{n+1} = g(X_n) + Z_{n+1}, \quad n=0,1,2,\ldots$$

where $g(\cdot)$ is a Borel measurable function and the sequence $\{Z_n\}$ is a sequence of zero mean independent random variables independent of the initial condition X_0. It is easily seen that in this case we will have $E\{X_{N+1}|X_N,\ldots,X_1\}=g(X_N)$ wp1.

It is clear that for any random process for which

$$E\{X_{N+1}|X_N,\ldots,X_1\} = \sum_{n=1}^{N} g(X_n)h_{N-n} \quad \text{wpl}, \tag{2}$$

a system of the form of Eq.(1) will produce the optimal estimate of X_{N+1}. As another example of a process for which the conditional expectation has the form of Eq.(2) consider the process generated by the following second order stochastic difference equation:

$$X_{n+2} = h_0 g(X_{n+1}) + h_1 g(X_n) + Z_{n+2}\ , \quad n=-1,0,1,2,\ldots, \tag{3}$$

where $g(\cdot)$ is a Borel measurable function and $\{Z_n\}$ is a sequence of zero mean independent random variables independent of the initial conditions X_{-1} and X_0. It can be easily seen that for this example, for any $N\geq 2$,

$$E\{X_{N+1}|X_N,\ldots,X_1\} = h_0 g(X_N) + h_1 g(X_{N-1}) \quad \text{wpl}.$$

Extension of this example to the case where Eq.(3) is a k-th order stochastic difference equation is obvious.

To obtain a characterization of a random process for which a form of Eq.(2) holds, we use a theorem due to Balakrishnan [4].

Theorem (Balakrishnan): Let $C_{N+1}(t_1,\ldots,t_{N+1})$ denote the joint characteristic function of the random variables $X_1,\ldots,X_{N+1}$. Assume that the moments of all orders of the random variables exist, so that $C_{N+1}(\ldots)$ has derivatives of all orders. Let D_k denote the differential operator $\partial(\cdot)/\partial it_k$, so that

$$D_k C_{N+1}(t_1,\ldots,t_{N+1}) = \frac{\partial}{\partial it_k} C_{N+1}(t_1,\ldots,t_{N+1})\ .$$

Let $P(x_1,\ldots,x_N)$ be a polynomial in N variables. Then a necessary and sufficient condition for

$$E\{(X_{N+1})^M|X_N,\ldots,X_1\} = P(X_1,\ldots,X_N) \quad \text{wpl}$$

is that

$$\frac{\partial^M}{\partial(it_{N+1})^M} C_{N+1}(t_1,\ldots,t_{N+1})\Big|_{t_{N+1}=0} = P(D_1,\ldots,D_N)\cdot C_{N+1}(t_1,\ldots,t_N,0).$$

Now, in the above theorem let M=1 and let $g(\cdot)$ be a polynomial of degree d, i.e.

$$g(x) = \sum_{j=0}^{d} a_j x^j\ , \tag{4}$$

and assume $P(x_1,\ldots,x_N)$ has the form

$$P(x_1,\ldots,x_N) = \sum_{n=1}^{N} h_{N-n} g(x_n) = \sum_{n=1}^{N} \sum_{j=0}^{d} h_{N-n} a_j (x_n)^j \quad .$$

Assume that the random variables in the process possess moments of all orders. Then a necessary and sufficient condition for Eq.(2) to hold, where $g(\cdot)$ is given by Eq.(4), is that

$$\frac{\partial}{\partial(it_{N+1})} C_{N+1}(t_1,\ldots,t_{N+1})\Big|_{t_{N+1}=0} = \sum_{n=1}^{N} \sum_{j=0}^{d} h_{N-n} a_j D_n^j C_{N+1}(t_1,\ldots,t_N,0) \quad .$$

This result is of limited practical usefulness, because one often does not have the necessary statistical information available.

III. SUBOPTIMAL PREDICTION

In the general case there will not exist a function $g(\cdot)$ and a weighting sequence $h_0,\ldots,h_{N-1}$ such that Eq.(2) is satisfied. However, it is quite reasonable to conjecture that in many cases it may be possible to determine a filter having the form of Eq.(1) with a mean squared error either significantly smaller than that associated with the optimal linear filter or very close to the mean squared error associated with the optimal filter.

Once we assume that the function $g(\cdot)$ that minimizes the mean squared error is known, the $g(X_n)$'s will be well defined random variables and the determination of the h_n's that minimize the mean squared error reduces to an application of the Projection Theorem, i.e. setting

$$E\left\{\left[X_{N+1} - \sum_{n=1}^{N} h_{N-n} g(X_n)\right] g(X_j)\right\} = 0 \quad , \quad j=1,\ldots,N,$$

and solving for the h_n's. To carry out this step we need to calculate the terms $E\{g(X_n)g(X_j)\}$ and $E\{X_{N+1}g(X_j)\}$. The difficult problem is the determination of the function $g(\cdot)$ that minimizes the mean squared error.

Notice that, in the optimization problem where the filter is constrained to be of the form in Eq.(1), only second order information (i.e. the family of bivariate distributions) is required. This is more statistical information than is required if we were doing optimal linear filtering, which requires second moment information. However, it is still considerably less statistical information than is required if we were doing optimal filtering, which requires statistical information pertaining to an (N+1)-st dimensional distribution.

In order to circumvent the difficult problem of determining the function $g(\cdot)$ to use in Eq.(1), we will parameterize $g(\cdot)$ and thus let the determination of $g(\cdot)$ simply depend upon finding the correct parameters. Doing so, we would then write the resulting mean squared error as a function of the parameters associated with $g(\cdot)$ and the weighting sequence of the linear filter. In this case, the mean squared error would be a function of K+N parameters, where K is the number of parameters associated with $g(\cdot)$. For example, let $g(\cdot)$ be given by

$$g(x) = \sum_{j=1}^{K} a_j b_j(x) \quad .$$

Then our estimate is given by

$$\hat{X}_{N+1} = \sum_{n=1}^{N} \sum_{j=1}^{K} h_{N-n} a_j b_j(X_n) \quad ,$$

and the resulting mean squared error is given by

$$E\left\{\left[X_{N+1}-\hat{X}_{N+1}\right]^2\right\} = E\left\{\left[X_{N+1}\right]^2\right\} - 2\sum_{n=1}^{N}\sum_{j=1}^{K} h_{N-n} a_j E\{X_{N+1} b_j(X_n)\} \qquad (5)$$

$$+ \sum_{n=1}^{N}\sum_{m=1}^{N}\sum_{j=1}^{K}\sum_{k=1}^{K} h_{N-n} h_{N-m} a_j a_k E\{b_j(X_n) b_k(X_m)\} \quad .$$

The functions $b_j(\cdot)$ should be determined so that there is considerable flexibility in the functional form of $g(\cdot)$ and also so that the expectations in Eq.(5) could be determined from the statistical information at hand. For example, if $b_j(x)=x^j$, then the necessary statistical information would consist of the higher order joint moments.

The next step might be to minimize Eq.(5) over the N+K parameters. This would result in N+K equations of third order polynomials in the parameters. This simultaneous optimization over all the parameters presents potential numerical problems. As an alternative to the simultaneous optimization over all the parameters, we will now describe an iterative technique.

The basic plan of the iterative technique is to consider the two sets of parameters separately and to iteratively optimize over one set of parameters while holding the other set fixed. This iterative technique results in the need to solve systems of linear equations, as opposed to the need to solve systems of equations in third order polynomials such as encountered in the effort to simultaneously optimize over all the parameters.

We will assume that the parametric form of $g(\cdot)$ is such that with the proper choice of parameters we could have $g(x)=x$. In this way the mean squared error that results will always be upper bounded by the mean squared error associated with the optimal linear filter.

The iterative technique is as follows:

Step 1. Determine the optimal weighting sequence $h_0,\ldots,h_{N-1}$ for the case where $g(x)=x$.

Step 2. Evaluate the resulting mean squared error.

Step 3. For this choice of $h_0,\ldots,h_{N-1}$, determine $a_1,\ldots,a_K$ so as to minimize the mean squared error.

Step 4. For this choice of $a_1,\ldots,a_K$, determine the optimal weighting sequence $h_0,\ldots,h_{N-1}$.

Step 5. Repeat Steps 3 and 4 until the improvement in the mean squared error is negligible.

The $a_1,\ldots,a_K$ and $h_0,\ldots,h_{N-1}$ that are obtained in Step 5 after the termination of the iterations determine the system. Step 1 and Step 4 make use of the Projection Theorem and result in $E\{X_{N+1}g(X_j)\}$ =

$\sum_{n=1}^{N} h_{N-n} E\{g(X_n)g(X_j)\}$, $j=1,\ldots,N$. Step 2 makes use of Eq.(5). Step 3 also makes use of Eq.(5) and results in

$$\sum_{n=1}^{N}\sum_{m=1}^{N} h_{N-n}h_{N-m}\left[2a_j E\{b_j(X_n)b_j(X_m)\} + \sum_{\substack{k=1\\k\neq j}}^{K} a_k E\{b_j(X_n)b_k(X_m)\}\right]$$
$$= 2\sum_{n=1}^{N} h_{N-n} E\{X_{N+1}b_j(X_n)\}\ ,\quad j=1,\ldots,K.$$

IV. EXAMPLES

In this section we consider a particular parametric form for the ZNL and a specific model for the random sequence. The iterative method described earlier is used in this case to determine a filter of the form of Eq.(1). We also determine the mean squared error resulting from use of the optimal filter and that resulting from use of the optimal linear filter. Performances of the filters are compared and it is seen that in several instances the improvement in mean squared error of the suboptimal filter over that of the optimal linear filter is a significant fraction of the corresponding improvement of the optimal filter over that of the optimal linear filter.

Assume that we have knowledge of the regression function

$$r(x) = E\{X_{N+1}|X_N=x\}\ . \tag{6}$$

Notice that if we choose $g(x)=r(x)$ and

$$h_n = \begin{cases} 1, & n=0 \\ 0, & n\neq 0 \end{cases}\ ,$$

then the estimate would be the same as that of the optimal filter based on the most recent observation. If we were to use the Projection Theorem to choose a different weighting sequence $\{h_n\}$, we might do better. It seems reasonable to expect that if we were to parameterize $g(\cdot)$ in such a way that by proper choice of the parameters we would have $g(x)=r(x)$, and then use this parameterization of the ZNL in the iterative technique described earlier, we might determine a system of the form of Eq.(1) exhibiting very good performance. This is how we will choose the ZNL in this section.

As a model for the random sequence $\{X_n, n=1,2,\ldots\}$ we will assume that

$$X_n = \left(Z_n\right)^{2k+1} \tag{7}$$

where $\{Z_n, n=1,2,\ldots\}$ is a zero mean stationary Gaussian process with unit variance and autocorrelation function $\rho(\cdot)$.

First we will derive an expression for the regression function of Eq.(6) when the random sequence is given by Eq.(7). Using results in [5], we have that

$$E\{X_{N+1}|X_N\} = E\left\{(Z_{N+1})^{2k+1}|Z_n\right\}$$
$$= \sum_{n=0}^{\infty}[\rho(1)]^n\, b_n\theta_n(Z_N)$$
$$= \sum_{n=0}^{\infty}[\rho(1)]^n\, b_n\theta_n\left((X_N)^{1/(2k+1)}\right)\ ,$$

where the series are mean square convergent, the constants $\{b_n\}$ are given by

$$b_n = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (x)^{2k+1} \theta_n(x) \exp\left(\frac{-x^2}{2}\right) dx \quad , \tag{8}$$

and θ_n is the n-th normalized Hermite polynomial given by

$$\theta_n(x) = \frac{(-1)^n}{\sqrt{n!}} \exp\left(\frac{x^2}{2}\right) \frac{d^n}{dx^n} \exp\left(\frac{-x^2}{2}\right) \quad .$$

We see from Eq.(8) that $b_n=0$ for $n>2k+1$ and, in fact, the b_n's can be obtained from the relation

$$(x)^{2k+1} = \sum_{n=0}^{2k+1} b_n \theta_n(x) \quad .$$

For example, for k=1,

$$b_n = \begin{cases} 3, & n=1 \\ \sqrt{6}, & n=3 \\ 0, & n\neq 1,3 \end{cases}$$

and r(x) is given by $r(x) = [\rho(1)]^3 x + 3\rho(1)\left(1-[\rho(1)]^2\right) x^{1/3}$.
For k=2,

$$b_n = \begin{cases} 15, & n=1 \\ 10\sqrt{6}, & n=3 \\ 2\sqrt{30}, & n=5 \\ 0, & n\neq 1,3,5 \end{cases}$$

and

$$r(x) = [\rho(1)]^5 x + 10[\rho(1)]^3 \left(1-[\rho(1)]^2\right) x^{3/5} + 15\rho(1)\left(1-[\rho(1)]^2\right)^2 x^{1/5}.$$

In general, for an arbitrary positive integer k, it is easily seen that r(·) has the form

$$r(x) = c_{k+1} x + c_k (x)^{(2k-1)/(2k+1)} + c_{k-1} (x)^{(2k-3)/(2k+1)} + \ldots + c_1 (x)^{1/(2k+1)} \quad ,$$

where the c_i's are constants that can be determined using the above procedure. Thus we choose the ZNL g(·) to be

$$g(x) = \sum_{i=1}^{k+1} a_i (x)^{(2i-1)/(2k+1)}$$

where the parameters a_i are to be determined by the iterative procedure. In utilizing the iterative procedure we encounter the need for the knowledge of moments and joint moments of $\{Z_n\}$ (see [6]), which are given by

$$E\left\{(Z_n)^p\right\} = \begin{cases} 1\cdot 3\cdot 5 \ldots (p-1) & \text{for p even} \\ 0 & \text{for p odd} \end{cases}$$

$$E\left\{(Z_n)^r(Z_{n+i})^s\right\} = \mu(r,s,i) = \begin{cases} (r+s-1)\rho(i)\mu(r-1,s-1,i)+(r-1)(s-1)\cdot \\ \qquad \left(1-[\rho(i)]^2\right)\mu(r-2,s-2,i) \text{ when} \\ \qquad (r+s) \text{ is even} \\ 0 \quad \text{when } (r+s) \text{ is odd} \quad . \end{cases} \tag{9}$$

Observing that $\mu(1,1,i)=\rho(i)$ and $\mu(2,2,i)=1+2[\rho(i)]^2$, all higher order joint moments can be calculated using Eq.(9).

In order to compare the performance of the suboptimal estimator with that of the optimal estimator, we have obtained expressions for the mean squared error associated with the optimal estimator. For the optimal system we are interested in

$$E\left\{(Z_{N+1})^{2k+1} | Z_N,\ldots,Z_1\right\} \quad .$$

Notice that this is the (2k+1)-st conditional moment and the conditional distribution has the functional form of a Gaussian distribution. Thus the minimum mean squared error follows using standard properties of the Gaussian distribution (see, for example, [7]). For k=1 we find that the minimum mean squared error is given by $15 - P_1^4\left[9E\{Y^2\} + 6P_1E\{Y^4\} + P_1^2\,E\{Y^6\}\right]$; and for k=2, the minimum mean squared error is given by $945 - P_1^6\left[225\,E\{Y^2\} + 300P_1E\{Y^4\} + 130P_1^2E\{Y^6\} + 20P_1^3E\{Y^8\} + P_1^4E\{Y^{10}\}\right]$. In these expressions P_1 is a constant and Y is a normal random variable with zero mean and variance γ^2. The constants P_1 and γ^2 are defined as follows. Assume without loss of generality that the correlation matrix R associated with $Z_1,\ldots,Z_{N+1}$ is positive definite (if it is not, the data can be reduced to achieve this result). Then P_1 is the reciprocal of the element in the lower right corner of R^{-1}. Denote the first N elements in the last row of R^{-1} as $r_1,\ldots,r_N$. Then

$$\gamma^2 = \sum_{i=1}^{N}(r_i)^2 + 2\sum_{m=1}^{N-1}\sum_{n=1}^{m} r_{N-n+1}r_{m-n+1}\rho(N-m) \quad .$$

The mean squared error associated with the optimal linear filter can be obtained in a straightforward fashion.

In the following tables results are presented comparing the suboptimal filter to the optimal filter and the optimal linear filter. Several correlation sequences for $\{Z_n\}$ are considered, both the third power and the fifth power of Z_n are used as models, and examples for two observations and five observations are given. In these tables L_1, L, and L_{min} are the mean squared errors resulting from the optimal linear filter, suboptimal filter using a ZNL, and the optimal filter, respectively. The quantity η_1 is the percent of decrease in L_1 when the suboptimal filter using a ZNL is employed, i.e. $\eta_1 = 100(L_1-L)/L_1$. The quantity η_2 is the percent of possible improvement in L_1 using the optimal filter, i.e. $\eta_2 = 100(L_1-L_{min})/L_1$. The quantity η_3 is the normalized percent of improvement over the linear filter given by the suboptimal filter using a ZNL, i.e. $\eta_3 = 100\,\eta_1/\eta_2 = 100(L_1-L)/(L_1-L_{min})$.

	$\rho(1)$	$\rho(2)$	$\rho(3)$	$\rho(4)$	$\rho(5)$
1	.75	.575	.45	.35885	.291
2	.885	.7887	.70762	.639	.5805
3	.55	.315	.187	.11445	.07183
4	.55	.395	.319	.26885	.23023
5	.425	.2375	.14675	.09448	.06207
6	.8333	.6666	.5	.3333	.1666
7	.5787	.2963	.125	.037	.00463
8	.4822	.1975	.0625	.0123	.00077

Table 1. Correlation sequences corresponding to Tables 2-5.

	L_1	L	L_{min}	η_1	η_2	η_3
1	9.1983	8.8614	8.8581	3.6	3.69	97.3
2	5.1744	5.0622	5.0599	2.16	2.21	97.6
3	12.5987	12.1084	12.108	3.89	3.89	99.8
4	12.3196	11.9216	11.8952	3.23	3.44	93.7
5	13.6849	13.2957	13.293	2.84	2.86	99.1
6	6.9247	6.6228	6.4926	4.36	6.23	69.8
7	12.2903	11.732	11.7259	4.54	4.59	98.8
8	13.3219	12.8142	12.8123	3.81	3.82	99.6

Table 2. Mean squared errors and percentages of improvement for k = 1.

	L_1	L	L_{min}	η_1	η_2	η_3
1	727.42	704.58	704.22	3.13	3.18	98.1
2	453.78	444.78	444.49	1.98	2.04	96.7
3	887.49	859.95	859.9	3.1	3.1	99.7
4	879.44	854.59	851.86	2.82	3.13	89.8
5	920.93	899.7	899.43	2.3	2.33	98.5
6	584.57	564.58	550.99	3.41	5.74	59.3
7	876.33	845.86	845.24	3.47	3.54	97.7
8	910.86	884.62	884.42	2.88	2.9	99.2

Table 3. Mean squared errors and percentages of improvement for k = 2.

	h_0	h_1	h_2	h_3	h_4	a_1	a_2
1	.6115	.0127	.008	.0059	.0094	1.519	.6811
2	.7899	.0084	.0064	.0051	.0132	.674	.862
3	.4026	.0093	.0049	.0029	.0024	2.7896	.4114
4	.3654	.0687	.0433	.0297	.028	2.4749	.4334
5	.2827	.0407	.0164	.0076	.0047	3.3779	.2656
6	.776	-.0234	-.0175	-.0111	-.0662	1.2015	.7635
7	.4476	-.028	-.018	-.01	-.0015	2.775	.4375
8	.3505	-.0247	-.0121	-.0032	-.0017	3.342	.3215

Table 4. The coefficients a_i of the nonlinearity $g(x) = a_2x + a_1\sqrt[3]{x}$ and the h_i's of the suboptimal system for k = 1.

	h_0	h_1	h_2	h_3	h_4	a_1	a_2	a_3
1	.4779	.0119	.0063	.0042	.0052	4.0527	3.7727	.493
2	.7065	.0097	.0067	.005	.0093	.7136	2.032	.7585
3	.2563	.0059	.0028	.0017	.0014	15.173	4.5019	.196
4	.2466	.0472	.0282	.019	.017	11.733	4.465	.1966
5	.162	.0227	.009	.0043	.0026	23.858	3.802	.0839
6	.6534	-.034	-.0234	-.0136	-.024	2.742	2.9562	.6302
7	.2864	-.0184	-.0096	-.0054	.0002	14.7841	4.5769	.2267
8	.2032	-.0139	-.0065	-.0019	.0008	22.373	4.2663	.128

Table 5. The coefficients a_i of the ZNL $g(x) = a_3x + a_2x^{3/5} + a_1x^{1/5}$ and the h_i's of the suboptimal system for k = 2.

	$\rho(1)$	$\rho(2)$
1	.9	.7
2	.8	.5
3	.8	.3
4	.7	.1

Table 6. Correlation sequences corresponding to Tables 7-8.

	h_0	h_1	a_1	a_2
1	1.2377	-.4974	.9333	.82983
2	.8837	-.3001	1.6639	.6923
3	1.095	-.6467	2.3987	.6089
4	.7927	-.4786	3.2982	.4545

Table 7. The coefficients a_i of the ZNL $g(x) = a_2x + a_1x^{1/3}$ and the h_i's of the suboptimal system for k = 1.

	I_1	I_∞	I_{min}	η_1	η_2	η_3
1	3.7487	3.494	3.1354	6.79	16.3	41.65
2	7.566	7.0273	6.7406	7.12	10.9	65.32
3	5.7804	4.371	1.0231	24.38	82.3	29.62
4	8.9825	7.1689	4.9674	20.19	44.7	45.16

Table 8. Mean squared errors and percentages of improvement for k = 1.

ACKNOWLEDGEMENT

D. Minoo-Hamedani was supported by the Air Force Office of Scientific Research under Grant AFOSR-76-3062 and by the Department of Defense Joint Services Electronics Program under Contract F49620-77-C-0101. G. L. Wise was supported by the Air Force Office of Scientific Research under Grant AFOSR-76-3062. N. C. Gallagher and T. E. McCannon were supported by the Air Force Office of Scientific Research under Grant AFOSR-78-3605.

REFERENCES

1. J.L. Doob, Stochastic Processes, John Wiley, New York, 1953.
2. T. Kailath, ed., Linear Least-Squares Estimation, Dowden, Hutchinson, and Ross, Stroudsburg, Pennsylvania, 1977.
3. I.F. Blake and J.B. Thomas, "On a Class of Processes Arising in Linear Estimation Theory," IEEE Trans. Inform. Theory, Vol. IT-14, pp.12-16, January 1968.
4. A.V. Balakrishnan, "On a Characterization of Processes for which Optimal Mean-Square Systems are of Specified Form," IRE Trans. Inform. Theory, Vol. IT-6, pp.490-500, September 1960.
5. G.L. Wise and J.B. Thomas, "A Characterization of Markov Sequences," Journal of the Franklin Institute, Vol. 299, pp.269-278, April 1975.
6. N.L. Johnson and S. Kotz, Distributions in Statistics: Continuous Multivariate Distributions, p.91, John Wiley, New York, 1972.
7. K.S. Miller, Multidimensional Gaussian Distributions, pp.21-22, John Wiley, New York, 1964.

A NOVEL APPROACH FOR DESIGNING NON-LINEAR DISCRETE TIME FILTERS: PART II

T.E. MCCANNON & N.C. GALLAGHER
School of Electrical Engineering
Purdue University
W. Lafayette, IN 47907

G.L. WISE & D. MINOO-HAMEDANI
Department of Electrical Engineering
University of Texas
Austin, Texas 78712

ABSTRACT

We propose two methods for designing nonlinear discrete time filters. The first method involves an iteration procedure that for simple cases, converges in one or two iterations. However, convergence problems in this approach for higher (> 1) time order filters leads to a second method which is based upon an augmented Hilbert subspace on which the orthogonality principle can be easily applied.

I. INTRODUCTION

In many problems, it can be shown that a non-linear filter either outperforms the linear filter or performs a function not possible with a linear filter. One example of this is the homomorphic processing of speech which utilizes a linear process followed by a non-linearity that is followed by another linear processor [1].

This paper is concerned with the non-linear prediction problem. We consider the system shown in Fig. 1,

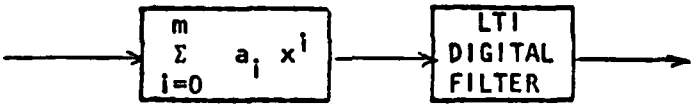


Fig. 1. Non-linear System Under Study.

where we investigate two different methods of design. In Section II, we consider an iterative scheme and give examples of its use. In Section III, we develop a new non-iterative technique motivated by the poor performance of the iterative scheme found in several non-trivial examples. It is also worthwhile to point out that in Part I of this paper, results are obtained based on complete knowledge of the process statistics, while in Part II we only assume that we have a finite sequence of samples from the random process. We also require the random process to be Wide Sense Stationary (WSS) and to have finite higher order moments.

II. ITERATIVE FILTER DESIGN PROCEDURE

We propose the following iterative procedure for determining the MMSE filter coefficients for the system of Fig. 1.

(1) Assume the non-linearity is not present and design the optimum (Wiener) linear filter.
(2) Keeping the unit pulse response of the linear filter constant, compute the polynomial coefficients required to minimize the mse.
(3) With the polynomial coefficients fixed, redesign the optimum linear filter with the polynomial non-linearity.
(4) Repeat Steps (2) and (3) until convergence.

Presented at the Sixteenth Annual Allerton Conference on Communication, Control, and Computing, October 4-6, 1978.

Consider the example where we have a second degree polynomial and a first order linear filter as shown in Fig. 2.

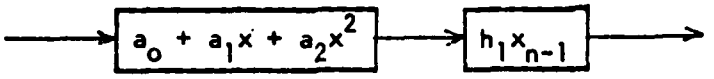


Fig. 2. Example of Non-linear System.

Step 1 tells us to design the optimum linear filter assuming the non-linearity is not present. This linear predictor is given by

$$\hat{x}_n = h_1 \, x_{n-1}$$

where $\hat{x}_n \equiv$ estimate of the n^{th} sample of the r.p. $x(t)$

$x_{n-j} \equiv$ actual value of the $(n-j)^{th}$ sample of $x(t)$

Using the orthogonality principle

$$E\{(x_n - h_1 \, x_{n-1}) \, x_{n-1}\} = 0$$

we find the optimum linear filter to be

$$h_1 = \frac{R_x(1)}{R_x(0)} \tag{1}$$

where $R_x(j) \equiv E\{x_n x_{n-j}\}$ and we have made use of the fact that the r.p. is W.S.S.

Step 2 tells us to compute a_o, a_1, and a_2 to minimize the mse keeping h_1 constant. This non-linear predictor is given by

$$\hat{x}_n = h_1 \, [a_o + a_1 x_{n-1} + a_2 x_{n-1}^2] \tag{2}$$

Using this expression in the mse equation

$$\text{mse} = E\{(\hat{x}_n - x_n)^2\}$$

we have that

$$\text{mse} = E\{[h_1(a_o + a_1 x_{n-1} + a_2 x_{n-1}^2) - x_n]^2\}$$

We minimize the mse with respect to the filter coefficients a_o, a_1, and a_2 by taking partial derivatives

$$\frac{\partial \text{mse}}{\partial a_o} = 0, \quad \frac{\partial \text{mse}}{\partial a_1} = 0, \quad \frac{\partial \text{mse}}{\partial a_2} = 0$$

Then the coefficients a_o, a_1 and a_2 that satisfy the following set of matrix equations are computed:

$$\begin{bmatrix} E\{x_n^2 x_{n-1}\} \\ E\{x_n x_{n-1}\} \\ E\{x_n\} \end{bmatrix} = \begin{bmatrix} h_1 \, E\{x_{n-1}^4\} & h_1 \, E\{x_{n-1}^3\} & h_1 \, E\{x_{n-1}^2\} \\ h_1 \, E\{x_{n-1}^3\} & h_1 \, E\{x_{n-1}^2\} & h_1 \, E\{x_{n-1}\} \\ h_1 \, E\{x_{n-1}^2\} & h_1 \, E\{x_{n-1}\} & h_1 \end{bmatrix} \begin{bmatrix} a_2 \\ a_1 \\ a_0 \end{bmatrix} \tag{3}$$

Step 3 tells us to compute the new optimum linear filter with a_o, a_1 and a_2 constant. The orthogonality principle together with Eq. (2) gives

$$E\{[x_n - h_1 \, (a_2 x_{n-1}^2 + a_1 x_{n-1} + a_0)] \, x_{n-1}\} = 0,$$

and we find the new linear filter to be

$$h_1' = \frac{E\{x_n x_{n-1}\}}{a_2 E\{x_{n-1}^3\} + a_1 E\{x_{n-1}^2\} + a_o E\{x_{n-1}\}} \tag{4}$$

If we solve Eq. (3) for a_o, a_1 and a_2 and substitute these values into Eq. (4) we find

$$h_1' = \frac{R_x(1)}{R_x(0)} = h_1 \tag{5}$$

where the second equality follows from Eq. (1). It is seen that for the non-linear predictor of Fig. 2, the iteration procedure converges in one iteration for an arbitrary W.S.S. r.p. x(t) with finite first, second and third order moments. As an example, consider the following signal

$$x_n = k_1 x_{n-1}^2 + k_2 + p_1 u_n$$

where u_n are iid, uniform $(-\frac{1}{2}, \frac{1}{2})$. We can easily obtain the optimum MMSE predictor [2], [3] for this signal by utilizing the conditional expectation

$$\begin{aligned} \hat{x}_n &= E\{x_n | x_{n-1}, x_{n-2}, \cdots\} \\ &= k_1 x_{n-1}^2 + k_2 \end{aligned} \tag{6}$$

since $E\{u_n | x_{n-1}, x_{n-2}, \cdots\} = 0$. If we let $p_1 = .005$, $k_1 = -1.74$ and $k_2 = 0.87$, the linear filter gives a mse = .293. Calculating the required moments needed for the solution of Eq. (3) empirically with a computer, we find the values of a_o, a_1 and a_2 to be

$$\begin{aligned} a_o &= -3.960261 \\ a_1 &= .005525 \\ a_2 &= 7.913872 \end{aligned}$$

By computer simulation, we find that the non-linear system gives a mse = 2×10^{-6}, a significant improvement over that obtained with the linear filter alone. It is possible to analytically solve for the optimum predictor coefficients; we find the values of a_o, a_1 and a_2 by equating Eqs. (2) and (6) and also using Eq. (1). The values are found to be

$$\begin{aligned} a_o &= -3.960017 \\ a_1 &= 0 \\ a_2 &= 7.920035 \end{aligned}$$

This result agrees very well with the computer simulation.

Next, consider the example as shown in Fig. 3.

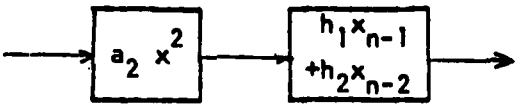


Fig. 3. Example of Non-linear System.

We again apply the iteration procedure as outlined by steps (1), (2) and (3) above. Applying the signal

$$x_n = k_1 x_{n-1}^2 + p_1 u_n \qquad (7)$$

where $\{u_n\}$ are iid, uniform $(-\frac{1}{2}, \frac{1}{2})$, we can easily show the procedure terminates after two iterations. However, if we use a general second order polynomial

$$a_2x^2 + a_1x + a_o$$

simulation with the signal of Eq. (7) indicates that convergence is very slow unless the initial choices of a_2, a_1, a_o, h_1 and h_2 are close to the optimum solutions, and then convergence occurs in 2 to 3 iterations. Simulation also shows strong dependence of the final solution on the initial choices of a_2, a_1, a_o, h_1 and h_2. In an attempt to force the solution to the optimum result for the general case, we propose the following modified iteration procedure.

(1) Set $h_1 = h_2 = \cdots = h_k = 1$

(2) Compute the polynomial coefficients required to minimize the mse.

(3) Design the optimum linear filter using the polynomial non-linearity.

(4) Repeat Steps (2) and (3) until convergence.

But even with the modified procedure, simulation indicates sluggish convergence. Fig. 4 demonstrates this convergence with plots of mse versus number of iterations for the polynomial non-linearity $a_2 x^2 + a_1 x + a_o$.

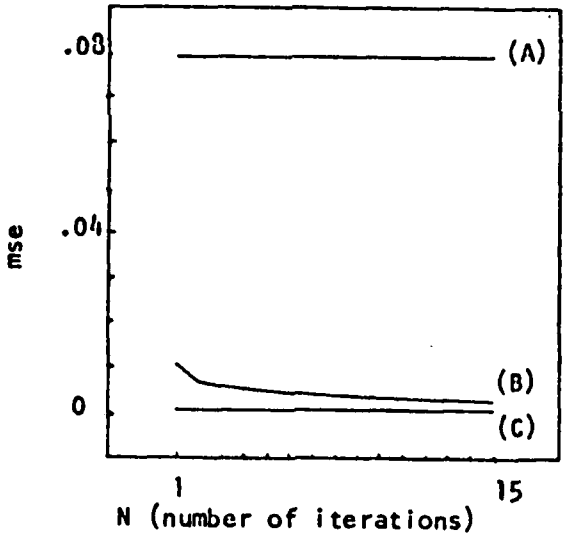


Fig. 4. Demonstration of Convergence. Curve (A) shows mse vs N where the initial values of h_1 and h_2 are the optimum linear filter coefficients. Curve (B) shows mse vs. N for the initial values $h_1 = h_2 = 1$. Curve (C) shows mse vs N where the initial values of h_1 and h_2 are the optimum non-linear coefficients.

It appears that this method only works well for very simple structures and for more general cases another type of design procedure is required.

III. NON ITERATIVE FILTER DESIGN METHOD

In Section II, we have studied an iterative procedure for the design of the non-linear predictor in Fig. 1. This system leads to a prediction better than that obtained from the linear predictor, although in many cases

the improvement is not significant. We would, however, like to retain the basic structure of Fig. 1, which can also be implemented as shown in Fig. 5 and 6.

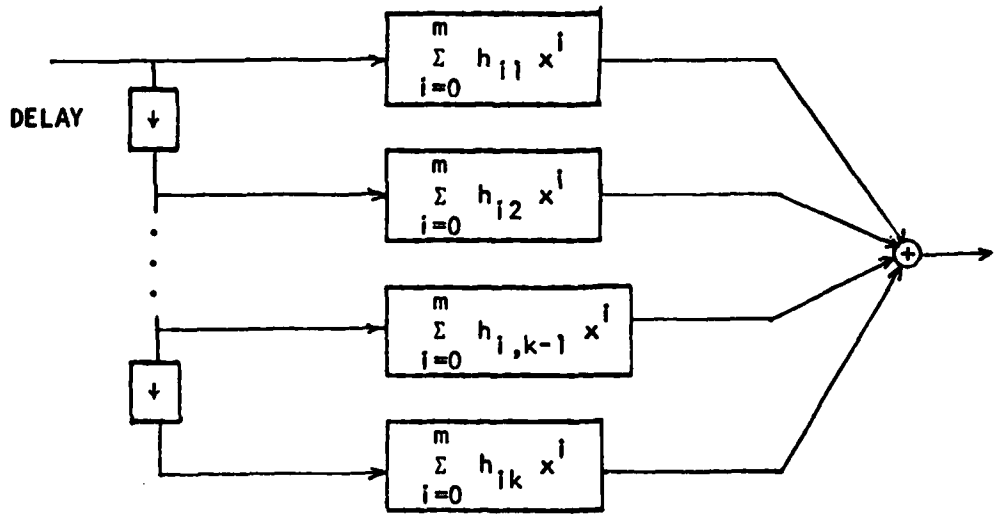


Fig. 5. Alternate Structure to the Non-linear System of Fig. 1.

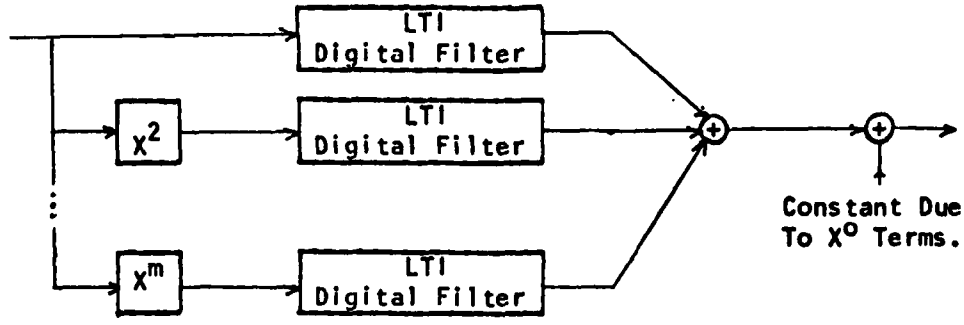


Fig. 6. Equivalent Structure to the Non-linear System of Fig. 5.

We can now express the non-linear predictor of Fig. 5 as

$$\hat{x}_n = \sum_{j=1}^{k} \left[\sum_{i=0}^{m} h_{ij} x_{n-j}^{i} \right] \tag{8}$$

Defining

$$h_o \triangleq \sum_{j=1}^{k} h_{oj}$$

where the h_{oj} are the constants multiplying the x_{n-j}^{o} terms, we can also write Eq. (8) as

$$\hat{x}_n = h_o + \sum_{j=1}^{k} \sum_{i=1}^{m} h_{ij} x_{n-j}^{i} \tag{9}$$

We now minimize the mse with respect to the coefficients h_{ij} and h_o where

$$\text{mse} = E\{(\hat{x}_n - x_n)^2\} \tag{10}$$

by setting

$$\frac{\partial \text{mse}}{\partial h_o} = 0, \quad \frac{\partial \text{mse}}{\partial h_{ij}} = 0, \qquad i = 1,2,\ldots,m \quad j = 1,2,\ldots,k$$

Consider the structure where m = h = 2. Substituting m = h = 2 into Eq. (9) we have the non-linear predictor

$$x_n = h_o + h_{11}x_{n-1} + h_{21}x_{n-1}^2 + h_{12}x_{n-2} + h_{22}x_{n-2}^2 \tag{11}$$

When we substitute Eq. (11) into Eq. (10) and minimize the mse by taking derivatives, we find the coefficients must satisfy.

$$\begin{bmatrix} 1 & E\{x_{n-1}\} & E\{x_{n-1}^2\} & E\{x_{n-2}\} & E\{x_{n-2}^2\} \\ E\{x_{n-1}\} & E\{x_{n-1}^2\} & E\{x_{n-1}^3\} & E\{x_{n-1}x_{n-2}\} & E\{x_{n-1}x_{n-2}^2\} \\ E\{x_{n-1}^2\} & E\{x_{n-1}^3\} & E\{x_{n-1}^4\} & E\{x_{n-1}^2x_{n-2}\} & E\{x_{n-1}^2x_{n-2}^2\} \\ E\{x_{n-2}\} & E\{x_{n-1}x_{n-2}\} & E\{x_{n-1}^2x_{n-2}\} & E\{x_{n-2}^2\} & E\{x_{n-2}^3\} \\ E\{x_{n-2}^2\} & E\{x_{n-1}x_{n-2}^2\} & E\{x_{n-1}^2x_{n-2}^2\} & E\{x_{n-2}^3\} & E\{x_{n-2}^4\} \end{bmatrix} \begin{bmatrix} h_o \\ h_{11} \\ h_{21} \\ h_{12} \\ h_{22} \end{bmatrix}$$

$$= \begin{bmatrix} E\{x_n\} \\ E\{x_nx_{n-1}\} \\ E\{x_nx_{n-1}^2\} \\ E\{x_nx_{n-2}\} \\ E\{x_nx_{n-2}^2\} \end{bmatrix} \tag{12}$$

It is seen that the solution requires knowledge of the various moments and cross moments. Since the r.p. is assumed W.S.S., we can apply well known procedures to estimate empirically these various moments. Now consider the example where the signal is generated by use of the equation

$$x_n = k_1x_{n-1}^2 + k_2 + p_1\, u_n;$$

$\{u_n\}$ are iid, uniform $(-\frac{1}{2}, \frac{1}{2})$. From Eq. (6), we know that the optimum predictor is given by

$$\hat{x}_n = k_1x_{n-1}^2 + k_2 \,. \tag{13}$$

This optimum result corresponds to the solution $h_o = k_2$, $h_{21} = k_1$, $h_{11} = h_{12} = h_{22} = 0$.

It is easily shown that this solution satisfies Eq. (12), and hence Eq. (12) leads to the optimum result for the signal in Eq. (13). Likewise, for the general polynomial signal of the form

$$x_n = \sum_{\alpha=0}^{S} \sum_{\beta=0}^{P} \gamma_{\beta\alpha} x_{n-\alpha}^{\beta} + p_1\, u_n \,, \tag{14}$$

Eq. (8) again leads to the optimum solution. These results can also be interpreted in an alternate way. First, define a Hilbert Space over the probability space and the set of r.v. x such that [4], [5]

$$E\{|x|^2\} < \infty$$

with the inner product defined as

$$< x, y > = E\{xy\}$$

We then generate the smallest subspace that contains the elements of the form

$$(x_n)^i - \mu_i$$

where x_n is the n^{th} sample of the r.p. $x(t)$ and μ_i is chosen so that

$$E\{(x_n)^i - \mu_i\} = 0 .$$

The condition

$$E\{|(x_n)^i - \mu_i|^2\} < \infty$$

implies all moments of the r.p. of interest upto and including the $(2m)^{th}$ moment ($m \triangleq$ highest degree polynomial used in the predictor) be finite. Using this augmented subspace, we then have a predictor for x_n, denoted by $\hat{x}_n$, given by

$$\hat{x}_n = \sum_{i=0}^{m} \sum_{j=1}^{k} h_{ij}(x_{n-j}^i - \mu_i) \tag{15}$$

Expanding Eq. (15) and grouping all the constant terms together, the predictor for x_n becomes

$$\hat{x}_n = h_o + \sum_{i=1}^{m} \sum_{j=1}^{k} h_{ij}\, x_{n-j}^i \tag{16}$$

which is identical to Eq. (8). We can use the orthogonality principle to determine the h_{ij}'s. Consequently, we must solve the following set of equations

$$E\{(x_n - h_o - \sum_{i=1}^{m} \sum_{j=1}^{k} h_{ij}x_{n-j}^i)x_{n-p}^\ell\} = 0 \qquad \ell = 0,1,2,\ldots,m \quad p = 1,2,\ldots,k \tag{17}$$

When $m = h = 2$, Eq. (17) leads to Eq. (12). Again consider the signal

$$x_n = -1.74\, x_{n-1}^2 + .87 + .005\, u_n ,$$

where $\{u_n\}$ are iid, uniform $(-\frac{1}{2}, \frac{1}{2})$. Simulation for the case $k = 1$ shows excellent agreement with the known optimum result. However, the determininant of the coefficient matrix vanishes for the case $k = 2$. This is explained by the observantion that the signal can also be represented as

$$x_n = -1.74\, x_{n-1}^2 + .87\,(\alpha+\beta) + .005\, u_n$$

where $\alpha+\beta = 1$. Since

$$x_{n-1} = -1.74\, x_{n-2}^2 + .87 + .005\, u_{n-1},$$

then

$$.87\beta = \beta x_{n-1} + 1.74\, \beta x_{n-2}^2 - .005\, \beta u_{n-1} ;$$

hence

$$x_n = -1.74\, x_{n-1}^2 + .87\alpha + \beta x_{n-1} + 1.74\, \beta x_{n-2}^2 + .005\, u_n - .005\, \beta u_{n-1}$$

We note that there are an infinite number of equivalent signal representations and therefore an infinite number of equivalent predictors. This leads to the following design procedure

(1) Set m. (highest desired polynomial degree)

(2) Set k=1. (k $\triangleq$ number of past samples used in the prediction)
(3) Solve for the h_{ij}.
(4) Set k=2. Compute the determinant of the coefficient matrix. If the determinant is zero, terminate; otherwise proceed to step (5).
(5) Solve for the h_{ij}.
(6) Continue incrementing k, either until the determinant vanishes or a desired value of k is reached.

We also note that functions other than polynomials can be used in the predictor. In this case, the predictor is of the form

$$\hat{x}_n = \sum_{i=1}^{m} \sum_{j=1}^{k} h_{ij} f_i(x_{n-j}),$$

where we assume the r.v. $f_i(x_{n-j})$ possesses the proper second moment properties. In addition, the $f_i(x)$ should be continuous and bounded over the range of arguments to insure that the augmented subspace is complete and the condition

$$E\{|f_i(x)|^2\} < \infty$$

is satisfied.

IV. SUMMARY AND CONCLUSIONS

In this paper, we investigate two methods of designing non-linear discrete-time filters. The first methods makes use of an iterative procedure, that is alternately computing the linear filter coefficients and the non-linearity coefficients. We show how this procedure performs by applying it to several examples. Because the resulting filter design is dependent on the initial conditions before iteration, this method is only applicable to certain problems. For example, this procedure appears acceptable when the starting point of the iteration is close to the optimum design. We then present a second non-iterative procedure that makes use of the orthogonality principle over an augmented subspace. The performance of the resulting design is tested by use of several examples and is shown to provide excellent results. This method appears to work well even when the general form of the optimum filter is not known a priori.

REFERENCES

[1] Lawrence R. Rabiner and Bernard Gold, Theory and Application of Digital Signal Processing, Prentice Hall, Englewood Cliffs, New Jersey, p. 689.

[2] Harry L. Van Trees, Detection Estimation and Modulation Theory, Part I, John Wiley and Sons, Inc., New York, New York, pp. 52-74.

[3] Mischa Schwartz and Leonard Shaw, Signal Processing: Discrete Spectral Analysis, Detection and Estimation, McGraw-Hill, pp. 275-314.

[4] Athanasios Papoulis, Probability, Random Variables and Stochastic Processes, McGraw-Hill, pp. 385-426.

[5] H. Cramer and M.R. Leadbetter, Stationary and Related Stochastic Processes, John Wiley & Sons, Inc., New York, p. 96.

ACKNOWLEDGEMENT

T.E. McCannon and N.C. Gallagher were supported by the Air Force Office of Scientific Research under grant AFOSR 78-3605. G. L. Wise was supported by the Air Force Office of Scientific Research under grant AFOSR 76-3062. D. Minoo-Hamedoni was supported by the Air Force Office of Scientific Research under grant AFOSR 76-3062 and by the Department of Defense Joint Services Electronics Program under contract F49620-77-C-0101.

QUANTIZATION OF BIVARIATE CIRCULARLY SYMMETRIC DENSITIES

J. A. BUCKLEW & N. C. GALLAGHER
School of Electrical Engineering
Purdue University
West Lafayette, IN 47907

ABSTRACT

The problem of quantizing a two dimensional random variable whose bivariate density has circular symmetry is considered in detail. Two quantization methods are considered, leading to polar and rectangular representations. A simple necessary and sufficient condition is derived to determine which of these two quantization schemes is best. If polar quantization is deemed best, the question arises as to the ratio of the number of phase quantizer levels to that of magnitude quantizer levels when the product of these numbers is fixed. A simple expression is derived for this ratio that depends only upon the magnitude distribution. Several examples of common circularly symmetric bivariate densities are worked out in detail using these expressions.

I. INTRODUCTION

Consider a two dimensional random variable X whose bivariate density is circularly symmetric and we desire to represent this quantity by a finite set of values. One possible representation of X leads to a Cartesian co-ordinate system expression wherein we individually quantize the two rectangular components of the random variable. Another common representation leads to a polar co-ordinate representation where we quantize the magnitude and phase angle of X. These two representations are mainly chosen for their computational feasibility and ease of implementation. Other authors have considered the general problem of multidimensional quantization; Zador [1] derives an expression for the minimum error achievable by a multidimensional quantizer for an arbitrary density, but no insight into the required quantizer structure is attained. Chen [2] describes a technique whereby one can use a recursive computer technique to solve for a "good" quantizer, but the optimality of the final solution is not assured. By constraining ourselves to circularly symmetric densities and also to either Cartesian or polar co-ordinate systems, it becomes possible to reduce the optimal two dimensional quantization problem to one dimension. Max [3] develops necessary conditions for the optimality of a one dimensional quantizer. Panter and Dite [4], give a formula for the asymptotic error to be expected for optimal mean square error quantizers (of sufficiently smooth input densities).

In Section II of this paper we obtain a simple criterion by which to determine whether polar format or rectangular format gives a smaller mean square quantization error. It is shown for some very important cases, notably for the Gaussian bivariate density, that polar format is asymptotically superior.

If polar format is to be used and the product $N = N_\theta N_r$ is fixed, where N_θ and N_r are the number of phase and magnitude quantization levels, respectively, the question arises as to the optimum ratio N_θ/N_r. We derive a simple expression for this ratio that depends only upon the magnitude density.

In Section III, we provide several examples of common circularly symmetric densities (e.g. marginal densities are Pearson II, Pearson VII, sinusoidal, and Gaussian) and we address the question of whether the rectangular or the polar format scheme gives a smaller quantization error.

Presented at the Sixteenth Annual Allerton Conference on Communication, Control, and Computing, October 4-6, 1978.

II. DEVELOPMENT

Consider the mean square quantization error E_p of a polar format representation,

$$E_p = \sum_{j=1}^{N_\theta} \sum_{i=1}^{N_r} \int_{c_{j-1}}^{c_j} \int_{a_{i-1}}^{a_i} |re^{j\theta} - b_i e^{jd_j}|^2 \frac{f_r(r)\, dr\, d\theta}{2\pi} . \tag{1}$$

Implicit use has been made of the fact that in circularly symmetric bivariate densities the magnitude random variable with probability density $f_r(\cdot)$ is independent of the uniformly distributed $[-\pi,\pi]$ phase random variable. The b_i and d_j are the output levels of the magnitude and phase quantizers corresponding to input levels lying in the intervals $(a_{i-1}, a_i]$ and $(c_{j-1}, c_j]$, respectively. It is shown in [5] that the optimal phase quantizer is the uniform quantizer. This allows us to simplify Eq. (1);

$$E_p = \sum_{i=1}^{N_r} \int_{a_{i-1}}^{a_i} \left[r^2 + b_i^2 - 2rb_i \frac{\sin \frac{\pi}{N_\theta}}{\frac{\pi}{N_\theta}}\right] f(r)\, dr . \tag{2}$$

Differentiating with respect to b_i, we find the optimum b_i is

$$b_i = \frac{\sin \frac{\pi}{N_\theta}}{\frac{\pi}{N_\theta}} \frac{\int_{a_{i-1}}^{a_i} rf(r)\, dr}{\int_{a_{i-1}}^{a_i} f(r)\, dr} . \tag{3}$$

The equation given by Max for the output levels b_i' of an optimal one dimensional magnitude quantizer is found in [3] to be

$$b_i' = \frac{\int_{a_{i-1}'}^{a_i'} rf(r)\, dr}{\int_{a_{i-1}'}^{a_i'} f(r)\, dr} , \tag{4}$$

where the optimal input interval endpoints a_i' (for the one dimensional case) satisfy

$$a_i' = \frac{b_i' + b_{i+1}'}{2} . \tag{5}$$

If we minimize Eq. (2) with respect to the a_i, we then arrive at the necessary condition (for the two dimensional case)

$$a_i = \frac{b_i + b_{i+1}}{2\left(\frac{\sin \frac{\pi}{N_\theta}}{\frac{\pi}{N_\theta}}\right)} = \frac{b_i' + b_{i+1}'}{2} = a_i' \tag{6}$$

This equation indicates that the quantizer interval endpoints for the

optimum magnitude quantizer in the two dimensional case is the same as the quantizer interval endpoints for the optimum one dimensional quantizer. From Eqs. (3) and (4) and the preceeding discussion, we have the following relationship between the output levels b_i' and b_i:

$$b_i' = \frac{\frac{\pi}{N_\theta}}{\sin \frac{\pi}{N_\theta}} b_i \quad . \tag{7}$$

Consequently, Eq. (2) becomes

$$E_p = E\{r^2\} - \left(\frac{\sin \frac{\pi}{N_\theta}}{\frac{\pi}{N_\theta}}\right)^2 \sum_{i=1}^{N_r} (b_i')^2 \int_{a_{i-1}}^{a_i} f(r)\, dr \quad , \tag{8}$$

where $E\{\cdot\}$ is the statistical expectation operator. In [6] it is shown that the mean square quantization error for a minimum mean square error quantizer is simply the input variance minus the output variance. If we denote by E_X^N the mean square quantization error produced by an optimal N level quantizer for the random variable X, we may rewrite Eq. (8) as

$$E_p = \left(\frac{\sin \frac{\pi}{N_\theta}}{\frac{\pi}{N_\theta}}\right)^2 E_r^{N_r} + \left(1 - \left(\frac{\sin \frac{\pi}{N_\theta}}{\frac{\pi}{N_\theta}}\right)^2\right) E\{r^2\} \quad . \tag{9}$$

Our problem now is one of characterizing the quantity E_X^N. Panter and Dite [4] give a formula for the expected error of a minimum mean square error quantizer with a large number of output levels and a smooth input density. This formula is

$$E_X^N = \frac{K_X}{N^2}$$

where

$$K_X = \frac{\left[\int_{-\infty}^{\infty} f(x)^{1/3}\, dx\right]^3}{12} \quad . \tag{10}$$

Roe [7] also derives some asymptotic formula which were later used by Wood [8] to rederive Eq. (10). Roe's formula depend on the truncation of a Taylor series expansion of the input density. Wood, in his formula, explicitly states that the input density and the first few derivates (up to order five in some cases) must exist and be continuous. Panter and Dite, in their derivation, require that as the input intervals become very small, the density function may be approximated as a constant over each interval. In [1] it is shown that a sufficient condition for Eq. (10) to hold is that $f(x)$ be Riemann integrable and that $E\{x^{2+\delta}\} < \infty$ for some $\delta > 0$, in general a much less severe restriction than continuity or differentiability.

We make use of the approximation

$$\left(\frac{\sin x}{x}\right)^2 \approx 1 - \frac{x^2}{3} \quad , \tag{11}$$

and of Eq. (10) in order to reduce Eq. (9) as

$$E_p \approx \left(1 - \frac{\pi^2}{3N_\theta^2}\right) \frac{K_r}{N_r^2} + \frac{2}{3}\frac{\pi^2}{N_\theta^2} \quad . \tag{12}$$

If we denote as N the total number of output levels allowed to represent the two dimensional random variable X, we have the following relation,

$$N = N_r N_\theta \quad . \tag{13}$$

Since $K_r > 0$, it is simple to show that $N_r = O(N^{1/2})$ and $N_\theta = O(N^{1/2})$ by differentiating Eq. (12) and solving for the optimal quantities. Making use of this fact and Eq. (13), we may, assuming sufficiently large N, write Eq. (10) as

$$E_p = \frac{K_r N_\theta^2}{N_\theta^2} + \frac{2}{3}\frac{\pi^2}{N_\theta^2} \quad . \tag{14}$$

This is then optimized with respect to N_θ and yields the optimal N_θ^2 as

$$N_\theta^2 = \left(\frac{2\pi^2}{3K_r}\right)^{1/2} N \quad . \tag{15}$$

This leads to the following expression for the minimal attainable asymptotic polar format error,

$$(E_p)_{opt} = \sqrt{\frac{2K_r}{3}}\,\frac{2\pi}{N} \quad . \tag{16}$$

Now consider the problem of optimally quantizing the random variable X in a rectangular format. The mean square quantization error, E_x, of this representation is given by

$$E_x = \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} \int_{g_{j-1}}^{g_j} \int_{e_{i-1}}^{e_i} \left[(x-f_i)^2 + (y-h_j)^2\right] f_{x,y}(x,y)\, dxdy \quad , \tag{17}$$

where N_x and N_y are the number of levels in each of the respective orthogonal random variables. The other notation should be clear. Equation (17) may be written as

$$E_x = \sum_{i=1}^{N_x} \int_{e_{i-1}}^{e_i} (x-f_i)^2 f_x(x)\, dx + \sum_{j=1}^{N_y} \int_{g_{j-1}}^{g_j} (y-h_j)^2 f_y(y)\, dy \quad . \tag{18}$$

By symmetry arguments (since $f_x(x) = f_y(x)$), we may argue that $N_x = N_y = N^{1/2}$. The quantizer that minimizes the above equation is simply the minimum mean square error quantizer for each of the two components. Therefore, again using Eq. (10) we have for large N.

$$E_x = \frac{2K_x}{N} \quad ,$$

where

$$K_x = \frac{\left[\int_{-\infty}^{\infty} f'(x)^{1/3}\, dx\right]^3}{12} \quad . \tag{19}$$

Comparing Eq. (19) and Eq. (16), we say that polar format is asymptotically better than rectangular format if and only if

$$\frac{2K_x}{N} > \sqrt{\frac{2K_r}{3}}\,\frac{2\pi}{N} \quad ,$$

or

$$K_x > \sqrt{\frac{2K_r}{3}}\,\pi \quad . \tag{20}$$

In other words, if the inequality is satisfied and the original input probability density is Riemann integrable, then we are guaranteed that there exists an N_o such that for every $N > N_o$, polar format quantization will perform better than rectangular format quantization.

If polar quantization is deemed best for a particular density, then what is the ratio N_θ/N_r that provides the smallest total error? This question is answered with the use of Eq. (15); we find

$$\left(\frac{N_\theta^2}{N}\right)_{opt} = \left(\frac{N_\theta}{N_r}\right)_{r\ opt} = \sqrt{\frac{2}{3K_r}}\,\pi \quad . \tag{21}$$

III. EXAMPLES

For our first example, we calculate the relevant parameters for a random variable whose marginal density is of Pearson Type VII. This distribution is a generalization of Students-t distribution. The bivariate density is

$$f(x,y) = \frac{\nu}{\pi}\,\frac{2^\nu(\nu-1)^\nu}{(2(\nu-1) + x^2 + y^2)^{\nu+1}} \quad , \quad -\infty < x,\ y < \infty \tag{22}$$

(with $\nu > 1$ in order to assure finite variance) and the marginal density appears as

$$f(x) = \frac{2^\nu(\nu-1)^\nu\,\Gamma(\nu+1/2)}{\sqrt{\pi}\,\Gamma(\nu)\,(2(\nu-1) + x^2)^{\nu+1/2}} \quad , \quad -\infty < x < \infty \tag{23}$$

where $\Gamma(\cdot)$ is the gamma function and where we have normalized the distribution so that $f(x)$ has unit variance. The magnitude density is always derived by substituting in r for $\sqrt{x^2 + y^2}$ in $f(x,y)$ and multiplying the result by $2\pi r$, as shown by a simple change of variable. Eq. (23) yields after some tedious algebra

$$K_x = \frac{[B(\frac{1}{2}\ ;\ \frac{\nu-1}{3})]^3}{12\ B(\frac{3}{2}\ ;\ \nu-1)} \quad , \tag{24}$$

where $B(\cdot\ ;\ \cdot)$ is the beta function. We perform similar operations with the magnitude density to yield

$$K_r = \frac{\nu(\nu-1)}{24}\,[B(\tfrac{2}{3}\ ;\ \tfrac{\nu-1}{3})]^3 \quad . \tag{25}$$

In Fig. 1 K_x (solid line) and $\sqrt{\frac{2K_r}{3}}\,\pi$ (dotted line) are plotted as a function of ν for values from 1.1 to 21.1. As shown by this graph, polar format is always asymptotically best for this class of distributions. An interesting point about this set of distributions is that in the limit as $\nu \to \infty$ Eq. (23) converges to a unit variance Gaussian density. Therefore, taking this limit in Eq. (24) and making use of Stirling's approximation,

we have

$$K_x \to \frac{\sqrt{3\pi}}{2} \cong 2.721 \quad . \tag{26}$$

Wood [8] estimates this number as 2.73 which is close to our derived value. From Eq. (25) we have similarly

$$K_r \to \frac{3}{8}\left(\Gamma\left(\frac{2}{3}\right)\right)^3 \cong .931 \quad , \tag{27}$$

which is the parameter for the Rayleigh distribution obtained in the limit. Using these two values in Eq. (20), we conclude that asymptotically polar formatting is better than rectangular formatting for Gaussian bivariate densities. As a matter of interest, when we substitute the value of K_r found in Eq. (27) into Eq. (21), we find the optimal ratio N_θ/N_r to be 2.659. Pearlman [9] using distortion rate theory states that this ratio should be > 2.596, which is in agreement with our result.

For the next example, consider distributions of the Pearson II class. The bivariate density is

$$f(x,y) = \frac{v(2(v+1) - (x^2 + y^2))^{v-1}}{\pi 2^v (v+1)^v} U(2(v+1) - (x^2 + y^2)) \quad , \tag{28}$$

where $v > 0$, and $U(\cdot)$ is the unit step function. The marginal density is

$$f(x) = \frac{\Gamma(v+1)(2(v+1) - x^2)^{v-\frac{1}{2}} U(2(v+1) - x^2)}{2^v (v+1)^v \sqrt{\pi}\, \Gamma(v+\frac{1}{2})} \quad . \tag{29}$$

For $v = 1/2$ we find that $f(x)$ has a uniform distribution. For $v = 1$, we have that the bivariate density is uniform over a circular region in the plane. Using Eq. (29), we find

$$K_x = \frac{[B(\frac{1}{2}\ ;\ \frac{2v+5}{6})]^3}{12\ B(\frac{3}{2}\ ;\ v+\frac{1}{2})} \quad . \tag{30}$$

From the magnitude density we derive that

$$K_r = \frac{v(v+1)}{24}\ [B(\tfrac{2}{3}\ ;\ \tfrac{v+2}{3})]^3 \quad . \tag{31}$$

In Fig. (2) can be seen a plot of K_x (solid line) and $\sqrt{\frac{2}{3}\ K_r\ \pi}$ (dotted line) as a function of v for values from 0 to 10. It should be noted that Eq. (29) also converges to a Gaussian density as $v \to \infty$. It is a simple matter to check that the expressions in Eqs. (30) and (31) indeed go to the correct limits. From the plot it can be seen that for values of v in the interval (0.0, .4) polar format is better. In the interval (.4, 3.635) it is seen that rectangular is better, and from 3.635 to ∞ polar again is better. It appears then that for the circularly symmetric bivariate density whose marginal density is uniform, we have the interesting result that rectangular format is asymptotically better than polar format.

In the theoretical development and in the examples considered so far, we have constrained the class of quantizers considered to two different types, the rectangular format and the polar format. In general, neither of these schemes will be optimal for an arbitrary two dimensional random variable with a circularly symmetric probability density. Zador [1] gives an expression for the asymptotic mean square error E_z of the optimal two dimensional mean square error quantizer. This equation is

$$E_z = C_z/N \ , \tag{32}$$

where

$$C_z = \frac{5}{18\sqrt{3}} \left[\int_{-\infty}^{\infty} \int f_{x,y}(x,y)^{1/2} \, dxdy \right]^2 \tag{33}$$

For the Pearson VII density $C_z = 4.0307\ v/(v-1)$, for the Pearson II density $C_z = 4.0307\ v/(v+1)$. Since in the limit as v becomes large, both of these classes of densities converge to the Gaussian, the smallest error attainable for a two dimensional normal random variable is approximately 4.0307/N. The best that we can do with a polar format representation is 4.95/N and the best that we can do with a Cartesian format representation is 5.442/N. There is certainly room for improvement here, however, the important thing to note is that the structure of the polar format quantizer is known while that of the theoretical optimum quantizer is not.

In section II it was stated that a sufficient condition for Eq. (10) to be valid is that the magnitude density function be Riemann integrable. For most density functions of interest in modeling physical systems this criterion is met. One group of densities that doesn't meet this condition is the set of atomic densities, i.e., densities for which probability mass is contained at a single point. In a circularly symmetric bivariate density, the phase must be uniformly distributed $[-\pi,\pi]$. The only quantity that can be discrete is the magnitude distribution, i.e. we may have "rings" of probability mass distributed in the plane. Suppose we have a single "ring" of probability mass, where the radius of the ring is 1, i.e.,

$$F(r) = U(r-1) \ , \tag{34}$$

where $F(\cdot)$ is the magnitude distribution function and $U(\cdot)$ is the unit step function. The rectangular component marginal density is the sinusoidal density

$$f(x) = \frac{U(1 - x^2)}{\pi\sqrt{1-x^2}} \ . \tag{35}$$

This density function is Riemann integrable, hence Eq. (10) and Eq. (19) are valid. This implies the rectangular format error is $O(N^{-1})$. Now consider the polar format case. For $N_r \geq 1$, $E_r^N = 0$. This implies the polar format error for large N is $O(N^{-2})$. Clearly then polar format is asymptotically better for this density. By extending this argument, we may say that if $P(r=0) \neq 1$, then for any bivariate circularly symmetric density with an atomic magnitude density with a finite number of atoms, polar format will give a smaller asymptotic mean square quantization error than rectangular format.

IV. SUMMARY

In this paper we have derived a simple criterion to determine whether rectangular format or polar format gives smaller mean square error for circularly symmetric densities. The optimal ratio of phase quantizer levels to magnitude quantizer levels is also derived. Several examples including the Gaussian case have been studied in detail.

It is interesting to note that polar format is not always better than rectangular format even for the case of densities with circular symmetry.

This research has been supported by the Air Force Office of Scientific Research under grant AFOSR 78-3605.

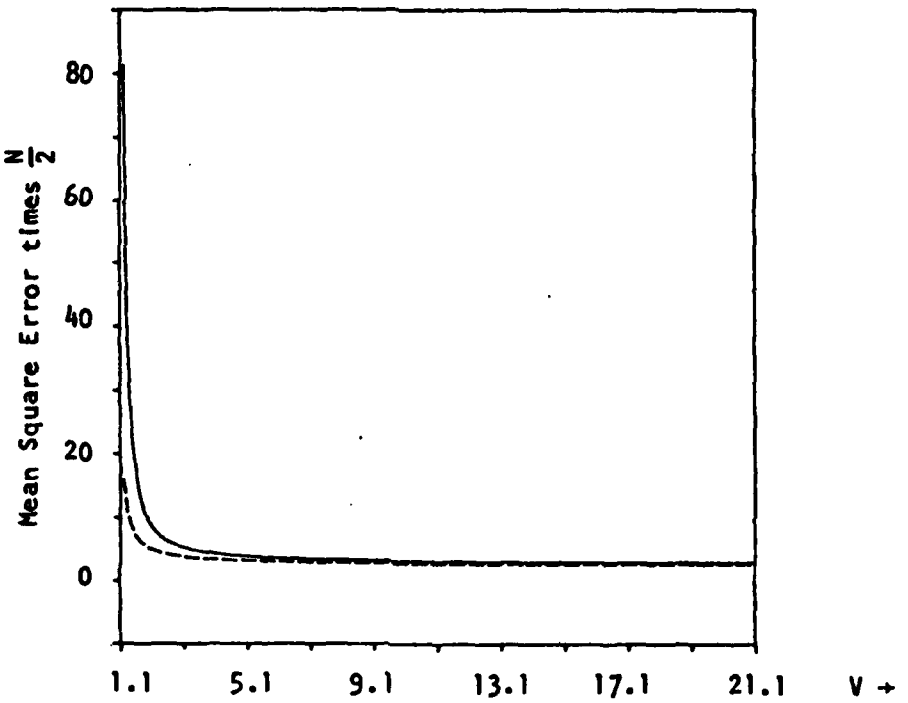


Fig. 1. The solid line is a plot of K_x as a function of V, the dotted line is a plot of $\sqrt{\frac{2K_r}{3}}\,\pi$ as a function of V for the Pearson VII density.

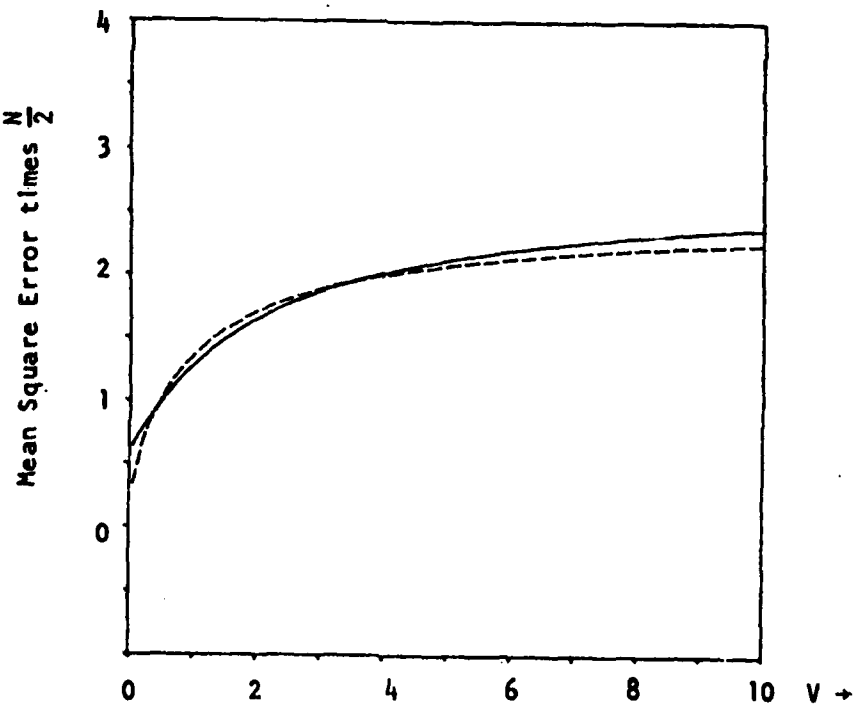


Fig. 2. The solid line is a plot of K_x as a function of V, the dotted line is a plot of $\sqrt{\frac{2K_r}{3}}\,\pi$ as a function of V for the Pearson II density.

REFERENCES

[1] P. Zador, "Development and evaluation of procedures for quantizing multivariate distributions," Ph.D. dissertation, Stanford University, Stanford, CA, 1964.

[2] D. Chen, "On two or more dimensional optimum quantizers," The Aloha System Tech. Rep. A71-4, Univ. of Hawaii, Honolulu, Hawaii, Jan. 1971.

[3] J. Max, "Quantizing for minimum distortion," IEEE Trans. Info. Theory, Vol. IT-6, pp. 7-12, Jan. 1960.

[4] P. F. Panter and W. Dite, "Quantization distortion in pulse count modulation with nonuniform spacing of levels," Proc. IRE Vol. 39, pp. 44-48, Jan. 1951.

[5] N. C. Gallagher, "Quantizing schemes for the discrete Fourier transform of a random time series," IEEE Trans. Info. Theory, Vol. IT-24, pp. 156-163, Mar. 1978.

[6] J. A. Bucklew and N. C. Gallagher, "A Note on Optimum Quantization," To appear in IEEE Trans. on Info. Theory.

[7] G. M. Roe, "Quantizing for minimum distortion," IEEE Trans. Info. Theory, Vol. IT-10, pp. 384-385, Oct. 1964.

[8] R. C. Wood, "On optimum quantization," IEEE Trans. Info. Theory, Vol. IT-5, pp. 248-252, Mar. 1969.

[9] W. A. Pearlman, "Quantization Error Bounds for Computer Generated Holograms," Tech. Rep. #65031-1, Stanford University Information Systems Laboratory, Stanford, CA, August 1974.

QUANTIZATION IN SPECTRAL PHASE CODING

Kerry D. Rines and Neal C. Gallagher, Jr.

School of Electrical Engineering
Purdue University
West Lafayette, Indiana 47907

ABSTRACT

Spectral Phase Coding (SPC) is a robust suboptimum digital encoding scheme utilizing the discrete Fourier transform. The quantization of the SPC sequence $\{\psi_p\}$ is examined as an effective quantization of the spectral magnitude and phase. A new encoding technique called Prequantized Spectral Phase Coding (PQSPC) is introduced. PQSPC exhibits the same robust characteristics as SPC with a reduction in MSE. For the case of a double-sided exponential input density this reduction in MSE is 47.5%.

I. INTRODUCTION

Spectral Phase Coding (SPC) is a robust suboptimum technique for coding a nonstationary or large dynamic range discrete-time series into digital form. In previous work [1], the performance of SPC in a mean squared error sense has been evaluated. However, limited insight is provided into the effects of the various SPC parameters on overall performance. In this paper, we investigate the effect of converting the spectral magnitude and phase of the discrete signal into the SPC sequence $\{\psi_p\}$ before quantization and transmission. In section II, density functions for the magnitude and phase errors at the receiver are obtained. These results suggest a method of improving the SPC encoding algorithm. In section III, a technique called Prequantizing is introduced. The addition of Prequantizing to SPC offers a substantial improvement in the overall system performance.

