USC–SIPI REPORT #138

Maximally Sparse Constrained Optimization for Signal Processing Applications

by

Brian Dean Jeffs

January 1989

Signal and Image Processing Institute
UNIVERSITY OF SOUTHERN CALIFORNIA
Department of Electrical Engineering-Systems
3740 McClintock Avenue, Room 404
Los Angeles, CA 90089-2564 U.S.A.

This Dissertation Is

Dedicated to

Karen Jeffs

Acknowledgments

I would like to thank Professor Richard Leahy for his continual encouragement and help. His assistance with this project was above the call of duty as an advising professor. Without his suggestions and input, this dissertation would not have been possible.

Special recognition must go to Karen Jeffs, who supported and encouraged me through the long years of effort. Her sacrifice and faith in my ability to accomplish this have made all the difference.

This work was supported in part by the Hughes Aircraft Company Doctoral Fellowship Program. This far sighted program provided the financial support and work schedule flexibility necessary for completion of this degree.

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

A new approach for vector space optimization is presented which enables solution of a number of signal processing problems, otherwise solvable only in a suboptimal sense. The method seeks a maximally sparse solution vector to a system of linear or quadratic inequality constraints. An optimal solution (typically not unique) contains the fewest possible nonzero terms consistent with the constraints. The relationship between l_p quasinorm minimization and sparse optimization is discussed, and it is proved that for some p, $0<p<1$, the constrained l_p minimum will be maximally sparse. The related $l_{1/q}$ cost function is shown to be a superior objective functional for deterministic sparse optimization, and to yield maximum a-posteriori estimates when random sources are distributed as generalized p-Gaussian.

Three new algorithms based on the $l_{1/q}$ cost are presented. The $l_{1/q}$ simplex search algorithm achieves strong local optimality by searching the set of vertices of the convex polytope formed by linear constraints. The "basic" solutions corresponding to these vertices are proved to include the optimum. The stochastic search algorithm finds globally optimum results using techniques of simulated annealing to direct the simplex search. The convex transformation gradient search finds strong local optima of the quadratically constrained problem by transformation to a space where gradient search techniques are more successful.

Three applications are presented as examples of problems best solved using the maximally sparse criterion. Neuromagnetic image reconstruction produces 3-D maps of brain neural currents by reconstruction from externally measured induced magnetic fields. Sparse optimization is shown to be superior to other reconstruction methods which obscure virtually all current dipole position detail. Seismic deconvolution of sparse reflectivity sequences is demonstrated using all three of the algorithms, and design of thinned beamforming arrays using the stochastic search algorithm is shown to yield great improvement over existing methods. The new algorithms provide the only available method for thinning of arbitrarily shaped arrays in an optimal sense.

CHAPTER 1: INTRODUCTION

1.1. Problem Definition

This dissertation addresses the problem of finding the maximally sparse solution vector to a system of linear inequality constraints, i.e. a feasible solution with the maximum number of zero valued elements. The obvious formulation of this problem is the nonlinear mathematical program:

$$\min_{\underline{x}} f(\underline{x}) = \sum_{i=1}^{N} \mathbf{I}(x_i) \quad such \quad that \quad |\mathbf{H}\underline{x} - \underline{b}| \leq \underline{\varepsilon} \tag{1.1}$$

$$where \quad \mathbf{H} \in R^{MxN}, \; and \; \mathbf{I}(x_i) = \begin{cases} 1 & x_i \neq 0 \\ 0 & x_i = 0 \end{cases}$$

Here $\underline{x}$ is the solution vector, $\mathbf{H}$ the system matrix, $\underline{b}$ the measurement vector, and $\underline{\varepsilon}$ the error constraint. No acceptable algorithm was found in the literature for solving (1), but a related nonlinear program was identified which yields maximally sparse results:

$$\min_{\underline{x}} g(\underline{x}) = \sum_{i=1}^{N} |x_i|^{1/q} \quad such \; that \quad |\mathbf{H}\underline{x} - \underline{b}| \leq \underline{\varepsilon}, \; q > 1 \tag{1.2}$$

It will be shown that for $q >> 1$, $g(\underline{x})$ approximates $f(\underline{x})$, and provides a more suitable objective for iterative maximally sparse optimization. With some restrictions on the value of q, optimal solutions to (1.2) are optimal for (1.1). Since for $0<p<1$ we have $[g(\underline{x})]^q = \|\underline{x}\|_{lp}$, the l_p quasi-norm of $\underline{x}$ for $p = 1/q$, solving eqn (1.2) is equivalent to

constrained l_p optimization. Though for p values in this range, optimization is difficult, it is a significant improvement over the indicator function representation of (1.1). Eqn (1.2) will be referred to as an $l_{1/q}$ program.

The primary concern of this work is the development of practical algorithms for finding maximally sparse solutions to real signal processing problems. Three algorithms based on the $l_{1/q}$ program are presented. The utility of this class of problems became apparent to the author while trying to reconstruct, from the externally measured magnetic fields, 3D neuromagnetic images of brain neuron currents [1]. With this and other applications, discussed in Chapters 4, 5, and 6, it was found that the common approaches for constrained optimization using l_2 or l_∞ norms, entropy maximization, or other objective functions produced poor results. In these cases the desired solution, which best represented the true underlying source, was very sparse. Results produced by the common optimization criteria were overly smoothed and showed no physical relationship to the true source, even though they were consistent with the sampled measurement data. It is suggested that a large class of signal processing and more general problems exists which would benefit from application of the minimum order criterion. In any problem where extremely "spiked" results are expected, or where the incremental cost of adding a nonzero term to the solution outweighs the cost of increasing an already nonzero term, the minimum order solution is desirable. Applications of the technique are presented for three signal processing problems, including neuromagnetic image reconstruction (NMI), seismic deconvolution problems, and sparse beamforming array design.

1.2. Motivation and Background

It is proposed that a potentially broad range of applications is untapped due to the lack of practical maximally sparse optimization algorithms in the digital signal processing (DSP) literature. Some papers are discussed below which deal with several related problems, including some which tend to sparse results, but few address optimality. The dearth of such applications is likely due to the difficulty in finding optimal solutions, rather than a lack of potential uses. In the short duration of this present research, the availability of a useable algorithm has led to several useful applications, and it is suggested that the surface has barely been scratched for a large class of related problems.

This dissertation concentrates on applications in various fields of DSP, but it is expected that many more applications exist in the areas of operations research and general optimization theory. To introduce the basic elements of a generic maximally sparse problem, the following simplified hypothetical resource allocation problem is presented. Consider an outdoor public warning alarm system design problem. There are N potential amplifier/loudspeaker sites, s_i, to provide reliable coverage for M listening locations, r_i. Further assume that we are in obstructed conditions where the expected signal path strength (1/attenuation), a_{mn} from speaker n to receiver m varies dramatically, and is effectively zero in same cases. The constraints are that each listening site must receive at least enough total acoustic energy (assuming incoherent summing) to cross the threshold, t, of reliable hearing, and must also be below the level, d, which could cause ear damage. It costs money to build higher powered amplifiers and speakers, but the cost of installation, cable runs, etc., make it desirable to use as few speaker sites as possible, regardless of how powerful we need to make them. If received power at all sites is given by $\underline{r}$ and transmitted power is $\underline{s}$, then the formulation is:

$$\underline{r} = \mathbf{A}\underline{s} \qquad \mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ \cdot & & & \cdot \\ a_{M1} & a_{M2} & \cdots & a_{MN} \end{bmatrix} \tag{1.3}$$

$$\min_{\underline{s}} f(\underline{s}) = \sum_{i=1}^{N} \mathrm{I}(s_i) \text{ such that } \begin{bmatrix} -\mathbf{A} \\ \mathbf{A} \end{bmatrix} \underline{s} \le \begin{bmatrix} -\underline{t} \\ \underline{d} \end{bmatrix} \quad \underline{s} \ge \underline{0}$$

A solution to this equation will yield a system design with the fewest possible speaker sites. For a specific example, if we let

$$\mathbf{A} = \begin{bmatrix} 4 & 3 & & \\ 2 & 4 & 2 & \\ & 4 & 1 & \\ & & 4 & \end{bmatrix}, \quad t = 1, \; d = 24$$

then some of the (not unique) maximally sparse solutions are:

$$\underline{s} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 6 \\ 0 \\ 6 \end{bmatrix}, \begin{bmatrix} 1/2 \\ 0 \\ 0 \\ 1/4 \end{bmatrix}, \begin{bmatrix} 0 \\ 1/3 \\ 0 \\ 1/4 \end{bmatrix} \cdots$$

Using $g(\underline{s})$ with $q = 2$, rather than $f(\underline{s})$, as the objective we obtain the unique optimum

$$\underline{s} = \begin{bmatrix} 0 \\ 1/3 \\ 0 \\ 1/4 \end{bmatrix}$$

Two classes of DSP problems have been identified to which the minimum order criterion may be usefully applied. The first class includes system design problems, as above, in which the dominant cost is the number of nonzero components required to meet a given set of constraints. An example of this class, which is treated below, is the design of an array beamformer to meet a given spatial response using the minimum number of array elements (Chapter 6). Other potential applications include the design of minimum

computation FIR filters [2]. The second type of application lies in the processing of signals. Results which demonstrate this class include seismic deconvolution [2] (Chapter 5) and imaging of current dipole sources as a means of locating neural activity in the human brain from external magnetic field measurements [1] (Chapter 4). Another interesting potential application is the reconstruction of star source images in radioastronomy [3].

Previous related work includes both mathematical optimization for concave cost functions, and applications where sparse solutions were sought. In [4], the basic behavior of linearly constrained l_p optimization problems for $0<p<1$ is discussed, including examples of how the solution changes in a stepwise fashion as p is varied over this range. The theory of quasi-Banach spaces based on the l_p quasinorm, $0<p<\infty$, has been studied, and is discussed, for example, in [5]. Equation (1.2) is also related to the linearly constrained concave minimization problem, for which a number of global optimization algorithms have been proposed, based on collapsing polytopes [6,20] and branch and bound procedures [6,7]. While these methods do achieve global optima, they are probably computationally infeasible for the large dimensions ($N \approx 100$) considered in this paper, due to the use of multiple nested linear programming subproblems and the assumptions necessary to apply them to the general form of eqn (1.2).

For blind deconvolution of seismic reflectivity data, several authors have discussed the need for optimization norms which increase sparseness or minimize the entropy of the solution. These authors have proposed the use of the "varimax," "parsimonious," and l_p norms with $p<1$ [8,9,10]. These examples are closely related to our problem, however the algorithms used do not in general find the global optimum, and are not directly applicable to the form of eqn (1.2).

In the area of sparse image reconstruction, linear programming has been applied successfully [11] but without explicitly seeking the maximally sparse image. Also, a technique of "beam subtraction" has been employed to restore sparse astronomical star field images [3]. Design of sparse beamforming arrays, or array "thinning," has also appeared in the literature [12] but the author knows of no published approach for optimally thinned arrays of arbitrary shape.

Additional historical information will be presented in detail with each application discussed in Chapters 4, 5, and 6.

1.3. Research Goals

The primary goals which have motivated this research effort are presented below, and include theoretical analysis, algorithm development, and application to engineering problems. These goals have been achieved to varying degrees, as discussed in the following Chapters.

1) **Investigate feasibility of maximally sparse optimization**

   The primary focus of this research is the investigation and characterization of the linearly constrained, minimum order (maximally sparse) optimization problem. Although the need for such an approach was suggested by difficulties encountered in NMI applications, it was apparent from study of the literature that little had been accomplished in the field, and the feasibility of minimum order optimization for problems of moderately large dimensionality was in question. Demonstration that the problem could indeed be solved efficiently was paramount to this effort.

2) **Develop a theoretical foundation for maximally sparse optimization**

   Before a useful algorithm can be developed, a theoretical understanding of the problem must be acquired. A goal of this project was to understand the nature of

the difficulties posed by maximally sparse optimization, relate the problem to other classes of optimization which were more tractable, and provide a theoretical, rather than heuristic, basis for algorithm design. The greater emphasis has been deterministic optimization, but analysis of related probabilistic models was also a priority.

3) **Develop practical algorithms for maximally sparse optimization**

Since no usable algorithms were found for the specific class of problems studied, a major goal was to develop algorithms which were significantly faster than an exhaustive search, and which gave globally optimum or nearly optimum results. These algorithms could not be application specific, but must be useable in a wide range of general related problems. They must deal efficiently with problems of moderately large dimensionality (several hundred variables).

4) **Demonstrate practical DSP applications using the algorithms**

This research grew out of the effort to develop a practical NMI reconstruction algorithm when it became apparent that none of the conventional algorithms were acceptable. In this sense, the theoretical and algorithm developments were application driven, and NMI was the first application considered for the new algorithms. The maximally sparse approach, and the algorithms developed, were novel enough to warrant investigation into a broader use of the technique. A major goal of the research was to validate the algorithms' usefulness, and demonstrate their scope of application, by successful use in solving several real engineering problems from different branches of DSP study. It is expected that a significant pool of potential applications exist, and hoped that this demonstration of practical

algorithms will motivate a broader interest in the community in a maximally sparse approach to problem solving.

For the applications presented in this dissertation, the specific goals are as follows. In NMI the primary goal is three dimensional localization of distributions of discrete current dipole sources in a conducting volume based on synthesized magnetic field measurements exterior to the volume. This requires a reconstruction algorithm which incorporates the known physics of the system and overcomes the blurring and position bias encountered with existing algorithms. For deconvolution problems the hope is to obtain more accurate estimation of sparse sources than is achieved by algorithms which do not make the maximally sparse assumption. For beamforming array design, the goal is to demonstrate an improved method for designing thinned arrays and for element placement in arbitrary shaped arrays.

1.4. Contributions of Completed Research

The research presented in this dissertation has made the following contributions and accomplishments in the fields of optimization theory and application, biomedical image reconstruction, and digital signal processing.

a) The importance of a class of linearly constrained, maximally sparse optimization problems was identified and demonstrated by theoretical analysis, algorithm development, and successful application to several signal processing problems. This class of problems has received little attention in the engineering literature, but it is felt that this research justifies more extensive study in the field.

b) The close relationship between maximally sparse optimization and l_p norm minimization was recognized, and a theorem was proved to show under what condi-

tions the two problems are equivalent. This relationship was a key finding because it is the basis for the sparse optimization algorithms presented here.

c) A "fundamental theorem for $l_{1/q}$ programming" was proved to show that for $q>1$, the $l_{1/q}$ minimization problem is optimized at the "basic" solutions of the linear constraint equations. This too was essential to development of the algorithms.

d) Three new or adapted algorithms have been developed for maximally sparse optimization. The $l_{1/q}$ simplex search, for linear inequality constraint problems, is an efficient finite time algorithm which uses a tableau pivoting approach similar to linear programming to find good locally optimal sparse solutions. The stochastic search algorithm applies simulated annealing techniques to achieve asymptotically optimal results while pivoting the simplex tableau. The convex transformation gradient search algorithm performs a nonlinear transformation on the system to enable use of gradient search techniques for sparse optimization of a quadratically constrained system.

e) A method has been proposed for reconstructing, from the externally sampled induced magnetic field, neuromagnetic images (NMI) of the electrical activity in the human brain . A mathematical model suitable for reconstruction was developed, and analysis performed to demonstrate the inherent ill-conditioned nature of the problem. Conventional reconstruction algorithms were analyzed and their inappropriateness for this application was demonstrated. The superior performance of a maximally sparse, minimum current dipole, approach was demonstrated by reconstructing simulated 3-D current distributions from simulated magnetic field measurements using the $l_{1/q}$ simplex search algorithm.

f) The use of maximally sparse optimization for deconvolution problems was demonstrated. Both the $l_{1/q}$ simplex search and the convex transformation gradient

search algorithms were used successfully in deconvolving synthesized seismic reflectivity sequence data, with results comparable to existing methods.

g) Design of thinned beamforming arrays was demonstrated using the $l_{1/q}$ simplex search and stochastic search algorithms. Examples are given which show significant improvement over other array thinning approaches found in the literature. Application to array element placement for arbitrarily shaped 3-D arrays is also demonstrated.

h) The stochastic search algorithm introduced here is the only currently available method for finding true maximally sparse results for beamforming array thinning and other signal processing applications. Optimally thinned array design had not been possible for other than special case array configurations. Earlier methods are either add-hoc, sub-optimal, or are limited to simple element configurations (e.g. small or linear). This algorithm uses an optimization theoretic approach which is much more general and powerful, and can be applied to other signal processing applications which require maximally sparse results and can be expressed in linear inequality constraint form. In seismic deconvolution based on minimizing some measure of sparseness [8,9,10,71], the typical algorithms used cannot produce globally optimum results, while the stochastic search is well suited to the problem.

This work has also been the source for one published [1] and one submitted journal article [13], and four conference papers [2,14,15,16].

1.5. Dissertation Organization

The remaining Chapters of this document are organized as follows. Chapter 2 discusses the theoretical aspects of maximally sparse optimization and presents two theorems which

are the basis for the optimization algorithms. It also examines the relationship between the deterministic approach which is the primary topic of this work, and a parameter estimation interpretation where the data is modeled as generalized p-Gaussian distributed random data. This probabilistic view gives some justification for accepting results from the three algorithms presented below when noise or other random data is involved.

Three algorithms for maximally sparse optimization are presented in Chapter 3. The theoretical development specific to each algorithm is discussed and detailed algorithm descriptions are given. The use of these algorithms is demonstrated in Chapters 4, 5, and 6, which discuss three major applications of maximally sparse optimization. Chapter 4 is a somewhat self-contained study of neuromagnetic image reconstruction. It includes a thorough discussion of biomagnetism fundamentals and a study of the effectiveness (or lack thereof) of other reconstruction algorithms. Current source reconstruction from simulated sample data is used to demonstrate effectiveness of the $l_{1/q}$ simplex search. Chapter 5 deals with seismic deconvolution problems, and includes demonstration of the $l_{1/q}$ simplex search and convex transformation gradient search algorithms. Chapter 6 shows how the $l_{1/q}$ simplex search and stochastic search algorithms can be used in the design of thinned beamforming arrays. Section 9 includes the appendices where proofs of theorems may be found.

CHAPTER 2: THEORETICAL FUNDAMENTALS OF MAXIMALLY SPARSE OPTIMIZATION

2.1. Objective Functionals, Minimum Order from l_p Optimization

As an optimization problem, eqn (1.1) is particularly difficult to solve. We are plagued with numerous local minima, and $f(\underline{x})$ is discontinuous and has zero gradient except at the discontinuities. In an effort to overcome these limitations, an approach to the minimum order problem is proposed which is based on generalized linearly constrained l_p optimization. Figure 2.1 illustrates the unit ball surfaces in R^2 space for the quasi-norm $\|\underline{x}\|_{l_p} = (\sum_{i=1}^{N} |x_i|^p)^{1/p}$ for values of p in the range $0 \leq p \leq \infty$. For $p \geq 1$ we have the conventional l_p norm, which is a convex functional and obeys the triangle inequality. Since for $p \geq 1$, the linear constraints in (1.2) form a convex set it is well known that any local minimum of $\|\underline{x}\|_{l_p}$ satisfying the constraints is a global optimum. Many efficient algorithms exist for solving such problems [17,18]. Of particular interest are the cases for values of p=1,2, and ∞, corresponding to linear, quadratic, and minimax objective, which form the basis of many widely used optimization procedures. However the resulting solutions for these do not achieve the sparse results of interest in this dissertation.

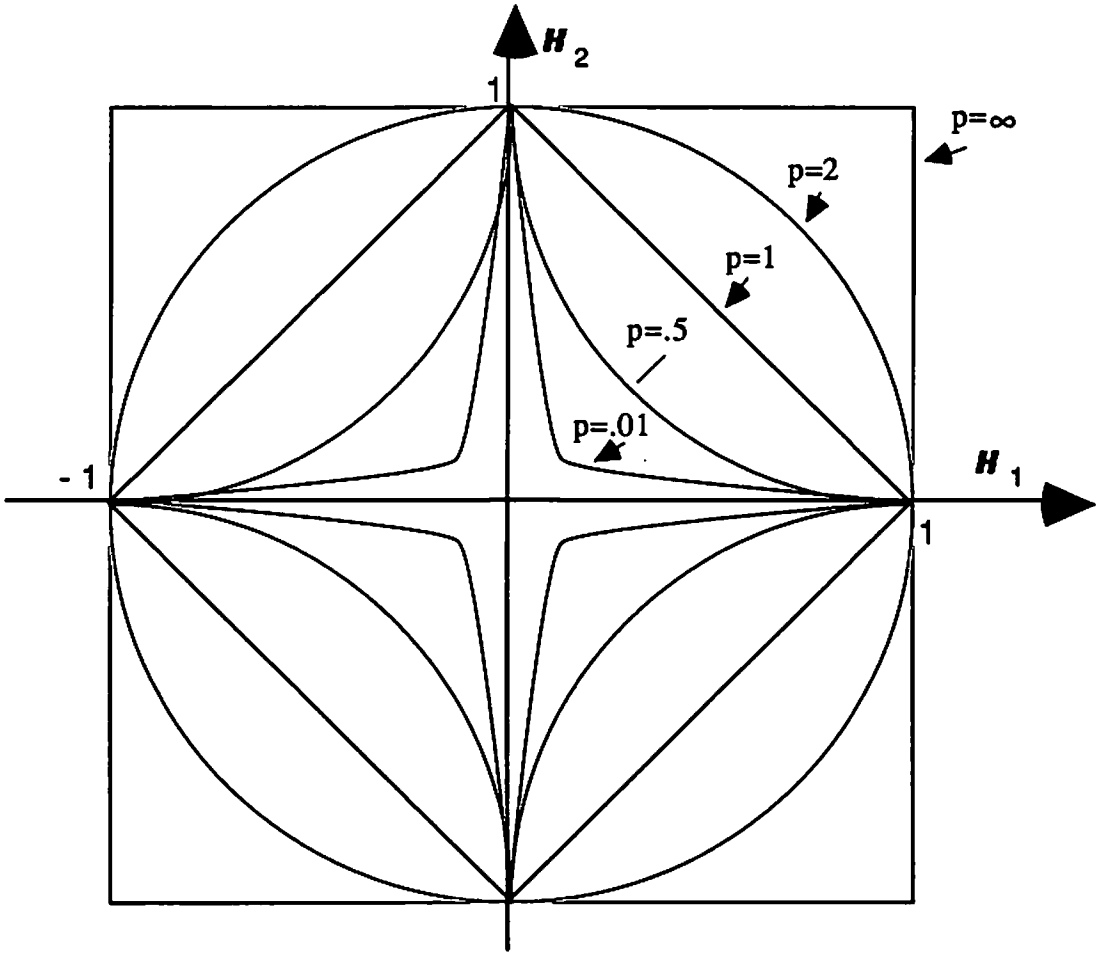


Figure 2.1. Unit Balls of the l_p Norm for Various p.
Note that as p approaches 0, the unit ball approaches the axes.

For $0<p<1$, l_p is only a quasi-norm [5], since the triangle inequality does not hold, and in fact the inequality is reversed for positive x_i . Over R^N, $\|x\|_{l_p}$ is neither convex nor concave, containing many strong local minima and presenting a difficult optimization problem. Large values of p result in smooth solutions, however, as $p \rightarrow 0$ the solutions tend to become more "spikey," or sparse [8].

The reason for this can be seen in Figure 2.1. As $p \rightarrow 0$, the curves in Figure 2.1 approach the x_1, x_2 axes, on which the unit ball lies for $f(\underline{x})$ in eqn (1.1). We identify minimum order optimization as a special case of generalized l_p optimization. Since with $g(\underline{x}) = (\|\underline{x}\|_{l_p})^p$ for $p = 1/q$, we have

$$\lim_{q \rightarrow \infty} g(\underline{x}) = \sum_{i=1}^{N} |x_i|^{1/q} = \sum_{i=1}^{N} \mathbf{I}(x_i) = f(\underline{x}) \tag{2.1}$$

This suggests that we may use (at least in the limiting case) eqn (1.2) instead of eqn (1.1) for sparse optimization. In the following section equivalence is proved for finite q. The utility of this observation is that for q finite, $g(\underline{x})$ eliminates some of the handicaps of $f(\underline{x})$. $g(\underline{x})$ is continuous everywhere and differentiable except at the axes. Gradients may be computed for all nonzero terms. This enables use of gradient search techniques at least for finding local minima. Section 3.1 presents a finite extreme point search algorithm which also benefits from the use of $g(\underline{x})$ rather than $f(\underline{x})$. For reasonably small values of q, $g(\underline{x})$ is computationally stable, and thresholds are not needed to handle inexact zero values. Also, $g(\underline{x})$, unlike $f(\underline{x})$, can provide some discrimination in cost between solutions of equal order, thus avoiding a stalled search at a point surrounded by "adjacent" solutions of equal cost. Adjacency relies on the concept of basic feasible solutions (BFS), presented in the following section, such that two BFS are defined as adjacent if they contain the same set of nonzero terms, except for one variable.

2.2. Fundamental Theorem of $l_{1/q}$ Programming

To facilitate the development, two specific forms of eqn (1.2) are introduced, for $\underline{\varepsilon} = 0$ and $\underline{\varepsilon} \neq 0$ as follows.

For $\underline{\varepsilon} = 0$ we have:

$$\min_{\underline{x}} \; g(\underline{x}) = \sum_{i=1}^{N} |x_i|^{1/q} \; s.t. \;\; \mathbf{H}\underline{x} - \underline{b} = \underline{0}, \;\; q > 1 \qquad (2.2a)$$

Note that unlike linear programming (LP), no positivity constraint on $\underline{x}$ is needed, as shown by theorem 1 below. This form may be applied directly to a version of the $l_{1/q}$ simplex algorithm which allows bipolar valued variables.

For $\underline{\varepsilon} > 0$ the form of (1.2) is

$$\min_{\tilde{\underline{x}}} \; g(\tilde{\underline{x}}) = \sum_{i=1}^{2N} (\tilde{x}_i)^{1/q} \; s.t. \;\; \tilde{\mathbf{H}}\,\tilde{\underline{x}} - \tilde{\underline{b}} = \underline{0}, \; \tilde{\underline{x}} \geq \underline{0}, \; q>1 \qquad (2.2b)$$

where

$$\tilde{\mathbf{H}} = \begin{bmatrix} \mathbf{H} & -\mathbf{H} & \mathbf{I} & \mathbf{0} \\ \mathbf{H} & -\mathbf{H} & \mathbf{0} & -\mathbf{I} \end{bmatrix}, \; \tilde{\underline{x}} = \begin{bmatrix} \underline{x}^+ \\ \underline{x}^- \\ \underline{s}^+ \\ \underline{s}^- \end{bmatrix}, \; \tilde{\underline{b}} = \begin{bmatrix} \underline{b} + \underline{\varepsilon} \\ \underline{b} - \underline{\varepsilon} \end{bmatrix}$$

$$\underline{x}^+, \underline{x}^- \in R^N, \; \underline{s}^+, \underline{s}^- \in R^M, \; \tilde{\underline{x}} \in R^{N'}, \; \tilde{\underline{b}} \in R^{M'}, \; N' = 2N+2M, \; M' = 2M$$

$\underline{s}^+$ and $\underline{s}^-$ are respectively slack and surplus variables as commonly employed in linear programming [18]. Note these variables are not included in the computation of $g(\tilde{\underline{x}})$. This form allows us to confine the search to the nonnegative orthant of the space $R^{N'}$. Theorem 1C below provides the equivalence of equations (2.2b) and (1.2) for $\underline{x} = \underline{x}^+ - \underline{x}^-$. The introduction of slack and surplus variables, positivity constraints, and use of a form of the $l_{1/q}$ simplex algorithm which allows pivots only to positive valued solutions, are needed to deal with the inequality in eqn (1.2).

In solving the $l_{1/q}$ program, we lack the convexity properties which for l_p programming yield a well posed problem, and imply global optimality from local optimality. Over R^N,

the cost $g(\underline{x})$ is nonlinear and neither convex nor concave. Our problem would appear hopeless, but for the fortunate fact that for $q \geq 1$ we can limit the candidates to a finite set of "basic" solutions, which are the same as those defined for linear programming. This is indicated by the "Fundamental Theorem of $l_{1/q}$ Programming" presented here, with proof in Appendix A. Basic solutions to eqns (2.2a) or (2.2b) satisfy the usual definition requiring $\underline{x}$ and $\underline{\tilde{x}}$ to meet the corresponding equality constraints, and contain at most M or M' nonzero components respectively. Additionally, we say any $\underline{\tilde{x}}$ of eqn (2.2b) is "properly basic" if it is basic and $(\underline{x}^+)^T (\underline{x}^-) = 0$.

Theorem 1A

Given a problem of the form (2.2a) or (2.2b), if a solution exists, then a basic solution exists.

Theorem 1B

If a global optimal solution to eqn (2.2a) exists, it is a basic solution, and is globally optimal to eqn (1.2) for $\underline{\varepsilon} = \underline{0}$

Theorem 1C

If a globally optimal solution to (2.2b) exists, then a properly basic globally optimal solution exists, furthermore, this solution implies $\underline{x} = \underline{x}^+ - \underline{x}^-$ is a globally optimal solution to eqn (1.2) for $\underline{\varepsilon} > \underline{0}$

The basic solutions of the forms (2.2a) and (2.2b), are isomorphic with the vertices of the convex polyhedron defined by their constraints in R^N and R^{2N+2M} space respectively [18]. Consequently, we can restrict our search to these vertices, since one must be the optimal solution. There are potentially $O\binom{N}{M}$ of these vertices, making an undirected search of even this finite set impractical for moderately large M and N. These properties motivate us to use a procedure similar to the linear programming simplex algorithm,

traversing the vertices while monotonically reducing the cost. Due to the nonlinear nature of the cost, modifications to the standard l_p algorithm are required, and globally optimal solutions are not assured. Sections 3.1 and 3.2 present algorithms which are based on the properties described by the fundamental theorem. Existing algorithms for this and similar problems are also presented.

