M1.9

A NONLINEAR SIMPLEX ALGORITHM FOR MINIMUM ORDER SOLUTIONS

Richard Leahy[†] , *Brian Jeffs*[†‡] *and Zhenyu Wu*[†]

[†]Signal and Image Processing Institute
Department of Electrical Engineering - Systems
University of Southern California
Los Angeles, CA 90089

[‡]Hughes Aircraft Company
7000 Village Drive
Buena Park, CA 90621

ABSTRACT

A nonlinear simplex algorithm is developed for solving a special class of linearly constrained optimization problems. The cost for this class of problem is a sequence of functionals indexed by the positive integers P such that for $P=1$ we have a standard linear programming problem and for $P \rightarrow \infty$ the cost function selects the solution with the minimum number of non-zero elements in the solution vector. Two theorems, analogous to the fundamental theorem of linear programming, are given as the basis for the nonlinear simplex algorithm. Applications of this technique are discussed for system design and signal processing problems in which an optimally sparse solution vector is desired. Examples of such problems include FIR filter design, placement of beamformer arrays and seismic deconvolution.

1. Introduction

In this paper we describe an algorithm for solving the linearly constrained optimization problem

$$\min_{\underline{Q} \in R^N} \quad g_P(\underline{Q}) \;\; s.t. \;\; \underline{B} = \mathbf{H}\,\underline{Q}\,, \;\; \underline{Q} \geq \underline{0}, \tag{1}$$

where $\mathbf{H} \in R^{M \times N}$ and $\underline{B} \in R^M$ define a set of linear equality constraints and $g_P(\underline{Q})$ denotes the sequence of functionals, indexed by the integers $P \geq 1$, defined as:

$$g_P(\underline{Q}) = \sum_{i=1}^{N} |Q_i|^{\frac{1}{P}} \tag{2}$$

The extension to inequality constraints is also addressed.

The algorithm is based on the interesting property that the solution to (1) will always lie in the set S of basic feasible solutions defined as

$$S = \{\underline{Q} \in R^N : \sum_{i=1}^{N} \mathrm{I}(Q_i) \leq M, \underline{B} = \mathbf{H}\underline{Q}, \underline{Q} \geq \underline{0}\} \tag{3}$$

where

$$\mathrm{I}[Q_i] = \begin{cases} 1 & Q_i \neq 0 \\ 0 & Q_i = 0 \end{cases}$$

Equivalently, any optimal feasible solution to (1) will have at most M nonzero components. This property is well known for the case $P=1$ since then (1) is a standard linear programming problem and the property described above results from the well known fundamental theorem of linear programming [1].

We have recently shown, [2], that an equivalent theorem holds for positive integer values of P with $P=1$ as a special case. The significance of this result is that, as in the case of linear programming, we can confine our search for the solution to problem (1) to the finite set S. A procedure for performing this kind of search is the well known simplex algorithm discussed below.

Our interest in the sequence $g_p(\underline{Q})$ was motivated by a problem in which we wish to find a solution to a set of linear constraints with the minimum number of non-zero elements in the solution vector. Since the P^{th} root of any strictly positive number converges to unity as $P \rightarrow \infty$ and all roots of zero are zero it follows that as P increases the cost asymptotically approaches the desired zero/one cost function:

$$\lim_{P \rightarrow \infty} g_P(\underline{Q}) = \sum_{i=1}^{N} \mathrm{I}(Q_i) \tag{4}$$

The reasons for using a finite value of P rather than the limit of the sequence lies in the convergence and stability of the algorithm.

A number of problems exist in which our primary goal is to select a solution vector with the minimum number of non-zero coefficients. These problems fall into two classes: system design and data processing. An example of the latter is the seismic deconvolution problem. A seismic source wavelet propagating through the earth will be reflected at the boundaries between two layers of differing acoustical impedance. These differences in impedance are characterized by the 'reflectivity sequence' [3] and are useful as an aid in geophysical exploration. The reflected wave measured at the surface can be modeled as the convolution of the incident wavelet with the reflectivity sequence. Applying the cost function in equ (4) to the deconvolution problem will yield the minimum number of layers for which there is evidence in the data.