II. SPECTRAL PHASE CODING QUANTIZATION ERROR

Spectral Phase Coding uses the discrete Fourier transform (DFT) to encode a discrete-time complex-valued random sequence $\{a_n\}_{n=0}^{M-1}$ for digital transmission. The SPC encoding and decoding algorithms are given here. A detailed explanation of the SPC procedure is available in [2]. The spectral magnitude A_p and the spectral phase θ_p of the discrete sequence are given below:

$$\{a_n\}_{n=0}^{M-1} \xrightarrow{\text{DFT}} \{A_p e^{i\theta_p}\}_{p=0}^{M-1} . \tag{1}$$

SPC encodes the magnitude and phase of the spectrum by forming the sequence $\{\psi_p\}_{p=0}^{2M-1}$ given by

$$\psi_p = \theta_p + \gamma_p$$

$$\psi_{p+M} = \theta_p - \gamma_p , \tag{2}$$

where $\gamma_p = \cos^{-1}\frac{A_p}{S}$

and $S = \max_p A_p \quad p = 0,1,\ldots,m-1.$

The quantized sequence $\{\hat{\psi}_p\}$ is transmitted and used at the receiver to recover the original discrete signal. The reconstructed discrete sequence is

$$\{\hat{a}_n\}_{n=0}^{M-1} \xrightarrow{\text{DFT}^{-1}} \{\tfrac{S}{2}(e^{i\hat{\psi}_p} + e^{i\hat{\psi}_{p+M}})\}_{p=0}^{M-1} . \tag{3}$$

This equation can be written in terms of the equivalent magnitude and phase components at the receiver.

$$\{\hat{a}_n\}_{n=0}^{M-1} \xrightarrow{\text{DFT}^{-1}} \{\tfrac{S}{2} e^{i\hat{\theta}_p}(e^{i\hat{\gamma}_p} + e^{-i\hat{\gamma}_p})\}_{p=0}^{M-1} \tag{4}$$

where $$\hat{\theta}_p = \tfrac{1}{2}(\hat{\psi}_p + \hat{\psi}_{p+M})$$

$$\hat{\gamma}_p = \tfrac{1}{2}(\hat{\psi}_p - \hat{\psi}_{p+M}) . \tag{5}$$

We define $\hat{\theta}_p$ and $\hat{\gamma}_p$ to be the effective quantization levels of θ_p and γ_p that result when $\{\hat{\psi}_p\}$ is formed. The effective quantization errors of θ_p and γ_p are defined in Eq. (6).

$$e_p = \theta_p - \hat{\theta}_p$$

$$d_p = \gamma_p - \hat{\gamma}_p . \tag{6}$$

The effective errors e_p and d_p are the result of using $\hat{\theta}_p$ and $\hat{\gamma}_p$ instead of θ_p and γ_p to reconstruct $\{a_n\}$ at the receiver.

It is also possible to reconstruct the discrete signal by sending quantized values of θ_p and γ_p directly. We define these quantized values to be $\tilde{\theta}_p$ and $\tilde{\gamma}_p$ and the resulting quantization error for this case is

$$n_p = \theta_p - \tilde{\theta}_p$$

$$m_p = \gamma_p - \tilde{\gamma}_p . \tag{7}$$

The two sets of errors in Eqs. (6) and (7) are compared to determine the effect of transmitting $\{\hat{\psi}_p\}$ rather than $\{\tilde{\theta}_p\}$ and $\{\tilde{\gamma}_p\}$ on the overall performance. We find the effective quantization errors e_p and d_p can be written as deterministic functions of the individual quantization errors n_p and m_p. This is first demonstrated with two simple examples. The quantizer has N levels uniformly spaced from 0 to 2π for both methods.

Example 1:

$N = 4$ with output levels $0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}$. Let $\theta_p = 0.6\,\pi$ and $\gamma_p = 0.4\,\pi$, then we find that

$$\tilde{\theta}_p = 0.5\,\pi \qquad \tilde{\gamma}_p = 0.5\,\pi$$

$$\psi_p = \pi \qquad \psi_{p+M} = 0.2\,\pi .$$

Upon quantizing the value for ψ_p and ψ_{p+M}, we have

Presented at the 1979 Conference on Information Sciences & Systems, The John Hopkins University, March, 1979. To be published in Proceedings of this Conference.

$$\hat{\psi}_p = \pi \qquad \hat{\psi}_{p+M} = 0.0 \ .$$

So,

$$\hat{\theta}_p = 0.5\,\pi \qquad \hat{\gamma}_p = 0.5\,\pi .$$

Consequently,

$$e_p = 0.1\,\pi \qquad n_p = 0.1\,\pi$$
$$d_p = -0.1\,\pi \qquad m_p = -0.1\,\pi \ .$$

In this example, the effective quantization errors are the same as the errors from direct quantization.

Example 2:

$N = 4$, $\theta_p = 0.7\,\pi$, $\gamma_p = 0.1\,\pi$; we find that

$$\tilde{\theta}_p = 0.5\,\pi \qquad \tilde{\gamma}_p = 0.0$$
$$\psi_p = 0.8\,\pi \qquad \psi_{p+M} = 0.6\,\pi \ .$$

So,

$$\hat{\psi}_p = \pi \qquad \hat{\psi}_{p+M} = 0.5\,\pi$$
$$\hat{\theta}_p = 0.75\,\pi \qquad \hat{\gamma} = 0.25\pi \ .$$

Consequently,

$$e_p = -0.05\,\pi \qquad n_p = 0.2\,\pi$$
$$d_p = -0.15\,\pi \qquad m_p = 0.1\,\pi \ .$$

In this case the effective quantization errors have different values than the direct quantization errors. We note that the difference between n_p and e_p is π/N and that the difference between m_p and d_p is also π/N.

A detailed comparison of the two sets of quantization errors is developed in the Appendix. The results are given below:

$$\begin{aligned} e_p = n_p \quad & , \ -\tfrac{\pi}{N} \le n_p + m_p \le \tfrac{\pi}{N} \text{ and } -\tfrac{\pi}{N} \le n_p - m_p \le \tfrac{\pi}{N} \\ = n_p + \tfrac{\pi}{N} & , \ -\tfrac{2\pi}{N} \le n_p + m_p \le -\tfrac{\pi}{N} \text{ or } -\tfrac{2\pi}{N} \le n_p - m_p \le -\tfrac{\pi}{N} \\ = n_p - \tfrac{\pi}{N} & , \ \tfrac{\pi}{N} \le n_p + m_p \le \tfrac{2\pi}{N} \text{ or } \tfrac{\pi}{N} \le n_p - m_p \le \tfrac{2\pi}{N} \ , \end{aligned} \tag{8}$$

and

$$\begin{aligned} d_p = m_p \quad & , \ -\tfrac{\pi}{N} \le n_p + m_p \le \tfrac{\pi}{N} \text{ and } -\tfrac{\pi}{N} \le n_p - m_p \le \tfrac{\pi}{N} \\ = m_p + \tfrac{\pi}{N} & , \ -\tfrac{2\pi}{N} \le n_p + m_p \le -\tfrac{\pi}{N} \text{ or } \tfrac{\pi}{N} \le n_p - m_p \le \tfrac{2\pi}{N} \\ = m_p - \tfrac{\pi}{N} & , \ -\tfrac{2\pi}{N} \le n_p - m_p \le -\tfrac{\pi}{N} \text{ or } \tfrac{\pi}{N} \le n_p + m_p \le \tfrac{2\pi}{N} . \end{aligned} \tag{9}$$

Examining Eq. (8), we see that the effective phase quantization error e_p is a function of both the magnitude and phase errors n_p and m_p. The same is true for the effective error d_p.

The distribution functions of e_p and d_p can be evaluated in terms of the joint density $f(n,m)$ by use of Eqs. (8) and (9). For $x \le 0$,

$$\begin{aligned} F_e(x) = P\{n_p \le x, \ -\tfrac{\pi}{N} \le n_p + m_p \le \tfrac{\pi}{N}, \ -\tfrac{\pi}{N} \le n_p - m_p \le \tfrac{\pi}{N}\} \\ + P\{n_p \le x + \tfrac{\pi}{N}, \ \tfrac{\pi}{N} \le n_p + m_p \le \tfrac{2\pi}{N}\} \\ + P\{n_p \le x + \tfrac{\pi}{N}, \ \tfrac{\pi}{N} \le n_p - m_p \le \tfrac{2\pi}{N}\} \ . \end{aligned}$$

This expression and a similar expression for $F_d(x)$ lead to the results below. For $-\frac{\pi}{N} \le x \le 0$,

$$\begin{aligned} F_e(x) = & \int_0^{\frac{\pi}{N}+x} \int_{m-\frac{\pi}{N}}^{x} f(n,m)\, dn\, dm \\ & + \int_{-\frac{\pi}{N}-x}^{0} \int_{-m-\frac{\pi}{N}}^{x} f(n,m)\, dn\, dm \\ & + \int_{-x}^{\frac{\pi}{N}} \int_{\frac{\pi}{N}-m}^{x+\frac{\pi}{N}} f(n,m)\, dn\, dm \\ & + \int_{-\frac{\pi}{N}}^{x} \int_{\frac{\pi}{N}+m}^{x+\frac{\pi}{N}} f(n,m)\, dn\, dm \end{aligned} \tag{10}$$

and

$$\begin{aligned} F_d(x) = & \int_{-\frac{\pi}{N}}^{x} \int_{-\frac{\pi}{N}-m}^{0} f(n,m)\, dn\, dm \\ & + \int_{-\frac{\pi}{N}}^{x} \int_{m+\frac{\pi}{N}}^{0} f(n,m)\, dn\, dm \\ & + \int_{0}^{\frac{\pi}{N}+x} \int_{-\frac{\pi}{N}+m}^{-\frac{\pi}{N}} f(n,m)\, dn\, dm \\ & + \int_{0}^{\frac{\pi}{N}+x} \int_{\frac{\pi}{N}-m}^{\frac{\pi}{N}} f(n,m)\, dn\, dm \ . \end{aligned} \tag{11}$$

We obtain similar results for $0 \le x \le \frac{\pi}{N}$.

The general results given in Eqs. (10) and (11) can now be used to determine the densities of the SPC error e_p and d_p. The properties of the DFT indicate that for a large block size M, θ_p and γ_p will be independent with θ_p uniformly distributed $(0, 2\pi)$. Therefore, we assume that n_p and m_p are statistically independent and that n_p is uniformly distributed $(-\pi/N, \pi/N)$. The densities of the equivalent quantization errors e_p and d_p for the SPC case are given here. For $-\pi/N \le x \le \pi/N$,

$$\begin{aligned} f_e(x) = \frac{N}{2\pi}\Big(& \int_0^{\frac{\pi}{N}-|x|} f(m)\, dm + \int_{-\frac{\pi}{N}+|x|}^{0} f(m)\, dm \\ & + \int_{|x|}^{\frac{\pi}{N}} f(m)\, dm + \int_{-\frac{\pi}{N}}^{-|x|} f(m)\, dm \Big) \end{aligned} \tag{12}$$

and

$$\begin{aligned} f_d(x) &= (1 + \tfrac{N}{\pi} x)[f_m(x) + f_m(\tfrac{\pi}{N} + x)], \quad x<0 \\ &= (1 - \tfrac{N}{\pi} x)[f_m(x) + f_m(x - \tfrac{\pi}{N})], \quad x>0. \end{aligned} \tag{13}$$

The density of m_p is, in general, dependent upon the statistics of the input signal. For a large number of quantization levels N, we can assume the density of m_p to be uniform $(-\pi/N, \pi/N)$. The resulting error densities are shown in Figure 1. The result is confirmed by computer simulation.

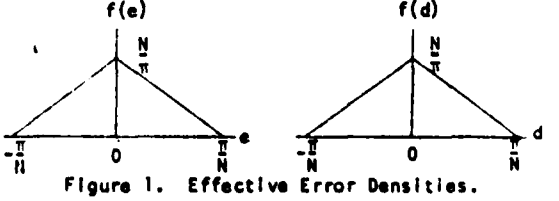


Figure 1. Effective Error Densities.

individual quantization of θ_p and γ_p would yield uniformly distributed error densities $(-\pi/N, \pi/N)$ for the case described above. Therefore we conclude that preparing θ_p and γ_p for digital transmission by using the sequence $\{\psi_p\}$ represents an important element in SPC performance.

Once we have evaluated the densities f(e) and f(d) the calculation of the mean squared quantization error for θ_p and γ_p is straight forward. We now present expressions for computing this quantization error directly. The expressions are obtained by computing the Fourier series expansion for the quantization error of ψ_p and using the result with Eqs. (2) and (5). The effective errors e_p and d_p are

$$e_p = -\frac{2}{N}\sum_{n=1}^{\infty}\frac{(-1)^n}{n}\sin nN\theta_p \cos nN\gamma_p \qquad (14)$$

and

$$d_p = -\frac{2}{N}\sum_{n=1}^{\infty}\frac{(-1)^n}{n}\cos nN\theta_p \sin nN\gamma_p \; . \qquad (15)$$

The mean squared error expressions for SPC are obtained by assuming θ_p and γ_p are independent and θ_p is uniformly distributed $(0, 2\pi)$. Thus the effective mean squared errors are

$$E\{e_p^2\} = \frac{1}{N^2}\sum_{n=1}^{\infty}\frac{1}{n^2}\left(1 + E\{\cos 2nN\gamma_p\}\right) \qquad (16)$$

and

$$E\{d_p^2\} = \frac{1}{N^2}\sum_{n=1}^{\infty}\frac{1}{n^2}\left(1 - E\{\cos 2nN\gamma_p\}\right) \; . \qquad (17)$$

We have investigated the sequence $\{\hat{\psi}_p\}$ in terms of the effective quantization of θ_p and γ_p. Effective quantization error densities and mean squared error expressions have been found. These results will be used in the following section to improve the SPC performance.

III. PREQUANTIZED SPECTRAL PHASE CODING

We have stated at the outset that SPC is a suboptimum technique for encoding discrete-time signals. The results from [1] indicate that for a fixed bit rate the number of magnitude quantization levels N_1, and the number of phase levels N_2, must be related by

$$N_2 = 2.596\, N_1 \qquad (18)$$

for optimum performance. In SPC, γ_p ranges from 0 to $\pi/2$ and θ_p ranges from 0 to 2π. Thus γ_p has only one-fourth the effective quantization levels of θ_p at the receiver. This suggests that an encoding tradeoff which decreases the MSE on γ_p at the expense of increasing the MSE on θ_p could improve the SPC performance. The previous results offer a method of obtaining the desired tradeoff.

The effective errors e_p and d_p have been shown to be functions of both n_p and m_p and thus they are functions of both θ_p and γ_p. Suppose θ_p' has a density function that minimizes the MSE on γ_p at the receiver. By shaping the density of θ_p to be that of θ_p' before forming the sequence $\{\psi_p\}$ we can lower the MSE on γ_p at the expense of increasing the MSE on θ_p. We determine θ_p' as given below. Using Eq. (15) we obtain

$$E\{d_p^2\} = \frac{4}{N^2}\sum_{n=1}^{\infty}\sum_{m=1}^{\infty}\frac{(-1)^{n+m}}{nm} \times E\{\cos nN\theta_p \cos mN\theta_p \sin nN\gamma_p \sin mN\gamma_p\} \; . \qquad (19)$$

Thus the MSE on γ_p assuming SPC statistics and a large number of quantization levels N is

$$E\{d_p^2\} \cong \frac{1}{N^2}\sum_{n=1}^{\infty}\frac{1}{n^2}\left(1 + E\{\cos 2nN\theta_p\}\right) \; . \qquad (20)$$

From Eq. (20), we find that $E\{d_p^2\}$ is minimized for

$$\theta_p = \theta_p' = K\frac{\pi}{N} + \frac{\pi}{2N} \; . \qquad (21)$$

Applying these results, we propose the following coding scheme called Prequantized Spectral Phase Coding (PQSPC). First obtain θ_p and γ_p as with SPC. θ_p is then input into a uniform quantizer with output levels $K\pi/N + \pi/2N$ for $K = 0, 1, \ldots, 2N-1$. This operation is called Prequantizing. The quantizer output θ_p' is then used to form the sequence $\{\psi_p\}$ and the rest of the procedure is identical to SPC.

The techniques acquired in Section II are applied to determine the effective error densities of PQSPC.

$$f_e(x) = \frac{N}{2\pi} \quad , \quad -\frac{\pi}{N} \le x \le \frac{\pi}{N} \qquad (22)$$

and

$$f_d(x) = f_m(x) + f_m\left(x + \frac{\pi}{N}\right), \quad -\frac{\pi}{2N} \le x \le 0$$
$$= f_m(x) + f_m\left(x - \frac{\pi}{N}\right), \quad 0 \le x \le \frac{\pi}{2N} \; . \qquad (23)$$

The tradeoff accomplished by Prequantizing can be seen by comparing the above densities to those evaluated in Figure 1. The MSE and range of d_p are both reduced by a factor of two at the expense of e_p.

The normalized MSE performance of PQSPC is presented in Table 1 for a number of computer simulations. The optimum unit variance Gaussian quantizer (O.G.Q.), the optimum uniform unit variance Gaussian quantizer (U.G.Q.), and SPC performances are also presented in Table 1 for comparison. All the quantizers have 32 levels and the block size for SPC and PQSPC is 64. N(A) is the Gaussian density and X(A) is the double-sided exponential density. The input densities are both zero mean with variance A.

In terms of normalized MSE, PQSPC offers an improvement over SPC of 20.4% for the Gaussian input densities, and 47.5% for the exponential densities. A desirable characteristic of SPC is its relative insensitivity to a change in signal power or statistics. Table 1 demonstrates that PQSPC shares this characteristic. In the unit variance Gaussian case where the optimum quantization scheme is given, the MSE of PQSPC is just double that of the optimum MSE. Further, for a significant change in the input signal power of statistics, PQSPC often out performs that same quantizer.

Table 1. A comparison of normalized MSE between the optimum unit variance Gaussian quantizer (O.G.Q.), the optimum uniform unit variance Gaussian quantizer (U.G.Q.), SPC, and Prequantized SPC (PQSPC).

Density	O.G.Q.	U.G.Q	SPC	PQSPC
N(1)	2.48 E-3	3.82 E-3	7.39 E-3	5.88 E-3
N(2)	6.76 E-3	1.23 E-2	7.39 E-3	5.88 E-3
N(4)	3.63 E-3	5.43 E-2	7.39 E-3	5.88 E-3
X(1)	1.81 E-2	2.65 E-2	2.78 E-2	1.46 E-2
X(2)	5.08 E-2	6.77 E-2	2.78 E-2	1.46 E-2
X(4)	1.13 E-1	1.40 E-1	2.78 E-2	1.46 E-2

IV. DISCUSSION

We began with an investigation of quantization in SPC. We have found error densities and MSE equations that completely characterize the quantization. The results of this investigation indicate that additional quantization can lead to improved MSE performance. This is an interesting concept as it does not follow simply from intuition. Using the concept of additional quantization, a technique called Prequantized Spectral Phase Coding is introduced. It is shown that PQSPC has the same properties as SPC with substantially reduced MSE. Finally, computer examples indicate that PQSPC is often superior to fixed quantization for nonstationary or large dynamic range signals.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the support of the Air Force Office of Scientific Research under grant AFOSR 78-3605.

APPENDIX

DERIVATION OF EQUIVALENT QUANTIZATION ERROR EXPRESSIONS

All quantization is to N levels uniformly spaced from 0 to 2π with output levels $K\ 2\pi/N$ for K=0,1,...,N-1. Using Eqs. (2) and (7) we write ψ_p in terms of the direct quantization levels $\tilde{\theta}_p$ and $\tilde{\gamma}_p$.

$$\psi_p = \tilde{\theta}_p + n_p + \tilde{\gamma}_p + m_p$$

$$\psi_{p+M} = \tilde{\theta}_p + n_p - \tilde{\gamma}_p - m_p \qquad \text{(A1)}$$

Since $\tilde{\theta}_p$ and $\tilde{\gamma}_p$ represent quantized values,

$$\tilde{\theta}_p + \tilde{\gamma}_p = k\,\frac{2\pi}{N}\ , \quad k \text{ an integer}\ .$$

Thus, an equivalent way of expressing ψ_p before quantization is

$$\psi_p = k\,\frac{2\pi}{N} + n_p + m_p \qquad \text{(A2)}$$

Note that $|n_p| \le \pi/N$, $|m_p| \le \pi/N$ and thus

$$-\frac{2\pi}{N} \le n_p + m_p \le \frac{2\pi}{N} \qquad \text{(A3)}$$

The quantization of ψ_p is now described.

$$\hat{\psi}_p = K\,\frac{2\pi}{N}\ , \quad -\frac{\pi}{N} \le n_p+m_p \le \frac{\pi}{N}$$
$$= K\,\frac{2\pi}{N} - \frac{2\pi}{N},\quad -\frac{2\pi}{N} \le n_p+m_p \le -\frac{\pi}{N}$$
$$= K\,\frac{2\pi}{N} + \frac{2\pi}{N},\quad \frac{\pi}{N} \le n_p+m_p \le \frac{2\pi}{N} \qquad \text{(A4)}$$

Recalling $K\,2\pi/N = \tilde{\theta}_p + \tilde{\gamma}_p$, we write Eq. (4) as

$$\hat{\psi}_p = \tilde{\theta}_p + \tilde{\gamma}_p\ , \quad -\frac{\pi}{N} \le n_p+m_p \le \frac{\pi}{N}$$
$$= \tilde{\theta}_p + \tilde{\gamma}_p - \frac{2\pi}{N},\quad -\frac{2\pi}{N} \le n_p+m_p \le -\frac{\pi}{N}$$
$$= \tilde{\theta}_p + \tilde{\gamma}_p + \frac{2\pi}{N},\quad \frac{\pi}{N} \le n_p+m_p \le \frac{2\pi}{N}\ . \qquad \text{(A5)}$$

Similarly,

$$\hat{\psi}_{p+M} = \tilde{\theta}_p - \tilde{\gamma}_p\ , \quad -\frac{\pi}{N} \le n_p-m_p \le \frac{\pi}{N}$$
$$= \tilde{\theta}_p - \tilde{\gamma}_p - \frac{2\pi}{N},\quad -\frac{2\pi}{N} \le n_p-m_p \le -\frac{\pi}{N}$$
$$= \tilde{\theta}_p - \tilde{\gamma}_p + \frac{2\pi}{N},\quad \frac{\pi}{N} \le n_p-m_p \le \frac{2\pi}{N}\ . \qquad \text{(A6)}$$

Using Eqs. (5) and (6) we write e_p in terms of $\hat{\psi}_p$ and $\hat{\psi}_{p+M}$

$$e_p = \theta_p - \hat{\theta}_p = \theta_p - \frac{1}{2}\,(\hat{\psi}_p + \hat{\psi}_{p+M})\ . \qquad \text{(A7)}$$

We examine three examples here for clarity and then state the general results.

Case 1:

$$-\frac{\pi}{N} \le n_p+m_p \le \frac{\pi}{N}\ ,\quad -\frac{\pi}{N} \le n_p-m_p \le \frac{\pi}{N}$$

Thus, $\hat{\psi}_p = \tilde{\theta}_p + \tilde{\gamma}_p$, $\hat{\psi}_{p+M} = \tilde{\theta}_p - \tilde{\gamma}_p$, and

$$e_p = \theta_p - \frac{1}{2}\,[(\tilde{\theta}_p + \tilde{\gamma}_p) + (\tilde{\theta}_p - \tilde{\gamma}_p)]$$
$$= \theta_p - \tilde{\theta}_p = n_p\ .$$

Case 2:

$$-\frac{2\pi}{N} \le n_p+m_p \le -\frac{\pi}{N}\ ,\quad -\frac{\pi}{N} \le n_p-m_p \le \frac{\pi}{N}$$

Thus, $\hat{\psi}_p = \tilde{\theta}_p + \tilde{\gamma}_p - \frac{2\pi}{N}$, $\hat{\psi}_{p+M} = \tilde{\theta}_p - \tilde{\gamma}_p$, and

$$e_p = \theta_p - \frac{1}{2}\,[(\tilde{\theta}_p + \tilde{\gamma}_p - \tfrac{2\pi}{N}) + (\tilde{\theta}_p - \tilde{\gamma}_p)]$$
$$= \theta_p - \tilde{\theta}_p + \frac{\pi}{N} = n_p + \frac{\pi}{N}\ .$$

Case 3:

$$\frac{\pi}{N} \le n_p+m_p \le \frac{2\pi}{N}\ ,\quad -\frac{\pi}{N} \le n_p-m_p \le \frac{\pi}{N}$$

Thus, $\hat{\psi}_p = \tilde{\theta}_p + \tilde{\gamma}_p + \frac{2\pi}{N}$, $\hat{\psi}_{p+M} = \tilde{\theta}_p - \tilde{\gamma}_p$, and

$$e_p = \theta_p - \frac{1}{2}\,[(\tilde{\theta}_p + \tilde{\gamma}_p + \tfrac{2\pi}{N}) + (\tilde{\theta}_p - \tilde{\gamma}_p)]$$
$$= \theta_p - \tilde{\theta}_p - \frac{\pi}{N} = n_p - \frac{\pi}{N}\ .$$

There are five possible pairings of $\hat{\psi}_p$ and $\hat{\psi}_{p+M}$ since

$$|n_p + m_p| > \frac{\pi}{N} \Rightarrow |n_p - m_p| < \frac{\pi}{N}$$

and $$|n_p - m_p| > \frac{\pi}{N} \Rightarrow |n_p + m_p| < \frac{\pi}{N}$$

show that four of the nine available pairings are not allowed. The complete results for e_p are given in Eq. (A8). The results for d_p given in Eq. (9) are obtained in a similar manner.

$$\begin{aligned}
e_p = n_p \quad &, -\frac{\pi}{N} \leq n_p + m_p \leq \frac{\pi}{N} \text{ and} \\
&\quad -\frac{\pi}{N} \leq n_p - m_p \leq \frac{\pi}{N} \\
= n_p + \frac{\pi}{N} \quad &, -\frac{2\pi}{N} \leq n_p + m_p \leq -\frac{\pi}{N} \text{ or} \\
&\quad -\frac{2\pi}{N} \leq n_p - m_p \leq -\frac{\pi}{N} \\
= n_p - \frac{\pi}{N} \quad &, \frac{\pi}{N} \leq n_p + m_p \leq \frac{2\pi}{N} \text{ or} \\
&\quad \frac{\pi}{N} \leq n_p - m_p \leq \frac{2\pi}{N} .
\end{aligned}$$

REFERENCES

[1] N.C. Gallagher, Jr., "Quantization Schemes of the DFT of a Random Time-Series," IEEE Trans. on Info. Theory, IT-24, pp. 156-163, (1978).

[2] N.C. Gallagher, Jr., "Spectral Phase Coding," Proc. of John Hopkins CISS, April 1976.

A Note on Optimal Quantization

JAMES A. BUCKLEW AND NEAL C. GALLAGHER, JR., MEMBER, IEEE

Abstract—For a general class of optimal quantizers the variance of the output is less than that of the input. Also the mean value is preserved by the quantizing operation.

I. INTRODUCTION

J. Max [1] is generally credited with being the first to consider the problem of designing a quantizer to minimize a distortion measure given that the input statistics are known. Max derives necessary conditions for minimizing the mean square quantization error. These results are summarized in the following equations:

$$y_j = \int_{x_{j-1}}^{x_j} xf(x)\,dx / P(x_{j-1} < x \leq x_j) \tag{1}$$

$$\frac{y_j + y_{j+1}}{2} = x_j \tag{2}$$

where $f(x)$ is the probability density of the variable to be quantized and $P(x_{j-1} < x \leq x_j)$ is the probability that x lies in the interval $(x_{j-1}, x_j]$. The y_j are output levels and the x_j are the break points where an input value between x_{j-1} and x_j is quantized to y_j. Fleisher [2] later gave a sufficient condition for Max's equations to be the optimal set.

Typically, the above equations are intractable except for simple input densities, causing some researchers to derive approximate formulae for some common densities. Roe [3] derives an approximation for the input interval endpoints assuming that the widths of these intervals are small, i.e., the number of output levels is large. Wood [4] derives a result which states, in effect, that the variance of the output of a minimum mean-square error quantizer should be less than the input variance. He also states that the significance of his result is that the signal and noise are dependent and that no pseudo-independence of the sort considered by Widrow [4] is possible.

However, Wood's derivation assumes the input density to be five times differentiable and that the quantizer input intervals be very small in order to truncate various Taylor series expansions. Furthermore, the derived expression for the output variance is dependent upon the input interval lengths and the input probability density function evaluated at the midpoints of these intervals.

In this note we derive a generalization of Wood's results that eliminates a number of his approximations and generalizes the results to apply to more than just Max quantizers.

Manuscript received May 5, 1978; revised September 5, 1978. This work was supported in part by the National Science Foundation under Grant ENG-7682426 and in part the Air Force Office of Scientific Research, Air Force Systems Command, USAF under Grant AFOSR-78-3605.

The authors are with the School of Engineering, Purdue University, West Lafayette, IN 47907.

0018-9448/79/0500-0365$00.75

II. Development

In the sequel it is assumed all random variables have finite second moments.

Property 1: The mean value of the output of a minimum mean-square error quantizer is equal to the mean value of the input.

Proof: Consider (3):

$$P(x_{j-1} < x < x_j) y_j = \int_{x_{j-1}}^{x_j} x f(x)\, dx. \tag{3}$$

Sum both sides of the equation from $j = 1$ to $j = N$. The result follows. Property 1 allows us to assume the input density has zero mean without loss of generality.

Property 2: The variance of the output of a minimum mean-square error quantizer is always less than or equal to the input variance. Furthermore the mean-square quantization error is given by the difference of the two.

Proof: Let us consider the mean-square error e between the quantizer's input and output:

$$e = \sum_{i=1}^{N} \int_{x_{i-1}}^{x_i} (y_i - x)^2 f(x)\, dx \tag{4}$$

where x_0 and x_N are the smallest and largest values taken on by the input density and may take on the values $-\infty$ and $+\infty$, respectively. Expanding the integrand and using the expression

$$\sum_{i=1}^{N} \int_{x_{i-1}}^{x_i} x^2 f(x)\, dx = E\{x^2\} = \sigma_x^2$$

and (3), we find that

$$e = \sigma_x^2 - \sum_{i=1}^{N} y_i^2 P(x_{i-1} < x < x_i) \tag{5}$$

where $E\{\cdot\}$ is the statistical expectation operator. But we notice that the last term on the right is the variance of the output, σ_y^2. Since $e > 0$, this implies

$$\sigma_x^2 \geq \sigma_y^2. \tag{6}$$

Property 3: The signal and quantization noise are always nonpositively correlated at the output of the minimum mean-square error quantizer.

Proof: Consider an additive noise model for the quantizing error; by Property 2,

$$E\{(x+n)^2\} = E\{x^2\} + 2E\{xn\} + E\{n^2\} < E\{x^2\}. \tag{7}$$

This implies that $E\{xn\} < 0$. Therefore, since x has mean zero, the correlation coefficient must be nonpositive.

Remark: The above proofs depend only upon the output levels being chosen as the conditional means of the input intervals. Therefore, the same theorem applies to the maximum entropy and equal interval quantizers when the output levels y_i are chosen as above. As indicated by an anonymous reviewer, Property 2 may also be easily derived by averaging the conditional mean square error over all the quantization intervals where we condition on the event of being in one particular quantization interval.

III. Discussion

Some interesting observations can be made when these results are compared with the recent papers of Wise et al. [6] and Sripad and Snyder [7]. In [6] it is shown that the rms bandwidth of any (stationary) Gaussian process must always increase on passing through a memoryless nonlinearity. By using the result of Wise et al., we can say that a Max quantizer operating on a stationary Gaussian input increases the rms bandwidth while simultaneously reducing the variance.

In [7], Sripad and Snyder develop necessary and sufficient conditions for the quantization error to have a uniform distribution. In addition, they derive sufficient conditions for the signal and quantization error to be uncorrelated given that the error is uniformly distributed. Consider the case where the random variable to be quantized is uniformly distributed. The Max quantizer for this case is the equal step size quantizer. It is found that the uniform input density satisfies the conditions for the quantization error to be uniform but fails the conditions for uncorrelatedness. The results contained herein confirm this result and in fact tell us the signal and noise are strictly negatively correlated.

Acknowledgment

We express our thanks to Ed Delp for posing the problem.

References

[1] J. Max, "Quantizing for minimum distortion," *IRE Trans. Inform. Theory*, vol. IT-6, pp. 7–12, Mar. 1960.
[2] P. E. Fleischer, "Sufficient conditions for achieving minimum distortion in a quantizer," *IEEE Intl. Conv. Rec.*, P. 1, pp. 104–111, 1964.
[3] G. M. Roe, "Quantizing for minimum distortion," *IEEE Trans. Inform. Theory*, vol. IT-10, pp. 384–385, Oct. 1964.
[4] B. Widrow, "A study of rough amplitude quantization by means of Nyquist sampling theory," *IRE Trans. Circuit Theory*, vol. CT-3, pp. 226–276, Dec. 1956.
[5] R. C. Wood, "On optimum quantization," *IEEE Trans. Inform. Theory*, vol. IT-5, pp. 248–252, Mar. 1969.
[6] G. L. Wise, A. P. Traganitis, and J. B. Thomas, "The effect of a memoryless nonlinearity on the spectrum of a random process," *IEEE Trans. Inform. Theory*, vol. IT-23, pp. 84–89, Jan. 1977.
[7] A. B. Sripad and D. L. Snyder, "A necessary and sufficient condition for quantization errors to be uniform and white," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 442–448, Oct. 1977.

Correction to 1976–1977 List of Reviewers

It has been brought to our attention that the name of C. E. Sundberg was inadvertently omitted from the 1976–1977 list of reviewers which appeared on pages 654–655 of the November 1977 issue of this Transactions. We sincerely apologize for this oversight.

SOME PROPERTIES OF UNIFORM STEP SIZE QUANTIZERS*

JAMES A. BUCKLEW and NEAL C. GALLAGHER, JR.
School of Electrical Engineering
Purdue University
West Lafayette, IN 47907

*This research was supported by The Air Force Office of Scientific Research under grant AFOSR 78-3605.

ABSTRACT

This paper treats some properties of the optimal mean square error uniform quantizer. It is shown for the OUQ that the mean square error (mse) is given by the input variance minus the output variance. It is shown that $\lim_{N\to\infty} \frac{mse}{\Delta^2/12} \geq 1$ with equality when the support of the random variable is contained in a finite interval. A class of probability densities which have the above limit greater than 1 is given. It is shown that $\lim_{N\to\infty} N^2\, mse = (b-a)^2/12$ where (b-a) is the measure of the smallest interval that contains the support of the input random variable.

In many problems arising in the evaluation or design of a control or communication system, it becomes necessary to predict the performance of a uniform quantizer. Uniform quantizers are of interest because they are usually the simplest quantizer structure to implement. The study of uniform quantization is also of interest because many noise processes in physical systems may be considered to be the noise produced by a uniform quantizing operation. For example the final position of a stepping motor or the line drawn by the pen of a computer plotting device under a continuous control may be considered to be corrupted by a uniform quantizing operation.

Because of the importance of these quantizers several authors have considered various properties of them. Widrow [1] shows that under certain conditions on the characteristic function of the input random variable, the quantization noise is uniformly distributed. Gish and Pierce [2] show that asymptotically the uniform quantizer is optimum in the sense of minimizing the output entropy subject to a fixed mean square error value. Sripad and Snyder [3] later extend Widrow's work to give a sufficient condition for when the quantization error is uniform and uncorrelated with the input random variable.

We will now state and prove some additional properties of these quantizers when we design them to minimize the mean square error. We may write down the analytic expression for the quantizer characteristic g(x) as,

$$g(x) = \begin{cases} a & x<q \\ a+(i+1)\Delta & q+i\Delta<x<q+(i+1)\Delta \quad i=0,\ldots,N-3 \\ a+(N-1)\Delta & x>(N-2)\Delta+q \end{cases} \tag{1}$$

where N is the number of output levels in the quantizer. We see that if x is less than q or greater than q + (N-2)Δ, then x is truncated to a and a + (N-1)Δ respectively. An important parameter of interest is the width of the nontruncation region which equals (N-2)Δ.

The quantizer characteristic g(x) must be optimized with respect to three parameters, q which fixes its position along the x axis, a which

Presented at the Seventeenth Annual Allerton Conference on Communications Control and Computing, October 10-12, 1979.

fixes its position along the y axis, and Δ which specifies the step size of the quantizer. Because it makes little sense to speak of minimizing the mean square error of a random variable with infinite variance, we will always assume $\int_{-\infty}^{\infty} x^2 f(x)\, dx < \infty$.

Property 1

The optimum uniform quantizer preserves the mean of the input random variable.

Proof:

Suppose $g(x)$ is the optimum uniform quantizer. Then we must have

$$\frac{\partial}{\partial \epsilon} \int (x - g(x) + \epsilon)^2 f(x)\, dx \Big|_{\epsilon=0} = 0. \tag{2}$$

This implies,

$$\int x f(x) = \int g(x) f(x). \tag{3}$$

[]

Property 2

For the optimum uniform quantizer we have that

$$a = q - \Delta/2.$$

Proof:

Suppose $g(x)$ is the optimum uniform quantizer. Then we must have

$$\frac{\partial}{\partial \epsilon} \int (g(x - \epsilon) - x)^2 f(x)\, dx \Big|_{\epsilon=0} = 0. \tag{4}$$

$$= \left[\frac{\partial}{\partial \epsilon} \int g(x-\epsilon)^2 f(x)\, dx - \frac{\partial}{\partial \epsilon} \int 2xg(x-\epsilon) f(x)\, dx\right] \Big|_{\epsilon = 0}. \tag{5}$$

$$= \frac{\partial}{\partial \epsilon}\left[\sum_{i=0}^{N-3} (a+(i+1)\Delta)^2 \int_{q+\epsilon+i\Delta}^{q+\epsilon+(i+1)\Delta} f(x)\, dx + a^2 \int_{-\infty}^{q+\epsilon} f(x)\, dx\right.$$

$$+ (a+(N-1)\Delta)^2 \int_{q+\epsilon+(N-2)\Delta}^{\infty} f(x)\, dx$$

$$- 2\left[\sum_{i=0}^{N-3} (a+(i+1)\Delta) \int_{q+\epsilon+i\Delta}^{q+\epsilon+(i+1)\Delta} xf(x)\, dx + a \int_{-\infty}^{q+\epsilon} xf(x)\, dx\right.$$

$$\left.\left. + (a+(N-1)\Delta) \int_{q+\epsilon+(N-2)\Delta}^{\infty} xf(x)\, dx \right] \Big|_{\epsilon=0} = 0. \tag{6}$$

$$= \left[\sum_{i=0}^{N-3} (a+(i+1)\Delta)^2 [f(q+\epsilon+(i+1)\Delta) - f(q+\epsilon+i\Delta)]\right.$$

$$\left. + a^2 f(q+\epsilon) - (a+(N-1)\Delta)^2 f(q+\epsilon+(N-2)\Delta)\right]$$

$$- 2\left[\sum_{i=0}^{N-3} (a+(i+1)\Delta)\ [(q+\varepsilon+(i+1)\Delta)\ f(q+\varepsilon+(i+1)\Delta)\right.$$

$$\left. - (q+\varepsilon+i\Delta)\ f(q+\varepsilon+i\Delta)\right] + a(q+\varepsilon)\ f(q+\varepsilon)$$

$$- (a+(N-1)\Delta)\ (q+\varepsilon+(N-2)\Delta)\ f(q+\varepsilon+(N-2)\Delta) \Big|_{\varepsilon=0} = 0. \qquad (7)$$

Simplifying this expression we obtain

$$(\Delta+2a-2q) \sum_{i=0}^{N-2} f(q+i\Delta) = 0.$$

The solution $\sum_{i=0}^{N-2} f(q+i\Delta) = 0$ is of no interest because without affecting the mean square error, we may always arbitrarily set $f(q+i\Delta) = 0, i = 0,\ldots,N-2$. Hence $\Delta+2a-2q = 0$ which is what we wish to prove.

[]

Property 3

The mean square error of an optimum uniform quantizer is given by the input variance minus the output variance.

Proof:

We again write the mean square error mse as

$$mse = E\{x^2\} - 2\ E\{xg(x)\} + E\{g(x)^2\}. \qquad (8)$$

We wish to optimize this equation with respect to Δ. Using $a = q-\Delta/2$ we first obtain

$$E\{xg(x)\} = \sum_{i=0}^{N-3} (q+(i+\tfrac{1}{2})\Delta)^2 \int_{q+i\Delta}^{q+(i+1)\Delta} xf(x)\ dx$$

$$+ (q-\Delta/2) \int_{-\infty}^{q} xf(x)\ dx + (q+(N-\tfrac{3}{2})\Delta) \int_{q+(N-2)\Delta}^{\infty} xf(x)\ dx \qquad (9)$$

and

$$E\{g(x)^2\} = \sum_{i=0}^{N-3} (q+(i+\tfrac{1}{2})\Delta)^2 \int_{q+i\Delta}^{q+(i+1)\Delta} f(x)\ dx$$

$$+ (q-\tfrac{\Delta}{2})^2 \int_{-\infty}^{q} f(x)\ dx + (q+(N-\tfrac{3}{2})\Delta)^2 \int_{q+(N-2)\Delta}^{\infty} f(x)\ dx. \qquad (10)$$

Now substitute Eq. (9) and Eq. (10) into Eq. (8); take the partial derivative with respect to Δ and set the result equal to zero. We find that

$$E\{xg(x)\} + qE\{g(x)\} = E\{g(x)^2\} + qE\{x\}. \qquad (11)$$

But $E\{g(x)\} = E\{x\}$ for the optimum quantizer. Hence $E\{xg(x)\} = E\{g(x)^2\}$ and we have for the mean square error mse

$$mse = E\{x^2\} - E\{g(x)^2\} \qquad (12)$$

which together with Property 1 finishes the proof.

□

Sripad and Snyder [3] show sufficient conditions for (x-g(x)) to be uniform and uncorrelated with x to be

$$\phi_x(\frac{2\pi n}{\Delta}) = \dot{\phi}_x(\frac{2\pi n}{\Delta}) = 0 \qquad n = \pm 1, \pm 2, \ldots \tag{13}$$

where $\phi_x(\omega)$ is the characteristic function of the input random variable x. Frequently in the analysis of a system corrupted by a uniform quantizing operation the assumption is made that the quantization noise is uncorrelated (sometimes independence is assumed) with the input. The next property demonstrates that this can't be done with the optimum uniform quantizer.

Property 4

Suppose the input probability density is Riemann integrable. Then the quantization noise can't be uncorrelated with the input for the optimum uniform quantizer.

Proof:

Without loss of generality assume E{X} = 0. Now suppose the converse to the property. This implies

$$E\{(x-g(x))x\} = E\{x^2\} - E\{g(x)x\} = 0 \tag{14}$$

But from Property 3

$$\begin{aligned} &E\{xg(x)\} = E\{g(x)^2\} \qquad \text{hence} \\ &E\{x^2\} - E\{g(x)^2\} = 0 \end{aligned} \tag{15}$$

But again from Property 3, the left hand side of Eq. (15) is the mean square error which implies a contradiction. That a probability density function is Riemann integrable necessarily implies that the mean square error for any finite number of output levels is greater than zero (i.e. f(x) has no delta functions).

□

We now state an obvious property which will be used in several subsequent proofs. The proof of property 5 follows from a simple application of the Lebesgue dominated convergence theorem.

Property 5

The mean square error approaches zero for the optimal uniform quantizer as the number of output levels approaches ∞.

Let I = [a,b] be the smallest interval such that $\int_a^b f(x)\,dx = 1$. Note that |a| or |b| may be infinite.

Property 6

Suppose f(x) is Riemann integrable. Then for the optimum uniform quan-

tizer, $\lim_{N\to\infty} (N-2)\Delta = b-a$.

Proof:

Suppose $\lim_{N\to\infty} (N-2)\Delta < b-a$. This implies for N sufficiently large that we are always truncating some finite amount of probability mass which means the mean square error can't go to zero which is a contradiction of the previous property. Hence we have the $\lim_{N\to\infty} (N-2)\Delta \geq b-a$.

Let us suppose $\lim_{N\to\infty} (N-2)\Delta > b-a$. Note that this makes sense only if the random variable is of finite support. Now for N large enough there is no truncation error. It is easy to show as will be done in the next property that for a quantizer with no truncation error, $\lim_{N\to\infty} \frac{mse}{\Delta^2/12} = 1$ for a Riemann integrable density function. So for N sufficiently large $(N-2)\Delta > C > b-a < \infty$. Then

$$1 = \lim_{N\to\infty} \frac{mse}{\Delta^2/12} \leq \lim_{N\to\infty} \frac{mse}{C^2/12(N-2)^2} \text{ or}$$

$$\lim_{N\to\infty} (N-2)^2 \, mse \geq \frac{C^2}{12} \tag{16}$$

Consider now a suboptimal quantizer whose input intervals are given by dividing up the interval I into N-2 equal subintervals. Denote the mean square error or this quantizer as mse_{SUB}, and its step size $\Delta_S = (b-a)/(N-2)$. This quantizer has no truncation error and hence

$$1 = \lim_{N\to\infty} \frac{mse_{SUB}}{\Delta_S^2/12} = \lim_{N\to\infty} \frac{mse_{SUB}}{(b-a)^2/12(N-2)^2} \text{ or}$$

$$\lim_{N\to\infty} (N-2)^2 \, mse_{SUB} = \frac{(b-a)^2}{12} < \frac{C^2}{12} \leq \lim_{N\to\infty} (N-2)^2 \, mse \tag{17}$$

which is a contradiction since we have found a suboptimal quantizer with a better mean square error than the optimal.

[]

Property 7

Suppose the density function is Riemann integrable and $(b-a) < \infty$. Then for the optimal uniform quantizer we have

$$\lim_{N\to\infty} \frac{mse}{\Delta^2/12} = 1.$$

Proof:

From property 6 we know that $\lim_{N\to\infty} (N-2)\,\Delta_0 = b-a < \infty$ where Δ_0 is the optimum Δ. We may design a suboptimum quantizer by dividing the interval I

(smallest interval such that $\int_a^b f(x)\,dx = 1$) into N-2 equal subintervals and use these subintervals as the breakpoints for our quantizer. We will denote the mean square error associated with this quantizer as mse_{SUB} and the step size $\Delta_S \triangleq (b-a)/(N-2)$.

Define

$$M_i \triangleq \sup_{X\in(q+i\Delta)\ q+(i+1)\Delta)} f(x)$$

and

$$m_i = \inf_{X\in(q+i\Delta)\ q+(i+1)\Delta)} f(x) .$$

Then since there is no truncation error for the suboptimal quantizer we have

$$\sum_{i=0}^{N-3} m_i \int_{q+i\Delta_S}^{q+(i+1)\Delta_S} (X-(q+(i+\tfrac{1}{2})\Delta_S))^2\,dx \le mse_{SUB}$$

and

$$mse_{SUB} \le \sum_{i=0}^{N-3} M_i \int_{q+i\Delta_S}^{q+(i+1)\Delta_S} (X-(q+(i+\tfrac{1}{2})\Delta_S))^2\,dx \tag{18}$$

or

$$\frac{\Delta_S^2}{12}\sum_{i=0}^{N-3} m_i\Delta_S \le mse_{SUB} \le \frac{\Delta_S^2}{12}\sum_{i=0}^{N-3} M_i\Delta_S \tag{19}$$

$$\lim_{N\to\infty}\sum_{i=0}^{N-3} m_i\Delta_S \le \lim_{N\to\infty}\frac{mse_{SUB}}{\Delta_S^2/12} \le \lim_{N\to\infty}\sum_{i=0}^{N-3} M_i\Delta_S \tag{20}$$

Now since f(x) is a density function and is Riemann integrable

$$\lim_{N\to\infty}\sum_{i=0}^{N-3} m_i\Delta_S = \lim_{N\to\infty}\sum_{i=0}^{N-3} M_i\Delta_S = 1$$

implies

$$\lim_{N\to\infty}\frac{mse_{SUB}}{\Delta_S^2/12} = 1. \tag{21}$$

Now $\lim_{N\to\infty}\frac{\Delta_S}{\Delta_0} = \lim_{N\to\infty}\frac{(N-2)\Delta_S}{(N-2)\Delta_0} = \frac{\lim_{N\to\infty}(N-2)\Delta_S}{\lim_{N\to\infty}(N-2)\Delta_0} = 1$, which gives automatically $\lim_{N\to\infty}\frac{\Delta_S^2}{\Delta_0^2} = 1$. Now for any quantizer whose nontruncation region covers the

support of the Riemann integrable density function in the limit as N approaches infinity, we may show as above that $\lim_{N\to\infty} \frac{mse}{\Delta^2/12} \geq 1$. This bound is arrived at by ignoring the truncation error and is true for finite or infinite support density functions. We now have the following string,

$$\lim_{N\to\infty} \frac{mse_{SUB}}{\Delta_0^2/12} = \lim_{N\to\infty} \left(\frac{mse_{SUB}}{\Delta_s^2/12}\right)\left(\frac{\Delta_s^2/12}{\Delta_0^2/12}\right)$$

$$= \left(\lim_{N\to\infty} \frac{mse_{SUB}}{\Delta_s^2/12}\right)\left(\lim_{N\to\infty} \frac{\Delta_s^2/12}{\Delta_0^2/12}\right) = 1 \qquad (22)$$

but

$$1 = \lim_{N\to\infty} \frac{mse_{SUB}}{\Delta_0^2/12} \geq \lim_{N\to\infty} \frac{mse_{OPTIMAL}}{\Delta_0^2/12} \geq 1 \qquad (23)$$

$$\text{Or } \lim_{N\to\infty} \frac{mse_{OPTIMAL}}{\Delta_0^2/12} = 1$$

which is what we wanted to prove.

[]

Zador [4] shows that if f(x) is Riemann integrable and $E\{x^{2+\delta}\} < \infty$ for same $\delta > 0$; then we have for the optimal nonuniform quantizer

$$\lim_{N\to\infty} N^2\, mse = ||f||_{1/3}/12$$

where $||f||_{1/3}$ is the $L_{1/3}$ norm. This result shows that for the nonuniform quantizer, the mean square error decreases on the order of $1/N^2$ for large N. Is there a similar property for the optimum uniform quantizer? We now give our next property.

Property 8

Suppose f(x) is Riemann integrable. Then for the optimum uniform quantizer $\lim_{N\to\infty} N^2\, mse = \frac{(b-a)^2}{12}$.

Proof:

Suppose $(b-a) = \infty$. Then $1 \leq \lim_{N\to\infty} \frac{mse}{\Delta^2/12} = \lim_{N\to\infty} \frac{(N-2)^2 mse}{(N-2)^2\Delta^2/12}$

$$= \frac{\lim (N-2)^2 mse}{\lim N^2\Delta^2/12} \qquad (24)$$

but $(N-2)^2\,\Delta^2 \to \infty$ which implies $\lim_{N\to\infty} (N-2)^2\, mse \to \infty$.

If $b-a < \infty$ then $\lim_{N\to\infty} \frac{mse}{\Delta^2/12} = 1$ or $\lim_{N\to\infty}(N-2)^2 mse = \lim_{N\to\infty} N^2 mse =$

$\frac{\lim (N-2)^2 \Delta^2}{12} = \frac{(b-a)^2}{12}$ which finishes the proof.

Discussion

We should note that not everyone employs the same definition of optimum uniform quantizer that we have used. For example Pearlman and Senge [5] have published tables of the optimal uniform Rayleigh quantizer. For their computations, they add the constraints $a = 0$ and that $q = \Delta/2$.

It is interesting to note that properties 1 and 3 are also shared by the optimal nonuniform quantizer as shown in [6]. As a further consequence of these two properties we find that for the N=2 case, the optimum uniform quantizer and the optimum nonuniform quantizer are identical.

Property 7 is one of the more interesting properties proved in this paper. A common approximation to the mean square error of a uniform quantizer has been $\Delta^2/12$. Consider the class of density functions given by

$$f(x) = \frac{(1 + \frac{\delta}{2})}{(1 + |x|)^{3+\delta}} \quad -\infty < x < \infty.$$

We easily see that $\delta = \text{Sup}\{\varepsilon: \int x^{2+\varepsilon} f(x)\, dx < \infty\}$. By straightforward minimization techniques one can show for this class of densities that

$$\lim_{N\to\infty} \frac{mse}{\Delta^2/12} = 1 + \frac{2}{\delta}.$$

Property 8 is of interest because it sets forth a basic difference between uniform and nonuniform quantizers. For the nonuniform quantizer we can expect the mean square error to be of the order $1/N^2$. We can expect this rate of convergence to zero for the uniform quantizer only if the probability density is of finite support. We may show for the optimal uniform Gaussian quantizer that the error is the same or larger than $\ell n\, N/N^2$.

REFERENCES

[1] B. Widrow, "Statistical analysis of amplitude quantized sampled data systems," Trans. Amer. Inst. Elec. Eng., Pt. 11, Applications and Industry, Vol. 79, pp. 555-568, Jan. 1960.

[2] H. Gish and J. N. Pierce, "Asymptotically efficient quantizing," IEEE Trans. Inform. Theory, Vol. IT-14, pp. 676-683, Sept. 1968.

[3] A. B. Sripad and D. L. Snyder, "A necessary and sufficient condition for quantization errors to be uniform and white," IEEE Trans. Acoustics, Speech, and Signal Processing, Vol. ASSP-25, pp. 442-448, Oct. 1977.

[4] P. Zador, "Development and evaluation of procedures for quantizing multivariate distributions," Ph.D. dissertation, Stanford University, Stanford, CA 1964.

[5] W. A. Pearlman and G. H. Senge, "Optimal quantization of the Rayleigh probability distribution," IEEE Trans. Communications, Vol. COM-27, pp. 101-112, Jan. 1979.

[6] J. A. Bucklew and N. C. Gallagher, Jr., "A note on optimum quantization," IEEE Trans. Info. Theory, Vol. IT-25, pp. 365-366, May 1979.

ON THE DETERMINATION OF REGRESSION FUNCTIONS

GARY L. WISE
Department of Electrical Engineering
University of Texas at Austin
Austin, Texas 78712

and

NEAL C. GALLAGHER, JR.
School of Electrical Engineering
Purdue University
West Lafayette, Indiana 47907

ABSTRACT

This paper is concerned with the determination of regression functions from only a partial characterization of the joint distribution. It is shown that statistical information consisting of various moments and joint moments is sufficient to characterize a regression function. An application to regression functionals is also considered.

I. INTRODUCTION

Let X and Y be random variables with Y integrable, i.e. $E\{|Y|\} < \infty$, and consider the regression function of Y on X,

$$m(x) = E\{Y|X=x\}.$$

As is well known, $m(\cdot)$ is a Borel measurable function, and it frequently arises in engineering applications. For example, if Y is a second order random variable, then the minimum mean squared error estimate of Y in terms of X is given by m(X) [1, pp. 77-78].

In some cases $m(\cdot)$ has a particularly simple form. For example, if X and Y are jointly Gaussian with respective means m_X and m_Y, respective variances $\sigma_X^2 > 0$ and σ_Y^2, and correlation coefficient ρ, then

$$m(x) = ax + b, \tag{1}$$

where $a = (\sigma_Y/\sigma_X)\rho$ and $b = m_Y - am_X$. However, in the case of jointly Gaussian random variables, m_X, m_Y, σ_X, σ_Y, and ρ completely determine the bivariate distribution of the two random variables.

In more general cases, the question arises as to how much information about the bivariate distribution is required to determine the regression function. If X and Y are two second order random variables that are separable in the sense of Nuttall [2], then the regression function $m(\cdot)$ has the form given by (1). However, knowing that two second order random variables are separable in the sense of Nuttall, and knowing the means, variances, and the correlation coefficient is not sufficient to determine the bivariate distribution of the two random variables. Notice that any two random variables whose bivariate characteristic function is elliptically symmetric are separable in the sense of Nuttall [3].

As we have seen, there exists a class of joint distributions such that the regression function can be determined knowing that the two random variables belong to that class and also knowing means, variances, and the correlation coefficient. However, it might seem reasonable to conjecture that in more general cases, the regular conditional distribution [4] of Y

Presented at the Seventeenth Annual Allerton Conference on Communication, Control, and Computing, October 10-12, 1979; to be published in the Proceedings of the Conference.

given X=x is required. Although the regular conditional distribution of Y given X=x is sufficient to determine m(x), in the next section we will see that it is never necessary.

In this paper we will be concerned with statistical information such that there can be only one regression function consistent with the given statistical information. In the next section we consider the regression of Y on a random variable and then on a random vector. Then in the following section we consider the regression functional, that is, the regression of Y on a random process.

II. DEVELOPMENT

Let Y be a second order random variable, let X be a random variable with compact support, and let $m(\cdot)$ be given by Eq. (1). Define the measure μ on the Borel sets of $\mathbb{R}$ by

$$\mu(A) = P(X \in A) ,$$

and let $\|\cdot\|$ denote the $L_2(\mu)$ norm. We will say that a polynomial has max degree N if the degree of the polynomial is no greater than N. We note that for any $\epsilon > 0$, if N is sufficiently large, there exists a polynomial of max degree N $P_N(x)$ such that

$$\| m - P_N \| < \epsilon. \tag{2}$$

That is, there exists a continuous function $h(\cdot)$ such that [5]

$$\| m - h \| < \epsilon/2 ,$$

and by the Weierstrass Theorem there exists a polynomial P_N of max degree N with N sufficiently large such that

$$\| h - P_N \| < \epsilon/2 .$$

Thus Eq. (2) follows by the triangle inequality. Hence there exists a sequence of polynomials $P_N(x)$ such that

$$P_N(x) \to m(x) \qquad \text{in } L_2(\mu) .$$

Let $Q_N(x)$ be the polynomial of max degree N that is closer to m(x) (in $L_2(\mu)$) than any other polynomial of max degree N. We note in passing that $Q_N(x)$ is uniquely defined a.e. $[\mu]$ by the Projection Theorem. That is, there may exist more than one representation of $Q_N(x)$ (i.e. with different coefficients) but they are all equal a.e. $[\mu]$. From the preceding observations, we have that

$$Q_N(x) \to m(x) \qquad \text{in } L_2[\mu] .$$

Express the polynomial $Q_N(x)$ as

$$Q_N(x) = \sum_{j=0}^{N} a_j(N)\, x^j .$$

It follows from the Projection Theorem that the $a_j(N)$ can be determined from the relation

$$E\left\{\left[m(X) - \sum_{j=0}^{N} a_j(N)\, X^j\right] X^k\right\} = 0, \qquad k = 0, 1, 2, \ldots, N .$$

This is equivalent to

$$E\{X^k Y\} = \sum_{j=0}^{N} a_j(N)\, E\{X^{j+k}\}, \quad k = 0, 1, 2, \ldots, N . \tag{3}$$

Thus we have seen that from a knowledge of

$$E\{X^k\} , \quad k = 1, 2, \ldots$$

and

$$E\{YX^k\} , \; k = 0, 1, 2, \ldots ,$$

we can construct a sequence of polynomials $Q_N(x)$ that converge in $L_2(\mu)$ to $m(x)$.

Now let X be an arbitrary random variable. Let g be an invertible Borel measurable function whose range is bounded. Define the random variable $\tilde{X}$ as $\tilde{X} = g(X)$, and the measure $\tilde{\mu}$ on the Borel sets of $\mathbb{R}$ by $\tilde{\mu}(A) = P(\tilde{X} \in A)$. From the above discussion, we see that

$$\tilde{m}(x) = E\{Y | \tilde{X} = x\}$$

is determined a.e.$[\tilde{\mu}]$ by the quantities

$$E\{\tilde{X}^k\} , \qquad k = 1, 2, \ldots \tag{4}$$

and

$$E\{Y\tilde{X}^k\} , \qquad k = 0, 1, 2, \ldots . \tag{5}$$

Let $\tilde{Q}_N(x)$ denote the polynomial of max degree N constructed in the preceding fashion. Then

$$\tilde{Q}_N(x) \to \tilde{m}(x) \qquad \text{in } L_2(\tilde{\mu}) .$$

Notice that $m(x) = \tilde{m}[g(x)]$. From a change of variables result [6, p. 182], we have that

$$\int_{g(\mathbb{R})} [\tilde{Q}_N(x) - \tilde{m}(x)]^2 \, \tilde{\mu}(dx) = \int_{\mathbb{R}} [\tilde{Q}_N[g(x)] - m(x)]^2 \, \mu(dx) .$$

Therefore, $\tilde{Q}_N[g(x)] \to m(x)$ in $L_2(\mu)$.

Now we will remove the restriction that Y be second order. Assume that Y is an integrable random variable and let

$$G_k(y) = \begin{cases} y \text{ if } |y| \le k \\ 0 \text{ if } |y| > k \end{cases} .$$

Then $G_k(Y)$ is a second order random variable and [1, p. 23]

$$E\{G_k(Y) | X=x\} \to E\{Y | X=x\} \quad \text{a.e.}[\mu] .$$

Since $|G_k(Y)-Y| \le |Y|$ and $|Y|$ is integrable, we have that $E\{G_k(Y)|X=x\} \to m(x)$ in $L_1(\mu)$ by the dominated convergence theorem [6, pp. 124-125].

Thus from a knowledge of the quantities in Eqs. (4) and (5) we can derive a sequence of estimates for $E\{G_k(Y)|X=x\}$ which converges in $L_2(\mu)$, and consequently in $L_1(\mu)$ (see, for example, [7]). Also, $E\{G_k(Y)|X=x\}$ converges to $E\{Y|X=x\}$ in $L_1(\mu)$. Thus, by a straightforward diagonalization procedure, we can derive a sequence of estimates which converges in $L_1(\mu)$ to $m(x)$. These results are summarized in the following theorem.

Theorem 1: Let Y be an integrable random variable, let X be an arbitrary random variable, and let g be an invertible Borel measurable function mapping the reals into a bounded set. Then the regression function m is determined a.e.[μ] by the quantities

$$E\{[g(X)]^k\} , \quad k = 1, 2, \ldots$$

and

$$E\{Y[g(X)]^k\} , \quad k = 0, 1, 2, \ldots .$$

Consider for the moment the case where X and Y are independent. In this case a solution to Eq. (3) is given by

$$a_0(N) = E\{Y\}$$

$$a_j(N) = 0 , \; j > 0 ,$$

and we get that $m(x) = E\{Y\}$.

Now consider the following two different bivariate density functions:

$$f_1(x,y) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{(y-\rho x)^2}{2\sigma^2}\right] I_{[0,1]}(x)$$

$$f_2(x,y) = x I_{[\rho x-1,\rho x+1]}(y)\; I_{[0,1]}(x),$$

where $\sigma > 0$, $\rho \in (-1,1)$, and I denotes the indicator function. Assuming that the density of (X,Y) is given by f_1, we find that

$$E\{X^k\} = \frac{1}{k+1}$$

$$E\{YX^k\} = \frac{\rho}{k+2} .$$

In this case, for $N \geq 1$, a solution to Eq. (3) is given by

$$a_1(N) = \rho \tag{6}$$

$$a_j(N) = 0 , \; j \neq 1 , \tag{7}$$

and we conclude that

$$m(x) = \rho x . \tag{8}$$

If we assume that the density of (X,Y) is given by f_2, we find that

$$E\{X^k\} = \frac{2}{k+2}$$

$$E\{YX^k\} = \frac{2\rho}{k+3} .$$

In this case, for $N \geq 1$, Eqs. (6) and (7) still satisfy Eq. (3), and the regression function is once again given by Eq. (8). Thus, in this example, the two pairs of marginal densities are not the same, the conditional densities of Y given X=x are not the same, and the moment sequences are not the same; however, the moment sequences are sufficient to characterize the conditional expectations, which are identical. Numerous other similar examples may easily be constructed.

Now we will consider the regression of Y upon a set of random variables. Let X be an arbitrary random vector taking values in $\mathbb{R}^n$, and let μ be defined on the Borel sets of $\mathbb{R}^n$ by

$$\mu(B) = P(X \in B) .$$

Lemma 1: If μ has compact support, then the class of all polynomials is dense in $L_2(\mu)$.

Proof: Let q be an arbitrary element in $L_2(\mu)$. For any $\varepsilon > 0$, there exists [5] a function $h: \mathbb{R}^n \to \mathbb{R}$ which is continuous and has compact support such that

$$\| q-h \| < \varepsilon/2 .$$

By the Stone-Weierstrass Theorem [8] there exists a polynomial p in n variables such that

$$\| h-p \| < \varepsilon/2 ,$$

and thus by the triangle inequality

$$\| p-q \| < \varepsilon .$$

QED

We recall that the degree of a monomial in n variables is the sum of the powers of the variables, and the degree of a polynomial is the degree of the monomial having the largest degree over all the monomials in the polynomial with nonzero coefficients. There are

$$C(n,d) = \binom{n+d-1}{d}$$

monomials of degree d in n variables [9].

Assume that Y is a second order random variable, and define m(x) by Eq. (1), where x is now an element of $\mathbb{R}^n$. Assume that μ has compact support. Let $Q_N(x)$ be the polynomial of max degree N which is closer, in the $L_2(\mu)$ norm, to m(x) than any other polynomial of max degree N.

Consider a monomial in n variables of degree d. There will be C(n,d) of them. Order them lexicographically by the powers of the components of x, and let $m_{jd}(x)$ denote the j-th monomial of degree d. Then $Q_N(x)$ can be expressed as

$$Q_N(x) = \sum_{d=0}^{N} \sum_{j=1}^{C(n,d)} a_{jd}(N) m_{jd}(x) .$$

It follows from the Projection Theorem that the coefficients $a_{jd}(N)$ are given by the solution to the following set of equations:

$$E\{Ym_{ik}(X)\} = \sum_{d=0}^{N} \sum_{j=1}^{C(n,d)} a_{jd}(N)\ E\{m_{jd}(X)\ m_{ik}(X)\}, \tag{9}$$

$k = 0, 1, \ldots, N$ and $i = 1, \ldots, C(n,k)$. If the coefficients $a_{jd}(N)$ satisfy Eq. (9), then it follows from Lemma 1 that

$$Q_N(x) \to m(x) \qquad \text{in } L_2(\mu) .$$

Now we remove the assumption that X has compact support and let X be an arbitrary random vector taking values in $\mathbb{R}^n$. Let g be an invertible Borel measurable function mapping $\mathbb{R}^n$ into a bounded subset of $\mathbb{R}^n$, and let $\tilde{X} = g(X)$. We see that

$$\tilde{m}(x) = E\{Y|\tilde{X}=x\}$$

is determined a.e.$[\tilde{\mu}]$, where $\tilde{\mu}(A) = \mu[g^{-1}(A)]$, by the quantities

$$E\{m_{jd}(\tilde{X})\}$$

and

$$E\{Ym_{jd}(\tilde{X})\}$$

for $d = 0, 1, 2, \ldots$ and $j = 1, \ldots, C(n,d)$. Let $\tilde{Q}_N(x)$ be the polynomial of max degree N determined in the preceding fashion. Then, similar to the development of Theorem 1, we can employ a change of variables result [6, p. 182] to conclude that

$$\tilde{Q}_N[g(x)] \to m(x) \qquad \text{in } L_2(\mu) .$$

A chopping argument as in the development of Theorem 1 allows us to remove the second order restriction on Y. Then a straightforward diagonalization procedure results in a sequence of estimates which converges to m(x) in $L_1(\mu)$. This result is summarized in the following theorem.

Theorem 2: Let Y be an integrable random variable, let X be an arbitrary random vector taking values in $\mathbb{R}^n$, and let g be an invertible Borel measurable function mapping $\mathbb{R}^n$ into a bounded subset of $\mathbb{R}^n$. Then the regression function m is determined a.e.$[\mu]$ by the quantities

$$E\{m_{jd}[g(X)]\} \quad \text{and} \quad E\{Ym_{jd}[g(X)]\}$$

for $d = 0, 1, 2, \ldots$ and $j = 1, \ldots, C(n,d)$.

III. REGRESSION FUNCTIONALS

As before, assume that Y is an integrable random variable, but now let T be an infinite subset of $\mathbb{R}$ and let $\{X(t), t \in T\}$ be a random process. Let S denote the space of all extended real valued functions defined on T, and let $\mathcal{B}(S)$ denote the σ-algebra on S generated by the class of all cylinders in S. Let $\mathcal{B}$ denote the Borel sets of $\mathbb{R}$. Then the regression functional

$$m[x(t), t \in T] = E\{Y | X(t) = x(t), t \in T\}$$

is a measurable function from $(S, \mathcal{B}(S))$ to $(\mathbb{R}, \mathcal{B})$ (see, for example, [10]).

Let μ be the measure induced on $\mathcal{B}(S)$ by $\{X(t), t \in T\}$. That is, for any cylinder C in S, $\mu(C) = P(\{X(t), t \in T\} \in C)$, and μ is extended to $\mathcal{B}(S)$ via Kolmogorov's Theorem (see, for example, [11]).

It follows from [1, pp. 21, 604] that there exists a countable subset of T, say $\tilde{T} = \{t_1, t_2, \ldots\}$, depending on the random variable Y, such that

$$E\{Y | X(t) = x(t), t \in T\} = E\{Y | X(t) = x(t), t \in \tilde{T}\} \text{ a.e.}[\mu] .$$

Let

$$M = E\{Y | X(t), t \in \tilde{T}\},$$

$$M_n = E\{Y | X(t_1), \ldots, X(t_n)\},$$

$$\mathcal{F} = \sigma\{X(t), t \in \tilde{T}\},$$

and

$$\mathcal{F}_n = \sigma\{X(t_1), \ldots, X(t_n)\}.$$

Then from the properties of iterated conditional expectations [1, p. 37], it follows that

$$E\{M_{n+1} | \mathcal{F}_n\} = M_n \quad \text{wp1} ,$$

and hence $\{M_n, \mathcal{F}_n, n \geq 1\}$ is a martingale. It follows from [1, p. 332] that $M_n \to M$ wp1. Since $E\{|M_n|\} \leq E\{|Y|\} < \infty$, it follows from a martingale convergence theorem [1, p. 319] due to Doob that $E\{|M_n - M|\} \to 0$. This is equivalent to

$$E\{Y | X(t_i) = x(t_i), i=1,\ldots,n\} \to E\{Y | X(t) = x(t), t \in \tilde{T}\}$$

in $L_1(\mu)$. Notice that Theorem 2 is applicable to $E\{Y | X(t_i) = x(t_i), i=1,\ldots,n\}$. Thus a straightforward diagonalization procedure results in a sequence of estimates which converges to $m[x(t), t \in T]$ in $L_1(\mu)$. This result is summarized in the following theorem.

Theorem 3: Let Y be an integrable random variable and let $\{X(t), t \in T\}$ be a random process. Let $\{g_n, n=1,2,\ldots\}$ be a sequence of functions where g_n is an invertible Borel measurable function from $\mathbb{R}^n$ to a bounded subset of $\mathbb{R}^n$. Assume that for all positive integers n and for all sets

of n points in T, say $t_1, \ldots, t_n$, the quantities

$$E\{m_{jd}(g_n[X(t_1), \ldots, X(t_n)])\}$$

and

$$E\{Ym_{jd}(g_n[X(t_1), \ldots, X(t_n)])\}$$

for $d = 0, 1, 2, \ldots$ and $j = 1, \ldots, C(n,d)$ are known. Then up to μ equivalence, there is only one possible regression functional $m[x(t), t \in T] = E\{Y|X(t) = x(t), t \in T\}$.

ACKNOWLEDGEMENT

This research was supported by the Air Force Office of Scientific Research, Air Force Systems Command, USAF, under Grants AFOSR-76-3062 and AFOSR-78-3605.

REFERENCES

1. J. L. Doob, Stochastic Processes, Wiley, New York, 1953.

2. A. H. Nuttall, "Theory and Application of the Separable Class of Random Processes," Technical Report 343, Research Laboratory of Electronics, Massachusetts Institute of Technology, May 26, 1958.

3. D. K. McGraw and J. F. Wagner, "Elliptically Symmetric Distributions," IEEE Trans. Inform. Th., Vol. IT-14, pp. 110-120, January 1968.

4. L. Breiman, Probability, p. 79, Addison-Wesley, Reading, Mass., 1968.

5. W. Rudin, Real and Complex Analysis, p. 71, McGraw-Hill, New York, 1974.

6. N. Dunford and J. T. Schwartz, Linear Operators Part I: General Theory, Interscience, New York, 1957.

7. M. Loève, Probability Theory, p. 164, Van Nostrand, New York, 1963.

8. J. Dieudonné, Foundations of Modern Analysis, p. 139, Academic Press, New York, 1969.

9. R. W. Brockett, "Lie Algebras and Lie Groups in Control Theory," in Geometric Methods in System Theory, D. Q. Mayne and R. W. Brockett, eds., Reidel, The Netherlands, 1973, pp. 43-82.

10. I. I. Gihman and A. V. Skorohod, The Theory of Stochastic Processes I, p. 34, Springer-Verlag, New York, 1974.

11. P. Billingsley, Probability and Measure, p. 433, Wiley, New York, 1979.

Quantization Schemes for Bivariate Gaussian Random Variables

JAMES A. BUCKLEW AND NEAL C. GALLAGHER, JR., MEMBER, IEEE

Abstract—**The problem of quantizing two-dimensional Gaussian random variables is considered. It is shown that, for all but a finite number of cases, a polar representation gives a smaller mean square quantization error than a Cartesian representation. Applications of the results to a transform coding scheme known as spectral phase coding are discussed.**

I. Introduction

CONSIDER a two-dimensional Gaussian random variable X with independent components. For many applications in signal processing and digital communications it is necessary to represent this quantity by a finite set of values. One possible representation of X is in Cartesian coordinates, obtained by individually quantizing the two rectangular components of X. An alternative representation, in polar coordinates, is obtained by quantizing the magnitude and phase angle of X.

In [1] experimental data are put forward to show that, in all of the cases treated, polar formatting is better than rectangular. The purpose of this paper is to give a more rigorous treatment of the problem and to ascertain which of the representations leads to a smaller mean square quantization error.