2.3. Equivalence Theorem for $l_{1/q}$ Optimization

If we must allow $q \rightarrow \infty$ before eqn (1.2) leads to a solution of eqn (1.1), then we cannot benefit from the practical advantages of $g(\underline{x})$ mentioned in section 2.1. Theorem 2 provides justification for minimum order optimization based on minimizing $g(\underline{x})$, by demonstrating that for a bounded solution set there exists a finite q_1 such that for all $q > q_1$, any solution to eqn (1.2) is a solution to eqn (1.1). As in linear programming, we define a basic feasible solution to an $M \times N$ system of linear equalities to be any solution containing at most M nonzero terms.

Theorem 2: Let S denote the set of all basic feasible solutions to $\mathbf{H}\underline{x} = \underline{b}$ s.t. $\mathbf{H} \in R^{M \times N}$. If the solutions in S are bounded, then $\Omega = \max_{\underline{x} \in \mathrm{S}} [l_\infty(\underline{x})]$ is finite.

Let $\varepsilon = \min_{x_i \in \mathrm{S}, 1 \le j \le N} \left\{ |x_{ij}| \; s.t. \; x_{ij} \neq 0 \right\}$. i.e. ε is the smallest nonzero magnitude of any element of any vectors in S.

Given $\varepsilon > 0$ and $\Omega < \infty$, if V is the set of all globally optimal solutions to

$$\min_{\underline{x}} f(\underline{x}) = \sum_{i=1}^{N} \mathbf{I}(x_i) \quad \text{such that} \quad \mathbf{H}\underline{x} = \underline{b} \tag{2.3}$$

with $r = f(\underline{x})$ for any $\underline{x} \in V$ (i.e. r is the optimal solution order), and U is the set of all globally optimal solutions to

$$\min_{\underline{x}} \; g(\underline{x}) = \sum_{i=1}^{N} |x_i|^{1/q} \quad \text{such that} \quad \mathbf{H}\underline{x} = \underline{b}\,, \; q>1 \tag{2.4}$$

then if $q \geq q_1$, U is a subset of V, where

$$q_1 = \frac{log\left(\frac{\Omega}{\varepsilon}\right)}{log\left(\frac{r+1}{r}\right)} \tag{2.5}$$

Proof of Theorem 2 is found in Appendix B. In other words, given an upper bound on vector elements of the basic solutions in S, and a lower bound on nonzero element magnitudes, eqn (2.5) yields a finite q which insures $l_{1/q}$ minimization will lead to a global solution of the maximally sparse problem of eqn. (1.1).

Eqn (1.2) therefore defines a class of problems, indexed by q whose solutions are increasingly sparse as q increases, until $q>q_1$, where an optimally sparse solution is given. For $q \leq 1$ the optimal $\underline{x}$ changes continuously as a function of q, but for $q>1$, there is a finite number of optimal solutions. A given $\underline{x}_{opt}$ will remain optimal over a range of q values, and as q increases, we step from one solution to another in a discrete fashion [4]. This behavior is shown in Figure 2.2 for the problem presented in detail in section 3.1.4, (see Table 3.1 and Figure 3.1). In this example there are five variables and four basic solutions, with $\underline{x}_1$ optimal for $1 \leq q \leq 1.64$, and $\underline{x}_3$ optimal for $q>1.64$.

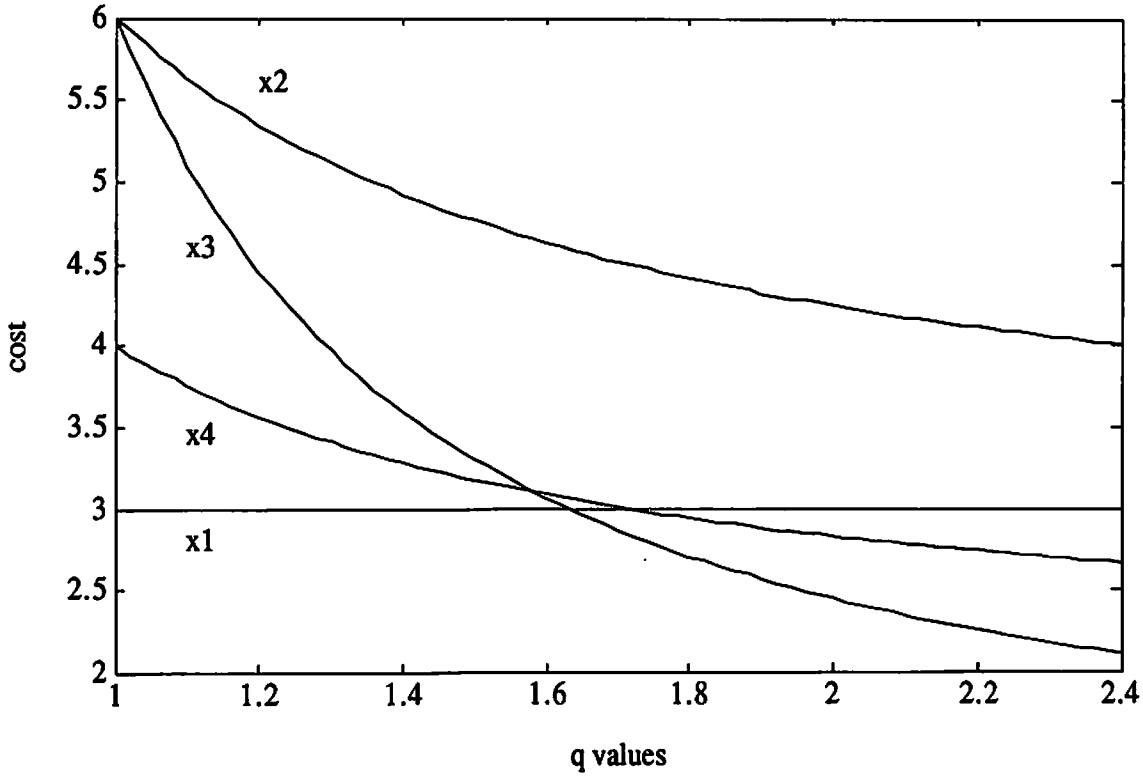


Figure 2.2. $l_{1/q}$ Costs for Basic Solutions as a Function of q.
The value of $g(\underline{x})$ is plotted for the basic solutions to the problem of eqn (3.10) and Figure 3.1. Note that only two different optimal solutions are found for all values of $q>1$.

It should be noted that since solutions to eqn (1.1) are not necessarily unique, and lacking any justification for accepting one over another, we are satisfied with any algorithm which will select one from the optimal set. Theorem 2 proves that solutions to eqn (1.2), for $q>q_1$, form a subset of solutions to eqn (1.1), so we accept any $l_{1/q}$ optimum. In order to improve the computational stability of an algorithm, we wish to use the smallest value of q which reasonably approximates $f(\underline{x})$. The q_1 as computed in Appendix B is a conservative upper bound, and in practice a much smaller value may often be used.

Figure 2.3 is an illustrative example of the effect of changing q on solving eqn (1.2) for

$$\mathbf{H} = \begin{bmatrix} 1 & 0 & -1 \\ 1 & .2 & 1 \end{bmatrix}, \underline{b} = \begin{bmatrix} 0 \\ 2 \end{bmatrix}, \underline{\varepsilon} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{2.6}$$

yielding the optimal solutions for various q as shown in Table 2.1. It is interesting to note that for this example the bound from Theorem 2 is correctly predicted with Ω=10, ε=1, and r=1 to be

$$q_1 = \log (10/1)/\log (2) = 3.32. \tag{2.7}$$

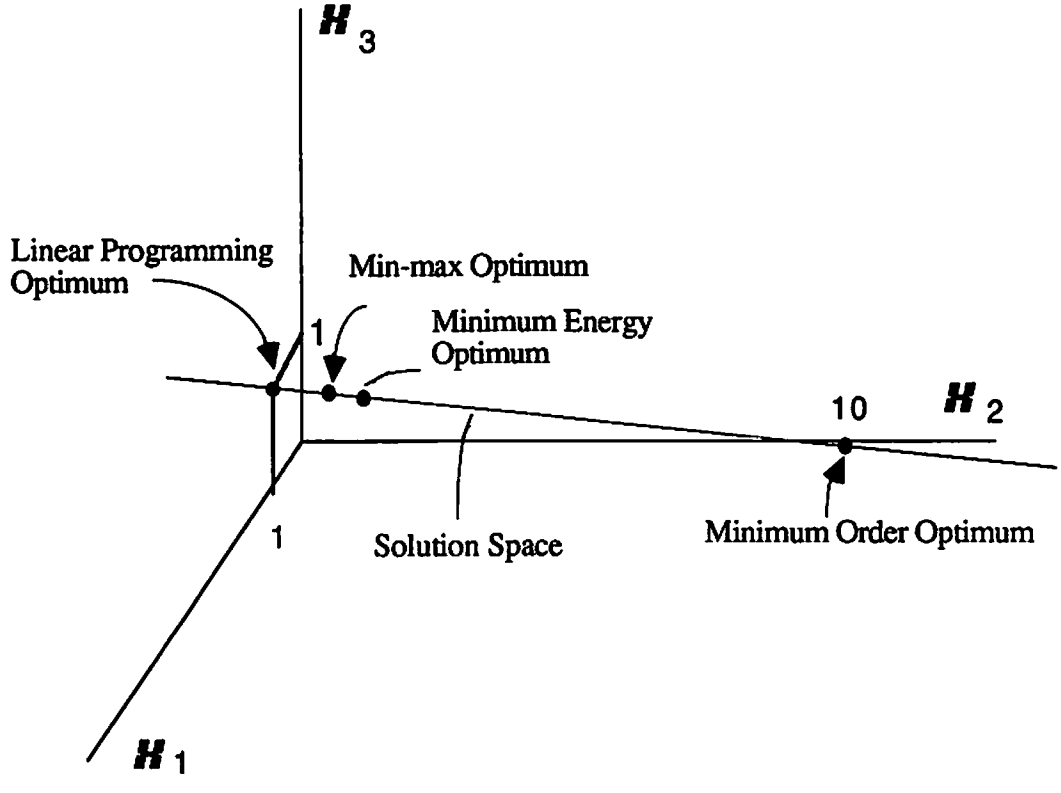


Figure 2.3. Example of $l_{1/q}$ Optimization for Various Values of q. Solutions to equation 2.6 show minimum order results for q>3.32.

Optimal Solutions v.s. q				
q values:	Solution Type:	x_1:	x_2:	x_3:
0	min-max	.91	.91	.91
.5	min energy	.98	.20	.98
$1 \le q \le 3.32$	linear programming	1.00	0.00	1.00
$q > 3.32$	min order (max sparse)	0.00	10.00	0.00

Table 2.1. Optimal Solutions to eqn (2.6) for Various Values of q.

2.4. Comparative Characteristics of $l_{1/q}$ Minimization

In this section the constrained $l_{1/q}$ minimization problem will be classified and characterized with respect to the broader field of general optimization problems. It will be shown that this problem is a member of one of the most difficult classes, and that the success of the algorithms presented in Chapter 3 is due to exploitation of unique properties of the specific objective function used.

2.4.1. Problem Classification

Of the endless variety of problems where one wishes to minimize some multivariate continuous objective functional over a given solution space, the majority of easily solved problems involve a quadratic objective and a convex constraint on the solution space [18,19]. For the general optimization problem,

$$\min_{\underline{x}} \; c(\underline{x}) \;\; \textit{such that} \;\; \underline{x} \in \Omega, \quad \Omega \subset R^N \tag{2.8}$$

Both the objective, $c(\underline{x})$, and the constraint space, Ω, can take many forms including linear, quadratic, convex, concave and nonlinear. The class of problem, and difficulty in finding global solutions are directly dependent on the forms of $c(\underline{x})$ and Ω. In general, the only problems for which finite global optimization algorithms exist across the class are the convex (including linear and quadratic) objectives with either convex (including linear) or no constraints. Some algorithms exist for global solution of special cases within a class. Beyond this, for other combinations of $c(\underline{x})$ and Ω, unless some special structure in the objective or constraint can be exploited, a global minimum may generally not be located in finite time. Examples of algorithms for cases related to $l_{1/q}$ minimization are given in Section 2.4.2.

The $l_{1/q}$ optimization problem of eqn (1.2) has a nonconvex, nonconcave objective function, with linear inequality constraints. As shown in Figure 2.4, the objective surface contains many ridges and valleys, and the numerous local minima are extremely strong, with infinite cost gradient at the minimum. These characteristics place $l_{1/q}$ minimization in the class of some of the most difficult problems. Our attempts at constrained gradient search optimization using a penalty method algorithm were completely unsuccessful due to the strong local minima.

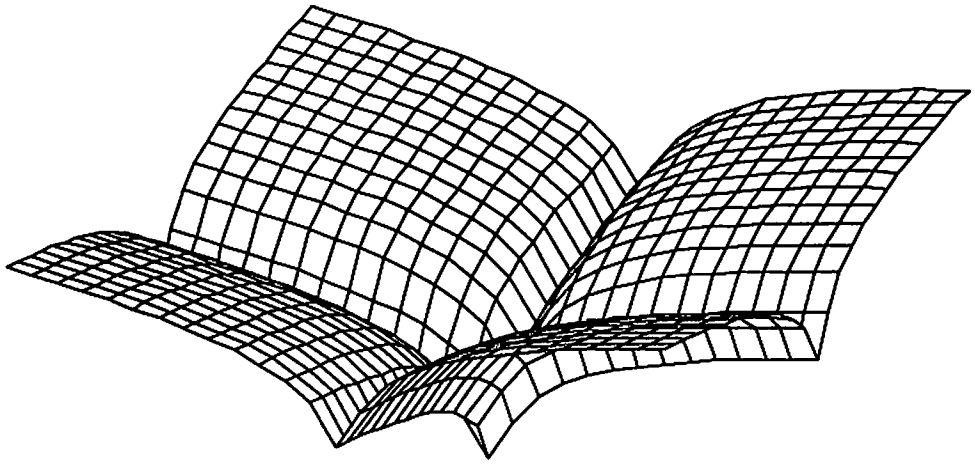


Figure 2.4. 2-D $l_{1/q}$ Cost Surface.
Values of $g(\underline{x})$ are plotted for for $q = 8$, $-2 \le x_1, x_2 \le 2$.

The fundamental theorem of section 2.2 enables us to transform continuous $l_{1/q}$ optimization to a combinatorial problem by limiting our search to the extreme points of the constraint space. The symmetry characteristics of $g(\underline{x})$ about each coordinate axis lead to an equivalent representation in R^{2N+2M} space, constrained to the positive only orthant, where $g(\underline{x})$ is strictly concave (see Appendix A). Since a concave function is minimized at the extreme points of a convex set, we may limit the search to the $\binom{N}{M}$ basic solutions of eqn (1.2). Although no proof is presented, it is believed that $l_{1/q}$ optimization (q>1) over the finite set of basic solutions, like the classical "traveling salesman problem," is a member of the class of *NP*-complete problems, for which no algorithm can guarantee solution in a number of iterations bounded by a power of N [21]. It is proved in [82] that the related problem of minimizing an arbitrary concave quadratic function over an arbitrary parallelepiped is *NP*-hard. Though the most simple minded exhaustive search algorithm

can deliver finite time (but not in our lifetime!) convergence in searching this set, algorithms with much better average performance are possible. The methods presented in Chapter 3 use the $l_{1/q}$ cost to direct an efficient search strategy through the basic solutions in locating either a global or good local minimum. Though eqn (1.1) could also be solved directly by searching the same basic solutions, $f(\underline{x})$ provides no information for a search direction when moving between solutions of equal order, and thus cannot provide the efficiencies of $g(\underline{x})$.

A related form of the problem, discussed in section 3.3 uses the nonlinear transformation

$$y_i = \frac{1}{q} ln(x_i) , \quad x_i > -\infty \tag{2.9}$$

to map the problem to a space where $g(\underline{y})$ becomes convex. This problem cannot be interpreted as combinatorial since Theorem 1 does not apply, and must be solved using conventional nonlinear constrained optimization algorithms. Here the continuous nature of $g(\underline{x})$, as opposed to $f(\underline{x})$ from eqn (1.1), is necessary for gradient computation in the descent algorithms, and permits use of exterior point search methods.

2.4.2. Related Existing Algorithms

For the general class of problems (of which $l_{1/q}$ minimization is a member) with arbitrary Ω, $c(\underline{x})$ nonconcave/nonconvex, and multiple local optima, heuristic algorithms exist which can yield acceptable results in some cases. The general strategy is to combine any appropriate constrained descent method (Newton's method, conjugate gradient, etc. [17,18]) with a scheme to repeatedly restart the search from enough initial points to provide some confidence that most of the significant local optima have been located. Figure 2.5a illustrates how the solution space may be partitioned with a uniform grid of starting

points for the descent algorithm. The best solution over all initial points is retained. If available, bounds on the first and second derivatives of the underlying cost surface can be used to guide grid spacing to correspond to the size of the major modes of the function. An often used variation on this basic approach is to randomly place the initial points with some specified average separation. Figure 2.5b illustrates the basic moves of a pattern search procedure which is often more efficient than the arbitrary restart approach [80]. When a local optimum is located (2), an "explore" move is executed to locate a distant point (3) outside of the current "valley," and a search along a line to this point finds the minimum cost location (4), from which a new descent search starts. The direction and length of the explore moves follow a pattern designed to provide thorough coverage of the cost surface [80]. Each of the above methods fails for $l_{1/q}$ minimization because of the extremely strong local minima encountered in any continuous search through the constraint space. A descent move from any starting point leads to the nearest basic solution since all basic solutions are local minima (see Appendix A, alternate proof). Using enough starting points to investigate each local minimum would be as costly as an exhaustive search.

In special cases where the steps of the iteration may be represented as a strongly ergodic, time-inhomogeneous Markov chain, the methods of stochastic relaxation (simulated annealing) may be used to guarantee asymptotic globally optimal results even with arbitrary $c(x)$ [29]. This approach is discussed in detail in section 3.2 and is the basis for the stochastic search algorithm presented there. Figure 2.5c illustrates how the algorithm produces a random move direction, favoring reduced cost, which can overcome a local optimum by allowing some "up-slope" moves.

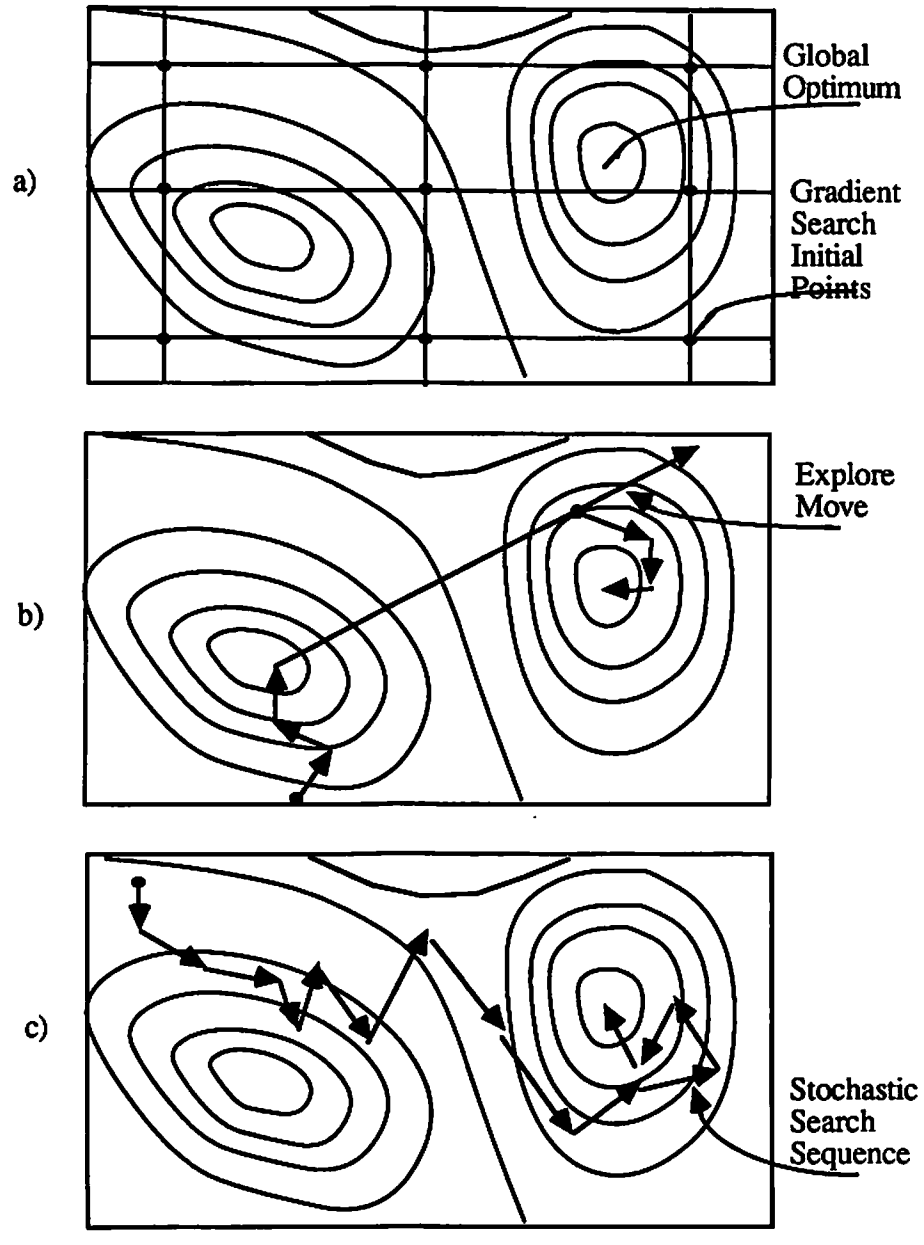


Figure 2.5. Comparison of Global Optimum Search Algorithms
a) Gradient descent restart method with initial solutions at each grid point for re-running the search. b) Pattern search method with explore move. c) Stochastic search descent path.

For $c(\underline{x})$ concave, and Ω a convex polytope given by linear inequality constraints, there have been some recent papers on methods for finding the global minimum much more efficiently than an exhaustive extreme point search [6,7,20,82]. When $l_{1/q}$ minimization is transformed to the equivalent positive orthant representation (see Appendix A), it is in the form addressed by these algorithms. Tuy first demonstrated such an algorithm in 1964, using a cutting plane method to successively subdivide a hypercone which initially contained the entire feasible set [83]. This algorithm was shown to be nonfinite and to suffer from cycling, but he [7] and others [6] have recently published modifications which overcome the problems and are quite efficient. Zwart [6] uses a method of successively enlarging a convex hull about extreme points until the exterior region contains no lower cost feasible points. All of these related methods require solutions of multiple linear programming subproblems at each step, work best when few local minima are present, and contain restrictions (non-degeneracy, inclusion of the origin in feasible set) which make the algorithms presented in Chapter 3 much more attractive for our problem.

Falk and Hoffman have presented a much improved concave minimization approach based on collapsing polytopes, which is more efficient and requires less computation per iteration [20]. Given a problem of the form

$$\min_{\underline{x}} \; c(\underline{x}) \quad s.t. \; S = \{ \underline{x} : A\underline{x} \le \underline{b} \}, \; \underline{x} \ge 0, \; A \in R^{M \times N}, \; M > N, \qquad (2.10)$$

$$c(\underline{x}) \quad \textit{concave}$$

we form the N+1 dimension system

$$D = \{ \underline{v} = (\underline{x}, y) : A\underline{x} + \underline{a}y \le \underline{b} \}, \; y \ge 0 \textit{ is scalar}, \qquad (2.11)$$

$$\underline{a} = (\|A_1\|, \|A_2\|, \cdots \|A_M\|)^T \quad \textit{where } A_m \textit{ is row } m \textit{ of } A$$

The feasible set S is thus a face of the polytope D, i.e. at $y = 0$. The algorithm forms successive convex hulls in D, which enclose S, by using selected vertices of D as the hull's extreme points. These hulls are reduced in size until the lowest cost extreme point is an element of S, which can occur only at a global optimum to (2.10). The algorithm steps are as follows:

1) *Initialize:* Find the initial solution $v^0 = (\underline{x}^0, y^0)$ in D by solving the linear program: minimize $y \in D$. Since this yields $y = \min_{1 \le i \le M} \left\{ \frac{b_i - A_i \underline{x}}{\|A_i\|} \right\}$, where the term in braces represents the Euclidean distance from $\underline{x}$ to the hyperplane $A_i \underline{x} = b_i$. The point $\underline{x}^0$ is the center of the largest hypersphere contained in S. Set $c_0 = \infty$, and let $\{(v^0, c_0)\}$ be the root node of the tree, T, used to tabulate the hull extreme points in D.

2) *Select a node of T_k to expand:* From the current tree, T_k, select the terminal node, v^t with minimum associated cost, c_t. If $v^t \in S$, i.e. if y=0, stop, the optimum is found.

3) *Step:* Expand the terminal node v^t by generating all its neighbors, $v^{t,i}$, by pivoting the tableau associated with D, such that $y^{t,i} < y^t$ and $v^{t,i} \notin T_k$. For each new $v^{t,i}$ compute the associated cost, $c_{t,i}$ at the point $w^{t,i}$ where the line connecting v^t and $v^{t,i}$ pierces the hyperplane S. The piercing point occurs where $y^{t,i} = 0$, and can be found using simple pivoting operations. Add each $\{(v^{t,i}, c_{t,i})\}$ to T_k and increment k. Go to 2).

This is the most promising existing algorithm studied, since iterations involve pivot operations only, rather than full linear programs, and demonstrated performance is significantly better than an exhaustive search. Though it promises global solutions, a number of problems made it impractical for the applications addressed in this dissertation. No

degeneracy is allowed and $M>N$ is required, both of which are violated by many of the examples shown in Chapters 4-6. The code implementation is also significantly more complex than the algorithms presented below. Future research could include adaptation of this algorithm to signal processing problems.

2.5. Probabilistic Interpretations of $l_{1/q}$ Optimization

Although the bulk of this dissertation treats the maximally sparse problem as deterministic, a probabilistic view gives some interesting insight which can provide a frame of reference for justifying the assumptions made and methods used in the rest of the dissertation. Since few real-world problems based on data measurement are truly deterministic, we are compelled to at least address the situation which includes random sources and noise. This section will provide a theoretical justification for broad use of the technique in the presence of stochastic signals. Optimal parameter estimation problems will be considered which are related to the deterministic constrained optimization covered in the remainder of the dissertation.

Much of the related work in the literature deals with blind deconvolution of seismic reflectivity sequences. In these problems, both the transmitted wavelet shape, and the reflectivity sequence corresponding to the reflections sites due to geological strata interfaces, are unknown. The deconvolution approach seeks the "simplest" possible representation for the wavelet and reflectivity sequence. Several authors have noted that the observed data is sparse, and non-Gaussian. A number of so-called "minimum entropy deconvolution" methods have been proposed which minimize a heuristic measure of the reflectivity sequence entropy, or sparseness. These include objective functions referred

to as the "varimax norm," the "parsimonious norm," and the "variable norm ratio" [8,9,10]. More recently the relationship of generalized p-Gaussian distributions and the l_p norm has been discussed, but the proposed algorithms require $p \geq 1$ for globally optimum solutions, and therefore cannot produce maximally sparse results.

2.5.1. Generalized p-Gaussian Probability Density Functions

We will investigate maximum likelihood and maximum a-posteriori (MAP) estimation in the presence of "generalized p-Gaussian" (gpG) distributed data. The gpG probability density function defines a family of distributions which can be used to characterize non-Gaussian sample data, and particularly sparse or spiky data. The gpG densities were introduced by Subbotin [22] in 1923, and used by Miller and Thomas [23] in 1972 for modelling non-Gaussian noise in detection theory. The density is defined for shape parameter p and variance σ^2 as:

$$f_x(x) = gpG(\mu,\sigma^2) = \frac{p}{2\Gamma(1/p)\,\gamma\,\sigma}\, e^{-\left(\frac{|x-\mu|}{\gamma\sigma}\right)^p} \qquad (2.12)$$

$$\gamma = \left[\frac{\Gamma(1/p)}{\Gamma(3/p)}\right]^{1/2} \qquad \Gamma() \textit{ is the gamma function}$$

For $p = 2$ and $p = 1$ this yields the familiar Gaussian and double exponential distributions respectively, and as $p \rightarrow \infty$ we have a uniform distribution. Figure 2.6 plots the bivariate density function for two iid gpG random variables. Of particular note is the similarity between these curves and the l_p norm unit balls shown in Figure 2.1, from which one

might infer a heuristic relationship between gpG distributions and l_p optimization. This apparent relationship will be verified below.