An example of a design problem which may be formulated as in (1) is digital filter design. The computational cost of implementing a digital FIR filter is determined primarily by the number of non-zero coefficients [4] rather than the order of the filter, thus it is desirable in some applications to design a filter which meets a given specification and has the minimum number of non-zero taps. The specification may be written in the form of a set of linear inequality constraints relating the filter coefficients and the desired frequency response. A similar formulation is also possible for the beamformer array design problem.

2. The Nonlinear Simplex Algorithm

Consider the set of solutions to $\mathbf{H}\underline{Q} = \underline{B}$. We shall call any $\underline{Q}$ which satisfies this linear constraint and the inequality $\underline{Q} \geq 0$ a 'feasible solution'. We may re-index the elements of $\underline{Q}$ and corresponding columns of $\mathbf{H}$ so as to select any M columns and group them into the first columns of $\mathbf{H}$. We shall call this first $M \times M$ submatrix $\mathbf{A}$. The corresponding elements of $\underline{Q}$ are called the basis variables and we can rewrite the equation as:

$$[\,\mathbf{A}\,|\,\mathbf{D}\,]\underline{Q} = \underline{B}, \quad where \quad [\,\mathbf{A}\,|\,\mathbf{D}\,] = \mathbf{H} \tag{5}$$

This work supported in part by the Hughes Aircraft Company Doctoral Fellowship Program and in part by the USC Faculty Research and Innovation Fund.

CH2561-9/88/0000-0745 $1.00 © 1988 IEEE

Any negative element of $\underline{B}$ and the corresponding row of $\mathbf{H}$ is multiplied by -1 to make $\underline{B} \geq \underline{0}$. Multiplying by $\mathbf{A}^{-1}$ (we may use any left pseudoinverse if $\mathbf{A}$ is singular) we can find a feasible solution directly

$$[\,\mathbf{I}\,|\,\mathbf{A}^{-1}\mathbf{D}\,]\underline{Q} = \mathbf{A}^{-1}\underline{B}\ , \quad \underline{Q}_B = \begin{bmatrix} \mathbf{A}^{-1}\underline{B} \\ 0 \\ . \\ 0 \end{bmatrix} \tag{6}$$

We call $\underline{Q}_B$ a 'basic' solution and it contains at least N-M zero terms, i.e. it belongs to the set S defined in equ (3). An adjacent basic solution is one that is formed by moving one variable out of the basis and one non-basic variable into the basis. This is accomplished by permuting the columns of $\mathbf{H}$ and corresponding elements of $\underline{Q}$ so as to swap a column in $\mathbf{A}$ with one in $\mathbf{D}$ and recomputing $\mathbf{A}^{-1}$ or equivalently by pivoting the matrix and measurement vector $\underline{B}$.

The following two theorems are the basis for the nonlinear simplex algorithm used in solving (1). They are equivalent to the fundamental theorem of linear programming, differing only in the form of the cost function.

> **Theorem 1:** *If there is a feasible solution to the problem (1) then there is a basic feasible solution to (1).*

The existence of feasible and basic feasible solutions is dependent only on the constraint equation $\mathbf{H}\underline{Q} = \underline{B}$, not on the cost functional. Therefore, this portion of the proof is identical to the linear programming case for which a proof is available in many texts [1].

> **Theorem 2:** *If there exists an optimal feasible solution $\underline{Q}^*$ to (1), then $\underline{Q}^*$ is also a basic feasible solution.*

The proof of this theorem differs from that for the linear case and is given in [2].

With the justification provided by the above theorems, we may solve problem (1) using an approach similar to the linear programming simplex algorithm which searches only the basic solutions to find the optimal solution. It is noteworthy that the ability to take this approach is entirely dependent on the particular cost function chosen. Although $g(\underline{Q})$ is neither linear nor convex, and has numerous local minima which would make most gradient based optimization techniques ineffective, the fundamental theorems above imply that it is particularly suited to a simplex search approach.

The modified simplex algorithm is very similar to the linear case. The major difference is in the rule for choosing the entering and leaving variables. In the linear case, a simplex tableau is formed which contains an additional row to allow efficient computation of the cost of any basic solution. Since our cost is nonlinear it must be specifically computed at each iteration. The choice of the entering and leaving variables is based on the computed cost. A move is made in a direction such that the cost is decreased monotonically. Thus if a solution is reached such that all adjacent solutions are of higher cost then we accept this solution as the optimal solution. With the exception of the case $P=1$, the solution is not globally optimal but locally optimal with respect to the adjacent basic solutions.