In the first section we will derive the exact error expression for the polar format. The second and third sections deal with computer simulations of the expression and compare the polar and rectangular formats. It is shown that, in almost all cases, the polar format gives a smaller quantization error.

If the polar format is to be used, the question arises as to the best ratio of the number of phase quantizer levels to the number of magnitude quantizer levels. Pearlman [2] used distortion rate theory to derive a bound for this expression. In the fourth section we derive an asymptotic expression that agrees with the Pearlman result and perform computer simulations showing the validity of this bound.

In the fifth section we apply the above results to a transform coding scheme, spectral phase coding (SPC). Theoretical arguments are given for the observed robustness of SPC, and an exact error expression is derived. Computer simulations are then made demonstrating the robustness of SPC.

Manuscript received November 18, 1977; revised December 18, 1978. This work was supported by the Air Force Office of Scientific Research, Air System Command, USAF, under grant AFOSR-78-3605.

The authors are with the School of Electrical Engineering, Purdue University, West Lafayette, IN 47907.

II. Development

Consider the mean square quantization error E_p of a polar format representation:

$$E_p = \sum_{j=1}^{N_\theta} \sum_{i=1}^{N_r} \int_{c_{j-1}}^{c_j} \int_{a_{i-1}}^{a_i} |r\exp(j\theta) - b_i \exp(jd_j)|^2 \cdot \frac{f_r(r)\,dr\,d\theta}{2\pi}. \quad (1)$$

where N_θ and N_r are the number of levels in the phase and magnitude quantizers, respectively. The b_i and d_j are the output levels of the magnitude and phase quantizers corresponding to input levels lying in the intervals $(a_{i-1}, a_i]$ and $(c_{j-1}, c_j]$, respectively. The function $f_r(r)$ is the input density of the magnitude which is Rayleigh distributed and independent of the random phase θ which is uniformly distributed over $[-\pi, \pi]$.

After squaring out the integrand and integrating over θ from c_{j-1} to c_j, we obtain

$$E_p = \sum_{j=1}^{N_\theta} \sum_{i=1}^{N_r} \int_{a_{i-1}}^{a_i} \left[(c_j - c_{j-1})\left[r^2 + b_i^2\right] - 2rb_i\left[\sin(c_j - d_j) - \sin(c_{j-1} - d_j)\right]\right] \frac{f(r)\,dr}{2\pi}. \quad (2)$$

Setting $\partial E_p / \partial d_j = 0$ leads to the equations

$$c_j - d_j = \frac{\pi}{N_\theta} = d_j - c_{j-1} \quad (3a)$$

$$c_j - c_{j-1} = \frac{2\pi}{N_\theta}, \quad (3b)$$

for $j = 1, \cdots, N_\theta$. It should be noted that these are simply the equations for a uniform quantizer. Consequently, the expression for mean square error becomes

$$E_p = \sum_{i=1}^{N_r} \int_{a_{i-1}}^{a_i} \left[r^2 + b_i^2 - 2rb_i \operatorname{sinc}(1/N_\theta)\right] f(r)\,dr, \quad (4)$$

where $\operatorname{sinc}(\cdot) = \sin\pi(\cdot)/\pi(\cdot)$. A differentiation with respect to b_i yields the optimum b_i as

$$b_i = \operatorname{sinc}(1/N_\theta) \frac{\int_{a_{i-1}}^{a_i} rf(r)\,dr}{\int_{a_{i-1}}^{a_i} f(r)\,dr}. \quad (5)$$

0018-9448/79/0900-0537$00.75

Substituting this value back into (4), we find

$$E_p = \overline{r^2} - \sum_{i=1}^{N_r} \operatorname{sinc}^2(1/N_\theta) \frac{\left[\int_{a_{i-1}}^{a_i} rf(r)\,dr\right]^2}{\int_{a_{i-1}}^{a_i} f(r)\,dr}. \quad (6)$$

where the upper bar indicates the statistical expectation operator. Let $E(N_r, r)$ denote the mean square quantization error produced by an optimal, one-dimensional, N_r output level, Rayleigh quantizer. It is shown in [3] that $E(N_r, r)$ is given by the difference between the variance of the quantizer input and the variance of the output. Hence $E(N_r, r)$ may be written as

$$E(N_r, r) = \overline{r^2} - \sum_{i=1}^{N_r} b_i'^2 \left[\int_{a_{i-1}'}^{a_i'} f(r)\,dr\right], \quad (7)$$

where the $\{a_i'\}$ are the quantizer input interval endpoints and the $\{b_i'\}$ are the quantizer output levels. Max [4] shows that the $\{b_i'\}$ and $\{a_i'\}$ satisfy

$$a_i' = \frac{b_i' + b_{i+1}'}{2} \quad (8a)$$

$$b_i' = \frac{\int_{a_{i-1}'}^{a_i'} rf(r)\,dr}{\int_{a_{i-1}'}^{a_i'} f(r)\,dr}. \quad (8b)$$

These equations may be written as

$$a_i' = \frac{\int_{a_{i-1}'}^{a_i'} rf(r)\,dr}{2\int_{a_{i-1}'}^{a_i'} f(r)\,dr} + \frac{\int_{a_i'}^{a_{i+1}'} rf(r)\,dr}{2\int_{a_i'}^{a_{i+1}'} f(r)\,dr}. \quad (9)$$

Minimizing (4) with respect to the a_i yields

$$a_i = \frac{b_i + b_{i+1}}{2\operatorname{sinc}(1/N_\theta)}, \quad (10)$$

and substituting (5) into the above gives

$$a_i = \frac{\int_{a_{i-1}}^{a_i} rf(r)}{2\int_{a_{i-1}}^{a_i} f(r)\,dr} + \frac{\int_{a_i}^{a_{i+1}} rf(r)\,dr}{2\int_{a_i}^{a_{i+1}} f(r)\,dr}. \quad (11)$$

which is identical to (9). Fleisher [5] shows that Max's conditions (i.e., (8a) and (8b)) are necessary and sufficient for the optimality of the Rayleigh quantizer. Thus we are assured that the solutions to (11) are unique, leading us to the conclusion that

$$a_i = a_i'.$$

The polar format error expression then becomes

$$E_p = \operatorname{sinc}^2(1/N_\theta) E(N_r, r) + \left(1 - \operatorname{sinc}^2(1/N_\theta)\right) \overline{r^2}. \quad (12)$$

If we assume bit rate limited signal transmission, then we must constrain the product of N_r and N_θ to be less than or equal to some constant, let us say N. To compare the rectangular and polar formats, it is assumed that the product of N_x and N_y, the number of output levels of the rectangular format quantizers, must also equal N. By use of symmetry arguments it may be shown that, for optimal rectangular format operation, N_x must equal N_y. Therefore,

$$N_x = N_y = N^{1/2}. \quad (13)$$

Let $E(N_x, g)$ denote the mean square quantization error produced by an optimal N_x output level Gaussian quantizer. The rectangular format error E_{rect} is given by

$$E_{\text{rect}} = 2E(N_x, g) = 2E(\sqrt{N}, g). \quad (14)$$

The problem is now to compare (12) with (14).

III. Exact Computer Simulation

In this section we make use of Max's [4] tabulated results for $E(N_x, g)$. Max gives values of this function from $N_x = 1$ to $N_x = 36$. We duplicate Max's work for the Rayleigh quantizer and obtain values for $E(N_r, r)$. Using an exhaustive search, we compute the smallest values of error obtainable for (12) and (14) for values of N from 1 to 2000. For all of these cases, there are only 31 values of N for which the rectangular format is better. These values

TABLE I
VALUES OF N WHERE RECTANGULAR FORMAT IS SUPERIOR TO POLAR FORMAT

Based upon exact expressions	Based upon approximate expressions
	1, 2, 3, 4,
6	6
8	8
9	9
12	12
13	13, 15
16	16
17	17
20	20
21	21
25	24, 25
26	26, 27, 28, 29, 30, 31, 32
35	35
36	36
37	37
38	38
42	42
43	43, 44, 48
49	49
50	50
51	51
56	56
57	57
58	58
59	59
63	63
64	64, 65, 66, 67
72	72
73	73
74	74, 81, 82, 83, 96, 99
100	
101	
	110, 111, 112, 113

TABLE II
A TABULATION OF THE RELATIVE EFFICIENCY $\eta = (E_p - E_r)/E_p$ OF POLAR QUANTIZATION OVER THAT OF RECTANGULAR QUANTIZATION, THE BEST NUMBER OF MAGNITUDE LEVELS N_r, AND THE BEST NUMBER OF RECTANGULAR FORMAT LEVELS N_x, AS A FUNCTION OF N

N	η	N_r	N_x	N	η	N_r	N_x	N	η	N_r	N_x
1	.000	1	1	51	3.801	4	7	101	.578	5	10
2	-.001	1	1	52	-3.544	4	7	102	-4.424	6	10
3	-28.572	1	1	53	-3.544	4	7	103	-4.424	6	10
4	-.005	1	2	54	-.990	4	6	104	-4.424	6	10
5	-16.226	1	2	55	-3.459	5	6	105	-4.424	6	10
6	2.468	1	2	56	1.613	4	7	106	-4.424	6	10
7	-4.084	1	2	57	1.613	4	7	107	-4.424	6	10
8	3.325	2	2	58	1.613	4	7	108	-6.469	6	9
9	22.679	1	3	59	1.613	4	7	109	-6.469	6	9
10	-.703	2	3	60	-5.316	5	7	110	-1.554	6	10
11	-.703	2	3	61	-5.316	5	7	111	-1.554	6	10
12	.620	2	3	62	-5.316	5	7	112	-1.554	6	10
13	.620	2	3	63	3.655	5	7	113	-1.554	6	10
14	-15.048	2	3	64	4.369	4	8	114	-6.812	6	10
15	-1.004	2	3	65	-1.544	5	8	115	-6.812	6	10
16	1.936	2	4	66	-1.544	5	8	116	-6.812	6	10
17	1.936	2	4	67	-1.544	5	8	117	-6.204	6	9
18	-6.640	2	4	68	-1.544	5	8	118	-6.204	6	9
19	-6.640	2	4	69	-1.544	5	8	119	-8.422	7	9
20	4.377	2	4	70	-6.538	5	7	120	-4.120	6	10
21	1.220	3	4	71	-6.538	5	7	121	-1.919	6	11
22	-.660	2	4	72	.689	5	8	122	-1.919	6	11
23	-.660	2	4	73	.689	5	8	123	-1.919	6	11
24	-2.631	3	4	74	.689	5	8	124	-1.919	6	11
25	6.493	3	5	75	-6.442	5	8	125	-1.919	6	11
26	6.493	3	5	76	-6.442	5	8	126	-6.151	6	11
27	-5.911	3	5	77	-6.442	5	8	127	-6.151	6	11
28	-5.911	3	5	78	-6.442	5	8	128	-6.151	6	11
29	-5.911	3	5	79	-6.442	5	8	129	-6.151	6	11
30	-1.022	3	5	80	-4.180	5	8	130	-2.146	6	10
31	-1.022	3	5	81	-.972	5	9	131	-2.146	6	10
32	-1.022	3	5	82	-.972	5	9	132	-1.865	6	11
33	-9.643	3	5	83	-.972	5	9	133	-4.015	7	11
34	-9.643	3	5	84	-2.232	6	9	134	-4.015	7	11
35	1.471	3	5	85	-6.499	5	9	135	-4.015	7	11
36	1.388	3	6	86	-6.499	5	9	136	-4.015	7	11
37	1.388	3	6	87	-6.499	5	9	137	-4.015	7	11
38	1.388	3	6	88	-2.789	5	8	138	-5.298	6	11
39	-4.288	3	6	89	-2.789	5	8	139	-5.298	6	11
40	-3.440	4	5	90	-1.765	5	9	140	-8.395	7	10
41	-3.440	4	5	91	-1.765	5	9	141	-8.395	7	10
42	3.885	3	6	92	-1.765	5	9	142	-8.395	7	10
43	3.885	3	6	93	-1.765	5	9	143	-2.599	7	11
44	-2.196	4	6	94	-1.765	5	9	144	-.750	7	12
45	-2.196	4	6	95	-6.090	5	9	145	-.750	7	12
46	-2.196	4	6	96	-8.522	6	9	146	-.750	7	12
47	-2.196	4	6	97	-8.522	6	9	147	-5.660	7	12
48	-1.140	4	6	98	-8.522	6	9	148	-5.660	7	12
49	3.801	4	7	99	-.593	6	9	149	-5.660	7	12
50	3.801	4	7	100	.578	5	10	150	-5.660	7	12

for N correspond in general to regions where N is a perfect square. Apparently, for values of N greater than 101, polar formatting is always the better of the two methods. The left column of Table I contains a listing of the 31 values of N for which rectangular format gives smaller error. Table II gives an indication of the relative efficiency of polar and rectangular formatting by tabulating $(E_p - E_r)/E_p$ for values of N from 1 to 150. Also in Table II may be found the best number of magnitude levels N_r (with N_θ = greatest integer less than N/N_r) and the best number of rectangular format levels N_x (with N_y = greatest integer less than N/N_x) for each value of N from 1 to 150. For values of N larger than 2000, we may make use of approximation methods.

IV. Approximate Computer Simulation

Wood [6] describes a technique whereby one can approximate the mean square error of an optimal quantizer for large N. He then gives an expression for the error of an N level Gaussian quantizer which agrees to within about one percent with the actual computed mean square error given by Max [4]. This error expression is

$$E(N_x, g) = \frac{2.73 N_x \sigma^2}{(N_x + 0.853)^3}. \tag{15}$$

Using Wood's approximations, we obtain for the Rayleigh density a similar error expression which also agrees well with the actual computed error. This error expression is

$$E(N_r, r) = \frac{0.9287 N_r \sigma^2}{(0.596 + N_r)^3}. \tag{16}$$

By use of these approximate error expressions, we again find the values of N where rectangular format gives smaller error than polar format. Computer simulations are run up to a value of $N = 10^6$. We find that for values of N greater than 113, polar format is always better.

Table I summarizes the results of the last two sections. In the first column we find the values of N for which the

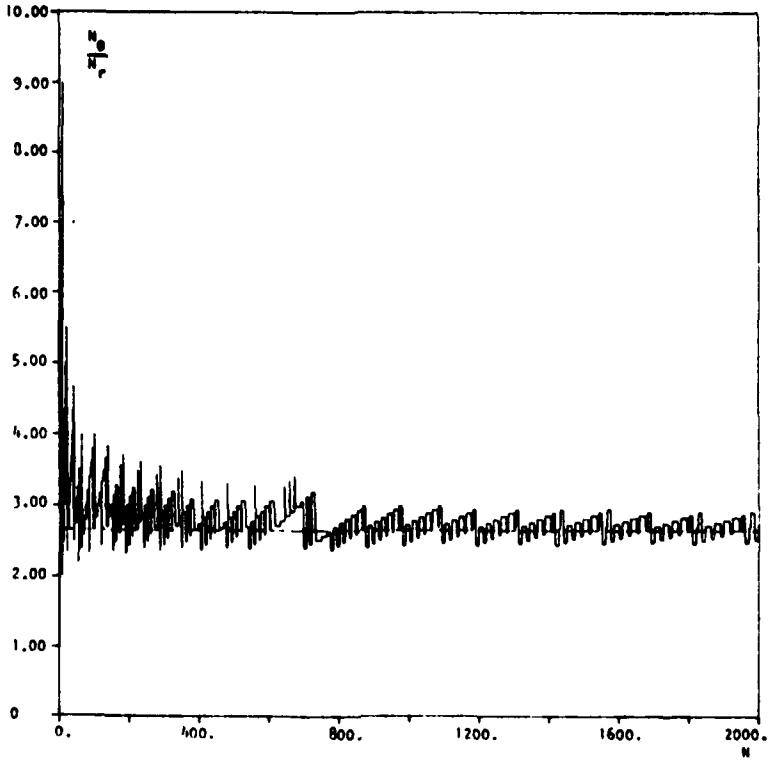


Fig. 1. Ratio of optimum number of phase quantizer levels to magnitude quantizer levels as a function of N.

Cartesian format error is smaller than the polar format error according to the exact error expressions. In the second column we find the values of N for which rectangular format is better than polar format according to the approximate error expressions. It can be seen that, in general, the approximate expressions are more pessimistic than the exact quantities.

V. Magnitude-Phase Information Comparison

An interesting problem that arises in using the polar representation is to find the best choice for the ratio of phase quantization levels to magnitude quantization levels. Pearlman [2] used distortion rate theory to obtain the ratio $N_\theta/N_r = 2.596$. We now give a somewhat different derivation.

We minimize (12), assuming N is large. We note that

$$\frac{\sin x}{x} \simeq 1 - \frac{x^2}{6} \tag{17}$$

and

$$\left(1 - \frac{x^2}{6}\right)^2 \simeq 1 - \frac{x^2}{3}. \tag{18}$$

Using these approximations, and (16) together with (12), we obtain

$$E_p \simeq \left(1 - (\pi/N_\theta)^2/3\right)\frac{0.9287 N_r}{(0.5965 + N_r)^3} + \tfrac{2}{3}(\pi/N_\theta)^2. \tag{19}$$

Assuming $(0.5965 + N_r)^3 \simeq N_r^3$, we substitute $N_r = N/N_\theta$ into (19), differentiate with respect to N_θ, and set the resulting expression equal to zero. Solving for N_θ, we find

$$N_\theta = 1.63 N^{1/2} \tag{20a}$$

or

$$\frac{N_\theta}{N_r} = 2.662, \tag{20b}$$

which agrees closely with the Pearlman bound. Fig. 1 shows a computer plot of the actual ratio plotted as a function of N. The dotted line is the value 2.662. Using this value in (19), it is a simple matter to show that, for large N, the polar format error is smaller than the rectangular format error.

VI. Applications to Spectral Phase Coding

From the preceeding sections, we know that if 3.33 bits or more per sample is to be used to quantize a white Gaussian sequence, it is better to pair the members of the sequence and quantize them in a polar format rather than simply quantizing the samples individually. We also know that the phase information is much more important than the magnitude information for minimizing the mean square quantization error.

Spectral phase coding (SPC) [1], [7] is one way in which we may make use of the above two properties. Consider some arbitrary data sequence $x_0, x_1, \cdots, x_{L-1}$, where in our examples we let $L = 4096$. The message sequence is divided into blocks of N samples; we consider the case

$N=32$. Each block of N terms is then divided in half, with the first $N/2$ terms forming the sequence $\{a_{1,n}\}_{n=0}^{N/2-1}$ and the second group of $N/2$ terms forming $\{a_{2,n}\}_{n=0}^{N/2-1}$. The complex-valued sequence $\{a_n\}_{n=0}^{N/2-1}$ is formed from

$$a_n = a_{1,n} + ia_{2,n}. \tag{21}$$

We then form the spectral sequence $\{A_p \exp(i\theta_p)\}$ from

$$A_p \exp(i\theta_p) = \sum_{n=0}^{(N/2)-1} a_n \exp(-i4\theta np/N), \qquad p = 0, \cdots, \frac{N}{2} - 1. \tag{22}$$

The SPC sequence $\{\psi_p\}_{p=0}^{N-1}$ is described by the following equations:

$$a_n = \frac{2}{N} \sum_{p=0}^{(N/2)-1} \frac{S}{2}\left[\exp(i\psi_p) + \exp(i\psi_{p+(N/2)})\right] \cdot \exp(i4\pi np/N), \qquad n = 0, \cdots, \frac{N}{2} - 1, \tag{23}$$

where

$$S = \max_p \{A_p\}, \tag{24a}$$

$$\psi_p = \theta_p + \phi_p, \tag{24b}$$

$$\psi_{p+(N/2)} = \theta_p - \phi_p, \tag{24c}$$

and $\phi_p = \cos^{-1}(A_p/S)$. Equation (24) describes the coding procedure and (23) the decoding procedure.

SPC is essentially a polar format representation of the discrete Fourier transform (DFT) of a random phase time series. In [8] the conditions under which the real and imaginary parts of the samples from the DFT tend to independent normal random variables are discussed. This is an asymptotic result, and it tells us that the magnitude of the DFT is Rayleigh and independent of the uniformly distributed phase. The uniform $(-\pi, \pi)$ distribution of the phase makes it a simple matter to quantize this quantity in an optimum fashion. Because of the relatively high phase information content, this case of quantization is important. Indeed, as is shown in Section IV, as long as the phase is optimally quantized, the quantizer characteristics for the magnitude component are much less important. In addition to the uniform phase property for the asymptotic case, we can show that in some special cases the phase has this property for small as well as large N.

Consider (22). We assume a_n can be represented as $r_n \exp(i\theta_n)$ where θ_n is uniform and independent of r_n for all θ_i, $i \neq n$. Under these assumptions, we have the following theorem.

Theorem: A_p is independent of θ_p, and θ_p is uniformly distributed for any arbitrary block size N.

Proof:

$$\mathrm{Re}\{A_p\} = \sum_{k=0}^{(N/2)-1} r_k \cos\phi_k \tag{25a}$$

$$\mathrm{Im}\{A_p\} = \sum_{k=0}^{(N/2)-1} r_k \sin\phi_k, \tag{25b}$$

where

$$\phi_k = \theta_k - \frac{4\pi}{N} kp. \tag{25c}$$

Consider the joint characteristic function of these two random variables:

$$\begin{aligned}
\Psi_N(\omega_1, \omega_2) &= E_{r_k} E_{\phi_k}\{\exp(j(\omega_1 \mathrm{Re}\{A_p\} + \omega_2 \mathrm{Im}\{A_p\}))\} \\
&= E_{r_k} E_{\phi_k}\left\{\exp\left(j\left[\omega_1 \sum_{k=0}^{(N/2)-1} r_k \cos\phi_k + \omega_2 \sum_{k=0}^{(N/2)-1} r_k \sin\phi_k\right]\right)\right\} \\
&= \frac{1}{(2\pi)^{(N/2)-1}} E_{r_k}\left\{\int_{-\pi}^{\pi} \exp\left(j\left(\sum_{k=0}^{(N/2)-1} r_k \cdot [\omega_1 \cos\phi_k + \omega_2 \sin\phi_k]\right)\right) d\phi_1 d\phi_2 \cdots d\phi_{(N/2)-1}\right\} \\
&= \frac{1}{(2\pi)^{(N/2)-1}} E_{r_k}\left\{\int_{-\pi}^{\pi} \exp\left(j \sum_{k=0}^{(N/2)-1} \sqrt{\omega_1^2 r_k^2 + \omega_2^2 r_k^2} \cdot \cos\left(\phi_k + \tan^{-1}\frac{\omega_2}{\omega_1}\right)\right) d\phi_1 \cdots d\phi_{(N/2)-1}\right\} \\
&= E_{r_k}\left\{\prod_{k=0}^{(N/2)-1} J_0\left((\omega_1^2 + \omega_2^2)^{1/2} r_k\right)\right\}
\end{aligned} \tag{26}$$

where E_{r_k} and E_{ϕ_k} are the expectation operators over the subscripted random variables. However, this is circularly symmetric. Using the properties of the two-dimensional Fourier transform, we know that the bivariate density must also be circularly symmetric. However, this can happen if and only if the magnitude is independent of the phase and the phase is uniformly distributed over a region of support 2π.

This theorem tells us that with the given assumptions, we can guarantee that the optimal transform phase quantizer is the uniform quantizer. In many cases, experimental data indicate that we are not far from the optimum result even when the conditions for the theorem do not hold for a particular sequence.

We now derive an expression for the quantizing error of the SPC representation. The ideal unquantized SPC representation is

$$A_p \exp(i\theta_p) = \frac{S}{2}\left[\exp(i\psi_p) + \exp(i\psi_{p+(N/2)})\right]. \tag{27}$$

To begin with, we assume that the phase terms $\{\psi_p\}$ are quantized to M equal step size quantization levels. From [2] we have

$$e^{i\hat{\psi}_p} = \sum_{m=-\infty}^{\infty} \mathrm{sinc}(m + 1/M) \exp(i(mM+1)\psi_p), \tag{28}$$

where $\hat{\psi}_p$ is the quantized version of $\hat{\psi}_p$. From experimental results it is found that quantization of the S parameter is negligible and will henceforth be ignored. The quantiza-

tion error E can now be expressed as

$$E = \hat{A}_p \exp(i\hat{\theta}_p) - A_p \exp(i\theta_p) \tag{29}$$

where $\hat{A}_p \exp(i\hat{\theta}_p)$ represents $A_p \exp(i\theta_p)$ using the quantized parameter $\hat{\psi}_p$. Using (28) in (29) we have

$$E = \frac{S}{2} \sum_{m \neq 0} \operatorname{sinc}(m + 1/M) \left[\exp(i(mM+1)\psi_p) + \exp(i(mM+1)\psi_{p+(N/2)})\right] + (1 - \operatorname{sinc}(1/M)) \frac{S}{2} (\exp(i\psi_p) + \exp(i\psi_{p+(N/2)})). \tag{30}$$

We square this quantity and take its expectation, using the following expressions derived in Appendix A:

$$\left| \frac{S}{2} \sum_{m \neq 0} \operatorname{sinc}(m + 1/M) \cdot \left[\exp(i(mM+1)\psi_p) + \exp(i(mM+1)\psi_{p+(N/2)})\right] \right|^2 = \frac{S^2}{2} \sum_{l \neq 0} \operatorname{sinc}^2(l + 1/M) \cdot \left[1 + \cos(2\phi_p(Ml+1))\right], \tag{31}$$

$$E\left\{ \left(\exp(i\psi_p) + \exp(i\psi_{p+(N/2)})\right) \sum_{m \neq 0} \operatorname{sinc}(m + 1/M) \cdot \left[\exp(-i(mM+1)\psi_p) + \exp(i(mM+1)\psi_{p+(N/2)})\right] \right\} = 0, \tag{32}$$

and

$$E\left\{ (1 - \operatorname{sinc}(1/M))^2 \left| \frac{S}{2} (\exp(i\psi_p) + \exp(i\psi_{p+(N/2)})) \right|^2 \right\} = (1 - \operatorname{sinc}(1/M))^2 E\{A_p^2\}, \tag{33}$$

where $E\{\cdot\}$ is the statistical expectation operator. Then

$$E_r = E\{(A_p^2)\}(1 - \operatorname{sinc}(1/M))^2 + \frac{S^2}{2} \sum_{l \neq 0} \operatorname{sinc}(l + 1/M) \cdot E\{1 + \cos[2(lM+1)\phi_p]\}. \tag{34}$$

From the Riemann-Lebesgue lemma [9] we know that, for large M, $E[\cos 2(lM+1)\phi_p] \ll 1$. Also,

$$\sum_{l \neq 0} \operatorname{sinc}^2(l + 1/M) = 1 - \operatorname{sinc}^2(1/M) = (1 - \operatorname{sinc}(1/M))(1 + \operatorname{sinc}(1/M)) \tag{35}$$

$$\approx 2(1 - \operatorname{sinc}(1/M)), \tag{36}$$

so that

$$E_r = E\{A_p^2\}(1 - \operatorname{sinc}(1/M))^2 + S^2(1 - \operatorname{sinc}(1/M)). \tag{37}$$

TABLE III
A COMPARISON OF NORMALIZED QUANTIZATION ERROR FOR AN SPC SEQUENCE AND AN OPTIMAL UNIT VARIANCE GAUSSIAN QUANTIZER FOR DIFFERENT PROBABILITY DENSITIES

Density	Error (Gaussian)	Error SPC
$N(0, 1)$	0.91 E-2	2.13 E-2
$N(0, 2)$	1.90 E-2	2.40 E-2
$N(0, 4)$	7.00 E-2	2.38 E-2
$U(-\frac{\sqrt{12}}{2}, \frac{\sqrt{12}}{2})$	0.73 E-2	2.08 E-2
$U(-\sqrt{12}, \sqrt{12})$	1.36 E-2	2.56 E-2
$U(-4, 4)$	3.50 E-2	2.43 E-2
$U(-5, 5)$	9.84 E-2	2.58 E-2
$X(2)$	3.50 E-2	12.60 E-2
$X(1)$	18.60 E-2	12.35 E-2
$X(.5)$	62.70 E-2	12.39 E-2

This error expression agrees extremely closely with computer simulations and with the error expression found in [1, eq. (22)] which is derived by a different method. The second term contributes the most to E_r.

We now present examples that make use of a sequence of 4096 zero mean, unit variance Gaussian random variables. We first form the SPC version of this sequence allowing four bits per SPC sample. The error expression in (37) predicts a mean square error of 2.2×10^{-2} per sample. The actual computed average error per sample for SPC block sizes of 32 is 2.3×10^{-2}. An optimal Max [4] quantizer would give an error of 0.91×10^{-2} per sample. By using SPC we create only a little over twice the minimum achievable error for this signal and this number of quantization levels. However, if the signal statistics change and the same quantizers are employed, what is the expected result?

Table III summarizes a number of computer simulations for Gaussian, double sided exponential, and uniform random variables coded using both the optimal unit variance Gaussian-Max quantizer and SPC. $N(0, A)$ is the zero mean, variance of A, Gaussian density; $U(-A/2, A/2)$ is the zero mean, variance of $A^2/12$, uniform density; and $X(A)$ is the zero mean, variance of $1/A^2$, double sided exponential density.

For this example, one can see that the large variance signals have lower quantizing error if coded with SPC. Because the Max-Gaussian quantizer has very small step sizes near the origin, we expect that it will produce small errors for those signals that have a large amount of probability in that region. The most striking characteristic of these results is the way the normalized SPC mean square error remains virtually constant for each particular distribution. SPC tracks variations in signal power very well.

VII. Conclusion

In this paper we have investigated in detail the optimum quantization of two-dimensional Gaussian random variables. Results are put forward to prove that, in general, polar format is superior to rectangular format. Applications of this to a coding scheme (SPC) are studied in order to explain why SPC seems to exhibit robustness with respect to variations in signal statistics and signal power.

Appendix A

We will now derive (31). Taking the square of the expression and moving the expectation operator through the sum leaves

$$\frac{S^2}{4}\sum_{m\neq 0}\sum_{l\neq 0}\operatorname{sinc}(m+1/M)\operatorname{sinc}(l+1/M)\cdot E\{\exp(iM(m-l)\psi_p)+\exp(iM(m\psi_p-l\psi_{p+(N/2)}))\cdot\exp(i(\psi_p-\psi_{p+(N/2)}))+\exp(-iM(l\psi_p-m\psi_{p+(N/2)}))\exp(-i(\psi_p-\psi_{p+(N/2)}))+\exp(iM(m-l)\psi_{p+(N/2)})\}. \quad \text{(A1)}$$

Assume that A_p is independent of θ_p. This means that the θ_p and ϕ_p used in the expressions for ψ_p and $\psi_{p+(N/2)}$ are also independent. Therefore, the expectation in (A1) is zero except for those terms where $l=m$. Consequently, this expression is equivalent to

$$\frac{S^2}{4}\sum_{l\neq 0}\operatorname{sinc}^2(l+1/M)\,E\{2+2\cos[(Ml+1)(\psi_p-\psi_{p+(N/2)})]\}. \quad \text{(A2)}$$

Because

$$\psi_p-\psi_{p+(N/2)}=2\phi_p, \quad \text{(A3)}$$

we have

$$\frac{S^2}{2}\sum_{l\neq 0}\operatorname{sinc}^2(l+1/M)[1+\cos 2\psi_p(Ml+1)]. \quad \text{(A4)}$$

Equation (32) is obtained by a similar argument. For (33), we recognize that

$$\frac{S}{2}(\exp(i\psi_p)+\exp(i\psi_{p+(N/2)}))=A_p\exp(i\theta_p). \quad \text{(A5)}$$

Therefore,

$$E\{(1-\operatorname{sinc}(1/M))^2|A_p\exp(i\theta_p)|^2\}=(1-\operatorname{sinc}(1/M))^2E\{A_p^2\}. \quad \text{(A6)}$$

References

[1] N. C. Gallagher, "Quatizing schemes for the discrete Fourier transform of a random time series," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 156–163, Mar. 1978.

[2] W. A. Pearlman, "Quantization error bounds for computer generated holograms," Stanford Univ. Inform. Syst. Lab., Stanford, CA, Tech. Rep #6503-1, Aug. 1974.

[3] J. A. Bucklew and N. C. Gallagher, "A note on optimum quantization," *IEEE Trans. Inform. Theory*, vol. IT-25, pp. 365–366, May, 1979.

[4] J. Max, "Quantization for minimum distortion," *IRE Trans. Inform. Theory*, vol. IT-6, pp. 7–12, Mar. 1960.

[5] P. E. Fleischer, "Sufficient conditions for achieving minimum distortion in a quantizer," *IEEE Int. Conv. Rec.*, Part I, pp. 104–111, 1964.

[6] R. C. Wood, "On optimum quantization," *IEEE Trans. Inform. Theory*, vol. IT-5, pp. 248–252, Mar. 1969.

[7] N. C. Gallagher, "Discrete spectral phase coding," *IEEE Trans. Inform. Theory*, vol. IT-22, pp. 622–624, Sept. 1976.

[8] N. C. Gallagher and B. Liu, "Statistical properties of the Fourier transform on random phase diffusers," *Optik*, vol. 42, pp. 65–86, Feb. 1975.

[9] H. L. Royden, *Real Analysis*. Toronto: MacMillan, 1968, pg. 90.

Two-Dimensional Quantization of Bivariate Circularly Symmetric Densities

JAMES A. BUCKLEW AND NEAL C. GALLAGHER, JR., MEMBER, IEEE

Abstract—**The problem of quantizing a two-dimensional random variable whose bivariate density has circular symmetry is considered in detail. Two quantization methods are considered, leading to polar and rectangular representations. A simple necessary and sufficient condition is derived to determine which of these two quantization schemes is best. If polar quantization is deemed best, the question arises as to the ratio of the number of phase quantizer levels to that of magnitude quantizer levels when the product of these numbers is fixed. A simple expression is derived for this ratio that depends only upon the magnitude distribution. Several examples of common circularly symmetric bivariate densities are worked out in detail using these expressions.**

I. Introduction

CONSIDER a two-dimensional random variable X whose bivariate density is circularly symmetric. We desire to represent this quantity by a finite set of values. One possible representation of X leads to a Cartesian coordinate system expression wherein we individually quantize the two rectangular components of the random variable. Another common representation leads to a polar coordinate representation where we quantize the magnitude and phase angle of X. These two representations are chosen mainly for their computational feasibility and ease of implementation. Other authors have considered the general problem of multidimensional quantization. Zador [1] derives an expression for the minimum error achievable by a multidimensional quantizer for an arbitrary density, but no insight into the required quantizer structure is attained. Chen [2] describes a recursive computer technique to solve for a "good" quantizer, but the optimality of the final solution is not assured. By constraining ourselves to circularly symmetric densities and also to either Cartesian or polar coordinate quantization schemes, it becomes possible to reduce the optimal two-dimensional quantization problem to one dimension. Max [3] develops necessary conditions for the optimality of a one-dimensional quantizer. Panter and Dite [4] give a formula for the asymptotic error to be expected for optimal mean square error quantizers (of sufficiently smooth input densities).

In Section II we obtain a simple criterion by which to determine whether polar format or rectangular format gives a smaller mean square quantization error. It is shown that for some very important cases, notably for the Gaussian bivariate density, the polar format is asymptotically superior.

If polar format is to be used and the product $N = N_\theta N_r$ is fixed, where N_θ and N_r are the number of phase and magnitude quantization levels, respectively, the question arises as to the optimum ratio N_θ / N_r. We derive a simple expression for this ratio that depends upon only the magnitude density.

In Section III we provide several examples of common circularly symmetric densities (e.g., marginal densities are Pearson II, Pearson VII, sinusoidal, and Gaussian), and we address the question of whether the rectangular or the polar format scheme gives a smaller quantization error.

Manuscript received September 12, 1978; revised April 2, 1979. This work was supported by the Air Force Office of Scientific Research under Grant AFOSR 78-3605.

The authors are with the Department of Electrical Engineering, Purdue University, West Lafayette, IN 47907.

II. Development

Consider the mean square quantization error E_p of a polar format representation of the two-dimensional random variable $x = r\exp[i\theta]$:

$$E_p = \sum_{j=1}^{N_\theta} \sum_{i=1}^{N_r} \int_{c_{j-1}}^{c_j} \int_{a_{i-1}}^{a_i} |re^{i\theta} - b_i e^{id_j}|^2 \frac{f_r(r)\,dr\,d\theta}{2\pi}. \quad (1)$$

Implicit use has been made of the fact that in circularly symmetric bivariate densities the magnitude random variable with probability density $f_r(\cdot)$ is independent of the uniformly distributed $[-\pi, \pi]$ phase random variable. The b_i and d_j are the output levels of the magnitude and phase quantizers corresponding to input levels lying in the intervals $(a_{i-1}, a_i]$ and $(c_{j-1}, c_j]$, respectively. Integrating over the θ variable, (1) becomes

$$\sum_{j=1}^{N_\theta} \sum_{i=1}^{N_r} \int_{a_{i-1}}^{a_i} \left[(r^2 + b_i^2)(c_j - c_{j-1}) - 2rb_i \cdot \left[\sin(c_j - d_j) - \sin(c_{j-1} - d_j) \right] \right] \frac{f_r(r)}{2\pi}\,dr. \quad (2)$$

It is shown in [5] that the optimal phase quantizer is the uniform quantizer. This means that $c_j - c_{j-1} = 2\pi / N_\theta$ and $c_j - d_j = -(c_{j-1} - d_j) = \pi / N_\theta$, for $j = 1, \cdots, N_\theta$. This allows us to simplify (2):

$$E_p = \sum_{i=1}^{N_r} \int_{a_{i-1}}^{a_i} \left[r^2 + b_i^2 - 2rb_i \frac{\sin \frac{\pi}{N_\theta}}{\frac{\pi}{N_\theta}} \right] f(r)\,dr. \quad (3)$$

Differentiating with respect to b_i, we find the optimum b_i

0018-9448/79/1100-0667$00.75

is

$$b_i = \frac{\sin\frac{\pi}{N_\theta}}{\frac{\pi}{N_\theta}} \frac{\int_{a_{i-1}}^{a_i} rf(r)\,dr}{\int_{a_{i-1}}^{a_i} f(r)\,dr}. \tag{4}$$

The equation given by Max for the output levels b_i' of an optimal one-dimensional magnitude quantizer is found in [3] to be

$$b_i' = \frac{\int_{a_{i-1}'}^{a_i'} rf(r)\,dr}{\int_{a_{i-1}'}^{a_i'} f(r)\,dr}, \tag{5}$$

where the optimal input interval endpoints a_i' (for the one-dimensional case) satisfy

$$a_i' = \frac{b_i' + b_{i+1}'}{2}. \tag{6}$$

If we minimize (3) with respect to the a_i, we arrive at the necessary condition (for the two-dimensional case)

$$a_i = \frac{b_i + b_{i+1}}{2\left(\frac{\sin\frac{\pi}{N_\theta}}{\frac{\pi}{N_\theta}}\right)} = \frac{b_i' + b_{i+1}'}{2} = a_i'. \tag{7}$$

This equation indicates that the quantizer interval endpoints for the optimum magnitude quantizer in the two-dimensional case is the same as the quantizer interval endpoints for the optimum one-dimensional quantizer. From (4) and (5) and the preceding discussion, we have the following relationship between the output levels b_i' and b_i:

$$b_i' = \frac{\frac{\pi}{N_\theta}}{\sin\frac{\pi}{N_\theta}} b_i. \tag{8}$$

Consequently, (3) becomes

$$E_\rho = E\{r^2\} - \left(\frac{\sin\frac{\pi}{N_\theta}}{\frac{\pi}{N_\theta}}\right)^2 \sum_{i=1}^{N_r} (b_i')^2 \int_{a_{i-1}}^{a_i} f(r)\,dr, \tag{9}$$

where $E\{\cdot\}$ is the statistical expectation operator. In [6] it is shown that the mean square quantization error for a minimum mean square error quantizer is simply the input mean square value minus the output mean square value. If we denote by E_x^N the mean square quantization error produced by an optimal N level quantizer for the random variable X, we may rewrite (9) as

$$E_\rho = \left(\frac{\sin\frac{\pi}{N_\theta}}{\frac{\pi}{N_\theta}}\right)^2 E\{r^2\} - \left(\frac{\sin\frac{\pi}{N_\theta}}{\frac{\pi}{N_\theta}}\right)^2 \sum_{i=1}^{N_r} (b_i')^2 \int_{a_{i-1}}^{a_i} f_r(r)\,dr + \left[1 - \left(\frac{\sin\frac{\pi}{N_\theta}}{\pi}\right)^2\right] E\{r^2\} = \left(\frac{\sin\frac{\pi}{N_\theta}}{\frac{\pi}{N_\theta}}\right)^2 E_r^{N_r} + \left[1 - \left(\frac{\sin\frac{\pi}{N_\theta}}{\frac{\pi}{N_\theta}}\right)^2\right] E\{r^2\}. \tag{10}$$

Our problem is now one of characterizing the quantity E_x^N. Panter and Dite [4] give a formula for the expected error of a minimum mean square error quantizer with a large number of output levels and a smooth input density. This formula is

$$E_x^N = \frac{K_x}{N^2} \tag{11}$$

where

$$K_x = \frac{\left[\int_{-\infty}^{\infty} f(x)^{1/3}\,dx\right]^3}{12}.$$

Roe [7] also derives some asymptotic formulas which were later used by Wood [8] to rederive (11). Roe's formulas depend on the truncation of a Taylor series expansion of the input density. Wood, in his formula, explicitly states that the input density and the first few derivates (up to order five in some cases) must exist and be continuous. Panter and Dite require that, as the input intervals become very small, the density function may be approximated as a constant over each interval. In [1] it is shown that a sufficient condition for (11) to hold is that $f(x)$ be Riemann integrable, a much less severe restriction then continuity or differentiability.

We make use of the approximation

$$\left(\frac{\sin x}{x}\right)^2 \simeq 1 - \frac{x^2}{3}, \tag{12}$$

and of (11) in order to reduce (10) to

$$E_\rho \simeq \left(1 - \frac{\pi^2}{3N_\theta^2}\right)\frac{K_r}{N_r^2} + \frac{2}{3}\frac{\pi^2}{N_\theta^2}, \tag{13}$$

where we assume $E\{r^2\} = 2$ (this implies unit variance rectangular marginal densities). If we let N be the total number of output levels allowed to represent the two dimensional random variable X, we have the relation.

$$N = N_r N_\theta. \tag{14}$$

Since $K_r > 0$, it is simple to show that $N_r = O(N^{1/2})$ and $N_\theta = O(N^{1/2})$ by differentiating (13) and solving for the optimal quantities. Making use of this fact and (14), we may, assuming sufficiently large N, write (13) as

$$E_\rho = \frac{K_r N_\theta^2}{N^2} + \frac{2}{3}\frac{\pi^2}{N_\theta^2}. \tag{15}$$

This is then optimized with respect to N_θ and yields the optimal N_θ^2 as

$$N_\theta^2 = \left(\frac{2\pi^2}{3K_r}\right)^{1/2} N. \tag{16}$$

This leads to the following expression for the minimal attainable asymptotic polar format error:

$$(E_p)_{\text{opt}} = \sqrt{\frac{2K_r}{3}}\,\frac{2\pi}{N}. \tag{17}$$

Now consider the problem of optimally quantizing the random variable X in a rectangular format. The mean square quantization error E_x of this representation is given by

$$E_x = \sum_{i=1}^{N_x}\sum_{j=1}^{N_y}\int_{g_{j-1}}^{g_j}\int_{e_{i-1}}^{e_i}\left[(x-f_i)^2+(y-h_j)^2\right] \cdot f_{x,y}(x,y)\,dx\,dy, \tag{18}$$

where N_x and N_y are the number of levels in each of the respective orthogonal random variables. The other notation should be clear. Equation (18) may be written as

$$E_x = \sum_{i=1}^{N_x}\int_{e_{i-1}}^{e_i}(x-f_i)^2 f_x(x)\,dx + \sum_{j=1}^{N_y}\int_{g_{j-1}}^{g_j}(y-h_j)^2 f_y(y)\,dy. \tag{19}$$

where we make use of the fact that the first term in the bracket in (18) depends only upon x and the second term depends only upon y. By symmetry arguments (since $f_x(x)=f_y(x)$), we may argue that $N_x = N_y = N^{1/2}$. The quantizer that minimizes the above equation is simply the minimum mean square error quantizer for each of the two components. Therefore, again using (11), we have for large N

$$E_x = \frac{2K_x}{N},$$

where

$$K_x = \frac{\left[\int_{-\infty}^{\infty} f(x)^{1/3}\,dx\right]^3}{12}. \tag{20}$$

Comparing (20) and (17), we say that polar format is asymptotically better than rectangular format if and only if

$$\frac{2K_x}{N} > \sqrt{\frac{2K_r}{3}}\,\frac{2\pi}{N},$$

or

$$K_x > \sqrt{\frac{2K_r}{3}}\,\pi. \tag{21}$$

In other words, if the inequality is satisfied and the original input probability density is Riemann integrable, then we are guaranteed that there exists an N_0 such that for every $N > N_0$, polar format quantization will perform better than rectangular format quantization.

If polar quantization is deemed best for a particular density, then what is the ratio N_θ/N_r that provides the smallest total error? This question is answered by the use of (16); we find

$$\left(\frac{N_\theta^2}{N}\right)_{\text{opt}} = \left(\frac{N_\theta}{N_r}\right)_{\text{opt}} = \sqrt{\frac{2}{3K_r}}\,\pi. \tag{22}$$

III. Examples

For our first example we calculate the relevant parameters for a random variable whose marginal density is of Pearson type VII. This distribution is a generalization of Student's t-distribution. The bivariate density is

$$f(x,y) = \frac{\upsilon}{\pi}\,\frac{2^\upsilon(\upsilon-1)^\upsilon}{(2(\upsilon-1)+x^2+y^2)^{\upsilon+1}}, \qquad -\infty < x,y < \infty \tag{23}$$

(with $\upsilon > 1$ to assure finite variance) and the marginal density appears as

$$f(x) = \frac{2^\upsilon(\upsilon-1)^\upsilon\Gamma(\upsilon+1/2)}{\sqrt{\pi}\,\Gamma(\upsilon)\left((2(\upsilon-1)+x^2)^{\upsilon+1/2}\right)}, \qquad -\infty < x < \infty \tag{24}$$

where $\Gamma(\cdot)$ is the gamma function and where we have normalized the distribution so that $f(x)$ has unit variance. The magnitude density is derived by substituting in r for $\sqrt{x^2+y^2}$ in $f(x,y)$ and multiplying the result by $2\pi r$, as shown by a simple change of variable. Equation (24) yields, after some tedious algebra,

$$K_x = \frac{\left[B\left(\frac{1}{2};\frac{\upsilon-1}{3}\right)\right]^3}{12B\left(\frac{3}{2};\upsilon-1\right)}, \tag{25}$$

where $B(\cdot;\cdot)$ is the beta function. We perform similar operations with the magnitude density to yield

$$K_r = \frac{\upsilon(\upsilon-1)}{24}\left[B\left(\frac{2}{3};\frac{\upsilon-1}{3}\right)\right]^3. \tag{26}$$

In Fig. 1 K_x (solid line) and $(2K_r\pi/3)^{1/2}$ (dotted line) are plotted as a function of υ for values from 1.1 to 21.1. As shown by this graph, the polar format is always asymptotically best for this class of distributions. An interesting point about this set of distributions is that, in the limit as $\upsilon\to\infty$, (23) converges to a unit variance Gaussian density. Therefore, taking this limit in (25) and making use of Stirling's approximation, we have

$$K_x \to \frac{\sqrt{3}\pi}{2} \simeq 2.721. \tag{27}$$

Wood [8] estimates this number as 2.73 which is close to our derived value. From (26) we have similarly

$$K_r \to \frac{3}{8}\left(\Gamma\left(\frac{2}{3}\right)\right)^3 \simeq 0.931, \tag{28}$$

which is the parameter for the Rayleigh distribution obtained in the limit. Using these two values in (21), we

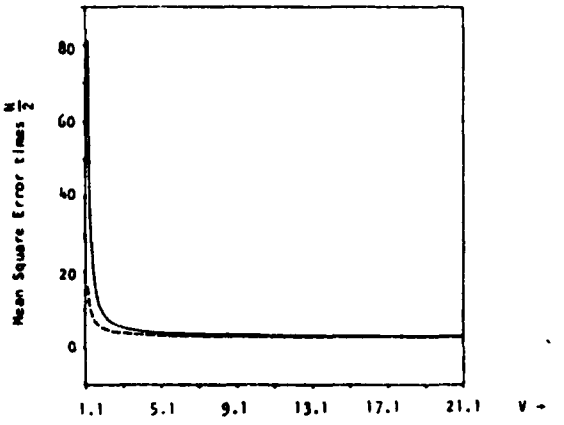


Fig. 1. Solid line is a plot of K_z as a function of V, and dotted line is a plot of $(2K_z\pi/3)^{1/2}$ as function of V for Pearson VII density.

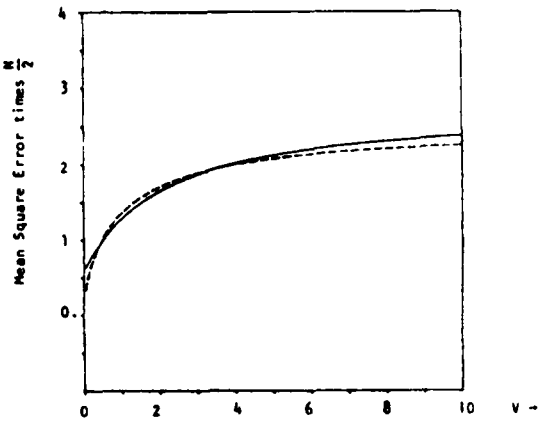


Fig. 2. Solid line is a plot of K_z as a function of V, and dotted line is a plot of $(2K_z\pi/3)^{1/2}$ as function of V for the Pearson II density.

conclude that asymptotically polar formatting is better than rectangular formatting for Gaussian bivariate densities. As a matter of interest, when we substitute the value of K_z found in (28) into (22), we find the optimal ratio N_θ/N_r to be 2.659. Pearlman [9] using distortion rate theory states that this ratio should be >2.596, which is in agreement with our result.

For the next example, consider distributions of the Pearson II class. The bivariate density is

$$f(x,y)=\frac{\upsilon(2(\upsilon+1)-(x^2+y^2))^{\upsilon-1}}{\pi 2^{\upsilon}(\upsilon+1)^{\upsilon}}\cdot U(2(\upsilon+1)-(x^2+y^2)), \quad (29)$$

where $\upsilon>0$, and $U(\cdot)$ is the unit step function. The marginal density is

$$f(x)=\frac{\Gamma(\upsilon+1)(2(\upsilon+1)-x^2)^{\upsilon-(1/2)}U(2(\upsilon+1)-x^2)}{2^{\upsilon}(\upsilon+1)^{\upsilon}\sqrt{\pi}\,\Gamma\left(\upsilon+\frac{1}{2}\right)}. \quad (30)$$

For $\upsilon=\frac{1}{2}$ we find that $f(x)$ has a uniform distribution. For $\upsilon=1$, we have that the bivariate density is uniform over a circular region in the plane. Using (30), we find

$$K_z=\frac{\left[B\left(\frac{1}{2};\frac{2\upsilon+5}{6}\right)\right]^3}{12B\left(\frac{3}{2};\upsilon+\frac{1}{2}\right)}. \quad (31)$$

From the magnitude density we derive that

$$K_r=\frac{\upsilon(\upsilon+1)}{24}\left[B\left(\frac{2}{3};\frac{\upsilon+2}{3}\right)\right]^3. \quad (32)$$

In Fig. 2 can be seen a plot of K_z (solid line) and $(2K_r\pi/3)^{1/2}$ (dotted line) as a function of υ for values from zero to ten. It should be noted that (30) also converges to a Gaussian density as $\upsilon\to\infty$. It is a simple matter to check that the expressions in (31) and (32) indeed approach the correct limits. From the plot it can be seen that for values of υ in the interval (0.0, 0.4) polar format is better. In the interval (0.4, 3.635) it is seen that rectangular is better, and from 3.635 to infinity polar again is better. It appears then that for the circularly symmetric bivariate density whose marginal density is uniform, we have the interesting result that rectangular format is asymptotically better than polar format.

In our analysis and in the examples considered so far we have constrained the class of quantizers considered to two different types, the rectangular format and the polar format. In general, neither of these schemes will be optimal for an arbitrary two-dimensional random variable with a circularly symmetric probability density. Zador [1] gives an expression for the asymptotic mean square error E_z of the optimal two-dimensional mean square error quantizer. This equation is

$$E_z=C_z/N, \quad (33)$$

where

$$C_z=\frac{5}{18\sqrt{3}}\left[\int_{-\infty}^{\infty}\int f_{x,y}(x,y)^{1/2}\,dx\,dy\right]^2. \quad (34)$$

For the Pearson VII density $C_z=4.0307\,\upsilon/(\upsilon-1)$, for the Pearson II density $C_z=4.0307\,\upsilon/(\upsilon+1)$. Since, in the limit as υ becomes large, both of these classes of densities converge to the Gaussian, the smallest error attainable for a two-dimensional normal random variable is approximately $4.0307/N$. The best that we can do with a polar format representation is $4.95/N$ and the best that we can do with a Cartesian format representation is $5.442/N$. There is certainly room for improvement here. However, the important thing to note is that the structure of the polar format quantizer is known while that of the theoretical optimum quantizer is not.

In Section II it was stated that a sufficient condition for (11) to be valid is that the magnitude density function be Riemann integrable. For most density functions of interest in modeling physical systems, this criterion is met. One group of densities that does not meet this condition is the set of atomic densities, i.e., densities for which probability mass is contained at a single point. In a circularly symmetric bivariate density, the phase must be uniformly distributed $[-\pi,\pi]$. The only quantity that can be discrete is the magnitude distribution, i.e., we may have "rings" of probability mass distributed in the plane. Suppose we have a single "ring" of probability mass, where the radius of the ring is one, i.e.,

$$F(r) = U(r-1), \tag{35}$$

where $F(\cdot)$ is the magnitude distribution function and $U(\cdot)$ is the unit step function. The rectangular component marginal density is the sinusoidal density

$$f(x) = \frac{U(1-x^2)}{\pi\sqrt{1-x^2}}. \tag{36}$$

This density function is Riemann integrable, hence (11) and (20) are valid. This implies the rectangular format error is $O(N^{-1})$. Now consider the polar format case. For $N_r > 1$, $E_r^{N_r} = 0$. This implies the polar format error for large N is $O(N^{-2})$. Clearly polar format is asymptotically better for this density. By extending this argument, we may say that if $P(r=0) \neq 1$, then for any bivariate circularly symmetric density with an atomic magnitude density with a finite number of atoms, polar format will give a smaller asymptotic mean square quantization error than rectangular format.

IV. Summary

In this paper we have derived a simple criterion to determine whether rectangular format or polar format gives smaller mean square error for circularly symmetric densities. The optimal ratio of phase quantizer levels to magnitude quantizer levels is also derived. Several examples including the Gaussian case have been studied in detail.

It is interesting to note that polar format is not always better than rectangular format even for the case of densities with circular symmetry.

References

[1] P. Zador, "Development and evaluation of procedures for quantizing multivariate distributions," Ph.D. dissertation, Stanford University, Stanford, CA, 1964.

[2] D. Chen, "On two or more dimensional optimum quantizers, *Record of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing Conference*, IEEE press, pp. 640–643, 1977.

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

[4] P. F. Panter and W. Dite, "Quantization distortion in pulse count modulation with nonuniform spacing of levels," *Proc. IRE*, vol. 39, pp. 44–48, Jan. 1951.

[5] N. C. Gallagher, "Quantizing schemes for the discrete Fourier transform of a random time series," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 156–163, Mar. 1978.

[6] J. A. Bucklew and N. C. Gallagher, "A Note on Optimum Quantization," *IEEE Trans. Inform. Theory*, vol. IT-25, pp. 365–366, May 1979.

[7] G. M. Roe, "Quantizing for minimum distortion," *IEEE Trans. Inform. Theory*, vol. IT-10, pp. 384–385, Oct. 1964.

[8] R. C. Wood, "On optimum quantization," *IEEE Trans. Inform. Theory*, vol. IT-5, pp. 248–252, Mar. 1969.

[9] W. A. Pearlman, "Quantization Error Bounds for Computer Generated Holograms," *Stanford Univ. Inform. Syst. Lab.*, Stanford, CA, Tech. Rep. #65031-1, Aug. 1974.

SOME RESULTS IN MULTIDIMENSIONAL QUANTIZATION THEORY*

James A. Bucklew
Electrical and Computer Engineering Department
University of Wisconsin, Madison, WI 53705

and

N. C. Gallagher, Jr.
Department of Electrical Engineering
Purdue University, West Lafayette, IN 47906

Abstract

This paper contains several results in multidimensional quantization theory. The first section gives a simplified derivation of a well known upper bound on the distortion introduced by a k-dimensional optimum quantizer. It is then shown that an optimum multidimensional quantizer preserves the mean vector of the input and that the mean square quantization error is given by the sum of the component variances of the input minus the sum of the variances of the output. Lastly, a general equation which can be used to evaluate the performance of multidimensional companders is derived. It is shown that the optimal compander must be conformal everywhere. An example is given to show that asymptotically optimal performance could be obtained through nonconformal companding schemes.

I. Introduction

Block or vector quantization deals with the representation of multidimensional elements with a finite discrete set of values. The values to be quantized may naturally fall into a k-dimensional representation; typical examples are complex numbers, positional coordinates, or state vectors. In other cases, k-dimensional vectors are formed from blocks of k samples taken from one dimensional signals. In 1964 Paul Zador published his Ph.D. dissertation which contains a number of very interesting results on the properties of optimal block quantizers for the r'th moment euclidean norm distortion measure [1]. Among Zador's contributions are the derivation of both upper and lower bounds on the distortion introduced by the optimal quantizer. These bounds are derived without actually finding the optimal quantizer. Unfortunately, at some points Zador's development is difficult to follow and alternate derivations and extensions by Gersho [2], and Yamada, et al. [3] have recently appeared. In Section II we present an alternate derivation of Zador's random quantization upper bound not treated in either [2] or [3].

In [4] Bucklew and Gallagher show that for one dimensional mean squared error distortion the optimum quantizer has the property that the mean value of the quantizer output equals the mean value of the input and also that the mean square quantization error equals the variance of the input minus the variance of the output. In [5] Bucklew and Gallagher prove that the same results hold for constant step size minimum mean squared error quantizers. In Section III we extend these properties to k-dimensional optimal block quantizers.

W. R. Bennett [6] was the first to model a nonuniform quantizer as a zero memory nonlinearity followed by a uniform quantizer in turn followed by the inverse of the first zero memory nonlinearity. This sequence of operations is generally referred to as companding. The word arises because the data is first "compressed", then quantized, then "expanded". As a consequence the first nonlinearity is generally referred to as the "compressor" and its inverse the "expander".

The fourth Section of this paper is an investigation of companding in several dimensions. In several dimensions the compressor characteristic is a mapping function

$$f : \mathbb{R}^k \to \bigtimes_{i=1}^{k} (0,1)$$

where X denotes the Cartesian cross product. $\bigtimes_{i=1}^{k} (0,1)$ is of course the k-dimensional hypercube. In the companding approach to optimal quantization, we have quantizer output levels distributed in the hypercube. We choose from these output levels the nearest neighbor (usually) to $f(\underline{x})$, where $\underline{x}$ is the input data vector. Our quantized output is then f^{-1} of this particular output level.

Our theory will hold for analog signal processing in several dimensions also. It happens that it doesn't matter whether the noise is quantization noise or any other kind of additive noise as long as the noise components in each channel are uncorrellated with one another. For example, let us denote the error vector caused by quantization in the hypercube as $(r_1, r_2, \ldots, r_k)^T$. Then the condition that is needed is $E(r_i r_j) = \sigma_r^2 \delta_{ij}$ where δ_{ij} is the Kronecker delta function. In a practical sense, this is not a very restrictive assumption. It may be shown, at least asymptotically (as the number of output levels in the hypercube approaches infinity), that the error vector in an optimal or random quantizer converges to a hyperspherically symmetric probability density which satisfies our above condition.

II. Random Quantization Upper Bound

In [2] Gersho provides a very readable derivation of Zador's expression for quantizer distortion. To improve continuity and readability we employ Gersho's notation; the quantizer input is a k dimensional random vector in R_k which is quantized to one of N levels $\underline{y}_1, \underline{y}_2, \ldots, \underline{y}_N$ in R_k. The space R_k is partitioned into N disjoint and exhaustive regions $S_1, S_2, \ldots, S_N$. The quantizer is defined by the function $Q(\underline{x})$, where for k-dimensional input value $\underline{x}$,

$$Q(\underline{x}) = \underline{y}_i, \text{ if } \underline{x} \in S_i. \tag{1}$$

Note that this definition does not require $\underline{y}_i \in S_i$, although in practice $\underline{y}_i$ is usually contained in S_i. The performance of the quantizer is measured by the distortion

$$D = \frac{1}{k} E(\|\underline{x} - Q(\underline{x})\|^r) \tag{2}$$

where $\|\cdot\|$ denotes the usual ℓ_2 norm, the operator $E(\cdot)$ denotes statistical expectation and the input $\underline{x}$ is a k dimensional random input vector. The case

PRINCETON 1980

where r=2 is the usual mean squared distortion. The expression derived by Zador and Gersho for the minimum distortion D_0 obtained by use of the best quantizer is

$$D_0 = N^{-\frac{r}{k}} C(k,r) \| p(\underline{x}) \|_{k/(k+r)}, \quad (3)$$

where

$$\| p(x) \|_\alpha = [\int [p(x)]^\alpha dx]^{1/\alpha},$$

and where the constant C(k,r), called the coefficient of quantization, is independent of the density $p(\underline{x})$ and is in general unknown. This expression is an asymptotic result valid only for large N. Two special cases for which the value of C(k,r) is known exactly are [2]

$$C(1,r) = \frac{1}{r+1} 2^{-r}, \quad (4)$$

and

$$C(2,2) = \frac{5}{36\sqrt{3}} . \quad (5)$$

Consider the density p(x) that has a constant value of one over the unit volume hypercube; then $\| p(x) \|_{k/(k+r)} = 1$. Consequently, Eq. (3) becomes

$$D_0 = N^{-\frac{r}{k}} C(k,r). \quad (6)$$

So, we see that by finding a bound on D_0 we also bound C(k,r). To find this bound we choose the quantizer output levels to have a random distribution uniformly distributed over the hypercube. For a particular input value $\underline{x}$, we find the closest output level and quantize to that value. Because this quantizer is not the optimum quantizer the associated distortion will bound from above the distortion for the optimum quantizer.

To begin, place at random N independent uniformly distributed k dimensional samples in the hypercube. These will be our output levels. We take the quantizer input $\underline{X}$ to have a uniform distribution over the hypercube. We also assume that N is sufficiently large so that there is a very small probability that the quantizer input is closer to an edge of the hypercube than to one of the output values. Suppose that an input value $\underline{x}$ has arrived and is sitting in the hypercube waiting to be quantized. The probability that one particular output value is within a distance ρ of this input sample is given approximately by the volume of a sphere of radius ρ about that sample point, or

$$\text{Prob (one particular output level is within } \rho \text{ of the input sample)} = V_k \rho^k, \quad (7)$$

where if V_k is volume of the unit radius sphere, then $V_k \rho^k$ is the volume of the sphere with radius ρ. We are interested in the closest output level to the input sample. We really want to know the probability that the closest output level is within a distance ρ of the input sample. To compute this probability, we combine classical order statistics with the result found in Eq. (7). By employing this approach, we compute the probability density $f(\rho)$ for the distance between the input sample and the nearest output level to be

$$f(\rho) = N[1 - V_k \rho^k]^{N-1} V_k k \rho^{k-1}. \quad (8)$$

By construction $\rho = \| \underline{x} - \underline{y}_i \|$, where $\underline{x}$ is the input value and $\underline{y}_i$ is the output value. Consequently,

$$E\{ \| \underline{x} - Q(\underline{x}) \|^r \} = E\{\rho^r\}, \quad (9)$$

so, by Eq. (2)

$$D = \frac{1}{k} E\{\rho^r\}$$
$$= \frac{1}{k} \int_0^\infty \rho^{r+k-1} N[1-V_k \rho^k]^{N-1} k V_k d\rho. \quad (10)$$

Make the change of variables $s = V_k \rho^k \leq 1$.

$$D \leq \frac{N}{k V_k^{r/k}} \int_0^1 s^{r/k} [1-s]^{N-1} ds,$$

or

$$D \leq \frac{N}{k V_k^{r/k}} \frac{\Gamma(\frac{k+r}{k}) \Gamma(N)}{\Gamma(N + \frac{k+r}{k})}, \quad (11)$$

where $\Gamma(\cdot)$ is the gamma function. For large N the following approximation is valid:

$$\frac{\Gamma(N)}{\Gamma(N + \frac{k+r}{k})} \approx N^{-\frac{k+r}{k}} . \quad (12)$$

Therefore,

$$D \leq \frac{N^{-r/k} \Gamma(1 + \frac{r}{k})}{k V_k^{r/k}} \quad (12)$$

Because $D \geq D_0$, we use Eq. (6) to write

$$C(k,r) \leq \frac{\Gamma(1 + \frac{r}{k})}{k V_k^{r/k}}, \quad (13)$$

which is Zador's random quantization upper bound.

III. Moment Properties of Optimum Quantizers

In [4] and [5] it is shown that for minimum mean squared error one dimensional quantizers that the mean of the input equals the mean of output and the distortion equals the variance of the input minus the output variance. It is shown that these properties apply with and without the equal-step-size constraint. In this section we generalize these results to the k dimensional case.

We are interested in the properties of quantizers designed to minimize the distortion defined by Eq. (2) for r = 2:

$$D = \frac{1}{k} E\{ \| \underline{X} - Q(\underline{X}) \|^2 \} . \quad (14)$$

Many constraints we impose on the quantizer can be imposed by the functional form of $Q(\underline{x})$; for example, the k dimensional version of the equal-step-size condition might require the regions $S_1, S_2, \ldots, S_N$ to have equal area and be jointly congruent. A variational approach is used in the derivation. Assume the optimum quantizer is $Q_0(\underline{x})$; so, an arbitrary quantizer characteristic can be represented as

$$Q(\underline{x}) = Q_0(\underline{x}) + \epsilon \, \delta Q(\underline{x}), \quad (15)$$

where ϵ is an arbitrary real variable and $\delta Q(\underline{x})$ is an arbitrary variation chosen so that $Q(\underline{x})$ satisfies all constraints imposed on the quantizer. We know that the optimum choice for ϵ is $\epsilon = 0$; it is this value that minimizes the distortion D. Thus, $\partial D / \partial \epsilon = 0$ at $\epsilon = 0$; so,

$$\left.\frac{\partial D}{\partial \epsilon}\right|_{\epsilon=0} = \frac{1}{k}\frac{\partial}{\partial \epsilon} E\{\|\underline{x} - Q(\underline{x})\|^2\}\Big|_{\epsilon=0} = 0 \quad (16)$$

or

$$= \frac{2}{k} E\{[\underline{x} - Q_0(\underline{x})]\delta Q^T(\underline{x})\} = 0, \quad (17)$$

where we note that $Q(\underline{x})$ is a vector valued function so that $\delta Q(\underline{x})$ represents an arbitrary variation and consider the case where each component of this variation vector equals one; consequently, Eq. (17) becomes

$$E\{\underline{x} - Q_0(\underline{x})\} = 0,$$

or

$$E\{\underline{x}\} = E\{Q_0(\underline{x})\} . \quad (18)$$

Now consider the case where $\delta Q(\underline{x}) = Q_0(\underline{x})$; then,

$$E\{[\underline{x} - Q_0(\underline{x})]Q_0^T(\underline{x})\} = 0,$$

or

$$E\{\underline{x}Q_0^T(\underline{x})\} = E\{\|Q_0(\underline{x})\|^2\} . \quad (19)$$

When the optimum quantizer is in use, the distortion D_0 is

$$\begin{aligned} D_0 &= \frac{1}{k} E\{\|\underline{x} - Q_0(\underline{x})\|^2\} \\ &= \frac{1}{k} E\{[\underline{x} - Q_0(\underline{x})][\underline{x} - Q_0(\underline{x})]^T\} \\ &= \frac{1}{k}\left[E\{\|\underline{x}\|^2\} - E\{\underline{x}Q_0^T(\underline{x})\} - E\{Q_0(\underline{x})\underline{x}^T\} + \right. \\ &\qquad \left. + E\{\|Q_0(\underline{x})\|^2\}\right] . \end{aligned} \quad (20)$$

We combine the results of Eq. (19) with Eq. (20) to produce

$$D_0 = \frac{1}{k}\left[E\{\|\underline{x}\|^2\} - E\{\|Q_0(\underline{x})\|^2\}\right]. \quad (21)$$

The results of Eq. (21) combine with those in Eq. (18) to provide the multidimensional extension of the one dimensional case found in [4] and [5]. We note here that this derivation is quite general. It applied to the unconstrained optimal quantizer as well as the equal volume congruent area (equal step size) quantizer because this constraint can be included directly into the functional form of $Q(\underline{x})$.

IV. Compander Error Derivation

Our data will be assumed to be k-dimensional samples from a probability density function $p(\underline{x})$, $\underline{x} \in R^k$. Denote D_p as the support of $p(\underline{x})$. Let $f : D_p \to \bigtimes_{i=1}^{k} (0,1)$ such that f is regular and onto.

We force f to be onto because if it wasn't, there would be code vectors in the hypercube that would never be used. This would imply that the quantizer would have to be suboptimal. We use this condition at only one point in the derivation as a constraint on the optimal compander. All equations derived up to that point are still valid without this restriction. We will sometimes represent this mapping as

$$f = (f_1(\underline{x}), f_2(\underline{x}), \ldots, f_k(\underline{x}))^T.$$

Let $\underline{r} = (r_1, r_2, \ldots, r_k)^T$ be the error vector in the hypercube. As stated above, under some fairly general conditions, $E\{r_i r_j\} = \frac{E\{r^2\}\delta_{ij}}{k}$ where δ_{ij} is the Kroeneker delta. Assuming very small distortion, a good approximation to the final error vector in the output is $(f^{-1})'(\underline{x})\underline{r}$. Let $\underline{y}$ be the variable in the hypercube. If $\underline{y} = f(\underline{x})$ then

$$p_y(\underline{y}) = \frac{p_x(f^{-1}(\underline{y}))}{|f'(f^{-1}(y))|} .$$

Therefore the final output mean square error (mse) may be written

$$\text{mse} = \int_{\bigtimes_{i=1}^{k}(0,1)} \underline{r}^T (f^{-1})'^T(f^{-1}(\underline{y}))(f^{-1})'(f^{-1}(\underline{y}))\underline{r}\,\frac{p_x(f^{-1}(\underline{y}))d\underline{y}}{|f'(f^{-1}(y))|}$$

Let $\underline{x} = f^{-1}(\underline{y})$ then $d\underline{x} = |(f^{-1})'(y)|d\underline{y}$ and note that $|(f^{-1})'(y)| = \frac{1}{|f'(f^{-1}(y))|}$ by the inverse mapping theorem [7]. Therefore, making these changes of variable, we obtain

$$\text{mse} = \int_{D_p} \underline{r}^T [f'(\underline{x})]^{-1T}[f'(\underline{x})]^{-1}\underline{r}\, p_x(\underline{x})d\underline{x},$$

again by the inverse mapping theorem. Denote $[f'(\underline{x})]^{-1T}[f'(\underline{x})]^{-1} = \Sigma^{-1}(\underline{x})$ and note this is a symmetric matrix for every $\underline{x}$. Therefore our problem is to optimize

$$\int_{D_p} \underline{r}^T \Sigma^{-1}(\underline{x})\underline{r} p_x(\underline{x})dx$$

Using a matrix identity the above integral becomes,

$$\int_{D_p} \text{tr}\{\Sigma^{-1}(\underline{x})\underline{r}\underline{r}^T\}p_x(\underline{x})dx .$$

Let us now take the expectation over the $\underline{r}$ variable which is independent of any other quantity in the integral (one can make a random coding argument to do this),

$$E\{rr^T\} = E\left\{\begin{pmatrix} r_1^2 & r_1r_2 & \cdots & r_1r_n \\ r_2r_1 & r_2^2 & & r_2r_n \\ \vdots & & & \vdots \\ r_nr_1 & \cdots & \cdots & r_1^2 \end{pmatrix}\right\} = \frac{E\{r^2\}}{k} I.$$

Therefore

$$\text{mse} = \frac{E\{r^2\}}{k}\int_{D_p} \text{tr}(\Sigma^{-1}(\underline{x}))p_x(\underline{x})dx . \quad (22)$$

This expression is of interest in its own right. $E\{r^2\}/k$ is the mean square error per sample suffered by the hypercube quantization. So the total error is a product of two terms operating independently of one another. Denote the eigenvalues of $\Sigma(\underline{x})$ as $\lambda_i^2(\underline{x})$ $(i = 1, \ldots, k)$. Then

$$\text{mse} = \frac{E\{r^2\}}{k}\sum_{i=1}^{k}\int \frac{p_x(\underline{x})}{\lambda_i^2(\underline{x})}d\underline{x} .$$

Since our map f is onto this implies

$$\int_{D_p} |f'(\underline{x})| = \int \prod_{i=1}^{k} \lambda_i(x)d\underline{x} = 1.$$

Let us minimize the mse subject to the above constraint. It is easy to show first of all that $\lambda_i(x) = \lambda(x)$ for every i. So now minimize

$$\int_{D_p} \frac{p(\underline{x})}{\lambda(\underline{x})^2}\,d\underline{x} \quad \text{subject to constraint}$$

$$\int_{R^k} \lambda(\underline{x})^k d\underline{x} = 1.$$

Let $\beta(\underline{x}) = \lambda(\underline{x})^k$, so minimize

$$\int \frac{p(\underline{x})}{\beta(\underline{x})^{2/k}}\,d\underline{x} \quad \text{where} \quad \int \beta(\underline{x})d\underline{x} = 1.$$

Gersho [2] shows that the optimal $\beta(\underline{x})$ is proportional to $p(\underline{x})^{\frac{1}{1+2/k}} = p^{k/k+2}(\underline{x})$. This implies $\lambda(\underline{x}) = p(\underline{x})^{\frac{1}{k+2}} / (\|p\|_{k/k+2})^{\frac{1}{k+2}}$. Using these eigenvalues, the mse $= \frac{E(r^2)}{k} \|p\|_{k/k+2}$. If an optimal k-dimensional uniform quantizer is implemented in the hypercube this equation gives the same error as Zador's optimum quantizer. Our condition for the optimal compressor, is all of the eigenvalues of the symmetric matrix

$$\sum(\underline{x}) = [f'(\underline{x})][f'(\underline{x})]^T$$

are the same; this implies there exists an orthonormal matrix $\phi(\underline{x})$ such that

$$\phi^T(\underline{x})\sum(\underline{x})\phi(\underline{x}) = \lambda^2(\underline{x})I$$

or

$$\sum(\underline{x}) = \lambda^2(\underline{x})I = [f'(\underline{x})][f'(\underline{x})]^T$$

which implies $\frac{[f'(\underline{x})]}{\lambda(\underline{x})}$ is an orthonormal matrix. Since we know what $\lambda(\underline{x})$ is, in principal we could solve for $f'(\underline{x})$ for every value of $\underline{x}$. Therefore our condition for an optimal compander is that $[f'(\underline{x})]/cp(\underline{x})^{\frac{1}{k+2}}$ be an orthogonal matrix for almost every value of $\underline{x}$ where $c = 1/(\|p\|_{k/k+2})^{\frac{1}{k+2}}$.