By adjusting the shape parameter, p, we may use this density function to produce a close match to the statistics of a remarkably wide range of sampled data distributions [8]. For the problem at hand, the primary range of interest is $0<p<1$, corresponding to $1<q<\infty$. Figure 2.7 shows a comparison between synthesized Gaussian and gpG time series data, where it is apparent that for small p, the sequence is much more 'spiky,' containing primarily small values with a few large outlying spikes. This sparse data is consistent with our view of how noisy samples from a maximally sparse source should appear. Indeed, several authors report excellent data modelling using low order gpG distributions [8,9,10,24]. In seismic reflectivity data produced from sonic impulse logging, the value of p ranges from .4 to 1.5; atmospheric processes are matched by $.1<p<.6$; and marine seismic and voiced speech data samples also require small values of p. The seismic reflectivity data is a particularly good example of a gpG model providing a close match to a problem known to be sparse in nature. The received data consists of the convolution of an explosive signal wavelet with a series of discrete, spatially separated, reflectivity spikes corresponding to the interfaces between geologic layers in the earth which cause reflections due to impedance mismatch. The nature of geologic strata suggests that the reflectivity series is sparse. With these examples as justification, we will proceed with the assumption that a gpG distribution with $p<1$ is a good model for sparse random data, and that it grows increasingly sparse as $p \rightarrow 0$.

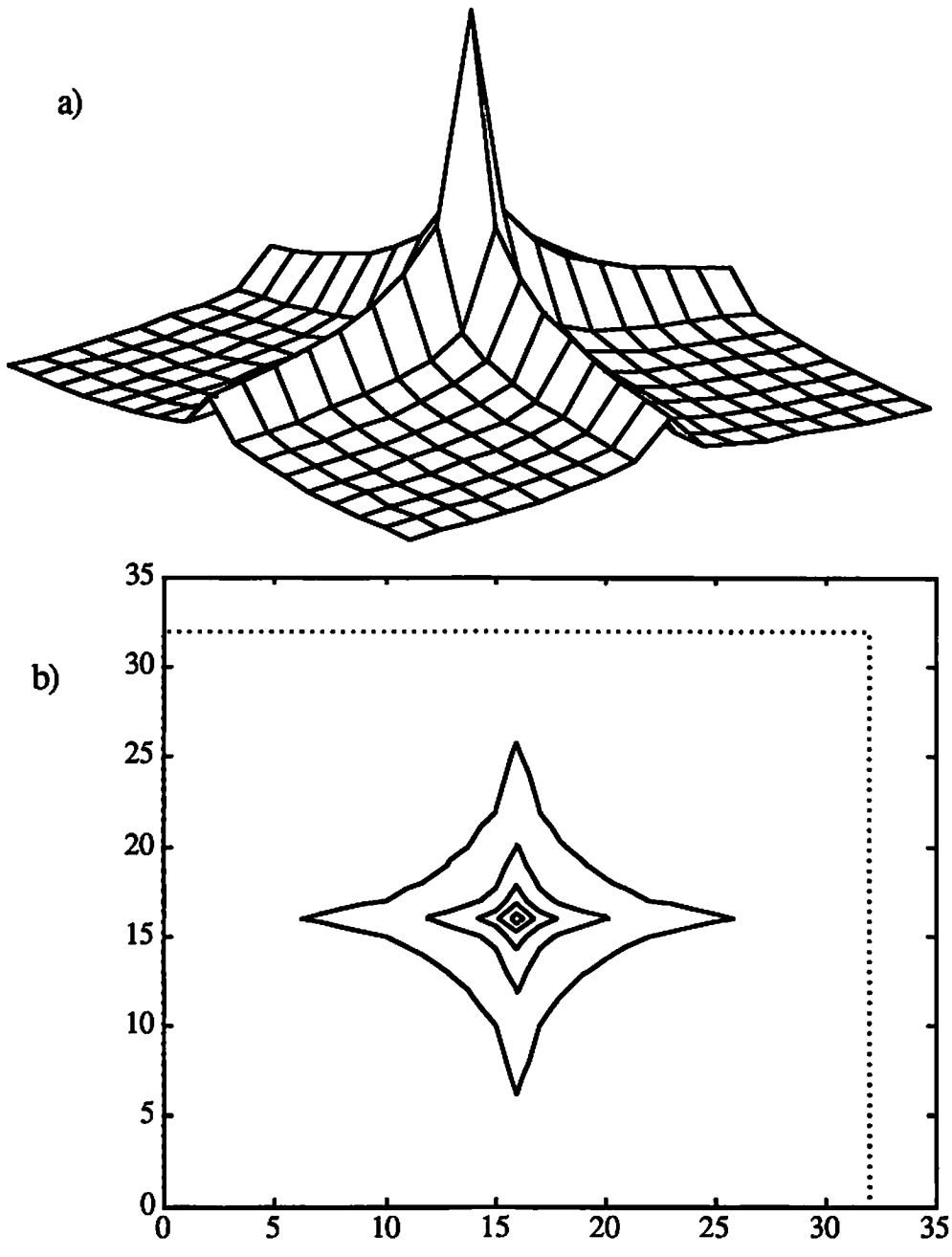
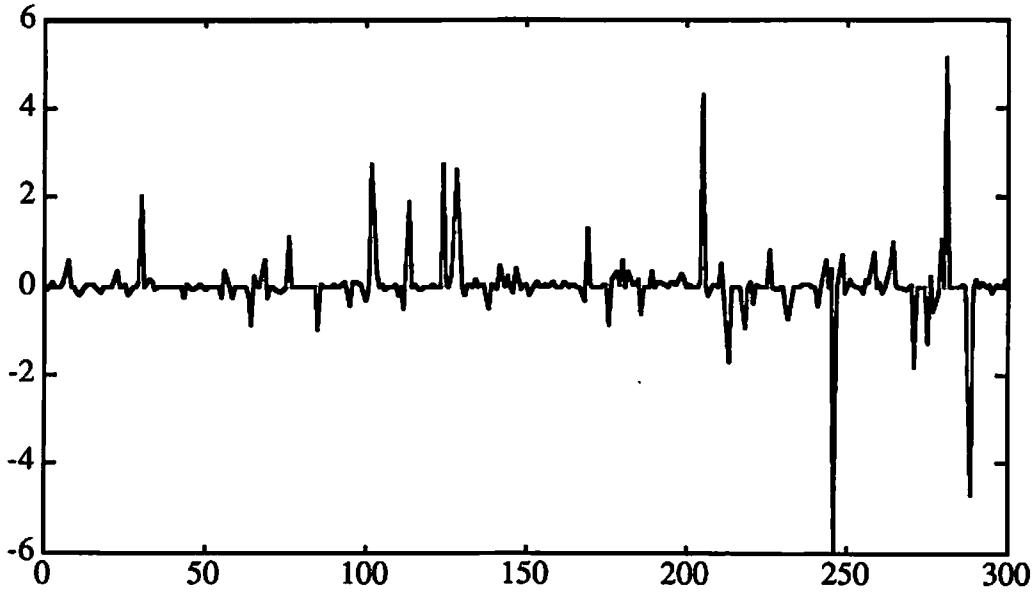
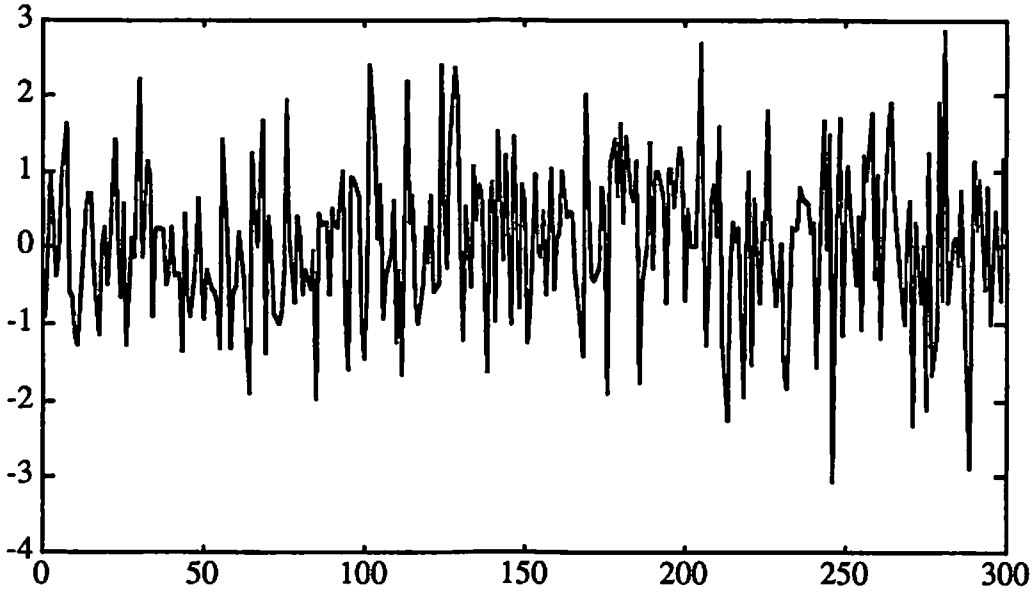


Figure 2.6. Generalized p-Gaussian Density Function Curves. a) Plot of the 2-D density function for independent gpG random variables with p=.4. b) Contour plot of the same function.



a)



b)

Figure 2.7. Comparison of Gaussian to Generalized p-Gaussian Data.

a) gpG data samples, p=.2, σ=1. b) Gaussian data, σ=1.

2.5.2. Maximum Likelihood Estimation with gpG Distributions

Maximum likelihood (ML) estimation of the mean and variance of gpG populations has been studied thoroughly by Pham and deFigureiredo in [24]. For an arbitrary value of p, μ_{ML} is difficult to formulate, requiring solution of complex nonlinear equations, but σ_{ML} is proportional the the l_p norm. This second fact leads to maximum likelihood solutions of unconstrained estimation problems by minimizing the l_p norm. Consider the system

$$\underline{y} = \mathbf{H}\underline{x} + \underline{n} \tag{2.13}$$

where $\underline{n}$ is zero-mean white additive generalized p-Gaussian white noise, $\underline{y}$ is the observed data, H is the system matrix, and $\underline{x}$, is the deterministic parameter to be estimated. This model has been used with gpG data in [8,24] for deconvolution by defining H to be the Toeplitz matrix representing the known convolutional sequence. The ML estimate of $\underline{x}$ is derived as follows:

$$l_x(\underline{y}|\underline{x}) \propto f_x(\underline{y}|\underline{x}) = \frac{p}{2\Gamma(1/p)\,\gamma\,\sigma}\, e^{-\sum_i \left(\frac{|y_i - \mathbf{H}x_i|}{\gamma\,\sigma}\right)^p} \tag{2.14}$$

$$= \min_{\underline{x}} \sum_i |y_i - H_i\underline{x}|^p = \min_{\underline{x}} \|\underline{y} - \mathbf{H}\underline{x}\|_{l_p} \Rightarrow \hat{\underline{x}}_{ML}$$

Thus, $\hat{\underline{x}}_{ML}$ for p-Gaussian data is found by minimizing the l_p norm of the error term. In general, for a given p this is solved using nonlinear programming techniques. For $p \geq 1$ a number of convex optimization methods may be used, including variants of linear programming. For $p < 1$ the problem is more difficult, and few global optimization techniques are found in the literature. For $p = 2$ we have the familiar closed form least squares solution: $\hat{\underline{x}}_{ML} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\underline{y}$.

2.5.3. MAP Estimation for Linearly Constrained gpG Problems

This study is primarily concerned with the linearly constrained problem of eqn (1.2). The ML approach above is unconstrained and assumes $\underline{y}$ is sparse for $p<1$, but does not impose a sparseness criterion on $\underline{x}$. We will attempt to formulate maximum a-posteriori (MAP) estimates of $\underline{x}$ which justify using the nonlinear program of eqn (1.2) in the presence of random signals. The results presented below are strictly dependent on the assumed model and the specific probability density functions cited, but do demonstrate a duality between deterministic sparse optimization and optimal parameter estimation for distributions which are likely to produce sparse $\underline{x}$.

The following simple signal model for noise corrupted measurements shall be adopted:

$$\underline{y} = \mathbf{H}\underline{x} + \underline{n} \tag{2.15}$$

$f_x(\underline{x}) = gpG(0,\sigma_x^2)$, *iid, white zero mean generalized p Gaussian,* $p<1$

$f_n(\underline{n}) =$ *as specified in the following cases:*

Case 1: Uniform iid noise,

$$f_n(\underline{n}) = U(-\varepsilon,\varepsilon) = \begin{cases} \frac{1}{2\varepsilon} & |\underline{n}| \le \varepsilon \\ 0 & otherwise \end{cases}$$

which leads to the conditional density:

$$f_y(\underline{y}\,|\,\underline{x}) = f_n(\underline{y}-\mathbf{H}\underline{x}) = \begin{cases} \frac{1}{2\varepsilon} & |\underline{y}-\mathbf{H}\underline{x}| \le \varepsilon \\ 0 & otherwise \end{cases} \tag{2.16}$$

The MAP estimate of $\underline{x}$ is then:

$$\hat{\underline{x}}_{MAP} = \max_{\underline{x}} \; f_x(\underline{x}\,|\,\underline{y}) = \max_{\underline{x}} \; f_y(\underline{y}\,|\,\underline{x})f_x(\underline{x}) \tag{2.17}$$

$$= \max_{\underline{x}} \begin{cases} \dfrac{p}{2\Gamma(1/p)\,\gamma\,\sigma}\; e^{-\sum_i \left(\frac{|x_i|}{\gamma\,\sigma}\right)^p} & \text{if } |\underline{y}-\mathbf{H}\underline{x}| \le \varepsilon \\ 0 & \text{otherwise} \end{cases}$$

$$= \max_{\underline{x}} \; e^{-\sum_i |x_i|^p} \quad \text{such that} \quad |\underline{y}-\mathbf{H}\underline{x}| \le \varepsilon$$

$$\hat{\underline{x}}_{MAP} = \min_{\underline{x}} \; \sum_i |\underline{x}|^p \quad \text{such that} \quad |\underline{y}-\mathbf{H}\underline{x}| \le \varepsilon$$

which corresponds exactly with the $l_{1/q}$ optimization problem of eqn (1.2). Thus, if the generalized p Gaussian distribution (p<1) is an acceptable model for $\underline{x}$, and the noise is uniform, we may use the algorithms presented below in sections 3.1 and 3.2 for optimal results. It may be argued however that uniformly distributed noise is atypical. We therefore consider the more realistic case of Gaussian noise.

Case 2: Gaussian noise,

$$f_n(\underline{n}) = N(0,\sigma_n^2), \; \textit{iid, white zero mean Gaussian} \tag{2.18}$$

$$f_y(\underline{y}\,|\,\underline{x}) = \frac{1}{(\sqrt{2\pi}\,\sigma_n)^M}\; e^{-\frac{(\underline{y}\,-\,\mathbf{H}\,\underline{x})^T(\underline{y}\,-\,\mathbf{H}\,\underline{x})}{2\sigma_n^2}}$$

which leads to the following MAP estimate of $\underline{x}$

$$\hat{\underline{x}}_{MAP} = \max_{\underline{x}} \; f_x(\underline{x}\,|\,\underline{y}) = \max_{\underline{x}} \; f_y(\underline{y}\,|\,\underline{x})f_x(\underline{x})$$

$$= \max_{\underline{x}} \; e^{-\frac{(\underline{y}\,-\,\mathbf{H}\,\underline{x})^T(\underline{y}\,-\,\mathbf{H}\,\underline{x})}{2\sigma_n^2}}\; e^{-\sum_i \left(\frac{|x_i|}{\gamma\sigma_x}\right)^p}$$

$$= \min_{\underline{x}} \frac{(\underline{y} - \mathbf{H}\underline{x})^T(\underline{y} - \mathbf{H}\underline{x})}{2\sigma_n^2} + \left(\frac{1}{\gamma\sigma_x}\right)^p \sum_i |x_i|^p \qquad (2.19)$$

Equation (2.19) is difficult at best to solve directly, but this form is well suited for application to the convex transformation gradient search algorithm presented in section 3.3. Rather than attempt an optimal solution, we use the known (or estimated) noise statistics to specify a confidence region about our solution. An upper bound, consistent with our uncertainty due to noise, is set on the first term of (2.19) and we then adjust $\underline{x}$ to minimize the second term. The resulting constrained optimization problem is (see eqn 3.23):

$$\min_{\underline{x}} \sum_i |x_i|^p \quad s.t. \quad (\mathbf{H}\underline{x} - \underline{y})^T (\mathbf{H}\underline{x} - \underline{y}) \le \varepsilon, \; x_i \ge 0, \; p<1 \qquad (2.20)$$

With σ_n^2 known, we may specify ε to give a fixed probability, α, that the true $\mathbf{H}\underline{x}$ lies within a distance ε from $\underline{y}$, i.e. select ε such that

$$P[\, \|\mathbf{H}\underline{x} - \underline{y}\|^2 < \varepsilon \,|\, \underline{y} \,] = \alpha \qquad (2.21)$$

$\hat{\underline{x}}$ is then the most sparse vector that maps into this confidence region given by the hypersphere of radius ε.

In deconvolution work reported in the literature, one major aspect of the problem was determining the value of p which produced the best gpG model fit to the observed data. An interesting fact for our problem is that since we are assuming a maximally sparse sequence, the value of p is not critical. Although the exact value of p may not be known for the observed gpG data, it has been shown above (section 2.3) that the minimum l_p solution remains unchanged over some range of p values less than 1. Also, Theorem 2 proves that we may use any p less than some finite p_1 and achieve the same solution,

therefore, the goodness of fit tests used in [8] for evaluating p are typically not needed in maximally sparse problems.

CHAPTER 3: MAXIMALLY SPARSE OPTIMIZATION ALGORITHMS

3.1. The $l_{1/q}$ Simplex Search Algorithm

3.1.1. Similarity to Linear Programming

With the justification provided by the fundamental theorem of $l_{1/q}$ programming, we can immediately recognize the similarity to linear programming. The basic solutions which must be searched for an optimum are identically those found in the corresponding linear program [18]. As in linear programming, the algorithm described below searches through the basic solutions contained in the simplex by pivoting between adjacent solutions while monotonically reducing the objective function. A "tableau" structure is also adopted, as in linear programming methods, to take care of the numerical bookkeeping during the search. The major difference lies in the objective function used, which in our case is nonlinear. This requires an entirely different method of selecting which adjacent solution to move to at each iteration. The proof of global optimality for linear programming also does not hold for the $l_{1/q}$ program, and we must be satisfied with locally optimal solutions, though experience with the algorithm indicates results are usually very nearly global solutions.

3.1.2. Formulation, Basic Solutions, the Tableau

As in linear programming, we begin by computing any basic feasible solution as a starting point for the simplex search. The two forms presented above describe the constraint

as a vector-matrix linear equality in R^{MxN} or $R^{M'xN'}$space respectively. In the following development we will use $\mathbf{H}\underline{x} = \underline{b}$ of form (2.2a), but $\tilde{\mathbf{H}}\,\tilde{\underline{x}} = \tilde{\underline{b}}$ from (2.2b) can be substituted. Selecting any M columns of **H** for which we can compute a left pseudoinverse, we permute the matrix columns and corresponding elements of $\underline{x}$ to place these as the first M columns. Partitioning **H** we have

$$[\,\mathbf{A}\;|\mathbf{D}\,] = \mathbf{H}\,, \qquad (3.1)$$
$$\textit{where}\;\; \mathbf{A} \in R^{M'xM'},\; \mathbf{A}^{\dagger} = \textit{left pseudoinverse of}\; \mathbf{A}$$

Multiplying by $\mathbf{A}^{\dagger}$ leads directly to the basic solution $\underline{x}_B$

$$[\,\mathbf{I}\,|\,\mathbf{A}^{\dagger}\mathbf{D}\,]\underline{x} = \mathbf{A}^{\dagger}\underline{b}\,,\; \underline{x}_B = [\,\mathbf{A}^{\dagger}\underline{b}\,,\,0,\,\cdots 0]^{\mathrm{T}} \qquad (3.2)$$

For form (2.2b), somewhat more care than indicated above is required in selecting columns for **A** to insure $\underline{x}_B \geq \underline{0}$. A linear programming phase one procedure may be used with either form to compute a non-negative initial basic solution if we first negate the elements of $\underline{b}$, and corresponding rows of **H** required to force $\underline{b} \geq \underline{0}$. We identify the variables x_i associated with columns of **A** as basic variables, and **A** as the basis. An adjacent basic solution is one that is formed by moving (pivoting) one variable out of the basis and one non-basic variable into the basis by swapping a column in **A** with one in **D**, adjusting $\underline{x}$ indices, and recomputing $\mathbf{A}^{\dagger}$.

Equation (3.2) suggests the structure of the "tableau" used in linear programming to facilitate the pivoting computations [18]. The tableau is formed by augmenting the matrix with the right hand side

$$\mathbf{Y} = [\, \mathbf{I} \,|\, \mathbf{A}^{\dagger}\mathbf{D} \,|\, \mathbf{A}^{\dagger}\underline{b} \,] \tag{3.3}$$

The reduced cost row of the l_p tableau does not appear in **Y** due to the nonlinear cost functional. Since the first M columns are always the identity matrix, they need not be stored. During pivoting, all remaining columns of **Y** are updated using the simple pivoting equations, or with a recursive product form of computing the new inverse $\mathbf{A}^{\dagger}$, both described in many texts [18,19]. Index vectors are maintained to keep track of which variables are in or out of the basis. At any iteration, $\mathbf{Y}^i$, the current basic solution can be read directly from the last column. Two tableaus are said to be adjacent if we can move from one to the other with a single pivot operation, or equivalently, if their sets of basic variables differ by one variable only.

For the $l_{1/q}$ simplex algorithm, we may view the set of tableaus as a connected graph with a node for each tableau satisfying (3.3).

Let **S** = the set of all basic feasible solutions (BFS) $\underline{x}$ to (1.2).

Let **T** = the set of all tableaus, $\mathbf{Y}^i$ associated with **S**.

Define a graph, G = (**T** , υ) where υ consists of pairs $(\mathbf{Y}^i , \mathbf{Y}^j) \in \upsilon$ iff tableau $\mathbf{Y}^j$ can be generated from $\mathbf{Y}^i$ by a single pivot operation (or vice versa).

Each element of **T** maps onto an element of **S** (possibly many-to-one due to degeneracy). This graph is connected and has properties discussed in depth in [25]. The fundamental theorem implies we may find an optimum by searching only the graph G and ignoring non-basic solutions. The algorithm performs this search by generating a sequence, $\{ \mathbf{Y}^i \,|i=1,2, \cdots i_{max}\}$ which traverses the graph along a path of monotonically non-increasing cost, halting when all adjacent solutions are of greater cost.

It is noteworthy that the ability to take this approach is entirely dependent on the particular cost function chosen. Although $g(Q)$ is neither linear nor convex, and has numerous local minima which would make most gradient based optimization techniques ineffective, the fundamental theorem proven above implies that it is particularly suited to a simplex search approach.

3.1.3. Detailed Algorithm Description

The procedure used in the $l_{1/q}$ simplex search for traversing the graph G is described below. At each iteration of the algorithm we chose an element of $\underline{x}$ as an "entering variable" to enter the basis and become nonzero and a "leaving variable" to be forced to zero. This process moves the trial solution to an adjacent basic solution and is accomplished by pivoting the **H** matrix and measurement vector $\underline{b}$. The selection of an entering variable is made such that at each iteration the cost is reduced. The leaving variable is chosen to be the one that first reaches zero value as the solution traverses along the edge of the solution space in the direction of the entering variable.

The steps of the algorithm are as follows:

1) **Find a starting basic solution.** The starting tableau, $\mathbf{Y}^0$, and initial basic solution for the $l_{1/q}$ search are formed by using the L.P. simplex method to solve the augmented phase one linear program:

$$\min_{\underline{x}} \underline{c}^T \underline{x}' \quad s.t. \quad [\,\mathbf{H}\,|\,\mathbf{I}\,]\,\underline{x}' = \underline{b}\,, \quad \underline{x}' = \begin{bmatrix} \underline{x} \\ \underline{a} \end{bmatrix}, \quad \underline{c} = \begin{bmatrix} \underline{0} \\ \underline{1} \end{bmatrix} \tag{3.4}$$

where $\underline{a}$ is the vector of M artificial variables used only in phase one, and $\underline{c}$ is the phase one cost. An immediately obvious basic solution to (3.4) is $\underline{x}' = \begin{bmatrix} \underline{x} \\ \underline{a} \end{bmatrix} = \begin{bmatrix} \underline{0} \\ \underline{b} \end{bmatrix}$, which places $\underline{a}$ in the basis. This is used as the initial basic solution for the linear programming simplex algorithm. Since all a_i have a cost of 1, and all x_i have a cost of zero, the solution to eqn (3.4) will drive all a_i out of the basis, replacing each with some x_i. The initial tableau, $\mathbf{Y}^0$, and basic solution for our $l_{1/q}$ simplex search are given by the final pivoted tableau of the phase one linear program. All columns corresponding to $\underline{a}$ are dropped, yielding an $M \times (N+1)$ tableau, $\mathbf{Y}^0$ in the standard form:

$$\mathbf{Y}^k = \begin{bmatrix} y^k_{1,1} & \cdot & y^k_{1,N} & y^k_{1,0} \\ \cdot & & \cdot & \cdot \\ y^k_{m,1} & \cdot & y^k_{m,N} & y^k_{m,0} \end{bmatrix} \tag{3.5}$$

with k being the iteration number, and the last column, $y^k_{1,0}$ through $y^k_{m,0}$ giving the current basic solution, corresponding to the reduced measurement vector $\underline{b}$. The columns of $\mathbf{Y}^k$ are reordered at each iteration so that the first m columns correspond to the variables (elements of $\underline{x}$) which are in the basis. $\mathbf{Y}^k$ is thus row reduced for columns 1 through m which contain a permuted identity matrix. The initial basic solution is:

$$x_{j_i} = \begin{cases} y^0_{i,0} & for \quad 1 \le i \le m \\ 0 & for \quad m+1 \le i \le N \end{cases} \tag{3.6}$$

where j_i is a permuting index which maps the original elements of $\underline{x}$ to the present corresponding column of Y^k. The solution at any step k is likewise read from this last column.

2) **Compute the present cost.** The $l_{1/q}$ cost of the present basic solution is computed as

$$C = \sum_{i=1}^{m} (y_{i,0}^{0})^{1/q}, \quad q >> 1 \tag{3.7}$$

3) **Select entering and leaving variables.**

Repeat for $s = m + 1$ to N until $C_s < C$:

$$\delta = \min_{1 \le i \le m} \left\{ \frac{y_{i,0}^{k}}{y_{i,s}^{k}} \quad if \quad y_{i,s}^{k} > 0 \right\} \tag{3.8}$$

$r_s = i \quad$ *at this minimum*

$$C_s = \sum_{j \neq r_s}^{m} (y_{j,0}^{k} - y_{j,s}^{k}\delta)^{1/p}$$

The entering variable is identified by the first s value for which $C_s < C$ and the leaving variable is the corresponding r_s.

4) **Test for algorithm termination.** If in step 3, δ could not be computed due to $y_{i,s}^{k} \le 0$ for all i and s then stop, the solution is unbounded. If after step 3, $C_s \ge C$ then stop, the present solution is optimum, otherwise let $C = C_s$ and continue.

5) **Pivot the tableau.** Form $\mathbf{Y}^{k+1}$ by pivoting $\mathbf{Y}^{k}$. For the s found in step 3) and $r = r_s$, bring the entering variable into the basis by pivoting on the $y^{k}_{r,s}$ element.

For $1 \le i \le m \quad 0 \le j \le N$:

$$y^{k+1}_{i,j} = y^{k}_{i,j} - \frac{y^{k}_{i,s}}{y^{k}_{r,s}} y^{k}_{r,j} \qquad i \neq r \tag{3.9}$$

$$y^{k+1}_{r,j} = \frac{y^{k}_{r,j}}{y^{k}_{r,s}}$$

After pivoting, reorder the columns of $\mathbf{Y}^{k+1}$ so that the first M correspond to basis variables.

6) **Repeat** steps 3) through 5).

This algorithm's approach is similar to one described in [4], although we have not seen it applied to the minimum order problem. As with linear programming, a practical computer algorithm must contain enhancements to deal with accumulated error from pivoting, find an initial solution, include an anti-cycling procedure, and handle systems of large order.

A major difference between this and the linear programming simplex algorithm is the method in which the entering variable is selected. For linear programming the reduced cost is computed efficiently for each possible entering variable by using a cost row which is pivoted with the rest of the tableau. Any row with a negative reduced cost entry can be used as an entering variable. For the $l_{1/q}$ simplex search algorithm, it is necessary at each iteration to actually project the solution to each possible adjacent basic solution and com-

pute the $l_{1/q}$ cost. The first adjacent solution with reduced cost is chosen to enter. This is less efficient, but the maximum possible additional computation is equivalent to one extra pivot of the tableau per iteration. The algorithm also requires that each $y_{i,j}$ value be available for computing adjacent costs, thus making the improved performance of product form inverses as used in the revised simplex method inappropriate since they do not explicitly store **Y** [18]. The advantages of computing $\mathbf{A}^{-1}$ as a running product rather than explicitly pivoting the tableau would be lost in needing to compute **Y** at each iteration. Also, since the cost for a feasible solution remains high everywhere except near basic or degenerate solutions, it is not likely that an approach like Karmarkar's method for linear programming would be usable. These new approaches to linear programming perform searches in the interior of the feasible solution space, where the $l_{1/q}$ cost, unlike a linear cost, can give misleading gradient information on a direction to move toward a solution [26].