Our experiments have shown that the 1/P simplex algorithm converges in approximately the same number of iterations as the linear algorithm would for a similar sized system, which implies that the 1/P simplex algorithm will converge in about twice the computer time required by a linear program due to the increased complexity in calculating the change in the cost. With a convergence time comparable to the $\mathbf{O}(5N)$ iterations of the LP simplex, this algorithm is dramatically more efficient than an exhaustive search.

The choice of P is an important factor in the behavior of the algorithm. Using the indicator function directly sometimes leads us to a global minimum, particularly when the solution is highly degenerate. However, there are advantages in using finite values of P. Firstly, all solutions for which the number of nonzero components is equal have the same cost for $P=\infty$ but differ for finite P. Thus if a lower order solution is not adjacent to all basic solutions of the same (higher) order, the algorithm may terminate too early for $P=\infty$ but is less likely to do so otherwise. Smaller values of P also lead to better numerical stability in the computation of the cost function. It should be noted that these observations are preliminary and we are currently working to place them on firmer theoretical ground.

3. Inequality Constraints

The above algorithm would be of very limited use if it were restricted to equality constraints and non-negative values of $\underline{Q}$. Fortunately, as with the linear simplex algorithm, it may be extended to include both positive and negative components in the solution vector and to allow inequality constraints.

Inequalities in the data are handled by the introduction of 'slack variables' [1] as follows. Consider the set of inequality constraints $|\underline{B}-\mathbf{H}\underline{Q}|<\delta$. The number of equations is doubled to accommodate simultaneous upper and lower bounds on the error and the system becomes:

$$\begin{bmatrix} \mathbf{H} & | & -\mathbf{I} & | & 0 \\ \hline \mathbf{H} & | & 0 & | & \mathbf{I} \end{bmatrix} \begin{bmatrix} \underline{Q} \\ \underline{S}^+ \\ \underline{S}^- \end{bmatrix} = \begin{bmatrix} \underline{B}+\underline{\delta} \\ --- \\ \underline{B}-\underline{\delta} \end{bmatrix} \tag{7}$$

where $\underline{S}^+$ and $\underline{S}^-$ are a set of slack variables which are allowed to take on any positive value and are not included in the computation of the cost. Similarly, both positive and negative values in the solution vector are allowed by doubling the length of the solution vector and replacing $\underline{B}=\mathbf{H}\underline{Q}$ with:

$$\begin{bmatrix} \mathbf{H} & | & 0 \\ \hline 0 & | & -\mathbf{H} \end{bmatrix} \begin{bmatrix} \underline{Q}^+ \\ --- \\ \underline{Q}^- \end{bmatrix} = \begin{bmatrix} \underline{B} \\ --- \\ \underline{B} \end{bmatrix} \tag{8}$$

The final solution is given by $\underline{Q}=\underline{Q}^+-\underline{Q}^-$. Due to the nature of the cost functions, Q_i^+ and Q_i^- cannot simultaneously be nonzero.

4. Minimum Computation Order FIR Filter Design

The Remez exchange algorithm is an efficient technique for computing equiripple approximations of finite order FIR filters to a desired frequency response [5]. The resulting filter has the minimum L_∞ error norm for any filter of fixed order and transition bandwidth. In the related fields of multidimensional filter and beamformer array design, the equiripple error criterion can also be used for optimal design, but algorithms for exact optimal solutions do not always exist, particularly for non-uniformly spaced samples or array elements [6,7].

A minimum length filter however may not be the most efficient filter from the point of view of the processor computational load. Symmetry in the filter coefficients can often be exploited to save processing time, and any filter tap with a zero coefficient eliminates a multiply and accumulate operation. Considering only savings due to zero valued taps, the minimum computation order filter would be that filter, of any length, which meets the response criteria and has the fewest non-zero tap weights. For beamforming arrays this would be equivalent to the optimally sparse array, the array with fewest elements. Since the response constraints can be expressed as a set of linear inequalities [5,8], this optimization problem can be formulated as a 1/P program and is well suited for the 1/P simplex algorithm. This algorithm has been implemented for the design of linear phase real symmetric FIR filters.