When k=2 this condition says that $f(\underline{x})$ must be conformal almost everywhere. Excluding sets of measure zero is an important point. Gersho points out (for the 2 dimensional case) that conformal maps do not exist for circularly symmetric probability densities. One consequence of this is the work by Heppes and Szuz [8] which shows that you can't tesselate a circular region with an arbitrary "surface distribution function" using regular hexagons. There must always be a "slit" where the tessellation fails. This "slit" however is a set of measure zero. It is only local conformality almost everywhere that we need, not global conformality.

We will now do an example illustrating the use of Eq. [22]. Suppose our input probability density $p(\underline{x})$ can be written as $\prod_{i=1}^{k} p(x_i)$. Let $C = 1/\int_{-\infty}^{\infty} p(x)^{\alpha}dx$ and our compressor function $f = (f_1(x_1), f_2(x_2), \ldots, f_k(x_k))^T$ where $f_i(x_i) = \int_{-\infty}^{x_i} p(x)^{\alpha}dx$. With little loss of generality, we will assume f is regular. It is obviously onto. Hence

$$[f'(\underline{x})] = \begin{pmatrix} Cp(x_1)^{\alpha} & & \cdots & 0 \\ 0 & Cp(x_2)^{\alpha} & & \vdots \\ 0 & & & \vdots \\ \vdots & & & 0 \\ 0 & \cdots & 0 & Cp(x_k)^{\alpha} \end{pmatrix}$$

$$[f'(\underline{x})]^{-1} = \begin{pmatrix} \frac{1}{Cp(x_1)^{\alpha}} & 0 & & 0 \\ 0 & \frac{1}{Cp(x_2)^{\alpha}} & & \vdots \\ \vdots & & & 0 \\ 0 & \cdots & 0 & \frac{1}{Cp(x_k)^{\alpha}} \end{pmatrix}$$

The eigenvalues of $\sum^{-1}(\underline{x})$ are $\frac{1}{C^2 p(x_i)^{2\alpha}}$, $i=1,\ldots,k$. So the error may be written

$$\text{mse} = \frac{E(r^2)}{k} \sum_{i=1}^{k} \int_{D_p} \frac{\prod_{i=1}^{k} p(x_j)\,d_x}{C^2 P(x_i)^{2\alpha}}$$

$$= \frac{E(r^2)}{C^2} \int_{-\infty}^{\infty} p(x)^{1-2\alpha}dx$$

$$= E(r^2)\left[\int_{-\infty}^{\infty} p(x)^{\alpha}dx\right]^2 \left[\int_{-\infty}^{\infty} p(x)^{1-2\alpha}dx\right]$$

Using Hölders inequality we may show that $\alpha = 1/3$ minimizes the error or

$$\text{mse} = E(r^2)\|p\|_{1/3}.$$

But looking at Zador's coefficient for the one dimension case (see Eq. 4) we have

$$\text{mse}_{1\text{-Dim}} = \frac{\|p\|_{1/3}}{12\,N^2}.$$

Therefore this compressor characteristic gives us the same error as the optimal 1 dimensional quantizer if in the hypercube we quantize with one dimensional uniform quantizers. We can quantize in the hypercube using optimal schemes for a coefficient of (as $K \to \infty$)

$$\text{mse} = \frac{\|p\|_{1/3}}{N^2 2\pi e}.$$

Therefore the best we may do with this compressor characteristic is a gain of $\frac{2\pi e}{12} \tilde{=} 1.42$ in signal to quantizing noise ratio, at the expense of implementing optimal uniform quantizers in the hypercube.

As a second example suppose again $p(\underline{x}) = \prod_{i=1}^{k} p(x_i)$. Suppose we choose the eigenvalues of $\sum(\underline{x})$ to be

$$\lambda_i^2(\underline{x}) = \left[\frac{\prod_{\substack{j=1 \\ j \neq i}}^{k} p(x_i)^{\frac{1}{k+1}}}{\int_{-\infty}^{\infty} p(x)^{\frac{k-1}{k+1}} dx} \right]^2$$

This obviously leads to a nonconformal map. We may using Eq. [22] now evaluate the error for such a compressor characteristic to obtain the mean square error to be

$$\text{mse} = E\{r^2\} \; \|p\|_{\frac{k-1}{k+2}} = E\{r^2\} \|p\|_{\frac{(k-1)}{(k-1)+2}}$$

which are the optimal coefficients for k-1 dimensional space. This implies the possibility of obtaining nonconformal mapping functions that will asymptotically give optimal results.

References

[1] P. Zador, Development and Evaluation of Procedures for Quantizing Multivariate Distributions, Ph.D. Dissertation, Stanford University, University Microfilm no. 64-9855.

[2] A. Gersho, "Asymptotically Optimal Block Quantization", IEEE Trans. Inform. Theory, Vol. IT-25, pp. 373-380, July 1979.

[3] Y. Yamada, S. Tazaki, and R. M. Gray, "Asymptotic Performance of Block Quantizers with Difference Distortion Measures", to appear in IEEE Trans. Inform. Theory.

[4] J. A. Bucklew and N. C. Gallaher, "A Note on Optimal Quantization", IEEE Trans. Inform. Theory, Vol. IT-25, pp. 365-366, May 1979.

[5] J. A. Bucklew and N. C. Gallagher, "Optimum Uniform Quantizers", to appear in Inform. Theory, Sept. 1980.

[6] W. R. Bennett, "Spectra of Quantized Signals", B.S.T.J., Vol. 27, pp. 446-472, July 1948.

[7] W. Fleming, Functions of Several Variables. Springer-Verlag, 1977.

[8] A. Heppes and P. Szüsz, "Bemerkung zu einer Arbeit von L. Fejes Toth", El. Math., 15, 1960, pp. 134-136.

*This work was partially supported by Grant AFOSR-78J605.

SOME RECENT DEVELOPMENTS IN QUANTIZATION THEORY*

(Invited Paper)

by

Neal C. Gallagher, Jr.
School of Electrical Engineering
Purdue University
West Lafayette, IN 47907

and

James A. Bucklew
Department of Electrical and
Computer Engineering
University of Wisconsin
Madison, Wisconsin 53705

Abstract

A critical review of many important developments in quantization theory is presented beginning with Bennett's 1948 paper [1]. The purpose of this study is to resolve some seemingly conflicting results. We then turn to a discussion block or vector quantizers. We show that minimum mean squared error block quantizers preserve the input mean in the output variable and that the error equals the variance of the input minus the variance of the output. We also illustrate a way by which the compander method of quantizer implementation may be extended to block quantizers.

I. Introduction

The quantization problem has been around for ages. The fact is that almost all real numbers must be quantized if they are to be represented by use of a finite number of digits. If we are to choose a real number at random, the probability is one that the number would need to be quantized for representation with a finite number of digits. Early modern work on quantization includes the work of Bennett [1], and Panter and Dite [2]. Bennett is the first to present an analysis of companding systems. A typical companding system is shown in Fig. 1, where the sy tem input is x and output is y.

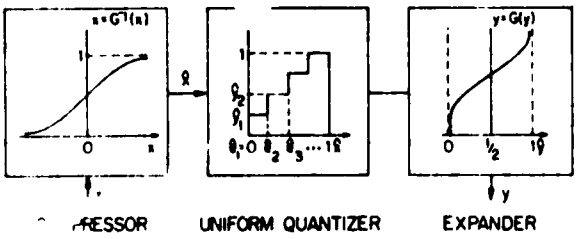


Figure 1 Typical Companding System

*This wor as supported in part by the Air Force Office Scientific Research under Grant AFOSR 78-3605.

The input is first compressed by the nonlinearity $G^{-1}(\cdot)$ whose output is uniformly quantized over the interval [0,1). It is this quantized value that may be transmitted over a communication link or stored in digital memory. When we require a true representation of this quantized value, this uniformly quantized value is expanded by the nonlinearity $G(\cdot)$. Bennett presents an expression for the mean square quantization error for a companding system in the asymptotic (large N) case. This work for further extended by Panter and Dite who studied the design of optimum non-uniform step size quantizers. They derived asymptotic expressions (large N) for finding the minimum mean squared error quantizer design.

In his studies, Bennett made a number of empirical observations concerning the statistical properties of quantized signals. These observations where given a theoretical foundation by Widrow [3]. Widrow showed that the instantaneous quantization error, which is a signal dependent error, can be treated as statistically independent from the signal and uniformly distributed over the quantizer step size (for equal step size quantizers) when the number of quantization levels is sufficiently large. In another often referenced paper Smith [4] further extended Panter and Dite's results and compares theoretical and experimental studies. These four papers by Bennett, Panter and Dite, Widrow, and Smith form the basis for subsequent work on quantization.

By 1957 there was still no exact solution for the optimum quantizer; however, during this time at Bell Labs, Lloyd [5] completed an unpublished technical memorandum in which he provides a method of solution for the optimum quantizer. It is unfortunate that Lloyds work was never published because it is Max's 1960 paper [6] that receives most of the acknowledgement for solving the optimum quantizer design problem. Max's paper is probably the most widely referenced paper on quantization. In their respective papers Lloyd and Max develop necessary conditions for the optimum quantizer; however, these conditions are not sufficient and they can be satisfied for non-optimum quantizers. In 1964 Fleischer [7] presented conditions under

To be presented at the 12th Annual Symposium on System Theory, May 19-20, 1980, Virginia Beach, Virginia.

which Max's results are also sufficient. Fleischer's conditions establish that Max's results are both necessary and sufficient for the optimum quantization of many random variables which have common distributions such as Gaussian or Rayleigh.

We move to 1977 when Sripad and Snyder [8] considered the correlation between the input signal and quantization noise. Their work actually represents a re-evaluation of Widrow's results. They developed necessary and sufficient conditions for the quantization error to be uniform and uncorrelated with the input. These conditions are very restrictive and are not satisfied by most common densities of interest. Although Sripad and Snyder do not actually do so, the flavor of their paper is to contradict the observations of Bennett and Widrow that the quantization noise behaves like uniformly distributed, uncorrelated (independent) additive noise. The difference between these apparently conflicting claims is that Sripad and Snyder are saying that quantization noise is usually not exactly uniformly distributed and uncorrelated with the signal, while Widrow is saying that although these properties are not exact they often are almost valid. Experimental evidence seems to verify Widrow's conclusions as being valid in most situations while Sripad and Snyder are cautioning us to be careful in applying Widrow's conclusions.

The work of Widrow, and Sripad and Snyder applies only to uniform step size quantizers with an infinite number of output levels; of course, real quantizers have only a finite number of output levels. Thus, for a real quantizer the error analysis may be divided into two parts: one part occurs when the input signal falls within the quantizer's range, called non-truncation error, and the other is called truncation error and occurs when the input signal falls beyond the quantizers range. The analysis of Widrow, and Sripad and Snyder implicitly assumes that the contribution of the truncation error can be made arbitrarily small by choosing the quantizer range to be arbitrarily large. For a quantizer with a finite number of output levels this is not possible because the quantizer error will increase in an unbounded manner as the quantizer range increases. If we turn attention to optimum uniform step size quantizers, where the quantizer step size is chosen so as to minimize the total error, we can study the optimum relationship between the truncation and non-truncation errors. We may then study the limiting behavior of the error as the number of output levels becomes large and determine the relative effect of the truncation error. In section II we will study the effect of truncation error and illustrate through an example the fact that truncation can not be ignored.

Optimum quantizers, both uniform step-size and non-uniform step-size, possess a number of interesting properties not proven until the 1979 paper of Bucklew and Gallagher [9]. Here it is shown that for the non-uniform step-size minimum mean square error quantizer the output mean value is equal to the input mean value. It is also shown that the quantizer's error is equal to the input variance minus the output variance. In an unpublished manuscript Bucklew and Gallagher prove that the minimum mean squared error uniform step size quantizer possesses these some two properties. By using these properties it can also be shown that correlation between the quantizer error and input signal is equal to minus the mean squared error. Consequently, for minimum mean squared error quantizers, the signal and noise are negatively correlated, but this correlation is near zero for quantizers with small error. In section III, we present a novel derivation of the aforementioned properties; this derivation is general and is valid for both the optimum non-uniform step-size and uniform step size quantizer.

To this point we have only discussed the quantization of scalar quantities. Often the data to be quantized naturally falls into a k-dimensional representation; typical examples are complex numbers, positional coordinates, or state vectors. In other cases, k-dimensional vectors are formed from blocks of k samples taken from one dimensional signals. The topic of block or vector quantization deals with the representation of multidimensional elements with a finite discrete set of values. In 1964 Zador published his Ph.D. dissertation which contains a number of very interesting results on the properties of optimal block or vector quantizers for the r'th moment euclidean norm distortion measure [10]. Among Zador's contributions are the derivation of both upper and lower bounds on the distortion introduced by the optimal quantizer. Unfortunately, at some points Zador's development is difficult to follow and alternate derivations and extensions by Gersho [11] in 1979, and Yamada et. al. [12] in 1980 have recently appeared. In section IV we present an alternate derivation of Zador's upper bound. Unfortunately, this work on vector quantizers provides very few clues on how to actually find the best quantizer and this remains an unsolved problem at present.

Some of the early work on the implementation vector quantizers actually occurred in the study of computer-generated holograms; see the work of Pearlman [13] and Gallagher [14] for references. The questions treated in this work concerns the representation of two-dimensional vectors in quantized polar format and quantized rectangular format. The reasoning behind this work is to investigate the relative merits of those two-dimensional quantizers that we know how to implement whereas we don't know the optimum implementation. In their 1978 paper Pearlman and Gray [15] employ an information theoretic approach to study the quantization of two-dimensional Gaussian vectors where the vector's X and Y components are independent, zero mean, and identically distributed. In particular they compare polar quantization against rectangular quantization. They show that, when the vector is in polar form, the phase component carries significantly more information than the magnitude component. As a result, the phase component should be quantized very finely in comparison to he magnitude component. Pearlman and Gray show that for a fixed number of output levels $N_P N_R$ = constant, the optimum ratio between the number of phase levels N_P and the number of magnitude levels N_P is approximately $N_G/N_R = 2.6$. In 1979 using a non-information theoretic approach Bucklew and Gallagher [16] rederive this same ratio and then generalize the analysis to circularly symmetric distributions [17]. It is found that in most, but not all, cases polar format quantization is better than rectangular format.

The problem of the design and implementation of optimum vector quantizers remains open. Sections IV, V, and VI of this paper will discuss some recent work toward the solution. In section IV we show that the optimum vector quantizer shares some common properties with the optimum scalar quantizer; in particular the mean value of the quantizer output equals the mean value of the input, and the mean squared error equals the input variance minus the output variance. Section V contains a simplified derivation of Zador's upper bound on the quantizer error, and section VI discusses the possibility of extending the companding concept to multidimensional quantization.

II. Truncation Errors in Optimum Uniform Step-Size Quantizers

Much of the work dealing with the properties of uniform step size quantizers assumes a nonzero step-size Δ with an infinite $N=\infty$ number of output levels. In other words the quantizer has infinite range and never reaches a saturation point which is the largest (or smallest) value to which an input may be quantized. If an input value falls between the largest and smallest saturation points, we say that it is within the quantizer's non-truncation region. If an input value falls beyond a saturation point, we say that the input falls within the truncation region and call this type of error truncation error. Practical quantizers have truncation errors and it is the tradeoff between the truncation and non-truncation errors that is optimized when we design a minimum error quantizer.

A common approximation for the mean squared error of a quantizer with step size Δ is $\Delta^2/12$. This approximation is derived under the assumption that the truncation error is negligible. In many non-pathological situations the contribution of the truncation error can not be ignored; this can be most simply illustrated through an example. Consider the probability density function

$$f(x) = \frac{1+\delta/2}{[1+|x|]^{3+\delta}} . \tag{1}$$

Let the non-truncation region be $(-T,T)$, where the value of T is approximately given by $T=N\Delta/2$. If the quantizer input falls in the region $[T,\infty)$, the output value is $T+\Delta/2$. If the input falls in the region $(-\infty,T]$, the output is quantized to $-T-\Delta/2$. If the input falls within the non-truncation region $(-T,T)$, then the mean squared is accurately approximated as $\Delta^2/12$. Because the density in (1) is even, we can write the following approximate expression for the mean squared error D

$$D = 2\int_{T}^{\infty}[x-T-\frac{\Delta}{2}]^2 f(x)dx + \frac{\Delta^2}{12}P(|X|<T) . \tag{2}$$

In the limit as the number of output levels N goes to infinity, the value of T also approaches infinity and the value of Δ goes to zero. Consequently, for large N, the second term in (2) is accurately represented by $\Delta^2/12$. In the first term of (2), for large values of T, the density in (1) may be approximated as

$$f(x) \approx \frac{1+\delta/2}{|x|^{3+\delta}} ; \tag{3}$$

therefore, for large T

$$D \approx 2\int_{T}^{\infty}[x-T-\frac{\Delta}{2}]^2 \frac{1+\delta/2}{|x|^{3+\delta}}dx + \frac{\Delta^2}{12}$$

$$= K_1T^{-\delta} + K_2T^{-(1+\delta)} + K_3T^{-(2+\delta)} + \frac{\Delta^2}{12}$$

where

$$\approx K_1T^{-\delta} + \frac{\Delta^2}{12} , \tag{4}$$

$$K_1 = -\frac{(2+\delta)}{\delta T^{\delta}}$$

$$K_2 = \frac{\Delta(2+\delta)}{1+\delta} - \Delta ,$$

and

$$K_3 = (\frac{\Delta}{2})^2 .$$

Equation (4) may be rewritten as

$$D \approx K_1(\frac{N\Delta}{2})^{-\delta} + \frac{\Delta^2}{12} . \tag{5}$$

This is an approximate expression valid for large N. To find the optimum value for Δ, we take the derivative of D with respect to Δ and set the resulting expression equal to zero. This yields

$$(\frac{N\Delta}{2})^{-\delta} = \frac{\Delta^2}{6\delta K_1} ;$$

consequently, the minimum D is given by

$$D = \frac{\Delta^2}{12}(1+\frac{2}{\delta}) , \tag{7}$$

Depending on the value of δ, the value of D can be significantly larger than the common $\Delta^2/12$ approximation. Loosely speaking, the validity of this approximation seems to depend upon the existence of higher order moments. If all moments exist, then the approximation appears to be asymptotically correct. If only a few moments exist, it does not seem to be a good approximation.

III. First and Second Moment Properties of Optimum Quantizers

At this point a general definition of a quantizer is required. First, the input signal space is partitioned into N disjoint and exhaustive regions $S_1,S_2,\ldots,S_N$. The quantizer function is defined by the function Q(x), where for input value x

$$Q(x) = y_i, \text{ if } x \in S_i . \tag{8}$$

Note that this definition does not require $y_i \in S_i$, although in practice y_i is usually contained in S_i.

The performance of the quantizer is measured by the mean squared distortion

$$D = E\{[X-Q(\ )]^2\}, \tag{9}$$

for random input X. Assume that the optimum quantizer characteristic is denoted by $Q_0(x)$. At this point, we may or may not add the restriction the $Q_0(x)$ represent an uniform step-size quantizer. This restriction may be represented in the functional form of Q(x) and $Q_0(x)$. Consider the quantizer function $Q(x) = Q_0(x) + \epsilon\delta Q(x)$, where $\delta Q(x)$ represents an arbitrary variation and ϵ is a real valued constant. It should be noted that the term $\epsilon\delta Q(x)$ must be such that Q(x) is a legitimate quantizer characteristic. If the uniform step size restriction is in place, then Q(x) must satisfy this restriction clearly $\epsilon = 0$ is the optimum choice for this parameter; thus,

$$\left.\frac{\partial D}{\partial \epsilon}\right|_{\epsilon=0} = 0 , \tag{10}$$

or

$$E\{[X-Q_0(x)]\delta Q(X)\} = 0 . \tag{11}$$

As proven to this point, the condition in (11) is only a necessary condition for the optimum quantizer. In order to prove that this condition is also sufficient we consider the error D for an arbitrary quantizer $Q(x) = Q_0(x) + \epsilon\delta Q(x)$.

$$\begin{aligned} D &= E\{[X-Q(X)]^2\} \\ &= E\{[X-Q_0(x)]^2\} - 2\epsilon E\{[X-Q_0(x)]\delta Q(X)\} \\ &\qquad + \epsilon^2 E\{[\delta Q(X)]^2\} . \end{aligned} \tag{12}$$

The first term in this expression is the error for the optimum quantizer $Q_0(x)$. The second term is zero by (11), and the third term must be non-negative. Consequently (11) is both a necessary and sufficient condition for an optimum.

We can use (11) to show that for the optimum quantizer the mean of the output equals the mean of the input. To do this we choose the arbitrary variation $\delta Q(x) = 1$; therefore by (11)

$$E\{X-Q_0(X)\} = 0. \tag{13}$$

If we choose $Q_0(x) = x$, then

$$E\{XQ_0(X)\} = E\{[Q_0(X)]^2\}, \tag{14}$$

and consequently,

$$E\{[X-Q_0(X)]^2\} = E\{X^2\} - E\{[Q_0(X)]^2\}, \tag{15}$$

and finally,

$$E\{[Q_0(X)]^2\} = E\{X^2\} - E\{[X-Q_0(X)]^2\}. \tag{16}$$

Equation (13) states that the mean squared error for the optimum quantizer equals the input variance minus the output variance. Equation (16) indicates that the correlation between the quantization error and the quantizer input is equal to the negative of the mean squared error.

IV. First and Second Moment Properties of the Optimum Vector Quantizer

As the second moment properties for the vector quantizer are similar to those properties discussed in the previous section for the scalar quantizer, we will only sketch their derivation. We consider the k-dimensional case where the distortion D is measured as

$$D = \frac{1}{k} E\{||\underline{X}-Q(\underline{X})||^2\}, \tag{17}$$

where $\underline{X}$ and $Q(\underline{X})$ are vector valued, and $||\cdot||$ denotes the usual Euclidean distance norm. Again a variational approach is employed, where an arbitrary quantizer function $Q(\underline{x})$ is written in terms of the optimal quantizer as

$$Q(\underline{x}) = Q_0(\underline{x}) + \epsilon\ \delta Q(\underline{x}),$$

for a vector-valued variation $\delta Q(\underline{x})$. As before, we take the derivative of D with respect to ϵ, and set the result equal to zero at $\epsilon=0$; the result is

$$E\{[\underline{X}-Q_0(\underline{X})]^T\delta Q(\underline{X})\} = 0, \tag{18}$$

where $[\ ]^T$ denotes the transpose of the column vector. This expression is both necessary and sufficient for the optimum quantizer $Q_0(\underline{X})$. The optimum vector quantizer also has the following properties analogous to those scalar quantizer properties found in (14), (15), and (16):

$$E\{[\underline{X}-Q_0(\underline{X})]^T Q_0(\underline{X})\} = 0 \tag{19}$$

$$\frac{1}{k}E\{||\underline{X}-Q_0(\underline{X})||\} = \frac{1}{k}[E\{||X||^2\} - E\{||Q_0(\underline{X})||^2\}], \tag{20}$$

and

$$\frac{1}{k}E\{[Q_0(\underline{X})-\underline{X}]^T\underline{X}\} = -\frac{1}{k}E\{||Q_0(\underline{X})-X||^2\}. \tag{21}$$

V. Zador's Random Quantization Bound

The quantizer input is a k dimensional random vector in R_k which is quantized to one of N levels $\underline{y}_1,\underline{y}_2,\ldots,\underline{y}_N$ in R_k. The space R_k is partitioned into N disjoint and exhaustive regions $S_1,S_2,\ldots,S_N$. The quantizer is defined by the function $Q(\underline{x})$, where for k-dimensional input value $\underline{x}$,

$$Q(\underline{x}) = \underline{y}_i, \quad \text{if } \underline{x} \in S_i.$$

The performance of the quantizer is measured by the distortion

$$D = \frac{1}{k} E\{||\underline{X} - Q(\underline{X})||^r\}$$

The case where r=2 is the usual mean squared distortion. The expression derived by Zador [10] and Gersho [11] for the minimum distortion D_o obtained by use of the best quantizer is

$$D_0 = N^{-\frac{r}{k}} C(k,r) ||p(\underline{x})||_{k/(k+r)}, \quad (22)$$

where $p(\underline{x})$ is the probability density for the input vector $\underline{X}$, and

$$||p(\underline{x})||_\alpha = [\int [p(\underline{x})]^\alpha d\underline{x}]^{1/\alpha} .$$

The constant $C(k,r)$, called the coefficient of quantization, is independent of the density $p(\underline{x})$ and is in general unknown. This expression is an asymptotic result valid only for large N. Two special cases for which the value of $C(k,r)$ is known exactly are [11]

$$C(1,r) = \frac{1}{r+1} 2^{-r},$$

and

$$C(2,2) = \frac{5}{36\sqrt{3}} .$$

Consider the density $p(x)$ that has a constant value of one over the unit volume hypercube; then, $||p(x)||_{k/(k+r)} = 1$. Consequently, Eq. (22) becomes

$$D_0 = N^{-\frac{r}{k}} C(k,r). \quad (23)$$

So, we see that by finding a bound on D_0 we also bound $C(k,r)$. To find this bound we choose the quantizer output levels to have a random distribution uniformly distributed over the hypercube. For a particular input value $\underline{x}$ we find the closest output level and quantize to that value. Because this quantizer is not the optimum, the associated distortion will bound from above the distortion for the optimum quantizer.

To begin, place at random N independent uniformly distributed k dimensional samples in the hypercube. These will be our output levels. We take the quantizer input $\underline{X}$ to have a uniform distribution over the hypercube. We also assume that N is sufficiently large so that there is a very small probability that the quantizer input is closer to an edge of the hypercube than to one of the output values. Suppose that an input value $\underline{x}$ has arrived and is sitting in the hypercube waiting to be quantized. The probability that one particular output value is within a distance ρ of this input sample is given approximately by the volume of a sphere of radius ρ about that sample point, or

$$\text{Prob (one particular output level is within } \rho \text{ of the input sample)} = V_k \rho^k, \quad (24)$$

where if V_k is volume of the unit radius sphere, then $V_k \rho^k$ is the volume of the sphere with radius ρ. We are interested in the closest level to the input sample. We want to know the probability that the closest output level is within a distance ρ of the input sample. To compute this probability, we combine classical order statistics with the result found in (23). By employing this approach, we compute the probability density $f(\rho)$ for the distance between the input sample and the nearest output level to be

$$f(\rho) = N[1 - V_k \rho^k]^{N-1} V_k k \rho^{k-1}. \quad (25)$$

Note that for large values of N this probability density goes to zero rapidly as ρ increases. By construction $\rho = ||\underline{x} - \underline{y}_i||$, where $\underline{x}$ is the input value and $\underline{y}_i$ is the output value. Consequently,

$$E[||\underline{X} - Q(\underline{X})||^r] = E[\rho^r]; \quad (26)$$

so,

$$D = \frac{1}{k} E[\rho^r]$$
$$= \frac{1}{k} \int_{\text{hypercube}} \rho^{r+k-1} N[1-V_k p_k]^{N-1} k\, V_k \, d\rho$$

If we make the change of variables $s = V_k \rho^k$, then we use the fact that $s \leq 1$ to write

$$D \leq \frac{N}{k\, V_k^{r/k}} \int_0^1 s^{r/k} [1-s]^{N-1} ds$$

$$= \frac{N}{k\, V_k^{r/k}} \frac{\Gamma(1+\frac{r}{k})\Gamma(N)}{\Gamma(N+1+\frac{r}{k})}, \quad (27)$$

where $\Gamma(\cdot)$ is the gamma function. For large N the following approximation is valid:

$$\frac{\Gamma(N)}{\Gamma(N+\frac{k+r}{k})} = N^{-\frac{k+r}{k}} .$$

Therefore,

$$D = \frac{N^{-r/k}\Gamma(1+\frac{r}{k})}{kV_k^{r/k}} \quad (28)$$

Because $D \geq D_0$, we use (22) to write

$$C(k,r) \leq \frac{\Gamma(1+\frac{r}{k})}{kV_k^{r/k}}, \quad (29)$$

which is Zador's random quantization upper bound.

VI. Companding in Several Dimensions

For one dimensional quantizers companding provides a method whereby asymptotically optimum quantizers may be implemented in a straightforward fashion. In several dimensions the compressor characteristic is a mapping function $f: R \to \prod_{i=1}^{k} (0,1)$, where $\prod_{i=1}^{k}$ denotes the Cartesian cross product. The set $\prod_{i=0}^{k} (0,1)$ is of course the k-dimensional hypercube. In the companding approach to optimal quantization, we have quantizer output levels distributed in the hypercube. We choose from these output levels the nearest neighbor to $f(\underline{x})$, where $\underline{x}$ is the input data vector. Our

quantized output is then f^{-1} of this particular output level. Denote the error vector caused by quantization in the hypercube as $(r_1, r_2, \ldots, r_k)^T$ and impose the condition that $E(r_i r_j) = \sigma_r^2 \delta_{ij}$, where δ_{ij} is the Kronecker delta. It may be shown that as the number of output levels N in the hypercube approaches infinity, that the error vector for an optimal quantizer converges to a hyperspherically symmetric probability density which satisfies the above condition. In addition, for large N there are an infinite number of quantizers each of which has approximately the same near optimum error. These quantizers may be generated as translations of one another within the hypercube. A simple way to visualize this fact is by use of the one dimensional companding system where the compressor function output is a uniformly distributed (0,1) random variable. We can form translations of the uniform quantizer and still obtain approximately the same mean squared error in the expander output. So, we may consider an ensemble of near optimum quantizers over the hypercube, where each quantizer approaches the optimum quantizer in the asymptotic (large N) case. By allowing us to choose (in an arbitrary fashion) from this ensemble of quantizers for each input vector $(x_1, x_2, \ldots, x_k)^T$, we can decouple the error vector $(r_1, r_2, \ldots, r_k)^T$ from the input so as to make this error vector approximately independent of the input vector. This procedure is analogous to the technique of assigning a random time origin to sampling operations in order to model the sampled signals as wide sense stationary processes.

Our data will be assumed to be k-dimensional samples from a probability density function $p(\underline{x})$, $\underline{x} \in R_k$. Denote S_p as the support of $p(\underline{x})$. Let $f: S_p \to \underset{i=1}{\overset{k}{\times}} (0,1)$ such tht f is regular and onto. We can represent this mapping as

$$f = (f_1(\underline{x}), f_2(\underline{x}), \ldots, f_k(\underline{x}))^T.$$

Let $\underline{r} = (r_1, r_2, \ldots, r_k)$ be the error vector in the hypercube. Assuming very small distortion a good approximation to the final error vector in the output is $(f^{-1})'(\underline{x})\underline{r}$, where $(f^{-1})'(\underline{x})$ represents the matrix of partial derivatives of the inverse operator f^{-1}. Let $\underline{y}$ be the variable in the hypercube. If $\underline{y} = f(\underline{x})$, then the probability density for $\underline{y}$ may be written in terms of the probability density for $\underline{x}$ as

$$p_y(\underline{y}) = \frac{p_x(f^{-1}(\underline{y}))}{|f'(f^{-1}(\underline{y}))|}.$$

Therefore the final output mean square error D may be written

$$D = \int_{\underset{i=1}{\overset{k}{\times}}} \underline{r}^T (f^{-1})'^T (f^{-1}(\underline{y}))(f^{-1})'(f^{-1}(\underline{y}))\underline{r} \frac{p_x(f^{-1}(\underline{y}))d\underline{y}}{|f'(f^{-1}(\underline{y}))|}.$$

Let $\underline{x} = f^{-1}(\underline{y})$, then $d\underline{x} = |(f^{-1})'(\underline{y})|d\underline{y}$ and note that $|(f^{-1})'(\underline{y})| = \frac{1}{|f'(f^{-1}(\underline{y}))|}$ by the inverse mapping theorem. Making these changes we can write

$$D = \int_{S_p} \underline{r}^T [f'(\underline{x})]^{-1T} [f'(\underline{x})]^{-1} \underline{r} p_x(\underline{x}) d\underline{x}.$$

Denote $[f'(\underline{x})]^{-1T}[f'(\underline{x})]^{-1} = \Sigma^{-1}(\underline{x})$ and note this is a symmetric matrix for every $\underline{x}$. Therefore our problem is to minimize

$$D = \int_{S_p} \underline{r}^T \Sigma^{-1}(\underline{x}) \underline{r} p_x(\underline{x}) d\underline{x}.$$

As discussed earlier, there is an ensemble of near optimum quantizers. If we now average the distortion D over this ensemble, we assume that the error vector $\underline{r}$ is sufficiently decoupled from the input vector so as to be treated as an independent random quantity. Consequently, we have

$$D = \int_{S_p} \mathrm{tr}\{\Sigma^{-1}(\underline{x}) \underline{r}\underline{r}^T\} p_x(\underline{x}) d\underline{x}.$$

$$= \sigma_r^2 \int_{S_p} \mathrm{tr}\{\Sigma^{-1}(\underline{x})\} p_x(\underline{x}) d\underline{x}.$$

So, the total error is a product of two terms operating independently of one another. Denote the eigenvalues of $\Sigma(\underline{x})$ as $\lambda_i^2(\underline{x})$ $(i = 1, \ldots, k)$. Then

$$D = \sigma_r^2 \sum_{i=1}^{k} \int P_x(\underline{x}) / \lambda_i^2(\underline{x}) d\underline{x}. \tag{30}$$

Consider, for the moment, a random vector with a uniform distribution over the hypercube $\underset{i=1}{\overset{k}{\times}} (0,1)$; the $f^{-1}(\cdot)$ function maps this vector to a vector in R_k with support S_p and density $|f'(\underline{x})|$. Therefore we have

$$\int_{S_p} |f'(\underline{x})| = \int \prod_{i=1}^{k} \lambda_i(\underline{x}) d\underline{x} = 1 \tag{31}$$

The problem now is to minimize the expression in (30) subject to the constraint in (31). We may do this in the following fashion: (1) Assume that except for $\lambda_j(\underline{x})$ all of the $\lambda_i(\underline{x})$ are the optimum choice. (2) Use a variational method to optimize $\lambda_j(\underline{x})$ subject to the constraint (31). The result is that $\lambda_i(x) = \lambda(x)$ for all i and that the optimum $\lambda(x)$ is

$$\lambda(x) = p(x)^{\frac{1}{k+2}} / (||p||_{k(k+2)})^{\frac{1}{k+2}}. \tag{32}$$

Using these eigenvalues, we find that the minimum error D_0 is given by

$$D_0 = \sigma_r^2 ||p||_{k(k+2)}.$$

If an optimal k-dimensional uniform quantizer is implemented in the hypercube (this determines the value of σ_r^2) this expression gives the same error as Zador's optimum quantizer [10]. Additional results on the properties and implementation of multi-dimensions companding systems are presented in an as yet unpublished paper by Bucklew [18].

References

1. W. R. Bennett, "Spectra of Quantized Signals," Bell Syst. Tech. J., Vol. 27, pp. 446-472, July 1948.
2. P. F. Panter and W. Dite, "Quantization in Pulse-Count Modulation with Nonuniform Spacing of Levels," Proc. IRE, Vol. 39, pp. 44-48, 1951.
3. B. Widrow, "A Study of Rough Amplitude Quantization by Means of Nyquist Sampling Theory," IRE Trans. Circuit Theory, Vol. CT-3, pp. 266-276, Dec. 1956.
4. B. Smith, "Instantaneous Companding of Quantized Signals," Bell Syst. Tech. J., Vol. 47, pp. 653-709, May 1957.
5. S. P. Lloyd, "Least Squares Quantization in PCM," unpublished memorandom, Bell Laboratories, 1957.
6. J. Max, "Quantization for Minimum Distortion," IRE Trans. Inform. Theory, Vol. IT-6, pp. 7-12, Mar. 1960.
7. P. E. Fleisher, "Sufficient Conditions for Achieving Minimum Distortion in a Quantizer," IEEE Int. Conv. Rec., Vol. I, 1964, pp. 104-11.
8. A. B. Sripad and D. L. Snyder, "A Necessary and Sufficient Condition for Quantization Errors to be Uniform and White," IEEE Trans. Acoust. Speech, Signal Processing, Vol. ASSP-25, pp. 442-448, Oct. 1977.
9. J. A. Bucklew and N. C. Gallagher, "A Note on Optimal Quantization," IEEE Trans. Inform. Theory, Vol. IT-25, pp. 365-366, May 1979.
10. P. Zador, Development and Evaluation of Procedures for Quantizing Multivariate Distribution, Ph.D. Dissertation, Stanford University, 1964, University Microfilms, Inc., Ann Arbor, Michigan, #64-9855.
11. A. Gersho, "Asymptotically Optimal Block Quantization," IEEE Trans. Inform. Theory, Vol. IT-25, pp. 373-380, July 1979.
12. Y. Yamada, S. Tazaki, and R. M. Gray, "Asymptotic Performance of Block Quantizers with Difference Distortion Measures," To appear in IEEE Trans. Inform. Theory.
13. W. A. Pearlman, Quantization Error Bounds for Computer Generated Holograms, Stanford Univ. Inform. Syst. Lab., Stanford, CA, Tech. Rep. #6503-1, Aug. 1974.
14. N. C. Gallagher, "Optimum Quantization in Digital Holography," Appl. Opt., Vol. 17, pp. 109-115, Jan. 1, 1978.
15. W. A. Pearlman and R. M. Gray, "Source Coding of the Discrete Fourier Transform," IEEE Trans. Inform. Theory, Vol. IT-24, pp. 683-692, Nov. 1978.
16. J. A. Bucklew and N. C. Gallagher, "Quantization Schemes for Bivariate Gaussian Random Variables," IEEE Trans. Inform. Theory, Vol. IT-25, pp. 537-543, Sept. 1979.
17. J. A. Bucklew and N. C. Gallagher, "Two-Dimensional Quantization of Bivariate Circularly Symmetric Densities," IEEE Trans. Inform. Theory, Vol. IT-25, pp. 667-671, Nov. 1979.
18. J. A. Bucklew, "Companding and Random Quantization in Several Dimensions," Submitted for publication.

PASSBAND AND STOPBAND PROPERTIES OF MEDIAN FILTERS*

N. C. Gallagher, Jr.
School of Electrical Engineering
Purdue University
W. Lafayette, IN 47907

G. L. Wise
Department of Electrical Engineering
University of Texas
Austin, Texas 78712

ABSTRACT

Median filtering is a signal smoothing technique that has been applied successfully in speech and picture processing. However the method suffers from the want of a formal theory by which filter properties may be studied. In this paper Necessary and sufficient conditions for a signal to be invariant under median filtering are derived. These conditions state that a signal must be locally monotone to pass through a median filter unchanged. It is proven that successive median filtering of a signal (i.e. the filtered output is itself filtered) reduces the original signal to an invariant signal called a root signal. For a signal of length L samples a maximum of $\frac{1}{2}(L-2)$ repeated filterings produces an root signal.

I. Introduction

In many signal processing applications a method called median filtering has achieved some very interesting results. One useful characteristic of median filtering is its ability to preserve signal edges while also filtering out impulses. Promising applications of median filtering are picture processing, and speech processing [1-3]. These applications employ the median filter as a signal smoother. The implementation of a median filter requires a very simple digital nonlinear operation. To begin, we take a sampled and quantized signal of length L; across this signal we slide a window that spans 2N+1 signal sample points. The filter output is set equal to the median value of these 2N+1 signal samples. The filter output is associated with the time sample at the center of the window. To account for start up and end effects at the two endpoints of the L-length signal, N samples are appended to the beginning and the end of the sequence. The appended samples are constant and equal in value to the first and last samples of the original sequence, respectively. As an example, consider the binary valued sequence of Fig. 1(a), where L=10 and N=1; the median filtered signal is plotted below the input signal. The appended values are marked as X's. Figure 1(b) illustrates the filtering of the same input signal as for Fig. 1(a) but we set N=2; we set N=3 for the example in Fig. 1(c). The signal of Fig. 1 passes undisturbed through the N=1 filter; however it is affected by the N=2 and N=3 filters. The signal would be reduced to a constant value by an N=4 filter.

*The research was supported by the Air Force Office of Scientific Research under grants AFOSR 78-3605 and AFOSR 76-3062.

The results illustrated in Fig. 1 suggest the concept of a filter "passband" and "stopband". The given signal is in the passband of the N=1 filter and the stopband of the N=4 filter. If we view the median filter as one that passes edges but not impulses, then edges for an N=1 filter may be impulses for an N=4 filter. But what about the N=2 and N=3 filters? Suppose the signal of Fig. 1 is filtered twice in succession by the N=2 filter; in other words, the filtered output is again filtered. The result is a constant output identical to that obtained by a single filtering with an N=4 filter. If the constant is filtered again, the output is the same as the filter input; the constant is invariant to median filtering. So, by filtering the original signal two times with an N=2 or N=3 filter we have a resulting signal that is invariant to successive filterings, the same result obtained by a single pass with the N=4

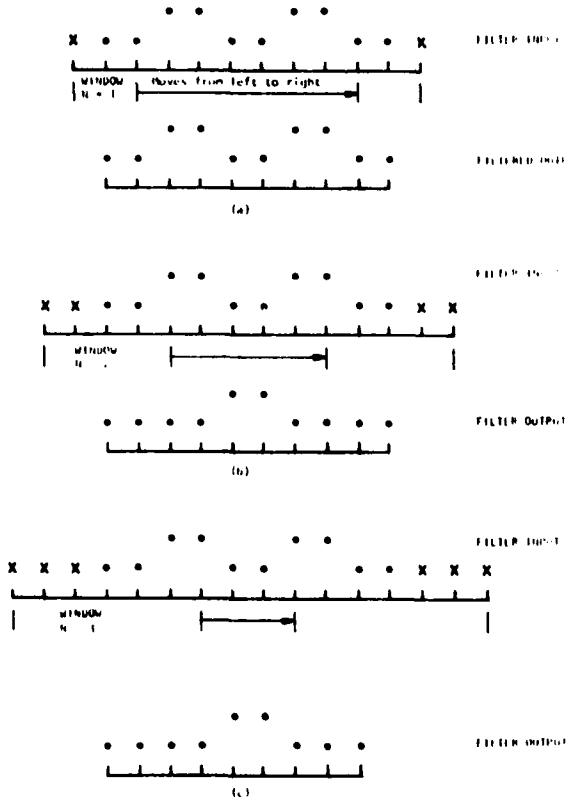


Fig. 1. Signal Filtered by Three Different Median Filters (a) N = 1, (b) N = 2, and (c) N = 3.

filter. Note that the input signal of Fig. 1 is invariant to repeated filtering with an N=1 filter. We see that signals which do not reside entirely within the filter "passband" can be reduced to their passband component by repeated filterings.

At present, there has been no proposed median filter design procedure. There is no method by which the filter window size can be designed to account for some special properties of the signal or noise; the only way of doing this is by trial and error. In this paper we initiate the development of a formal theory for median filters. We will formalize the concepts of filter passband and stopband. We described desirable signal characteristics for signals employed in median filtering and show how some types of noise can be completely removed by median filtering and how other types can not be removed. These results will be presented through the development of a formal theory of median filtering. In section II we present some basic definitions that allow us to precisely state and prove a number of interesting results.

II. Theory for Median Filtering

In order to give a precise statement for the theorems presented later in the section a number of definitions are necessary. We will always be working with a sample length L where each sample is quantized to one of K different values. The filter window length is the number of consecutive samples considered when computing the running median. We will always take the window length to be an odd integer (2N+1) for N=0,1,2,... As noted earlier, our convention is that the filter output at position i is the median value obtained when position i is in the center of the window. We define the following signal characteristics:

1. A constant neighborhood is at least N+1 consecutive identically valued points
2. An edge is a constant neighborhood whose last point is the first point of a monotonic change whose last point is the first point of another constant neighborhood having a different constant value from the first constant neighborhood.
3. An impulse is a constant neighborhood followed by at least one but no more than N points which are then followed by another constant neighborhood having the same value as the first constant neighborhood. The two boundary points of these at most N points do not have the same value as the two constant neighborhoods.
4. An oscillation is a sequence of points which is not part of a constant neighborhood an edge or an impulse.

Of particular interest is the class of signals that can pass through the filter unchanged as well as the class of signals that are completely removed by filtering. Assume that an L-length signal is filtered with a 2N+1 window. As noted previously, we always append to the beginning of the signal an additional N constants equal in value to the first sample of the signal. Similarly, N constant points are appended to the end of the L-length signal. By doing this, we assure that when the initial signal's first or last sample is in the center of the window, the median filter output equals this sample value. For a signal to pass through a median filter unchanged means that the central sample value for each window position is itself the median of the samples within the window.

Consider a signal unchanged by median filtering. Assume that the window increments from sample to sample moving from left to right across the signal and that the window is now centered at the second signal sample of the original signal. We know that the N points to left of center have the same constant value. If they equal the value of the center point, then it (the center point) must be the median. If they are less than the value of the center point, then the N points to the right of center must be all greater than or equal to the central value. If the N points to left are greater in value than the central point, then the N points to the right are all less than or equal to the center value. Thus note that the leftmost N+2 points in the window form a monotone sequence of points. Increment the windown another sample to the right, so that the window is now centered at the third signal sample. The leftmost N+1 samples in the window form a monotone sequence. Assume that the N leftmost points in the window are not greater than (respectively, not less than) the center point. Then, since the center point is the median value of the points in the window, the N rightmost points in the window must be not less than (respectively, not greater than) the center point. Thus we see once again that the leftmost N+2 points in the window form a monotone sequence. Increment the window another sample to the right. By applying the same argument as before, we again find that the N+2 leftmost points in the window form a monotone sequence. Indeed, a straightforward inductive argument proves that the leftmost N+2 points in the window form a monotone sequence regardless of the window position. Recalling that the appended signal has N constant points appended to he right of the original signal, we see that the appended signal is such that any consecutive N+2 points must be monotone. Thus a signal invariant to median filtering must be such that the appended signal contains only constant neighborhoods and edges.

Now assume that the appended signal contains only constant neighborhoods and edges. If the center of the window is at any signal sample, then the points in the window are either monotone or non-monotone. If the points are monotone, then the signal sample at the center of the window is not changed by the median filter. If they are non-monotone, then the window must be centered on a point in the constant neighborhood shared by two edges. Of the 2N+1 points in the window, at least N+1 of them are equal to the center point, and thus the center point is unchanged by median filtering.

These observations are formalized in the following theorem.

Theorem 1. Given a length L, K valued, sequence to be median filtered with a 2N+1 window, a necessary and sufficient condition for the signal to be invariant under median filtering is that the appended signal consist only of constant neighborhoods and edges*.

*It has recently come to our attention that S. Tyan has proven a version of this theorem in an unpublished manuscript. We have not seen a copy of this manuscript at can only speculate as to its contents.

The following corollary is a direct result of this theorem.

Corollary. For a median filter invariant signal to contain both regions of increase and decrease, the points of increase and decrease must be separated by a constant neighborhood (at least N+1 consecutive identical points).

As a result of this theorem it is possible to construct signals that are invariant to median filtering. Also, given the space of all length-L, K-valued signals S it is possible to identify all those signals invariant to median filtering with a 2N+1 window. We will call these signals the roots of the filter, and this set of signals is denoted as R_N. Note that R_N S for any N and that we have the following lemma.

Lemma 1: For an L-length K-valued set of signals S, the root sets R_N are nested such that ...

R_{N+1} R_N ... $R_0 = S$.

Proof. If a signal is invariant to a filter of window length 2(N+1) + 1, then each neighborhood of N+3 samples is monotone. Consequently each neighborhood of length N+2 is monotone and the signal is invariant to a filter window of length 2N+1; i.e. R_{N+1} R_N. It is trivial to verify that a window of length 1 reproduces any signal exactly upon filtering because the median value of a set containing just one point is the value of that point; thus, $R_0 = S$.

We have established that for a given filter window 2N+1 and a signal set S, there exist a root set R_N of signals invariant to filtering. For a given L-length signal s we represent the median filtered version of s by $f_N(s)$ for a 2N+1 size window. We represent by $f_N^{(2)}(s)$ the twice filtered signal:

$$f_N^{(2)}(s) = f_N[f_N(s)].$$

We define $f_N^{(n)}(s)$ as the n-times filtered signal:

$$f_N^{(n)}(s) = f_N[f_N^{(n-1)}(s)].$$

If $s = f_N(s)$, then s is a root of the filter. We next prove that for any signal s there exists an n such that $f_N^{(n)}(s) = r$, where r is a root.

Suppose we are given an L-length signal s that is not a root. Recall that N constant points are appended to the beginning of the signal. By construction, the first original signal point is the median of the interval for which it is the central point. As we slide the window from left to right across the signal, the first point to move (i.e. where the window's central point is not the median) must, by definition, be either a point contained in an impulse or oscillation. Suppose it is an impulse. By construction an impulse has two constant neighborhoods of equal value on either side, and every point in the impulse is filtered to this constant value by one pass of the filter window. Suppose the first point to be moved is contained in an oscillation. Let p be the last point unaffected by the median filter, and assume the filter is centered at this point. Then the leftmost N+2 points must be monotone as seen in the proof of Theorem 1. Assume without loss of generality that they are monotone nondecreasing. Assume that the window is now centered at the point p+1. By hypothesis, this point must change in value. Recall that the leftmost N points are not greater in value than the center point. If the N rightmost points were greater than or equal to the center value, then this value at p+1 would be the median. Thus, at least one point to the right of center must have a value less than that of p+1. Thus there are N+1 points in the window not greater in value than the center point, and the center point changes. Therefore it changes downward in value. Note that it can never achieve a value less than the value of the immediately preceeding constant neighborhood because there are always at least N+1 points contained in the window including p+1 itself whose values are all greater than or equal to the constant neighborhood.

So we see that the first point that changes under filtering is preceded by but not necessarily adjacent to an invariant constant neighborhood, and the point is contained either in an impulse or oscillation. We also see that upon filtering, the value of this point moves closer to the value of the constant neighborhood. There are two possibilities: the value of point p equals the value of p+1, or the value of point p+1 is greater than that of p. In addition, it can be shown that the value of point p+1 is greater than the value of point p. Suppose that the two points have the same value. As the window increments from position p to p+1 one point moves out of the window on left side and another point moves into the window on the right. The point that moves out on the left has a value less than or equal to that of point p+1. Because we know that the filtered value of p+1 is less than the original value, the point that moves in on the right side must also have a value less than that of p+1, otherwise the value of p+1 cannot decrease. If the value of point p+1 is the same as that of p then there remain N points in the window less than or equal to the value at p+1 (and at p) and also N points in the window greater than or equal to the value at p+1; consequently, point p+1 is the median and would not change. Thus, the value of the first point to change must be greater than its predecessor.

Recall what is known concerning the last consecutive point p that is invariant to filtering. The N points in the window to the left of the center point p are all less than or equal to p in value; the N points to the right of p are all greater than or equal to p in value. When the next point p+1, is centered in the window there will be at least N points less than or equal to p in value and at least N+1 points greater than or equal to p in value. Therefore the median value can not be less than the value of p. For convenience we summarize this as the following.

Observation 1: The first point to change value during a median filtering operation must be on the opposite side of its predecessor than the most recent constant neighborhood, and this point upon filtering moves toward its predecessor but does not move past its predecessor.

Continuing in this fashion, consider the point following p+1; that is, p+2. Note that the value

of p+2 is greater than or equal to the value of p. As the window is incremented to the right, p+2 is centered in the window and a point moves out of the window on the left. A new point enters the window on the right. The value of this point must be either greater than that of p or less than or equal to the value of p. If it is less than or equal to the value of p, then there are at least N-1 points in the window with values less than or equal to p and at least N+1 points with values greater than or equal to p. Consequently, p+2 can not be filtered to a value less than p. If the value of the new point is greater than that of p, then trivially, the filtered value of p+2 can not be less than that of p. The same reasoning can be applied to points p+3, p+4, ..., p+N. For convenience, we summarize this as the following.

Observation 2: After filtering, the N rightmost points in the window centered at p must all have values equal to that of p or on the opposite side of the value of p than the most recent constant neighborhood.

Consequently the value of p is always invariant to median filtering, and, in addition the same argument applies to any other (invariant) point to the left of p. Also, the point p+1 has one of two possible filtered values, as follows.

Observation 3: Of all the values in the window centered at p+1, the filtered value of p+1 is either the value of p or the closest value to p on the opposite side as the most recent constant neighborhood.

By using an argument similar to that just presented we reason that the filtered values of p+2 through p+N are greater than or equal to the filtered value of p+1. If the filtered value of p+1 is the same as the value of p, then point p+1 is invariant to filtering on the next pass of the window because it is not greater than the value of p. Suppose, however, that the filtered value of point p+1 is greater than that of p. We must re-examine the pre-filtered point values. When p+1 is in window center, the N+1 rightmost points must all have values greater than that of p including the rightmost point p+N+1. As a result, when p+N+1 is in window center, the leftmost N+1 points have values greater than that of p and the filtered value of p+N+1 must be greater than that of p. Consequently, on the second pass of the window, after all the points have been filtered once, when point p+1 is in window center, the N leftmost points are all less in value than that of p+1, and the rightmost N points all have values greater than or equal to that of p+1. Thus, p+1 is the median of the window and does not change value upon the second filtering. This yields the following.

Observation 4: The first point to change value on a median filtering operation remains invariant upon additional filter passes.

When the observation is made that the median filtering operation is independent of whether the window moves from right to left or left to right across the signal, we see that the properties of the first point to change value apply also to the last point in the signal to change value. Because of the appended constant valued points to the front and back of the L-length signal, the first and last signal points are invariant to filtering. Thus at most $\frac{1}{2}(L-2)$ window passes are required to reduce the signal to a root. As a result of the previous discussion we have the following theorem for an L-length signal.

Theorem 2. Upon successive median filter window passes any non-root signal will become a root after a maximum of $\frac{1}{2}(L-2)$ successive filterings. Also, any non-root signal can not repeat, and the first point to change value on any pass of the filter window will remain constant upon successive window passes.

To illustrate this characteristic of median filtering consider the binary valued L=8 signal of Fig. 2. This signal will be repeatedly filtered by use of a window length of 3 samples. The appended constant terms are marked with x's. We see that $\frac{1}{2}(L-2) = 3$ window passes are required to reduce this signal to a root

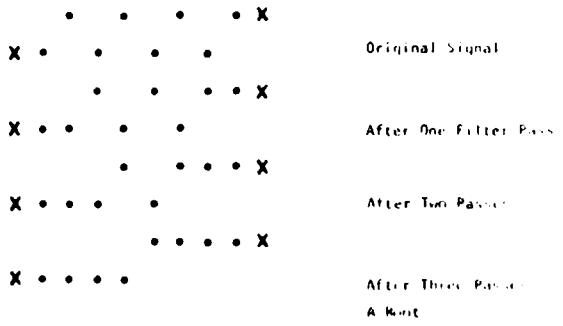


Fig. 2 Result of Repeated Median Filtering

To this point, it has always been assumed that the signal is quantized to K levels for an L-length signal this requirement is not needed because an L-length signal can have at most L different values even if the signal samples are not quantized to specific values. Thus, we can always bound K from above by the value of L and all results stated in this paper apply to unquantized signals.

III. Discussion

The development in the preceeding section suggests a number of interesting results. First, we note that every signal in the space of signals, $s \in S$, can be filtered to a unique root with a bounded number of repeated filterings. Thus, the elements of the root set R_N partition S as illustrated in Fig. 3 where it is shown how the signal space

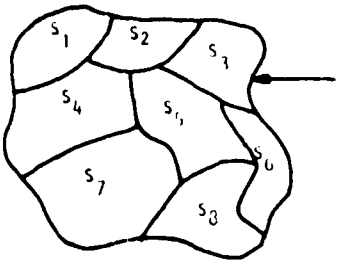


Fig. 3. Partition of the Signal Space S by Eight Roots.

is partitioned by a root set with eight elements, where upon repeated filtering every signal $s \in S_3$ is filtered to root $r_3 \in R_N$ and so on; we will call S_i the ancestor set of root r_i. If a signal s requires l filter passes to reach the root r_3 we say that s is an l-th generation ancestor of r_3. We know from Theorem 2 that any root has at most $\frac{1}{2}(L-2)$ ancestral generations and we know that the root of a signal depends on the filter window size, i.e., a root for a window of size 3 may not be a root for a window of size 5, although a root for a size 5 window is always a root for a size 3 window. In a loose sense, median filters are a type of lowpass filter with an increasingly narrow passband as the window size increases.

The application of median filtering to signal smoothing problems introduces an interesting twist to the concepts of signal and noise. A median filter has no design parameters other than window size. It can not be designed to accomodate special signal or noise characteristics. In the extreme case a filter can completely remove a signal component leaving only noise. It seems desirable that a noise-free signal be a root signal in order that it is invariant to median filtering. If the root signal has added noise, then it may or may not be possible to remove the noise by filtering. Noise that can be filtered is noise that changes the signal in such a way that the noisy signal is an ancestor of the same root. This noise can be removed with repeated filtering. However, if the noisy signal is now the ancestor of a different root, then it can not be removed by repeated filtering. This property of either perfect signal recovery or false signal recovery points to yet another application of median filtering - channel coding. For this application the root set R corresponds to an alphabet set. The transmitted code can contain either roots or ancestors. In either case decoding is accomplished through repeated filtering.

In this paper we have established several fundamental theoretical properties of median filters. We have presented necessary and sufficient conditions for a signal to be invariant to median filtering and we call these signals roots of the filter. We have also shown that repeated filtering of any signal results in a root signal and have established the maximum number of filtering operations required to reach a root. As a result of the theory developed in this paper a better understanding of the potential applications as well as the limitations of these filters is achieved.

References

1. T. S. Huang, G. J. Yang, and G. Y. Yang, "A Fast Two Dimensional Median Filtering Algorithm", IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-27, pp. 13-18, Feb. 1979.

2. N. S. Jayant, "Average- and Median-Based Smoothing Techniques for Improving Digital Speech Quality in the Presence of Transmission Errors", IEEE Trans. Commun., Vol. COM-24, pp. 1:43-1045, Sept. 1976.

3. L. R. Rabiner, M. R. Sambus, and C. E. Schmidt, "Applications of a Nonlinear Smoothing Algorithm to Speech Processing", IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-23, pp. 552-557, Dec. 1975.

ROOT-SIGNAL SET ANALYSIS FOR MEDIAN FILTERS

G. R. Arce and N. C. Gallagher, Jr.
School of Electrical Engineering
Purdue University
West Lafayette, Indiana 47907

ABSTRACT

Invariant signals to median filters are called roots of the signal. A tree structure for the roots of a binary signal is obtained. Showing a state propagation property from which a state diagram is obtained. The number of roots R(n) for a signal of length n and a window filter 2*s-1 is exactly represented by the difference equation R(n)=R(n-1) + R(n-s). A general solution is obtained in a Z domain approach, and in a transformation approach.

SUMMARY

Many properties of a median filter may be described in terms of the so called root signals. A signal invariant to the filter is called a root.

In this paper, a tree structure of the roots is modeled and implemented graphically. This structure has very attractive properties such as symmetry as well as a predictable pattern of state propagation. Each state in the tree generates other states, not necessarily of the same kind; then, the new states generate another group of states and so the tree structure follows. The repetition of states in a tree is a function of the length of the signal, and the number of different kinds of states is a function of the filter window size. At each stage along the tree, each state yields a number of roots.

On the binary signal we obtain 4 different states, states A & D yield 2 roots, and states B & C yield 1 root each. The relation for the number of roots is: $R(n+1) = 2*(A(n)+D(n)) + B(n) + C(n)$, where n represents the signal length. For the binary case the difference equation for R(n) can be shown to be: $R(n+s)=R(n+s-1) + R(n)$, where s depends on the window size. The solution of the difference equation is obtained with a state equation approach. Let: $R(k) = X1(k); R(k+1)=X2(k); R(k+2)=X3(k), \ldots, R(k+s-1)=Xs(k)$. By solving the vector state equation $\underline{X}(k+1)=[\underline{A}]^k \underline{x}(k)$ we obtain the solution: $R(k)=[\,1\ 0\ 0\ 0\ \ldots\ 0]\underline{X}(k)$. Therefore a solution to $[\underline{A}]^k$ is necessary, where the $\underline{A}$ matrix has the form of a bottom companion matrix. The characteristic polynomial for the $\underline{A}$ matrix of size s by s is: $f(X)=X^s - X^{s-1} - 1$. Using Sturm's theorem, we can see that the characteristic function has distinct eignevalues only. Two different approaches are used to obtain A^k. One approach used the Z domain, $A^k = Z^{-1}\{(zI-A)^{-1} z\}$, the other approach uses a similarity transformation: $X=MQ$ where $\underline{A}^k = \underline{M}\ \underline{D}^k\ \underline{M}^{-1}$ and $\underline{D} = \underline{M}^{-1}\ \underline{A}\ \underline{M}$. A closed form solution is then obtained showing that the number of roots for a signal of length k is a linear combination of the eigenvalues raised to the kth power. The Z domain approach yields the result:

$$R(k_o) = \lim_{z\to\infty} \frac{1}{k_o} \left[\frac{-Zd}{dZ}\right]^{k_o} \frac{1}{Z^s - Z^{s-1} - 1} \left[Z^{s-1}(Z-1), Z^{s-2}(Z-1), \ldots, Z(Z-1), Z\right] \underline{X}(o)$$

where k_o is a specific signal length. In the paper we analyze in detail every point touched in this summary.

The authors gratefully acknowledge the support of the Air Force Office of Scientific Research under grant AFOSR 78-3605.

Presented at the Eighteenth Annual Allerton Conference on Communication, Control, and Computing, October 8-10, 1980.

Some Properties of Uniform Step Size Quantizers

JAMES A. BUCKLEW, MEMBER, IEEE, AND NEAL C. GALLAGHER, JR., MEMBER, IEEE

Abstract—**Some properties of the optimal mean-square error uniform quantizer are treated. It is shown that the mean-square error (mse) is given by the input variance minus the output variance. Furthermore $\lim_{N\to\infty} \text{mse}/(\Delta^2/12) \geq 1$, where N is the number of output levels and Δ (a function of M) is the step size of the uniform quantizer, with equality when the support of the random variable is contained in a finite interval. A class of probability densities is given for which the above limit is greater than one. It is shown that $\lim_{N\to\infty} N^2 \cdot \text{mse} = (b-a)^2/12$, where $(b-a)$ is the measure of the smallest interval that contains the support of the input random variable.**

In many problems arising in the evaluation or design of a control or communication system it is necessary to predict the performance of a uniform quantizer. Uniform quantizers are of interest because they are usually the simplest to implement and because many noise processes in physical systems may be considered as the noise produced by a uniform quantizing operation. For example, the final position of a stepping motor or the line drawn by the pen of a computer plotting device under a continuous control may be considered to be corrupted by a uniform quantizing operation.

Because of the importance of these quantizers several authors have considered their properties. Widrow [1] shows that under certain conditions the quantization noise is uniformly distributed. Gish and Pierce [2] show that asymptotically the uniform quantizer is optimum in the sense of minimizing the output entropy subject to a fixed mean-square error. Morris and Vandelinde [3] show the uniform quantizer to be minimax. Sripad and Snyder [4] later extended Widrow's work to give a sufficient condition for the quantization error to be uniform and uncorrelated with the input.

We now prove some additional properties of these quantizers when they are designed to minimize the mean-square error (mse). We may write down the analytic expression for the quantizer characteristic $g(x)$ as

$$g(x) = \begin{cases} a, & \text{if } x < q, \\ a + (i+1)\Delta, & \text{if } q + i\Delta < x < q + (i+1)\Delta, \\ & \text{for } i = 0, \cdots, N-3 \\ a + (N-1)\Delta, & \text{if } x > (N-2)\Delta + q, \end{cases} \tag{1}$$

where N is the number of output levels. We see that if x is less than q or greater than $q+(N-2)\Delta$, x is truncated to a or $a+(N-1)\Delta$, respectively. An important parameter of interest is the measure of the nontruncation region, $(N-2)\Delta$.

The quantizer characteristic ($g(x)$ must be optimized with respect to three parameters, q which fixes its position along the x axis, a which fixes its position along the y axis, and Δ (a function of N) which specifies the step size of the quantizer. Because it makes little sense to speak of minimizing the mean-square error of a random variable with infinite variance, we will always assume $\int_{-\infty}^{\infty} x^2 f(x)\,dx < \infty$.

Property 1: The minimum mean-square error uniform quantizer preserves the mean of the input random variable.

Proof: Suppose $g(x)$ is the optimum uniform quantizer. Then

$$\frac{\partial}{\partial \epsilon} \int (x - g(x) + \epsilon)^2 f(x)\,dx \Big|_{\epsilon=0} = 0, \tag{2}$$

which implies

$$\int x f(x)\,dx = \int g(x) f(x)\,dx. \tag{3}$$

□

Property 2: For the optimum uniform quantizer

$$a = q - \Delta/2.$$

Proof: Suppose $g(x)$ is the optimum uniform quantizer. Then

$$0 = \frac{\partial}{\partial \epsilon} \int (g(x-\epsilon) - x)^2 f(x)\,dx \Big|_{\epsilon=0} \tag{4}$$

$$= \frac{\partial}{\partial \epsilon} \Bigg[\sum_{i=0}^{N-3} (a + (i+1)\Delta)^2 \int_{q+\epsilon+i\Delta}^{q+\epsilon+(i+1)\Delta} f(x)\,dx + a^2 \int_{-\infty}^{q+\epsilon} f(x)\,dx + (a + (N-1)\Delta)^2 \int_{q+\epsilon+(N-2)\Delta}^{\infty} f(x)\,dx \Bigg]$$

$$- 2 \Bigg[\sum_{i=0}^{N-3} (a + (i+1)\Delta) \int_{q+\epsilon+i\Delta}^{q+\epsilon+(i+1)\Delta} x f(x)\,dx + a \int_{-\infty}^{q+\epsilon} x f(x)\,dx + (a + (N-1)\Delta) \int_{q+\epsilon+(N-2)\Delta}^{\infty} x f(x)\,dx \Bigg] \Big|_{\epsilon=0} \tag{5}$$

$$= \Bigg[\sum_{i=0}^{N-3} (a + (i+1)\Delta)^2 [f(q+\epsilon+(i+1)\Delta) - f(q+\epsilon+i\Delta)] + a^2 f(q+\epsilon) - (a + (N-1)\Delta)^2 f(q+\epsilon+(N-2)\Delta) \Bigg]$$

$$- 2 \Bigg[\sum_{i=0}^{N-3} (a + (i+1)\Delta)[(q+\epsilon+(i+1)\Delta) f(q+\epsilon+(i+1)\Delta) - (q+\epsilon+i\Delta) f(q+\epsilon+i\Delta)] + a(q+\epsilon) f(q+\epsilon) - (a + (N-1)\Delta)(q+\epsilon+(N-2)\Delta) f(q+\epsilon+(N-2)\Delta) \Bigg] \Big|_{\epsilon=0}. \tag{6}$$

Simplifying this expression we obtain

$$(\Delta + 2a - 2q) \sum_{i=0}^{N-2} f(q + i\Delta) = 0.$$

The solution $\sum_{i=0}^{N-2} f(q+i\Delta) = 0$ corresponds to a trivial solution because without affecting the mean-square error, we may always arbitrarily set $f(q+i\Delta) = 0, i = 0, \cdots, N-2$. Hence $\Delta + 2a - 2q = 0$ which is what we wish to prove. □

Property 3: The mean-square error of an optimum uniform quantizer is given by the input variance minus the output variance.

Proof:

$$\text{mse} = E(g(x) - x)^2$$

$$= E\{x^2\} - 2E\{x g(x)\} + E\{g(x)^2\}. \tag{7}$$

We wish to optimize this expression with respect to Δ. Using

Manuscript received April 23, 1979; revised October 25, 1979. This work was supported by the Air Force Office of Scientific Research under Grant AFOSR 78-3605. This paper was presented at the 1979 Allerton Conference on Information Sciences and Systems, Monticello, IL, October 10-12, 1979.

J. A. Bucklew was with the School of Electrical Engineering, Purdue University, West Lafayette, IN. He is now with the Electrical and Computer Engineering Department, University of Wisconsin, Madison, WI 53706.

N. C. Gallagher, Jr. is with the School of Electrical Engineering, Purdue University, West Lafayette, IN 47907.

0018-9448/80/0900-0610$00.75

$a = q - \Delta/2$ we first obtain

$$E\{xg(x)\} = \sum_{i=0}^{N-3}\left(q+\left(i+\frac{1}{2}\right)\Delta\right)^2\int_{q+i\Delta}^{q+(i+1)\Delta} xf(x)\,dx + (q-\Delta/2)\int_{-\infty}^{q} xf(x)\,dx + \left(q+\left(N-\frac{3}{2}\right)\Delta\right)\cdot\int_{q+(N-2)\Delta}^{\infty} xf(x)\,dx \tag{8}$$

and

$$E\{g(x)^2\} = \sum_{i=0}^{N-3}\left(q+\left(i+\frac{1}{2}\right)\Delta\right)^2\int_{q+i\Delta}^{q+(i+1)\Delta} f(x)\,dx + \left(q-\frac{\Delta}{2}\right)^2\int_{-\infty}^{q} f(x)\,dx + \left(q+\left(N-\frac{3}{2}\right)\Delta\right)^2\cdot\int_{q+(N-2)\Delta}^{\infty} f(x)\,dx. \tag{9}$$

Substitute (9) and (10) into (8); take the partial derivative with respect to Δ and set the result equal to zero. We find that

$$E\{xg(x)\} + qE\{g(x)\} = E\{g(x)^2\} + qE\{x\}, \tag{10}$$

but $E\{g(x)\} = E\{x\}$ for the optimum quantizer. Hence $E\{xg(x)\} + E\{g(x)^2\}$ and

$$\text{mse} = E\{x^2\} - E\{g(x)^2\} \tag{11}$$

which together with Property 1 completes the proof. □

Sripad and Snyder [4] show that a sufficient condition for $x - g(x)$ to be uniform and uncorrelated with x is

$$\phi_x\left(\frac{2\pi n}{\Delta}\right) = \dot{\phi}_x\left(\frac{2\pi n}{\Delta}\right) = 0 \qquad \text{for } n = \pm 1, \pm 2, \cdots, \tag{12}$$

where $\phi_x(\omega)$ is the characteristic function of the input random variable x and $\dot{\phi}_x(\omega) = d\phi_x(\omega)/d\omega$. Frequently in the analysis of a system corrupted by a uniform quantizing operation it is assumed that the quantization noise is uncorrelated with (or sometimes independent of) the input. The next property demonstrates that this cannot be done with the optimum uniform quantizer.

Property 4: Suppose the input probability density is Riemann-integrable. Then the quantization noise is never uncorrelated with the input for the optimum uniform quantizer.