3.1.4. Adjacency Graph Representation and Degeneracy Issues

A feasible solution to eqn (1.2) with fewer than M nonzero components is termed a degenerate solution [18,25]. It follows then, that our original problem, (1.1), involves a search for the maximally degenerate solution, and for $q>q_1$ we seek the same solution for eqn (1.2). In conventional linear programming, degeneracy is handled as a nuisance and so-called "anti-cycling" procedures, mentioned above, are employed to avoid related problems. Here, degeneracy is one means of attaining the sparse solutions we seek, particularly for the equality constraint problems without slack variables.

The problems associated with degeneracy arise from the fact that zero valued basic variables may be interchanged with non-basic variables resulting in a different tableau, but identical corresponding solutions in $\underline{x}$. In the non-degenerate case, there is a one-to-one correspondence between each tableau and its corresponding basic solution. Degenerate solutions have a many-to-one relationship. A degenerate solution is overdetermined, with more than one constraint equation active at once [25]. It is important to consider the effect on algorithm performance of having a finite set of tableaus associated with a single basic solution.

Consider a degenerate BFS, $\underline{x}_j \in S$, and its representation in the graph G described above. For each such solution we define a degeneracy subgraph, DG^j containing all nodes which map onto $\underline{x}_j$ and their interconnecting arcs. Equation (3.10) defines a system with 5 variables and 3 constraints which has degenerate solutions is shown in Figure 3.1 and Table 3.1.

$$\mathbf{H}\underline{x} = \underline{b}, \quad \mathbf{H} = \begin{bmatrix} 1 & 0 & 0 & 1/2 & 1/6 \\ 0 & 1 & 0 & 1/2 & 1/6 \\ 0 & 0 & 1 & -1/2 & 1/6 \end{bmatrix}, \quad \underline{b} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \tag{3.10}$$

Basic Feasible Solutions:							
	x_1:	x_2:	x_3:	x_4:	x_5:	Order of Degeneracy	Tableaus in Degen. Subgraph
$\underline{x}_1$	1	1	1	0	0	0	(1)
$\underline{x}_2$	2	2	0	-2	0	0	(2)
$\underline{x}_3$	0	0	0	0	6	2	(3),(5),(6),(7),(9)
$\underline{x}_4$	0	0	2	2	0	1	(4),(8)

Table 3.1. Comparison of the Degeneracy of the Feasible Basic Solutions to eqn (3.10).

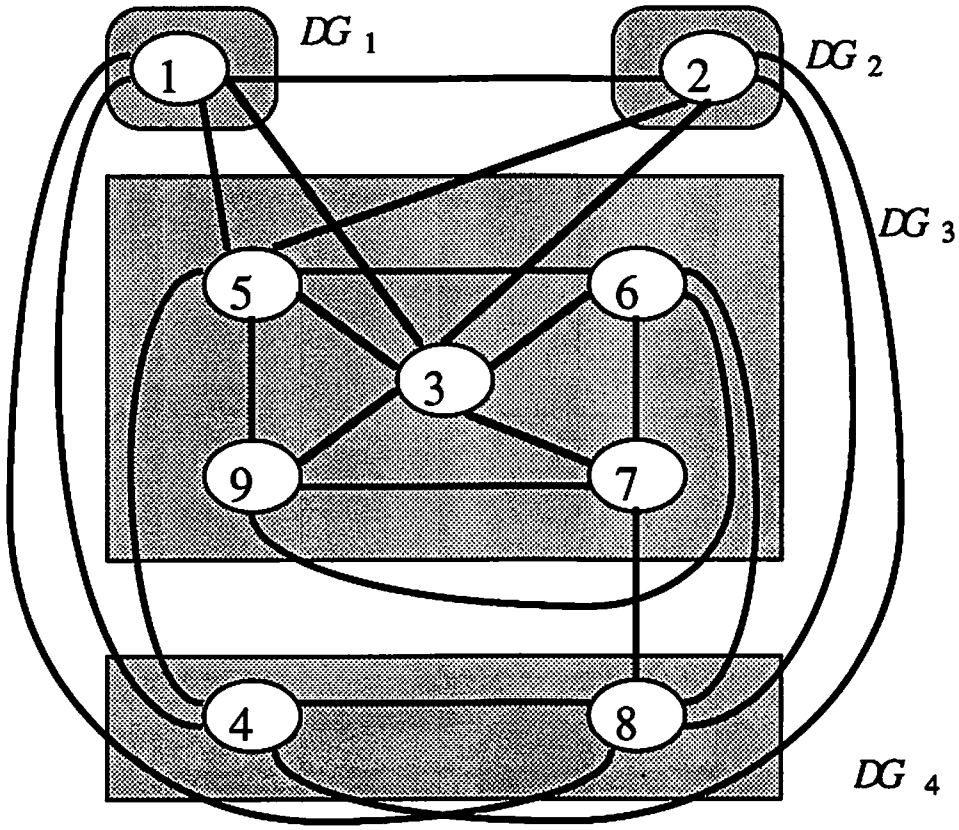


Figure 3.1. A Graph Tableaus for eqn (3.10) Showing Degeneracy. Nodes of this graph correspond to the tableaus for eqn (3..8.1) and the solutions in Table 3.1. Each shaded area, DG_i, represents the degeneracy subgraph associated with BFS x_i.

The set of tableaus and basic feasible solutions are shown in Table 3.1 and the associated graph, including the degeneracy subgraphs associated with the degenerate solutions, are shown in Figure 3.1. Note that the tableaus in DG_3 are not all mutually adjacent. This Figure also illustrates the increased number of paths to a degenerate solution. The number of nodes in DG^j increases extremely rapidly as the order of degeneracy increases, and it is likely that many, if not most, of these nodes are not mutually adjacent [25]. This

poses a problem for the $l_{1/q}$ simplex algorithm if such a subgraph is encountered prior to reaching the optimal solution. All nodes in DG^j are of the same cost, yet we must traverse the subgraph to insure access to any lower cost nodes in G which are adjacent to distant nodes of DG^j. One cannot use a single pivot operation to reach all BFS's connected to the degeneracy subgraph associated with $\underline{x}_j$. Procedures must be employed to insure that we do not terminate prematurely in DG^j without accessing all external nodes connected to the subgraph, of potentially lower cost, and that we do not "cycle" endlessly in DG^j. A number of approaches to this problem have been used, including a very simple anti-cycling procedure due to Bland, perturbation methods [18,27], and algorithms to find the minimum spanning tree for DG^j to guarantee no cycling and access to each external connected point [25]. The stochastic search algorithm described in section 3.3 randomizes the selection of adjacent nodes to pivot to, and thus eliminates the problem of cycling or missing lower cost adjacencies at a degenerate solution. Degeneracy of the optimal solution however provides advantages which contribute to the success of the algorithm. As the $l_{1/q}$ simplex algorithm traverses the graph from an arbitrary starting point, it can find the global optimum only if a path of monotonic decreasing cost exists through adjacencies. With the many to one mapping of the nodes of DG^j onto $\underline{x}_j$, there are many more paths from an arbitrary point in G to a node mapped to a degenerate solution $\underline{x}_j$ than would be found for a non-degenerate solution. Thus the optimum is adjacent to more nodes and it is less likely to be an isolated minimum (see the example of Figure 3.1).

The algorithm's operational modes differ for forms (2.2a) and (2.2b). With (2.2a), the degeneracy properties described above dominate algorithm performance, but with eqn (2.2b), cost reduction is obtained primarily by pivoting zero cost slack variables into the basis. Few examples of degenerate solution results have been observed for form (2.2b).

3.1.5. Algorithm Performance

The author's experiments have shown that the $l_{1/q}$ simplex algorithm converges in approximately the same number of iterations as the linear algorithm would for a similar sized system. Since the cost at adjacent tableaus can be computed without pivoting the entire tableau, the processing load involved in computing costs at all adjacencies in step 2) above is equivalent to a single pivot operation. Thus computation time is approximately twice that of the linear programming simplex method overall. With a convergence time comparable to the $\mathbf{O}(5N)$ iterations of the simplex method, this algorithm is dramatically more efficient than an exhaustive search.

Performance of the $l_{1/q}$ simplex search algorithm has been evaluated using both simple low order arbitrary linear systems and simulations of the reconstruction problems presented in chapters 4, 5, and 6. Initial tests used simple synthesized system matrices, **H**, constructed with known rank and condition number. In all tested configurations of **H**, including orthogonal, ill conditioned nonsingular, and rank deficient matrices, the $l_{1/q}$ search outperformed all other algorithms in correctly reconstructing sources involving a small number of nonzero terms.

In general, the algorithm does not guarantee convergence to a global optimum. However, the author has found that surprisingly, in virtually every case, acceptably sparse solutions were found. The reasons for this observed performance is not fully understood, but the increased adjacency due to the degeneracy of optimal solutions provides more paths through G which lead to the optimum. It was shown by exhaustive evaluation [81] that due to degeneracy alone, the algorithm is guaranteed optimal results for an $M \times N$ system in the following special cases:

1) $M<N$ and $N = 2, 3, 4,$ or 5.

2) $M = N - 1$

3) Any system with an order 1 solution (i.e. only one x_i nonzero).

For $N>5$, $M<N-1$, or solution order >1, the analysis gets difficult, but the trend may continue. For example, any basic solution in an $N=6$ system is never more than two pivots from a global optimum.

Another possible explanation, requiring further analysis, for the algorithm's favorable performance is the cost gradient structure of the graph G. By eliminating all but the basic solutions from our search, it appears that the extreme cost gradients and ridges are eliminated. The cost structure of G may be inherently better behaved than the entire continuous solution set, with many monotonically decreasing paths to low order nodes. Perhaps some algebraic analysis of H and $\underline{b}$ could provide a bound on deviation from the optimum for the algorithm.

The $l_{1/q}$ search algorithm is much more efficient than the globally optimum stochastic search algorithm described in section 3.3, so if a good, near optimal solution is acceptable, this algorithm is recommended.

3.2. The Stochastic Search Algorithm

The algorithm described in this section uses the techniques of simulated annealing [28,29] to arrive at globally optimal solutions to problem (1.2). The algorithm still limits its search to the graph of basic solutions and traverses the graph across adjacent node arcs. A Markov chain is generated using the Metropolis algorithm [28,29] to randomly select from the adjacent solutions.

3.2.1. Overview of the Stochastic Relaxation Technique

Stochastic relaxation, or simulated annealing as it is often called, has been used by a number of authors in recent years for a wide range of combinatorial optimization problems of high dimension [21,28,29,30]. Kirkpatrick, et al. first used simulated annealing in solving function partitioning, wire routing, and other computer design problems, and demonstrated favorable algorithm performance with the classical traveling salesman problem. Geman and Geman applied the technique to image restoration by use of a Markov random field model. The power of the technique lies in its ability to overcome the "frustration" component of an optimization problem and thus avoid being trapped in a local minimum. By randomizing the direction of each iterative step, these algorithms can climb "hills" of increasing cost while on average reducing cost, so that in the limit they find a globally optimal minimum.

The term "simulated annealing" is coined by analogy to the statistical mechanics process of gradually cooling a material (e.g. metallic alloy etc.) until the individual atoms, or domains, reach a joint state of globally minimum energy, thus forming a crystalline structure. Too rapid cooling may produce a higher energy glass-like metastable form by being trapped in an atomic configuration of local minimum energy. The probability of reaching a given configuration, r_i, is given by Boltzmann's probability factor, $exp\left\{\frac{-E\{r_i\}}{k_B T}\right\}$, where $E\{r_i\}$ is the energy of the configuration, k_B is Boltzmann's constant, and T is the temperature.

Kirkpatrick, et al. observed the similarity between annealing and random search algorithms for combinatorial problems. They replaced the energy term with the cost function

to be minimized and included a normalizing factor, $z(T)$, to yield the Gibb's probability measure $\frac{1}{z(T)} exp\left\{\frac{-c(\underline{\chi}_i)}{T}\right\}$. The Metropolis algorithm was then used to simulate the random state changes in the atomic system by randomizing the selection of transitions in the search algorithm in accordance with the Gibb's probability. This produced a time-inhomogeneous Markov chain of state changes with the probabilities of the succeeding state depending only on the present state and on T. Although the temperature has no physical significance for the general combinatorial problem, it is retained as the parameter which controls the rate of convergence to the solution. For high "temperatures," T, the distribution is nearly uniform, with all states equally likely, but as T is decreased, the modes of the distribution corresponding to the cost function surface become more prominent, and the instantaneous state tends to be of lower cost. Geman and Geman [28] proved that if a satisfactory "annealing schedule" is used for reducing T at each step, this heuristically satisfying technique does in fact converge in the limit to the global minimum with probability one, and Mitra, et al. [29] refined and generalized this proof, and analyzed the finite time performance of the algorithm.

3.2.2. Markov Chain Representation of the Simplex Graph

In order to apply the stochastic relaxation technique, we must prove that we may define a Gibb's measure, and use the Metropolis algorithm, to randomize traversal of the simplex graph, G, and produce a time-homogeneous Markov chain for fixed T. This done, we may use the results of [29] to infer that for the given annealing schedule, the time-inhomogeneous chain is strongly ergodic, and thus converges to the global optimum.

Using the graph notation described above, let $N(\mathbf{Y}^i) \subset \mathbf{T}$ denote the set of neighbors in G to the node $\mathbf{Y}^i$. $\mathbf{Y}^j \in N(\mathbf{Y}^i)$ iff $\mathbf{Y}^j$ is an adjacent node (reached by a single pivot opera-

tion) of $\mathbf{Y}^i$. Let $\mathbf{Y}^i$ be mapped onto the BFS $\underline{x}_j \in \mathbf{S}$, along with all other nodes contained in the degeneracy subgraph DG_j associated with $\underline{x}_j$.

The simplex algorithm described in section 3.1 generates the pair of sequences

$$\{ \mathbf{Y}^n \mid n = 1,2,\cdots n_{\max} \} \rightarrow \{ \underline{x}^n \mid n = 1,2,\cdots n_{\max} \} \tag{3.11}$$

according to the rule $\mathbf{Y}^{(n+1)} \in N(\mathbf{Y}^N)$ such that $g(\underline{x}^{N+1}) \leq g(\underline{x}^N)$, where $\{\underline{x}^N\}$ may have repeated elements due to degeneracy. Thus the cost is non-increasing at each iteration and the sequence $\mathbf{Y}^n$ is deterministic. The sequence terminates at the first element $\underline{x}_j$ such that all tableaus adjacent to the degeneracy subgraph DG_j of $\underline{x}_j$ have corresponding solutions of higher cost.

Since each element of the sequence $\{\mathbf{Y}^n\}$ belongs to the neighborhood of the previous element, as defined by the graph G, it follows that randomizing the choice of $\mathbf{Y}^{n+1}$ from the neighbors of $\mathbf{Y}^n$ will produce a Markov chain, since for any randomizing rule

$$P(\mathbf{Y}^{n+1} \mid \mathbf{Y}^n, \mathbf{Y}^{n-1}, \ldots \mathbf{Y}^0) = P(\mathbf{Y}^{n+1} \mid \mathbf{Y}^n) \tag{3.12}$$

The key to use of the simulated annealing algorithm is to choose a Gibb's measure, $P_T(\mathbf{Y})$ and randomize the updating rule (choice of tableau) such that for a fixed temperature parameter, T, the resulting quasi-stationary Markov chain is homogeneous [29]. The temperature parameter T is introduced to simulate the annealing process so that as $T \rightarrow 0$, the measure $P_T(\mathbf{Y})$ becomes concentrated at the tableaus corresponding to the global minima of $g(\underline{x})$ for all $\underline{x} \in \mathbf{S}$. The Gibb's measure is chosen to reflect the cost function of our problem:

$$P_T(\mathbf{Y}) = \frac{1}{z(T)} \exp\left[\frac{-g(\mathbf{Y})}{T}\right], \quad \mathbf{Y} \in \mathbf{T} \tag{3.13}$$

where $z(T)$ acts like the partition function of statistical mechanics to normalize the P_T such that $\sum_{\mathbf{Y}\in T} P_T(\mathbf{Y})=1$ and $g(\mathbf{Y})$ denotes the cost $g(\underline{x})$ of eqn (1.2) where $\underline{x}$ is the BFS associated with Y. For a time homogeneous Markov chain, the Gibb's measure must obey [29]

$$P_T(\mathbf{Y}^i) = \sum_{\mathbf{Y}^j\in \mathbf{T}} \mathbf{P}_T(i,j) P_T(\mathbf{Y}^j) \text{ for all } \mathbf{Y}^i, \mathbf{Y}^j \in \mathbf{T} \tag{3.14}$$

where $\mathbf{P}_T$ denotes the one step transition probability matrix. The choice of the updating rule explicitly determines the matrix $\mathbf{P}_T$. The Metropolis algorithm is used for the following development, and a proof is given that the resulting matrix $\mathbf{P}_T$ satisfies (3.14) with $P_T(\mathbf{Y})$ as defined in (3.13). Each point $\mathbf{Y}^i$ has a neighborhood, $N(\mathbf{Y}^i)$, with cardinality $K_i = |N(\mathbf{Y}^i)|$. Let $\delta(\mathbf{Y}^i)$ denote the degeneracy degree of $\mathbf{Y}^i$ [25], i.e. the number of zero valued elements in the BFS $\underline{x}^i$ corresponding to tableau $\mathbf{Y}^i$, and $\delta_{max} = \max_{\mathbf{Y}^i\in T} [\delta(\mathbf{Y}^i)]$. Then $|K_i| \le (\delta_{max}+1)(N-M) = |K|_{max}$ is the maximum number of adjacent tableaus [25]. Let

$$P_T(i,j) = z(i,j) \min\left[1, \exp\left(-\frac{g(\mathbf{Y}^j)-g(\mathbf{Y}^i)}{T}\right)\right] \tag{3.15}$$

where

$$z(i,j) = \begin{cases} \dfrac{1}{K_{\max}} & \text{if } \mathbf{Y}^j \in N(\mathbf{Y}^i) \\ 1 - \sum\limits_{j'\neq i} P_T(i,j') & \text{if } i=j \\ 0 & \text{if } \mathbf{Y}^j \notin N(\mathbf{Y}^i),\ i\neq j \end{cases} \tag{3.16}$$

The Metropolis algorithm [28,29] which produces this transition probability is described below. We first verify that eqn (3.15) satisfies eqn (3.14). For (3.14) to hold, it is sufficient to verify the detailed balance equation [29,30]:

$$P_T(\mathbf{Y}^i)\,\mathbf{P}_T(i,j) = P_T(\mathbf{Y}^j)\,\mathbf{P}_T(j,i) \quad \textit{for all} \quad \mathbf{Y}^i, \mathbf{Y}^j \in \mathbf{T} \tag{3.17}$$

Case 1 For $j \neq i$, $\mathbf{Y}^j \notin N(\mathbf{Y}^i)$ from eqn (3.16), $z(i,j)=0$ hence $\mathbf{P}(i,j)=0$. If $\mathbf{Y}^j \in N(\mathbf{Y}^i)$ then $\mathbf{Y}^i \in N(\mathbf{Y}^j)$, thus both sides are identically zero.

Case 2: For $j=i$, equality clearly holds.

Case 3: For $\mathbf{Y}^j \in N(\mathbf{Y}^i)$, rewriting eqn (3.17) we have

$$\frac{P_T(\mathbf{Y}^i)}{P_T(\mathbf{Y}^j)} = \frac{\mathbf{P}_T(j,i)}{\mathbf{P}_T(i,j)} \tag{3.18}$$

from eqn (3.13)

$$\frac{P_T(\mathbf{Y}^i)}{P_T(\mathbf{Y}^j)} = \exp\left(-\frac{g(\mathbf{Y}^i)-g(\mathbf{Y}^j)}{T}\right) \tag{3.19}$$

From eqn (3.15)

$$\frac{\mathbf{P}_T(j,i)}{\mathbf{P}_T(i,j)} = \frac{z(j,i)}{z(i,j)} \frac{min\,\{1,\ \exp-[g(\mathbf{Y}^i) - g(\mathbf{Y}^j)]\}}{min\,\{1,\ \exp-[g(\mathbf{Y}^j) - g(\mathbf{Y}^i)]\}} \tag{3.20}$$

and since 1 is minimum in either the numerator or the denominator, both cases yield

$$\exp\left(-\frac{g(\mathbf{Y}^i) - g(\mathbf{Y}^j)}{T}\right) = \frac{P_T(\mathbf{Y}^i)}{P_T(\mathbf{Y}^j)} \tag{3.21}$$

thus eqn (3.17) and (3.14) are satisfied.

3.2.3. Algorithm Description

To generate the corresponding time homogeneous Markov chain, we would use the following implementation of the Metropolis algorithm [29]:

1) Select initial basic feasible tableau, $\mathbf{Y}^n$, n=0.

2) Let $|N(\mathbf{Y}^n)| = K_n$, then select a tableau $\mathbf{Y}^j \in [N(\mathbf{Y}^n) \cup \mathbf{Y}^n]$ according to the probability:

$$Pr(\mathbf{Y}^j) = \begin{cases} 1/K_{max} & if\ \mathbf{Y}^j \in N(\mathbf{Y}^n) \\ 1 - K_n/K_{max} & if\ \mathbf{Y}^j = \mathbf{Y}^n \end{cases} \tag{3.22}$$

3) Compute cost $g(\mathbf{Y}^j)$ for new candidate tableau.

If $\Delta g = g(\mathbf{Y}^j) - g(\mathbf{Y}^n) \le 0$

Then: $\mathbf{Y}^{n+1} = \mathbf{Y}^j$

Else: generate random variable $r \sim \mathbf{U}[0,1]$

If $r \le \exp\left(\frac{-\Delta g}{T_n}\right)$

Then: $\mathbf{Y}^{n+1} = \mathbf{Y}^j$.

Else: $\mathbf{Y}^{n+1} = \mathbf{Y}^n$

4) n=n+1, go to 2).

Additionally, the algorithm requires an annealing schedule to monotonically decrease T_n at each step. As $T_n \to 0$ the Gibb's measure, $P_T(\mathbf{Y})$, will converge to the limit

$$P_0(\mathbf{Y}^\infty) = \begin{cases} \frac{1}{|\mathbf{T}^*|} & \mathbf{Y}^\infty \in \mathbf{T}^* \\ 0 & \mathbf{Y}^\infty \notin \mathbf{T}^* \end{cases} \tag{3.23}$$

where $\mathbf{T}^* = \{\mathbf{Y} \in \mathbf{T} : g(\mathbf{Y}) \leq g(\mathbf{Y}^i) \text{ for all } \mathbf{Y}^i \in \mathbf{T}\}$. For global convergence of the simulated annealing algorithm, i.e. the sequence $\underline{x}^n$ converges with probability 1 to some element $\underline{x}^* \in \mathbf{S}^* \to \mathbf{T}^*$, it is necessary that the time inhomogeneous Markov chain is strongly ergodic. The algorithm described above, produces a strongly ergodic Markov chain, provided the annealing schedule is of the form [29]

$$T_n = \frac{\gamma}{\log(n+n_0+1)} \qquad n=0,1,2, \cdots \tag{3.24}$$

where n_0 is any parameter $1 \leq n_0 \leq \infty$, and $\gamma \geq rL$. The graph radius, $r = \min_{\mathbf{Y}^i \in \mathbf{T}^\dagger} \max_{\mathbf{Y}^j \in \mathbf{T}} d(\mathbf{Y}^i, \mathbf{Y}^j)$, where $d(\mathbf{Y}^i, \mathbf{Y}^j)$ is the distance between the two nodes measured as the number of edges in the minimum length path from $\mathbf{Y}^i$ to $\mathbf{Y}^j$, and $\mathbf{T}^\dagger$ is $\mathbf{T}$ less all nodes of locally maximum cost. L is a Lipshitz-like constant which is a measure of the maximum cost differential, $|g(\mathbf{Y}^j) - g(\mathbf{Y}^i)|$ over all possible pairs, i and j, of neighboring nodes in G. Though (3.24) is a powerful theoretical result sufficient for guaranteed optimal convergence, in real applications these parameters may be hard to obtain, and if known, would be too large for reasonable computation times. In both [28] and [29] it is indicated that much smaller values of γ may be used with acceptable results.

Since the simulated annealing algorithm exhibits only asymptotic convergence, and yet the graph G has only a finite number of nodes, this algorithm is of practical importance only if its finite time behavior yields improved solutions over the deterministic simplex search. In [29] the finite time behavior of the algorithm is analyzed, and a bound computed for the deviation between the optimal cost and the finite time cost. For a finite sequence, the terminal result is not guaranteed to be the lowest cost solution in the sequence due to the random search. Therefore, truncated procedures can improve performance by maintaining memory storage of the lowest cost solution achieved up to the current iteration. This storage is updated only when the current result is better than all previous points in the sequence. As discussed below, our experiments verify that the truncated sequence produces improved results over the deterministic simplex search.

3.2.4. Algorithm Performance

The stochastic search algorithm was applied successfully to seismic deconvolution and sparse array design problems, as described in Chapters 5 and 6. In both applications, examples of lower order solutions than those obtained with the $l_{1/q}$ simplex search were demonstrated, thus validating the algorithm's utility. In some cases however, no improvement was achieved, but no stochastic search result was less sparse than that of an $l_{1/q}$ simplex search. It is proposed that when equivalent results arise from the two algorithms, it is not due to failure of the stochastic search, but to the usually high quality solutions found by the the $l_{1/q}$ simplex search. Though in the worst case the $l_{1/q}$ simplex search may theoretically find only a poor local minimum, as Kirkpatrick, et al. argued [21], when the size of the optimization problem is large, it is the average performance that dominates, not the worst case. Verification of finding a global optimum has been diffi-

cult to establish, since most problems of significant size have unknown optima. It is very difficult to synthesize a problem with a known optimal solution which is also complex enough that both algorithms do not find the solution in a few iterations. Validation has therefore consisted of demonstrating improved solutions over those obtained with other algorithms.

The major algorithm drawback is the need for very many iterations to meet the slow annealing schedule needed to insure optimal convergence (eqn (3.24)). This is particularly troublesome when traversing the graph, since error accumulates at each pivot move. A problem requiring several hundred iterations with the $l_{1/q}$ simplex search will retain high precision, but the stochastic search would require many thousands of steps, introducing unacceptable accumulative error. This necessitates periodic "reinversion" to directly compute, from the original matrix, the inverse of the current basis.

Three types of annealing schedules have been used for updating T: 1) eqn (3.24) was used directly, but with very small γ, 2) exponential decay, and 3) manual control by the operator was used to reduce temperature stepwise and control the number of iterations at each setting. Method 1) was found to be prohibitively slow, even for small γ, but both 2) and 3) were used successfully. For a typical problem, the $l_{1/q}$ simplex search was first computed for comparison, then method 2) or 3) was used repeatedly until a temperature decay rate was found which improved the solution but terminated in reasonable time.

3.3. The Convex Transformation Gradient Search Algorithm

3.3.1. Quadratic Constraint Minimum Order Problems

In some applications the constraints may be better expressed as an upper bound on the l_2 norm of the deviation of $\mathbf{H}\underline{x}$ from the desired vector $\underline{b}$, as in the quadratic form:

$$\min_{\underline{x}} g(\underline{x}) = \sum_{i=1}^{N} |x_i|^{1/q} \; s.t. \; (\mathbf{H}\underline{x} - \underline{b})^T (\mathbf{H}\underline{x} - \underline{b}) \leq \varepsilon, \; x_i \geq 0, \; q>1 \quad (3.25)$$

The convex transformation gradient search presented here is an efficient algorithm which can handle systems of large dimension. Though global optimality of the result is not assured, as with the $l_{1/q}$ simplex algorithm, it produces locally optimal solutions which in practice achieve low order.

If the hyper-volume defined by the constraint $(\mathbf{H}\underline{x} - \underline{b})^T (\mathbf{H}\underline{x} - \underline{b}) \leq \varepsilon$ does not contain the origin, then the globally optimal solution must lie on its surface. The surface is smooth and contains no isolated extreme points, so we must search a continuous surface, rather than a finite set of points for the optimum. For problems of this form we may not define a finite set of basic solutions, and therefore cannot rely on a simplex search approach. Attempts at straightforward constrained optimization gradient search techniques are doomed by the numerous extremely strong local minima of the objective function $g(\underline{x})$.

3.3.2. A Convexity Transformation for Objective Function Regularization

The convex transformation approach maps the system into a space which eliminates the extremely steep sided local minima, and improves the computational and numerical aspects of a gradient search by giving us a convex cost objective functional.

For each point $\underline{x}$ in the original space we define the isomorphic one-to-one mapping:

$$\{\underline{x} \mid \underline{x} \in R^N, x_i > 0\} \leftrightarrow \{\underline{y} \mid y_i = \frac{1}{q} ln(x_i), \underline{y} \in R^N, y_i > -\infty\} \quad (3.26)$$

Equation (3.25) then becomes

$$\inf_{\underline{y}} h(\underline{y}) = \sum_{i=1}^{N} e^{y_i} \quad s.t. \ (\mathbf{H}e^{q\underline{y}} - \underline{b})^T(\mathbf{H}e^{q\underline{y}} - \underline{b}) \leq \varepsilon, \text{ and } y_i > -\infty \quad (3.27)$$

where $e^{\underline{y}}$ denotes point by point exponentiation of each element of a vector $\underline{y}$.