If we sample the desired spectral response at M equal spaced frequencies then the real, symmetric FIR filter coefficients which achieve the desired response at these points are related by the linear system $\underline{H} = \mathbf{W}\,\underline{h}$ where $\underline{h}$ is half of the symmetric filter impulse response (the other half being redundant), $\underline{H}$ is the vector of half of the desired frequency response samples and **W** is the discrete Cosine transform (DCT) kernel:

$$W_{m,n} = \begin{cases} 2\cos(2\pi mn/M) & , \; 0 \le m \le \frac{M}{2}, 1 \le n \le \frac{M}{2} - 1 \\ \cos(2\pi mn/M) & , \; 0 \le m \le \frac{M}{2}, \;\; n = 1, \frac{M}{2} \end{cases} \tag{9}$$

The number of spectral samples, M, is chosen to be larger than the longest allowable filter. The ripple constraint for each band is expressed by setting an upper and a lower bound at each frequency sample, which yields the simultaneous linear inequalities:

$$\underline{H} - \underline{\delta}^{-} \le \mathbf{W}\underline{h} \le \underline{H} + \underline{\delta}^{+} \tag{10}$$

Where $\underline{\delta}^{-}$ and $\underline{\delta}^{+}$ are the maximum allowable lower and upper deviations, respectively, from the desired response. We may express this in a form suitable for the 1/P simplex algorithm with a variation of the basic tableau to allow for the dual inequalities and positive or negative $\underline{h}$ values, as follows:

$$\begin{bmatrix} \mathbf{W} & | -\mathbf{W} & | -\mathbf{I} & | \; 0 \\ \hline \mathbf{W} & | -\mathbf{W} & | \; 0 & | \; \mathbf{I} \end{bmatrix} \begin{bmatrix} \underline{h}^{+} \\ \underline{h}^{-} \\ \underline{S}^{+} \\ \underline{S}^{-} \end{bmatrix} = \begin{bmatrix} \underline{H} + \underline{\delta}^{+} \\ --- \\ \underline{H} - \underline{\delta}^{-} \end{bmatrix} \tag{11}$$

Where $\underline{S}^{+}$ and $\underline{S}^{-}$ are slack variables associated with the upper and lower inequality bounds respectively.

The 1/P programming method has been used to design a number of lowpass, highpass, and bandpass filters. Fig. 1 shows the spectral response, and design constraints, for a bandpass filter designed using the 1/P simplex algorithm. This filter requires 21 delay taps, but has only 15 non-zero coefficients. Fig. 2 shows an equiripple filter designed to the same response constraints which requires 17 non-zero coefficients. This example is typical of design cases for simple filters; the minimum computation order filter reduces the number of non-zero taps by 2 to 4. This validates the concept that more computationally efficient filters exist, however, the savings are not large enough to justify their use. It is possible to create a complex arbitrary frequency response constraint that would demonstrate large improvements over the equiripple design, but we have found no cases where any such spectral response would be of any practical value. One must also be aware that since the frequency samples are not necessarily placed at the ripple maxima, there may be leakage between the sample points which exceeds the constraint. We have found that it is necessary to evaluate the response of the filter after it is designed to determine the actual ripple limits, but have found no cases where it was worse than that of an equiripple filter with the same number of non-zero taps.

It is anticipated that the 1/P programming approach will more appropriately be applied to related problems where optimal algorithms do not exist. A common approach to line array beamformer design is the Chebyshev weighting which produces equal ripple sidelobe levels [10]. Many of the same techniques and algorithms are used in line array and multidimensional filter design as in 1-D FIR filter design. For two dimensional arrays and filters, the problem is more difficult since no factorization theorem exists to allow polynomial approximation simultaneously in two dimensions, but good approximate solutions exist for equally spaced square and rectangular arrays and filters [6,7] The problem of optimal design for randomly spaced arrays, or 3-D conformal arrays which must follow some 3-D curve, is not well understood. The same is true for non-square sample grids for 2-D filters [6]. One widely used array design technique involves projecting the array element positions onto a plane and then weighting the elements with a planar equiripple approximation [9]. This can produce poor results in many cases. The 1/P programming design approach is not affected by the non-uniform sample spaces or array configuration, and the performance of the algorithm for 1-D filters suggests that we may be able to achieve near optimum results in these other problems. In particular, the 1/P solution for conformal or random array design should yield a maximally sparse array. These potential applications will be addressed in our continued analysis of the 1/P simplex algorithm.