Proof: Without loss of generality assume $E\{X\} = 0$. Suppose the converse holds. This implies

$$E\{(x - g(x))x\} = E\{x^2\} - E\{g(x)x\} = 0, \tag{13}$$

but from Property 3

$$E\{xg(x)\} = E\{g(x)^2\},$$

$$E\{x^2\} - E\{g(x)^2\} = 0. \tag{14}$$

But, again from Property 3, the left side of (14) is the mean-square error. This is a contradiction, since a Riemann-integrable probability density function necessarily implies that the mean-square error for any finite number of output levels is greater than zero (i.e., $f(x)$ has no delta functions). □

We now state an obvious property which will be used in several subsequent proofs.

Property 5: The mean-square error for the optimal uniform quantizer approaches zero as the number of output levels approaches infinity.

Proof: The mean-square error is given by $E\{(g(x) - x)^2\}$, and for this to approach zero it is sufficient that $g(x)$ approach x in mean-square. Consider a quantizer with the parameters $\Delta = 1/\sqrt{N-2}$ and $q = -(N-2)\Delta/2$. The width of the nontruncation region is $(N-2)\Delta = \sqrt{N-2}$. Hence as N becomes large the width of the nontruncation region approaches infinity and delta approaches zero. It is a simple matter to show that $\lim_{N\to\infty} g(x) = x$ everywhere. Since $(g(x) - x)^2 \le x^2 + \Delta^2$ and $\int_{-\infty}^{\infty}(x^2 + \Delta^2)f(x)\,dx < \infty$, this implies

$$\lim_{N\to\infty}\int_{-\infty}^{\infty}(g(x) - x^2)f(x)\,dx = \int_{-\infty}^{\infty}\lim_{N\to\infty}(g(x) - x)^2 f(x)\,dx = 0$$

by the Lebesgue dominated convergence theorem. This quantizer is in general suboptimal, which implies that an optimal quantizer must have even smaller mean-square error for each N, and hence its error must also go to zero. □

As a consequence of the above property, it is easy to show $\lim_{n\to\infty}\Delta = 0$ for the optimal uniform quantizer.

Let (a,b) be the smallest interval such that $\int_a^b f(x)\,dx = 1$. Note that either $|a|$ or $|b|$ may be infinite.

Property 6: Suppose $f(x)$ is Riemann-integrable. Then, for the optimum uniform quantizer, $\lim_{N\to\infty}(N-2)\Delta = b - a$.

Proof: Suppose $\lim_{N\to\infty}(N-2)\Delta < b - a$. This implies that for N sufficiently large we are always truncating some finite amount of probability mass, and so the mean-square error cannot go to zero. This contradicts the previous property. Hence $\lim_{N\to\infty}(N-2)\Delta \ge b - a$.

Suppose $\lim_{N\to\infty}(N-2)\Delta > b - a$. This makes sense only if the random variable is of finite support. So for N large enough there is no truncation error. In the Appendix it is shown that for a family of quantizers with no truncation error $\lim_{N\to\infty}\text{mse}/(\Delta^2/12) = 1$ for a Riemann-integrable density function. So, for N sufficiently large, $(N-2)\Delta > C > b - a < \infty$. Then

$$1 = \lim_{N\to\infty}\frac{\text{mse}}{\Delta^2/12} \le \lim_{N\to\infty}\frac{\text{mse}}{C^2/12(N-2)^2},$$

or

$$\lim_{N\to\infty}(N-2)^2\text{mse} \ge \frac{C^2}{12}. \tag{15}$$

Consider a suboptimal quantizer whose input intervals are obtained by dividing the interval (a,b) into $N-2$ equal subintervals. Denote the mean-square error of this quantizer by mse_{SUB} and its step size by $\Delta_S = (b-a)/(N-2)$. This quantizer has no truncation error and hence

$$1 = \lim_{N\to\infty}\frac{\text{mse}_{\text{SUB}}}{\Delta_S^2/12} = \lim_{N\to\infty}\frac{\text{mse}_{\text{SUB}}}{(b-a)^2/12(N-2)^2},$$

$$\lim_{N\to\infty}(N-2)^2\text{mse}_{\text{SUB}} = \frac{(b-a)^2}{12} < \frac{C^2}{12} < \lim_{N\to\infty}(N-2)^2\text{mse}, \tag{16}$$

which is a contradiction since we have found a suboptimal quantizer with a better mean-square error than the optimal one. □

Bennett [5] shows that the mean-square error of a uniform quantizer is approximately $\Delta^2/12$, assuming that the truncation error is negligible. This is not always the case and in the discussion we will give examples for which Bennett's approximation may be very poor indeed. There are some special cases where Bennett's approximation does hold. The next property deals with one such case.

Property 7: Suppose the density function is Riemann-integrable and $b - a < \infty$. Then for the optimal uniform quantizer we have

$$\lim_{N\to\infty}\frac{\text{mse}}{\Delta^2/12} = 1.$$

Proof: From Property 6 $\lim_{N\to\infty}(N-2)\Delta_0 = b - a < \infty$ where Δ_0 is the optimum Δ. We may design a suboptimum quantizer by dividing the interval (a,b) into $N-2$ equal subintervals and using these subintervals as the breakpoints for our quantizer. We denote the mean-square error associated with this quantizer by mse_{SUB} and the step size by $\Delta_S \triangleq (b-a)/(N-2)$. This quantizer has no truncation error. Hence from the Appen-

dix

$$\lim_{N\to\infty} \frac{\text{mse}_{\text{SUB}}}{\Delta_S^2/12} = 1. \tag{17}$$

Now

$$\lim_{N\to\infty} \frac{\Delta_S}{\Delta_0} = \lim_{N\to\infty} \frac{(N-2)\Delta_S}{(N-2)\Delta_0} = \frac{\lim_{N\to\infty}(N-2)\Delta_S}{\lim_{N\to\infty}(N-2)\Delta_0} = 1,$$

implying $\lim_{N\to\infty}\Delta_S^2/\Delta_0^2 = 1$. For any quantizer whose nontruncation region covers the support of the Riemann-integrable density function in the limit as N approaches infinity, we show in the Appendix that $\lim_{N\to\infty}\text{mse}/(\Delta^2/12) \geq 1$. This bound is arrived at by ignoring the truncation error and is true for density functions with finite or infinite support. Then

$$\lim_{N\to\infty} \frac{\text{mse}_{\text{SUB}}}{\Delta_0^2/12} = \lim_{N\to\infty} \left(\frac{\text{mse}_{\text{SUB}}}{\Delta_S^2/12}\right)\left(\frac{\Delta_S^2/12}{\Delta_0^2/12}\right)$$

$$= \left(\lim_{N\to\infty} \frac{\text{mse}_{\text{SUB}}}{\Delta_S^2/12}\right)\left(\lim_{N\to\infty} \frac{\Delta_S^2/12}{\Delta_0^2/12}\right) = 1, \tag{18}$$

but

$$1 = \lim_{N\to\infty} \frac{\text{mse}_{\text{SUB}}}{\Delta_0^2/12} \geq \lim_{N\to\infty} \frac{\text{mse}_{\text{OPTIMAL}}}{\Delta_0^2/12} \geq 1, \tag{19}$$

or

$$\lim_{N\to\infty} \frac{\text{mse}_{\text{OPTIMAL}}}{\Delta_0^2/12} = 1,$$

which is what we wanted to prove. □

In the above property we have shown that the truncation error is negligible for the optimum uniform quantizer, if the density function has finite support. This is not true, however, for arbitrary uniform quantizers on these densities. It is easy to design a sequence of uniform quantizers (indexed by N) such that $\lim_{N\to\infty}\text{mse} = 0$, $\lim_{N\to\infty}\Delta = 0$ but $\lim_{N\to\infty}\text{mse}/(\Delta^2/12) \neq 1$.

Zador [6] shows that if $f(x)$ is Riemann-integrable and $E\{x^{2+\delta}\} < \infty$ for some $\delta > 0$ then for the optimal nonuniform quantizer

$$\lim_{N\to\infty} N^2 \cdot \text{mse} = \|f\|_{1/3}/12$$

where $\|f\|_{1/3}$ is the $L_{1/3}$ norm. This result show that for the nonuniform quantizer the mean-square error decreases like $1/N^2$ for large N. Is there a similar property for the optimum uniform quantizer? Not always.

Property 8: Suppose $f(x)$ is Riemann-integrable. Then for the optimum uniform quantizer $\lim_{N\to\infty} N^2 \cdot \text{mse} = (b-a)^2/12$.

Proof: If $b - a < \infty$ then

$$1 \geq \lim_{N\to\infty} \frac{\text{mse}}{\Delta^2/12} = \lim_{N\to\infty} \frac{(N-2)^2\text{mse}}{(N-2)^2\Delta^2/12}$$

$$= \frac{\lim(N-2)^2\text{mse}}{\lim N^2\Delta^2/12} \tag{20}$$

but $(N-2)^2\Delta^2 \to \infty$ which implies $\lim_{N\to\infty}(N-2)^2\text{mse} \to \infty$.

If $b - a < \infty$ then $\lim_{N\to\infty}\text{mse}/(\Delta^2/12) = 1$ or $\lim_{N\to\infty}(N-2)^2\text{mse} = \lim_{N\to\infty} N^2\text{mse} = (12)^{-1} \cdot \lim(N-2)^2\Delta^2 = (b-a)^2/12$ which completes the proof. □

Discussion

We should note that not everyone uses our definition of the optimum uniform quantizer. For example, Pearlman and Senge [7] have published tables of the optimal uniform Rayleigh quantizer. For their computations they add the constraints $a = 0$ and $q = \Delta/2$.

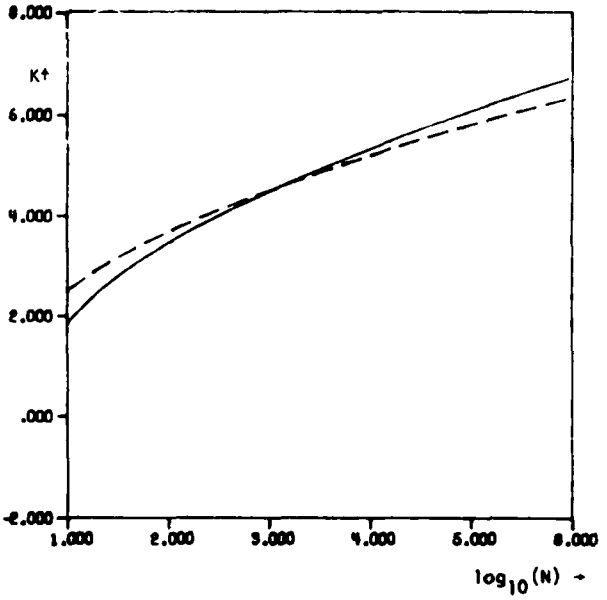


Fig. 1. K (solid line) and $D(N)$ (dashed line) plotted as a function of $\log_{10}(N)$.

It is interesting to note that Properties 1 and 3 are also shared by the optimal nonuniform quantizer as shown in [8]. As a further consequence of these two properties we find that, for the $N = 2$ case, the optimum uniform quantizer and the optimum nonuniform quantizer are identical.

Property 7 is one of the more interesting properties proved in this correspondence. A common approximation to the mean-square error of a uniform quantizer has been $\Delta^2/12$. Consider the class of density functions given by

$$f(x) = \frac{\left(1+\frac{\delta}{2}\right)}{(1+|x|)^{3+\delta}}, \qquad -\infty < x < \infty.$$

We easily see that $\delta = \text{Sup}\{\epsilon: \int x^{2+\epsilon} f(x)\,dx < \infty\}$. By straightforward minimization techniques one can show for this class of densities that

$$\lim_{N\to\infty} \frac{\text{mse}}{\Delta^2/12} = 1 + \frac{2}{\delta}.$$

Property 8 is of interest because it sets forth a basic difference between uniform and nonuniform quantizers. For the nonuniform quantizer we can expect the mean-square error to be of the order of $1/N^2$. We can expect this rate of convergence to zero to hold for the uniform quantizer only if the probability density has finite support. As an example consider the Gaussian case. The Gaussian probability density is of infinite support yet has extremely light tails. We may write down an expression for the mean-square error of a Gaussian random variable and solve for the optimum Δ for a specific N. Let us set $\Delta = 2\sigma K/(N-2)$ where K is a function of N and σ is the standard deviation. We find that, for large N, K is given by the following transcendental equation:

$$\frac{2K\sqrt{\frac{\pi}{2}}}{N-2}\left[\frac{erf\left(\frac{K}{\sqrt{2}}\right)}{6} + 2\left(\frac{N-1}{2}\right)^2 erfc\left(\frac{K}{\sqrt{2}}\right)\right]$$

$$= e^{-K^2/2}\left[N - 1 + \frac{K^2}{3(N-2)}\right].$$

This equation may be solved on a computer by a standard Newton-Raphson search. In Fig. 1 plot K as a function of N for values of N from 10 to 1000000. The dotted line is put in as a reference and is given by $D(N) = 1.7 \ln 36N/\pi$. It can be shown that $\lim_{N\to\infty} D(N)/K < \infty$. We conclude that the mean-square error in a uniform Gaussian quantizer is of the same or larger order than $(\ln N)/N^2$.

Appendix

Consider a sequence of quantizers $\{g_N(x)\}_{N=1}^{\infty}$, where N is the number of output levels, Δ_N is the step size, and I_N is the nontruncation region of $g_N(x)$. The measure of I_N is $(N-2)\Delta_N$. Suppose the input probability density function $f(x)$ is Riemann-integrable, and denote the support of $f(x)$ by supp f. Define $\text{mse}_N \triangleq E\{(x - g_N(x))^2\}$.

Lemma 1: Suppose $I_N \to \text{supp } f$ as $N \to \infty$ (i.e., if $x \in \text{supp } f$ then there exists an N_0 such that $x \in I_n$ for $n > N_0$) and $\lim_{N\to\infty} \Delta_n = 0$. Then $\lim_{N\to\infty} \text{mse}_N/(\Delta_N^2/12) \geq 1$. Furthermore if supp $f \subset I_N$ for all N and $\lim_{N\to\infty} \Delta_N = 0$ then $\lim_{N\to\infty} \text{mse}_N/(\Delta_N^2/12) = 1$.

Proof: Define

$$M_i \triangleq \sup_{x \in (q+i\Delta_N,\, q+(i+1)\Delta_N)} f(x)$$

$$m_i \triangleq \inf_{x \in (q+i\Delta_N,\, q+(i+1)\Delta_N)} f(x)$$

Then

$$\sum_{i=0}^{N-3} m_i \int_{q+i\Delta_N}^{q+(i+1)\Delta_N} \left(x - \left(q + \left(i + \frac{1}{2}\right)\Delta_N\right)\right)^2 dx \leq \text{mse}_N$$

and

$$\text{mse}_N \leq \sum_{i=0}^{N-3} M_i \int_{q+i\Delta_N}^{q+(i+1)\Delta_N} \left(x - \left(q + \left(i + \frac{1}{2}\right)\Delta_N\right)\right)^2 dx + \text{TE}_N$$

where TE_N is the truncation error. Thus

$$\frac{\Delta_N^2}{12} \sum_{i=0}^{N-3} m_i \Delta_N \leq \text{mse}_N \leq \frac{\Delta_N^2}{12} \sum_{i=0}^{N-3} M_i \Delta_N + \text{TE}_N.$$

If $I_N \to \text{supp } f$ as $N \to \infty$ and $\lim_{N\to\infty} \Delta_N = 0$ then, since $f(x)$ is Riemann-integrable, $\lim_{N\to\infty} \sum_{i=0}^{N-3} m_i \Delta_N \to 1$, which proves the first part of the lemma. If supp $f \subset I_N$ for every N then $\text{TE}_N = 0$ for every N, and since $\lim_{N\to\infty} \Delta_N = 0$ and $f(x)$ is Riemann-integrable, again $\sum_{N\to\infty}^{N-3} M_i \Delta_N \to 1$, which proves the second part of the lemma. □

References

[1] B. Widrow, "Statistical analysis of amplitude quantized sampled data systems," *Trans. AIEE Applications and Industry*, pt. II, vol. 79, pp. 555-568, Jan. 1960.

[2] H. Gish and J. N. Pierce, "Asymptotically efficient quantizing," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 676-683, Sept. 1968.

[3] J. M. Morris and V. D. Vandelinde, "Robust quantization of discrete-time signals with independent samples," *IEEE Trans. Commun.*, vol. COM-22, no. 12, pp. 1897-1901, Dec. 1974.

[4] A. B. Sripad and D. L. Snyder, "A necessary and sufficient condition for quantization errors to be uniform and white," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. ASSP-25, pp. 442-448, Oct. 1977.

[5] W. R. Bennett, "Spectra of quantized signals," *Bell Syst. Tech. J.*, vol. 27, pp. 446-472, 1948.

[6] P. Zador, "Development and evaluation of procedures for quantizing multivariate distributions," Ph.D. dissertation, Stanford Univ., Stanford, CA, 1964.

[7] W. A. Pearlman and G. H. Senge, "Optimal quantization of the Rayleigh probability distribution," *IEEE Trans. Commun.*, vol. COM-27, pp. 101-112, Jan. 1979.

[8] J. A. Bucklew and N. C. Gallagher, Jr., "A note on optimum quantization," *IEEE Trans. Inform. Theory*, vol. IT-25, pp. 365-366, May 1979.

SOME MODIFICATIONS TO THE MEDIAN FILTER PROCESS AND THEIR PROPERTIES†

T. A. NODES AND N. C. GALLAGHER, JR.
School of Electrical Engineering
Purdue University
W. Lafayette, IN 47907

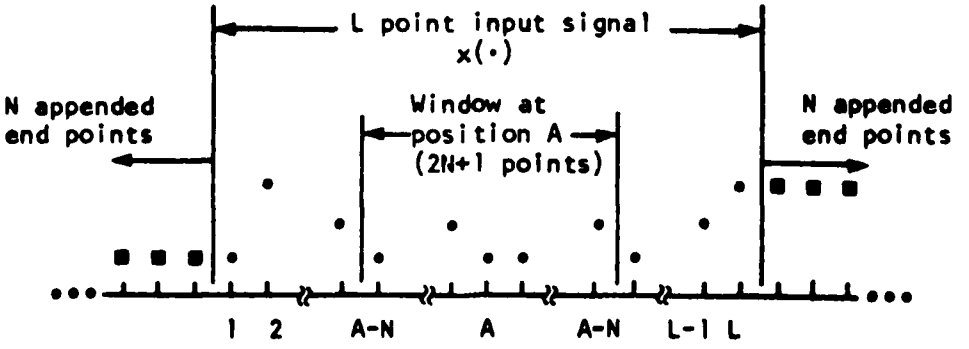
ABSTRACT

Some modifications of the median filter are given and their properties are derived. In addition, some results for standard median filters are given. It is shown that for non median nth ranked-order operations, repeated application of the operation will reduce any signal to a constant. Also, it is proved that the output of a recursive median filter is invarient to subsequence passes by the same filter.

I. INTRODUCTION

Median filtering, a method of signal processing which is easily implemented on a digital computer, has been used with success in many applications. These applications include picture processing and speech processing[1,2,3,4] where it is employed to smooth the signal. Further potentially useful properties can be obtained from slight modifications of the median process. We have investigated several such modifications and present the properties of two of them. In section II, we look at the nth ranked-order operation, which is a generalization of the median process. In section III, we study the recursive median operation, which incorporates previous output values into the median decision process. Finally, in section IV we introduce some other possible modifications to median filters. First, however, a review of the standard median filter is in order.

Median filtering is a discrete time process in which a 2N+1 points wide window is stepped across an input signal (see Fig. 1). At each step, the points inside the window are ranked according to their values, and the median value (mid-point) of the ranked set is taken as the output value of the filter for each window position. At both ends of the signal, N end points are appended to allow the filter to reach the edges of the signal.



The output of the median filter, Y(A) is given by

$$Y(A) = \text{the median value of } \{x(A-N),\ldots,x(A-1),x(A),x(A+1),\ldots,x(A+N)\}$$

Figure 1: The Median Filter

†The authors gratefully acknowledge the support of the Air Force Office of Scientific Research under Grant AFOSR 78-3605.

Presented at the Eighteenth Annual Allerton Conference on Communications, Control and Computing, October 8-10, 1980.

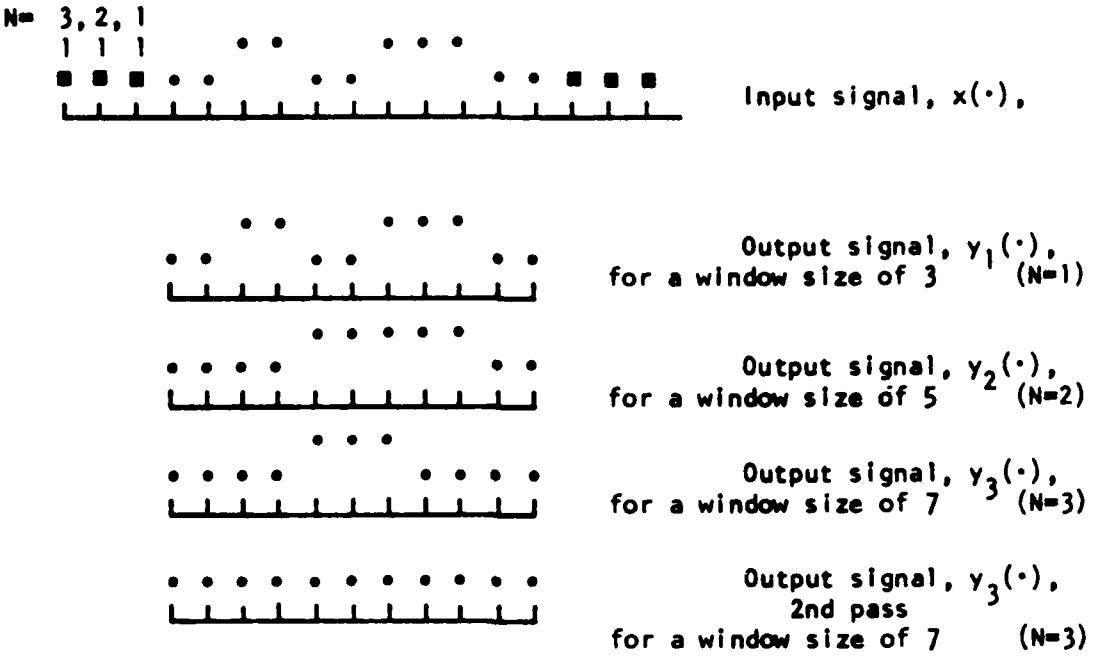


Figure 2: Effects of window size on a median filtered signal

The value of the front endpoints is equal to the value of the first point of the signal, and the value of the rear endpoints is equal to last point of the signal. As an example of this process, consider Fig. 2. Here, a binary signal of length eleven (the ■'s represent the appended endpoints) is median filtered by three different window widths $N = 1$ $(2N+1=3)$, $N = 2$ $(2N+1=5)$, and $N = 3$ $(2N+1=7)$. Notice, for the N=1 case, the signal is unperturbed, while for the N=2 and N=3 cases, the amount of structure in the signal is reduced. A number of signal structures which can be used to define the properties of median filters, can now be defined.

A constant neighborhood is a region of at least N+1 consecutive points all of which are identically valued.

An edge is a monotonically rising or falling set of points surrounded on both sides by constant neighborhoods.

An impulse is a set of N or less points whose values are different from the surrounding regions and whose surrounding regions are identically valued constant neighborhoods.

A root is a signal which is not modified by filtering.

Gallagher and Wise[5] have shown that, while impulses are eliminated by median filtering, constant neighborhoods and edges are unperturbed, and in fact, only signals composed solely of constant neighborhoods and edges are roots to the median filter. Again referring to Fig. 2, note that the signal is a root of the N=1 median filter but not for filters with N greater than one. However, after one pass of the N=2 filter or two passes of the N=3 filter the resulting outputs are roots of their respective filters. In fact, Gallagher and Wise have also proven that any signal of length L is reduced to its root after at most $\frac{1}{2}\cdot(L-2)$ successive passes by *any* median filter. Furthermore, any root of a median filter with a particular window size is also a root of any median filter with a smaller window size.

II. Nth RANKED-ORDER OPERATIONS

If instead of the median valued point the value of the nth largest point in the filter window is passed to the output at each step, then a general set of operations, called nth ranked-order operations, is found. More formally, the output of the nth ranked-order operation at position A is

$$Y(A) = \text{the nth largest value of } \{x(A-N),\ldots,x(A-1),x(A),x(A+1),\ldots,x(A+N)\}$$

This set of operations includes the median filter case, $n=N+1$, and many of the properties for all values of n are similar to the properties of the median filter. The non-median nth ranked-order operations have potential applications in areas such as peak detection with impulse rejection and digital A.M. detection (see Fig. 3).

The nth ranked-order operation can also be defined by the decision rule used to select the output value at each step. For 2N+1 points inside the window, the nth ranked point, x(a), is the point such that there are at least n points with values less than or equal to x(a) and at least $2N+1-(n-1)=2N+2-n$ points with values greater than or equal to x(a). A number of properties of the nth ranked-order operation can now be developed.

Property 1: A point, X(t), is unchanged (y(t) = x(t)) by an nth ranked-order operation if two conditions are met. The point, x(t), is located in a constant region, and x(t)'s position is restricted to $b+N-a \leq t \leq c-[|N+1-n|+a]$ where a is any nonnegative integer of value less than $N+1-|N+1-n|$ and b and c are the positions of the two endpoints of the constant region

Proof:

Assume that the two conditions given above are met. Now, let a = 0. The constant region must now extend to at least N points left (decreasing t) of x(t) and $|N+1-n|$ points right of x(t) for a total of at least $1+N+|N+1+n|$ points of value x(t) inside the window. Furthermore, if $a \neq 0$, then the constant region will extend 'a' fewer points to the left of x(t) but 'a' more points to the right, thus, maintaining a total of at least $N+1+|N+1+n|$ constant valued points inside the window. This means that if $N+1 \geq n$ then at least $1+N+|N+1-n| = 2N+2-n$ ($\geq n$) points inside the window have values equal

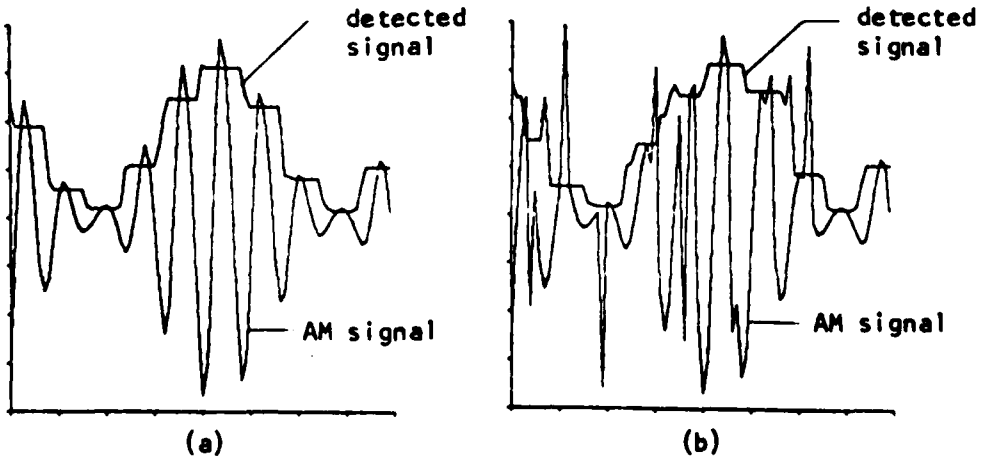


Figure 3: A.M. Detection of a 5KHz tone on a 31KHz carrier and sampled at 250KHz using an 8th ranked-order operation with a window size of 9

(a) original signal (b) signal corrupted with impulse noise

to x(t). Thus, x(t) meets the decision rule, and y(t) = x(t). Likewise, if N+1 < n, then $1+N+|N+1-n| = n(\geq 2N+2-n)$, and again y(t) = x(t).

Property 2: A rising impulse like signal of width less than 2N+2-n points or a falling impulse like signal of width less than n points will be eliminated.

Proof:

i) If a rising impulse has fewer than 2N+2-n points, then no point of the impulse can ever meet the second decision criterion. Thus, no output points will have values equal to the value of the impulse.

ii) Likewise, if a falling impulse has fewer than n points, then no point of the impulse can ever meet the first decision criterion, and no output points will have values equal to the value of the impulse.

The definitions previously given for the median case may now be generalized for all the nth ranked-order cases.

A constant neighborhood is a region of at least $N+1+|N+1-n|$ consecutive points all of which are identically valued.

An impulse is a set of points whose values are different from the surrounding regions and whose surrounding regions are identically valued constant neighborhoods. If the values of this set of points are greater than the surrounding neighborhoods, then the impulse contains less than 2N+2-n points, and if the values of the impulse are less than the surrounding regions, then the impulse contains less than n points.

The definitions for the edge and the root are unchanged. Note that, property 2 can be restated as "impulses are eliminated by nth order operations". Using these definitions, further properties can be developed. Due to lack of space, however, many of these properties are presented without proof.

Property 3: Upon each pass of an nth ranked-order operation, every edge of a signal will be moved to the left (advanced) by

$\text{sgn}[\text{edge}] \cdot (n-N-1)$ points

where $\text{sgn}[\text{edge}] = \begin{cases} +1 & \text{if } x(t) \leq x(t+1) \\ -1 & \text{if } x(t) \geq x(t+1) \end{cases}$ For t ranging over all positions in the edge

Property 4:

Any constant region of 2N+2-n or more points surrounded by constant neighborhoods of lesser values will be changed in width by $2 \cdot (n-N-1)$ points after being passed through an nth ranked-order operator.

Any constant region of n or more points surrounded by constant neighborhoods of greater values will after being operated on be changed in width by $2 \cdot (N+1-n)$ points.

As can be seen from the above properties, for n greater than N+1 the maximum valued signal segment (or the minimum if n is less than N+1) which is not an impulse tends to expand its coverage with each pass of a non-median operator. Thus, under repeated operations, a signal tends to be reduced to a constant. That this is true for any signal is shown in the following properties.

Property 5: Only constant signals are invariant to nth ranked-order operations if n is not equal to N+1.

Property 6: If n is not equal to N+1, then repeated passes of an nth ranked-order process will reduce any finite length signal to a constant.

The output of an nth ranked-order operation at position Z is not influenced by input points more than N points ahead (>Z+N) or N points behind (<Z-N) Z. This suggests a method by which long signals could be segmented and the ranked-order operations on each segment carried out in parallel.

i) Append the start and stop points as usual
ii) Divide the signal into overlapping segments. Each overlap is 2N+1 elements wide.
iii) Perform the normal nth ranked-order operation independently on each segement.
iv) After each operation replace the last N points of each segment (except the last segment) with the N+2 through the 2N+1 points of the following segment. Also, replace the first N element of each signal segment (except the first segment) with the elements from the 2N+1 through N+2 positions preceding the end of the prior signal segment.

Now, the signal is the same as it would be had the processing been done before the segmentation. Thus, further processing can now be done, or the segments can be recombined to form the final output signal.

A signal may be formed from independent identically distributed, iid, sample points of a random process. Such a signal would be formed if white noise were sampled to form the input signal. For this type of signal, results from order statistics[6] may be used to obtain the first order distribution, $F_y(\cdot)$, and the density, $f_y(\cdot)$, of the output of an nth ranked-order operation. If the distribution, $F_x(\cdot)$, and the density, $f_x(\cdot)$, of the input are known, then $f_y(\cdot)$ and $F_y(\cdot)$ are given by

$$f_y(x) = \frac{(2N+1)!}{(n-1)!1!(2N+1-n)!}\left[F_x^{n-1}(x)(1-F_x(x))^{(2N+1-n)}\, f_x(x)\right] \quad \text{property \#7}$$

$$F_y(x) = \sum_{K=n}^{2N+1} \frac{(2N+1)!}{K!(2N+1-K)!} F_x^K(x)\,(1-F_x(x))^{2N+1-K} \quad \text{property \#8}$$

where 2N+1 is the window size.

Kuhlman and Wise[7] will present further statistical analysis of the median filtering of independent identically distributed random processes in the next paper. However, the above formulas can immediately be used to prove that the statistical median of an iid process is preserved under standard median filtering.

Property 9: A median filter, $x(\cdot) \rightarrow y(\cdot)$; with an input of iid sample points will transform the distribution of the input, $F_x(\cdot) \rightarrow F_y(\cdot)$, symmetrically about 0.5. That is, for any ℓ such that $F_x(\ell) \rightarrow F_y(\ell)$, then $(1-F_x(\ell)) \rightarrow (1-F_y(\ell))$.

Property 10: The statistical median of a signal of iid sample points is preserved upon median filtering, or given ℓ such that $F_x(\ell) = 0.5$, then $F_y(\ell) = 0.5$.

Also recall that if the density of the input, $f_x(\cdot)$, is symmetric, then the

mean, $E_x\{\ \}$, and the median are equal. Therefore, by properties 9 and 10, the mean of an iid sample point signal whose density is symmetric is also preserved under median filtering. However, in general, the actual median point and the average of a particular signal will not be preserved.

Recursive Operations

Now consider replacing, at every step, the leftmost N points in the moving window with the previous N output points, and apply the same decision rule as was previously given for the nth ranked-order operation to obtain the next output value. This produces a recursive nth ranked-order operation which can be more formally stated as follows.

$Y(A)$ = the nth largest value of $\{Y(A-N),\ldots,Y(A-1),X(A),X(A+1),\ldots,X(A+N)\}$

Where X(A) and Y(A) are the values of the input and the output respectively at position A. The properties of these operations are similar to those of standard nth ranked-order operations. Most notably, they have the same set of roots.

Property 11: A signal is invariant to recursive filtering if and only if it is invariant to standard filtering.

Proof:

If a signal is invariant to an operation, $X(\cdot) \rightarrow Y(\cdot)$, then $X(k) = Y(k)$ for all k. Therefore, if a signal is invariant, then standard and recursive operations use the same points in the decision rule, and they must produce the same resulting signal.

However, the same signal will not in general reduce to the same root under recursive and standard operations. This is illustrated by an example for the median (n=N+1) filter case in Fig. 4. One may notice that under noisy conditions, the recursive filter tends to maintain a higher correlation between points in its output than does its non-recursive counterpart. This is further illustrated in figure 5 which compares the autocorrelation of the output for recursive and standard median filters with independant uniformly [0,1] distributed input points. These autocorrelation functions were obtained experimentally from a sequence of 2,200 random points. Thus, these filters may be useful in cases where more stringent filtering without a wider window is required.

One of the most interesting characteristics of the recursive operations

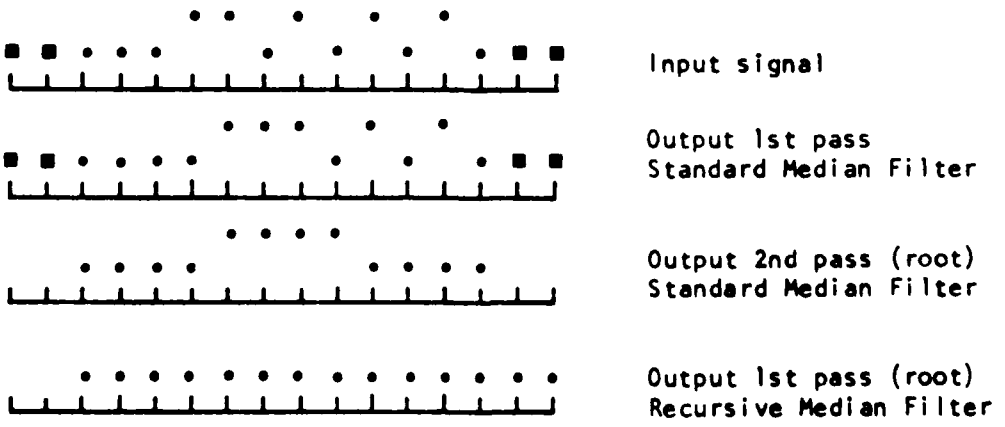


Figure 4: Recursive vs Standard Median filters with a window width of 5 (N=2)

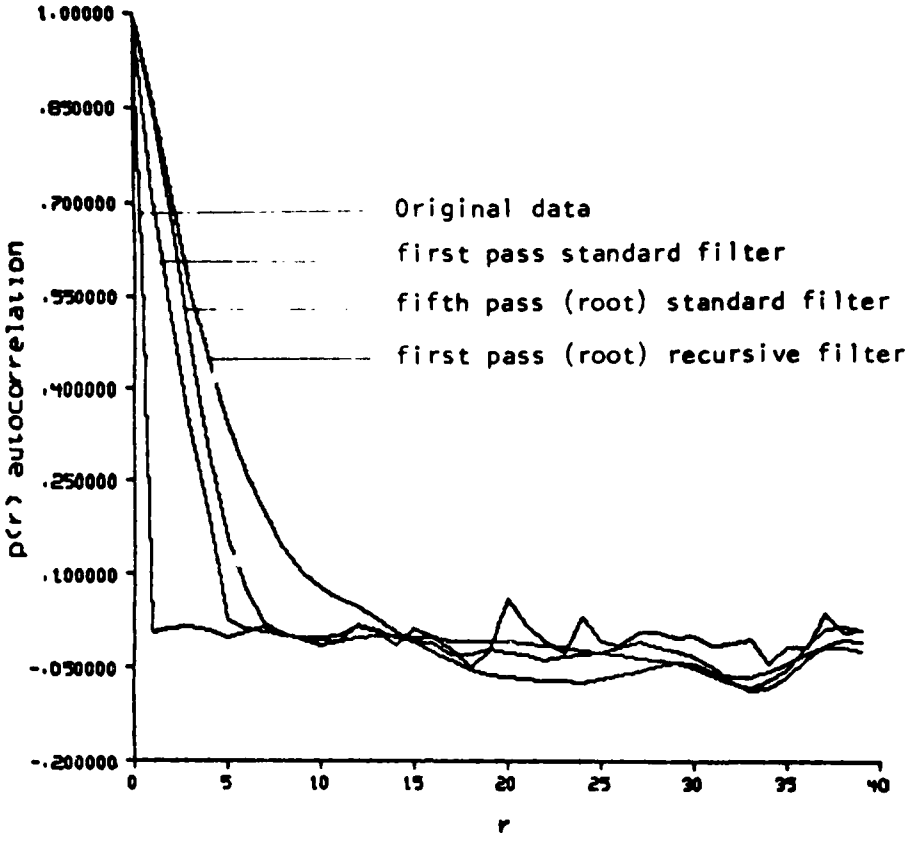


Figure 5: Autocorrelation function of standard and recursive median filters for a window width of five.

is that the root of a signal for a particular recursive process can always be found after the first pass of the operation. Recursive ranked-order operations are therefore potentially useful in areas, such as peak detection and coding operations, which require finding the root of a signal quickly. The following two properties prove this characteristic.

Property 12: Any signal will be reduced to a root after one pass of a recursive median filter (n=N+1).

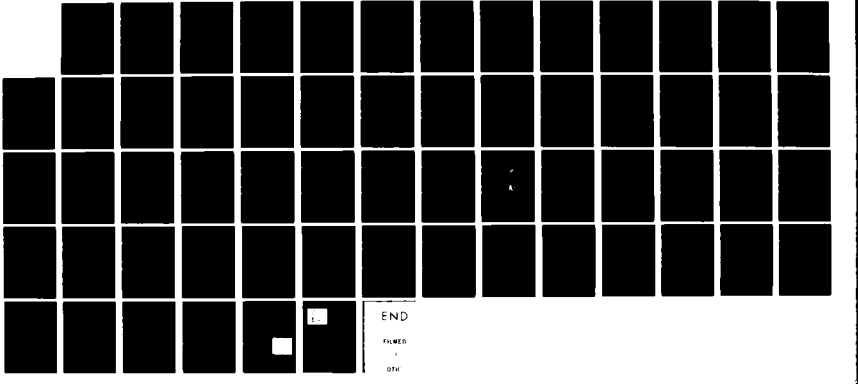
Property 13: If n≠N+1, then the last computed output value of a signal being operated on by a recursive nth ranked-order operation is the value of the signal root for that operator. For n > N+1 (n < N+1) this value is the value of the maximum (minimum) value to survive the first filter pass.

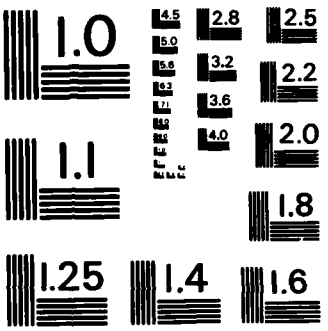
Other Functions

In addition to the above mentioned operations, many more variations of the median filter exist. Many of these other variations also have properties which may be useful in signal processing. We have studied several such modifications and present some of them here. Many of these modifications were obtained by defining a set of signal roots with certain desirable characteristics; then, we developed an operation which would have as many members of this set as possible for its own roots. Unfortunately, we have not, as yet, found a systematic method of determining an operation which will have any particular set of roots. Nevertheless, this approach does appear to hold promise.

AD-A121 294 THE ANALYSIS OF DESIGN OF ROBUST NONLINEAR ESTIMATORS AND ROBUST SIGNAL C..(U) PURDUE UNIV LAFAYETTE IN SCHOOL OF ELECTRICAL ENGINEERING N C GALLAGHER 2/2

UNCLASSIFIED 16 SEP 82 AFOSR-TR-82-0933 AFOSR-78-3605 F/G 12/1 NL





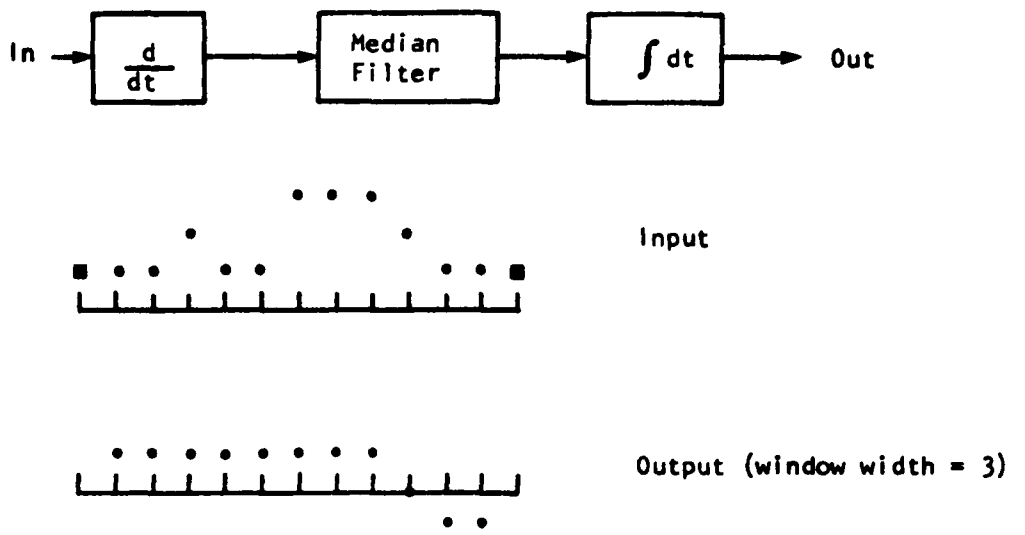


Figure 6: Example of a linear-median filter using a differentiator, integrator pair and a median filter with a window width of 3 (N=1)

One modification to the median filter, which Tukey [1] and Rabiner [2] have already utilized with promising results, is that of combining linear and median operations together. This allows one to greatly extend the number of available effects by utilizing some of the many linear operators whose properties are already well known. As an example of such an operation, consider figure 6. Here, a signal is differentiated, median filtered, and finally integrated. This operation has many of the same properties as a median filter alone. However, due to the differentiation, any slope of extent less than N+1 points will be seen by the median filter as an impulse and, thus, eliminated. Therefore, roots of this operation cannot contain sharp edges.

Another method of varying the median filter is to weight some positions of the window more heavily than others. This could be done by duplicating certain positions of the window. If the center position, for example, were to be weighted by three, then the output at position A would be given by

$$Y(A) = \text{the median value of } \{X(A-N),\ldots,X(A),X(A),X(A),\ldots,X(A+N)\}$$

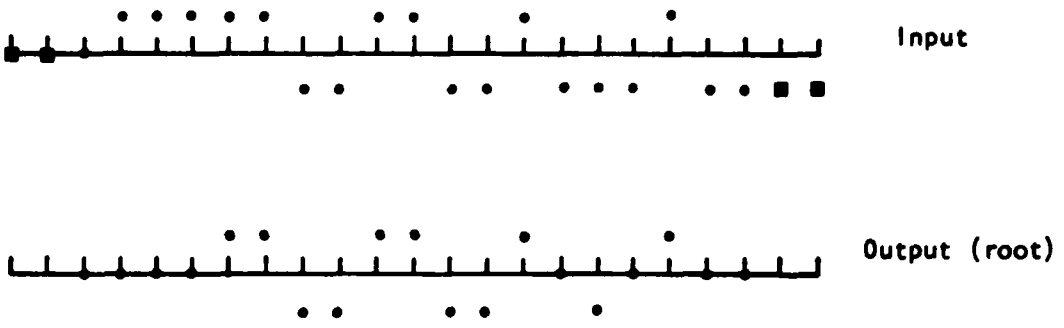
Yet, another modification would be to allow the value of a given position of the window to be a linear function of the points (possibly all of them) inside the window. Thus, the output at position A would be

$$Y(A) = \text{the median value of } \{f_1\{X(A-N),\ldots,X(A+N)\},\ldots,f_m\{X(A-N),\ldots,X(A+N)\}\}$$

where m is the number of values used in the decision process. A simple example combining the previous two modifications is given in figure 7. In this example, the points inside the window are first scaled by either -1, 0, or +1; then, the center position is weighted by three, and the median operation is carried out. The roots of this operation are zero or those segments of periodicity 4 ($X(i) = X(i \pm 4)$) which are symmetric about zero. Thus, with some modifications, median type filters can be designed for a wide range of different roots, including some periodic type signals.

Conclusion

In this paper, we have examined several variants of the median filter. We have found that the set of nth ranked-order operations is a generaliza-



$Y(A)$ = the median value of $\{-1 \cdot x(A-2), 0 \cdot x(A-1), x(A), x(A), x(A), 0 \cdot x(A+1), -1 \cdot x(A+2)\}$

Figure 7: Median filtering with a modified window function

tion of the median filter, and that they all have many similar characteristics. However, the non-median operators will, after repeated passes, reduce any signal to a constant. In contrast, the recursive median process retains the same set of roots as a standard median filter, though the same signal may not reduce to the same root under both operations. However, the recursive median filter reduces any signal to a root in just one pass, and thus, may be useful where high speed root determination is required. We have also reviewed some examples of other types of modified median operations, including combined linear, median functions and filters with modified windows.

REFERENCES

1. J. W. Tukey, "Nonlinear (Nonsuperposable) Methods for Smoothing Data," in Cong. Rec., 1974 EASCON, p. 673.

2. L. R. Rabiner, M. R. Sambur, and C. E. Schmidt, "Applications of a Nonlinear Smoothing Algorithm to Speech Processing," IEEE Trans. Acust., Speech, and Signal Processing, vol. ASSP-23, pp. 552-557, Dec. 1975.

3. N. S. Jayant, "Average and Median Based Smoothing Techniques for Improving Digital Speech Quality in the Presence of Transmission Errors," IEEE Trans. on Commun., vol. COM-24, pp. 1043-1045, Sept. 1976.

4. T. S. Huang, G. T. Yang, and G. Y. Tange, "A Fast Two-Dimensional Median Filtering Algorithm," IEEE Trans. Acoust., Speech, and Signal Processing, vol. ASSP-27, pp. 13-18, Feb. 1979.

5. N. C. Gallagher, Jr. and G. L. Wise, "Passband and Stepband Properties of Median Filters," Proceedings of the Princeton Conference on Information Sciences and Systems, March 1980.

6. H. A. David, Order Statistics, (1970), Wiley, New York.

7. F. Kuhlman and G.L. Wise, "On Spectral Characteristic of Median Filtered Independent Data", Allerton Conference on Communication, Control, and Computing, October 1980.

THE DESIGN OF MULTIDIMENSIONAL QUANTIZERS USING PREQUANTIZATION

Kerry D. Rines and Neal C. Gallagher, Jr.
School of Electrical Engineering
Purdue University
West Lafayette, Indiana 47907

ABSTRACT

A novel approach to the design of multidimensional quantizers is presented. This technique is used to design optimum uniform multidimensional quantizers that can be operated in real time. The quantizers are easily implemented using zero memory nonlinearities, linear transformations and univariate uniform step size quantizers.

I. INTRODUCTION

There is considerable interest in the use of multidimensional quantizers for the encoding of analog sources. Much of this interest has been generated from a theoretical standpoint. The multivariate quantization results of Zador [1] point to the advantages of multidimensional quantizers over univariate quantizers at high bit rates. Simply stated, the results indicate that the optimum per sample distortion decreases as the dimension of the quantizer increases. Therefore the potential exists to improve the performance of digital encoders by replacing univariate quantizers with multidimensional quantizers.

Recently the design of optimum multidimensional quantizers has been addressed. Computer algorithms for designing optimum quantizers of two or more dimensions have been presented by many authors, such as Linde et al [2]. The optimum quantizers are implemented using a search procedure to choose, from a specified output set, the output that is the smallest distance from the input. This implementation of the optimum quantizer may be difficult or impossible to operate in real time at high bit rates. In contrast the univariate uniform step size quantizer is a zero memory device that can be operated in real time. To date the easy implementation and real time operation of the univariate uniform step size quantizer has outweighed the theoretical advantages of using multidimensional quantizers in the design of digital encoders.

In this paper we present a novel approach to the design of multidimensional quantizers called prequantization. The design is illustrated in Figure 1 where a zero memory nonlinearity called a prequantizer precedes a specified multidimensional quantizer.

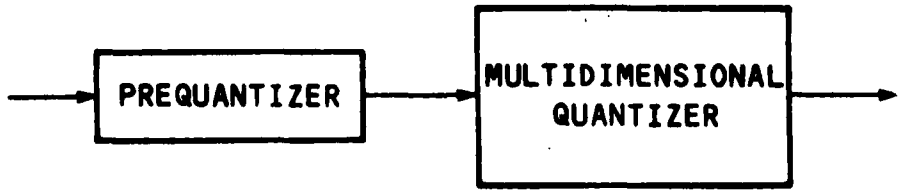


Figure 1. Multidimensional Quantizer Design using Prequantization.

Presented at the Eighteenth Annual Allerton Conference on Communications, Control and Computing, October 8-10, 1980.

This design is similar in some respects to the companding design of nonuniform univariate quantizers first proposed by Bennett [3]. In the univariate case a nonuniform quantizer may be difficult to implement directly. However, with companding we can design a nonuniform quantizer using a uniform step size quantizer, an invertible nonlinearity and the inverse nonlinearity. Similarly, prequantization can be used to design many multidimensional quantizers. Prequantization enables us to design these quantizers using a simple multidimensional quantizer, which is easy to implement and operate in real time, along with a zero memory nonlinearity. We illustrate the usefulness of prequantization with three examples.

In a recent paper Gersho [4] considers the partitioning of optimum uniform multidimensional quantizers. He states that the optimum uniform two-dimensional quantizer is the hexagonal quantizer. In three dimensions, Gersho argues that the truncated octahedral quantizer is very likely to be the optimum uniform three-dimensional quantizer. The analog of the truncated octahedron is considered for four dimensions. The resulting quantizer is not known to be optimal for four dimensions, but does have a lower per sample distortion than the three dimensional truncated octahedral quantizer. In this paper we present the designs for these three quantizers using prequantization. In each case the design is easy to implement and the quantizer can operate in real time. The real time operation of these quantizers for high bit rates is a significant result and demonstrates the important practical applications for prequantization. We begin in section II with a discussion of the prequantization design procedure.

II. PREQUANTIZATION

The design of k-dimensional quantizers using prequantization is illustrated in Figure 1. The design consists of a nonlinearity called a prequantizer preceding a specified k-dimensional quantizer. The implementation of this design approach takes place in two steps. First a k-dimensional quantizer meeting a specified criterion is chosen. In this paper we are interested in real time operation, therefore we specify that the quantizer be able to operate in real time. Examining Figure 1, we require that the real time (specified) quantizer have the same set of output values as the quantizer we wish to design. This is the only constraint placed on the choice of the real time quantizer. Free to choose from all quantizers satisfying the output constraint, we choose a real time quantizer that is easy to implement. The ability to exercise some control over the choice of the k-dimensional quantizer is one of the advantages of this design procedure.

The second step in the implementation is the design of the prequantizer. The role of the prequantizer is to complete the mapping of the input variables into the desired output values. The real time k-dimensional quantizer can be characterized by the mapping of its input space into its output values. This mapping is usually described by a partitioning of the input space, where all the input vectors contained within one partition are mapped into the same output vector. Since the real time quantizer is chosen based only on its output values, we do not expect its partitioning to be the same as the partitioning of the quantizer being designed. It is the prequantizer which is used to obtain the partitioning specified by the desired quantizer design. The prequantizing function maps a partition specified by the quantizer being designed into a partition of the real time quantizer that corresponds to the specified output. Once the prequantizing function is determined the k-dimensional

quantizer design is complete. We illustrate the design procedure with a simple example.

Consider the design of a univariate quantizer with input x and output $\hat{x}$ as described in (1).

$$\hat{x} = n\,\Delta \;\; ; \;\; n\,\Delta - \frac{\Delta}{4} \leq x < n\,\Delta + \frac{3\Delta}{4}. \tag{1}$$

Using the prequantization procedure, we first choose a quantizer that is easy to implement and has the same output set as given in (1). We choose the uniform step size quantizer given by

$$\hat{y} = n\,\Delta \;\; ; \;\; n\,\Delta - \frac{\Delta}{2} \leq y < n\,\Delta + \frac{\Delta}{2}. \tag{2}$$

We now determine the prequantizing function that must precede the quantizer in (2) to complete the design. Observe that quantizing $y = x - \frac{\Delta}{4}$ in (2) is identical to quantizing x in (1). Thus the prequantizing function is simply $f(x) = x - \frac{\Delta}{4}$ and the design of the quantizer in (1) is complete.

III. HEXAGONAL QUANTIZATION

Gersho has argued that the optimum uniform two-dimensional quantizer is the hexagonal quantizer. The design of a hexagonal quantizer using prequantizing is given here. First we attempt to find a two-dimensional quantizer that can be easily implemented and has the same set of output values as the hexagonal quantizer. One quantizer meeting these requirements is a scaled version of the diamond quantizer given below.

Let the inputs to the two-dimensional quantizer be x and y. The variables x and y are first encoded into two new variables w and z by the linear transformation,

$$\begin{aligned} w &= x + \sqrt{3}\,y \\ z &= x - \sqrt{3}\,y. \end{aligned} \tag{3}$$

The variables w and z are quantized separately by univariate quantizers with a uniform step size Δ. The outputs of the two-dimensional quantizer are then obtained using the linear transformation,

$$\begin{aligned} \hat{x} &= \frac{1}{2}(\hat{w} + \hat{z}) \\ \hat{y} &= \frac{1}{2\sqrt{3}}(\hat{w} - \hat{z}). \end{aligned} \tag{4}$$

The position of this quantizer in the hexagonal quantizer design is shown in Figure 2 and the partitioning of the scaled diamond quantizer is given in Figure 3. Having chosen the two-dimensional quantizer given in (3) and (4) we now turn to the design of the prequantizer.

The prequantizer must map the hexagonal region corresponding to each output into the scaled diamond shaped region corresponding to that same output. Consider the hexagonal partitioning shown in Figure 4.

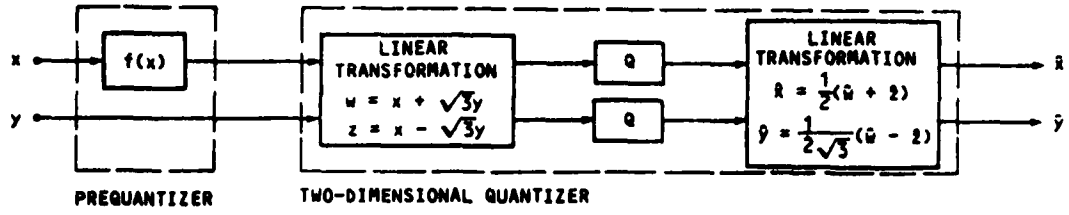


Figure 2. Prequantization design for the hexagonal quantizer. The quantizer Q has uniform step-size Δ.

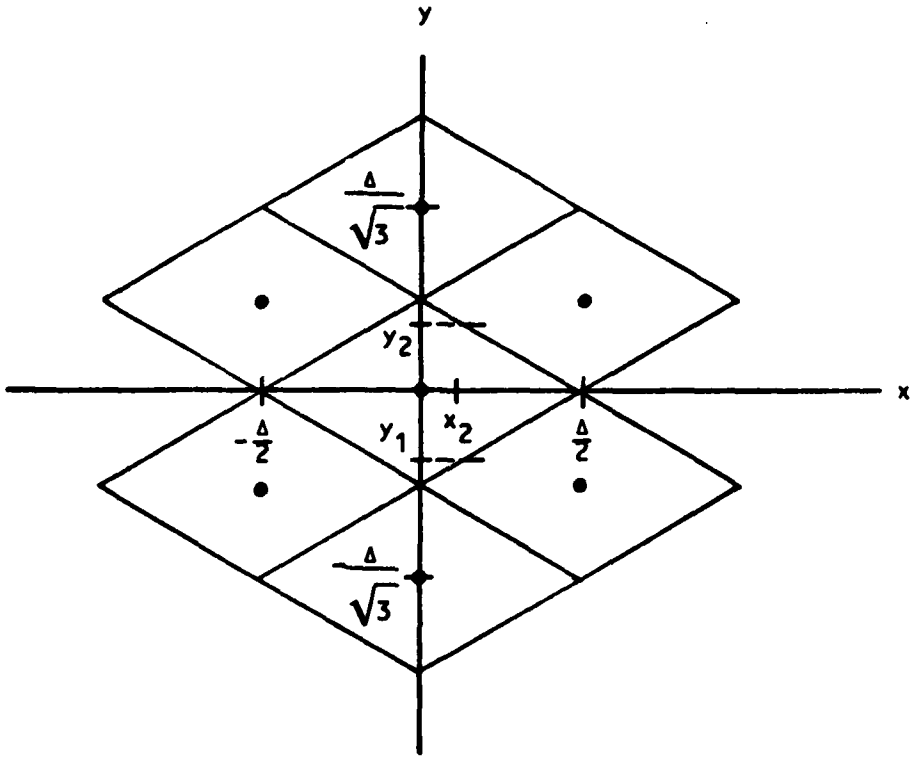


Figure 3. Partitioning of the scaled diamond quantizer.

Assume x is fixed and the pair (x,y) is contained within a given hexagonal partition. We now pose the question, does there exist a value x' such that the pair (x',y) is contained within the corresponding diamond partition for all values of y? This approach is illustrated with the following example. Let $x = x_1$ as shown in Figure 3 and let y be in the range $-\frac{\Delta}{2\sqrt{3}}$ to $\frac{\Delta}{2\sqrt{3}}$. In Figure 4 we observe that the hexagonal quantizer output will be (0,0) for all input pairs in the set $\{(x_1,y) : y_1 \leq y \leq y_2\}$. Similarly in Figure 3 we observe that the scaled diamond quantizer output will be (0,0) for all input pairs in the set $\{(x_2, y) : y_1 \leq y \leq y_2\}$. Therefore if $x_2 = f(x_1)$, the quantizer in Figure 2 will behave like the hexagonal quantizer for all input pairs in the set $\{(x_1,y) : \frac{-\Delta}{2\sqrt{3}} \leq y \leq \frac{\Delta}{2\sqrt{3}}\}$. In fact, we can show that the quantizer

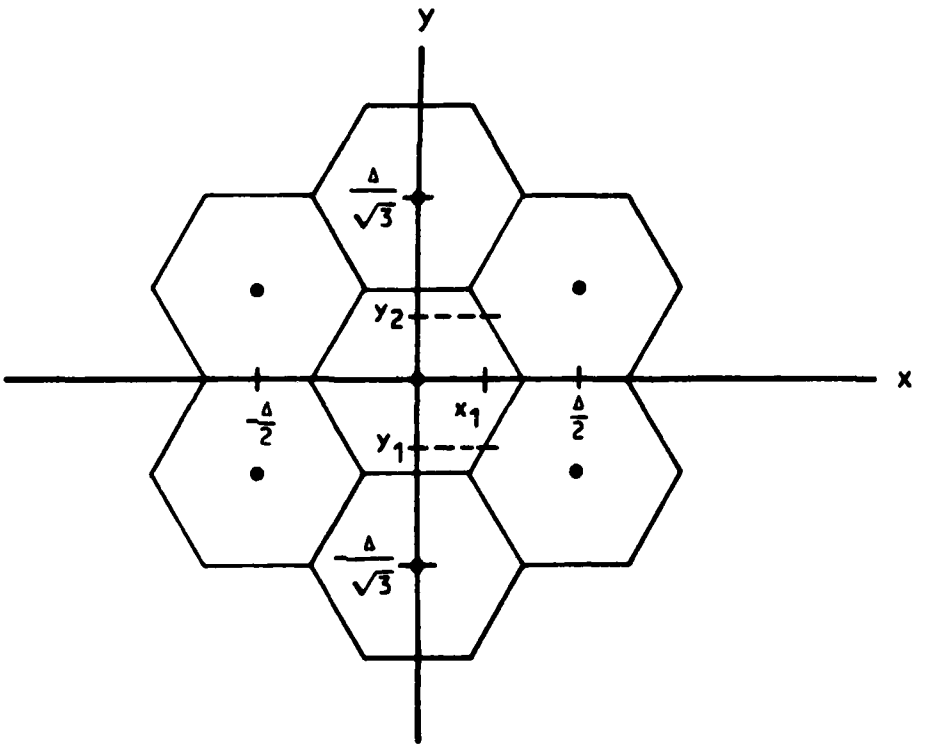


Figure 4. Partitioning of the hexagonal quantizer.

in Figure 2 behaves like the hexagonal quantizer for all inputs in the set $\{(x_1,y) : -\infty \leq y \leq \infty\}$ when $x_2 = f(x_1)$. Repeating this example for all possible values of x_1, we obtain a *prequantizing function* that maps the hexagonal region corresponding to each output into the scaled diamond shaped region corresponding to that same output. The prequantizing function is given in (5).

$$f(x) = n\frac{\Delta}{2} \qquad ; n\frac{\Delta}{2} - \frac{\Delta}{3} \leq x \leq n\frac{\Delta}{2} + \frac{\Delta}{3} \tag{5}$$

$$= 3x - (2n+1)\frac{\Delta}{2} \; ; n\frac{\Delta}{2} + \frac{\Delta}{3} \leq x \leq (n+1)\frac{\Delta}{2} - \frac{\Delta}{3}.$$

IV. RESULTS IN HIGHER DIMENSIONS

In this section we present the design of the optimum (or near optimum) uniform quantizers for three and four dimensions. Each of these quantizers use in their designs a two-dimensional quantizer termed the diamond quantizer. The algorithm for the diamond quantizer is as follows. Let the inputs to the two-dimensional quantizer be x and y. The variables x and y are first encoded into two new variables w and z by the linear transformation,

$$\begin{aligned} w &= x + y \\ z &= x - y. \end{aligned} \tag{6}$$

The variables w and z are quantized separately by univariate quantizers with a uniform step size Δ. The outputs of the diamond quantizer are then obtained from a linear transformation of the quantized variables $\hat{w}$ and $\hat{z}$ given by

$$\hat{x} = \frac{1}{2}(Q + \hat{z})$$

$$\hat{y} = \frac{1}{2}(Q - \hat{z}). \tag{7}$$

The outputs $\hat{x}$ and $\hat{y}$ will be multiples of $\frac{\Delta}{2}$ for all possible inputs. A useful property of the diamond quantizer is that if either input x or y is a multiple of $\frac{\Delta}{2}$, its quantized value $\hat{x}$ or $\hat{y}$ will be that same multiple of $\frac{\Delta}{2}$. Therefore if the output of one diamond quantizer $\hat{x}$ is used as the input to a second diamond quantizer, the output of the second diamond quantizer will also be $\hat{x}$. Using this property we are able to design quantizers of higher dimensions by cascading diamond quantizers. The results of these designs are now given.

Gersho states that the truncated octahedral quantizer is very likely the optimum three dimensional quantizer. This quantizer is defined by a tessellation of a truncated octahedron specified by the set $\{(x_1,x_2,x_3) : |x_1|+|x_2|+|x_3| < \frac{3\Delta}{2} ; |x_i| < \frac{\Delta}{2}, i=1,2,3 \}$. The design of this quantizer is given in Figure 5.

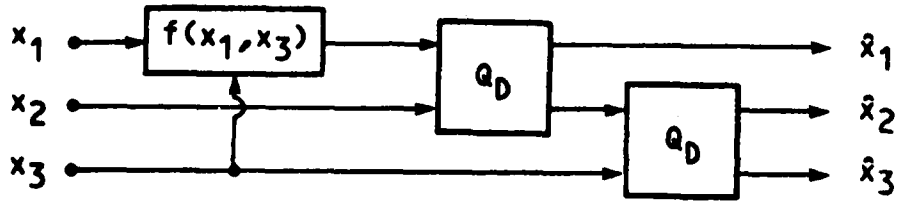


Figure 5. The truncated octahedral quantizer design using prequantization. Q_D is the diamond quantizer.

The prequantizing function is given in (8) where $e = |x_3| \bmod(0,\frac{\Delta}{2})$. For $e \leq \frac{\Delta}{4}$,

$$\begin{aligned} f(x_1,x_3) &= n\frac{\Delta}{2} && ; n\frac{\Delta}{2} - \frac{\Delta}{4} + e \leq x_1 \leq n\frac{\Delta}{2} + \frac{\Delta}{4} - e \\ &= x_1 - \frac{\Delta}{4} + e && ; n\frac{\Delta}{2} + \frac{\Delta}{4} - e \leq x_1 \leq (n+1)\frac{\Delta}{2} \\ &= x_1 + \frac{\Delta}{4} - e && ; (n-1)\frac{\Delta}{2} \leq x_1 \leq n\frac{\Delta}{2} - \frac{\Delta}{4} + e. \end{aligned} \tag{8}$$

A similar result is obtained for $\frac{\Delta}{4} \leq e \leq \frac{\Delta}{2}$.

The four dimensional analog of the truncated octahedral quantizer is defined by the tessellation of the polytope specified by the set $\{(x_1,x_2,x_3,x_4) : |x_1|+|x_2|+|x_3|+|x_4| < 2\Delta ; |x_i| \leq \frac{\Delta}{2}, i=1,2,3,4\}$. For convenience we will call this quantizer the 4-d uniform quantizer. The design of the 4-d uniform quantizer is shown in Figure 6.

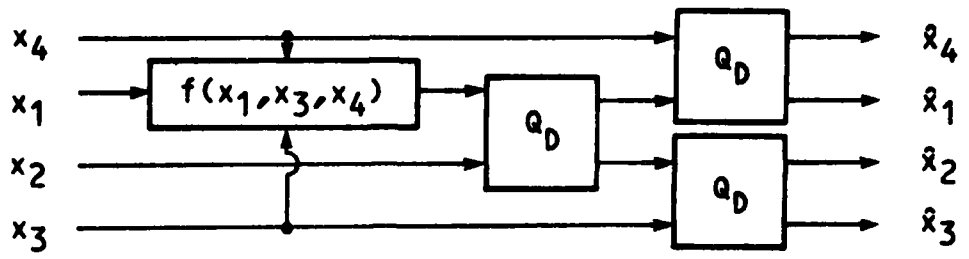


Figure 6. The 4-d uniform quantizer using prequantization. Q_D is the diamond quantizer.

The prequantizing function is given in (9) where $z = |x_3| \bmod(0, \frac{\Delta}{2})$, $w = |x_4| \bmod(0, \frac{\Delta}{2})$ and $e = z+w$. For $e \leq \frac{\Delta}{2}$,

$$\begin{aligned} f(x_1, x_3, x_4) &= n\,\frac{\Delta}{2} \quad ; (n-1)\,\frac{\Delta}{2} + e \leq x_1 \leq (n+1)\frac{\Delta}{2} - e \\ &= x_1 - \frac{\Delta}{2} + e \; ; (n+1)\,\frac{\Delta}{2} - e \leq x_1 \leq (n+1)\,\frac{\Delta}{2} \\ &= x_1 + \frac{\Delta}{2} - e \; ; (n-1)\,\frac{\Delta}{2} \leq x_1 \leq (n-1)\,\frac{\Delta}{2} + e. \end{aligned} \tag{9}$$

A similar result is obtained for $\frac{\Delta}{2} \leq e \leq \Delta$.

A comparison of the normalized mean-squared error performance of the uniform univariate and multidimensional quantizers is given in Table 1. The results were obtained by computer simulation using 30,000 samples uniformly distributed $(-\frac{1}{2}, \frac{1}{2})$. The output alphabet of each quantizer was assigned one hundred quantization levels per input sample.

Dimension	Quantizer	nmse ($\times 10^{-5}$)
1	uniform step-size	9.99
2	hexagonal	9.66
3	truncated octahedral	9.48
4	4-d uniform	9.17

V. DISCUSSION

In this paper we have presented a new approach to the design of multidimensional quantizers. The usefulness of the prequantization approach has been demonstrated by the design of three optimum (or near optimum) uniform multidimensional quantizers. In each example the quantizer can be implemented using a zero memory nonlinearity, linear transformations, and univariate uniform step-size quantizers. As a result the computation time of each quantizer is independent of the output alphabet size. Therefore, these quantizers are both easy to implement and are able to operate in real time even at very high bit rates.

The prequantization design approach is also compatible with the design of nonuniform multidimensional quantizers. In [4] Gersho generalizes the companding technique for the design of nonuniform univariate quantizers to the design of nonuniform multidimensional quantizers. Bucklew [5] shows that an optimum k-dimensional quantizer can be designed using an

optimum uniform k-dimensional quantizer, which is preceded by a multivariate invertible nonlinearity and followed by the inverse nonlinearity. Therefore the nonlinear prequantizing function used in optimum uniform k-dimensional quantizers is compatible and may even be of an advantage when the companding approach is applied to multidimensional quantizers.

ACKNOWLEDGMENT

The authors gratefully acknowledge the support of the Air Force Office of Scientific Research under grant AFOSR 78-3605.

REFERENCES

[1] P. Zador, Development and Evaluation of Procedures for Quantizing Multivariate Distributions, Ph.D. Dissertation, Stanford University, 1964, University Microfilm No. 64-9855.

[2] Y. Linde, A. Buzo, R.M. Gray, "An Algorithm for Vector Quantizer Design," IEEE Trans. Comm., Vol. COM-28, pp.84-95, January 1980.

[3] W.R. Bennett, "Spectra of Quantized Signals," B.S.T.J., Vol. 27, pp. 446-472, July, 1948.

[4] A. Gersho, "Asymptotically Optimal Block Quantization," IEEE Trans. on Inform. Theory, Vol. IT-25, pp. 373-380, July, 1979.

[5] J. A. Bucklew, "Companding and Random Quantization in Several Dimensions," to be published.

A novel approach for the computation of orthonormal polynomial expansions

Gary L. Wise (1), Neal C. Gallagher (2)

ABSTRACT

In this paper we present a novel technique for the computation of orthonormal polynomial expansions. The proposed method is very straightforward; given a function to be expanded in a polynomial series, we first use the FFT to compute a vector of Fourier coefficients. Then, using a change of basis transformation, we go from the Fourier coefficients to the polynomial coefficients. Convergence properties for this new approach are investigated.

1. INTRODUCTION

Two common ways of representing functions have been polynomial and trigonometric expansions. In much of science and engineering the trigonometric Fourier expansion has dominated over the generalized Fourier series expansions in applications. One advantage of the trigonometric series over the polynomial series is ease of coefficient computation by use of the fast-Fourier-transform (FFT) algorithm; compared to the FFT, coefficient computation for polynomial expansions can be cumbersome and time-consuming. In this paper we derive a simple change-of-basis transformation that maps a trigonometric series to a polynomial series.

These transformations have enabled us to develop an efficient algorithm for the computation of orthonormal polynomial expansions. The basic plan of these algorithms is to create a vector of Fourier coefficients by use of the FFT; this vector is then multiplied by a transformation matrix, resulting in a vector of polynomial coefficients. This approach can offer a saving in computation time over the standard integral formula for computing these polynomial coefficients. Section 2 contains the derivation of the elements of the transformation matrix, and in section 3 a numerical example is presented.