Since e^{y_i} is a strictly convex function, and sums of convex functions are convex, $h(\underline{y})$ is a strictly convex functional over $\underline{y}$. Note that although we have to restrict $x_i \neq 0$ (i.e. $y_i > -\infty$), we may allow x_i to be arbitrarily close to zero, and thus consider low order solutions as having the largest possible number of elements within an ε neighborhood of zero. This transformation is similar to one used in solving the geometric programming problem by transformation to a convex program [31,32]. In our case however, analysis of the Hessian matrix for the constraint indicates that it is in general not positive definite, and therefore the original convex set defined by the constraint becomes nonconvex. This complication prohibits us from proving optimal convergence for a descent algorithm, and requires us to use a nonlinear constraint optimization algorithm. This penalty is offset by

the regularization performed on the objective function which is necessary to even consider using a descent algorithm. $h(\underline{y})$ has no strong minima or cost ridges to sidetrack a descent algorithm, and $\nabla h(\underline{y}) \rightarrow \underline{0}$ near the optimal solution, while at the same point $\nabla g(\underline{x}) \rightarrow \pm\infty$.

3.3.3. The Schittkowski Nonlinear Optimization Algorithm

Equation (3.27) has been solved successfully for a number of sample problems using an exterior point, nonlinear constrained optimization algorithm due to Schittkowski [33], while this and other algorithms failed with the original form, eqn (3.25). This algorithm is available in the IMSL library of computer utility programs and uses successive quadratic programming method to solve the general nonlinear programming problem.

We define the constraint as $s(\underline{y}) = \varepsilon - (\mathbf{He}^{q\underline{x}} - \underline{b})^T(\mathbf{He}^{q\underline{x}} - \underline{b}) \geq 0$, $\underline{y}_l \leq \underline{y} \leq \underline{y}_u$, with $\underline{y}_l$ and $\underline{y}_u$ selected to insure computation of $s(\underline{y})$ remains within the limits of machine precision. The algorithm uses iterative formulation and solution of quadratic programming subproblems by quadratic approximation of the Lagrangian and by linearizing the constraints as follows:

$$\min_{\underline{d} \in R^N} \frac{1}{2}\underline{d}^T B_k \underline{d} + \nabla h(\underline{y}_k)^T \underline{d} \tag{3.28}$$

subject to:

$$\nabla s(\underline{y}_k)^T \underline{d} + s(\underline{y}_k) \geq 0, \quad \underline{y}_l - \underline{y}_k \leq \underline{d} \leq \underline{y}_u - \underline{y}_k$$

where B_k is the positive definite approximation of the Hessian at iteration k, and $\underline{y}_k$ is the current iterate. With $\underline{d}_k$ the solution of the subproblem, a line search is used to find the new point $\underline{y}_{k+1}$,

$$\underline{y}_{k+1} = \underline{y}_k + \lambda \underline{d}_k, \quad \lambda \in (0,1] \tag{3.29}$$

such that the augmented Lagrange merit function [33] is reduced at the new point. The gradients for our system are computed at each iteration as:

$$\nabla h(\underline{y}_k) = e^{\underline{y}_k} \tag{3.30}$$

$$\nabla s(\underline{y}_k) = 2qe^{q\underline{y}} \bullet \left[\mathbf{H}^T\underline{b} - \mathbf{H}^T\mathbf{H}\, e^{q\underline{y}} \right]$$

where • indicates point-by-point vector multiplication. When optimality is not achieved, B_k is updated according to the modified BFGS formula [34]. At each iteration, cost function and constraint error evaluations, and their gradient vectors, are computed from the current estimate of $\underline{y}$. Algorithm termination is accomplished when an error measure of the Kuhn-Tucker conditions is sufficiently small . Execution times in our experiments compare favorably with the $l_{1/q}$ simplex algorithm for similar sized systems. By using a least squares solution as a search starting point, the algorithm has been successful in yielding very low order solutions for systems as large as 100 or more variables. A penalty method approach [17,18] for constrained optimization has also been used successfully in solving eqn (3.27). Chapter 5 contains examples of applying the convex transformation gradient search.

CHAPTER 4: MAXIMALLY SPARSE OPTIMIZATION FOR NEUROMAGNETIC IMAGING

4.1. Problem Definition

The aim of Neuromagnetic Image (NMI) reconstruction is the production of three dimensional vector maps of neuron current densities arising from brain activity. These image maps are produced by measuring the extracranial induced magnetic field around the skull and performing a numeric inverse reconstruction to infer the underlying neuron currents. The ability to measure this very weak magnetic field is provided by Superconducting Quantum Interference Device (SQUID) detectors, which have been used for a number of years in the study of biomagnetic phenomena [35]. Evoked responses to sensory stimuli (visual, auditory, somatic) and some higher brain functions have been studied as neuromagnetic sources and hold potential for NMI [35,36,37].

For NMI, the brain volume of interest is divided into a grid of 3-D volume or voxel cells, each of which will be assigned a vector value which will represent the average directional current flow through the cell's volume at a given instant of time. The magnetic field is measured at the skull surface on a grid of sample points as in Figure 4.1 by moving the SQUID to each new position and repeating the stimulus, or by using an array of detectors to get a simultaneous measurement. From these sampled measurements, the 3-D source image estimate is reconstructed. NMI thus does not attempt to precisely locate actual individual neuron currents, but produces an image of the average current for all neurons within individual discrete fixed voxels in the brain.

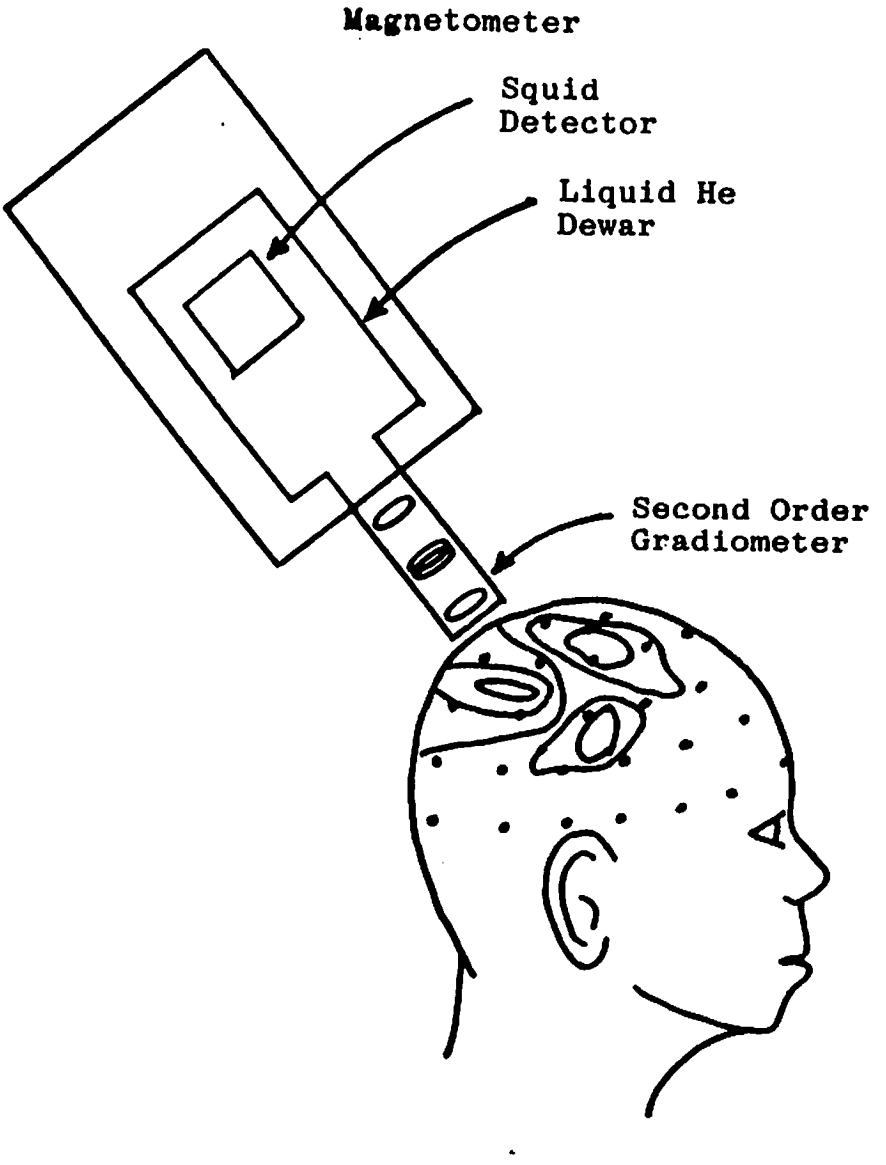


Figure 4.1. NMI Magnetic Field Sampling Using SQUID Detector

Manbir Singh et al. first demonstrated NMI with two dimensional reconstructions of sources constrained to lie in a single plane whose depth was adjusted to produce the lowest squared error in the solution [38]. They produced images of the visually evoked response from data acquired from human subjects using a single channel SQUID. The case of reconstruction in a 3-D volume rather than on a plane is considerably more difficult since there is no unique solution to a current source in a closed volume even if the external magnetic field is known everywhere [39]. For example, in a spherical volume conductor like the brain, any radial current flow, or current distributions with self spherical symmetry, produces no external magnetic field. It is the intent of this study to extend the 2-D results to the 3-D case and identify algorithms which will lead to meaningful solutions.

NMI is an attractive medical imaging mode for several reasons. It is a non-invasive passive technique using only biologically induced fields. It will provide a method of functional rather than structural imaging of the brain which, though it may not be high resolution, can provide information not available from other techniques. The methods of NMI may also be useful for other applications of current imaging in closed volumes, such as cardiac magnetic imaging and geomagnetic surveys.

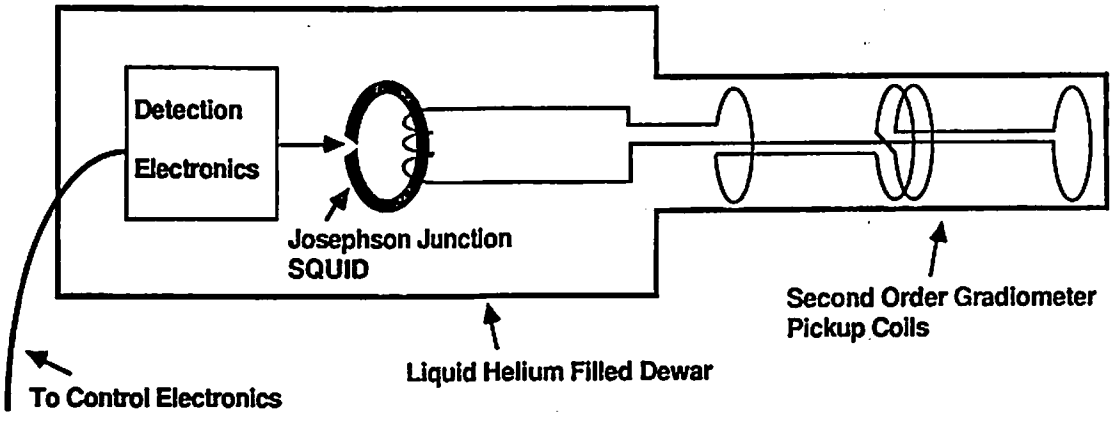


Figure 4.2. Schematic Representation of SQUID Construction

4.2. Previous Work Related to NMI

4.2.1. The SQUID Detector

Biomagnetic fields of the heart were detected in shielded rooms with wire loop magnetometers by Baule and Mcfeen more than twenty years ago [40], but the area of biomagnetic research did not flourish until the development of the Superconducting QUantum Interference Device (SQUID) in the early 1970's. Using a SQUID detector in a shielded room, Cohen measured magnetic fields of the brain and many other organs [41,42]. The SQUID uses a liquid helium cooled superconducting pickup coil and Josephson junction device as a very sensitive low noise magnetic detector with ability to measure fields from d.c. to several kilohertz. A noise threshold of 10^{-13} $T/Hz^{1/2}$ (Tesla per square root Hertz), which is in the range needed for biomagnetic research, is attainable in urban locations [43]. Another key to widespread use of the SQUID is the application of a second order gradiometer pickup coil configuration to reject all signals but those with second and higher order spatial gradients. This rejects distant interfering sources due to the weak gradients they would have while near sources are detected. The gradiometer consists of three sets of counterwound coils as shown in Figure 4.2. A typical system consists of a set of magnetic pickup coils wound in the gradiometer configuration, the SQUID detector, associated electronics, and a dewar to encase the coils and SQUID in liquid helium in order to keep them at superconducting temperatures. In a closed superconducting coil the magnetic flux through the coil is constant, so any external magnetic field threading the coils causes a current to flow to oppose the external field. This current is coupled to the SQUID detector which serves as a low noise, high gain, current to voltage amplifier.

The SQUID biomagnetometer in this configuration provides sufficient sensitivity and noise immunity to make neuromagnetic measurements in an unshielded laboratory environment possible [36,37]. However, shielded rooms are still useful in reducing the noise level and are used by many researchers, [35] and since the brain's fields are among the weakest of those generated by the body, it is still necessary to do signal averaging to increase signal to noise level. This is accomplished by repeating the stimulus for an evoked response to synchronize a series of data sampling windows while the SQUID is held stationary. Since spontaneous brain activity and external noise are uncorrelated with the stimulus, a point-by-point averaging across the data sampling windows reduces the interference level, while the synchronized evoked waveform adds coherently. The expected measurement accuracy for NMI uses will require both signal averaging and shielded room data acquisition. SQUID systems are commercially available in several configurations including single channel and five or seven channel array devices with typical pickup coil diameters in the range of one to two centimeters. The need to use liquid Helium to maintain superconductivity is a costly inconvenience which may soon be overcome. There is much active research now that is identifying high temperature (liquid Nitrogen and above) superconducting materials. It is hoped that this will lead to less costly instruments and to larger array SQUID devices which will be a major factor in opening up new applications to magnetic imaging. IBM recently announced that it is working on a prototype high temperature SQUID.

4.2.2. Early Neuromagnetic Studies

With the use of SQUID detectors, a number of measurable biomagnetic sources in humans have been observed, including fields of the heart, eye, skeletal muscles, lungs (due

to contaminations), fetus, and brain. In most cases, the incentive to seek measurable magnetic fields was provided by the existing base of knowledge of electrical potentials produced by these sources, as with the electrocardiogram (ECG) and electroencephalogram (EEG). Figure 4.3 plots the relative electrical and magnetic field strengths for many of the important biomagnetic sources. It can be seen that the cardiogram is by far the strongest signal and that the brain fields are the weakest, with the evoked responses the weakest of these. This indicates the technical challenge inherent in imaging using neuromagnetic sources where even other biological sources can produce major levels of interference and why the neuromagnetic fields have been some of the most recent to be measured.

The first fields of the brain to be observed were due to the spontaneous activity known as the alpha, beta, and delta rhythms which had been studied in electroencephalography [44]. The study and recording of these time domain magnetic waveforms has been called magnetoencephalography (MEG) and still constitutes an active area of research in neuromagnetism [45]. The magnetic signals have been used simultaneously with EEG measurements to provide complimentary information, and it has been found that the magnetic fields can provide a better localization of the source [46]. It is theorized that the improved localization is because the skull provides an insulating barrier which spreads the potential out over the scalp while the magnetic field is essentially unaffected due to the skull's low permeability.

Brenner, Kaufman, and Williamson first demonstrated measurement of the somatic and visually evoked fields in 1978, and since then these, and the auditory evoked response, have been widely studied [35,36,37]. The somatic evoked field is produced at the somatisensory cortex of the brain with an electrical shock to a finger, the visual field at the

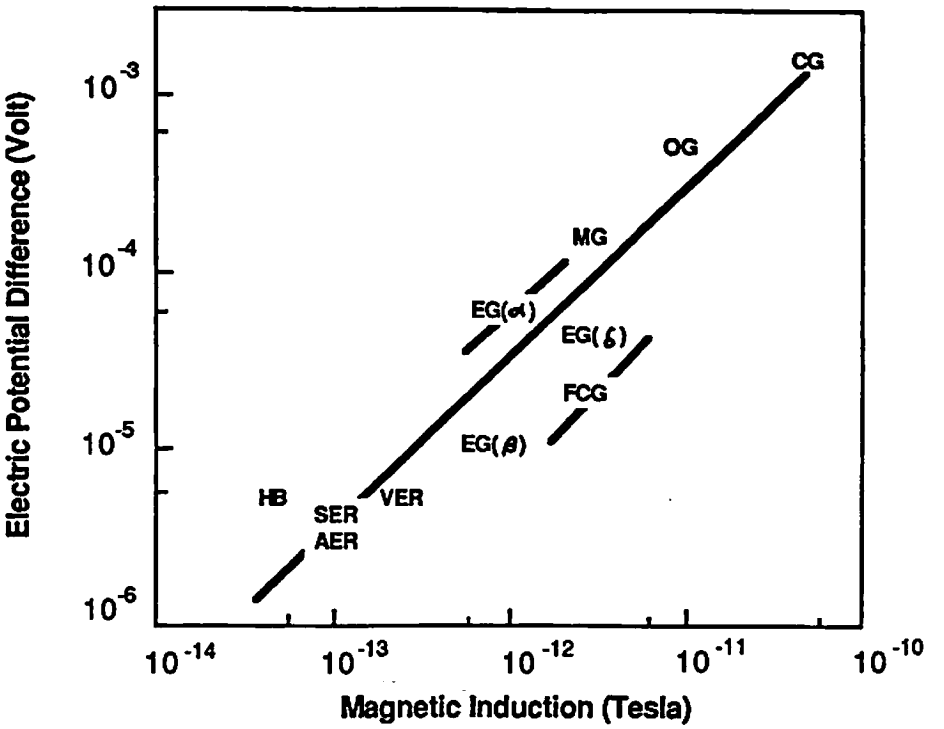


Figure 4.3. Comparison of Bioelectric Skin Voltages and Biomagnetic Fields.
Cardiogram = CG, Oculogram = OG, Myogram = MG, Fetal Cardiogram = FCG, Encephalogram = EG, Alpha Rhythm = EG(α), Delta Rhythm = EG(δ), Beta Rhythm = EG(β), Visually Evoked Response = VER, Somatic Evoked Response = SER, Auditory Evoked Response = AER. (after Williamson [43]).

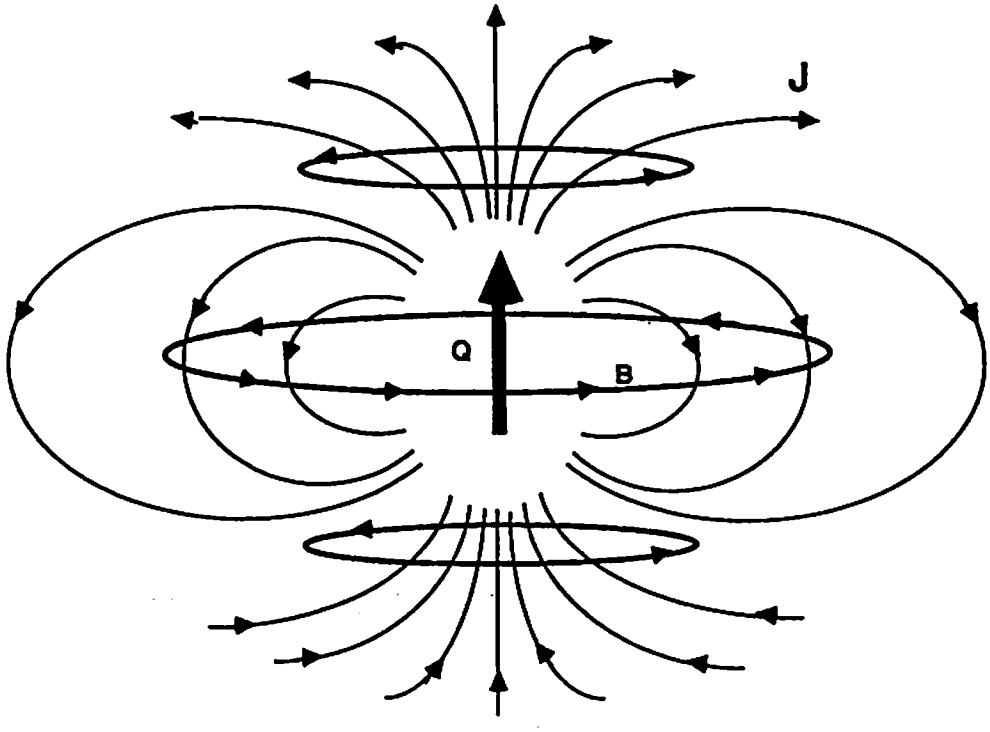


Figure 4.4. Volume Currents, J, and Induced Magnetic Field, B, from a Current Dipole Q.

visual cortex by repeatedly changing a regular geometric pattern or grating in the field of view, and the auditory field at the auditory cortex in response to short tones. They are of particular interest for NMI because images can be used to identify functional regions involved in these responses and their extent, and because these fields are induced by an external stimulus which can be repeated so as to allow synchronization of signals acquired during movement of the SQUID to the many sample grid positions. More recently, an MEG approach has been used to identify the epileptiform spike waveform associated with epilepsy, and other abnormal brainwave functions have been studied using SQUID's [47]. Presently the leading edge of neuromagnetic research is the measurement of fields related to higher brain functions. E.Flynn and D. Arthur of the Los Alamos National Laboratory Physics Division are presently studying the "attention wave" which is a measure of loss of mental concentration on a repeated task related to auditory stimulus. Perception and event related brain potentials associated with higher brain functions such as sentence processing and visual interpretation are being studied and are leading to corresponding neuromagnetic analysis [48,49]. Also, study of neuromagnetic fields related to learning has been proposed by the Los Alamos group. It is felt that these higher brain functions will have more complex and distributed underlying neuron current structures, and will therefore be best analyzed by an imaging technique which can reveal the detail of this structure.

4.2.3. Neuromagnetic Source Models

The brain currents of interest for NMI are the neuron intracellular currents in the cortex. There is an overall characteristic laminar and columnar organization in the cerebral cortex with 6 or more distinct horizontal regions and the majority of the cells extending verti-

cally. Over 70% of the neuron cells of the cortex are of the pyramidal cell type and are organized into repetitive vertically oriented (with respect to the local cortex surface) columnar units of about 200-300 μm in diameter [50,51]. The columnar units contain many cells but they have been shown to operate or "fire" in unison. The direction of current flow, either up or down, depends on whether inhibitory or exitory stimulus is received at a cell's synapses. The dendrites that interconnect local cells, and the afferent fibers which provide longer interconnections in the white matter layer beneath the cortex, carry currents that are not vertical, but which have been shown to be of low enough level so that the intracellular currents in the pyramid cells remain dominant. This orderly structure of the cortex lends itself well to a model of the neuro-current sources as sets of current dipoles. A current dipole is a vector quantity with a spatial orientation and a dipole moment, the orientation being the direction of a linear current flow between a point source and sink, and the moment being the product of the length of the current path and the current magnitude. The magnetic field produced by a current dipole is shown in Figure 4.4 and has a magnitude of

$$B_\phi = \mu_0 Q \sin \frac{\theta}{4\pi r^2} \tag{4.1}$$

where Q is the dipole moment, θ and r the angle and distance to the magnetic measurement point, and μ_0 the permeability of freespace. It is notable that the field drops off less rapidly with distance than the $\frac{1}{r^2}$ loss for a magnetic dipole. Based on the above physiological arguments, most researchers in neuromagnetism have used the current dipole model for brain activity rather than continuous current fields or magnetic dipoles, even when gross brain currents are averaged over a large region several centimeters in cross section [43]. The majority of neuromagnetic research has assumed the source could be represented with a single equivalent current dipole [35,43,52,53] but recently more

complex multiple dipole sources have been shown to be required to fit the measured data [54,55].

It has also been common to model the brain and skull as a nonconducting sphere filled with a homogeneous conducting medium or fluid in which the current dipoles are embedded. The source and sink of the dipole create an electric field in the conducting medium which leads to a primary dipole current path and volume currents as shown in Figure 4.4. These volume currents are considered noise sources since it is the primary impressed current that is of interest. The assumption of a homogeneous conducting sphere model however offers some important simplifications. In such a geometry, the radial component of any current source contributes neither to the tangential nor radial components of the surface magnetic field [*51]. Also, the volume currents induced by any dipole will not contribute to the radial magnetic field. This means that the tangential component of the primary current dipole can be measured without interference from volume currents by observing the radial magnetic field. The skull, however, is not a sphere, and the brain and its fluids do not present a homogeneous conductor, so there had been some doubt as to the validity of this model. However, in [56] Barth et al. reported a recent experiment using a real cadaver head with an implanted current dipole to demonstrate that the spherical volume conductor model was a good fit, even with large anomalies in the conducting volume.

4.2.4. Previous Solution Methods

Much of the neuromagnetic data gathered is displayed in direct methods such as the MEG or isofield maps [45,57]. With the MEG, data is displayed as a plot of the time series measured at each SQUID location. Isofield maps plot contour lines on the skull of the

amplitudes of the normal component of the magnetic field at a single instant by interpolating between the SQUID sample grid points as shown in Figure 4.1. The amplitudes shown in these maps are a static representation of a time sample of a single component of an evoked response. Although useful, neither of these approaches draw inferences on the underlying source locations; we are more interested in the inverse solution.

4.2.4.1. Dipole Fitting

The most widely used inverse solution is the single dipole fit. This technique assumes that the entire magnetic field is produced by a single equivalent current dipole. The solution is often obtained by simply measuring the distance, d, between the positive and negative extreme field points in the isofield map [58]. The dipole source lies midway between them, perpendicular to the connecting line, and at a depth equal to $\sqrt{\frac{d}{2}}$. This simplistic approach seldom yields accurate solutions for real data. A more sophisticated approach to the single dipole solution is the moving dipole fit algorithm [52,53]. The solution is obtained as a six dimensional parameter estimation problem where 3-D position, orientation, and amplitude are determined. These parameters are adjusted in a continuous fashion until the solution which minimizes the squared error between the forward projected data and the measured data is found. This approach has been extended to two or three dipoles, but becomes computationally intractable beyond that. Arthur and Flynn [59] have enhanced the moving dipole fit by computing rectangular confidence regions around the location in which the source must lie. They showed non-overlapping confidence regions for different auditory evoked responses indicating the change in location was not due to noise in the data.

Often the measured fields are a very poor match to a single equivalent dipole [54]. There have been a few attempts to deal with solutions involving multiple dipoles. One approach

is to perform a "multipole expansion" where a single dipole is fit in a least squares sense to the data using eqn (4.2.1), and the forward projected field from this dipole is subtracted from the data to obtain the error term [60,61]. This process is repeated with additional dipoles until the remaining error reaches an acceptable level. The resulting distribution, however, can be very different from the actual source. For example, two parallel dipoles would be represented as an infinite series of dipoles of decreasing magnitude, the largest located midway between the two. In one experiment, twenty dipoles of fixed location and orientation have been used to provide a least squares solution to cardiac current sources [62]. More than twenty magnetic measurements were taken to provide an overdetermined system. In each of these multipole approaches the problem interpretation was one of parameter estimation for a limited parameter set rather than image reconstruction.

4.2.4.2. Image-like Solutions

As mentioned above, Singh, et al. first demonstrated neuromagnetic imaging by reconstructing visually evoked response data on a single plane [38]. They made the simplifying assumptions that all sources were contained in a single plane and that they were all oriented parallel or antiparallel. The plane was divided into equal size square pixels, each of which was allowed to contain a single dipole. A linear matrix system equation was developed to relate the dipole amplitudes to the magnetic measurements by using a discrete form of the Biot-Savart equation (this will be discussed in section 4.3). For a given assumed reconstruction plane depth they computed the solution image using an additive algebraic reconstruction technique (ART) algorithm. This iterative algorithm will converge to the minimum norm solution if the data are consistent [63]. For noisy data, ART will converge to a solution where the difference between the measurements and the for-

ward projected solution is on the order of the measurement noise. To locate the unknown depth of the dipole plane, Singh computed reconstructions at a series of depths to find the one with least error between measurement data and the forward projected solution. This technique appears usable when the source is constrained to a single plane, but the author has shown that it may bear no relationship to an actual 3-D source [1].

An alternative formulation was proposed by Dallas in [64] based on Maxwells equations for non-time-varying fields:

$$\nabla \times B = \mu_0 J \quad , \quad \nabla \cdot B = 0 \tag{4.2}$$

Taking the Fourier transforms of (4.2) yields a set of linear equations relating the current and magnetic fields. By sampling the Fourier transform of the two fields and decomposing the magnetic field into two regions, the measurement region and a "forbidden region" over which the field cannot be measured, a large set of linear equations can be formed. The unknowns in the equation are the samples of the Fourier transforms of the current field and the magnetic field in the forbidden region. In the algorithm, the reconstruction volume and measurement region are discretized into sample cells as in our model. This formulation has the advantage that it provides simultaneous reconstruction of the internal current and magnetic fields. Dallas has demonstrated successful two dimensional reconstructions of simulated data. The Fourier space solution approach however does not eliminate the ill-posed nature of the problem . The reconstructed images using the Fourier domain approach appear to be of minimum norm type and should lead to the same problems if applied to 3-D distributions that the 2-D ART reconstruction does.

4.3. Models for Static NMI

For NMI reconstruction we adopt the simple physical model of Figure 4.5. A spheroid shaped reconstruction volume represents the interior of the skull which will contain a 3-D distribution of neuron currents. Measurements of the external magnetic field are taken at points on the sampling surface represented by the hemispherical shell. These points correspond to the positions on the skull where measurements with the SQUID gradiometer are taken. Our goal is to infer the current distribution from these measurements.