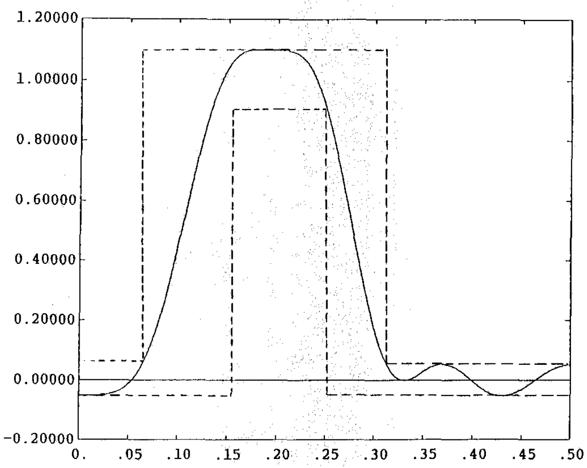

Fig. 1 Frequency response of the bandpass FIR filter designed using the 1/P simplex algorithm. The filter is of order 21 with 15 nonzero coefficients. The specification is marked with a dotted line.

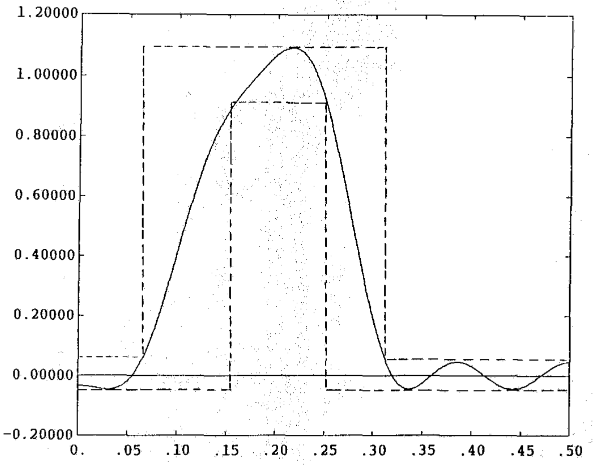

Fig. 2 Frequency response of the bandpass FIR filter designed to the specification of Fig. 2 using the equiripple design algorithm in [5].

1. Seismic Deconvolution

The 1/P simplex algorithm was also applied to the seismic deconvolution problem. A seismic signal was simulated by convolving a reflectivity sequence with a causal wavelet generated by sampling the impulse response of a 4th order ARMA filter, Fig. 3. The resulting signal was corrupted with 20dB additive Gaussian noise.

The linear system was set up to include slack variables in a similar manner to that described in section 5 to allow for noise in the data. The formulation differs slightly since rather than constraining each error in the error vector, $\underline{\Psi}=(\underline{B}-\mathbf{H}\underline{Q})$, the sum of the magnitude of the components of $\underline{\Psi}$ was bounded using a single additional constraint equation. The effect of this modification is to allow larger errors at a few sample points provided the total error is not too large. The result of applying the nonlinear simplex algorithm to this problem is shown in Fig. 4. Note that all of the main events are recovered at the correct locations and at almost the correct amplitude.

The second test used an identical source wavelet and reflectivity sequence as above but was modified to include an additional backscatter term by adding white Gaussian noise, with a 20dB SNR, to the reflectivity prior to convolution with the wavelet. This is introduced to model the effect of scattering of the wavelet between layer boundaries [3]. It should be noted we are primarily interested in recovering the major events only. The resulting data was then corrupted with a 20dB white Gaussian noise. The result of applying the new algorithm to this data is shown in Fig. 5. Again all of the main events are detected in the correct location with approximately the correct amplitude Note also that a large number of the remaining elements of the sequence are zero as expected.

6. References

[1] D. Luenberger, *Linear and Nonlinear Programming,* 2nd Edition, Addison-Wesley, 1984.