2. POLYNOMIAL EXPANSIONS

Assume that $H(x)$ is an $L_2[-T,T]$ function (where T is finite), and therefore possesses a Fourier series expansion convergent in $L_2[-T,T]$. Thus we may write

$$H(x) \approx \sum_{n=-\infty}^{\infty} h_n \exp\left(-\frac{in\pi x}{T}\right)$$

where

$$h_n = \frac{1}{2T} \int_{-T}^{T} H(x) \exp\left(\frac{in\pi x}{T}\right) dx.$$

We also assume that

$$\int_{-T}^{T} [H(x)]^2 w(x)\, dx < \infty,$$

where $w(x)$ is a nonnegative weight function integrable over $[-T,T]$. Let $\theta_n(x)$ denote an nth order polynomial, and assume that $\{\theta_n(x)\}_{n=0}^{\infty}$ is a set of polynomials that is orthonormal and complete in $L_2[-T,T]$ with respect to the weight function $w(x)$. Therefore, we can express $H(x)$ as

$$H(x) = \sum_{n=0}^{\infty} a_n \theta_n(x),$$

where

$$a_n = \int_{-T}^{T} H(x)\, \theta_n(x)\, w(x)\, dx. \qquad (1)$$

Define the truncated Fourier series as

$$H_M(x) = \sum_{|m| \le M} h_m \exp\left(-\frac{im\pi x}{T}\right).$$

Notice that

$$\left| a_n - \int_{-T}^{T} \sum_{|m| \le M} h_m \exp\left(-\frac{im\pi x}{T}\right) \theta_n(x)\, w(x)\, dx \right|^2$$

$$= \left| \int_{-T}^{T} [H(x) - H_M(x)]\, \theta_n(x)\, w(x)\, dx \right|^2$$

$$\le \int_{-T}^{T} [H(x) - H_M(x)]^2 w(x)\, dx \int_{-T}^{T} [\theta_n(y)]^2 w(y)\, dy.$$

(1) G. L. Wise, Department of Electrical Engineering, University of Texas at Austin, Austin, Texas 78712, USA.

(2) N. C. Gallagher, School of Electrical Engineering, Purdue University, West Lafayette, Indiana 47907, USA.

The integral with respect to y equals one by definition. Since the integral with respect to x is finite, we know that for any $\epsilon > 0$, there exists a K such that

$$\int_E [H(x) - H_M(x)]^2 w(x)\, dx < \epsilon,$$

where

$$E = \{ x : w(x) > K \},$$

and therefore

$$\int_{-T}^{T} [H(x) - H_M(x)]^2 w(x) dx < K \int_{-T}^{T} [H(x) - H_M(x)]^2 dx + \epsilon.$$

The first term can be made arbitrarily small by choosing M sufficiently large.
Thus, we see that

$$a_n = \sum_{m=-\infty}^{\infty} h_m c_{mn}, \tag{2}$$

where

$$c_{mn} = \int_{-T}^{T} \exp(-\frac{im\pi x}{T})\, \theta_n(x)\, w(x)\, dx,$$

and where the convergence is uniform in n. Consequently, (2) may be written as

$$a = h C \tag{3}$$

where h is the row vector of Fourier series coefficients, a is the row vector of polynomial coefficients, and C is the matrix whose mn-th element is c_{mn}.
After uniform sampling of the function H(x), we can compute the vector of polynomial coefficients in the following manner. In practice, a finite number of elements for h are computed by use of the FFT algorithms. Then we perform the vector multiplication indicated by (3). For example, h will be a $2M + 1$ dimensional row vector, a will be an L dimensional row vector, and C will be a $(2M + 1) \times L$ matrix. Because all computations must be performed using only a finite number of terms, we are concerned with the convergence of the resulting coefficients $a_n(2M+1)$ to the correct coefficients a_n given by (1). We see from the above derivation that this convergence is uniform in n, where we have neglected aliasing errors associated with the FFT and machine computation errors. In the remainder of this paper, it will be assumed that all computations are done with $2M+1$ such sample points of H(x).
Notice that if we take T = 1 and

$$w(x) = (1 - x)^{\alpha} (1 + x)^{\beta}, \tag{4}$$

where $\alpha > -1$ and $\beta > -1$, the resulting $\theta_n(x)$ are the normalized Jacobi polynomials given by

$$\theta_n(x) = \frac{P_n^{(\alpha,\beta)}(x)}{\sqrt{k_n}},$$

where [4, p. 284, #3.191–1]

$$P_0^{(\alpha,\beta)}(x) = 1,$$

$$k_0 = \frac{\Gamma(\alpha+1)\,\Gamma(\beta+1)\,2^{\alpha+\beta+1}}{\Gamma(\alpha+\beta+2)}$$

and for $n \geq 1$ [2, p. 169]

$$P_n^{(\alpha,\beta)}(x) = 2^{-n} \sum_{m=0}^{n} \begin{bmatrix} n+\alpha \\ m \end{bmatrix} \begin{bmatrix} n+\beta \\ n-m \end{bmatrix} (x-1)^{n-m} (x+1)^m$$

and

$$k_n = \frac{2^{\alpha+\beta+1}\,\Gamma(n+\alpha+1)\,\Gamma(n+\beta+1)}{(2n+\alpha+\beta+1)\,\Gamma(n+1)\,\Gamma(n+\alpha+\beta+1)}.$$

In this case the elements c_{mn} of the transformation matrix C may be calculated using a method suggested by Yao and Thomas [5] (there is an error in equation (32) in [5]). Utilizing this method we obtain

$$c_{mn} = 2\pi\, \phi_n(-m\pi),$$

where

$$\phi_n(t) = D(n,\alpha,\beta)\, t^{\frac{-\alpha-\beta-2}{2}}\, M_{\frac{\alpha-\beta}{2},\frac{2n+\alpha+\beta+1}{2}}(2it), \tag{5}$$

$M_{r,s}(t)$ is the Whittaker function [1, p. 264] given by

$$M_{r,s}(t) = e^{-t/2}\, t^{2s+1}\, {}_1F_1(\tfrac{1}{2} - r + s;\, 2s + 1;\, t),$$

and

$$D(n,\alpha,\beta) = \frac{\Gamma(n+\alpha+1)\,\Gamma(n+\beta+1)\,2^{\frac{\alpha+\beta}{2}}}{2\pi\sqrt{k_n}\,\Gamma(n+1)\,\Gamma(2n+\alpha+\beta+2)\,(i)^{\frac{\alpha+\beta+2}{2}}}.$$

For $\alpha = \beta$, we obtain the normalized Gegenbauer polynomials, and in this case (5) becomes

$$\phi_n(t) = \frac{(i)^n \sqrt{\pi}\,\Gamma(n+\beta+1)\,\Gamma(n+\beta+3/2)\,2^{2n+2\beta+3/2} J_{n+\beta+1/2}(t)}{2\pi\sqrt{k_n}\,\Gamma(2n+2\beta+2)\,\Gamma(n+1)\,t^{\beta+1/2}}. \tag{6}$$

Some special classes of normalized Gegenbauer polynomials are the normalized Legendre polynomials, both kinds of Chebyshev polynomials, and Tesseral polynomials. Applications of the above method for Legendre polynomials may be found in [3].

3. AN EXAMPLE

In this section we present an example of the above method using Chebyshev polynomials of the first kind. Let T = 1 and let $\alpha = \beta = -1/2$ in (4).
This results in $\theta_n(x)$ being the normalized nth Chebyshev polynomial of the first kind. The Chebyshev polynomials of the first kind can be defined by

$$T_{n+1}(x) = 2x\, T_n(x) - T_{n-1}(x)$$

$$T_0(x) = 1 \tag{7}$$

$$T_1(x) = x.$$

The resulting normalized polynomials are given by

$$\theta_0(x) = \frac{1}{\sqrt{\pi}}$$

$$\theta_n(x) = \sqrt{\frac{2}{\pi}}\, T_n(x), \quad n \geq 1.$$

The elements of the transformation matrix are found to be

$$c_{mn} = \sqrt{2\pi}\ (i)^n J_n(-m\pi).$$

Therefore, we have that

$$a_n = \sqrt{2\pi} \sum_{m=-\infty}^{\infty} h_m (i)^n J_n(-m\pi).$$

We consider the special case where the function H(x) is real valued. Using the relations

$$h_{-m} = h_m^*$$

and

$$J_n(-m\pi) = (-1)^n J_n(m\pi),$$

we have

$$a_0 = \sqrt{2\pi}\, h_0 + 2\sqrt{2\pi} \sum_{m=1}^{\infty} RE(h_m) J_0(m\pi)$$

$$a_n = 2\sqrt{2\pi}\,(-1)^{n/2} \sum_{m=1}^{\infty} RE(h_m) J_n(m\pi), \quad n \text{ even}, \ n \neq 0$$

$$a_n = 2\sqrt{2\pi}\,(-1)^{\frac{n-1}{2}} \sum_{m=1}^{\infty} IM(h_m) J_n(m\pi), \quad n \text{ odd}. \tag{8}$$

We now present an example of the computation of the Chebyshev polynomial coefficients. The function H(x) is

$$H(x) = \theta_1(x) + \theta_2(x)$$

$$= \sqrt{\frac{2}{\pi}}\,(2x^2 + x - 1).$$

The Chebyshev coefficients are computed by use of (8); selected coefficients a_n are found in tables 1 and 2 for the cases $N = 4096$ and $N = 8192$, respectively, where N is the number of equally spaced samples used in the FFT.

TABLE 1. Selected values for $\{a_n\}$ with M = 50, 75, and 100; N = 4096.

M	50	75	100	True value
a_0	-5.38×10^{-3}	-5.93×10^{-3}	-6.57×10^{-3}	0
a_1	0.886	0.907	0.919	1
a_2	0.995	0.994	0.993	1
a_3	-0.113	-9.27×10^{-2}	-8.05×10^{-2}	0
a_4	-5.19×10^{-3}	-5.76×10^{-3}	-6.42×10^{-3}	0
a_5	-0.111	-9.17×10^{-2}	-7.98×10^{-2}	0
a_{49}	-9.28×10^{-3}	6.22×10^{-3}	1.42×10^{-2}	0

TABLE 2. Selected values for $\{a_n\}$ with M = 50, 75, and 100; N = 8192.

M	50	75	100	True value
a_0	-3.17×10^{-3}	-3.23×10^{-3}	-3.46×10^{-3}	0
a_1	0.886	0.907	0.919	1
a_2	0.997	0.997	0.997	1
a_3	-0.113	-9.27×10^{-2}	-8.05×10^{-3}	0
a_4	-3.09×10^{-3}	-3.15×10^{-3}	-3.38×10^{-3}	0
a_5	-0.111	-9.16×10^{-2}	-7.98×10^{-2}	0
a_{49}	-9.29×10^{-3}	6.22×10^{-3}	1.42×10^{-2}	0

4. DISCUSSION

We have proposed in this paper a novel approach for computing polynomial expansions from equally spaced samples. The computation involved in this procedure falls into three categories :

(1) Compute the transformation matrix C; this computation need be done once and the result stored in computer memory. The same matrix C is used for the expansion of all functions;

(2) Given the function H(x) to be expanded into the polynomial series, compute the Fourier series expansion of H(x) by use of the FFT. This provides a vector of Fourier coefficients;

(3) Finally, multiply this vector by the matrix C to produce a vector of polynomial coefficients.

The major sources of computation error with this procedure are error in the FFT, and truncation error in matrix multiplication (finite - rather than infinite - vectors and matrix); these errors can be reduced by choosing larger values of N in the FFT and M in the matrix multiplication. If great accuracy is required, then large values for M and N may be required.

In examining the computation time required to evaluate polynomial coefficients we will ignore the computation of the transformation matrix C. If this matrix is recomputed each time a different function is expanded in polynomials, then the computation time for C must be considered. For our purposes, we assume that C is stored in memory. The matrix multiplication requires 2M + 1 multiplications and 2M additions for each coefficient; if L coefficients are computed, we then have a total of L(2M + 1) multiplications. In many cases, the equations will simplify as in (8). The FFT routine for computation of the Fourier coefficients requires $(N/2)\log_2(N)$ multiplications (for radix 2 FFT).

As a comparison to the approach proposed herein, consider the computations necessary to evaluate the integral of (1). First, we partition the interval for numerical evaluation of the integral. We then use a recursion relation such as that in (7) to generate values of $\theta_n(x)$ for the chosen partition points. Next a numerical evaluation procedure such as the trapezoidal rule is used to evaluate the integral. If the error in the evalua-

tion is not small enough, the procedure is repeated with a finer partitioning. It may be necessary to iterate several times. This general computation procedure is necessary for each coefficient; hence, if we want a total of L coefficients, we must evaluate L integrals in this manner. The actual computer time taken in evaluating coefficients in this manner varies greatly from one set of computer code to another. One may argue advantages for either technique of coefficient computation; it is possible for direct integral evaluation to take less time than the FFT procedure provided a fortuitous partitioning is made; however, we have found the FFT-matrix multiplication technique to be particularly simple and efficient. For comparison purposes consider the example of section 3. We evaluated the coefficients a_0, a_1, a_2, a_3, a_4, a_5, and a_{49} using the trapezoidal rule and Simson's rule, where we took the interval to be $[-0.99999, 0.99999]$. In table 3 we used 601 points and in table 4 we used 1201 points. Notice that since seven coefficients are being evaluated, these correspond respectively to 4207 and 8407 samples, and tables 1 and 2 correspond respectively to 4096 and 8192 samples.

TABLE 3. Selected values for $\{a_n\}$ using 601 samples.

Coefficient	Trapezoidal rule	Simpson's rule	True value
a_0	0.177	0.282	0
a_1	1.23	1.40	1
a_2	1.25	1.40	1
a_3	0.230	0.399	0
a_4	0.251	0.399	0
a_5	0.231	0.399	0
a_{49}	0.289	0.458	0

TABLE 4. Selected values for $\{a_n\}$ using 1201 samples.

Coefficient	Trapezoidal rule	Simpson's rule	True value
a_0	6.85×10^{-2}	0.130	0
a_1	1.11	1.18	1
a_2	1.10	1.18	1
a_3	0.112	0.183	0
a_4	9.72×10^{-2}	0.183	0
a_5	0.112	0.184	0
a_{49}	0.136	0.206	0

5. ACKNOWLEDGEMENT

This research was supported by the Air Force Office of Scientific Research, Air Force Systems Command, USAF, under Grants AFOSR-76-3062, AFOSR-78-3605, and AFOSR-81-0047, and also by the National Science Foundation under Grant ENG-76-82426. Appreciation is expressed to F. Kuhlmann for assistance with the programming.

REFERENCES

1. ERDÉLYI A., MAGNUS W., OBERHETTINGER F. and TRICOMI F. G. : *Higher transcendental functions*, vol. 1, McGraw-Hill, New York, 1953.
2. ERDÉLYI A., MAGNUS W., OBERHETTINGER F. and TRICOMI F. G. : *Higher transcendental functions*, vol. 2, McGraw-Hill, New York, 1953.
3. GALLAGHER N. C., WISE G. L. and ALLEN J. W. : *A novel approach for the computation of Legendre polynomial expansions*, IEEE Trans. Acoustics, Speech, and Signal Processing, ASSP-26 (1978), pp. 105-106.
4. GRADSHTEYN I. S. and RYZHIK I. W. : *Table of integrals, series, and products*, Academic Press, New York, 1965.
5. YAO K. and THOMAS J. B. : *On band-limited properties of Fourier transform pairs of some special functions*, Proc. 3rd Ann. Allerton Conf. on Circuit and System Theory, Monticello, IL, Oct. 1965, pp. 299-309.

Some Results on the Median Filtering of Signals and Additive White Noise†

T. A Nodes and N. C. Gallagher, Jr.

School of Electrical Engineering
Purdue University
West Lafayette, Indiana 47907

Abstract

The first order distribution of the output of a median filter when filtering a known signal plus additive white noise has been derived and is presented along with some examples. In addition, two programs have been written to aid in the design of median filters for the additive white impulse noise case and some of these results are tabulated.

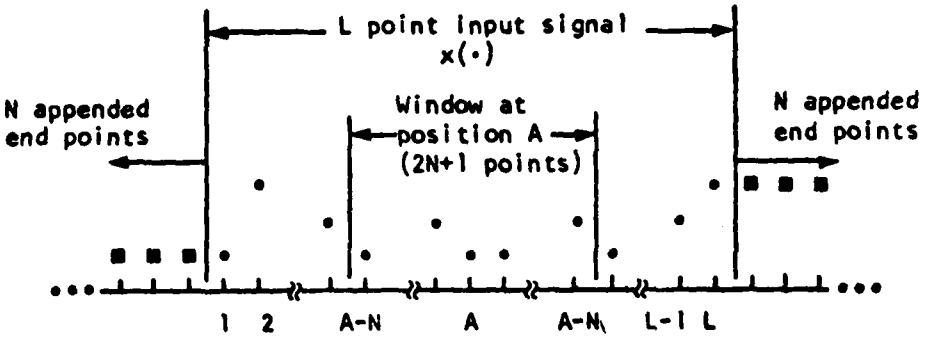
I. Introduction

Median filtering, a method of signal processing which is easily implemented on a digital computer, has been used with success in many applications. These applications include picture processing and speech processing[1,2,3,4] where it is employed to smooth the signal. Previous work in developing the properties of the median filter has been limited to the filtering of deterministic[5] and white noise[8] (i.i.d.) signals. Unfortunately, due to the nonlinearity of the median process, the analysis of the important signal plus additive noise case is not a direct extension of these simpler cases. In this paper, we present some results on the filtering of signals plus additive white noise. Specifically, we have derived the first order output distribution for an arbitrary given signal and noise distribution. This along with several examples is presented in the second part of the paper. In addition, we present some results on the effects of additive impulse noise on median filtered signals. First, however, a review of the standard median filter is in order.

Median filtering is a discrete time process in which a 2N+1 points wide window is stepped across an input signal (see Fig. 1). At each step, the points inside the window are ranked according to their values, and the median value (mid-point) of the ranked set is taken as the output value of the filter for each window position. At both ends of the signal, N end points are appended to allow the filter to reach the edges of the signal. The value of the front endpoints is equal to the value of the first point of the signal, and the value of the rear endpoints is equal to last point of the signal. As an example of this process, consider Fig. 2. Here, a binary signal of length eleven (the ■'s represent the appended endpoints) is median filtered by three different window widths N = 1 (2N+1=3), N = 2 (2N+1=5), and N = 3 (2N+1=7). Notice, for the N=1 case, the signal is unperturbed, while for the N=2 and N=3 cases, the amount of structure in the signal is reduced. A number of signal structures which can be used to define the properties of median filters can now be defined.

†The authors gratefully acknowledge the support of the Air Force Office of Scientific Research under grant AFOSR 783605.

Presented at the Eighteenth Annual Allerton Conference on Communications, Control and Computing, September 30 - October 2, 1981.



The output of the median filter, Y(A) is given by

Y(A) = the median value of {x(a-N),...,x(A-1),x(A),x(A+1),...,x(A+N)}

Fig.1: The Median Filter

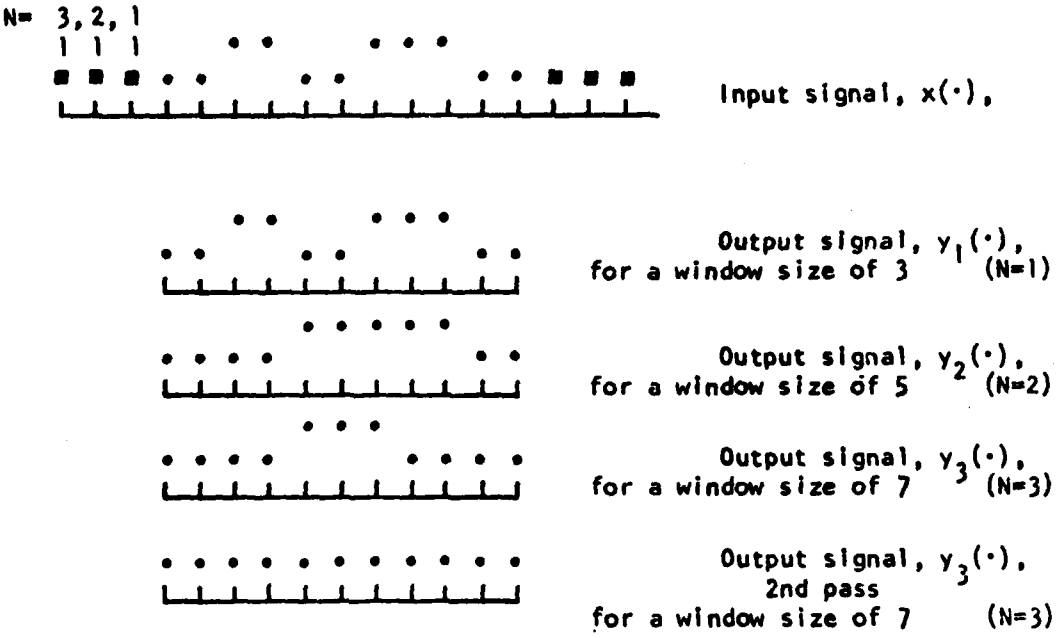


Fig.2: Effects of window size on a median filtered signal

A constant neighborhood is a region of at least N+1 consecutive points all of which are identically valued.

An edge is a monotonically rising or falling set of points surrounded on both sides by constant neighborhoods.

An impulse is a set of N or less points whose values are different from the surrounding regions and whose surrounding regions are identically valued constant neighborhoods.

A root is a signal which is not modified by filtering.

Gallagher and Wise [5,6] have shown that, while impulses are eliminated by median filtering, constant neighborhoods and edges are unper-

turbed, and in fact, only signals composed solely of constant neighborhoods and edges are roots to the median filter. Again referring to Fig. 2, note that the signal is a root of the N=1 median filter but not for filters with N greater than one. However, after one pass of the N=2 filter or two passes of the N=3 filter the resulting outputs are roots of their respective filters. In fact, Gallagher and Wise have also proven that any signal of length L is reduced to its root after at most $\frac{1}{2}\cdot(L-2)$ successive passes by any median filter. Furthermore, any root of a median filter with a particular window size is also a root of any median filter with a smaller window size.

For i.i.d. (white) random signals, Kulman and Wise[8] have derived the second order statistics of the median filter. They further show that for all the distributions which they have investigated, which include most of the common ones, the median filter has a low pass effect on the signal spectrum, and thus increases the correlation. In fact, this is also often true with more general signals; however, due to the nonlinear nature of the filter there are cases where the second moment bandwidth of a signal is actually increased upon median filtering and thus the correlation decreased. Thus, one must use some care in applying the low pass assumption to median filters.

II. Output Distribution

Section I reviewed much of the previous work on properties of median filtered deterministic and i.i.d. signals. As stated earlier, the more general case of filtering signals plus additive noise is much more difficult to analyze. In this section, the first order distribution of the output of a median filter with a known signal and additive white noise input is given. This is used in program Dis to compute some statistics of the output of the median filter several examples of which are given.

If the output of median filter at position m has a distribution of $F_y(q,m)$ and the input a distribution of $F_x(q,i) = F_{noise}(q - s_i)$ where s_i = signal at position i, then the output distribution is

$$F_y(q,m)=\sum_{k=1}^{N+m-k}\ \sum_{a_k=a_{k-1}+1}^{N+m}\left\{\left[\prod_{i\in c_{1,\cdots,a_k}}[1-F_x(q,i)]\right]\times\left[\prod_{i\in\bar{a}_{1,\cdots,a_k}}(F_x(q,i))\right]\right\}+\prod_{i=m-N}^{m+N}F_x(q,i)$$

where

$2\cdot N + 1$ = window width

$\{a_1,a_2,\cdots,a_k\} \cup \{a'_1,a_2,\cdots,a_k\} = \{1,2,\cdots,(2\cdot N+1)\}$

$$\sum_{i=a}^{b} f(i) = \begin{cases} f(a) + \cdots + f(b) & \text{if } a \le b \\ 1 & \text{if } a > b \end{cases}$$

This result comes about from combining all possible combinations of the points inside the window such that at least N+1 of them have values $\le q$. It is straightforward to extend this result to obtain the first order output distribution for any arbitrary input (any arbitrary random process) if the (2N+1)th order distribution is known at every position, however, this result is somewhat cumbersome and is not presented here.

The above equation was incorporated into program Dis to compute the value of the first order median filter output distribution, $F_y(q,m)$, for a signal plus white noise input. This is then used to numerically evaluate some of the statistical properties of the output at each position m. Specifically, Dis computes the value of the mean, E $\langle Y \rangle$, the standard deviation, σ_y, the mean square error, M.S.E. ($= E\langle(y_i - s_i)^2\rangle$), and the absolute error, A.E. ($= E\langle|y_i - s_i|\rangle$) at every position. These terms may then be plotted as in Fig. 4 through Fig. 7 or averaged over the signal and tabulated as in Tables 1 and 2. These examples illustrate some of the effects of median filtering signals plus noise. Two different distributions (impulsive and gaussian) both with the same two noise power levels are used in these examples. The impulse noise used is double sided symmetric with heights of ± 3 and probabilities $P^+ = P^- = 0.001$ and 0.05 for noise powers of $\sigma_n^2 = 0.018$ and 0.90 respectively. Likewise, the gaussian noise powers are also $\sigma_n^2 = 0.018$ and 0.90. For comparison, results are also given for windowed averaging filters.

First, consider a constant signal. The results from a constant signal indicate the effects that the noise distribution by itself has on the filter output. The results for several such cases using averaging and median filters are given in Table 1. It can be seen that the Average Filter does somewhat better than the median filter when filtering gaussian noise. This is expected since for a set window width the Average Filter is the optimum M.S.E. estimator in this case. However, when impulse noise is present the median filter reduces the output noise power by orders of magnitude more than the Average filter. This is due to the ability of the median filter to totally eliminate low probability high power impulses which is not possible with linear systems. In fact, it can be shown that for a fixed window width the median filter is the optimum MAP estimator in this case. In general, for constant signals median filters have been found to out-perform averaging type filters when the tails of the additive noise density are extensive[9] compared to the gaussian case. Also certain types of general signals are particu-

Table 1: Mean square error of median and average filter outputs with constant signal plus noise inputs. Window width=2N+1

Input Additive Noise	Average Filter			Median Filter		
	n=1	n=3	n=5	n=1	n=3	n=5
Impulse						
σ_{in}^2=0.018	6.000E-3	2.573E-3	1.637E-3	5.396E-5	6.285E-10	8.2612E-15
σ_{in}^2=0.90	3.000E-1	1.286E-1	8.182E-2	1.305E-1	3.484E-3	1.044E-4
Gaussian						
σ_{in}^2=0.018	6.000E-3	2.573E-3	1.637E-3	8.909E-3	4.610E-3	3.215E-3
σ_{in}^2=0.900	3.000E-1	1.286E-1	8.182E-2	4.046E-1	1.902E-1	1.243E-1

larly suitable for median filtering irregardless of the noise distribution.

As pointed out earlier, many systems generate signals which are not amenable to the general spectrum separation techniques that ease the design of linear filters. Often this is due to the presence of sharp edges in an otherwise low frequency signal. Such structures tend to be roots to median filters making the median filter a good alternative for smoothing such signals. One such signal is used here to illustrate the effects of median filtering these signals when additive white noise is present. This signal ranges from -2 to 2 and consists of edges and constant neighborhoods. Figures 3 through 6 plot the filter output expected value and standard deviation ($E\{Y_i\}$, $E\{Y_i\} + \sigma_{Y_i}$ and $E\{Y_i\} - \sigma_{Y_i}$) at each position as solid lines and the original uncorrupted signal as a dashed line. For comparison, the results for a windowed average filter are shown in Fig. 3. As with the median filters, the window width = $2 \cdot N + 1$.

As illustrated above, the median filter does an excellent job of eliminating impulses (see also Table 3). However, with non-constant signal structures, other types of errors become prevalent when impulse noise is present. Foremost among these is edge jitter. This effect is present even at low noise levels and is not reduced by using larger windows as illustrated in Fig. 4a and Fig. 5a. This effect will be further discussed in Section III. Fig. 5 also shows the effects of filtering with larger windows. The final peak of the signal is only five points wide instead of the six (= N + 1) necessary to pass through an N=5 median filter unperturbed. Fig. 5 also illustrates another error form which occurs when the width of a plateau or valley approaches N+1 points. One or two impulses of the correct sign located within such a plateau will cause the whole plateau to drop to the closest point below it, which can be a substantial change.

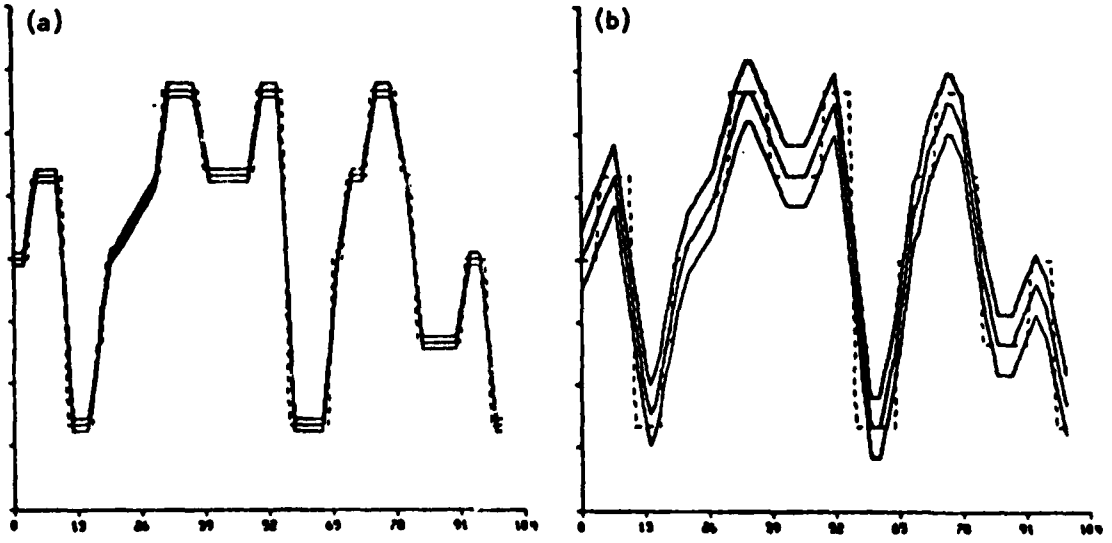


Fig. 3: For output, y, of an averaging filter, with input signal plus noise the E{y}(——), the E{y} + σy(——), the E{y} - σy(——), and the input signal (----) are plotted for a) N=1 and PN=0.018 and b) N=3 and PN=0.90 where PN=input noise power and the window width=2 N+1

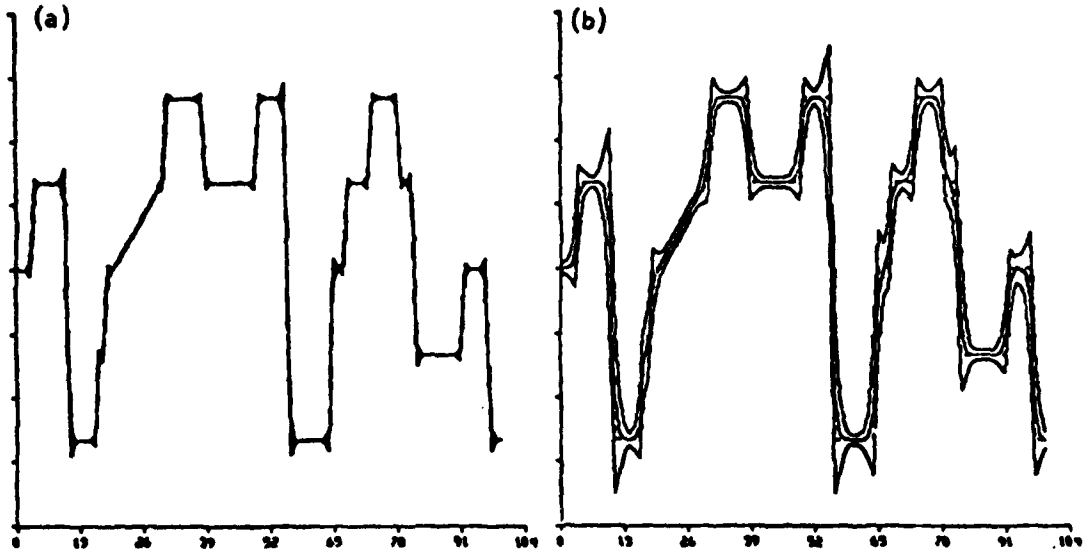


Fig. 4: For output, y, of a median filter with an input signal plus impulsive noise, the E{y}(——), the E{y} + σy(——), the E{y} - σy(——), and the input signal (----) are plotted for a) N=3 and PN=0.018(P+=P-=0.001 and Height(Imp.)=±3) and b) N=3 and PN=0.90(P+=P-=0.05 and Height(Imp.)=±3) where PN=input noise power and the window width=2·N+1=7

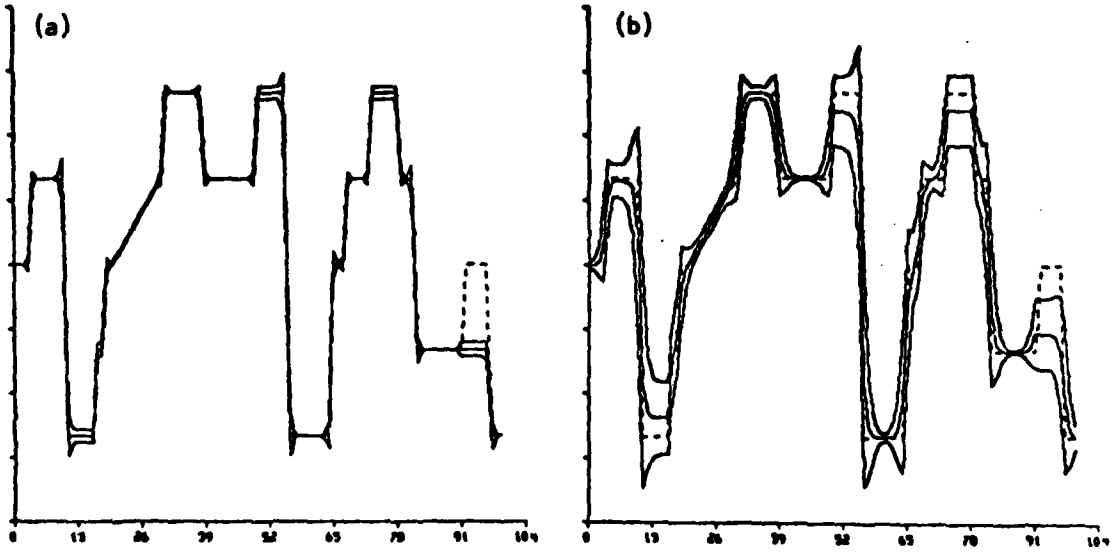


Fig. 5: For output, y, of a median filter with an input signal plus impulsive noise, the E{y}(——), the E{y} + σy(——), the E{y} - σy(——),and the input signal (----) are plotted for a) N=5 and PN=0.018(P+=P-=0.001 and Height(Imp.)=±3) and b) N=5 and PN=0.90(P+=P-=0.05 and Height(Imp.)=±3) where PN=input noise power and the window width=2·N+1=11

Conversely, when gaussian noise is present, quite different results are obtained. As can be seen from Fig. 6, in this case the Std. Dev. of the output is much more smooth and constant, than with impulse noise, and the plots more closely resemble the results of the Average Filter much more closely than before. This is further illustrated in Fig. 7 which plots the density of the output of the N=3 median filter at position 34 (as reviewed in Fig. 6). Notice that while it is shifted and the Std. Dev. reduced, it is still fairly smooth, symmetrical, and bell shaped (although the tails do exhibit some assymmetry which is unobserv-

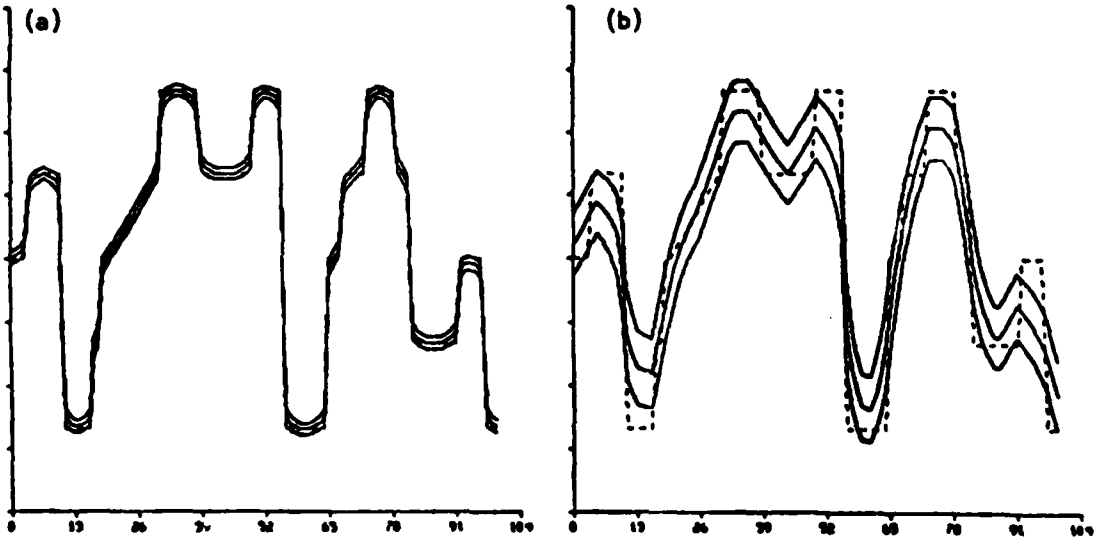


Fig. 6: For output, y, of a median filter with an input signal plus gaussian noise, the E{y}(——), the E{y} + σy(——), the E{y} - σy(——), and the input signal (----) are plotted for a) N=3 and PN=0.018 and b) N=5 and PN=0.90 where PN=input noise power and the window width=2·N+1=11

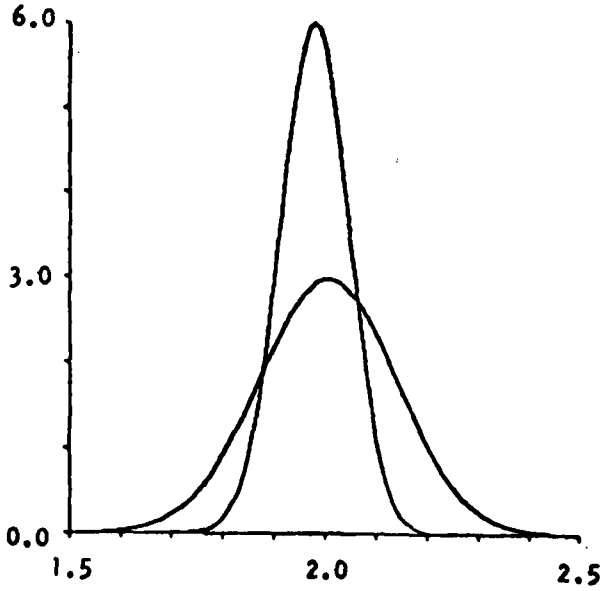


Fig. 7: The output density (upper curve) of an N=3 median filter with an input of signal plus gaussian noise (density: lower curve) with PN=0.018 at position 34 (see Fig. 6)

able in Fig. 7). This is due to the fact that gaussian noise perturbs almost all the input points by a small amount rather than just a few by a large amount as is the case with impulse noise. However, the median filter tracks the signal more closely than the Average filter does. A summary of the average M.S.E.s for the above filter is given in Table 2.

Table 2: Average mean square error, $\overline{M.S.E.}$, of filter output filtering a 100 point signal and additive noise (see Fig. 3 through Fig. 7)

Input Additive Noise	Average Filt.		Median Filt.	
	n=1	n=3	n=3	n=5
Impulse				
σ^2=0.018	1.125E-1	2.514E-1	3.202E-3	5.590E-2
σ^2=0.90	4.043E-1	3.774E-1	1.502E-1	2.640E-1
Gaussian				
σ^2=0.018	1.125E-1	2.514E-1	1.333E-2	———
σ^2=0.900	4.043E-1	3.774E-1	3.553E-1	4.280E-1

III. Impulse Noise

The special case of signal plus white impulsive noise is of particular interest as the median filter appears to perform especially well in reducing this type of noise. As pointed out above and in Table 3, this is due to the fact that the probability of an impulse being transferred to the output of a median filter is small. And while this is the predominate error form for constant signals, when more signal structure is added other types of errors take over. The problem of edge jitter appears to be particularly significant. This was shown in Section two with output standard deviation plots and can be qualitatively explained as follows. As pointed out above N+1 impulses of the same sign must be inside the filter window in order for the output to assume the value of an impulse. However, if a signal edge is being filtered, then an edge point, x(t) can be shifted by $j < N+1$ positions, $Y(t+j) = x(t)$, by the simple presence of j impulses of the correct signs within N+1 positions of t. Narrow ($\leq$ N+1 positions wide) plateaus and valleys are also susceptible to impulses; however, these structures are much less common than edges in most signals.

The distribution of edge jitter, $j(y(t+j) = x(t))$ has been derived. The equations, however, are rather untractable and do not lead to any particular insight into the process; thus, they will not be presented here. The distribution was incorporated into program Edg which was used to compile Tables 4 and 5. Table 4 lists the standard deviation of the edge jitter, j, for a number of different window sizes (window width = 2N + 1) and double sided impulse probabilities. Table 3 should be used in conjunction with Table 4 since the possibility of an impulse at the output is not incorporated into the standard deviation computation. Note if the edge has only two states, then, as seen in Table 4, the mean square error contributed by each edge is approximately doubled by increasing N from 1 to 5 if the probability of impulse is $P_+ = P_- = 0.05$, and this ratio decreases with decreasing $P_+ = P_-$. This

Table 3: Prob. of an impulse at the output of a median filter

p+=p-= Prob. of + or - impulse at the input vs. Prob. of impulse at the output with wind. =2n+1

p+=p-	n=1	n=2	n=3	n=4	n=5	n=6
1.00e-05	5.99990e-10	1.99994e-14	6.99969e-19	2.51985e-23	9.23933e-28	3.43167e-32
1.00e-04	5.99900e-08	1.99940e-11	6.99692e-15	2.51853e-18	9.23327e-22	3.42900e-25
1.00e-03	5.99000e-06	1.99401e-08	6.96926e-11	2.50535e-13	9.17296e-16	3.40212e-18
2.00e-03	2.39200e-05	1.59043e-07	1.11019e-09	7.97054e-12	5.82814e-14	4.31684e-16
5.00e-03	1.48750e-04	2.46282e-06	4.27994e-08	7.64906e-10	1.39222e-11	2.56679e-13
1.00e-02	5.90002e-04	1.94100e-05	6.69956e-07	2.37774e-08	8.59369e-10	3.14599e-11
2.00e-02	2.32007e-03	1.50713e-04	1.02619e-05	7.18246e-07	5.11873e-08	3.69481e-09
5.00e-02	1.37569e-02	2.15378e-03	3.52202e-04	5.91515e-05	1.01136e-05	1.75157e-06
1.00e-01	5.02024e-02	1.48470e-02	4.55957e-03	1.43575e-03	4.60250e-04	1.49544e-04
2.00e-01	1.65473e-01	8.78393e-02	4.80858e-02	2.69196e-02	1.53382e-02	8.86622e-03

Table 4: Std. deviation of edge jitter of median filtered signals with additive Imp. noise

p+=p-= Prob. of + or - impulse vs. Std. Dev. with wind. =2n+1

p+=p-	n=1	n=2	n=3	n=4	n=5	n=6
1.00e-05	6.32446e-03	7.74608e-03	8.94441e-03	1.00002e-02	1.09546e-02	1.18323e-02
1.00e-04	1.99970e-02	2.44986e-02	2.82885e-02	3.16275e-02	3.46462e-02	3.74222e-02
1.00e-03	6.31507e-02	7.75748e-02	8.95770e-02	1.00150e-01	1.09709e-01	1.18499e-01
2.00e-03	8.91745e-02	1.09867e-01	1.26871e-01	1.41846e-01	1.55385e-01	1.67835e-01
5.00e-03	1.40362e-01	1.74446e-01	2.01505e-01	2.25292e-01	2.46795e-01	2.66569e-01
1.00e-02	1.97008e-01	2.48301e-01	2.87106e-01	3.21015e-01	3.51655e-01	3.79831e-01
2.00e-02	2.74400e-01	3.55038e-01	4.12025e-01	4.60868e-01	5.04877e-01	5.45333e-01
5.00e-02	4.14095e-01	5.72890e-01	6.77901e-01	7.61845e-01	8.35481e-01	9.02651e-01
1.00e-01	5.40000e-01	8.12157e-01	1.00338e+00	1.15004e+00	1.27171e+00	1.37858e+00
2.00e-01	6.40000e-01	1.07016e+00	1.43122e+00	1.73486e+00	1.99263e+00	2.21482e+00

Table 5: Amount of jitter with 90% certainity (Prob(jitter>k)<0.1) for Sig. and Imp. noise

p+=p-= prob. of + or - impulse vs. k (jitter limit with 90% certainty) with wind. =2n+1

p+=p-	n=1	n=2	n=3	n=4	n=5	n=6
1.00e-05	0	0	0	0	0	0
1.00e-04	0	0	0	0	0	0
1.00e-03	0	0	0	0	0	0
2.00e-03	0	0	0	0	0	0
5.00e-03	0	0	0	0	0	0
1.00e-02	0	0	0	0	1	1
2.00e-02	0	1	1	1	1	1
5.00e-02	1	1	1	1	1	1
1.00e-01	1	1	2	2	2	2
2.00e-01	Impulse	2	3	3	3	4

increase, then, must be reconciled with a corresponding decrease in the probability of an impulse at the output by a factor of 1,300 for the same parameters as above. Further information can be obtained from Table 5 which gives the amount of jitter with 90% certainty. The listing "impulse" in this table indicates that the probability of the output assuming the value of an impulse is greater than 10%. The use of these tables in conjunction with the deterministic properties developed by Gallagher and Wise[11] should greatly facilitate the design of median filters used in filtering signals with additive impulse noise as they help to quantify the various trade offs available in such designs.

IV. Conclusion

The first order distribution of the output of a median filter for a signal plus white noise input was presented. Using this, the statistics of several examples with impulsive and gaussian noise were computed and given. These illustrate some of the properties of median filtering. Edge jitter and narrow plateau jitter are seen to be the dominate error modes for impulse noise. For gaussian additive noise, the output more closely resembles that of an average filter but with a larger standard deviation and closer tracking of edges. For the additive impulse noise case, some statistical properties of the edge jitter is tabulated. These results should aid in the design of median filters since they illustrate many of the properties the designer can expect from these filters in the important signal plus white noise case. However, much more work needs to be done in this area to develop easier to use and more general descriptions of the properties of the filter while retaining some quantitative ability.

References

1. J. W. Tukey, "Nonlinear (Nonsuperposable) Methods for Smoothing Data," in Cong. Rec., 1974 EASCON, p. 673.
2. T. S. Huang, G. T. Yang, and G. Y. Tang, "A Fast Two-Dimensional Median Filtering Algorithm," IEEE Trans. Accoust., Speech, and Signal Processing, vol. ASSP-27, pp. 13-18, Feb. 1979.
3. L. R. Rabiner, M. R. Sambur, and C. E. Schmidt, "Applications of a Nonlinear Smoothing Algorithm to Speech Processing," IEEE Trans. Accoust., Speech, and Signal Processing, Vol. ASSP-23, pp. 552-557, Dec. 1975.
4. N. S. Jayant, "Average and Median Based Smoothing Techniques for Improving Digital Speech Quality in the Presence of Transmission Errors," IEEE Trans. on Commun., Vol. COM-24, pp. 1043-1045, Sept. 1976.
5. N. C. Gallagher, Jr. and G.L. Wise, "Passband and Stepband Properties of Median Filters," IEEE Trans. Accoust. Spch. Sig. Proc., to be published.
6. S-G. Tyan, It has come to our attention that S-G. Tyan has also proved a number of these properties. We have not seen the proofs and can only speculate as to their form.
7. H. A. David, Order Statistics, (1970), Wiley, New York.
8. F. Kuhlmann and G. L. Wise, "On Spectral Characteristic of Median Filtered Independent Data," IEEE Trans. on Commun., Vol. COM-29, No. 9, pp. 1374 Sept. 1981.
9. Aulx F. Velleman, "Definition and Comparison of Robust Nonlinear Data Smoothing Algorithms," Journal of the American Statistical Assoc., pp. 609, Sept. 1980.

On a Class of Random Processes Exhibiting Optimal Nonlinear One-Step Predictors

T. E. McCANNON AND NEAL C. GALLAGHER, MEMBER, IEEE

Abstract—Two classes of random processes that exhibit one-step predictors with optimal nonlinear minimum mean-squared error (MMSE) are discussed, and conditions for membership to one of these classes are given. Examples of each class are presented, and the optimal one-step predictors are given.

I. Introduction

The problem of designing minimum mean-squared error (MMSE) prediction filters is often complicated by the absence of prior information on the mathematical structure of the optimum predictor. Historically it has often been assumed that the optimum implementable predictor is linear, and such well-known techniques as Wiener–Hopf spectral factorization or the orthogonality principle are applied to determine the optimum prediction filters. With the advent of modern digital technology, nonlinear functions are often easily implemented, and hence a renewed interest in optimal nonlinear-prediction theory has arisen.

We have previously presented [1] two methods of designing nonlinear MMSE predictions filters where we have assumed a polynomial nonlinearity followed by a linear filter. For both of these design methods, all that is required is knowledge of a finite number of moments and cross moments of the given random process. Wise and Gallagher [2] have shown that knowledge of certain moments is sufficient to specify the conditional expectation. In this case, the optimum nonlinear-prediction filter is given by a polynomial in the sample observations.

In this correspondence we point out two classes of kth order stationary random processes $\{X_n\}$ possessing as their optimum MMSE estimate

$$E\{X_{n+1}|X_n,\cdots,X_{n-k+1}\} = f(X_n,\cdots,X_{n-k+1}),$$

where we assume $f(\cdot)$ to be a Borel measurable function that can be of a nonpolynomial form. The first class we consider corresponds to the random process being represented by a nonlinear stochastic-difference equation

$$X_{n+1} = f(X_n,\cdots,X_{n-k+1}) + U_{n+1}. \tag{1}$$

The second class corresponds to the output obtained from passing a known process through an invertible zero-memory nonlinearity (ZNL). Such a process is of the form $X_n = g(Z_n)$, where we know the form of the predictor for the $\{Z_n\}$ process. This class of problems is of particular interest, because the best predictor for X_n does not, in general, involve finding the best prediction for Z_n and using it as the input valve for $g(\cdot)$. The form of the optimal X_n predictor can be quite complicated.

II. Class I

We define a kth-order stationary random process $\{X_n\}$ as being a Class I random process if and only if $\{X_n\}$ can be represented by a stochastic difference equation in the form of (1), where $\{U_n\}$ are independent identically distributed (i.i.d.) zero-mean random variables with a marginal density given by $P_u(\cdot)$. Clearly the conditional expectation of X_{n+1}, given the infinite past of the process, is

$$E\{X_{n+1}|X_n, X_{n-1},\cdots\} = f(X_n, X_{n-1},\cdots,X_{n-k+1}).$$

Writing the Chapman–Kolmogorov equation for the kth variate densities, we obtain

$$P_x^{(n+1)}(x_{n+1},\cdots,x_{n-k+2}) = \int\cdots\int q_{n+1}(x_{n+1},\cdots,x_{n-k+2}|z_n,\cdots,z_{n-k+1}) \cdot P_x^{(n)}(z_n,\cdots,z_{n-k+1})\, dz_n\cdots dz_{n-k+1}, \tag{2}$$

where $P_x^{(n)}(x_n,\cdots,x_{n-k+1})$ is the joint density of $(X_n,\cdots,X_{n-k+1})$ for the nth sampling instant, and $q_{n+1}(x_{n+1},\cdots,x_{n-k+2}|x_n,\cdots,x_{n-k+1})$ is the conditional density of $(X_{n+1},\cdots,X_{n-k+2})$, given $(X_n,\cdots,X_{n-k+1})$ for the $(n+1)$ sampling instant. From (1), we obtain

$$P_x^{(n+1)}(x_{n+1},\cdots,x_{n-k+2}) = \int P_u[x_{n+1} - f(x_n,\cdots,x_{n-k+2}, z_{n-k+1})] \cdot P_x^{(n)}(x_n,\cdots,x_{n-k+2}, z_{n-k+1})\, dz_{n-k+1}. \tag{3}$$

Since we have assumed that $\{X_n\}$ is kth-order stationary, the kth variate densities are independent of n, and (3) can be rewritten as

$$P_x(t, s_1,\cdots,s_{k-1}) = \int P_u[t - f(s_1,\cdots,s_k)]\, P_x(s_1,\cdots,s_k)\, ds_k. \tag{4}$$

We now state the following property.

Property 1: If $\{X_n\}$ is a kth-order stationary random process and is representable in the form

$$X_{n+1} = f(X_n,\cdots,X_{n-k+1}) + U_{n+1},$$

where $\{U_n\}$ are i.i.d. zero-mean random variables, then the kth

Manuscript received June 26, 1980. This work was supported by the Air Force Office of Scientific Research under Grant AFOSR-78-3605.

The authors are with the School of Electrical Engineering, Purdue University, West Lafayette, IN 47904.

0018-9448/81/0900-0652$00.75 ©1981 IEEE

variate density satisfies the integral equation

$$P_x(t, s_1, \cdots, s_{k-1}) = \int P_u[t - f(s_1, \cdots, s_k)]\, P_x(s_1, \cdots, s_k)\, ds_k.$$

Our convention is that lower case variables represent the realizations of upper case random variable. Furthermore, the optimum MMSE prediction filter is given by

$$E\{X_{n+1} | x_n, \cdots, x_{n-k+1}\} = f(x_n, \cdots, x_{n-k+1}).$$

An equivalent representation can be derived in a straightforward manner. We begin by noting that the $(k+1)$st variate density is given by

$$P_x(t, s_1, \cdots, s_k) = P_u[t - f(s_1, \cdots, s_k)]\, P_x(s_1, \cdots, s_k). \quad (5)$$

We then perform the expectation of $[t \exp i \Sigma_j^k S_j p_j]$, where i is the complex constant, as follows:

$$E\left\{t \exp i \sum_{j=1}^{k} S_j P_j\right\}$$

$$= \int \cdots \int \left[t \exp i \sum_{j=1}^{k} s_j p_j\right] P_x(t, s_1, \cdots, s_k)\, dt \cdots ds_k$$

$$= \int \cdots \int \left[\exp i \sum_{j=1}^{k} s_j p_j\right] \left[t P_u[t - f(s_1, \cdots, s_k)]\, dt\right]$$

$$P_x(s_1, \cdots, s_k)\, ds_1 \cdots ds_k. \quad (6)$$

Via (1) we see that

$$E\{X_{n+1} | X_n, \cdots, X_{n-k+1}\} = \int t P_u[t - f(X_n, \cdots, X_{n+k+1})]\, dt.$$

Hence we can write

$$E\left\{t \exp i \sum_{j=1}^{k} S_j p_j\right\} \int \cdots \int f(s_1, \cdots, s_k) \left[\exp i \sum_{j=1}^{k} s_j p_j\right]$$

$$\cdot P_x(s_1, \cdots, s_k)\, ds_1 \cdots ds_k$$

$$= E\left\{f(S_1, \cdots, S_k) \exp i \sum_{j=1}^{k} S_j p_j\right\}. \quad (7)$$

Thus we have (7) equivalent to (5). Note that (7) involves the characteristic function and as such requires knowledge of the kth variate distribution. However, there may be circumstances where (7) may be easier to apply than (5).

The expression in (7) is a generalization of a result presented by Balakrishnan [3] for polynomial nonlinearities $Q(\cdot)$, i.e.,

$$E\left\{t \exp i \sum_{j=1}^{k} S_j p_j\right\} = E\left\{Q(S_1, \cdots, S_k) \exp i \sum_{j=1}^{k} S_j p_j\right\}, \quad (8)$$

where $Q(\cdot)$ is the optimum MMSE estimator.

III. Class II

Define a stationary random process $\{Z_n\}$ such that

$$X_n = g(Z_n), \qquad n = \{\cdots, -2, -1, 0, 1, 2, \cdots\}, \quad (9)$$

where $g(\cdot)$ is an invertible function. This particular relation is of interest because the random process $\{Z_n\}$ might possess a simple MMSE one-step predictor. For example, suppose that $\{X_n\}$ is such that we can find a $g(\cdot)$ for which $\{Z_n\}$ is Gaussian. We then know that the MMSE one-step predictor on $\{Z_n\}$ is linear. We wish to investigate the best predictor for the $\{X_n\}$ process.

First, it is necessary to define what we will call optimal MMSE estimators and suboptimal MMSE estimators. We consider an MMSE estimator to be optimal if

$$E\left\{\left(Y - E\{Y | x_1, \cdots, x_j\}\right)^2\right\}$$

$$= E\left\{\left(Y - E\{Y | x_1, \cdots, x_j, \cdots\}\right)^2\right\};$$

that is, one cannot do better even if more information is available. Similarly we consider an MMSE estimator to be suboptimal if

$$E\left\{\left(Y - E\{Y | x_1, \cdots, x_j\}\right)^2\right\}$$

$$> E\left\{\left(Y - E\{Y | x_1, \cdots, x_j, \cdots\}\right)^2\right\};$$

that is, one can do better with more information available.

Consider the case where $\{Z_n\}$ is in Class I, as discussed in the previous section. The known optimal one-step predictor is

$$E\{Z_{n+1} | z_n, \cdots, z_{n-k+1}\} = f(z_n, \cdots, z_{n-k+1}).$$

We know from (1) that the conditional density of Z_{n+1}, given $(Z_n, \cdots, Z_{n-k+1})$, is given by

$$q(z_{n+1} | z_n, \cdots, z_{n-k+1}) = p_u[z_{n+1} - f(z_n, \cdots, z_{n-k+1})], \quad (10)$$

such that

$$E\{g(Z_{n+1}) | z_n, \cdots, z_{n-k+1}\}$$

$$= \int g(z_{n+1}) p_u[z_{n+1} - f(z_n, \cdots, z_{n-k+1})]\, dz_{n+1}. \quad (11)$$

Employing a change of variable, we can rewrite (11) as

$$E\{g(Z_{n+1}) | z_n, \cdots, z_{n-k+1}\}$$

$$= \int g[u + f(z_n, \cdots, z_{n-k+1})]\, p_u(u)\, du. \quad (12)$$

If we assume that $g(\cdot)$ can be written in the Taylor series

$$g(x + a) = \sum_{i=0}^{\infty} \frac{g^{(i)}(a)}{i!} x^i, \quad (13)$$

then we can rewrite (12) into the form

$$E\{g(Z_{n+1}) | z_n, \cdots, z_{n-k+1}\}$$

$$= \sum_{i=0}^{\infty} g^{(i)}[f(z_n, \cdots, z_{n-k+1})] \frac{1}{i!} u^i P_u(u)\, du. \quad (14)$$

Defining

$$a_i \equiv \frac{1}{i!} \int u^i p_u(u)\, du,$$

and using the fact that $g(\cdot)$ is an invertible function, we can rewrite (14) into the desired predictor for $\{X_n\}$:

$$E\{X_{n+1} | x_n, \cdots, x_{n-k+1}\}$$

$$= \sum_{i=0}^{\infty} a_i g^{(i)}\left\{f[g^{-1}(x_n), \cdots, g^{-1}(x_{n-k+1})]\right\} \quad (15)$$

Note that (15) is valid only when $\{Z_n\}$ belongs to Class I considered in Section II. This implies that (15) can be completely determined because the coefficients a_i correspond directly to knowledge of the marginal moments of the white driving process.

Suppose that the random process $\{Z_n\}$ is characterized by a suboptimal MMSE one-step predictor of the form

$$E\{Z_{n+1}|z_n,\cdots,z_{n-k+1}\} = h(z_n,\cdots,z_{n-k+1}). \quad (16)$$

where in the previous example we assumed the optimal predictor to be of this form. In this case, we do not know the conditional density corresponding to (10). If we define $\{\epsilon_n\}$ as a random process denoting the error at each sampling instant between the optimal estimate and the suboptimal estimate and let $P_{\epsilon_n}(\cdot)$ denote its marginal density at the nth sampling instant, we can then write the conditional density of Z_{n+1}, given $(Z_n,\cdots,Z_{n-k+1})$, as

$$q(z_{n+1}|z_n,\cdots,z_{n-k+1}) = \int r(z_{n+1}|z_n,\cdots,z_{n-k+1},\epsilon_{n+1})P_{\epsilon_{n+1}}(\epsilon_{n+1})\,d\epsilon_{n+1}, \quad (17)$$

where $r(\cdot|\cdot)$ denotes the conditional density of Z_{n+1} given $(Z_n,\cdots,Z_{n-k+1})$ and ϵ_{n+1}. Because $h(\cdot)$ is an MMSE estimator, $E\{\epsilon_{n+1}\} = 0$. We can then write the recursive relation

$$Z_{n+1} = h(Z_n,\cdots,Z_{n-k+1}) + \epsilon_{n+1} + P_{n+1} \quad (18)$$

describing the process $\{Z_n\}$, where $\{P_n\}$ is a zero-mean white driving process. Equation (18) is then the suboptimal analog to (1). From (18) we can write

$$r(z_{n+1}|z_n,\cdots,z_{n-k+1},\epsilon_{n+1}) = p_p[z_{n+1} - h(z_n,\cdots,z_{n-k+1}) - \epsilon_{n+1}]. \quad (19)$$

Substituting (19) into (17), we then have

$$q(z_{n+1}|z_n,\cdots,z_{n-k+1}) = \int p_p[z_{n+1} - h(z_n,\cdots,z_{n-k+1}) - \epsilon_{n+1}]P_{\epsilon_{n+1}}(\epsilon_{n+1})\,d\epsilon_{n+1}.$$

We compute $E\{g(Z_{n+1})|z_n,\cdots,z_{n-k+1}\}$ as before and write

$$E\{g(Z_{n+1})|z_n,\cdots,z_{n-k+1}\} = \int\int g(z_{n+1})p_p[z_{n+1} - h(z_n,\cdots,z_{n-k+1}) - \epsilon_{n+1}] \cdot P_{\epsilon_{n+1}}(\epsilon_{n+1})\,d\epsilon_{n+1}\,dz_{n+1}. \quad (20)$$

Employing a change of variable, we put (20) into the form

$$E\{g(Z_{n+1})|z_n,\cdots,z_{n-k+1}\} = \int g[p + h(z_n,\cdots,z_{n-k+1})]\int p_p(p-\epsilon)P_{\epsilon_{n+1}}(\epsilon)\,d\epsilon\,dp. \quad (21)$$

Again, if we assume that $g(\cdot)$ can be expanded into a Taylor series, remembering that $g(\cdot)$ is a invertible function, and upon defining

$$b_{n+1,i} = \frac{1}{i!}\int\int p^i P_p(p-\epsilon)P_{\epsilon_{n+1}}(\epsilon)\,d\epsilon\,dp, \quad (22)$$

we obtain

$$E\{X_{n+1}|x_n,\cdots,x_{n-k+1}\} = \sum_{i=0}^{\infty} b_{n+1,i}g^{(i)}\{h[g^{-1}(x_n),\cdots,g^{-1}(x_{n-k+1})]\}. \quad (23)$$

If we assume that the marginal error density is independent of n, (23) can be simplified to

$$E\{X_{n+1}|x_n,\cdots,x_{n-k+1}\} = \sum_{i=0}^{\infty} c_i g^{(i)}\{h[g^{-1}(x_n),\cdots,g^{-1}(x_{n-k+1})]\}, \quad (24)$$

where we set $b_{n+1,i} = c_i$. In most cases, the marginal density $P_{\epsilon_{n+1}}(\cdot)$ will be difficult, if not impossible, to obtain. For this reason, (23) and (24) should only be interpreted as providing a functional form for the prediction filter, and the coefficients $b_{n+1,i}$ and c_i should be obtained through some procedure which minimizes the quantity

$$E\left\{\left(X_{n+1} - E\{X_{n+1}|x_n,\cdots,x_{n-k+1}\}\right)^2\right\}.$$

IV. Examples

A. Class I Random Process ($k = 1$)

Consider the random process characterized by the nonlinear stochastic difference equation

$$X_{n+1} = f(X_n) + U_{n+1},$$

where $\{(U_n\}$ are i.i.d. zero-mean random variables. For the case where $k = 1$, (7) becomes

$$P_x^{(n+1)}(x_{n+1}) = \int P_u[x_{n+1} - f(x_n)]P_x^{(n)}(x_n)\,dx_n.$$

We now make use of the following theorem proved in the Appendix.

Theorem: If the random process $\{X_n\}$ can be characterized by

$$X_{n+1} = f(X_n) + U_{n+1},$$

where

1) $P_u(\cdot)$ is strictly positive and uniformly continuous on a finite closed support Ω_u, and
2) f: $\Omega \to \Omega_f$ such that $\{v: v = u + \tilde{f},\ u \in \Omega_u,\ \tilde{f} \in \Omega_f\} \subset \Omega$ and $f(\cdot)$ is continuous on Ω_u,

then the densities $P_x^{(n)}(x_n)$ converge to a steady state limiting density $P_x(x_n)$ with finite closed support Ω_x. Because the marginal densities possess steady state limits, the random process $\{X_n\}$ is asymptotically first-order stationary. Hence if $\{X_n\}$ satisfies the conditions of this theorem, then $\{X_n\}$ belongs to Class I with $k = 1$.

B. Class II Random Process

We consider a particular example of Class II. We assume that $\{Z_n\}$ is a zero-mean Gaussian random process and that we have a suboptimal characterization implying that either (23) or (24) applies. Consider

$$h(z_n,\cdots,z_{n-k}) = c_k z_n. \quad (25)$$

Applying the orthogonality principle to obtain the coefficient c_k, we find that the suboptimal one-step predictor is given by

$$h(Z_n,\cdots,Z_{n-k}) = \rho Z_n,$$

where

$$\rho \equiv \frac{E\{Z_{n+1}Z_n\}}{E\{Z_n^2\}}.$$

Therefore, the random variables $(Z_{n+1} - \rho Z_n)$ are uncorrelated with the random variable Z_n at each sampling instant. For this

example (18) becomes

$$Z_{n+1} = \rho Z_n + \epsilon_{n+1} + P_{n+1}, \tag{26}$$

where we require $P_n \sim N(0, \sigma_p^2)$. Consequently, $\epsilon_n \sim N(0, \sigma_{\epsilon_n}^2)$, and at a fixed sampling instant Z_n and $(\epsilon_{n+1} + P_{n+1})$ are independent. Since ϵ_{n+1} can only depend on $\{P_i, i \le n\}$, P_{n+1} and ϵ_{n+1} are independent as a result of the whiteness of $\{P_n\}$. Because $\{Z_n\}$ is stationary, the prediction error variance must be independent of n, and since σ_p^2 is a constant, $\sigma_{\epsilon_n}^2 = \sigma_\epsilon^2$ must also be a constant.

Recall from (22) that

$$c_i = \frac{1}{i!} \int\int p^i P_p(p - \epsilon) P_\epsilon(\epsilon)\, d\epsilon\, dp. \tag{27}$$

If we rewrite (27) into

$$c_i = \frac{1}{i!} \int p^i \left[\int P_p(p - \epsilon) P_\epsilon(\epsilon)\, d\epsilon \right] dp,$$

we notice that the integral within the brackets corresponds to the density of the sum, $P_{n+1} + \epsilon_{n+1}$. Hence

$$\int P_p(p - \epsilon) P_\epsilon(\epsilon)\, d\epsilon = \frac{1}{\sqrt{2\pi}\sqrt{\sigma_p^2 + \sigma_\epsilon^2}} \exp\left[-p^2/2\left(\sigma_p^2 + \sigma^2\right)\right],$$

and (27) becomes

$$c_i = \frac{1}{i!} \int p^i \frac{1}{\sqrt{2\pi}\sqrt{\sigma_p^2 + \sigma_\epsilon^2}} \exp\left[-p^2/2\left(\sigma_p^2 + \sigma_\epsilon^2\right)\right].$$

The moments of a zero-mean normal random variable are [6]

$$c_i = \begin{cases} \frac{1}{i!}[1 \cdot 3 \cdots (i-1)]\left(\sigma_p^2 + \sigma_\epsilon^2\right)^{1/2}, & i \text{ even} \\ 0, & i \text{ odd.} \end{cases} \tag{28}$$

Because Z_n and $(\epsilon_{n+1} + p_{n+1})$ are independent, we can use (26) to obtain the expression

$$\sigma_z^2 = \rho^2 \sigma_z^2 + \left(\sigma_p^2 + \sigma_\epsilon^2\right), \tag{29}$$

where we have used the fact that $\{Z_n\}$ is stationary. Substituting (29) into (28), we obtain for the predictor coefficients

$$c_i = \begin{cases} \frac{1}{i!}[1 \cdot 3 \cdots (i-1)](1 - \rho^2)^{i/2} \sigma_z^i, & i \text{ even} \\ 0, & i \text{ odd.} \end{cases} \tag{30}$$

Finally, by substitution of (25) into (24) and using the nonlinearity $g(z) = z^5$, we obtain for the form of the predictor

$$E\{X_{n+1} | x_n\} = c_0 \rho^5 x_n + 5c_1 \rho^4 x_n^{4/5} + 20c_2 \rho^3 x_n^{3/5} + 60c_3 \rho^2 x_n^{2/5} + 120c_4 \rho x_n^{1/5} + 120c_5.$$

From (30) we obtain the coefficient values

$$c_0 = 1,$$
$$c_2 = \tfrac{1}{2}(1 - \rho^2)\sigma_z^2,$$
$$c_4 = \tfrac{1}{8}(1 - \rho^2)^2 \sigma_z^4,$$
$$c_1 = c_3 = c_5 = 0.$$

We thus obtain for the MMSE one-step predictor

$$E\{X_{n+1} | x_n\} = 15\rho(1 - \rho^2)^2 \sigma_z^4 x_n^{1/5} + 10\rho^3(1 - \rho^2)\sigma_z^2 x_n^{3/5} + \rho^5 x_n$$

Appendix
Proof of Theorem

Theorem: If the random process $\{X_n\}$ can be characterized by

$$x_{n+1} = f(X_n) + U_{n+1},$$

where

1) $P_u(\cdot)$ is strictly positive and uniformly continuous on a finite closed support Ω_u and
2) f: $\Omega \to \Omega_{\bar{f}}$ such that $\{v: v = u + \bar{f},\ u \in \Omega_u,\ \bar{f} \in \Omega_{\bar{f}}\} \subseteq \Omega$, and $f(\cdot)$ is continuous on Ω_u, then the densities $P_X^{(n)}(x_n)$ converge to a steady state limiting density $P_x(x_n)$ with finite closed support Ω_x.

Proof: We need to show

1) Ω_x bounded and closed and
2) $P_u(\cdot)$ is strictly positive and regular

at which point we can then apply the results due to Feller [5].

1) Ω_x Bounded and Closed: Consider first $\Omega_n \equiv \{\text{support of } P_x^{(n)}(\cdot)\}$. We show by induction that Ω_n is bounded and closed for all n. For $n = 0$, f: $\Omega_u \to \Omega_{\bar{f}}$, and because Ω_u is bounded and closed and $f(\cdot)$ is continuous on Ω_u then $\Omega_{\bar{f}}$ is bounded and closed. Now,

$$\Omega_1 = \left\{v: v = u + \bar{f},\ u \in \Omega_u, \bar{f} \in \Omega_{\bar{f}}\right\}.$$

To prove Ω_1 is bounded, first assume that Ω_1 is not bounded; then there exists $\{\bar{f}_j\} \in \Omega_{\bar{f}}$ such that $\bar{f}_j \to \infty$ and $\{u_j\} \in \Omega_u$ such that $u_j \to \infty$ or both so that $v_j = (u_j + \bar{f}_j) \to \infty$. But $\Omega_{\bar{f}}$ and Ω_u are bounded. Hence Ω_1 is bounded.

To prove Ω_1 is closed, for any $\bar{f}_0 \in \Omega_{\bar{f}}$ there exists $\{\bar{f}_j\} \in \Omega_{\bar{f}}$ such that $\{\bar{f}_j\} \to \bar{f}_0$, and for any $u_0 \in \Omega_u$ there exists $\{u_j\} \in \Omega_u$ such that $u_j \to u_0$. Form the sequence $v_j = u_j + \bar{f}_j$; $v_j \to v_0 = \bar{f}_0 + u_0$. But $v_0 \in \Omega_1$, for all such sequences in Ω_1. Hence Ω_1 is closed.

Assume Ω_n is bounded and closed. By the argument above, Ω_{n+1} is bounded and closed. Hence Ω_n is bounded and closed for all n.

From condition (2), we know that f: $\Omega_n \to \Omega_{\bar{f}}$ such that $\{v: v = u + \bar{f}, u \in \Omega_u, \bar{f} \in \Omega_{\bar{f}}\} \subseteq \Omega_n$, and that $\Omega_{n+1} = \{v: v = u + \bar{f}, u \in \Omega_n, \bar{f} \in \Omega_{\bar{f}}\}$. Hence $\Omega_{n+1} \subseteq \Omega_n \subseteq \Omega_{n-1} \subseteq \cdots \subseteq \Omega_1 \subseteq \Omega_u$. The support of $P_x(\cdot)$ is given by

$$\Omega_x = \Omega_u \bigcap \left(\bigcap_{n=1}^{\infty} \Omega_n \right).$$

Since Ω_n is bounded and closed for all n and Ω_u is bounded and closed by assumption, then Ω_x is bounded and closed. Also, since $\Omega_x \subseteq \Omega_u$ and $P_u(\cdot) > 0$ and uniformly continuous on Ω_u, then $P_u(\cdot) > 0$ and uniformly continuous on Ω_x.