The relationship between a continuous vector current field and its induced magnetic field at a point r under in space is given by the vector integral form of the Biot-Savart Law:

$$B(\underline{r}) = \frac{\mu_0}{4\pi}\int \frac{J(\underline{r}') \times (\underline{r}-\underline{r}')}{|\underline{r}-\underline{r}'|^3} d^3r' \qquad (4.3)$$

where $J(\underline{r}')$ denotes the vector current density at $\underline{r}'$ and μ the magnetic permeability of the medium, which we approximate with μ_0, the permeability of free space. Although the brain's current field consists of discrete firings of individual neuron cells, the high density of neurons in brain tissue and the inability of present instrumentation to resolve single cell current flow make this continuous field model an accurate one. However, we can take only a finite number of magnetic field measurements, and in order to reduce the dimensionality of the problem and express this nonlinear relationship as a set of linear equations we approximate eqn (4.3) with a discrete sum. The vector current field is replaced by a finite number of current dipoles, $Q(\underline{r}_i)$ located in a three dimensional grid, where a dipole's orientation and magnitude are determined by integrating the current field over the volume cell (voxel) surrounding the dipole. Equation (4.3) becomes:

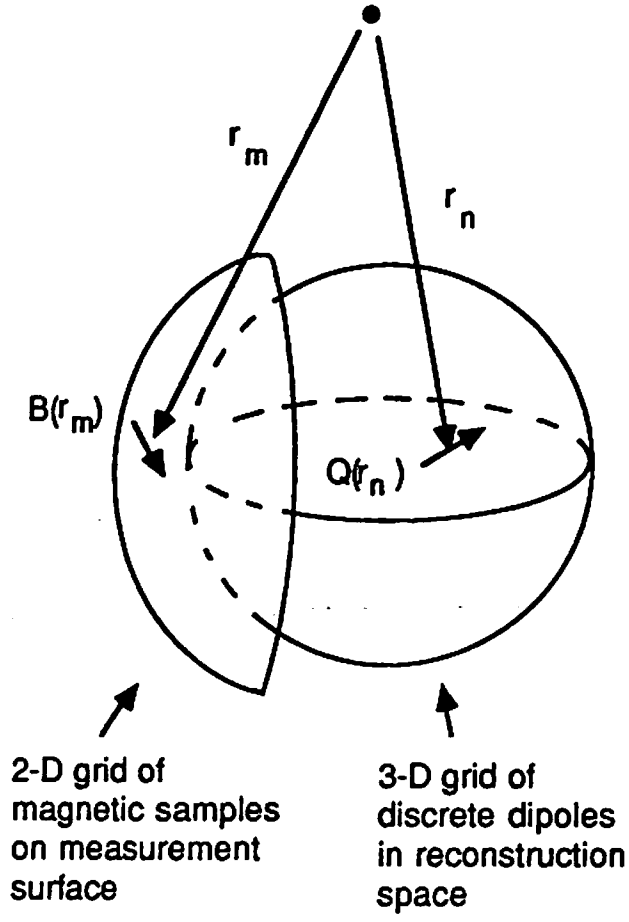


Figure 4.5. Basic Physical Model for Neuromagnetic Imaging.

$$B(\underline{r}_m) = \frac{\mu_0}{4\pi}\sum_{n=1}^{N}\frac{Q(\underline{r}_n)\times(\underline{r}_m-\underline{r}_n)}{|\underline{r}_m-\underline{r}_n|^3} \tag{4.4}$$

where N is the total number of dipole cells and m and n are the indices on the discrete sample points in space. This equation can be rewritten in vector-matrix form as the linear system

$$\underline{B} = W\,\underline{Q} \tag{4.5}$$

where

$$\underline{B} = [B_{x1}\ , B_{y1}\ , B_{z1}, \cdots B_{xM}\ , B_{yM}, B_{zM}\ ]^T$$

$$B_{xm} = x\ component\ of\ \ B(\underline{r}_m)$$

$$\underline{Q} = [Q_{x1}\ , Q_{y1}\ , Q_{z1}, \cdots Q_{xM}\ , Q_{yM}, Q_{zM}\ ]^T$$

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_{1,1} & \cdots & \mathbf{W}_{1,N} \\ \vdots & & \vdots \\ \mathbf{W}_{M,1} & \cdots & \mathbf{W}_{M,N} \end{bmatrix}\ ,\ \ a\ (3m\ by\ 3N)\ \ block\ matrix$$

$$\mathbf{W}_{m,n} = \frac{\mu_0}{4\pi|\underline{r}_m-\underline{r}_n|^3}\begin{bmatrix} 0 & r_{z:m,n} & -r_{y:m,n} \\ -r_{z:m,n} & 0 & r_{x:m,n} \\ r_{y:m,n} & -r_{x:m,n} & 0 \end{bmatrix}$$

$$r_{x:m,n} = x\ component\ of\ \ (\underline{r}_m-\underline{r}_n)$$

With the addition of an independent noise term, $\underline{v}$, to represent measurement error, the system is expressed in a form suitable for applying digital image reconstruction tech-

niques, i.e. we may solve the inverse problem of finding $\underline{Q}$ given the measurement vector $\underline{B}$ where:

$$\underline{B} = W\,\underline{Q} + \underline{v} \tag{4.6}$$

The formulation of the entries in matrix **W** can be adjusted to compensate for the fact that a squid gradiometer does not provide an exact measurement of the point flux density. Methods for doing so are discussed in section 4.2.

The commonly used model of the brain and skull as a nonconductive spherical casing filled with a homogeneous conductive medium in which current dipole activity exists [35,43] enables the introduction of several simplifications and constraints. As previously mentioned, volume currents do not contribute to the external field normal to the surface [35,65], and dipoles aligned with the sphere's radii, and radially symmetric dipole distributions produce no measurable external magnetic field [43]. The "invisibility" of these sources makes their inclusion in a reconstruction solution meaningless, so we may neglect them.

The skull is not a perfect sphere, however Barth's experiment has shown that the measurements on the skull show little deviation from that predicted by a spherical model. Also, most areas of the skull can be fit to a spherical segment with a local radial center. These considerations enable us to utilize the constraints offered by the spherical conductor model even when our reconstruction space is not exactly spherical.

The brain region of primary interest for NMI is the cortex. As we have seen, neurons within the cortex, and thus the current paths, are arranged predominantly normal to the local surface [50]. If the structure of the cortex surface shape can be mapped a-priori (e.g. by magnetic resonance imaging) we can assume the locally normal orientation to

provide an additional constraint on the source dipoles and simplification to the formulation. This approach could also be used in other current field reconstructions where the currents flow directions a known, but not their magnitudes.

These constraints are incorporated in the system by modifying eqn (4.4) as follows:

$$B(\underline{r}_m) = \frac{\mu_0}{4\pi}\sum_{n=1}^{N}\frac{|Q_n|\,\underline{z}_n(\underline{r}_n)\times(\underline{r}_m-\underline{r}_n)}{|\underline{r}_m-\underline{r}_n|^3} \tag{4.7}$$

where $\underline{z}_n(\underline{r}_n)$ denotes the unit vector representing the orientation of the n^{th} current dipole of magnitude $|Q_n|$. Expressing this as a vector-matrix equation, we modify (4.5) as follows: let $\underline{Q}= D\,\underline{Q}'$ where **D** is a tri-diagonal matrix of the known direction cosines of constrained dipole orientations and $\underline{Q}'$ is an N element vector of the dipole magnitudes.

$$\underline{B} = WD\,\underline{Q}' + \underline{v} \tag{4.8}$$

$$\mathbf{D} = \begin{bmatrix} \alpha_1 & 0 & \cdot & \cdot & \cdot & 0 \\ \beta_1 & 0 & & & & 0 \\ \gamma_1 & 0 & & & & 0 \\ 0 & \alpha_2 & & & & \cdot \\ 0 & \beta_2 & & & & \cdot \\ 0 & \gamma_2 & & & & 0 \\ \cdot & 0 & & & & \alpha_N \\ \cdot & \cdot & & & & \beta_N \\ 0 & 0 & \cdot & \cdot & \cdot & \gamma_N \end{bmatrix}$$

$$\underline{Q}' = [\,|Q_1|,\ |Q_2|,\cdots |Q_N|\,]^T$$

then eqn (4.5) becomes:

$$\underline{B} = \boldsymbol{H}\,\underline{Q}' + \underline{\nu} \tag{4.9}$$

$$\mathbf{H} = \mathbf{WD} = \begin{bmatrix} \mathbf{H}_{1,1} & \cdots & \mathbf{H}_{1,N} \\ \vdots & & \vdots \\ \mathbf{H}_{M,1} & \cdots & \mathbf{H}_{M,N} \end{bmatrix}$$

$$\mathbf{H}_{m,n} = \frac{\mu_0}{4\pi \left| \underline{r}_m - \underline{r}_n \right|^3} \begin{bmatrix} \beta_n r_{z:m,n} - \gamma_n r_{y:m,n} \\ -\alpha_n r_{z:m,n} + \gamma_n r_{x:m,n} \\ \alpha_n r_{y:m,n} - \beta_n r_{x:m,n} \end{bmatrix}$$

Note that (4.9) includes the measured data in $\underline{B}$ for all three vector components of the field at each sample point. If only normal measurements are taken, then the projection of the normal field onto these rectangular vector components is used. This formulation and model permit significant reduction in the dimensionality of the problem, but in general still do not lead to a unique solution.

4.4. Inverse Solution Feasibility

The NMI problem has several physical restrictions which limit the ability to produce high quality reconstructed images, and ultimately provide the motivation for using the "minimum dipole" approach. These difficulties include:

(i) The ill-posed nature of the system equations.

(ii) The resolution characteristics of the SQUID gradiometer.

(iii) The high noise level of the background magnetic fields compared to the neuromagnetic field

4.4.1. System Equation Considerations

Consider the linear problem $\underline{B}= \mathbf{H}\,\underline{Q}' + \underline{v}$ of eqn (4.6) where the system matrix **H** is of dimension $3M\times N$. We will investigate the properties of this matrix. Expanding **H** using the singular value decomposition [66]:

$$H = U\,\Lambda^{1/2}\,V^T \tag{4.10}$$

where **U** and **V** are matrices of the eigenvectors of $\mathbf{HH}^T$ and $\mathbf{H}^T\mathbf{H}$ respectively, with Λ the diagonal matrix of corresponding eigenvalues. From the orthogonality of the columns $\underline{U}_i$ and $\underline{V}_i$ of **U** and **V** we can rewrite the expansion as [66]:

$$\mathbf{H} = \sum_{i=1}^{R} \lambda_i^{1/2}\,\underline{U}_i\,\underline{V}_i^T \tag{4.11}$$

$$\underline{B} = \sum_{i=1}^{R} \lambda_i^{1/2}\,\underline{U}_i\,\underline{V}_i^T\,\underline{Q}' + \underline{v}$$

where R is the rank of **H**. We can form the pseudoinverse $\mathbf{H}^\dagger$, of **H**, by inverting each nonzero λ_i and write the least squares minimum norm solution of (4.9) as

$$\underline{Q}^\dagger = \sum_{i=1}^{R} \frac{1}{\lambda_i^{1/2}}\,\underline{V}_i\,\underline{U}_i{}^T\underline{B} \tag{4.12}$$

In (4.11) the component of $\underline{B}$ due to the projection of $\underline{Q}'$ through $\underline{U}_i\underline{V}_i^T$ is weighted by $\lambda_i^{1/2}$ thus if $\lambda_i^{1/2}$ is small the resulting contribution to the data is small. In the pseudo inverse however, $\frac{1}{\lambda_i^{1/2}}$ will be large and hence the component of the noise vector projected

through $\underline{V}_i \underline{U}_i^T$ will be disproportionately amplified. To avoid this problem it is common to truncate the summation in (4.12) to sum over $P \leq R$ values so as to reduce the error in $\underline{Q}^\dagger$ due to measurement noise. This truncation however produces another type of error in the solution image by reducing the resolution and fine detail available from the small singular values. An optimal truncated pseudoinverse solution is obtained when P is chosen to minimize the sum of noise error plus resolution error [66]. This optimal truncation index, under the assumption that $\underline{Q}$ and $\underline{v}$ are independent, white Gaussian vectors, is given by:

$$P_{opt} = \max_i \left[i \;\middle|\; \lambda_i \geq \frac{E\{\underline{v}^T \underline{v}\}}{E\{\underline{B}^T \underline{B}\}} \right] \tag{4.13}$$

with the λ_i's ordered in a descending manner. The number of terms P used in the summation of eqn (4.12) determines the possible dimension of the solution. As P is increased towards R, the dimension of the solution, and hence potential resolution, is increased, but at the cost of increased sensitivity to noise.

This development of an optimum truncated pseudoinverse solution suggests that much can be learned about the stability, and attainable resolution and dimensionality of an image solution by analyzing the singular values of the system matrix **H**. The **H** obtained from several configurations of reconstruction volume and sampling surfaces has been analyzed. Figure 4.6b is a plot of the square of the ordered singular values for **H** obtained from a sampling surface and a reconstruction space as shown in Figure 4.6a. It can be seen that λ_i drops off rapidly, that **H** is not of full rank, and that 90% of the "energy" is contained in the first 10 values. Since the number of independent features in the source which may be recovered in the reconstruction is limited by the number of significant singular values, we should not expect successful reconstructions of an arbitrarily

dense distribution. If however we assume a-priori knowledge that the source distribution is maximally sparse, with fewer sources than large singular values, then full resolution reconstruction may be possible. Figures 4.6c and 4.6d show similar results for sample points on concentric nested hemispheres and sampling on a more widely separated grid, respectively, indicating little additional information is gained by additional external samples.

4.4.2. SQUID Resolution

Usable resolution of the SQUID biomagnetometer is limited by pickup coil size, gradiometer sensitivity as a function of distance, and data acquisition time. Pickup coil diameter and the gradiometer coil configuration affect resolution and sensitivity as a function of range. Figure 4.7 shows biomagnetometer response to an isolated dipole as a function of lateral distance and axial range, as compared to the ideal magnetic point detector. This non-ideal response characteristic must be accounted for in the system matrix **H** by modifying the expression for the submatrixes **W.**

A usable approximation is obtained if coil diameter is neglected and **W** is formed as a sum of matrices corresponding to the different coil positions in the gradiometer. In our device there are three coils with axial spacing between them of 50 mm, the center one is counter-wound and contains twice as many turns as the other two. With this configuration, **W** is computed as follows:

$$\mathbf{W} = \mathbf{W}^1 - 2\mathbf{W}^2 + \mathbf{W}^3 \qquad (4.14)$$

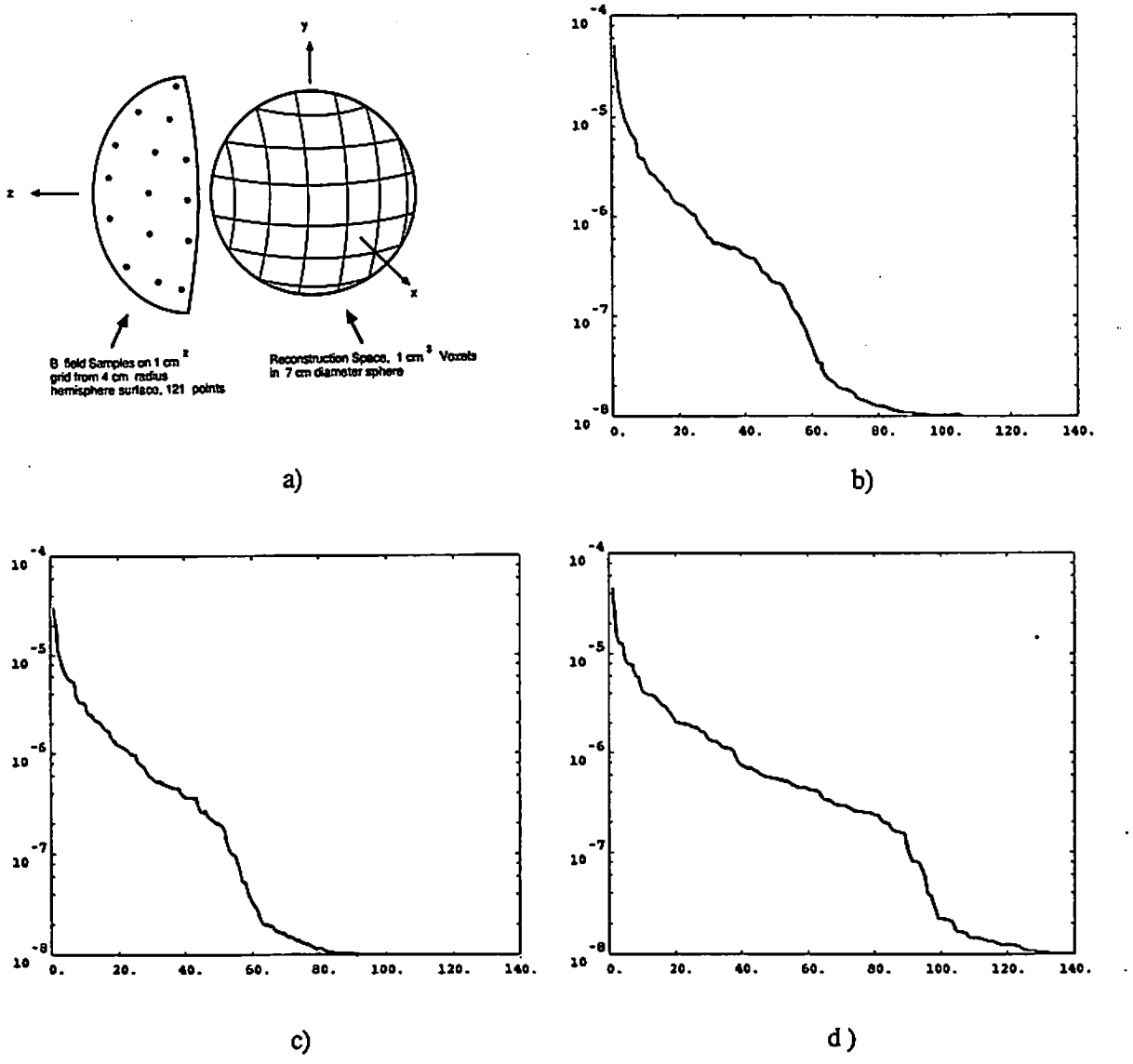


Figure 4.6. Singular Value Analysis for NMI Hemispherical Sampling. a) Geometry of reconstruction space and sample surface used for SVD analysis of the system Matrix. b) Plot of squared singular values for geometry of 4.5a. c) Squared singular values for double layer sampling. d) Squared singular values for reduced resolution ($2cm^3$ voxels) in reconstruction space.

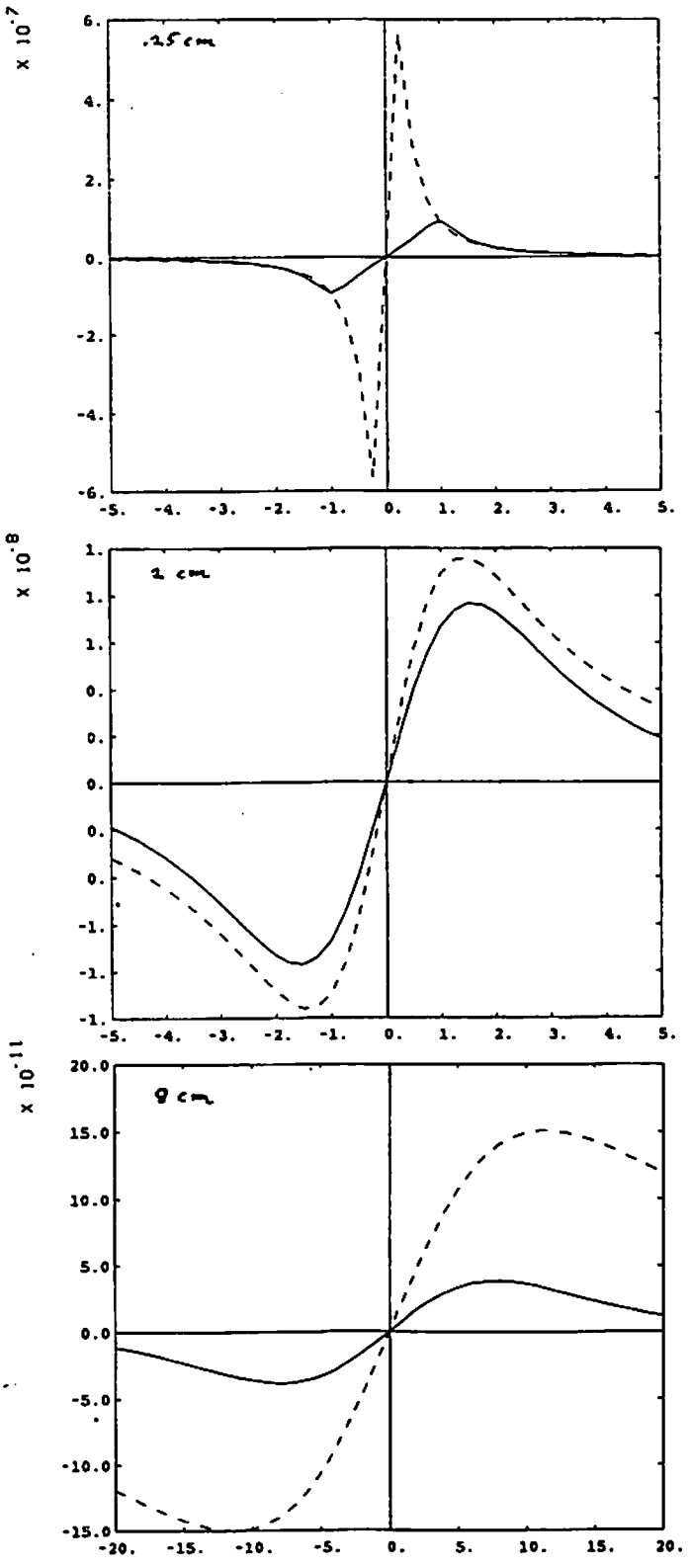


Figure 4.7. Gradiometer Response and Flux Density vs. Lateral Position for a Single Current Dipole at an Axial Depth of: a) .25*cm*, b) 2.0*cm*, c) 16.0*cm*.

where $\mathbf{W}^1$, $\mathbf{W}^2$, and $\mathbf{W}^3$ are matrices computed separately as in eqn (4.5) for the three coils, but using the centers of the three coils respectively when computing the position vectors $\underline{r}_m$. A more accurate expression for $\mathbf{W}^i$, i=1,2,3 can be found by averaging each submatrix found in (4.5) over the cross section area of each coil, i.e.

$$\mathbf{W}^i_{m,n} = \frac{4}{\pi\, \mathrm{d}^2} \int_{\phi^i_m} \mathbf{W}_{m,n}\, d\underline{r}_m \tag{4.15}$$

Where $\mathbf{W}^i_{m,n}$ is the m,n th submatrix in $\mathbf{W}^i$, ϕ^i_m is the integration surface, corresponds to the disk enclosed by the i th pickup coil for sample position $\underline{r}_m$, and d is the diameter of the coils.

The logistics of data acquisition also impose a practical limit on the number of data points which can be gathered. Functional neuromagnetic fields of the brain are weak enough that it is essential to perform time averaging between successive sets of data for noise suppression. For measurements from evoked responses, ten or more sample windows, each approximately one second long and synchronized to the patient stimulus, are needed at each sample point. With a single channel SQUID, this acquisition time and the time required to reposition the SQUID detector imply that data from the small 17 by 17 cell grid used in some of the experiments could take several hours to gather. The seven channel array devices are an improvement, but we look to future large array SQUID detectors to eliminate this restriction on resolution.

4.4.3. Noise and Background Magnetic Fields

The functional neuromagnetic fields of the brain which are of interest in neuromagnetic imaging are on the low end of the detectable scale. Somatically, visually, and auditory evoked fields have been measured at levels near .1 pT [43], and as shown by Figure 4.3 this is orders of magnitude below some other nearby biomagnetic fields. The earth's steady state magnetic field has an amplitude at 40 degrees latitude of approximately 50 μT and interference from commercial power and nearby feromagnetic objects can cause serious noise and interference problems. The second-order gradiometer configuration of the pickup coils can attenuate the zero and first order gradients of these fields to acceptable levels, but care must still be taken to reduce local interference. This background noise level will set a limit on the system sensitivity, and thus on the achievable resolution when dealing with the weak signals of interest.

4.5. Poor Conventional Reconstruction Results

A major motivation for considering the maximally sparse minimum dipole approach for NMI was the unsatisfactory results obtained using more convectional optimization criteria which permit solutions with nonzero components in every element. There are several well known algorithms used in image reconstruction problems where solutions to ill posed underdetermined linear systems are required [63]. These algorithms use a cost criterion to select a single solution from the infinite possible solutions to the underdetermined system where $M < N$. Given the data vector $\underline{B}$ and assuming there is no noise, then a single solution can be obtained by solving the general constrained optimization problem

$$\min_{\underline{Q} \in R^N} \quad [g(\underline{Q}) \ \textit{subject to} \ \ \underline{B} = \mathbf{H}\underline{Q} \ , \ \underline{Q} \geq 0 \ ] \tag{4.16}$$

where $g(\underline{Q})$ denotes some functional on the solution vector. The choice of the cost function will determine the class of solutions. Two different cost functionals, associated algorithms, and analysis of simulated experimental reconstruction results will be discussed in the following sections.

4.5.1. Minimum Norm, Additive ART Algorithm

A common optimization cost function is the minimum norm solution, with $g(\underline{Q}) = \underline{Q}^T \underline{Q}$, which can be found by numerous pseudoinversion and quadratic optimization techniques. An additive version of the Algebraic Reconstruction Technique (ART), which is an iterative algorithm popular in medical imaging, was used to evaluate NMI performance. The algorithm steps are as follows:

let $\underline{H_i}^T$ be the i th row of $\mathbf{H}$, then for iteration step k:

$$\underline{Q}^{k+1} = \underline{Q}^k + \frac{B_i - \underline{H_i}^T \underline{Q}^k}{\underline{H_i}\ \underline{H_i}^T} \underline{H_i} \quad for\ k = 0,1,\cdots k_e \ , \ i = k_{modM} + 1 \tag{4.17}$$

$$e_k = \underline{B} - \mathbf{H}\ \underline{Q}^k$$

Iterations terminate at k_e when the error e_k drops below a predetermined limit. The constrained ART algorithm was also evaluated for the case of solving eqn (4.9) with prior dipole orientation information.

The minimum norm approach favors smooth solutions, and it tends to force the solution dipoles as close as possible to the SQUID detectors since the field falls off as the inverse square of the distance between the detector and source. $g(\underline{Q})$ is minimized when smaller

current dipole magnitudes are located near the detectors to yield equivalent magnetic field measurements. The bias toward a solution near the detector and the underdetermined nature of the system make depth resolution difficult or impossible with a straight-forward minimum norm solution.

4.5.2. Maximum Entropy Solution

Another technique is to produce the maximum entropy solution where we minimize the functional

$$g(\underline{Q}) = \sum_{i=1}^{N} \frac{Q_i}{\|\underline{Q}\|_{l1}} \ln \frac{Q_i}{\|\underline{Q}\|_{l1}}, \quad Q_i \geq 0, \quad \|\underline{Q}\|_{l1} = \sum_{i=1}^{N} Q_i \tag{4.18}$$

This technique is favored by many researchers [67,68] as it yields the maximally uniform image consistent with the data.

There are a number of algorithms for maximum entropy restoration. Most require the elements of **H** to be non-negative [63,69], which does not meet our needs. Non-normalized algorithms are unacceptable due to the nonlinear effect on the solution caused by scaling the data. An algorithm based on techniques described in [69] was used for a normalized maximum entropy reconstruction with a general system matrix **H**. The iterative steps are as follows:

$$\underline{Q} = [\ 1.0,\ 1.0, \cdots]^T \tag{4.19}$$

$$Q_j^{k+1} = Q_j^k exp\ [\omega(\mathbf{H}^T \underline{\lambda}^k)_j]$$

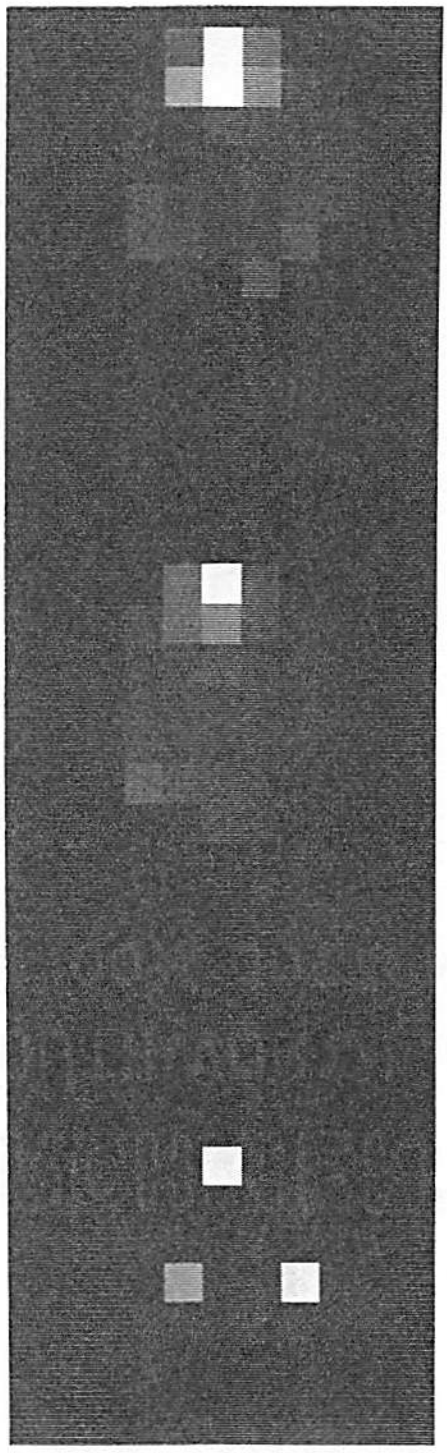
$$\underline{\lambda}^k = \frac{(\underline{B} - \underline{H}\ \underline{Q}^k)}{\|\underline{Q}\|_{l1}} \qquad for\ k = 0,\ 1, \cdots k_e\ ,\ i = k_{modM} + 1$$

ω is a relaxation constant to control convergence. As above, iterations are terminated when e_k drops below an acceptable limit. The entropy expression is defined only for $Q_j^k \geq 0$, so we are required to use our dipole orientation constraint model of eqn (*8) so that Q becomes a vector of non-negative magnitudes. This maximum entropy approach has shown in our simulation experiments to suffer from the same bias toward near detector solutions as the minimum norm image.