[2] R. Leahy and B. Jeffs, "A DSP algorithm for minimum order solutions" Proc. 21st Asilomar Conf. Signals, Syst. Comp., Nov. 1987.

[3] J. Mendel, *Optimal Seismic Deconvolution,* Academic Press, New York, 1983.

[4] G. Boudreaux and T. Parks, "Thinning Digital Filters: a Piecewise-Exponential Approximation Approach", IEEE Trans. Acoust. Speech Signal Proc., Vol ASSP-31, pp 105-112, 1983.

[5] T.W. Parks and J.H McClellan, "Chebyshev Approximation for Nonrecursive Digital Filters with Linear Phase," IEEE Trans. Circuit Theory, CT-19, no. 2, Mar. 1972, pp 189-94.

[6] D.E. Dudgeon, R.M. Mersereau, *Multidimensional Digital Signal Processing,* Prentice-Hall, Englewood Cliffs, 1984.

[7] S. W. Autrey, "Approximate synthesis of non-separable design responses for rectangular arrays", IEEE Proc. Anten. Prop., Vol AP-35, pp 907-912, 1987.

[8] L. R. Rabiner, B. Gold, *Theory and Application of Digital Signal Processing,* Prentice-Hall, Englewood Cliffs, 1975.

[9] C. A. Balanis, *Antenna theory, analysis and design,* Harper and Row, New York, 1982.

[10] A. L. Van Buren, "Improved Performance of receiving arrays in the presence of localized near field noise sources", Jou. Acoust. Soc. Amer., Vol. 69, pp 681-688, 1981.

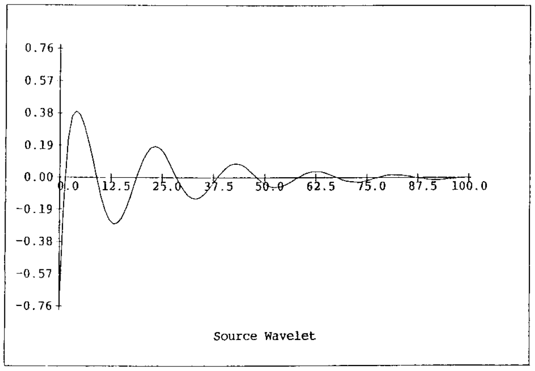

Fig. 3 Impulse response of the 4th order ARMA wavelet used in the seismic deconvolution example.

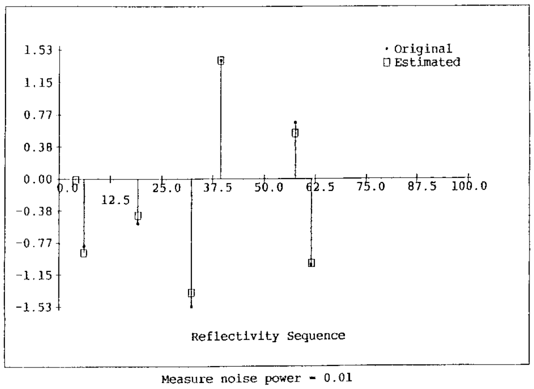

Fig. 4 Reconstruction of the reflectivity sequence from the data in Fig. 5 using the 1/P algorithm. Note that all the events are recovered.

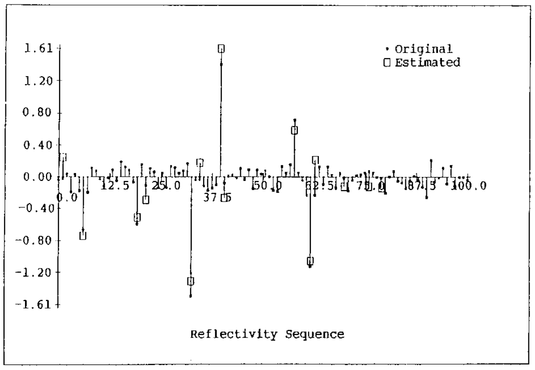

Fig. 5 Reconstruction of the reflectivity sequence from a second set of data in which a 10dB SNR backscatter sequence is added to the reflectivity sequence before convolving with the wavelet in Fig. 4. The data was also corrupted with 20dB SNR additive noise.