2) The Kernel ($P_u(\cdot)$) Is Strictly Positive and Regular: We have already shown $P_u(\cdot) > 0$ and uniformly continuous on Ω_x.

Definition (Feller): The kernel is regular if the family of transforms $P_x^{(n)}(\cdot)$ are equicontinuous whenever $P_x^0(\cdot)$ is uniformly continuous in Ω_x.

We note that $P_x^{(0)}(\cdot) = P_u(\cdot)$. Hence $P_x^{(0)}$ is uniformly continuous on Ω_x. We have that

$$P_x^{(n)}(\phi) = \int_{\Omega_{n-1}} P_u\left[\phi - f(z)\right] P_x^{(n-1)}(z)\, dz,\ \phi \in \Omega_n$$

Look at the expression of $\phi - f(z)$. We have that $\phi \in \Omega_n$, which is bounded and closed, and f: $\Omega_{n-1} \to \Omega_{\bar{f}}$, which is bounded and closed. Define

$$\Omega_{p_n} = \left\{p: p = \phi - \bar{f}, \phi \in \Omega_n, \bar{f} \in \Omega_{\bar{f}}\right\}.$$

By the same argument we used to show Ω_n is bounded and closed when Ω_{n-1} is bounded and closed, we can state Ω_{p_n} is bounded

and closed. Since Ω_{p_n} is compact and $P_u(\cdot)$ is uniformly continuous, then there exist $a_{n_0} \in \Omega_{p_n}$ such that $P_u(a_{n_0}) = \sup_{x \in \Omega_{p_n}} P_u(x)$.

Define $M = \max_n(|P_u(a_{n_0})|)$. Then because

$$|P_x^{(n)}(\cdot)| \leq \int_{\Omega_{n-1}} |P_u[\phi - f(z)]| \cdot |P_x^{(n-1)}(z)|\, dz,$$

we have

$$|P_x^{(n)}(\cdot)| \leq M \int_{\Omega_{n-1}} |P_x^{(n-1)}(z)|\, dz = M, \qquad \text{for all } n.$$

Recalling that

$$P_x^{(n)}(\phi) = \int_{\Omega_{n-1}} P_u[\phi - f(z)] P_x^{(n-1)}(z)\, dz,$$

we can immediately write

$$|P_x^{(n)}(\phi') - P_x^{(n)}(\phi'')| \leq \int_{\Omega_{n-1}} |P_u[\phi' - f(z)] - P_u[\phi'' - f(z)]| \cdot |P_x^{(n-1)}(z)|\, dz.$$

Define $W = \max_n \int_{\Omega_n} dz < \infty$. Pick an arbitrary $z_0 \in \Omega_{n-1}$. Give $\epsilon > 0$. Let $\delta = \delta(\epsilon) > 0$ such that $|\phi' - \phi''| \leq \delta$. Then $|P_u[\phi' - f(z_0)] - P_u[\phi'' - f(z_0)]| \leq \epsilon/MW$, because $P_u(\cdot)$ is uniformly continuous on Ω_n, for all n. Hence $|P_x^{(n)}(\phi') - P_x^{(n)}(\phi'')| \leq (\epsilon/MW)\, MW = \epsilon$, for all n and $P_x^{(n)}(\cdot)$; therefore, $P_x^{(n)}(\cdot)$ are equicontinuous and the kernel $P_u(\cdot)$ is regular. We now appeal to the following theorems:

Theorem 3 [Feller]: Every strictly positive regular kernel on a bounded closed interval is ergodic;

Theorem 4 [Feller]: A strictly positive regular kernel is ergodic if and only if it possesses a strictly positive stationary probability distribution; where $P_x(x)$ has support Ω_x which is bounded and closed.

References

[1] T. E. McCannon, N. C. Gallagher, G. L. Wise, and D. Minoo-Hamedani, "A novel approach for designing nonlinear discrete time filters: Part II," in *Proc. 16th Ann. Allerton Conf. Communication, Control and Computing*, Oct. 4-6, 1978.

[2] G. L. Wise and N. C. Gallagher, "On the determination of regression functions," in *Proc. 17th Ann. Allerton Conf. Communication, Control and Computing*, Oct. 10-12, 1979.

[3] A. V. Balakrishnan, "On a characterization of processes for which optimal mean-square systems are of specified form," *IRE Trans. Inform. Theory*, vol. IT-6, pp. 490-500, Sept. 1960.

[4] H. E. Henry and P. M. Schultheiss, "The analysis of certain nonlinear feedback systems with random inputs," *IRE Trans. Inform. Theory*, vol. IT-8, pp. 25-29, July 1962.

[5] W. Feller, *An Introduction to Probability Theory and Its Applications*. New York: Wiley, 1965, pp. 270-272.

[6] A. Papoulis, *Probability, Random Variables and Stochastic Processes*. New York: McGraw-Hill, 1965, pp. 147-148.

[7] H. L. Royden, *Real Analysis*. New York: Macmillian, 1968.

A Theoretical Analysis of the Properties of Median Filters

NEAL C. GALLAGHER, JR., MEMBER, IEEE, AND GARY L. WISE, MEMBER, IEEE

Abstract—Necessary and sufficient conditions for a signal to be invariant under a specific form of median filtering are derived. These conditions state that a signal must be locally monotone to pass through a median filter unchanged. It is proven that the form of successive median filtering of a signal (*i.e.*, the filtered output is itself again filtered) eventually reduces the original signal to an invariant signal called a root signal. For a signal of length L samples, a maximum of $\frac{1}{2}(L-2)$ repeated filterings produces a root signal.

Manuscript received December 7, 1979; revised April 16, 1981. This work was supported by the Air Force Office of Scientific Research under Grants AFOSR 78-3605, AFOSR 76-3602, and AFOSR 81-0047.

N. C. Gallagher, Jr. is with the School of Electrical Engineering, Purdue University, W. Lafayette, IN 47907.

G. L. Wise is with the Department of Electrical Engineering, University of Texas, Austin, TX 78712.

I. Introduction

In many signal processing applications, a method called median filtering has achieved some very interesting results. One useful characteristic of median filtering is its ability to preserve signal edges while filtering out impulses. Promising applications of median filtering are picture processing and speech processing [1]-[3]. The implementation of a median filter requires a very simple digital nonlinear operation. To begin, we take a sampled and quantized signal of length L; across this signal we slide a window that spans $2N+1$ signal sample points. The filter output is set equal to the median value of these $2N+1$ signal samples, and is associated with the time sample at the center of the window.

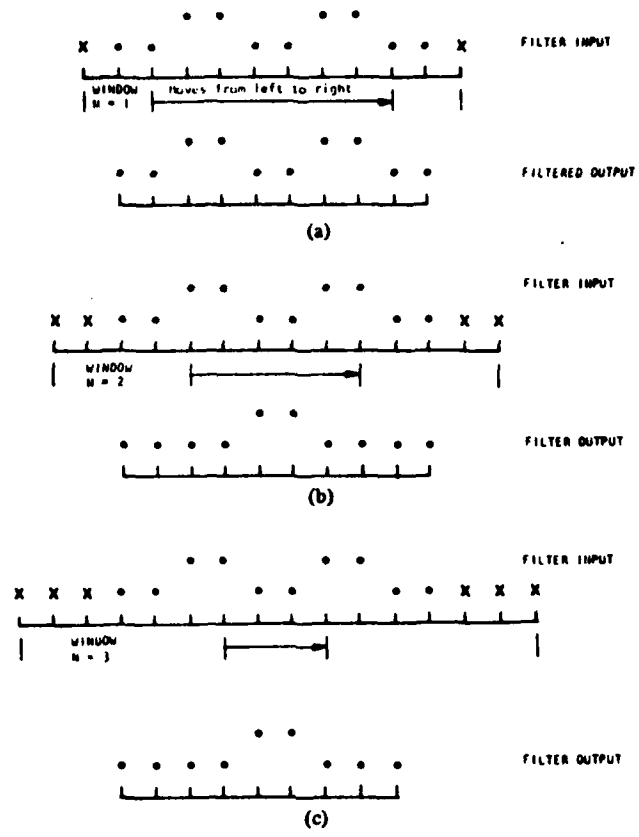


Fig. 1. Signal filtered by three different median filters: (a) $N = 1$, (b) $N = 2$, and (c) $N = 3$.

In one form of median filtering, to account for startup and end effects at the two endpoints of the L-length signal, N samples are appended to the beginning and the end of the sequence. The appended samples are constant and equal in value to the first and last samples of the original sequence, respectively. For other ways of treating the start-up problem that gives less emphasis to the first and last values encountered, see [4, p. 221].

As an example, consider the binary valued sequence of Fig. 1(a) where $L = 10$ and $N = 1$; the median filtered signal is plotted below the extended input signal. The appended values are marked as X's. Fig. 1(b) illustrates the filtering of the same input signal as for Fig. 1(a), but we set $N = 2$; we set $N = 3$ for the example in Fig. 1(c). The signal of Fig. 1 passes undisturbed through the $N = 1$ filter; however, it is affected by the $N = 2$ and $N = 3$ filters. The signal would be reduced to a constant value by an $N = 4$ filter.

The results illustrated in Fig. 1 suggest the concept of a filter "passband" and "stopband." The given signal is in the passband of the $N = 1$ filter and the stopband of the $N = 4$ filter. If we view the median filter as one that passes edges but not impulses, then edges for an $N = 1$ filter may be impulses for an $N = 4$ filter. But what about the $N = 2$ and $N = 3$ filters? Suppose the signal of Fig. 1 is filtered twice in succession by the $N = 2$ filter; in other words, the filtered output is again filtered. The result in this specific instance is a constant output identical to that obtained by a single filtering with an $N = 4$ filter. If the constant is filtered again, the output is the same as the filter input; the constant is invariant to median filtering. So, by filtering this particular original signal two times with an $N = 2$ or $N = 3$ filter, we have a resulting signal that is invariant to successive filterings, the same result obtained by a single pass with the $N = 4$ filter. Note that the signal input signal of Fig. 1 is invariant to repeated filtering with an $N = 1$ filter. We call such a signal a root of the median filter. We see that signals which do not reside entirely within the filter "passband" can be reduced to their passband component by repeated filterings.

In this paper, we will formalize the concepts of filter passband and stopband. We described desirable signal characteristics for signals employed in median filtering, and show how some types of noise can be completely removed by median filtering and how other types cannot be removed. These results will be presented through the development of a formal theory of median filtering. In Section II we present some basic definitions that allow us to precisely state and prove a number of interesting results. The reader concerned only with results may wish to proceed to Section III.

II. Theory for Median Filtering

In order to give a precise statement for the theorems presented later in the section, a number of definitions are necessary. We will always be working with a sample length I, where each sample is quantized to one of K different values. The fil-

ter window length is the number of consecutive samples considered when computing the running median. We will always take the window length to be an odd integer ($2N + 1$) for $N = 0, 1, 2, \cdots$. As noted earlier, our convention is that the filter output at position L is the median value obtained when position L is in the center of the window. We define the following signal characteristics.

1) A *constant neighborhood* is at least $N + 1$ consecutive identically valued points such that the constant neighborhoods and edge together are monotone.

2) An *edge* is a monotonic region between two constant neighborhoods of different value. The connecting monotonic region cannot contain any constant neighborhood.

3) An *impulse* is a constant neighborhood followed by at least one, but no more than N points which are then followed by another constant neighborhood having the same value as the first constant neighborhood. The two boundary points of these at most N points do not have the same value as the two constant neighborhoods.

4) An *oscillation* is a sequence of points which is not part of a constant neighborhood, an edge, or an impulse.

Of particular interest is the class of signals that can pass through the filter unchanged, as well as the class of signals that are completely removed by filtering. Assume that an L-length signal is filtered with a $2N + 1$ window. As noted previously, we always append to the beginning of the signal an additional N constants equal in value to the first sample of the signal. Similarly, N constant points are appended to the end of the L-length signal. By doing this, we assure that when the initial signal's first or last sample is in the center of the window, the median filter output equals this sample value. For a signal to pass through a median filter unchanged means that the central sample value for each window position is itself the median of the samples within the window.

Consider a signal that is unchanged by median filtering. Assume that the window increments from sample to sample moving from left to right across the signal and that the window is now centered at the second signal sample of the original signal. We know that the N points to the left of center have the same constant value. If they equal the value of the center point, then it (the center point) must be the median. If they are less than the value of the center point, then the N points to the right of center must be all greater than or equal to the central value. If the N points to the left are greater in value than the central point, then the N points to the right are all less than or equal to the center value. Thus, note that the leftmost $N + 2$ points in the window form a monotone sequence of points. Increment the window another sample to the right, so that the window is now centered at the third signal sample. The leftmost $N + 1$ samples in the window form a monotone sequence. Assume that the N leftmost points in the window are not greater than (respectively, not less than) the center point. Then, since the center point is the median value of the points in the window, the N rightmost points in the window must be not less than (respectively, not greater than) the center point. Thus, we see once again that the leftmost $N + 2$ points in the window form a monotone sequence. Increment the window another sample to the right. By applying the same argument as before, we again find that the $N + 2$ leftmost points in the window form a monotone sequence. Indeed, a straightforward inductive argument proves that the leftmost $N + 2$ points in the window form a monotone sequence regardless of the window position. Recalling that the extended signal has N constant points appended to the right of the original signal, we see that the extended signal is such that any consecutive $N + 2$ points must be monotone. Thus, a signal invariant to median filtering must be such that the extended signal contain only constant neighborhoods and edges.

Now assume that the extended signal contains only constant neighborhoods and edges. If the center of the window is at any signal sample, then the points in the window are either monotone or nonmonotone. If the points are monotone, then the signal sample at the center of the window is not changed by the median filter. If they are nonmonotone, then the window must be centered on a point in the constant neighborhood shared by two edges. Of the $2N + 1$ points in the window, at least $N + 1$ of them are equal to the center point, and thus the center point is unchanged by median filtering.

These observations are formalized in the following theorem.

Theorem 1—Given a length-L, K-valued sequence to be median filtered with a $2N + 1$ window, a necessary and sufficient condition for the signal to be invariant under median filtering is that the extended signal consist only of constant neighborhoods and edges.[1]

The following corollary is a direct result of this theorem.

Corollary—For a median-filter-invariant signal to contain both regions of increase and decrease, the points of increase and decrease must be separated by a constant neighborhood (at least $N + 1$ consecutive identical points).

As a result of this theorem, it is possible to construct signals that are invariant to median filtering. Also, given the space of all length-L, K-valued signals S, it is possible to identify all those signals invariant to median filtering with a $2N + 1$ window. We will call these signals the roots of the filter, and this set of signals is denoted as R_N. Note that $R_N \subset S$ for any N, and that we have the following lemma.

Lemma 1: For an L-length, K-valued set of signals S, the root sets R_N are nested such that

$$\cdots R_{N+1} \subset R_N \subset \cdots \subset R_0 = S.$$

Proof: If a signal is invariant to a filter of window length $2(N + 1) + 1$, then each neighborhood of $N + 3$ samples is monotone. Consequently, each neighborhood of length $N + 2$ is monotone and the signal is invariant to a filter window of length $2N + 1$; i.e., $R_{N+1} \subset R_N$. It is trivial to verify that a window of length 1 reproduces any signal exactly upon filtering because the median value of a set containing just one point is the value of that point; thus, $R_0 = S$.

We have established that, for a given filter window $2N + 1$ and a signal set S, there exists a root set R_N of signals invariant to filtering. For a given L-length signal s, we represent the median-filtered version of s by $f_N(s)$ for a $2N + 1$ size window. We represent by $f_N^{(2)}(s)$ the twice filtered signal

[1] It has recently come to our attention that S. Tyan has proven a version of this theorem in an unpublished manuscript. We have not seen a copy of this manuscript and can only speculate as to its contents.

$$f_N^{(2)}(s) = f_N[f_N(s)].$$

We define $f_N^{(n)}(s)$ as the n-times filtered signal

$$f_N^{(n)}(s) = f_N[f_N^{(n-1)}(s)].$$

If $s = f_N(s)$, then s is a root of the filter. We next prove that for any signal s there exists an n such that $f_N^{(n)}(s) = r$ where r is a root.

Suppose that we are given an L-length signal s that is not a root. Recall that N constant points are appended to the beginning of the signal. By construction, the first original signal point is the median of the interval for which it is the central point. As we slide the window from left to right across the signal, the first point to move (i.e., where the window's central point is not the median) must, by definition, be either a point contained in an impulse or oscillation. Suppose that it is an impulse. By construction, an impulse has two constant neighborhoods of equal value on either side, and every point in the impulse is filtered to this constant value by one pass of the filter window. Suppose that the first point to be moved is contained in an oscillation. Let p be the location of the last point unaffected by the median filter, and assume that the filter is centered at this point. Then the leftmost $N + 2$ points must be monotone as seen in the proof of theorem 1. Assume without loss of generality that they are monotone nondecreasing. Assume that the window is now centered at the point $p + 1$. By hypothesis, this point must change in value. Recall that the leftmost N points are not greater in value than the center point. If the N rightmost points were greater than or equal to the center value, then this value at $p + 1$ would be the median. Thus, at least one point to the right of center must have a value less than that at $p + 1$. Thus, there are $N + 1$ points in the window not greater in value than the center point, and the center point changes. Therefore, it changes downward in value. Note that it can never achieve a value less than the value of the immediately preceding constant neighborhood because there are always at least $N + 1$ points contained in the window, including that at $p + 1$ itself, whose values are all greater than or equal to the constant neighborhood.

So we see that the first point that changes under filtering is preceded by, but not necessarily adjacent to, an invariant constant neighborhood, and the point is contained either in an impulse or oscillation. We also see that upon filtering, the value of this point moves closer to the value of the constant neighborhood. There are two possibilities: the value of point p equals the value of point $p + 1$, or the value of point $p + 1$ is greater than that at p. In addition, it can be shown that the value of point $p + 1$ is greater than the value of point p. Suppose that the two points have the same value. As the window increments from position p to $p + 1$, one point moves out of the window on the left side and another point moves into the window on the right. The point that moves out on the left has a value less than or equal to that of point $p + 1$. Because we know that the filtered value at $p + 1$ is less than the original value, the point that moves in on the right side must also have a value less than that at $p + 1$; otherwise, the value at $p + 1$ cannot decrease. If the value of point $p + 1$ is the same as that of p, then there remain N points in the window less than or equal to the value at $p + 1$ (and at p) and there also remain N points in the window greater than or equal to the value at $p + 1$; consequently, point $p + 1$ is the median and would not change. Thus, the value of the first point to change must be greater than its predecessor.

Recall what is known concerning the last consecutive point p that is invariant to filtering. The N points in the window to the left of the center point p are all less than or equal to p in value; the N points to the right of p are all greater than or equal to p in value. When the next point, $p + 1$, is centered in the window, there will be at least N points less than or equal to p in value and at least $N + 1$ points greater than or equal to p in value. Therefore, the median value cannot be less than the value of p. For convenience we summarize this as the following.

Observation 1: The value of the first point to change value during a median-filtering operation must be on the opposite side of its predecessor than the most recent constant neighborhood, and the value of this point upon filtering moves toward the value of its predecessor, but does not move past this value.

Continuing in this fashion, consider the point $p + 2$, which follows point $p + 1$. Note that the value at $p + 2$ is greater than or equal to the value at p. As the window is incremented to the right, $p + 2$ is centered in the window and a point moves out of the window on the left. A new point enters the window on the right. The value of this point must be either greater than that at p or less than or equal to the value at p. If it is less than or equal to the value of p, then there are at least $N - 1$ points in the window with values less than or equal to that at p and at least $N + 1$ points with values greater than or equal to that at p. Consequently, $p + 2$ cannot be filtered to a value less than that at p. If the value of the new point is greater than that at p, then, trivially, the filtered value at $p + 2$ cannot be less than that at p. The same reasoning can be applied to points $p + 3, p + 4, \cdots, p + N$. For convenience, we summarize this as the following.

Observation 2: After filtering, the N rightmost points in the window centered at p must all have values equal to that at p or on the opposite side of the value at p than the most recent constant neighborhood.

Consequently, the value at p is always invariant to median filtering, and, in addition, the same argument applies to any other (invariant) point to the left of p. Also, the point $p + 1$ has one of two possible filtered values, as follows.

Observation 3. Of all the values in the window centered at $p + 1$, the filtered value at $p + 1$ is either the value at p or the closest value to the value at p on the opposite side from the most recent constant neighborhood.

By using an argument similar to that just presented, we reason that the filtered values at $p + 2 \rightarrow p + N$ are greater than or equal to the filtered value at $p + 1$. If the filtered value at $p + 1$ is the same as the value at p, then point $p + 1$ is invariant to filtering on the next pass of the window because it is not greater than the value at p. Suppose, however, that the filtered value at point $p + 1$ is greater than that at p. We must reexamine the prefiltered point values. When $p + 1$ is in the window center, the $N + 1$ rightmost points must all have values greater than that at p including the rightmost point $p + N + 1$. As a result,

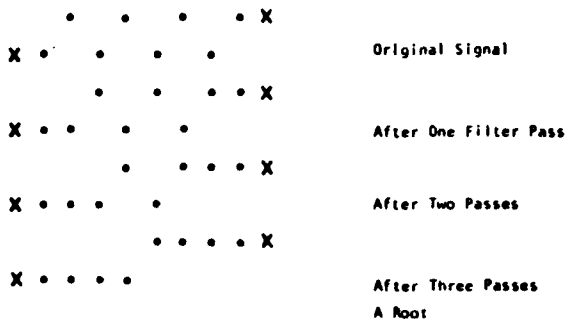


Fig. 2. Result of repeated median filtering.

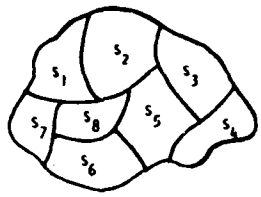


Fig. 3. Partition of the signal space S by eight roots.

when $p + N + 1$ is in the window center, the leftmost $N + 1$ points have values greater than that at p and the filtered value at $p + N + 1$ must be greater than that of p. Consequently, on the second pass of the window, after all the points have been filtered once, when point $p + 1$ is in the window center, the N leftmost points are all less in value than that at $p + 1$, and the rightmost N points all have values greater than or equal to that at $p + 1$. Thus, $p + 1$ is the median of the window and does not change value upon the second filtering. This yields the following.

Observation 4: The first point to change value on a median-filtering operation remains invariant upon additional filter passes.

When the observation is made that the median-filtering operation is independent of whether the window moves from right to left or left to right across the signal, we see that the properties of the first point to change value apply also to the last point in the signal to change value. Because of the appended constant valued points to the front and back of the L-length signal, the first and last signal points are invariant to filtering. Thus, at most, $\frac{1}{2}(L - 2)$ window passes are required to reduce the signal to a root. As a result of the previous discussion, we have the following theorem for an L-length signal.

Theorem 2—Upon successive median-filter window passes, any nonroot signal will become a root after a maximum of $\frac{1}{2}(L - 2)$ successive filterings. Also, any nonroot signal cannot repeat, and the first point to change value on any pass of the filter window will remain constant upon successive window passes.

To illustrate this characteristic of median filtering, consider the binary valued $L = 8$ signal of Fig. 2. This signal will be repeatedly filtered by use of a window length of 3 samples. The appended constant terms are marked with x's. We see that $\frac{1}{2}(L - 2) = 3$ window passes are required to reduce this signal to a root.

To this point, it has always been assumed that the signal is quantized to K levels for an L-length signal. This requirement is not needed because an L-length signal can have, at most, L different values even if the signal samples are not quantized to specific values. Thus, we can always bound K from above by the value of L, and all results stated in this paper apply to unquantized signals.

It should be noted that the value of the appended constant points is not important for the key results of Theorems 1 and 2 to be true with only slight modification to their proofs. It is only important that these values be constant. It is possible to assign nonconstant values to these points such that Theorem 1 does not hold true. Finally, we also note that Theorems 1 and 2 represent median-filter properties that have been observed in the past without proof [4, p. 212].

III. Discussion

The theory developed in the preceding sections provides a number of interesting results. First, we note that every signal in the space of signals, $s \in S$ can be filtered to a unique root with a bounded number of repeated filterings. Thus, the elements of the root set R_N partition S as illustrated in Fig. 3 where it is shown how the signal space is partitioned by a root set with eight elements, whereupon repeated filtering every signal $s \in S_3$ is filtered to root $r_3 \in R_N$ and so on; we will call S_i the ancestor set of root r_i. If a signal s requires L filter passes to reach the root r_3, we say that s is an Lth generation ancestor of r_3. We know from Theorem 2 that any root has, at most, $\frac{1}{2}(L - 2)$ ancestral generations, and we know that the root of a signal depends on the filter window size, i.e., a root for a window of size 3 may not be a root for a window of size 5, although a root for a size 5 window is always a root for a size 3 window. In a loose sense, median filters are a type of low-pass filter with an increasingly narrow passband as the window size increases.

The application of median filtering to signal smoothing problems introduces an interesting twist to the concepts of signal and noise. A simple median filter has no design parameters other than window size, so long as we append N values to each

end in the way discussed. It cannot be designed to accommodate special signal or noise characteristics. In the extreme case, a filter can completely remove a signal component, leaving only noise. It seems desirable that a noise-free signal be a root signal in order that it is invariant to median filtering. If the root signal has added noise, then it may or may not be possible to remove the noise by filtering. Noise that can be filtered is noise that changes the signal in such a way that the noisy signal is an ancestor of the same root. This noise can be removed with repeated median filtering. However, if the noisy signal is now the ancestor of a different root, then it cannot be removed by repeated median filtering. This property of either perfect signal recovery or false signal recovery points to yet another application of this form of median filtering--channel coding. For this application, the root set R corresponds to an alphabet set. The transmitted code can contain either roots or ancestors. In either case, decoding is accomplished through repeated filtering.

In this paper, we have established several fundamental theoretical properties of one form of median filters. We have presented necessary and sufficient conditions for a signal to be invariant to median filtering, and we call these signals roots of the filter. We have also shown that repeated filtering of any signal results in a root signal, and have established the maximum number of filtering operations required to reach a root. As a result of the theory developed in this paper, a better understanding of the potential applications, as well as the limitations of these filters, is achieved.

ACKNOWLEDGMENT

The authors would like to acknowledge the many helpful comments of J. Tukey, which improved the readability of this paper.

REFERENCES

[1] T. S. Huang, G. J. Yang, and G. Y. Yang, "A fast two-dimensional median filtering algorithm," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 13-18, Feb. 1979.

[2] N. S. Jayant, "Average- and median-based smoothing techniques for improving digital speech quality in the presence of transmission errors," *IEEE Trans. Commun.*, vol. COM-24, pp. 1043-1045, Sept. 1976.

[3] L. R. Rabiner, M. R. Sambur, and C. E. Schmidt, "Applications of a nonlinear smoothing algorithm to speech processing," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 552-557, Dec. 1975.

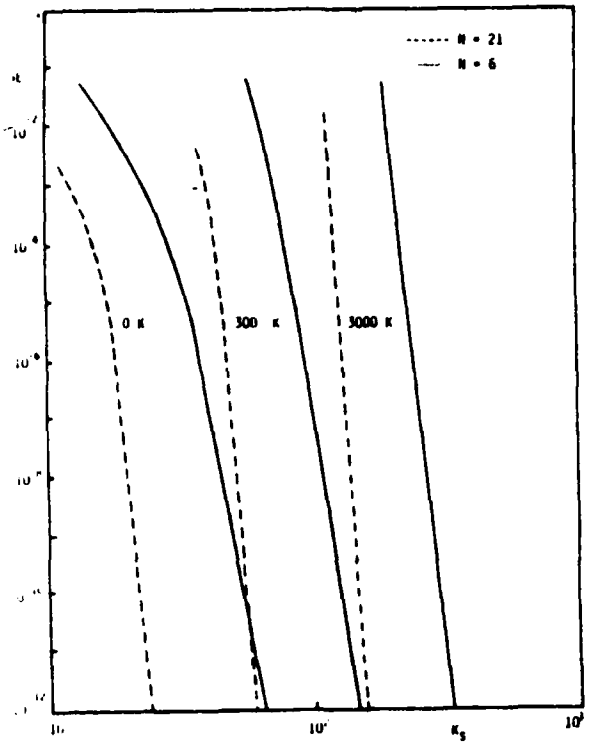
[4] J. W. Tukey, *Exploratory Data Analysis*. Reading, MA: Addison-Wesley, 1977.

Neal C. Gallagher, Jr. (S'72-M'75) received the Ph.D. degree in electrical engineering in 1974 from Princeton University, Princeton, NJ.

After being a member of the faculty of Case Western Reserve University, Cleveland, OH, he joined Purdue University, W. Lafayette, IN, in 1976, where he is an Associate Professor. He has publications in the areas of numerical analysis, digital signal processing, source coding, and optical information processing. He is Past President of the Central Indiana, Central Illinois, Chicago, and South Bend section of the IEEE Information Theory Group. He has consulted for industry and government in the areas of real-time signal processing, spectral estimation, and holography.

Gary L. Wise (S'69-S'72-M'74) was born in Texas City, TX, on July 29, 1945. He received the B.A. degree summa cum laude from Rice University, Houston, TX, in 1971 with a double major in electrical engineering and mathematics, and the M.S.E., M.A., and Ph.D. degrees in electrical engineering from Princeton University, Princeton, NJ, in 1973, 1973, and 1974, respectively.

He is currently an Associate Professor in the Departments of Electrical Engineering and Mathematics at the University of Texas, Austin. His research interests include random processes, statistical communication theory, and signal processing.



7. Frame error probability versus signal pulse count K_s and receiver noise temperature. Thermal-noise limited APD receiver. MAP (weighted linear and quadratic) decoder. N = encoding frames, $M = 100$, $K_b = 1$, $= 1/2$, $R = 50\ \Omega$, $B = 10^9$ Hz, $\epsilon = 0.028$.

es hold

$$P_k < \sum\sum_{y_{r(m)},\, y_{r(m')} \forall r} \prod_{r=1}^{k} P\{y_{r(m)} \mid K_s + K_b\} \cdot P\{y_{r(m')} \mid K_b\} \cdot [\Gamma_k]^{1/2} \quad \text{(A3)}$$

$$< \sum\sum_{\text{all } y_{r(m)},\, y_{r(m')}} \prod_{r=1}^{k} \left[P\{y_{r(m)} \mid K_s + K_b\} P\{y_{r(m')} \mid K_b\}\right]^{1/2} \cdot \left[P\{y_{r(m)} \mid K_b\} P\{y_{r(m')} \mid K_s + K_b\}\right]^{1/2} \quad \text{(A4)}$$

The first inequality holds because $(\Gamma_k)^{1/2}$ is still greater than one, while the second is valid because we are extending the sums over larger set. Inequality (A4) can be rewritten as

$$P_k < \prod_{r=1}^{k} \sum_{\forall y_{r(m)}} \left[P\{y_{r(m)} \mid K_s + K_b\} P\{y_{r(m)} \mid K_b\}\right]^{1/2} \cdot \sum_{\forall y_{r(m')}} \left[P\{y_{r(m')} \mid K_s + K_b\} P\{y_{r(m')} \mid K_b\}\right]^{1/2}$$

$$= \prod_{r=1}^{k} \left\{\sum_{y_r} \left[P\{y_r \mid K_s + K_b\} \cdot P\{y_r \mid K_b\}\right]^{1/2}\right\}^2. \quad \text{(A5)}$$

Due to the fact that the channel is memoryless, we have finally

$$P_k < P_0^k, \quad \text{(A6)}$$

where P_0 is (19) of the text.

References

[1] R. M. Gagliardi, and S. Karp, *Optical Communications*. New York: Wiley, 1976.

[2] M. Ross, *Laser Receivers*. New York: Wiley, 1966.

[3] N. Sorensen, and R. Gagliardi, "Performance of optical receivers with avalanche photodetection," *IEEE Trans. Commun.*, vol. COM-27, pp. 1315–1321, Sept. 1979.

[4] W. Peterson, and E. Weldon, *Error Correcting Codes*, 2nd ed. Cambridge, MA: MIT Press, 1972, p. 70.

[5] R. McIntyre, "The distribution of gains in uniformly multiplying avalanche photodiodes: Theory," *IEEE Trans. on Electron Devices*, vol. ED-19, pp. 703–712, June 1972.

[6] J. Conradi, "The distribution of gains in uniformly multiplying avalanche photodiodes: Experimental," *IEEE Trans. Electron Devices*, vol. ED-19, pp. 713–718, June 1972.

[7] D. Webb, R. McIntyre, and J. Conradi, "Properties of avalanche photodiodes," *RCA Rev.*, vol. 35, pp. 234–278, June 1974.

[8] H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Part I*. New York: Wiley, 1968.

[9] A. J. Viterbi, "Convolutional codes and their performance in communication systems," *IEEE Trans. Commun. Technol.*, vol. COM-19, pp. 751–772, Oct. 1971.

[10] R. Gagliardi and G. Prati, "On Gaussian error probabilities in optical receivers," *IEEE Trans. Commun.*, vol. COM-26, pp. 1742–1747, Sept. 1980.

Properties of Minimum Mean Squared Error Block Quantizers

NEAL C. GALLAGHER, JR., MEMBER, IEEE, AND JAMES A. BUCKLEW

Abstract—Two results in minimum mean square error quantization theory are presented. The first section gives a simplified derivation of a well-known upper bound to the distortion introduced by a k-dimensional optimum quantizer. It is then shown that an optimum multidimensional quantizer preserves the mean vector of the input and that the mean square quantization error is given by the sum of the component variances of the input minus the sum of the variances of the output.

I. Introduction

Block or vector quantization deals with the representation of multidimensional elements with a finite discrete set of values. The values to be quantized may naturally fall into a k-dimensional representation; typical examples are complex numbers, positional coordinates, or state vectors. In other cases, k-dimensional vectors are formed from blocks of k samples taken from one-dimensional signals. In 1964 Zador published a number of very interesting results on the properties of optimal block quantizers for the rth moment Euclidean norm distortion measure [1]. Among Zador's contributions are the derivation of both upper and lower bounds on the distortion introduced by the optimal quantizer. These bounds are derived without actually finding the optimal quantizer. Unfortunately, at some points Zador's development is not easy to follow, and alternate derivations and extensions by Gersho [2] and Yamada *et al.* [3] have recently appeared. In Section II we present an alternate derivation of Zador's random quantization upper bound not treated in either [2] or [3].

In [4] Bucklew and Gallagher show that for one-dimensional mean squared error distortion the optimum quantizer has the property that the mean value of the quantizer output equals the mean value of the input and also that the mean square quantization error equals the variance of the input minus the variance of the output. In [5] Bucklew and Gallagher prove that the same results hold for constant step-size minimum mean squared error quantizers. In Section III we extend these properties to k-dimensional optimal block quantizers.

Manuscript received December 30, 1980. This work was supported by the Air Force Office of Scientific Research under Grant AFOSR 78-3605.

N. C. Gallagher, Jr., is with the School of Electrical Engineering, Purdue University, West Lafayette, IN 47907.

J. A. Bucklew is with the Department of Electrical and Computer Engineering, University of Wisconsin, Madison, WI 53705.

0018-9448/82/0100-0105$00.75 ©1981 IEEE

II. Random Quantization Upper Bound

In [2] Gersho provides a very readable derivation of Zador's expression for quantizer distortion. To improve continuity and readability we employ Gersho's notation. The quantizer input is a k-dimensional random vector x in $\mathcal{R}_k$ which is quantized to one of N levels $y_1, y_2, \cdots, y_N$ in $\mathcal{R}_k$. The space $\mathcal{R}_k$ is partitioned into N disjoint and exhaustive regions $S_1, S_2, \cdots, S_N$. The quantizer is defined by the function $Q(x)$ defined by $Q(x) = y_i$, if $x \in S_i$. Note that this definition does not require that $y_i \in S_i$, although in practice y_i is usually contained in S_i. The performance of the quantizer is measured by the distortion

$$D = \frac{1}{k} E\{\|X - Q(X)\|^r\}$$

where $\|\cdot\|$ denotes the usual Euclidean distance norm, the operator $E\{\cdot\}$ denotes statistical expectation, and the input X is a k-dimensional random input vector. The case where $r = 2$ is the usual mean squared distortion. The expression derived by Zador and Gersho for the minimum distortion D_0 obtained by use of the best quantizer is

$$D_0 = N^{-r/k} C(k, r) \| p(x) \|_{k/(k+r)}, \tag{1}$$

where

$$\| p(x) \|_a = \left[\int [p(x)]^a \, dx \right]^{1/a},$$

and where the constant $C(k, r)$, called the coefficient of quantization, is independent of the density $p(x)$ and is in general unknown. This expression is an asymptotic result valid only for large N. Two special cases for which the value of $C(k, r)$ is known exactly are [2]

$$C(1, r) = \frac{1}{r+1} 2^{-r},$$

and

$$C(2, 2) = \frac{5}{36\sqrt{3}}.$$

Consider the density $p(x)$ having a constant value of one over the unit volume hypercube; then $\| p(x) \|_{k/(k+r)} = 1$. In this case (1) becomes

$$D_0 = N^{-r/k} C(k, r). \tag{2}$$

So, we see that by finding a bound on D_0 we also bound $C(k, r)$. To find this bound we choose the quantizer output levels to have a random distribution uniformly distributed over the hypercube. For a particular input value x, we find the closest output level and quantize to that value. Because this quantizer is not the optimum quantizer, the associated distortion will bound from above the distortion for the optimum quantizer.

To begin, place at random N independent uniformly distributed k-dimensional samples in the hypercube. These will be the output levels. We take the quantizer input X to have a uniform distribution over the hypercube. We also assume that N is sufficiently large so that there is a very small probability that the quantizer input is closer to an edge of the hypercube than to one of the output values. Suppose that an input value x has arrived and is sitting in the hypercube waiting to be quantized. The probability that one particular output value is within a distance ρ of this input sample is given approximately by the volume of a sphere of radius ρ about that sample point, or

$$\Pr(\text{one particular output level is within } \rho \text{ of the input sample}) = V_k \rho^k,$$

where if V_k is volume of the unit radius sphere, then $V_k \rho^k$ is the volume of the sphere with radius ρ. We are interested in the closest output level to the input sample. To compute the probability that the closest output level is within a distance ρ of the input sample, we combine classical order statistics with the result found in [3]. By employing this approach, we compute the probability density $f(\rho)$ for the distance between the input sample and the nearest output level to be

$$f(\rho) = N\left[1 - V_k \rho^k\right]^{N-1} V_k k \rho^{k-1}.$$

Note that for large values of N this probability density goes to zero rapidly as ρ increases. By construction $\rho = \| x - y_i \|$, where x is the input value and y_i is the output value. Consequently,

$$E\{\|X - Q(X)\|^r\} = E\{\rho^r\};$$

so,

$$D = \frac{1}{k} E\{\rho^r\}$$

$$= \frac{1}{k} \int_{\text{hypercube}} \rho^{r+k-1} N\left[1 - V_k \rho^k\right]^{N-1} k V_k \, d\rho.$$

Make the change of variables $s = V_k \rho^k$ and use the fact that $s \le 1$ to write

$$D \le \frac{N}{k V_k^{r/k}} \int_0^1 s^{r/k} [1 - s]^{N-1} \, ds = \frac{N}{k V_k^{r/k}} \frac{\Gamma\left(1 + \frac{r}{k}\right)\Gamma(N)}{\Gamma\left(N + 1 + \frac{r}{k}\right)},$$

where $\Gamma(\cdot)$ is the gamma function. For large N the following approximation is valid:

$$\frac{\Gamma(N)}{\Gamma\left(N + \frac{(k+r)}{k}\right)} \simeq N^{(k+r)/k}.$$

Therefore,

$$D = \frac{N^{-r/k} \Gamma\left(1 + \frac{r}{k}\right)}{k V_k^{r/k}}.$$

Because $D \ge D_0$, we use (2) to write

$$C(k, r) \le \frac{\Gamma\left(1 + \frac{r}{k}\right)}{k V_k^{r/k}},$$

which is Zador's random quantization upper bound.

III. Moment Properties of Optimum Quantizers

In [4] and [5] it is shown that, for minimum mean squared error one-dimensional quantizers, the mean of the input equals the mean of output and the distortion equals the variance of the input minus the output variance. These properties are shown to apply with and without the equal step-size constraint. In this section we generalize these results to the k-dimensional case.

We are interested in the properties of quantizers designed to minimize the distortion defined by (2) for $r = 2$:

$$D = \frac{1}{k} E\{\|X - Q(X)\|^2\}.$$

Many constraints we might impose on the quantizer can be imposed by the functional form of $Q(x)$; for example, the k-dimensional version of the equal step-size condition might require the regions $S_1 S_2, \cdots, S_N$ to have equal volume and be congruent. We had originally employed a variational approach to obtain the results of this section; however, an alternate approach, suggested by an anonymous reviewer, provides more intuition into quantizer structure. So, we employ his method.

To begin, we define the parameters P_i and X_i as follows:

$$P_i = \int_{S_i} P(x)\,dx,$$

and

$$x_i = P_i^{-1} \int_{S_i} xP(x)\,dx. \tag{3}$$

We note that partition $\{S_i\}_{i=1}^N$ need not be the optimum partition. Consider two different quantizers defined over the partition $\{S_i\}_{i=1}^N$: one with output value X_i and one with output value Y_i. These quantizer functions are represented as $Q_0(X) = X_i$ and $Q(X) = Y_i$, respectively. It will be shown that the quantizer $Q_0(X)$ is optimum for the given partition. We have that

$$E\{\|X - Q(X)\|^2\} = \sum_{i=1}^{N} \int_{S_i} (x - x_i + x_i - y_i)^2 P(x)\,dx. \tag{4}$$

By (3), we have

$$\int_{S_i} (x - x_i)(x_i - y_i)P(x)\,dx = 0;$$

therefore, (4) becomes

$$E\{\|X - Q(X)\|^2\} = E\{\|X - Q_0(X)\|^2\} + \sum_{i=1}^{N} P_i \|x_i - y_i\|^2. \tag{5}$$

The expression in (5) illustrates that the quantizer $Q_0(X)$ produces an error no larger than any other quantizer $Q(X)$ for a given partition. Also, by (3) we see that the mean of the quantizer outputs equals the mean value of the input; this follows by

$$\sum_{i=1}^{N} P_i x_i = \sum_{i=1}^{N} \int_{S_i} xp(x)\,dx = \int xP(x)\,dx, \tag{6}$$

where the left side is the mean of the output and the right side the mean of the input. It can also be shown that the quantizer error equals the variance of the input minus the variance of the output. Consider the input variance

$$\begin{aligned} E\{\|X - E\{X\}\|^2\} &= E\{\|X - Q_0(X) + Q_0(X) - E\{X\}\|^2\} \\ &= E\{\|X - Q_0(X)\|^2\} + E\{\|Q_0(X) - E\{X\}\|^2\}, \end{aligned} \tag{7}$$

where as before the cross terms are zero. The right side of (7) is simply the sum of the quantizer error and the output variance.

Equations (6) and (7) specify the first and second moment properties of the optimum quantizer; these properties follow regardless of the optimality of the partition. In addition, it is noteworthy that the optimum quantizer is not unique. A simple example serves to illustrate this point. Consider a two-dimensional circularly symmetric input density. Any rotation of a minimum error quantizer is also a minimum error quantizer. The same property holds for one-dimensional quantizers, where it is possible to have more than one minimum error quantizer.

VII. Summary

This correspondence contains two results dealing with the properties of k-dimensional minimum mean squared error quantizers. We have established necessary conditions for optimum quantizers. These conditions are used to show that for k-dimensional quantizers the mean value of the input is preserved in the output and that the mean squared error equals the input variance minus the output variance. Also, a simplified derivation of Zador's random quantization upper bound is developed.

References

[1] P. Zador, "Development and evaluation of procedures for quantizing multivariate distributions," Ph.D. dissertation, Stanford Univ., Stanford, CA, 1964; University Microfilms Inc. no. 64-9855.

[2] A. Gersho, "Asymptotically optimal block quantization," *IEEE Trans. Inform. Theory*, vol. IT-25, pp. 373–380, July 1979.

[3] Y. Yamada, S. Tazaki, and R. M. Gray, "Asymptotic performance of block quantizers with difference distortion measures," *IEEE Trans. Inform. Theory*, vol. IT-26, pp. 6–14, Jan. 1980.

[4] J. A. Bucklew and N. C. Gallagher, "A note on optimal quantization," *IEEE Trans. Inform. Theory*, vol. IT-25, pp. 365–366, May 1979.

[5] "Some properties of uniform step size quantizers," *IEEE Trans. Inform. Theory*, vol. IT-26, pp. 610–613, Sept. 1980.

The Binary Multiplying Channel—A Coding Scheme that Operates Beyond Shannon's Inner Bound Region

J. PIETER M. SCHALKWIJK, SENIOR MEMBER, IEEE

Abstract—Blackwell's binary multiplying channel is well known as an example of a two-way channel for which Shannon's inner and outer bounds to the capacity region differ. A deterministic coding scheme is given which outperforms the inner region for this channel. Dueck had earlier obtained an analogous result for another type of two-way channel.

I. Introduction

Shannon [1] derived inner and outer bounds to the capacity region of the two-way channel (TWC). A TWC (see Fig. 1) is a discrete memoryless channel with finite input and output alphabets and defined by a matrix $[P(y_1, y_2 \mid x_1, x_2)]$ of transition probabilities. Shannon's inner bound region equals the convex hull of the region of rate pairs $(I(X_1; Y_2 \mid X_2), I(X_2; Y_1 \mid X_1))$, where the input distribution $P(x_1, x_2)$ is allowed to vary over all product distributions $P(x_1, x_2) = P(x_1)P(x_2)$. Likewise, the Shannon outer bound is the convex hull of the region of rate pairs $(I(X_1; Y_2 \mid X_2), I(X_2; Y_1 \mid X_1))$, where the input distribution $P(x_1, x_2)$ is no longer restricted to be of the product type.

Blackwell's binary multiplying channel (BMC), which is a TWC satisfying $Y_1 = Y_2 = X_1X_2$, is an example of a simple TWC for which the inner and outer regions differ. In Fig. 8 we have reproduced from [1] the boundary G_i of the inner region and the boundary G_o of the outer region for the BMC. (See [1] for explicit equations specifying these regions.) We show that each point on the third curve in Fig. 8 can be achieved by a certain deterministic coding scheme. Consequently the inner region for the BMC is not the capacity region. (An analogous result had been obtained earlier by Dueck [2] for a TWC which was not a BMC.) For the sake of simplicity, in the next section we first describe the coding scheme which achieves the point on our curve for which $R_1 = R_2$.

II. The Coding Strategy

The senders try to send information that without loss of generality can be taken as the location of a subinterval [3], [4], of

Manuscript received October 6, 1980; revised April 15, 1981. This paper was presented at the 1981 International Symposium on Information Theory, Santa Monica, CA, Feb. 9–12.

The author is with the Department of Electrical Engineering, Eindhoven University of Technology, Den Dolech 2, P.O. Box 513, 5600 MB Eindhoven, The Netherlands.

0018-9448/82/0100-0107$00.75

NONUNIFORM MULTIDIMENSIONAL QUANTIZATION

by

Kerry D. Rines

TASC
8301 Greensboro Dr., Suite 1200
McLean, VA 22102

and

Neal C. Gallagher, Jr.

School of Electrical Engineering
Purdue University
West Lafayette, IN 47907

and

James A. Bucklew

Department of Electrical and Computer Engineering
University of Wisconsin
Madison, Wisconsin 53705

We have shown in a previous paper that an optimum quantizer can be designed for the random vector $\underline{X}$, when $\underline{X}$ is uniformly distributed. However, finding an optimum quantizer when $\underline{X}$ has an arbitrary density function is in general very difficult. Thus in this paper we consider the design of near-optimum quantizers for $\underline{X}$ when the density is nonuniform. The results show that if we allow the number of quantization levels to be large, we can obtain a distortion performance arbitrarily close to the distortion of the optimum quantizer. The results also provide a useful tool for the companding design of optimum quantizers discussed later in the paper.

I. Introduction

A number of authors [1]-[4] have examined the advantages of multidimensional quantization over univariate quantization. Unfortunately multidimensional quantizers are difficult to design and must usually be implemented using a search procedure. The disadvantage of a search implementation is that the storage and computation requirements increase with the number of quantization levels and the dimension of the quantizer. In a previous paper [5], we present a method called prequantization for the design of optimum uniform multidimensional quantizers without the drawbacks of a search. Now in this paper we extend Bennett's companding results [6] to k-dimensions for the design of nonuniform multidimensional quantizers. These new methods also avoid problems associated with a search.

II. Piecewise Companding

Let $p(\underline{x})$ be the probability density function (pdf) of the vector $\underline{X}$. We begin by constructing a density $g(\underline{x})$ that is a piecewise constant approximation of $p(\underline{x})$. Let S be the compact support of both $p(\underline{x})$ and $g(\underline{x})$. We partition S into M compact regions each denoted by C_i and with area (measure) m_i for $i = 1,2,\ldots,M$. The density is then defined as

$$g(\underline{x}) = \frac{p_i}{m_i} \qquad \forall \underline{x} \in C_i \quad , \quad i = 1, 2, \ldots, M$$

where

$$p_i \triangleq \int_{C_i} p(\underline{x})d\underline{x} \ .$$

Now compare the quantization of the random vectors $\underline{X}$ and $\underline{Y}$ where $\underline{Y}$ has the density $g(\underline{x})$. We define Q_o as the optimum quantizer for $\underline{X}$ given $p(\underline{x})$ and Q_g as the optimum quantizer for $\underline{Y}$ given $g(\underline{x})$. Zador's equation for the minimum per sample distortion of $\underline{X}$ is

$$D_o = \frac{1}{k} E(||\underline{X}-Q_o(\underline{X})||_2^r) = C(k,r)N^{-r/k}||p||_{k/k+r} \quad (1)$$

where

$\underline{X}$ k-dimensional vector

$Q_o(\underline{X})$ quantized output of Q_o

N number of quantization levels (assumed

C(k,r) constant dependent only on k and r

$$||p||_\alpha \left[\int p(\underline{x})^\alpha d\underline{x}\right]^{1/\alpha} .$$

Similarly the optimum distortion corresponding to the random vector $\underline{Y}$ is

$$D = \frac{1}{k} E(||\underline{Y}-Q_g(\underline{Y})||_2^r) = C(k,r)N^{-r/k}||g||_{k/k+r} \ . \quad (2)$$

Using (1) and (2) and Bennett's integral for mismatched quantizers, we can show that a near-optimum quantizer for the random vector $\underline{X}$ can be designed by finding an optimum quantizer for a random vector with the density $g(\underline{x})$. As the approximation of $p(\underline{x})$ by $g(\underline{x})$ becomes more accurate ($M \to \infty$), the distortion approaches the optimum distortion. Given this background, we now examine the design of optimum quantizers for random vectors with piecewise constant densities.

We design the optimum quantizer for the density $g(\underline{x})$ by finding the number of quantization levels that must be assigned to each partition C_i. The

This work was presented at the 1982 Conference on Information Sciences and Systems, Princeton University.

first step is to rewrite Zador's distortion equation in (2) as

$$D = \sum_{i=1}^{M} p_i\, C(k,r) \left[\frac{m_i^{k/k+r}}{N\, p_i^{k/k+r}} \sum_{j=1}^{M} p_j^{\frac{k}{k+r}}\, m_j^{\frac{r}{k+r}} \right]^{r/k} . \quad (3)$$

Second, we examine the minimum distortion D_i in each partition C_i. The distortion is defined as

$$D_i \triangleq \frac{1}{k} E\{ \|\underline{X} - Q_g(\underline{X})\|_2^r \mid \underline{X} \epsilon C_i \} .$$

We can write the density for each partition as

$$g_i(\underline{x}) = g(\underline{x} \mid \underline{x} \epsilon C_i) = \frac{1}{m_i} \quad ; \quad \underline{x} \epsilon C_i$$

$$= 0 \quad ; \quad \underline{x} \notin C_i .$$

Let N_i be the number of quantization levels assigned to the partition C_i. We note that the total number of quantization levels is N and therefore

$$\sum_{i=1}^{M} N_i \leq N . \quad (4)$$

Again using Zador's expression we find

$$D_i = C(k,r)\, N_i^{-r/k}\, \|g_i(\underline{x})\|_{k/k+r}$$

$$= C(k,r)\, N_i^{-r/k} \left[\frac{1}{m_i^{k/k+r}}\, m_i \right]^{\frac{k+r}{k}}$$

$$= C(k,r) \left(\frac{m_i}{N_i}\right)^{r/k} . \quad (5)$$

Since the density function of each partition is uniform, we can achieve the optimum distortion in (5) by using the optimum k-dimensional uniform quantizers described in [5]. The total distortion D_T can be written as the expected value of the D_i's in (5) and thus

$$D_T = \sum_{i=1}^{M} p_i D_i$$

$$= \sum_{i=1}^{M} p_i C(k,r) \left(\frac{m_i}{N_i}\right)^{r/k} . \quad (6)$$

Recall that D in (3) represents the optimum quantization distortion. Thus by setting $D_T = D$ we can solve for the optimum assignment of the quantization levels N_i. One solution is

$$\frac{m_i}{N_i} = \frac{m_i^{k/k+r}}{N\, p_i^{k/k+r}} \sum_{j=1}^{M} p_j^{k/k+r}\, m_j^{r/k+r}$$

and therefore

$$N_i = \frac{N\, p_i^{k/k+r}\, m_i^{r/k+r}}{\sum_{j=1}^{M} p_j^{k/k+r}\, m_j^{r/k+r}} . \quad (7)$$

We now have a method called piecewise companding for designing near-optimum quantizers for $\underline{X}$ given $p(\underline{x})$. With this method the support S is first partitioned into M regions and then each region is quantized using an optimum uniform multidimensional quantizer with the number of quantization levels specified in (7).

III Optimal Companding

A number of important properties of optimal companding have been examined in the literature. However, to the authors' knowledge an example of an optimum k-dimensional compander has never been presented. In this section we construct an optimum 2-dimensional quantizer using companding. The example adds insight into the companding problem and suggests general guidelines for the companding design of optimum k-dimensional quantizers.

Bennett [6] was the first to use companding to design a nonuniform 1-dimensional quantizer. The structure of a typical companding system is shown in Figure 1. The input is first compressed by the nonlinearity f(x) and quantized with a uniform quantizer. The uniformly quantized value is then expanded by the nonlinearity $f^{-1}(x)$. Bennett's work was later extended by Panter and Dite [7]. Panter and Dite derive an expression that can be used to design the optimum companding functions given the input density function and assuming N is large. As a result it is a relatively simple task to design a companding system for an optimum nonuniform 1-dimensional quantizer.

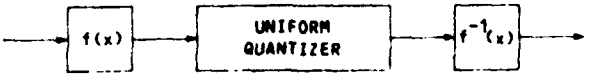


Figure 1 Typical companding system.

In [8] Bucklew shows that the companding design can be extended to k dimensions. For k dimensions, the uniform quantizer in Figure 1 becomes the optimum uniform k-dimensional quantizer. Similarly the compressor and expander functions become k-dimensional invertible nonlinearities. Bucklew shows that the optimum compressor and expander functions must be conformal almost everywhere. As it turns out, this restriction severely limits our ability to design optimum companding systems. However, using the results of Section II and the idea of conformality, we can construct an example of an optimum compander.

In practice we would be given a density function and asked to design the optimum compander. To construct this example we consider the problem in reverse. First we choose a compander that satisfies the conformality constraints and then we find the probability density function for which the compander is optimum.

Let (U,V) be a random vector with the density function p(u,v). For convenience let the support of p(u,v) be the set $S = \{(u,v): 1 \leq u^2 + v^2 \leq e^{2\pi}, v > 0\}$ as shown in Figure 2. Now consider the 2-dimensional conformal map $W = e^Z$ where $w = u + iv$ and $z = x + iy$. We define the compressor function as

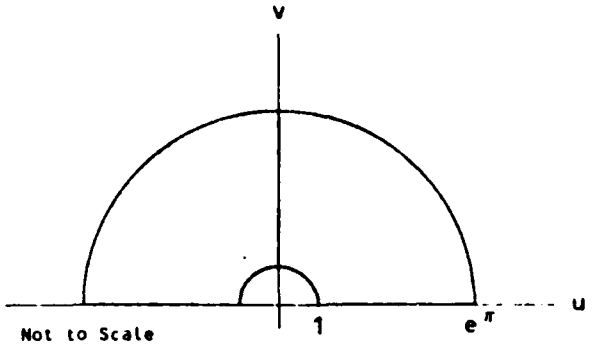


Figure 2 Support of density p(u,v).

$$x = \ln \sqrt{u^2 + v^2}$$
$$y = \tan^{-1} \frac{v}{u} \tag{8}$$

and we define the expander function as

$$u = e^x \cos y$$
$$v = e^x \sin y \, . \tag{9}$$

The vector (U,V) is mapped into the square $\times_{i=1}^{2} (0,\pi)$ in the Z-plane by the compressor function in (8). The resulting vector (X,Y) is quantized using the optimum uniform 2-dimensional (hexagonal) quantizer. Then the output from the hexagonal quantizer is mapped back into the W-plane using the expander function in (9). We now must find the density p(u,v) for which this quantizer is optimum.

We begin with a piecewise companding design for the unknown density p(u,v). The support S of p(u,v) is partitioned into M equal-sized regions C_i, each with area $m_i = \Delta u \Delta v$. Using (7), the optimum number of quantization levels for each partition is given by

$$N_i = N \frac{p_i^{1/2}}{\sum_{j=1}^{M} p_j^{1/2}} \tag{10}$$

where

$$p_i = \int_{C_i} p(u,v)dudv \, .$$

We implement the piecewise compander as follows. First, we find the partition that contains the random vector (U,V). Then for each partition C_i, (U,V) is quantized using an N_i-level hexagonal quantizer.

We compare this implementation of the piecewise compander with the companding system described in (8) and (9). The compressor function in (8) maps each partition C_i into a new partition C_i' in the $\times_{i=1}^{2} (0,\pi)$ square of the Z-plane. The partition C_i' is then quantized using the hexagonal quantizer discussed above. Let N_i' be the number of quantization levels contained within C_i'. For N large, we can consider the hexagonal quantization levels to be uniformly distributed within $0 \le x,y \le \pi$. Thus, N_i' will given by the ratio of the area of C_i' to the total area of the square. If we let $m_i' = k_i \Delta x \Delta y$ be the area of C_i', the number of levels N_i' is given by

$$N_i' = \frac{N}{\pi^2} k_i \Delta x \Delta y \, . \tag{11}$$

The expander function in (9) maps the N_i' quantization levels in C_i' into the partition C_i in the W-plane. Since the mapping is nonlinear, the quantization levels will no longer be in the form of a hexagonal lattice. However, the quantization will be approximately hexagonal when the area of C_i is small.

We now assume there exists a density p(u,v) continuous almost everywhere, such that the number of quantization levels N_i in (10) is equal to N_i'. Thus for N large and $\Delta u \Delta v$ small, the distortion of the companding system in (8) and (9) is approximately equal to the distortion of the piecewise compander. Setting $N_i = N_i'$ we obtain

$$\frac{N}{\pi^2} k_i \Delta x \Delta y = \frac{N p_i^{1/2}}{\sum_{j=1}^{M} p_j^{1/2}} \, . \tag{12}$$

We can rewrite this expression as follows. As stated above that the compressor function in (8) maps C_i onto C_i' for all i. Then by definition,

$$\int_{C_i} J_{xy}(u,v)dudv = \int_{C_i'} dxdy$$
$$= k_i \Delta x \Delta y$$

where $J_{xy}(u,v)$ is the Jacobian of the transformation in (8). Using this result and the definition of p_i in (10), we can rewrite (12) as

$$\int_{C_i} J_{xy}(u,v)dudv = \frac{\pi^2 [\int_{C_i} p(u,v)dudv]^{1/2}}{\sum_{j=1}^{M} [\int_{C_j} p(u,v)dudv]^{1/2}} \, . \tag{13}$$

Recall from section II that in the limit as $M \to \infty$ and $\Delta u \Delta v \to 0$, the distortion of the piecewise companding system approaches the distortion of the optimum quantizer for p(u,v). We can also show that for this same limiting relation, the distortion of the companding system in (8) and (9) is equal to the distortion of the piecewise compander. Therefore, the companding system in (8) and (9) will be an optimum quantizer for the density p(u,v) that satisfies (13) in the limit as $\Delta u \Delta v \to 0$. Dividing both sides of (13) by $\Delta u \Delta v$ and taking the limit as $\Delta u \Delta v \to 0$ we find

$$J_{xy}(u,v) = \frac{\pi^2 p^{1/2}(u,v)}{\int_S p^{1/2}(u,v)\,du\,dv}$$

$$= K\, p^{1/2}(u,v) \qquad (14)$$

where K is a constant.

Computing the Jacobian of the compander in (8), we find the companding system is optimal for the density

$$p(u,v) = \frac{2/(\pi - \pi e^{-2\pi})}{(u^2+v^2)^2} \; ; \; 1 \le u^2 + v^2 \le e^{2\pi} \; , \; v > 0$$

$$= 0 \qquad ; \text{elsewhere} .$$

IV. Summary

We have discussed the design of optimum and near-optimum quantizers for random vectors with nonuniform density functions. For the design of near-optimum quantizers a piecewise companding approach was presented. While not optimum, quantizers using piecewise companding can be designed for random vectors having ay given k-dimensional density function.

The use of k dimensional companding systems for optimum nonuniform quantization was also examined. Extending the results in (14) we find that a necessary condition on the Jacobian of the optimum compressor function is

$$J_{\underline{x}}(\underline{u}) = K\, p^{k/k+r}(\underline{u})$$

where r is the power of the distortion measure. While these results add to our understanding of optimal companding, they also suggest that it may be possible to design an optimum companding system for all but a few k dimensional densities. This further underscores the importance of the piecewise companding technique.

References

[1] P. Zador, Development and Evaluation of Procedures for Quantizing Multivariate Distributions, Ph.D. Dissertation, Stanford University, 1964, University Microfilm No. 64-9855.

[2] A. Gersho, "Asymptotically Optimal Block Quantization," IEEE Trans. on Inform. Theory, Vol. IT-25, pp. 373-380, July 1979.

[3] Y. Yamada, S. Tazaki, and R. M. Gray, "Asymptotic Performance of Block Quantizers with Difference Distortion Measures," IEEE Trans. on Inform. Theory, Vol. IT-26, pp. 6-14, January 1980.

[4] J. A. Bucklew, "Upper Bounds to the Asymptotic Performance of Block Quantizers," to appear in IEEE Trans. on Inform. Theory.

[5] K. D. Rines and N. C. Gallagher, Jr., "The Design of Multidimensional Quantizers using Prequantization," Proceedings of the Eighteenth Annual Allerton Conference, pp. 446-453, October 1980.

[6] W. R. Bennett, "Spectra of Quantized Signals," B.S.T.J., Vol. 27, pp. 446-472, July 1948.

[7] P. F. Panter and W. Dite, "Quantization in Pulse-Count Modulation with Nonuniform Spacing of Levels," Proc. IRE, vol. 39, pp. 44-48, 1951.

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

The authors wish to acknowledge partial support by the Air Force Office of Scientific Research under grant AFOSR-78-3605.

A Note on the Computation of Optimal Minimum Mean-Square Error Quantizers

J. A. BUCKLEW AND N. C. GALLAGHER, JR.

Abstract—**This paper considers the problems associated with computing optimal minimum mean-square error quantizers. Most computational methods in current use are iterative. These iterative schemes are extremely sensitive to initial conditions. Various methods of obtaining good initial conditions are presented and discussed.**

I. INTRODUCTION

In his classic paper of 1960, Max presents an iterative scheme for the computation of one-dimensional minimum mean-squared error quantization characteristics [1]. In addition, he solves for the optimum Gaussian quantizer for up to 36 output levels. In [2], Gallagher uses Max's method in the computation of optimum Rayleigh quantizer parameters, and in [3] Paez and Glisson use the same method to compute the optimum Laplacian quantizer later recomputed by Adams and Giesler [4]. Max's algorithm is very simple to program into a digital computer, and we view this simplicity as a good reason for using his method. However, one problem that arises with this algorithm is its failure to always converge to the optimum solution when the number of quantizer output levels is large. The reason for this is that the initial guess for starting the iteration must be increasingly precise as the number of quantizer levels becomes large. So, for a 64-level quantizer, Max's algorithm will not converge to the optimum solution unless the initial guess for the first output level is very close to the true value. This difficulty has prompted others to employ more sophisticated optimization methods in the solution for optimum quantizers. For example, Pearlman and Senge [5] use a vector space optimization technique that is a combination of the steepest descent and Newton-Raphson methods to solve for the optimum Rayleigh quantizer. It is not our purpose to detract from this and similar methods that do work well, but in our view, if the starting point problem can be solved, Max's method is the preferred method of solution. In Section II we discuss several methods for choosing the iteration's initial condition very accurately, and we have demonstrated convergence of Max's algorithm for at least 10 000 output levels and present numerical examples in Section III.

Paper approved by the Editor for Data Communication Systems of the IEEE Communications Society for publication without oral presentation. Manuscript received January 5, 1981; revised April 27, 1981. This work was supported by the Air Force Office of Scientific Research under Grant AFOSR 78-3605.

J. A. Bucklew is with the Department of Electrical and Computer Engineering, University of Wisconsin, Madison, WI 53706.

N. C. Gallagher, Jr. is with the School of Electrical Engineering, Purdue University, West Lafayette, IN 47907.

II. THE COMPUTATION OF OPTIMUM ONE-DIMENSIONAL QUANTIZERS

A common method for implementing one-dimensional quantizers is the companding method as discussed by Smith [6]. The companding method is straightforward: the input signal x with probability density $p(x)$ first enters the invertible nonlinearity $g(x)$, called the compressor; then it goes into a uniform quantizer over the range $[0, 1]$, and upon reconstruction it passes through the expansion nonlinearity $g^{-1}(x)$. For minimum mean-squared error quantization, the asymptotically optimum compressor function is given by

$$g(x) = \left[\int_{-\infty}^{\infty} [p(y)]^{1/3} dy\right]^{-1} \int_{-\infty}^{x} [p(y)]^{1/3}\, dy. \quad (1)$$

In Max's classic 1960 paper an iterative method is presented whereby the exact quantizer parameters can be computed for finite N.

Max's algorithm provides a method for the solution of the equations

$$e_i = (y_i + y_{i-1})/2, \qquad i = 2, \cdots, N \quad (2a)$$

and

$$\int_{e_i}^{e_{i+1}} (x - y_i) p(x)\, dx = 0, \qquad i = 1, \cdots, N \quad (2b)$$

where the output levels of the quantizer are denoted $y_1, y_2, \cdots, y_N$ and the internal breakpoints as $e_1, e_2, \cdots, e_{N+1}$. Typically, endpoint values e_1 and e_{N+1} are known *a priori* and the first step of Max's procedure is to choose a value for y_1 with which to solve (2b) for the value e_2. We then use this value in (2a) to find y_2 and use this to find e_3 in (2b), and so on. The last integral over (e_N, e_{N+1}) can be used to determine

0090-6778/82/0100-0298$00.75

the accuracy of the initial guess for y_1. If the last integral is zero within a specified error, we use the computed parameters to specify the quantizer; if not, we make a new guess for y_1 and begin the procedure again. Details on how to modify the initial guess for y_1 are not specified by Max.

We have computed quantizers using Max's method for several densities. It has been our observation that the convergence properties of Max's algorithm are greatly dependent on the initial guess for y_1. Let y_{1N} denote the first output level for an optimum N level quantizer. Intuitively, if the first guess at y_{1N} (call it $\hat{y}_{1N}$) is very close to $y_{1(N+1)}$, then Max's algorithm tries to converge to the $N+1$ level quantizer. A consideration of Max's method indicates that the first N steps of the algorithm are the same for the N or $N+1$ level quantizers. Although never reported in the literature, it is our understanding that this phenomenon has been widely observed [7].

As an aside, we remark that the conditions presented in (2) are not sufficient conditions to specify the optimum quantizer; they are only necessary. However, in 1965 Fleisher [8] showed that if

$$\frac{d^2}{dx^2}\left[\ln p(x)\right] < 0$$

then the expressions in (2) are both necessary and sufficient for the specification of the minimum mean-squared error quantizer, and their solution provides us with the unique optimum quantizer.

We now describe two similar methods for generating a good initial condition. First, note that the initial condition can be a guess at the value for y_1 or a guess for the value of any y_i, $i = 1, \cdots, N$ wherever we choose to begin the iteration. The first method is a modified version of an estimation method by Panter and Dite [9] and Roe [10]. The second method employs a companding model to produce the iteration starting point. Both methods grow more precise as the number of quantization levels N increases. Each method, however, requires computation to generate an initial value; the complexity of this computation varies depending on the distribution of the variable to be quantized.

In the first method we use the asymptotic level density $\lambda(x)$ for the minimum mean-squared error quantizer. $\lambda(x)\Delta x$ is approximately the ratio of the number of output levels in a region Δx about x to the total number of output levels N. This function is the first derivative of the compressor function $g(x)$ in (1):

$$\lambda(x) = g'(x) = \left[p(x)\right]^{1/3}\left[\int_{-\infty}^{\infty}\left[p(y)\right]^{1/3}dy\right]^{-1} \tag{3}$$

Smith [6] shows that this function has the property that for adjacent output levels y_i and y_{i+1},

$$y_{i+1} - y_i \cong \frac{1}{N\lambda(y)}, \qquad \text{for } y \in [y_i, y_{i+1}] \tag{4}$$

when the number of output levels is large. As an aside, we remark that our compressors always have unity range. Smith allows more generality in his formulas. The best way to illustrate the use of (4) is through an example. Suppose that $p(x)$ is a zero-mean symmetric density (no Dirac delta functions), that N is even, and that a unique optimum quantizer exists. The initial condition for the Max iteration is a guess for first output level greater than zero. We will call this level $y_{N/2}$. We first make the observation that the output levels must be symmetric about the origin. Also, for large N, the distance between the breakpoint at zero and $y_{N/2}$ approximately equals

$$y_{\frac{N}{2}} \cong \frac{1}{2N\lambda(y_{\frac{N}{2}})} \tag{5}$$

The solution of this equation provides the initial guess for $y_{N/2}$. This basic procedure can be used with modifications for N even or odd with most common probability densities. Some numerical examples are provided in the next section.

The second method uses the companding function to work backwards from the known uniform quantizer over $[0, 1]$ in order to estimate the initial output level. In fact, the method provides a reasonable approximation to the entire quantizer. An N level uniform quantizer on $[0, 1]$ has output levels

$$\hat{y}_i = \frac{2i-1}{2N}, \qquad i = 1, \cdots, N. \tag{6}$$

Therefore, the compandor approximation is simply

$$y_i \cong g^{-1}(\hat{y}_i) = g^{-1}\left(\frac{2i-1}{2N}\right). \tag{7}$$

For the purpose of identification, we will refer to the first method of (5) as the λ-approximation and the second as the g-approximation. In hindsight these two methods seem obvious; however, they have apparently not been widely used.

III. NUMERICAL EXAMPLES

In this section we provide some examples using the λ- and g-approximations to estimate the initial input interval endpoint of a Max quantizer. The asymptotically optimum mean-square error companding characteristic is given by

$$\frac{\int_{-\infty}^{x} p(y)^{1/3}\,dy}{\int_{-\infty}^{\infty} p(y)^{1/3}\,dy} = g(x)$$

where $p(y)$ is our input probability density.

The first example we consider is when $p(y)$ is the Gaussian unit variance, zero mean, probability density. $g(x)$ is then given by $\frac{1}{2}(1 + \text{erf}\,(x/\sqrt{6}))$; hence, $g^{-1}(y) = \sqrt{6}\ \text{erf}^{-1}(2y - 1)$. Using this equation, our expression for the initial positive input interval endpoint of an N output level quantizer is $x_{1\lambda} = \sqrt{6}\ \text{erf}^{-1}(2(N/2 + 1) - 1)$.

The λ-approximation requires us to solve the equations (using a standard Newton-Raphson search)

$$x_{1\lambda} = \frac{1}{N\lambda(x_{1\lambda})} \qquad \text{for } N \text{ even}$$

$$x_{1\lambda} = \frac{1}{2N\lambda(x_{1\lambda})} \qquad \text{for } N \text{ odd}$$

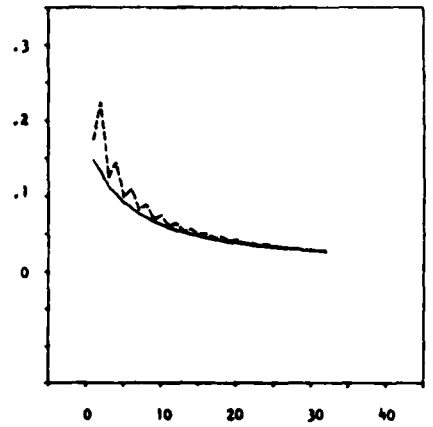


Fig. 1. P_g (solid line) and P_λ (dotted line) plotted as a function of N for the Gaussian density.