4.5.3. Simulation Results

Simulated noiseless magnetic field point measurements were used in a comparison of the algorithms discussed above. Three disjoint current dipoles of differing magnitude within a sphere were modeled, and noiseless measurements on a hemispherical surface surrounding them were computed. The sphere was 3 cm in radius centered at (0, 0, 0) and divided into 1 cm^3 voxels. The simulated magnetic samples were taken on a hemisphere of radius 4 cm with $z \geq 0$. In reconstruction, all dipoles were constrained to be in the $+x$ direction. The three dipoles were located at (x,y,z) coordinates (1, -1, 1), (1, -1, -2), and (1, 2, 0) with magnitudes 1.0, 1.5, and 2.0 respectively. This set of sources was chosen to demonstrate problems associated with the minimum norm and maximum entropy solutions, i.e. their inability to resolve depth from the measurement surface.

A single plane (x=1) of the original source and reconstructed 3-D images from the ART and maximum entropy algorithms are shown in Figure 4.8. The magnetic sample hemisphere surface surrounds the left half of the images. Since all three sources are located in the x=1 plane, this allows a comparison of the depth resolution for each algorithm. It can be seen that the minimum norm and the maximum entropy images shift energy toward the



a)

b)

c)

Figure 4.8. Noiseless Image Reconstruction for a 3 Dipole Source. x=1 plane shown. Intensity is $+x$ component of Dipole Field. a) ART (minimum norm) solution. b) Maximum entropy solution. c) Original source distribution.

sphere surface and blur the dipole locations. The maximum entropy entirely misses the deeper lying dipoles. The ART image produced a smoothed cluster of intensity near the (1, 2, 0) dipole, but the maximum voxel, not shown, is at (2, 2, 0) which is near the sphere edge. The maximum entropy criterion produced a slightly less disperse solution, but still had a smoothed cluster around (1, 3, 0). It did however produce a more uniform field of near zero values away from this cluster than did the ART algorithm.

Figure 4.9 shows results of reconstructing a larger more complex source distribution. The image sphere is 13 cm in diameter and the measurement hemisphere is in the $z > 0$ half-plane with radius 8 cm. Figure 4.9a and shows planar slices through the original source while 4.9b and 4.9c give the corresponding minimum norm ART and Max Entropy reconstructions. The source contains a 2 by 2 by 20 cell bar of current running diagonally through the sphere and a 7 cell diameter solid disk of current lying in the $z = 2$ plane. Note the complete loss of internal detail in the reconstructions.

The fact that these results differ so dramatically from the true source confirms the need to select the model and algorithm best suited to our current knowledge of the physical processes involved in neural activity. It is clear from the above examples that the minimum norm or maximum entropy methods result in unacceptable solutions. This conclusion motivates our effort to identify a new class of solutions which will more closely represent an underlying current.

a)

b)

c)

Figure 4.9. Reconstruction of Bar and Disk Source in a 13cm Diameter Sphere. a) x=1 (left) and z=2 (right) planes of the source distribution. b) ART reconstruction. c) Maximum entropy reconstruction.

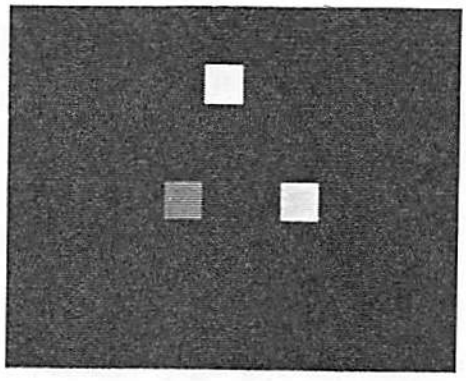
4.6. Minimum Dipole, Maximally Sparse Solutions

With the intent of overcoming the errors in the reconstructed images described above, a new alternative class of solutions was proposed which attempts to minimize the number of dipoles and expresses the problem as an $l_{1/q}$ optimization.

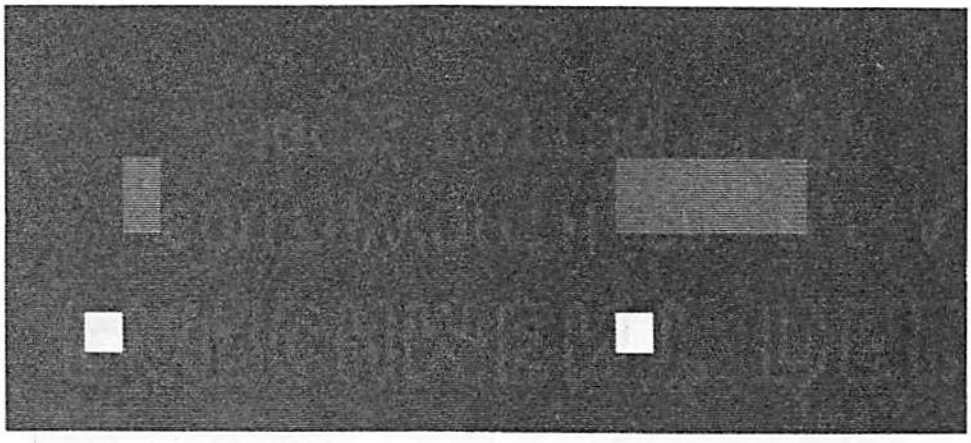
This formulation is appealing since choosing the minimum number of dipoles is maximally noncommittal in the sense that unless we have good reason to believe otherwise, the minimum order solution is the least activity which could have given rise to the data. For the NMI problem, it is proposed that the number of dipoles required to create the magnetic field be the measure of source complexity we wish to minimize. It is hypothesized that this minimum dipole solution is most likely to correspond to the physical source distributions we will study. If the minimum dipole representation is a reasonable model of the actual current distribution, then this approach also enables resolving current dipoles in depth by removing the underdetermination in the system.

Figure 4.10 shows results of reconstruction of simulated NMI measurement data with dipoles constrained to be oriented in the positive x direction. The image in Figure 4.10a is an $l_{1/q}$ search reconstruction of the source used in Figure 4.8. Note that it exactly locates and scales the three dipoles and that it did so in an elapsed time of about 20 minutes as compared to 3 days for the exhaustive search. Figure 4.10b and c show the x and z axis projections respectively of the reconstruction form another source involving 11 dipoles. This image used the same spherical reconstruction space and hemispherical sample surface centered on the positive z axis as in Figures 4.8 and 4.10a. The solution image shown exactly matches the original source to within 5 significant digits.

Figure 4.11a shows the x and y axis projections of a 20 dipole source. Figure 4.11b shows the corresponding planes of the $l_{1/q}$ SA solution for 5.11a, in which a line of five dipoles was replaced with a single dipole at (1,0,-2), producing a 16 dipole, lower order equivalent of the original. Note that the replacement dipole is located precisely in the center of the original current line. Although this solution differs from the original, it maintains all the major features and properly locates the centers of activity. Since the induced magnetic fields of Figure 4.11a and 4.11b are identical, there is no justification without prior knowledge to presume more dipole sites are involved than in the least order solution of 4.11b. The system matrix used for each image in Figures 14 and 15 involved 41 measurements and 125 dipole voxels and the algorithm execution time was approximately the same for each, while an image like 4.11b would be unattainable with the exhaustive search because of the number of iterations being proportional to $N!$. Though the exhaustive search could find Figure 4.10a, it could take ten orders of magnitude as many iterations to find 4.11b.

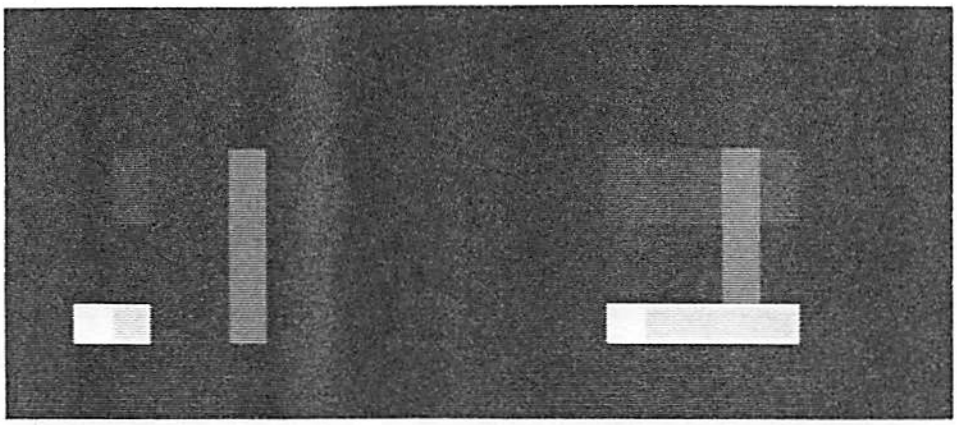


a)

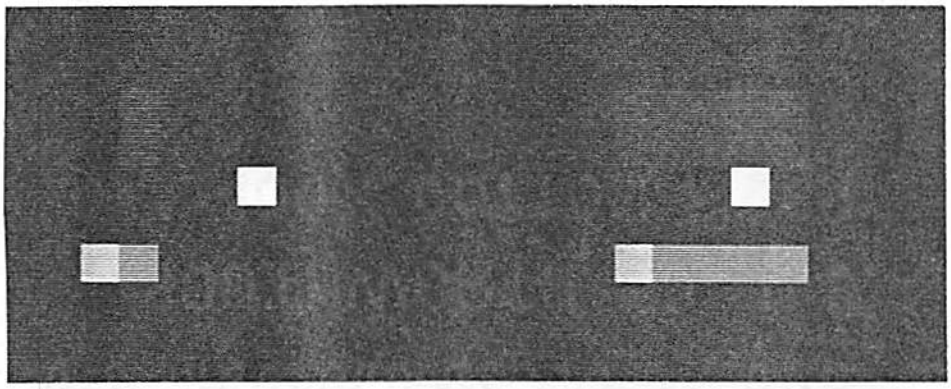


b)

Figure 4.10. Exact $l_{1/q}$ Simplex Search Algorithm Solutions.
a) 3 dipole source of Figure 4.8. b) An 11 dipole source, projection along x axis (left) and projection along z axis (right).



a)



b)

Figure 4.11. $l_{1/q}$ Simplex Search Reconstruction of a 20 Dipole Source. a) x axis (left) and y axis (right) projections of the original source. b) Reconstruction results. Note solution is lower order than source.

CHAPTER 5: SPARSE OPTIMIZATION FOR SEISMIC DECONVOLUTION

5.1. Seismic Deconvolution

Seismic deconvolution is a widely used reconstruction technique for recovering subsurface geological strata location information from reflected acoustic signals. This sounding method, known as reflection seismology, is most often used to aid in oil exploration by generating maps, or image-like reconstructions, of the deduced strata positions over a large enough region to permit identification of geologic structures likely to contain oil deposits. The techniques are also applied to well logging [8] and marine seismic exploration [70]. A typical field instrumentation arrangement is shown schematically in Figure 5.1. A large acoustic vibrational source (often explosive) is activated at the surface to generate a wave which propagates down through the strata, and the acoustic impedance discontinuities at the layer boundaries cause reflections which are in turn received by the array of geophones placed along the surface. By using a linear array of phones, multiple reflection ray paths can be detected and a 2-D "slice" of Earth covered by these rays can be mapped.

A simple discrete model of this process is used as the basis for many digital seismic deconvolution algorithms [70]. The following assumptions are made to support the model: a) the reflection process is approximately linear so that superposition holds, b) the reflection angle is approximately normal, c) negligible ray bending occurs, d) sound propaga-

tion speed is constant, and e) the reflection sites are at discrete locations rather than smoothly distributed. Experimental data indicates these are reasonably accurate assump

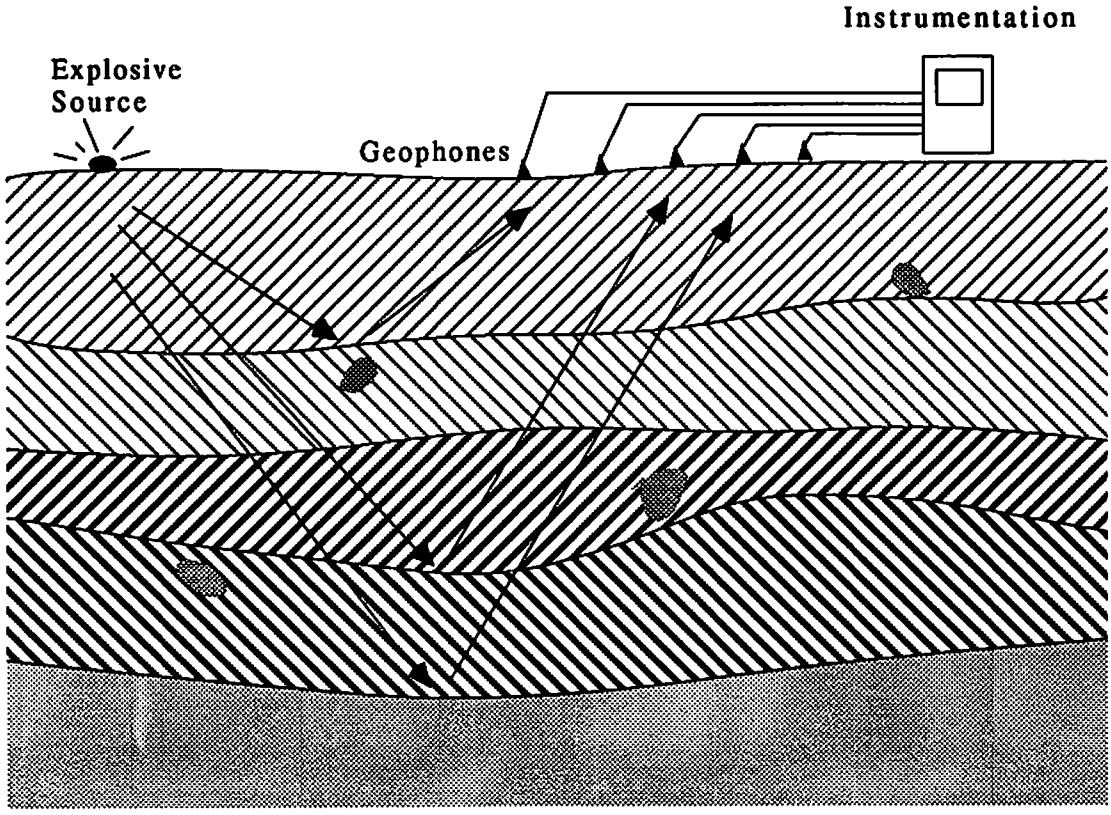


Figure 5.1. Basic Equipment Configuration for Reflection Seismography

tions, so we may represent the signal received by each phone as:

$$z(k) = \sum_{j=1}^{k} \mu(j)V(k\text{-}j) + n(k) \qquad (5.1)$$

where z is the observed signal, μ is the reflectivity sequence representing the discrete impulse response from the Earth, V is the acoustic source signal, and n is the measurement noise which incorporates the effects of sensor noise, distributed backscatter, and model error. The index, j, has units of time corresponding to propagation distance to a reflection site. V usually has much longer duration than the separation between nonzero $\mu(j)$ samples, so the convolutional from of eqn (5.1) causes corresponding components in $z(k)$ to overlap and obscure reflection site locations. A deconvolution procedure is thus needed to recover μ from z, and must also deal with the corrupting effect of the noise term n. The class of seismic deconvolution problem is determined by how many of the terms and parameters associated with eqn. (5.1) are known a-priori. If all parameters, including V, must be estimated from the received data, z, we have a problem of "blind deconvolution" [8,9,10,70]. Often though, it is possible to obtain independent, reasonably uncorrupted, measurements or estimates of the source wavelet and of the noise statistics so that the deconvolution algorithm need only estimate μ. The techniques described in sections 5.2 and 5.3 assume that an accurate model of the source wavelet and the second moment of the noise signal are known.

Our interest in seismic deconvolution lies in the typically sparse nature of recovered sequences $\mu(j)$, which makes it a likely candidate for maximally sparse optimization. A simple view of the Earth's structure, consisting of layers or bands of homogeneous material separated by abrupt boundaries where reflections occurs, suggests a heuristic expectation that reflectivity sequences will contain large impulses corresponding to these boundaries, separated by many near zero samples. Several authors [8,9,10] have confirmed this sparse characteristic in real data, (see section 2.5) and have used the generalized p-Gaussian (gpG) distribution with p in the range of .4 to 1.5 to model the reflectiv-

ity sequence. $\mu(j)$ is assumed to contain independent, white samples with gpG zero mean distribution, and n is modelled as either white Gaussian or gpG.

$$f_\mu(\mu(j)) = \frac{p}{2\Gamma(1/p)\,\gamma\,\sigma}\, e^{-\left(\frac{|\mu(j)|}{\gamma\,\sigma}\right)^p} \tag{5.2}$$

$$\gamma = \left[\frac{\Gamma(1/p)}{\Gamma(3/p)}\right]^{1/2} \qquad E\{\mu(i)\mu(j)\} = \begin{cases} \sigma^2 & for\ \ i=j \\ 0 & for\ \ i\neq j \end{cases}$$

This model is particularly well suited to the $l_{1/q}$ norm based algorithms presented in this work due to their relationship to maximum likelihood estimation (see section 2.5). Mendel [70] used a different Bernoulli-Gaussian model to represent the sparse, impulsive data. The Bernoulli component expresses the (small) probability of a reflection boundary occurring at each of the discrete depth (time) samples, while the Gaussian component controls the amplitude of the reflection. Let $b(j)$ be a sequence of independent, Bernoulli distributed samples, and $r(j)$ be an equal length sequence of independent zero mean Gaussian variates. $\mu(j)$ can then be simply represented as:

$$\begin{aligned} &\mu(j) = r(j)\, b(j), \qquad 1\le j\le N \\ &f_b(b) = (1-\lambda)\,\delta(b) + \lambda\delta(b-1) \\ &f_r(r) = N(0,\sigma), \qquad yielding:\ \ E\{\mu^2(j)\} = \sigma^2\lambda \end{aligned} \tag{5.3}$$

Typical values of λ=.05 and σ=.30 are given in [70]. Using this model a number of successful techniques were developed for both the blind deconvolution and known wavelet cases, including minimum variance, maximum likelihood, and maximum a-posteriori estimation methods [70]. Due to the independence of b and r, the problem is separable so the optimal solution when V is known is given by first computing the maximum likelihood estimates of the reflection locations, $b(j)$, and then their amplitudes, $r(j)$.

A number of heuristic "minimum entropy" optimization approaches have been proposed for blind seismic deconvolution. These problems are related to the approach of section 5.2, and some utilize the gpG distribution model for $\mu(j)$ [8,9,10]. These methods are minimum entropy only in the sense that the desired reconstruction is as "simple" as possible, where simple is defined as consisting "of a few large spikes of unknown sign or location separated by nearly zero terms" [71]. This is assumed to be a maximally structured result, and to be the opposite of maximum entropy solutions which are the most smooth or unstructured. Deconvolution involves solving for the inverse filter corresponding to the wavelet by minimizing a heuristic "norm" measure of sparseness. A variety of these "norms" have been used as objective functions, including the varimax norm ratio proposed by Wiggins [71]:

$$v(\underline{x}) = \sum_{j=1}^{m} \frac{\sum_{i=1}^{n} x_{ij}^{4}}{\left[\sum_{i=1}^{n} x_{ij}^{2}\right]^{2}} \tag{5.4}$$

with j indexing the geophone channels, and i the time samples. The parsimonious norm, a generalization proposed by Claerbout is:

$$v(\underline{x}) = \sum_{j=1}^{m} \frac{\sum_{i=1}^{n} |x_{ij}|^{\alpha}}{\left(\sum_{i=1}^{n} |x_{ij}|^{2}\right)^{\alpha/2}} \tag{5.5}$$

Gray used the variable norm ratio:

$$v(\underline{x}) = log \prod_{j=1}^{m} \frac{\left(\frac{1}{n}\sum_{i=1}^{n} |x_{ij}|^{\alpha_1}\right)^{n/\alpha_1}}{\left(\frac{1}{n}\sum_{i=1}^{n} |x_{ij}|^{\alpha_2}\right)^{n/\alpha_2}} \tag{5.6}$$

Each of these functionals is scale invariant and favors sparse solutions, but not in any optimal sense. While methods based on these norms as cost functions have met with some success, they are nonconvex, and since the methods are formulated as unconstrained optimization problems [9], the search is carried out over a continuous parameter space, converging to a local optimum. These methods do not adapt well to the case in which the source wavelet is known and the problem can be formulated in terms of a set of linear constraints, since local minima are not confined to the vertices of the convex polytope where the optimal solutions lie, as proved in section 2.2. It is proposed that the $l_{1/q}$ cost function is superior since it was shown in Theorem 2 (section 2.3) to yield maximally sparse solutions, and is optimum for maximum likelihood estimation when the gpG distribution model applies.

The generalized p-Gaussian model is used in section 5.2 below for synthesis of the data sequence and as the basis of the algorithm design, while in section 5.3 a Bernoulli-Gaussian model was used. The $l_{1/q}$ simplex search algorithm assumes no statistical model other than for the noise term. The reflectivity sequence is interpreted as deterministic and the maximally sparse configuration of elements consistent with the measurement data is sought. The results obtained by Mendel's algorithm using the Bernoulli-Gaussian model are presented in section 5.3 to compare with the deterministic results. All measurement data presented in the algorithm demonstrations was synthesized using computer models.

The transmit wavelet used in both sections following was synthesized using a representative fourth order ARMA model taken from [70]:

$$V(z) = \frac{-.76286+1.5884z^{-1}-.82356z^{-2}+0.0002224z^{-3}}{1 - 2.2633z^{-1}+1.77734z^{-2}-.49803z^{-3}+.045546z^{-4}} \quad (5.7)$$

A digital filter was implemented with these coefficients and driven with an impulse to generate the wavelet samples used below and plotted in figure 5.2.

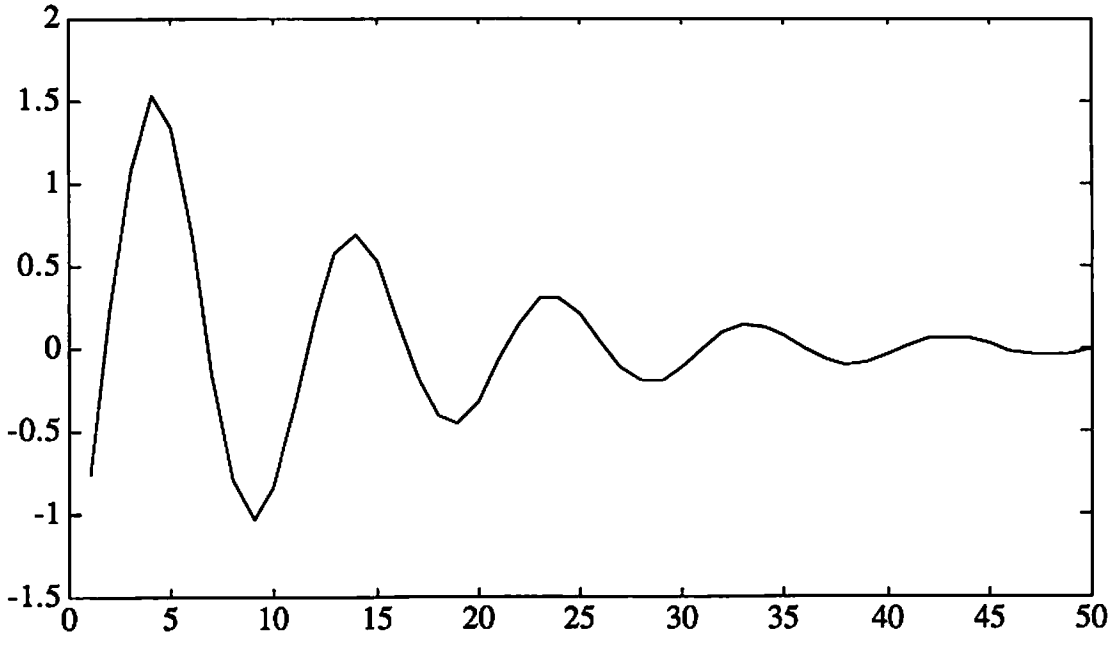


Figure 5.2. Fourth Order ARMA Wavelet Used in Seismic Simulations.

5.2. Convex Transform Gradient Search Solutions for Seismic Deconvolution

In this section, the convex transform gradient search algorithm described in section 3.3 is applied to seismic deconvolution using the reflectivity sequence model of eqn (5.2).

Based on the arguments of section 2.5.3, the algorithm is well suited for near optimal solution, in a maximum a-posteriori sense, when using this model, and is preferred over the simplex search approach. Various shape parameters, p, in the range of .2 to .5 were used for the gpG reflectivity sequence, and Gaussian measurement noise is assumed.

Using matrix notation, for a single geophone with M samples, and N depth sites we have:

$$\underline{z} = \mathbf{V}\underline{\mu} + \underline{n} \tag{5.8}$$

where $\mathbf{V}$ *is the* $M \times N$ *Toeplitz convolution matrix, i.e.*

$$v_{ij} = \begin{cases} V(i-j+1) & \text{for } N \geq i \geq j,\ j \leq M \\ 0 & \text{otherwise} \end{cases}$$

n_i *distributed iid* $N(0,\sigma_n)$

μ_i *distributed iid* $gpG(0,\sigma_\mu)$

We assume that V and σ_n are known. The maximum a-posteriori estimate of $\underline{\mu}$, as given in eqn (2.19) as:

$$\hat{\underline{\mu}}_{MAP} = \min_{\underline{\mu}} \ \frac{(\underline{z} - \mathbf{V}\underline{\mu})^T(\underline{z} - \mathbf{V}\underline{\mu})}{2\sigma_n^2} + \left(\frac{1}{\gamma\sigma_\mu}\right)^p \sum_i |\mu_i|^p \tag{5.9}$$

We cannot solve this directly, but we can guess at a reasonable lower bound on the first term since σ_n is known and the n_i are independent. We then concentrate on minimizing the second term. Letting $\varepsilon \approx M\sigma_n$, , we express (5.9) in a form suitable for the algorithm:

$$\min_{\underline{\mu}} \ \sum_i |\mu_i|^p \ \ s.t. \ \ (\underline{z} - \mathbf{V}\underline{\mu})^T (\underline{z} - \mathbf{V}\underline{\mu}) \leq \varepsilon,\ \mu_i \geq 0,\ p<1 \tag{5.10}$$

Letting $q = 1/p$, performing the convex transformation from $\underline{\mu}$ to $\underline{u}$, and doubling the number of variables allow for a bipolar representation the the transformed space, we have:

$$\hat{\underline{u}} = \inf_{\underline{u}} \sum_{i=1}^{N} e^{u_i} \quad s.t. \ (V e^{q\underline{u}} - \underline{z})^T (V e^{q\underline{u}} - \underline{z}) \le \varepsilon, \text{ and } u_i > -\infty \qquad (5.11)$$

where

$$V = [\, V \,|\, -V \,], \quad \underline{\mu}^{\dagger} = \begin{bmatrix} \underline{\mu}^{+} \\ \underline{\mu}^{-} \end{bmatrix}, \quad \mu_i = (\mu_i^{+} - \mu_i^{-}), \quad u_i = \frac{1}{q} ln\, (\mu_i^{\dagger})$$

This is solved using the Schittkowski or other nonlinear constrained optimization algorithm, after which the inverse transformation is performed.

Figure 5.3 shows the synthesized data used in this deconvolution example. The reflectivity sequence shown in Figure 5.3a has a shape parameter p=.2, and was produced by transforming computer generated uniform variates to gpG samples. For any random variable, x, with a continuous cumulative distribution function, $F_x(.)$, $y = F_x(\underline{x})$ is uniformly distributed, $U[0,1]$. Since $F_x(.)$ is monotonically increasing for a continuous distribution, the inverse function, $F^{-1}_x(.)$ always exists, and $z = F_x^{-1}(y)$, for y a uniform random variable, has the same distribution as x. If $F_x^{-1}(.)$ can be computed for the gpG density, we can synthesize gpG random samples, z_i by transforming uniform samples, y_i.

The gpG density function cannot be integrated in closed form, so $F_x(.)$ was obtained by numerically integrating gpG(x) on a dense sample grid. The inverse function was implemented using a table lookup and linear interpolation, which provided an efficient

generation of a gpG sample from each uniform sample for any desired value of p.