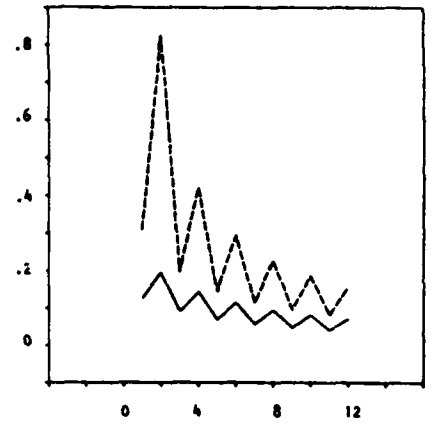


Fig. 2. P_g (solid line) and P_λ (dotted line) plotted as a function of N for the Laplacian density.

where

$$\lambda(x_{1\lambda}) = \frac{(2\pi)^{1/6}}{(6\pi)^{1/2}}\, p(x_{1\lambda})^{1/3}.$$

Since Max tabulated the actual values of the input interval endpoints, we may compute the quantities

$$P_g \triangleq \left| \frac{x_{1g} - x_{act}}{x_{act}} \right|$$

and

$$P_\lambda \triangleq \left| \frac{x_{1\lambda} - x_{act}}{x_{act}} \right|$$

for various values of N where x_{act} is the actual tabulated value.

In Fig. 1 we see P_g (solid line) and P_λ (dotted line) plotted as a function of N for values of N from 5 to 36. As may be seen from the figure, the g-approximation is better for all these vaues of N. Furthermore, the λ-approximation does not have a solution for $N = 4$, which is an additional drawback of using this approximation in low N regions.

We now perform the same computations for the Laplacian ($p(y) = \exp\{-|y|\}/2$) and Rayleigh ($p(y) = y \exp\{-y^2/2\}$) probability densities. In Fig. 2 we plot P_g (solid line) and P_λ (dotted line) for values of N from 5 to 16 for the Laplacian density. Again, the g-approximation is best for all values of N and, furthermore, the λ-approximation has no solution when $N = 4$.

In Fig. 3 we see plots of P_g (solid line) and P_λ (dotted line) for values of N from 2 to 36 for the Rayleigh distribution. For every value except $N = 2$, the g-approximation is better than the λ-approximation. The plot of P_g is noisy because calculation of x_{1g} for this density required a large numerical integration which was very sensitive to the number of samples used in the summation.

We should note that Max quantizers have been computed for the Rayleigh and the Gaussian densities using both $x_{1\lambda}$ and x_{1g} as the estimate for the initial interval endpoint. With no convergence problems, quantizers of 10 000 and 200 output

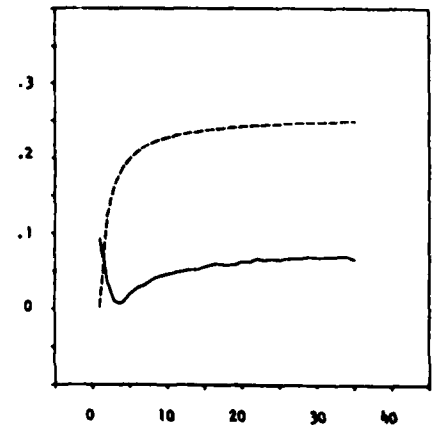


Fig. 3. P_g (solid line) and P_λ (dotted line) plotted as a function of N for the Rayleigh density.

levels have been computed for the Gaussian and Rayleigh probability densities, respectively. In practice, we find that both methods give sufficiently good estimates to allow quick convergence to the correct quantizer. A typical value is 200 iterations for a 1000 level Gaussian quantizer with the last level specified to 10^{-5} accuracy. We conclude that the x_{1g} estimate is a better approximation in most cases, but the $x_{1\lambda}$ estimate is often substantially easier to compute.

REFERENCES

[1] J. Max, "Quantizing for minimum distortion," *IRE Trans. Inform. Theory*, vol. IT-6, pp. 7–12, Mar. 1960.
[2] N. C. Gallagher, "Optimum quantization in digital holography," *Appl. Opt.*, vol. 17, pp. 109–115, Jan. 1, 1978.
[3] M. D. Paez and T. H. Glisson, "Minimum mean-squared error quantization in speech PCM and DPCM systems," *IEEE Trans. Commun.*, vol. COM-20, pp. 225–230, Apr. 1972.
[4] W. C. Adams and C. E. Giesler, "Quantizing characteristic for signals having Laplacian amplitude probability density function," *IEEE Trans. Commun.*, vol. COM-26, pp. 1295–1297, Aug. 1978.
[5] W. A. Pearlman and G. H. Senge, "Optimal quantization of the Rayleigh probability distribution," *IEEE Trans. Commun.*, vol. COM-27, pp. 101–112, Jan. 1979.
[6] B. Smith, "Instantaneous companding of quantized signals," *Bell Syst. Tech. J.*, vol. 36, pp. 653–709, May 1957.
[7] E. Delp and J. A. Bucklew, mutual correspondence, 1977.
[8] P. E. Fleisher, "Sufficient conditions for achieving minimum distortion in a quantizer," in *IEEE Int. Conv. Rec.*, 1964, pp 104–111.

[9] P. F. Panter and W. Dite, "Quantization distortion in pulse-count modulation with nonuniform spacing of levels," *Proc. IRE*, vol. 39, pp. 44–48, Jan. 1951.
[10] G. M. Roe, "Quantizing for minimum distortion," *IRE Trans. Inform. Theory*, vol. IT-10, pp. 384–385, Oct. 1964.

On the Design of Nonlinear Discrete-Time Predictors

T. E. McCANNON, MEMBER, IEEE, NEAL C. GALLAGHER, MEMBER, IEEE, D. MINOO-HAMEDANI, AND GARY L. WISE, MEMBER, IEEE

Abstract—The problem of minimum mean-squared error prediction of a discrete-time random process using a nonlinear filter consisting of a zero-memory nonlinearity followed by a linear filter is studied. Classes of random processes for which the best predictor is realizable using a nonlinear filter of the above form are discussed. For those random processes for which the best predictor is not realizable using the above nonlinear filter, an iterative procedure is presented for finding a suboptimal nonlinear filter;

Manuscript received November 17, 1980. This work was supported in part by the Air Force Office of Scientific Research under Grants AFOSR 78-3605, AFOSR 76-3062, AFOSR 81-0047, and AFOSR 76-3062, and in part by the Department of Defense Joint Services Electronics Program under Grant F4962-77-C0101.

T. E. McCannon and N. C. Gallagher are with the School of Electrical Engineering, Purdue University, West Lafayette, IN 47907.

G. W. Wise is with the Department of Electrical Engineering, University of Texas at Austin, Austin, TX 78712.

D. Minoo-Hamedani was with the Department of Electrical Engineering, University of Texas. He is now with Bell Laboratories, Holmdel, NJ 07733.

0018-9448/82/0300-0366$00.75

special attention is directed to the case where the nonlinearity is a polynomial. Also, a noniterative approach based on nonlinear regression is presented.

I. Introduction

In this correspondence we consider a second-order random process $\{X_n, n = 1, 2, \cdots\}$, and we are interested in predicting the random variable X_{N+1} from an observation of $X_1, \cdots, X_N$. Our estimate is denoted by $\hat{X}_{N+1}$, and we wish to choose it so as to minimize the mean-squared error.

It is well-known [1, pp. 77–78] that the optimal estimate of X_{N+1} in terms of $X_1, \cdots, X_N$ is given by the conditional expectation

$$\hat{X}_{N+1} = E\{X_{N+1} \mid X_N, \cdots, X_1\}.$$

In general, this is a Borel measurable function of $X_1, \cdots, X_N$. In many cases an exact expression for this quantity is difficult to obtain. Often we do not have the necessary statistical information to evaluate such a quantity. In such cases, we might restrict the form of the estimator to be linear and apply well-known techniques [2] for its determination. Linear estimation can also be thought of as applying the projection theorem [1, pp. 150–155] and projecting X_{N+1} onto the linear manifold generated by the observations $X_1, \cdots, X_N$. Clearly, in this case the only statistical information required is the second-moment characteristics of the random process.

In an attempt to improve estimation performance, we propose to modify or augment this subspace so as to have a larger signal component present within the subspace. A linear method cannot alter the subspace in the manner required to achieve the desired behavior; however, a nonlinear system can modify the subspace. So, we begin by restricting our estimate $\hat{X}_{N+1}$ to be of a form that is expressible as the output of a system consisting of a time-invariant zero-memory nonlinearity (ZNL) followed by a linear filter. The ZNL is characterized by a Borel measurable function $g(\cdot)$ such that $g(X_1), \cdots, g(X_N)$ are second-order random variables. We can now form our estimate of X_{N+1} as a linear combination of the $g(X_i)$ by projecting X_{N+1} onto the linear manifold generated by the modified observations $g(X_1), \cdots, g(X_N)$. If the weighting sequence of the linear filter is given by $h_0, \cdots, h_{N-1}$, then the estimate is given by

$$\hat{X}_{N+1} = \sum_{n=1}^{N} g(X_n) h_{N-n}. \tag{1}$$

We wish to determine a function $g(\cdot)$ and a set of coefficients $h_0, \cdots, h_{N-1}$ so that the resulting mean-squared error is minimized and is at least as good as that of the optimal linear filter. Similar system structures have been employed in certain detection applications [3].

We note that the purpose of the linear filter is simply to implement the projection operation. The purpose of the ZNL is to modify the observations in such a way that the resulting linear manifold contains a large component of X_{N+1} so that the error associated with the projection is small. We note that in working with a nonlinear system of this type, no statistical knowledge of the random process beyond that contained in the family of bivariate distributions is ever required. In some cases even less statistical knowledge suffices. For example, if the ZNL is chosen to be a polynomial, then the required statistical knowledge of the random process reduces to the family of certain joint higher order moments of the random process.

Since we know only the second-moment characteristics of the random process, the widest class of systems over which we could optimize is the class of linear systems. Thus, to do better than is possible using linear prediction, we must have more statistical knowledge of the random process than its second-moment characteristics. Therefore, since the ZNL serves the purpose of modifying the closed linear manifold onto which X_{N+1} is projected, and since the resulting prediction scheme never requires statistical knowledge of the random process beyond that contained in the family of bivariate distributions, a nonlinear predictor of this type seems reasonable.

In Section II we consider some cases where the optimal estimate has the form of (1). In the general case, the optimal predictor will not have the form of (1), and thus a predictor of this form will be suboptimal. This situation is discussed in Section III, where an iterative scheme is presented for determining suboptimal predictors. In Section IV examples are given to illustrate the method. Finally, in Section V a noniterative approach utilizing a modified ZNL structure is considered.

II. Optimal Prediction

In this section we consider some cases where the optimal filter has the form of (1). Whenever the optimal filter is linear, then it obviously has the form of (1) with $g(x) = x$. The class of spherically invariant random processes [4] admits linear solutions, with the most well-known examples being the Gaussian processes.

It is clear that the performance of the filter given by (1) can always be made at least as good as that of the optimal linear filter. In some cases the filter given by (1) can be optimal while the optimal linear filter is useless. For example, let $X_n = P_n(U)$ where U is a random variable uniformly distributed over $[-1, 1]$ and $P_n(\cdot)$ is the nth Legendre polynomial [5]. In this case, the sequence $\{X_n, n = 1, 2, \cdots\}$ is a sequence of uncorrelated zero-mean random variables, and the optimal linear filter yields an estimate which is zero. However, for $g(x) = P_{N-1}(x)$ and

$$h_n = \begin{cases} 1, & n = N-1 \\ 0, & n \neq N-1, \end{cases}$$

the filter of (1) gives the estimate $\hat{X}_{N+1} = X_{N+1}$. Numerous examples similar to this can easily be constructed.

When the process is a (first-order) Markov process, it is well-known [1, pp. 81–83] that $E\{X_{N+1} \mid X_N, \cdots, X_1\} = E\{X_{N+1} \mid X_N\}$ with probability one (wp1). Thus a system of the form of (1) with a ZNL given by $g(x) = E\{X_{N+1} \mid X_N = x\}$ and a weighting sequence given by

$$h_n = \begin{cases} 1, & n = 0 \\ 0, & n \neq 0 \end{cases}$$

will yield the optimal estimate of X_{N+1}.

Markov processes serve as models for many physical phenomena that arise in practice. Often they are obtained as the solution of first-order stochastic difference equations of the form

$$X_{n+1} = g(X_n) + Z_{n+1}, \qquad n = 0, 1, 2, \cdots,$$

where $g(\cdot)$ is a Borel measurable function and the sequence $\{Z_i\}$ is a sequence of zero-mean independent random variables that are independent of the initial condition X_0. It is easily seen that in this case we will have $E\{X_{N+1} \mid X_N, \cdots, X_1\} = g(X_N)$ wp1.

Clearly, for any random process for which

$$E\{X_{N+1} \mid X_N, \cdots, X_1\} = \sum_{n=1}^{N} g(X_n) h_{N-n} \text{ wp1}, \tag{2}$$

a system of the form of (1) will produce the optimal estimate of X_{N+1}. As another example of a process for which the conditional expectation has the form of (2), consider the process generated by the following second-order stochastic difference equation

$$X_{n+2} = h_0 g(X_{n+1}) + h_1 g(X_n) + Z_{n+2}, \qquad n = 0, 1, 2, \cdots, \tag{3}$$

where $g(\cdot)$ is a Borel measurable function and $\{Z_n\}$ is a sequence of zero-mean independent random variables independent of the initial conditions X_1 and X_0. It can be easily seen that for this example, for an $N \geq 2$,

$$E\{X_{N+1} \mid X_N, \cdots, X_1\} = h_0 g(X_N) + h_1 g(X_{N-1}) \text{ wp1}$$

Extension of this example to the case where (3) is a kth order stochastic difference equation is obvious.

III. Suboptimal Prediction

In the general case there will not exist a function $g(\cdot)$ and a weighting sequence $h_0,\cdots,h_{N-1}$ such that (2) is satisfied. However, it is quite reasonable to conjecture that in many cases it may be possible to determine a filter having the form of (1) with a mean-squared error either significantly smaller than that associated with the optimal linear filter or very close to the mean-squared error associated with the optimal filter.

If the function $g(\cdot)$ that minimizes the mean-squared error is known, the $g(X_n)$ will be well-defined random variables and the determination of the h_n that minimize the mean-squared error reduces to an application of the projection theorem that is, setting

$$E\left\{\left[X_{N+1}-\sum_{n=1}^{N}h_{N-n}g(X_n)\right]g(X_j)\right\}=0,\qquad j=1,\cdots,N, \tag{4}$$

and solving for the h_n. To carry out this step we need to calculate the terms $E\{g(X_n)g(X_j)\}$ and $E\{X_{N+1}g(X_j)\}$. In practice, the determination of the function $g(\cdot)$ that minimizes the mean-squared error is a difficult problem.

Notice that, in the optimization problem where the filter is constrained to be of the form in (1), only second-order information (i.e., the family of bivariate distributions) is required. This is more statistical information than would be required if we were doing optimal linear filtering, which requires only second-moment information. However, it is still considerably less statistical information than would be required if we were doing optimal filtering, which requires statistical information pertaining to an $(N+1)$-dimensional distribution.

In order to circumvent the difficult problem of determining the function $g(\cdot)$ to use in (1), we will sacrifice some degree of optimality and parameterize $g(\cdot)$, thus letting the determination of $g(\cdot)$ simply depend upon finding the correct parameters. Doing so, we then write the resulting mean-squared error as a function of the parameters associated with $g(\cdot)$ and the weighting sequence of the linear filter. In this case, the mean-squared error would be a function of $K+N$ parameters, where K is the number of parameters associated with $g(\cdot)$. For example, let $g(\cdot)$ be given by

$$g(x)=\sum_{j=1}^{K}a_jb_j(x).$$

Then our estimate is given by

$$\hat{X}_{N+1}=\sum_{n=1}^{N}\sum_{j=1}^{K}h_{N-n}a_jb_j(X_n),$$

and the resulting mean-squared error is given by

$$\begin{aligned}E\left\{\left[X_{N+1}-\hat{X}_{N+1}\right]^2\right\}&=E\left\{\left[X_{N+1}\right]^2\right\}-2\sum_{n=1}^{N}\sum_{j=1}^{K}h_{N-n}a_jE\left\{X_{N+1}b_j(X_n)\right\}\\&\quad+\sum_{n=1}^{N}\sum_{m=1}^{N}\sum_{j=1}^{K}\sum_{k=1}^{K}h_{N-n}h_{N-m}a_ja_kE\left\{b_j(X_n)b_k(X_m)\right\}.\end{aligned} \tag{5}$$

The functions $b_j(\cdot)$ should be determined so that there is considerable flexibility in the functional form of $g(\cdot)$ and also so that the expectations in (5) could be determined from the statistical information at hand. For example, if $b_j(x)=x^j$, then the necessary statistical information would consist of the higher order joint moments.

The next step might be to minimize (5) over the $N+K$ parameters. This would result in $N+K$ equations of third-order polynomials in the parameters. This simultaneous optimization over all the parameters presents potential numerical problems. As an alternative to the simultaneous optimization over all the parameters, we describe an iterative technique.

The basic plan of the iterative technique is to consider the two sets of parameters separately and to iteratively optimize over one set of parameters while holding the other set fixed. This iterative technique results in the need to solve systems of linear equations, as opposed to the need to solve systems of equations in third-order polynomials such as encountered in the effort to simultaneously optimize over all the parameters.

We will assume that the parametric form of $g(\cdot)$ is such that with the proper choice of parameters we could have $g(x)=x$. In this way the mean-squared error that results will always be upper bounded by the mean-squared error associated with the optimal linear filter.

The iterative technique is as follows.

- Step 1: Determine the optimal weighting sequence $h_0,\cdots,h_{N-1}$ for the case where $g(x)=x$.
- Step 2: For this choice of $h_0,\cdots,h_{N-1}$, determine $a_1,\cdots,a_K$ so as to minimize the mean-squared error.
- Step 3: For this choice of $a_1,\cdots,a_K$, determine the optimal weighting sequence $h_0,\cdots,h_{N-1}$.
- Step 4: Repeat Steps 2 and 3 until the improvement in the mean-squared error is negligible.

At each stage of execution the algorithm provides a system design whose mean-square estimation error is no larger than that for the previous step of the algorithm.

The $a_1,\cdots,a_K$ and $h_0,\cdots,h_{N-1}$ that are obtained in Step 4 after the termination of the iterations determined the system. Step 1 and Step 3 make use of the projection theorem and result in $E\{X_{N+1}g(X_j)\}=\sum_{n=1}^{N}h_{N-n}E\{g(X_n)g(X_j)\}$, $j=1,\cdots,N$. Step 2 makes use of (5) and results in

$$\begin{aligned}&\sum_{n=1}^{N}\sum_{l=1}^{N}h_{N-n}h_{N-l}\left[2a_jE\left\{b_j(X_n)b_j(X_l)\right\}+\sum_{\substack{p=1\\p\neq j}}^{K}a_pE\left\{b_j(X_n)b_p(X_l)\right\}\right]\\&\quad=2\sum_{n=1}^{N}h_{N-n}E\left\{X_{N+1}b_j(X_n)\right\},\qquad j=0,1,\cdots,k.\end{aligned}$$

IV. Examples

In this section we consider a particular parametric form for the ZNL and a specific model for the random sequence. The iterative method described earlier is used in this case to determine a filter of the form of (1). We also determine the mean-squared error resulting from use of the optimal filter and that resulting from use of the optimal linear filter. Performance results for these filters are compared, and it is seen that in several instances the improvement in mean-squared error of the suboptimal filter over that of the optimal linear filter is a significant fraction of the corresponding improvement of the optimal filter over that of the optimal linear filter.

Assume that we have knowledge of the regression function for stationary $\{X_n\}$:

$$r(x)=E\{X_{N+1}\mid X_N=x\}. \tag{6}$$

Notice that if we choose $g(x)=r(x)$ and

$$h_n=\begin{cases}1, & n=0\\0, & n\neq 0,\end{cases}$$

then the estimate would be the same as that of the optimal filter based on the most recent observation. If we were to use the projection theorem to choose a different weighting sequence $\{h_n\}$, we might do better. It seems reasonable to expect that if we were to parameterize $g(\cdot)$ so that by proper choice of the parameters we would have $g(x) = r(x)$, and then to use this parameterization of the ZNL in the iterative technique described earlier, we might determine a system of the form of (1) exhibiting very good performance. This is how we will choose the ZNL in this section.

As a model for the random sequence $\{X_n, n = 1, 2, \cdots\}$ we assume that

$$X_n = (Z_n)^{2q+1}, \tag{7}$$

where $\{Z_n, n = 1, 2, \cdots\}$ is a zero-mean stationary Gaussian process with unit variance and autocorrelation function $\rho(\cdot)$. First we derive an expression for the regression function (6) when the random sequence is given by (7). Using results in [6], we have that

$$\begin{aligned} E\{X_{N+1} \mid X_N\} &= E\{(Z_{N+1})^{2q+1} \mid Z_N\} \\ &= \sum_{n=0}^{\infty} [\rho(1)]^n b_n \theta_n(Z_N) \\ &= \sum_{n=0}^{\infty} [\rho(1)]^n b_n \theta_n((X_N)^{1/(2q+1)}), \end{aligned}$$

where the series are mean-square convergent, the constants $\{b_n\}$ are given by

$$b_n = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (x)^{2q+1} \theta_n(x) \exp\left(-\frac{x^2}{2}\right) dx, \tag{8}$$

and θ_n is the nth normalized Hermite polynomial given by

$$\theta_n(x) = \frac{(-1)^n}{\sqrt{n!}} \exp\left(\frac{x^2}{2}\right) \frac{d^n}{dx^n} \exp\left(-\frac{x^2}{2}\right).$$

We see from (8) that $b_n = 0$ for $n > 2q + 1$ and, in fact, the b_n can be obtained from the relation

$$(x)^{2q+1} = \sum_{n=0}^{2q+1} b_n \theta_n(x).$$

For example, for $q = 1$ we have

$$b_n = \begin{cases} 3, & n = 1 \\ \sqrt{6}, & n = 3 \\ 0, & n \neq 1, 3 \end{cases}$$

and $r(x)$ is given by $r(x) = [\rho(1)]^3 x + 3\rho(1)(1 - [\rho(1)]^2)x^{1/3}$. For $q = 2$,

$$b_n = \begin{cases} 15, & n = 1 \\ 10\sqrt{6} & n = 3 \\ 2\sqrt{30}, & n = 5 \\ 0, & n \neq 1, 3, 5, \end{cases}$$

and

$$r(x) = [\rho(1)]^5 x + 10[\rho(1)]^3(1 - [\rho(1)]^2)x^{1/3} + 15\rho(1)(1 - [\rho(1)]^2)^2 x^{1/5}.$$

In general, for an arbitrary positive integer q, it is easily seen that $r(\cdot)$ has the form

$$r(x) = c_{q-1} x + c_q (x)^{1/(2q+1)} + c_{q-1} (x)^{(2q-1)/(2q+1)} + \cdots + c_1 (x)^{1/(2q+1)},$$

where the c_i are constants that can be determined using the above procedure. Thus we choose the ZNL $g(\cdot)$ to be

$$g(x) = \sum_{i=1}^{q+1} a_i (x)^{(2i-1)/(2q+1)},$$

where the parameters a_i are to be determined by the iterative procedure. In utilizing the iterative procedure we encounter the need for the knowledge of moments and joint moments of $\{Z_n\}$ (see [7]), which are given by

$$E\{(Z_n)^p\} = \begin{cases} 1 \cdot 3 \cdot 5 \cdots (p-1) & \text{for } p \text{ even} \\ 0, & \text{for } p \text{ odd} \end{cases}$$

$$\begin{aligned} &E\{(Z_n)^r (Z_{n+i})^s\} \\ &\quad = \mu(r, s, i) \\ &\quad = \begin{cases} (r+s-1)\rho(i)\mu(r-1, s-1, i) + (r-1)(s-1) \\ \quad \cdot (1 - [\rho(i)]^2)\mu(r-2, s-2, i), & \text{for } (r+s) \text{ even} \\ 0, & \text{for } (r+s) \text{ odd.} \end{cases} \end{aligned} \tag{9}$$

Observing that $\mu(1, 1, i) = \rho(i)$ and $\mu(2, 2, i) = 1 + 2[\rho(i)]^2$, all higher order joint moments can be calculated using (9).

In order to compare the performance of the suboptimal estimator with that of the optimal estimator, we have obtained expressions for the mean-squared error associated with the optimal estimator. For the optimal system we are interested in

$$E\{(Z_{N+1})^{2q+1} \mid Z_N, \cdots, Z_1\}.$$

Notice that this is the $(2q + 1)$ conditional moment, and the conditional distribution has the functional form of a Gaussian distribution. Thus the minimum mean-squared error follows using standard properties of the Gaussian distribution (see, for example, [8]). For $q = 1$ we find that the minimum mean-squared error is of the form $15 - P_1^4[9E\{Y^2\} + 6P_1 E\{Y^4\} + P_1^2 E\{Y^6\}]$; and for $q = 2$, the minimum mean-squared error is of the form $945 - P_1^6[225 \quad E\{Y^2\} + 300P_1 E\{Y^4\} + 130P_1^2 E\{Y^6\} + 20P_1^3 E\{Y^8\} + P_1^4 E\{Y^{10}\}]$. In these expressions P_1 is a constant, and Y is a normal random variable with zero mean and variance γ^2. The constants P_1 and γ^2 are defined as follows. Assume without loss of generality that the correlation matrix R associated with $Z_1, \cdots, Z_{N-1}$ is positive definite (if it is not, the data can be reduced to achieve this result). Then P_1 is the reciprocal of the element in the lower right corner of R^{-1}. Denote the first N elements in the last row of R^{-1} as $r_1, \cdots, r_N$. Then

$$\gamma^2 = \sum_{i=1}^{N} (r_i)^2 + 2 \sum_{m=1}^{N-1} \sum_{n=1}^{m} r_{N-n+1} r_{m-n+1} \rho(N-m).$$

The mean-squared error associated with the optimal linear filter can now be obtained in a straightforward fashion

In Tables I-VIII results are presented comparing the suboptimal filter to the optimal filter and the optimal linear filter. Several correlation sequences for $\{Z_N\}$ are considered, both the third power and the fifth power of Z_n are used as models, and examples for two observations and five observations are given. In these tables L_1, L, and $L_{\min}$ are the mean-squared errors resulting from the optimal linear filter, suboptimal filter using a ZNL, and the optimal filter, respectively. The quantity n_1 is the percent of decrease in L_1 when the suboptimal filter using a ZNL is employed, i.e., $n_1 = 100(L_1 - L)/L_1$. The quantity n_2 is the percent of possible improvement in L_1 using the optimal filter, i.e., $n_2 = 100(L_1 - L_{\min})/L_1$. The quantity n_3 is the normalized percent of improvement over the linear filter given by the suboptimal filter using a ZNL, i.e., $n_3 = 100n_1/n_2 = 100(L_1 - L)/(L_1 - L_{\min})$.

TABLE I
CORRELATION SEQUENCES CORRESPONDING TO TABLES II-V

	ρ(1)	ρ(2)	ρ(3)	ρ(4)	ρ(5)
1	.75	.575	.45	.35885	.291
2	.885	.7887	.70762	.639	.5805
3	.55	.315	.187	.11445	.07183
4	.55	.395	.319	.26885	.23023
5	.625	.2375	.14675	.09648	.06207
6	.8333	.6666	.5	.3333	.1666
7	.5787	.2963	.125	.037	.00463
8	.4822	.1975	.0625	.0123	.00077

TABLE II
MEAN-SQUARED ERRORS AND PERCENTAGES OF IMPROVEMENT FOR $q = 1$

	L_1	L	L_{min}	η_1	η_2	η_3
1	9.1983	8.8614	8.8581	3.6	3.69	97.3
2	5.1744	5.0622	5.0599	2.16	2.21	97.6
3	12.5987	12.1084	12.108	3.89	3.89	99.8
4	12.3196	11.9216	11.8952	3.23	3.44	93.7
5	13.6849	13.2957	13.295	2.84	2.86	99.1
6	6.9247	6.6228	6.4926	4.56	6.23	69.8
7	12.2903	11.732	11.7259	4.54	4.59	98.8
8	13.5219	12.8142	12.8123	3.91	3.82	99.6

TABLE III
MEAN-SQUARED ERRORS AND PERCENTAGES OF IMPROVEMENT FOR $q = 2$

	L_1	L	L_{min}	η_1	η_2	η_3
1	727.62	704.58	704.22	3.13	3.18	98.1
2	455.78	444.78	444.69	1.98	2.04	96.7
3	887.49	859.95	859.9	3.1	3.1	99.7
4	879.44	856.59	851.86	2.82	3.13	89.8
5	920.93	899.7	899.43	2.3	2.33	98.5
6	584.57	556.58	550.99	5.41	5.74	59.3
7	876.33	845.86	845.24	3.47	3.54	97.7
8	910.86	884.62	884.42	2.88	2.9	99.2

TABLE IV
COEFFICIENTS a_i OF NONLINEARITY $g(x) = a_2 x + a_1 \sqrt[3]{x}$ AND h_i OF SUBOPTIMAL SYSTEM FOR $q = 1$

	h_0	h_1	h_2	h_3	h_4	a_1	a_2
1	.6115	.0127	.008	.0059	.0094	1.519	.6811
2	.7899	.0084	.0064	.0051	.0132	.674	.962
3	.4026	.0093	.0049	.0029	.0026	2.7896	.4114
4	.3654	.0687	.0433	.0297	.028	2.6769	.4334
5	.2827	.0407	.0164	.0076	.0047	3.5779	.2656
6	.776	-.0234	-.0175	-.0111	-.0062	1.2015	.7615
7	.4478	-.029	-.013	-.01	-.0015	2.775	.4375
8	.3505	-.0267	-.0121	-.0032	-.0017	3.362	.3215

TABLE V
COEFFICIENTS a_i OF ZNL $g(x) = a_1 x + a_2 x^{1/3} + a_3 x^{1/5}$ AND h_i OF SUBOPTIMAL SYSTEM FOR $q = 2$

	h_0	h_1	h_2	h_3	h_4	a_1	a_2	a_3
1	.6779	.0119	.0063	.0042	.0052	4.0527	3.7727	.495
2	.7065	.0097	.0067	.005	.0095	.7136	2.032	.7585
3	.2563	.0059	.0028	.0017	.0014	15.173	4.5019	.196
4	.2466	.0672	.0282	.019	.017	11.733	4.465	.1966
5	.162	.0227	.009	.0043	.0026	23.858	3.802	.0839
6	.6534	-.034	-.0234	-.0136	-.024	2.742	2.9562	.6302
7	.2864	-.0184	-.0096	-.0054	.0002	14.7841	4.5769	.2267
8	.2032	-.0139	-.0065	.0019	.0008	22.373	4.2663	.128

TABLE VI
CORRELATION SEQUENCES CORRESPONDING TO TABLES VII, VIII

	ρ(1)	ρ(2)
1	.9	.7
2	.8	.5
3	.8	.3
4	.7	.1

TABLE VII
COEFFICIENTS a_i OF ZNL $g(x) = a_2 x + a_1 x^{1/3}$ AND h_i OF SUBOPTIMAL SYSTEM FOR $q = 1$

	h_0	h_1	a_1	a_2
1	1.2577	-.4974	.9333	.82985
2	.9837	-.3001	1.6639	.6923
3	1.095	-.6467	2.3987	.6089
4	.7927	-.4796	3.2982	.4545

TABLE VIII
MEAN-SQUARED ERRORS AND PERCENTAGES OF IMPROVEMENT FOR $q = 1$

	L_1	L	L_{min}	η_1	η_2	η_3
1	3.7487	3.494	3.1354	6.79	16.3	41.65
2	7.566	7.0275	6.7406	7.12	10.9	65.52
3	5.7804	4.371	1.0231	24.58	82.3	29.62
4	8.9825	7.1689	4.9674	20.19	44.7	45.16

As mentioned earlier, the functions $b_j(\cdot)$ should be determined that considerable flexibility exists in the functional form of) For example, if X_{N-1} has a nonzero mean, then choosing ie of the $b_j(\cdot)$ to be constant would enable the mean to be ibtracted out and thus decrease the mean-squared error. In this se, for example, Step 1 of the algorithm should be replaced th the following: determine the optimal weighting sequence $\ldots, h_{N-1}$ for the case where $g(x) = x + 1$. In this case, Step will result in the best affine filter (i.e., linear plus a constant), as posed to the best linear filter.

As we also mentioned earlier, the functions $b_j(\cdot)$ should be osen such that the expectations in (5) could be determined on the statistical information at hand. To once again test this thod of nonlinear prediction, we simulated the following difrence equation driven by white noise and empirically estimated e necessary expectations from the simulated quantities:

$$X_{n+1} = -1.74X_n^2 + 0.005U_{n+1},$$

here the sequence $\{U_n\}$ is a sequence of independent random variables uniformly distributed on $[-1/2, 1/2]$. Letting $g(x) = c_0 + c_1 + c_2x^2$, we see that it is possible to realize the best predictor with a nonlinear system of the form under consideration. We took $N = 2$ and empirically estimated the expectations occurring in (5). After one iteration of the algorithm, the empirically estimated mean-squared error was reduced from 0.085 to 0.00031.

V. AN ALTERNATE DESIGN APPROACH

In the preceding, we considered an iterative procedure for the design of the nonlinear predictor. In this section we will consider a genralization of that concept which results in a noniterative procedure. Recall that the purpose of the ZNL was to modify the linear manifold onto which X_{N-1} is projected. The purpose of the linear filter was simply to implement the projection onto the linear manifold generated by $g(X_1), \cdots, g(X_N)$. If the ZNL were allowed change, then the possibility exists of choosing the ZNL such that a larger component of X_{N-1} lies within the linear manifold spanned by its output.

In the earlier case with a single ZNL we have sacrificed some degree of optimality by parameterizing the ZNL and then letting the determination of $g(\cdot)$ depend upon finding the correct parameters. In this situation, the mean-squared error was a function of $N + K$ parameters. If we now allow for N such ZNL's in the system, then the mean-squared error will be a function of $N(K + 1)$ parameters. It may appear at first glance that we have now made the problem much more complex, due to the introduction of more parameters. Hoever, as we shall see shortly, this alternative approach will result in a noniterative design procedure.

With N ZNL's the estimate is given by

$$\hat{X}_{N+1} = \sum_{n=1}^{N} g_n(X_n) h_{N,n}$$

where N ZNL's are given by

$$g_n(x) = \sum_{j=1}^{K} a_{n,j} b_j(x).$$

In this case, if we let $\tilde{a}_{n,j} = a_{n,j} h_{n,N}$, then the ZNL $g_n(\cdot)$ could be replaced by

$$\tilde{g}_n(x) = \sum_{j=1}^{K} \tilde{a}_{n,j} b_j(x),$$

and the linear filter could be replaced by an accumulator, and the mean-squared error will be a function of NK parameters. In the sequel we will take this approach. Thus our estimate is now of the form

$$\hat{X}_{N+1} = \sum_{n=1}^{N} \sum_{j=1}^{K} a_{n,j} b_j(X_n), \tag{10}$$

and we wish to determine the parameters $\{a_{n,j}\}$. The minimum mean-squared error estimate of this form is given by projecting X_{N+1} onto the linear manifold generated by the NK random variables $\{b_j(X_n)\}$. Thus the parameters $\{a_{n,j}\}$ are given as a solution to

$$BA = C, \tag{11}$$

where A is a KN-dimensional column vector of the parameters $\{a_{n,j}\}$ ordered lexicographically, B is a $KN \times KN$ matrix whose general term is of the form $E\{b_i(X_j)b_k(X_m)\}$ where the lexicographic order of i and j denote the column and the lexicographic order of k and m denotes the row, and C is a KN-dimensional column vector made up of the terms $E\{X_{N+1}b_i(X_n)\}$ ordered lexicographically in j and n. We note that if the parameters $\{a_{n,j}\}$ are such that (11) is satisfied, then the resulting estimate given by (10) is the minimum mean-squared error estimate, and by the projection theorem it is uniquely defined up to probability-one equivalence. That is, more than one solution to (11) may exist, however, for any number of solutions to (11), the resulting estimates are all equal with probability one. Also, the projection theorem guarantees that at least one solution to (11) exists.

As a specific example, we might choose $b_i(x) = x^{i-1}$. In this case, the matrix B will consist of various moments and cross moments of the set of random variables.

To compare the two methods, we simulated the following difference equation:

$$X_{n+1} = 0.87 + 1.74X_n^2 + 0.13X_{n-1} + 0.05U_{n-1},$$

where the U_n were independent random variables uniformly distributed over $[-1/2, 1/2]$. We set $N = 2$, $K = 3$, and $b_i(x) = x^{i-1}$. The necessary moments and cross moments were empirically estimated from the simulated quantities. The iteration procedure using a single ZNL yielded an estimate given by

$$\hat{X}_3 = 0.903677g(X_2) + 0.003506g(X_1),$$

where

$$g(x) = 1 + 0.097418x - 1.856364x^2.$$

The noniterative procedure using N ZNL's yielded an estimate given by

$$\hat{X}_3 = 0.828810 + 0.046003X_2 + 1.742086X_2^2 + 0.132142X_1 - 0.080110X_1^2.$$

If the actual moments and cross moments had been used in the noniterative procedure, then for this example the exact minimum mean-squared error estimate, given by

$$\hat{X}_3 = -0.87 + 1.74X_2^2 + 0.13X_1,$$

would have resulted. The resulting mean-squared errors were empirically estimated from the simulated quantities and are given by 0.003503 and 0.000205 for the iterative and noniterative procedures, respectively. The actual minimum mean-squared error for this problem is $2.5/12000 \simeq 0.0002083$.

VI. Summary

We investigated the design of nonlinear discrete-time prediction filters. We motivated our approach through the concept of modifying or augmenting the subspace generated by the observations in such a way so as to have a larger signal component present within this augmented subspace. The form of the system under study was that of a zero-memory nonlinearity followed by a linear time-invariant filter (ZNL-LTI). We have shown that in many cases, where the optimum nonlinearity is known, the ZNL-LTI structure produces nearly optimum results. Finally, an extension to the use of several ZNL's was considered.

References

[1] J. L. Doob, *Stochastic Processes*. New York: Wiley, 1953.

[2] T. Kailath, Ed., *Linear Least-Square Estimation*. Stroudsburg, PA: Dowden, Hutchinson, and Ross, 1977.

[3] J. H. Miller and J. B. Thomas, "Detectors for discrete-time signals in non-Gaussian noise," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 241-250, Mar. 1972.

[4] I. B. Blake and J. B. Thomas, "On a class of processes arising in linear estimation theory," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 12-16, Jan. 1968.

[5] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*. New York: Dover, 1964.

[6] G. L. Wise and J. B. Thomas, "A characterization of Markov sequences," *J. Franklin Inst.*, vol. 299, pp. 269-278, Apr. 1975.

[7] N. L. Johnson and S. Kotz, *Distributions in Statistics: Continuous Multivariate Distributions*. New York: Wiley, 1972, p. 91.

[8] K. S. Miller, *Multidimensional Gaussian Distributions*. New York: Wiley, 1964, pp. 21-22.

The Design of Two-Dimensional Quantizers using Prequantization

KERRY D. RINES, MEMBER, IEEE, AND NEAL C. GALLAGHER, JR., MEMBER, IEEE

Abstract—The theoretical advantages of two-dimensional quantization over univariate quantization have been studied in the literature. However, in many cases there is no known implementation for the two-dimensional quantizer that can operate in real time. A new approach to the design of two-dimensional quantizers is presented. This technique, called prequantization, is used to design two-dimensional quantizers that operate in real time. The importance of prequantization is demonstrated by the design of the optimum uniform two-dimensional (hexagonal) quantizer. Additional examples are given to illustrate the flexibility of this design approach.

I. Introduction

THE USE OF two-dimensional quantizers for encoding analog sources has been of increasing interest in recent years. Two-dimensional quantizers can offer advantages in the design of both optimum and suboptimum quantizers. These advantages may be offset by the difficulty in implementing many two-dimensional quantizers. In this paper we present a new approach to the design of two-dimensional quantizers called prequantization. We show that for a number of examples prequantization simplifies the quantizer implementation and/or improves the quantizer performance.

Manuscript received Feb. 19, 1980; revised March 12, 1981. This work was supported by the Air Force Office of Scientific Research under Grant AFOSR 78-3605

K. D. Rines was with the School of Electrical Engineering, Purdue University, West Lafayette, IN. He is now with The Analytic Sciences Corporation, McLean Operation, 8301 Greensboro Drive, Suite 1200, McLean, VA, 22102.

N. C. Gallagher, Jr., is with the School of Electrical Engineering, Purdue University, West Lafayette, IN 47907.

The design of two-dimensional quantizers for optimum quantization is one area of interest. Consider the random sequence $x_1, x_2, x_3, \cdots$ where the x_i are all independent and identically distributed. The traditional approach to quantizing this sequence is to perform the quantization one sample at a time using a one-dimensional quantizer. Much of the early work in quantization theory has addressed this problem. As a result the design and implementation of optimum one-dimensional quantizers is straightforward. In addition these quantizers are often able to operate at high source rates. These properties make one-dimensional quantization an attractive choice for quantizing the above sequence. The advantage of quantizing the independent identically distributed (i.i.d.) sequence in two or more dimensions is discussed by Zador [1]. Simply stated, these results indicate that the minimum obtainable per sample distortion decreases as the quantizer dimension is increased. Therefore, the potential exists to improve the performance of digital encoders by replacing one-dimensional quantizers with two-dimensional quantizers.

Zador's results include derivations of both the upper and lower bounds on the distortion obtained when using an optimum quantizer. Unfortunately, these results do not provide insight into the structure of the quantizer. The design and implementation of optimum two-dimensional quantizers remains a largely unsolved problem. Recently the design of two-dimensional quantizers has been addressed. Computer algorithms for designing optimum quantizers of two or more dimensions have been presented

0018-9448/82/0300-0232$00.75

by many authors, such as Linde *et al.* [2]. The algorithms specify the optimum set of output vectors for the quantizer. The optimum quantizer can then be implemented using a search procedure. Having specified the output set, the search is used to choose the output vector that is the smallest distance from the input vector. However, this implementation of the optimum quantizer may be difficult or impossible to operate at high bit rates. Thus we are left with the following dilemma. We can use a one-dimensional quantizer that is easy to implement and suffer a high level of distortion or we can improve the distortion by using a two-dimensional quantizer and accept the difficulties in the implementation. To date the easy implementation of one-dimensional quantizers has outweighted the theoretical advantages of using two-dimensional quantizers.

In Section III we consider the design of the optimum uniform two-dimensional quantizer. Gersho [3] has stated that the optimum uniform two-dimensional quantizer is the hexagonal quantizer. Using prequantization we construct a simple design for the hexagonal quantizer which can operate in real time. For our purposes we say that a quantizer can operate in real time if the quantizer can operate at approximately the same source rates as a one-dimensional quantizer. Thus the prequantization design of the hexagonal quantizer allows us to take advantage of the performance improvements available with two-dimensional quantizers while maintaining the easy implementation characteristic of one-dimensional quantizers. This hexagonal quantizer design is a significant result and demonstrates the potential practical applications of prequantization.

The design of suboptimum two-dimensional quantizers has also been studied in the literature. This interest has been motivated by the numerous examples in which the data are physically generated in groups of two. These studies note the difficulty in designing optimum quantizers and explore the advantages of using suboptimum two-dimensional quantizers. One example of data that are generated in pairs is samples from a complex-valued discrete Fourier transform. The design of suboptimum two-dimensional quantizers for the discrete Fourier transform (DFT) has been studied by Pearlman and Gray [4] and Gallagher [5].

In Sections IV and V we examine two examples of suboptimum two-dimensional quantizers. The quantizers are then redesigned using the prequantization approach. In each case, the addition of prequantization substantially reduces the mean-squared error performance of the quantizer. These results further emphasize the usefulness of prequantization.

II. Prequantization

The design of two-dimensional quantizers using prequantization is illustrated in Fig. 1. The design consists of a nonlinearity called a prequantizer preceding a two-dimensional quantizer called an output quantizer. This design approach is analogous to the implementation of a

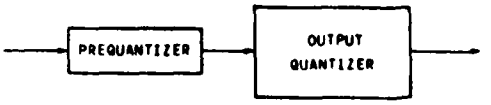


Fig. 1 Two-dimensional quantizer design using prequantization

quantizer using a search procedure. Let the quantizer to be designed be described by a partitioning of the input space, where all the input vectors contained within one cell of the partition are mapped to the same output vector. The first step in implementing a search is to define the set of allowable quantizer output vectors. Then for each input vector a search is conducted to find the output vector assigned to that input vector by the partitioning.

Similarly the first step in designing a quantizer using prequantization is to define the set of output vectors. This is done using a two-dimensional quantizer that is called the output quantizer. Thus we must determine the set of output vectors specified by the quantizer being designed and then build a two-dimensional quantizer with that same set of output vectors. The problem of building the output quantizer is somewhat simplified in the prequantization approach since there are no constraints on how the output quantizer partitions the input space.

The second step in the quantizer design is to require that for each input vector the proper output vector is assigned. For the quantizer being designed, let A_i be an output vector and S_i be the set of all input vectors contained in the cell of the partition corresponding to A_i. Similarly A_i is also an output vector of the output quantizer, and we let T_i be the set of all input vectors contained in the cell of the partition corresponding to A_i. A nonlinearity called a prequantizer is used to map S_i into T_i for all i. Thus the prequantization design maps S_i into A_i by first mapping S_i into T_i with the prequantizer and then mapping T_i into A_i using the output quantizer. This prequantization design procedure is illustrated with a simple example.

Consider the design of the two-dimensional quantizer shown in Fig. 2. This quantizer has no significance other than its usefulness in this example. Using the prequantization procedure we must first build an output quantizer that defines the same output set as in Fig. 2. The output quantizer can be designed very simply using two univariate equal-step-size quantizers. The partitioning of the output quantizer is shown in Fig. 3. Having defined the output vector set with the output quantizer, we now turn to the design of the prequantizer. We observe that each partition in Fig. 2 can be mapped into the corresponding cell in Fig. 3 by letting $y' = y$ and $x' = x - \Delta/4$. Thus the prequantizer that completes the design of the quantizer in Fig. 2 is given by

$$\begin{aligned} y' &= y \\ x' &= x - \frac{\Delta}{4} \end{aligned} \tag{1}$$

One advantage of using the prequantization design approach is that often the quantizer can operate in real time. Again we define a real-time quantizer as a quantizer that

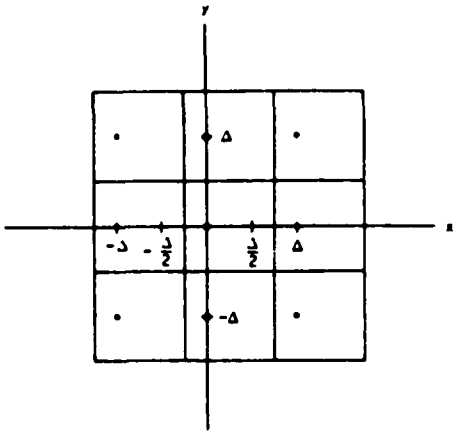


Fig. 2. Partitioning of a two-dimensional quantizer.

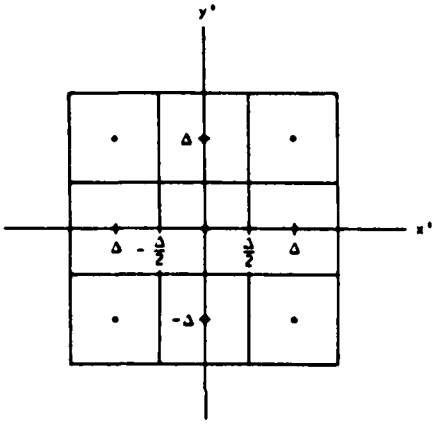


Fig. 3. Partitioning of output quantizer.

can operate at approximately the same source rates as a one-dimensional quantizer. In a number of examples the output quantizer can be implemented using a combination of one-dimensional quantizers and as a result can operate in real time. It is also useful to note that the prequantizer is defined only as a nonlinear mapping and may or may not be a quantizer. This differs from the term pre-quantizer used in the literature which refers to one quantizer preceding another quantizer.

III. Hexagonal Quantization

Gersho has argued that for independent samples (at high bit rates) the optimum uniform two-dimensional quantizer is the hexagonal quantizer. The design of a hexagonal quantizer using prequantization is given here. First we attempt to build a two-dimensional output quantizer that can be easily implemented and operate in real time. One quantizer meeting these requirements is a scaled version of the diamond quantizer given below.

Let the inputs to the two-dimensional output quantizer be x and y. The variables x and y are first encoded into two new variables w and z by the linear transformation

$$w = x + \sqrt{3}\,y$$
$$z = x - \sqrt{3}\,y. \tag{2}$$

The variables w and z are quantized separately by univariate quantizers with a uniform step-size Δ. The outputs of the output quantizer are then obtained using the linear transformation

$$\hat{x} = \frac{1}{2}(\hat{w} + \hat{z})$$
$$\hat{y} = \frac{1}{2\sqrt{3}}(\hat{w} - \hat{z}). \tag{3}$$

The position of this quantizer in the hexagonal quantizer design is shown in Fig. 4 and the partitioning of the scaled diamond quantizer is given in Fig. 5. Having chosen the output quantizer as defined in (2) and (3), we now turn to the design of the prequantizer.

The prequantizer must map the hexagonal region corresponding to each output into a scaled diamond region corresponding to that same output. Consider the hexagonal partition shown in Fig. 6. Assume x is fixed and the pair (x, y) is contained within a given hexagonal partition. We now pose a question: does there exist a value x' such that the pair (x', y) is contained within the corresponding diamond partition for all values of y? This approach is illustrated with the following example. Let $x = x_1$ as shown in Fig. 6 and let y be in the range $-\Delta/2\sqrt{3}$ to $\Delta/2\sqrt{3}$. In Fig. 6 we observe that the hexagonal quantizer output will be (0, 6) for all input pairs in the set $\{(x_1, y): y_1 \le y \le y_2\}$. Similarly in Fig. 5 we observe that the scaled diamond quantizer output will be (0, 0) for all input pairs in the set $\{(x_2, y): y_1 \le y \le y_2\}$. Therefore, if $f(x_1) = x_2$, the quantizer in Fig. 4 will behave like the hexagonal quantizer for all input pairs in the set $\{(x_1, y): -\Delta/2\sqrt{3} \le y \le \Delta/2\sqrt{3}\}$. In fact, we can show that the quantizer in Fig. 4 behaves like the hexagonal quantizer for all inputs in the set $\{(x_1, y): -\infty \le y \le \infty\}$ when $f(x_1) = x_2$. Repeating this example for all possible values of x_1, we obtain a prequantizing function that maps the hexagonal region corresponding to each output into a scaled diamond-shaped region corresponding to that same output. The resulting prequantizer function is given in (4).

$$f(x) = \begin{cases} n\dfrac{\Delta}{2}, & n\dfrac{\Delta}{2} - \dfrac{\Delta}{6} \le x \le n\dfrac{\Delta}{2} + \dfrac{\Delta}{6} \\ 3x - (2n+1)\dfrac{\Delta}{2}, & \\ & n\dfrac{\Delta}{2} + \dfrac{\Delta}{6} \le x \le (n+1)\dfrac{\Delta}{2} - \dfrac{\Delta}{6}. \end{cases} \tag{4}$$

IV. Prequantized Spectral Phase Coding

Spectral phase coding (SPC) is a robust suboptimum technique for coding a nonstationary or large dynamic range discrete-time series into digital form. SPC utilizes the discrete Fourier transform and a two-dimensional quantizer to obtain its robust characteristics. The SPC algorithms are given here, while a detailed explanation of SPC is available in [6]. The input is a discrete-time complex-valued random sequence $\{a_n\}_{n=0}^{N-1}$. The spectral magnitude A_p and the spectral phase θ_p of the discrete sequence are given

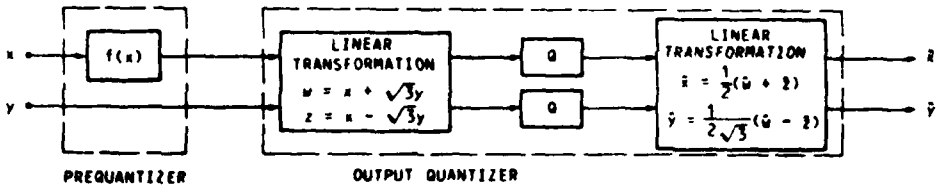


Fig. 4. Prequantization design for hexagonal quantizer. Quantizer Q has uniform step-size Δ.

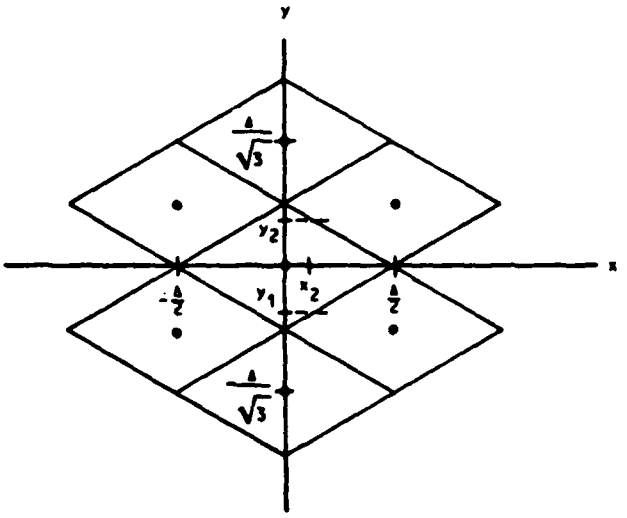


Fig. 5. Partitioning of scaled diamond quantizer

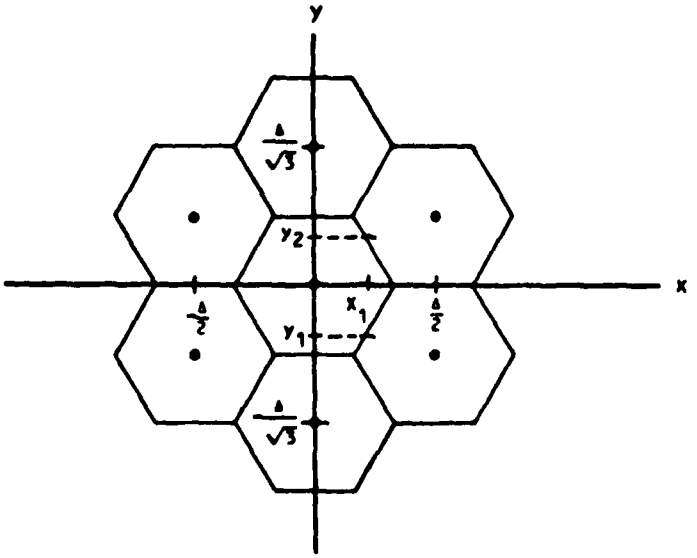


Fig. 6. Partitioning of hexagonal quantizer

below:

$$\{a_n\}_{n=0}^{M-1} \xrightarrow{\text{DFT}} \{A_p e^{i\theta_p}\}_{p=0}^{M-1}. \tag{5}$$

SPC encodes the magnitude and phase of the spectrum by forming the sequence $\{\psi_p\}_{p=0}^{2M-1}$ given by

$$\begin{aligned} \psi_p &= \theta_p + \gamma_p \\ \psi_{p+M} &= \theta_p - \gamma_p. \end{aligned} \tag{6}$$

where

$$\gamma_p = \cos^{-1}\frac{A_p}{S}$$

and

$$S = \max_p A_p,$$

where the maximum is taken over $p = 0, 1, \cdots, M-1$. The

quantized sequence $\{\hat{\psi}_p\}$ is transmitted and used at the receiver to recover the original discrete signal. The reconstructed discrete sequence is

$$\{\hat{a}_n\}_{n=0}^{M-1} \overset{\text{DFT}^{-1}}{\leftarrow} \left\{\frac{S}{2}\left(e^{i\hat{\psi}_p} + e^{i\hat{\psi}_{p+M}}\right)\right\}_{p=0}^{M-1}. \tag{7}$$

This equation can be rewritten in terms of the quantized magnitude and phase components at the receiver

$$\{\hat{a}_n\}_{n=0}^{M-1} \overset{\text{DFT}^{-1}}{\leftarrow} \left\{\frac{S}{2}e^{i\hat{\theta}_p}\left(e^{i\hat{\gamma}_p} + e^{-i\hat{\gamma}_p}\right)\right\}_{p=0}^{M-1}, \tag{8}$$

where

$$\hat{\theta}_p = \frac{1}{2}\left(\hat{\psi}_p + \hat{\psi}_{p+M}\right)$$
$$\hat{\gamma}_p = \frac{1}{2}\left(\hat{\psi}_p - \hat{\psi}_{p+M}\right). \tag{9}$$

Examining (6) and (9) we see that the variables θ_p and γ_p are quantized by a two-dimensional quantizer called a diamond quantizer. SPC utilizes the discrete Fourier transform along with the diamond quantizer to code the possibly nonstationary random sequence $\{a_n\}$ into a well-behaved uniformly bounded sequence $\{\psi_p\}$. In many cases the sequence $\{\psi_p\}$ is uniformly distributed from zero to 2π. As a consequence $\{\psi_p\}$ is quantized using a uniform step-size quantizer.

Since SPC is a suboptimum quantizer we ask the question: does there exist a prequantizing function that can improve the SPC performance? The results from [4] and [7] indicate that for polar quantization (at high bit rates) the number of magnitude quantization levels N_1, and the number of phase levels N_2, must be related by

$$N_2 \simeq 2.6\, N_1 \tag{10}$$

for optimum performance. In SPC, γ_p ranges from zero to $\pi/2$ and θ_p ranges from zero to 2π. Thus γ_p has only one-fourth the effective quantization levels of θ_p. If γ_p is simply rescaled to range from zero to π, $\{\hat{a}_n\}$ cannot be uniquely recovered from the sequence $\{\hat{\psi}_p\}$. However, using prequantization the quantizer can be redesigned to minimize the mean square error (mse) on γ_p and improve the SPC performance.

We begin by defining the quantization errors for ψ_p and ψ_{p+M} as

$$\alpha_p = \psi_p - \hat{\psi}_p,$$
$$\alpha_{p+M} = \psi_{p+M} - \hat{\psi}_{p+M}. \tag{11}$$

Assume the quantization takes place using an N-level equal step-size quantizer. Then using a Fourier series expansion, we can write

$$\alpha_p = -\frac{2}{N}\sum_{n=1}^{\infty}\frac{(-1)^n}{n}\sin nN\psi_1,$$
$$\alpha_{p+M} = -\frac{2}{N}\sum_{n=1}^{\infty}\frac{(-1)^n}{n}\sin nN\psi_2. \tag{12}$$

We now define the quantization errors for θ_p and γ_p as

$$e_p = \theta_p - \hat{\theta}_p,$$
$$d_p = \gamma_p - \hat{\gamma}_p. \tag{13}$$

Solving for θ_p and γ_p in (6) and using this result with (9) and (11) in (13) we obtain

$$e_p = \frac{1}{2}(\alpha_p + \alpha_{p+M}),$$
$$d_p = \frac{1}{2}(\alpha_p - \alpha_{p+M}). \tag{14}$$

Then substituting (12) into (14) and using a trigonometric identity we can write

$$e_p = -\frac{2}{N}\sum_{n=1}^{\infty}\frac{(-1)^n}{n}\sin nN\theta_p\cos nN\gamma_p,$$
$$d_p = -\frac{2}{N}\sum_{n=1}^{\infty}\frac{(-1)^n}{n}\cos nN\theta_p\sin nN\gamma_p. \tag{15}$$

Thus the mse on γ_p is

$$E\{d_p^2\} = \frac{4}{N^2}\sum_{n=1}^{\infty}\sum_{m=1}^{\infty}\frac{(-1)^{n+m}}{nm}$$
$$\cdot E\{\cos nN\theta_p\cos mN\theta_p\sin nN\gamma_p\sin mN\gamma_p\}. \tag{16}$$

For a large number of quantization levels N, the mse on γ_p becomes

$$E\{d_p^2\} \simeq \frac{1}{N^2}\sum_{n=1}^{\infty}\frac{1}{n^2}\left(1 + E\{\cos 2nN\theta_p\}\right). \tag{17}$$

From (17) we find that $E\{d_p^2\}$ is minimized for

$$\theta_p = \theta_p' = k\frac{\pi}{N} + \frac{\pi}{2N}, \tag{18}$$

where

$$k = 0, 1, \cdots, 2N-1.$$

Applying these results, we propose the following coding scheme called prequantized spectral phase coding (PQSPC). First obtain θ_p and γ_p as with SPC. The values $\{\theta_p\}$ are then quantized with output levels $k\pi/N + \pi/2N$ for $k = 0, 1, \cdots, 2N-1$. The quantizer output $\{\theta_p'\}$ is then used to form the sequence $\{\psi_p\}$ and the rest of the procedure is identical to SPC. Figs. 7 and 8 depict the quantization region shapes for SPC and PQSPC, respectively.

In [7] SPC was compared with the optimum unit variance Gaussian quantizer (O.G.Q.). We now present a similar comparison to evaluate the performance of PQSPC. The normalized mse performances of the optimum unit variance Gaussian quantizer, SPC and PQSPC are compared in Figs. 9 and 10. All the quantizers have 32 levels (5 bits/sample) and the block size for SPC and PQSPC is 64. In Fig. 9 the normalized mse of the three quantizers with a zero-mean Gaussian input is given as a function of the input variance. The normalized mse of the quantizers with a zero-mean Laplacian input is given as a function of the input variance in Fig. 10.

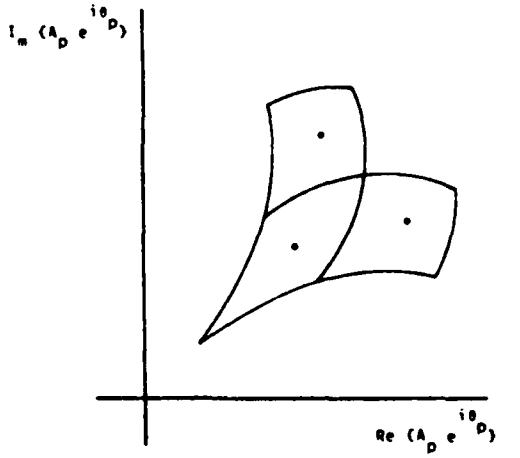


Fig. 7. Shape of quantizing regions for $A_p e^{i\theta_p}$ using SPC.

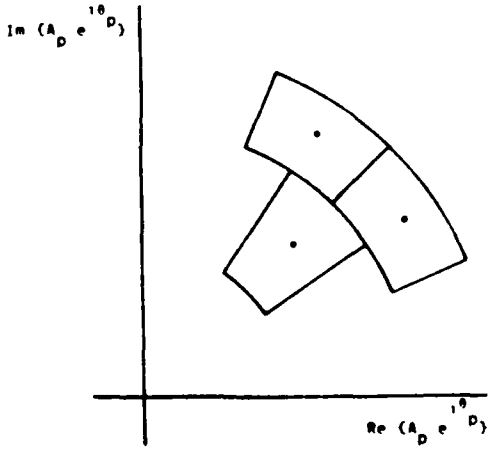


Fig. 8. Shape of quantizing regions for $A_p e^{i\theta_p}$ using PQSPC

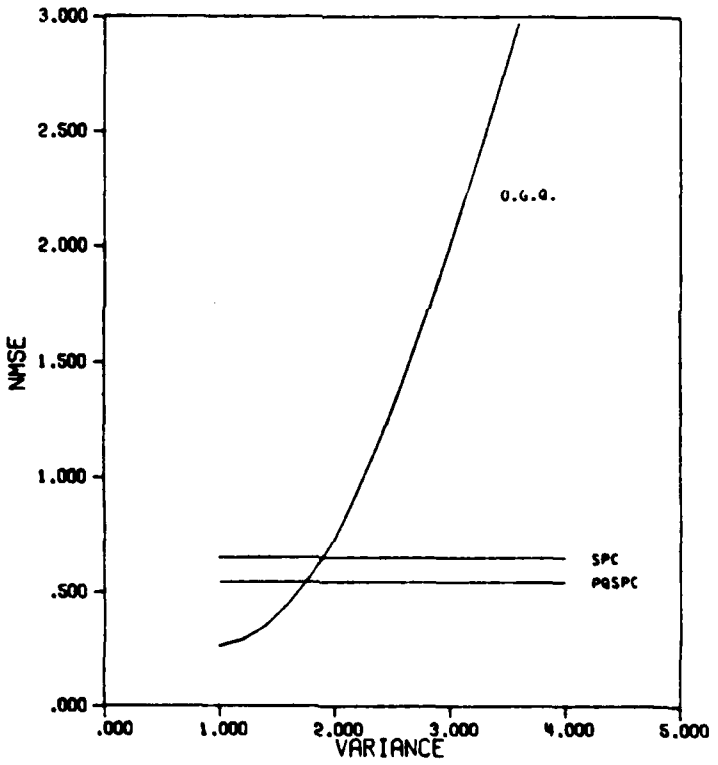


Fig. 9. Comparison of normalized mse between optimum unit variance Gaussian quantizer, SPC, and prequantized SPC with a zero-mean Gaussian input.

In terms of normalized mse, PQSPC offers an improvement over SPC of 16.3 percent for the Gaussian input densities and 16.0 percent for the Laplacian densities. The improvement for nonsymmetric input densities can be even more dramatic. In the case of the one side exponential density PQSPC offers a 47.5 percent reduction in normalized mean-squared error over that of SPC. A desirable characteristic of SPC is its relative insensitivity to a change in signal power or statistics. Figs. 9 and 10 demonstrate that PQSPC shares this characteristic. In fact, the normalized mse of PQSPC remains constant for any change in the signal variance and changes only 1.4 percent when the input statistics are changed from Gaussian to Laplacian.

V. Hsueh–Sawchuk Holograms

The wide applicability of prequantization is further illustrated by considering an example from computer-generated holography. In this section we present the results of using prequantization in Hsueh–Sawchuk computer-gener-

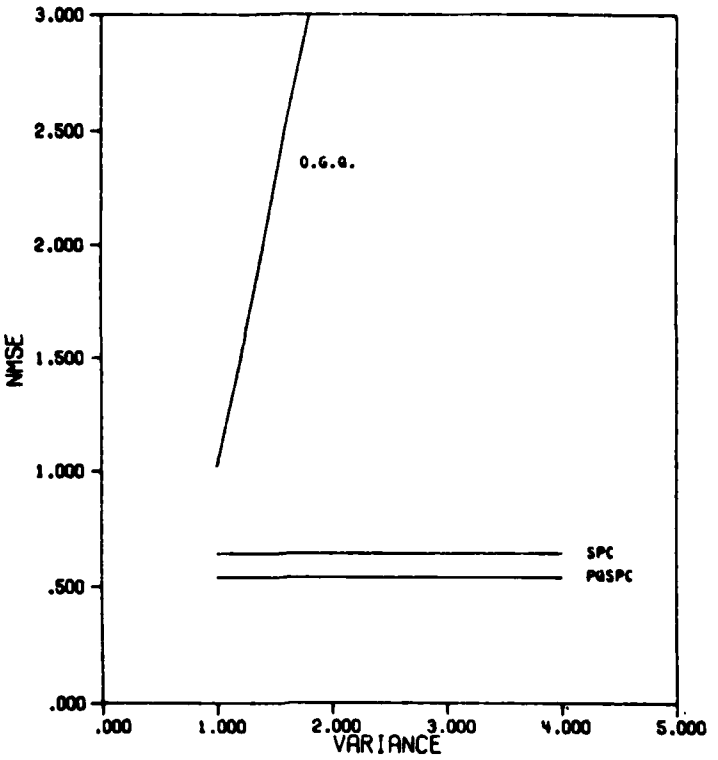


Fig. 10. Comparison of normalized mse between optimum unit variance Gaussian quantizer, SPC, and prequantized SPC with a zero-mean Laplacian input.

ated holograms. A detailed analysis of prequantization in Hsueh-Sawchuk holograms is given in [9] and a good summary of computer-generated holography is available in [10].

The Hsueh-Sawchuk hologram encodes the discrete Fourier transform of the desired holographic image into a binary pattern. This binary pattern is then written onto the hologram using a pattern generator with finite resolution. The finite resolution of the pattern generator can be modeled as a quantizer. Thus the complex-valued discrete Fourier transform of the holographic image is effectively quantized by a two-dimensional quantizer. This quantization can be improved by using prequantization.

The normalized mean square quantization error for the Hsueh-Sawchuk hologram in Fig. 11 is 6.82×10^{-2}. This compares with a mean square error of 5.25×10^{-2} for the prequantized Hsueh-Sawchuk hologram. Thus the quantization error is improved 23 percent by the addition of prequantization. The improved quantization error can also be seen by comparing Figs. 11 and 12. The quantization error can be approximated as a white additive noise which appears as the high frequency background noise in the holograms. We see the prequantized hologram in Fig. 12 has less background noise than the hologram in Fig. 11. Thus the prequantization has reduced the quantization error without any harmful effects on the holographic image itself.

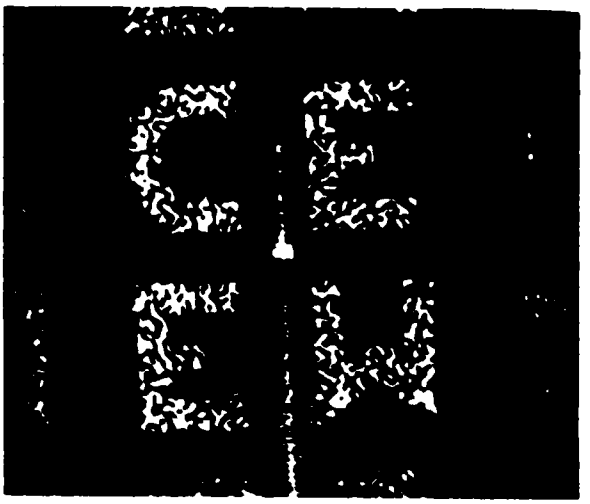


Fig. 11. Hsueh-Sawchuk hologram.

IV. Discussion

We have presented a new approach to the design of two-dimensional quantizers. The usefulness of the prequantization approach has been demonstrated in three examples. The hexagonal quantizer design is of particular importance. The prequantization design makes the use of the hexagonal quantizer with its theoretical advantages

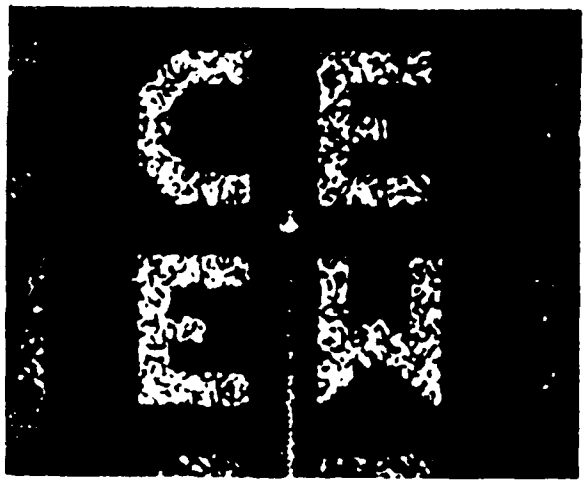


Fig. 12. Hsueh-Sawchuk hologram with prequantization.

more practical. Existing two-dimensional quantizers were examined in the two other examples. In each case prequantization reduced the quantization error while retaining the other important system characteristics.

At this stage the work on the prequantization design approach is incomplete. Presently there are no guidelines as to how or when prequantization can be used to design two-dimensional quantizers. However, the results presented here indicate that this approach may deserve some consideration whenever a two-dimensional quantizer is to be implemented.

References

[1] P. Zador, "Development and evaluation of procedures for quantizing multivariate distributions," Ph.D. dissertation, Stanford University, CA, University Microfilm no. 64-9855, 1964.

[2] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. COM-28, pp. 84–95, Jan. 1980.

[3] A. Gersho, "Asymptotically optimal block quantization," *IEEE Trans. Inform. Theory*, vol. IT-25, pp. 373–380, July 1979.

[4] W. A. Pearlman and R. M. Gray, "Source coding of the discrete Fourier transform," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 683–692, Nov. 1978.

[5] N. C. Gallagher, Jr., "Quantizing schemes for the discrete Fourier transform of a random time-series," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 156–163, Mar. 1978.

[6] ——, "Spectral phase coding," *Proc. of John Hopkins CISS*, Apr. 1976.

[7] J. A. Bucklew and N. C. Gallagher, Jr., "Quantization schemes for bivariate Gaussian random variables," *IEEE Trans. Inform. Theory*, vol. IT-25, pp. 537–543, Sept. 1979.

[8] J. Max, "Quantizing for minimum distortion," *IRE Trans. Inform. Theory*, vol. IT-6, pp. 7–12. Mar. 1960.

[9] K. Rines and N. C. Gallagher, Jr., "Reducing quantization error in Hsueh–Sawchuk holograms," *Applied Optics*, vol. 20, pp. 2008–2010, June 1981.

[10] W. H. Lee, "Computer generated holograms: Techniques and applications," in *Progress in Optics*, vol. 16, Amsterdam, The Netherlands: North Holland, 1977.

END

FILMED