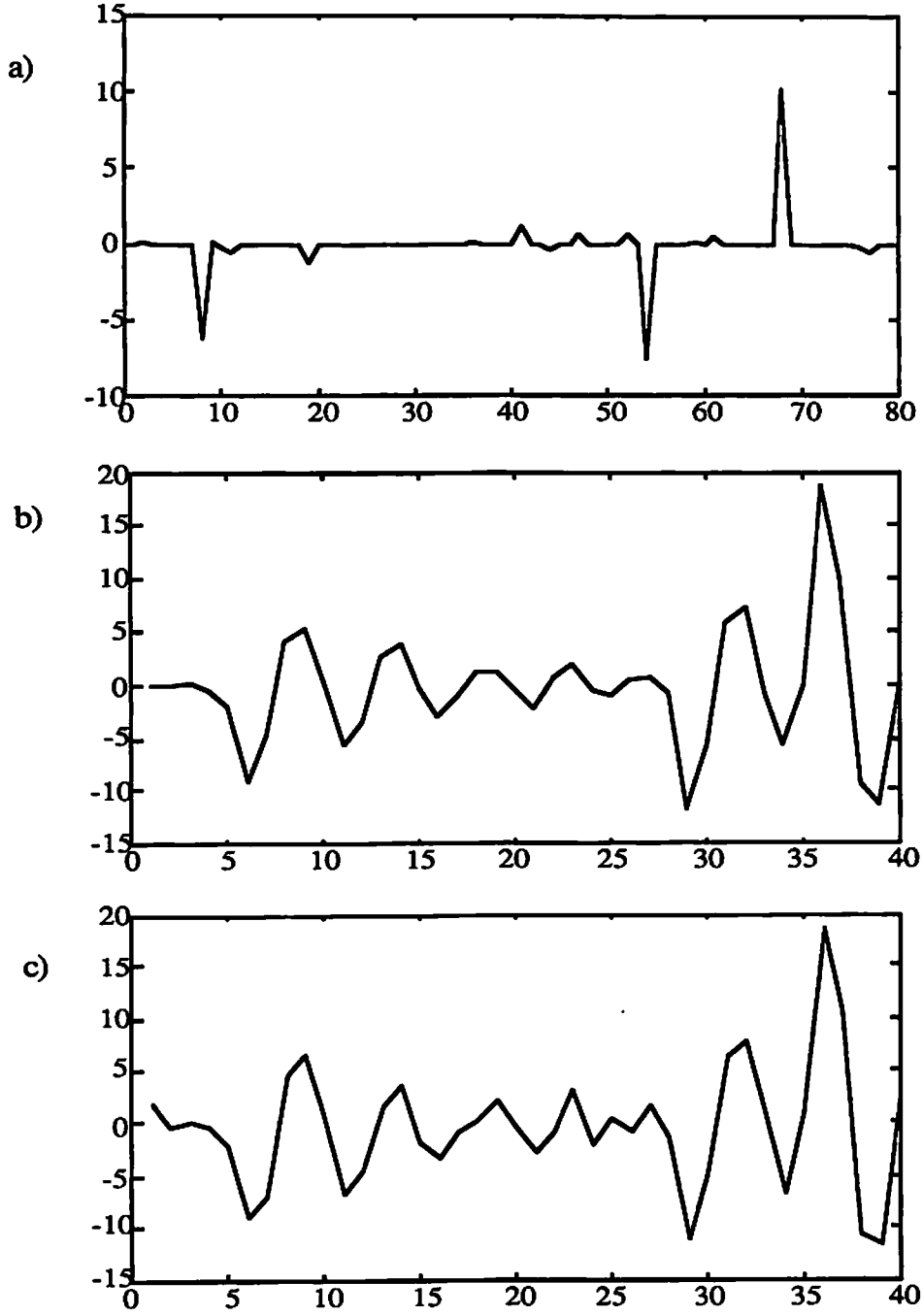


Figure 5.3. Simulated gpG Seismic Reflectivity Sequence and Received Data. a) Generalized p-Gaussian reflectivity sequence, p=.2, 80 samples. b) Noiseless received data after convolving with Figure 5.2, decimate by 2. c) Gaussian noise added, $\sigma = 1.0$.

Figure 5.3b shows the noiseless reflectivity sequence resulting from convolving 5.3a with the wavelet of Figure 5.2. Figure 5.3c is the same received data with Gaussian measurement noise added with σ_n=1.0.

The convex search algorithm has shown particular improvement over least squares and pseudoinverse solutions in underdetermined and extremely noisy cases. To simulate an underdetermined problem, after convolving with the ARMA wavelet, the received data sequences were decimated by a factor of two to generate Figures 5.3b and c. These last two sequences, consisting of 40 samples each, were the measurement data used to reconstruct the 80 sample reflectivity sequences of Figures 5.4 and 5.5. The Moore-Penrose pseudoinverse solutions for noiseless and σ_n=1.0 cases are shown respectively in Figures 5.4a and 5.4b, while Figures 5.5a and 5.5b give the convex transformation gradient search results .

It is notable that these solutions retain all of the major spikes of the original sequence, while forcing the majority of the other samples to zero. There is significant improvement over the pseudoinverse results in terms of fewer false detections, which is particularly useful for noisy data since there is no justification for retaining smaller valued reflection points when the noise level is high enough to have produced equally prominent spikes. Both Figure 5.4c and 5.4d required approximately 100 iterations of the Schittkowski algorithm and used a 40×160 system matrix V. In each of the examples, the pseudoinverse solution was used as the initial μ guess to start the Schittkowski algorithm. For the noisy case, an $\varepsilon = 35 \approx M\sigma_n$ was used, and convergence to this value was obtained, while for the noiseless case $\varepsilon = 10$ was used. Since prior knowledge of the value of p was not assumed, the maximally sparse approach was taken by using arbitrarily chosen smaller p values in the algorithm than were likely to be required by any typical gpG seis-

mic reflectivity data. For the noiseless data $p = .125$ ($q = 8$) was used, and $p = .143$ ($q = 7$) for the noisy data.

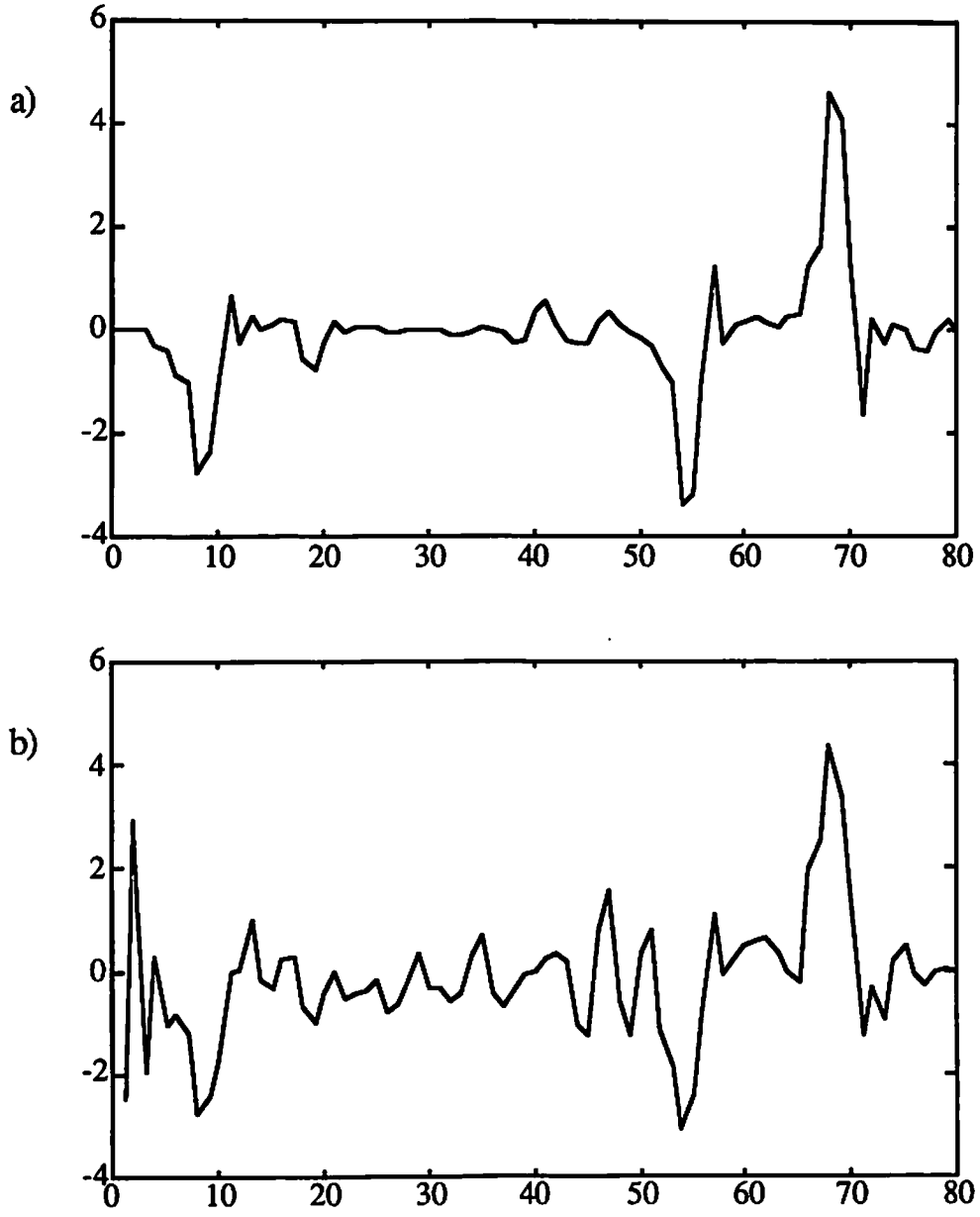


Figure 5.4. Pseudoinverse Deconvolutions of gpG Seismic Data. a) Moore-Penrose pseudoinverse result for deconvolution of noiseless sequence of Figure 5.3b, compare with Figure 5.3a. b) Result using noisy sequence of Figure 5.3c, compare with Figure 5.3a.

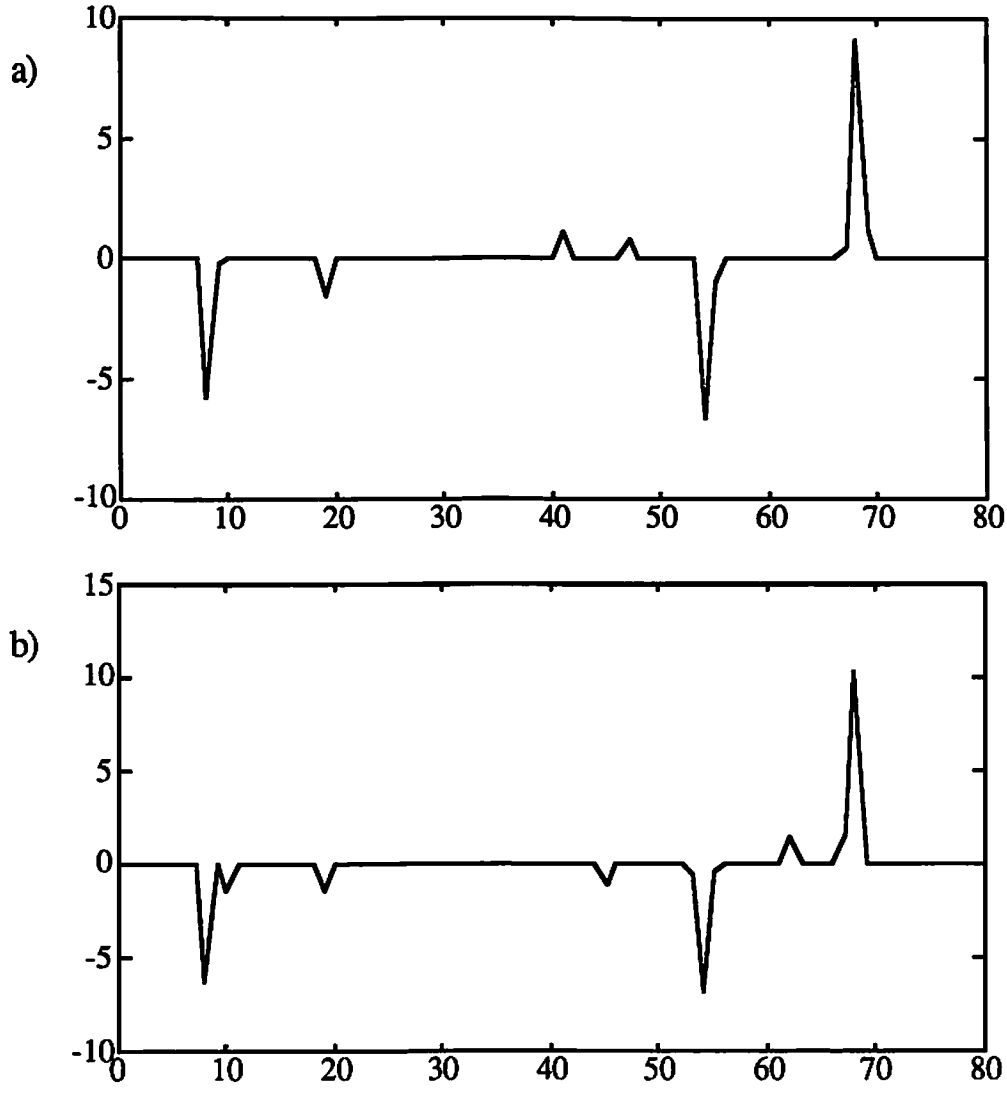


Figure 5.5. Convex Transformation Gradient Search Deconvolution of gpG Seismic Data.
a) Result for deconvolution of noiseless sequence of Figure 5.3b, compare with Figure 5.3a. b) Result using noisy sequence of Figure 5.3c, compare with Figure 5.3a.

5.3. Seismic Deconvolution using the $l_{1/q}$ Simplex Search

The following application of the $l_{1/q}$ simplex search algorithm to seismic deconvolution was conceived and implemented by Mr. Zhenyu Wu in collaboration with the author, and is reported in [2] and [13]. It is included here as another example of the excellent performance of the algorithm in sparse data environments. As before, the problem addressed here assumes that the source wavelet is known, and that a rough estimate of the combined backscatter and measurement noise level is available.

For the simplex search, the linear system formulation of equation (5.9) is shown in eqn (5.12). We cannot implement a constraint on the squared error term as was done in section 5.2, but this formulation differs from the other $l_{1/q}$ simplex search problems presented, because an l_1 norm constraint on the error is used rather than individual error bounds on each measurement, z_i.

$$V\underline{\mu} = \underline{z} \;:\; \begin{bmatrix} \mathbf{V} & -\mathbf{V} & \mathbf{I} & \mathbf{0} & 0 \\ \mathbf{V}^{\ddagger} & -\mathbf{V}^{\ddagger} & \mathbf{0} & -\mathbf{I}^{\ddagger} & 0 \\ \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{1} & 1 \end{bmatrix} \begin{bmatrix} \underline{\mu}^{+} \\ \underline{\mu}^{-} \\ \underline{s}^{+} \\ \underline{s}^{-} \\ s \end{bmatrix} = \begin{bmatrix} \underline{z}^{\ddagger} \\ \underline{z}^{\ddagger} \\ c \end{bmatrix} \tag{5.12}$$

Here $\underline{\mu} = \underline{\mu}^{+} - \underline{\mu}^{-}$, and ‡ indicates that some rows are negated to force all right-hand-sides to be non-negative. The last row of the system matrix, V, places a constraint on the sum of the plus and minus slack variables, i.e. $\sum_{i=1}^{M} |\varepsilon_i| = \sum_{i=1}^{M} (s_i^{+} + s_i^{-}) \le c$. This is equivalent to the problem

$$\min_{\underline{\mu}} \sum_{i=1}^{N} |\mu_i|^{1/q} \quad such\ that \quad \| \mathbf{V}\underline{\mu} - \underline{z} \|_{l_1} \le c\,,\ q > 1 \tag{5.13}$$

A suitable choice of c could be found from an estimate of the variance σ_n.

The seismic signal was simulated by convolving the reflectivity sequence in Figure 5.6a, generated using a Bernoulli Gaussian model (l=1/16, σ^2 =1), with the ARMA wavelet of Figure 5.2. The resulting signal, corrupted with i.i.d. Gaussian noise at an SNR of 10 dB, is shown in Figure 5.6b. A second set of data, shown in Figure 5.6c, was generated to incorporate convolutional backscatter [70]. In this case an i.i.d Gaussian sequence, of variance σ_b^2 =0.01, was added to the reflectivity sequence prior to convolution with the wavelet and i.i.d Gaussian noise was added to the resulting data at an SNR of 10dB. It is noteworthy that a backscatter component is implicit in the gpG model used in section 5.2 due to the numerous small valued samples between isolated large spikes.

The results of deconvolution are shown for Mendel's optimal seismic deconvolution method (OSD) are shown in Figures 5.6d and 5.6e for the with and without backscatter cases respectively. Event detection was performed using the 'single most likely replacement' detector [70]. The $l_{1/q}$ simplex search results are shown in Figures 5.6f and 5.6g. The square boxes in the graphs show the locations of the detected events, the solid lines the locations of the actual events. In the backscatter case, we are interested only in the larger events, the smaller 'events' being due to backscatter from small scatterers and not of primary interest in this problem. When applying the OSD method we assumed only knowledge of the wavelet parameters; all variances and the Bernoulli parameter were estimated from the data.

It is interesting to note that there is very little difference between the results obtained from the two approaches, even though the minimum variance deconvolution was based on the exact statistical model by which the data was generated, while the new method uses only the $1/q$ cost function and a rough approximation of the expected l_1 error. In the case for

data plus noise the OSD method produces a spurious event at about the 80th sample point (Figure 5.6c) which is not present in the $l_{1/q}$ simplex solution. In the case with backscatter, the OSD performs slightly better and detects only 'true' events while the new method also detects several of the larger 'backscatter events'. In both cases however, the number of events detected may be altered by modifying either the level of the threshold in the event detection in OSD [70] or the upper bound, c, on the l_1 norm of the residual error.

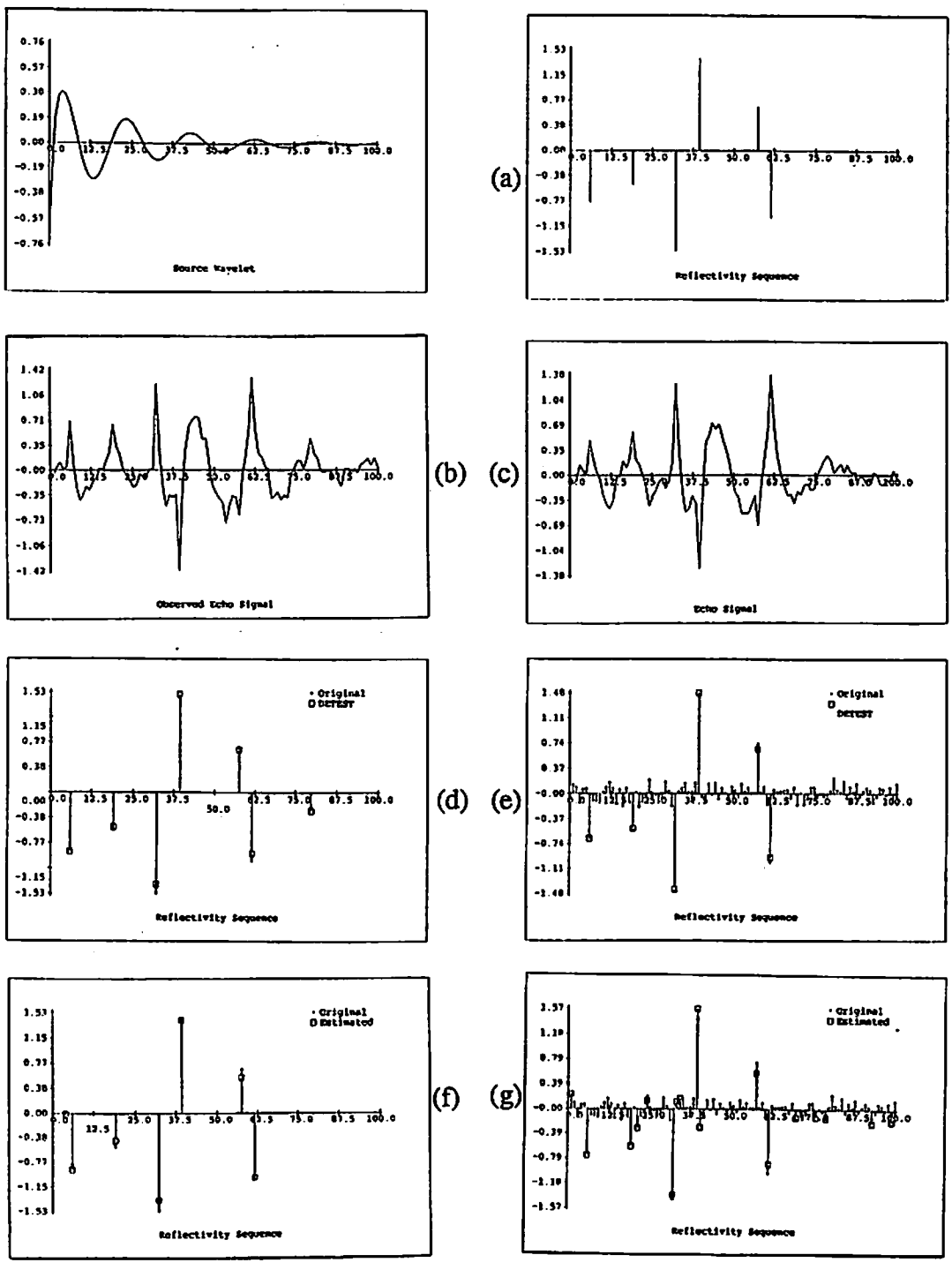


Figure 5.6. Seismic Deconvolution of Bernoulli-Gaussian Data. a) Reflectivity Sequence. b) Received data, 10 dB SNR. c) Received data with backscatter. d) OSD deconvolution of b. e) OSD deconvolution of c. f) $l_{1/q}$ Simplex deconvolution of b. g) $l_{1/q}$ Simplex deconvolution of c.

CHAPTER 6: SPARSE ARBITRARY BEAMFORMINNG ARRAY DESIGN

In this chapter the problems of array element shading and placement are considered for arbitrarily shaped symmetric 3-D arrays in narrow-band phased beamformer operation. The goal is to achieve "thinned," data independent (i.e. not statistically optimum) array designs which have spatial response patterns comparable to more conventionally designed arrays which use more elements. The "least elements" optimality criterion allows us to reduce the beamforming processing load by identifying the unnecessary phones for a given beam; a task which is difficult using other design approaches. Also, if we use a fine grid of potential element locations, then the minimum order solution can be used for optimal element placement analysis.

6.1. Beamforming Fundamentals

Array beamforming may be viewed as spatial-temporal filtering designed to extract a desired signal in a specific direction from other interfering signals and noise in a three dimensional propagation medium. Classical beamforming designs used in communications, RADAR, SONAR, and other acoustic and electromagnetic applications typically utilize a linear or planar array of sensor elements to form a strong response from the direction of interest, while rejecting signals from other directions. However, it is not necessary to maintain linear configurations, or to require uniform spacing since, as pointed out by Van Veen and Buckley, "there is no compelling reason to space sensors regularly.

Sensor locations provide additional degrees of freedom in designing a desired response . . . " [78]. Utilizing these degrees of freedom can be very difficult, due to the multidimensional nature of spatial sampling and the complex relationship between desired response characteristics and array geometry. The maximally sparse optimization technique will provide a methodology for addressing this problem.

We shall consider design of arbitrarily shaped narrow-band beamforming arrays with the basic architecture shown in Figure 6.1, but which contain the least possible number of elements for a desired spatial response.

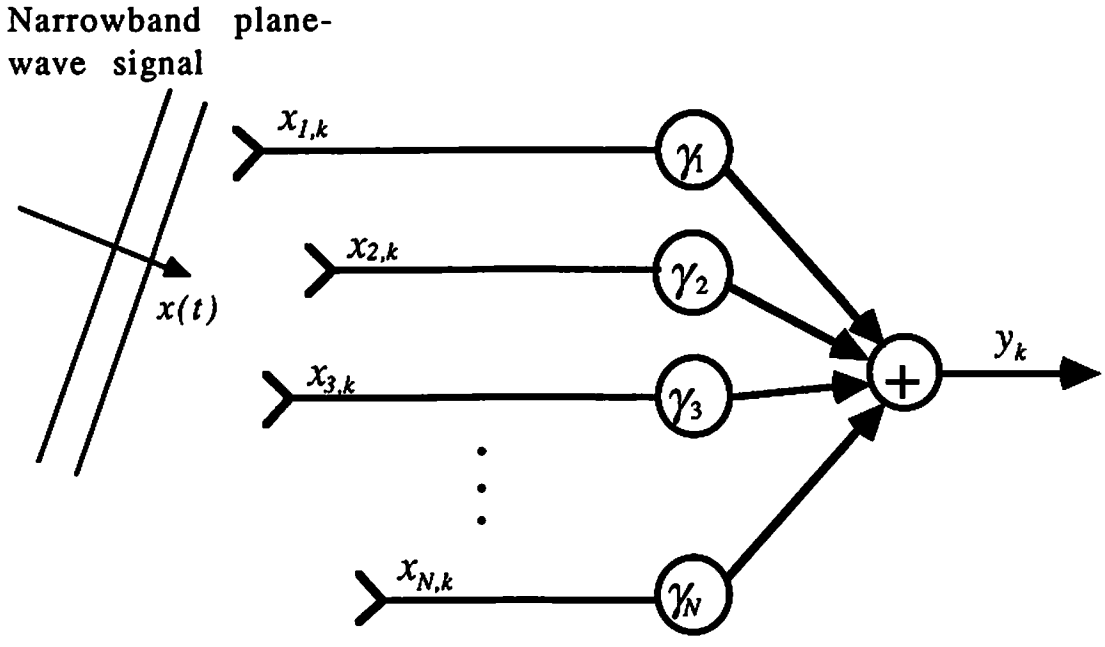


Figure 6.1. Narrowband Arbitrary Beamformer Architecture.

The processing in Figure 6.1 may be expressed as:

$$y_k = \underline{\gamma}^T \underline{x}_k \tag{6.1}$$

where $\underline{x}_k$ is the vector of array data samples at time index k, $\underline{\gamma}$ the vector of complex element beamforming coefficients, and y_k the kth time sample at the beamformer output. This structure is best suited to cases where the signal of interest is narrowband with a known center frequency, ω, since only at the design frequency do the phase shifts in $\underline{\gamma}$ correspond to element phase shifts due to differing plane wave propagation distances across the array. The spatial magnitude response, $R(s,\omega)$, at frequency ω, and any azimuth and elevation specified by a unit direction vector, s, is given by

$$R(s,\omega) = \left| \sum_{j=1}^{N} \gamma_j e^{j\frac{\omega}{c}(r_j \cdot s)} \right| \tag{6.2}$$

where c is the wave propagation speed, r_j is the position vector for the jth array element, and " $\cdot$ " indicates vector dot product.

Once the element positions are given, the primary focus of most array design problems is to solve eqn (6.2) for a $\underline{\gamma}$ which satisfies a specified $R(s,\omega)$, or some other optimization criterion. In statistically optimum beamforming, it is the average response relative to some desired signal characteristic which is optimized. A classical example is the maximum signal to noise ratio design [78]. Let $\underline{x}_k = \underline{s}_k + \underline{n}_k$ where $\underline{s}$ is a narrowband signal at direction s_0, and $\underline{n}_k$ is the noise signal with known covariance matrix $\mathbf{R}_n = E\{\underline{n}\,\underline{n}^T\}$. The signal to noise ratio at the beamformer output, $\frac{E\{(\underline{\gamma}^T\underline{s})^2\}}{E\{(\underline{\gamma}^T\underline{n})^2\}}$, is maximized when:

$$\underline{\gamma} = \alpha\, \mathbf{R}_n^{-1}\underline{v}\,, \qquad v_j = e^{-j\frac{\omega}{c}(r_j \cdot s_0)} \tag{6.3}$$

Here α is an arbitrary scaling constant, and $\underline{n}$ and $\underline{s}$ are assumed zero mean. This type of design adjusts the beam response, $R(s,\omega)$, in any fashion necessary to maximize the

SNR. Response nulls are placed on plane wave noise signals, and the sidelobes can be arbitrarily high in directions with little noise.

In data independent beamforming, which is the approach used in the examples below, no prior knowledge of the existing noise and interference field is assumed, so the beamformer is designed to provide a response which will perform acceptably in a variety of noise environments. $R(s,\omega)$ is specified so as to have maximum response in the direction of interest, and maximum attenuation in all other directions. Figure 6.2 illustrates the method we shall use to set upper and lower constraints for $R(s,\omega)$ on a finite grid of sample directions, s_i. Sample spacing must be fine enough to eliminate sidelobe leakage between samples, and is a function of the array aperture size. Conventional methods of data independent design include [78] windowing and inverse transformation of the desired spatial response, minmax design with the Remez exchange algorithm [84], spatial response sampling and linear weighted least squares, and pattern search nonlinear optimization algorithms [76,80]. For equally spaced line arrays it is possible to use transversal filter design techniques, but for arbitrary shaped arrays this is not possible. For example, Chebyshev tapered line arrays offer optimally low uniform sidelobe levels for a given mainlobe width, but for non-square planar arrays, and virtually all arbitrary arrays, the Chebyshev polynomial cannot be factored accordingly, and thus can not be used [85]. The l_{1q} optimization method however functions independently of the array configuration.

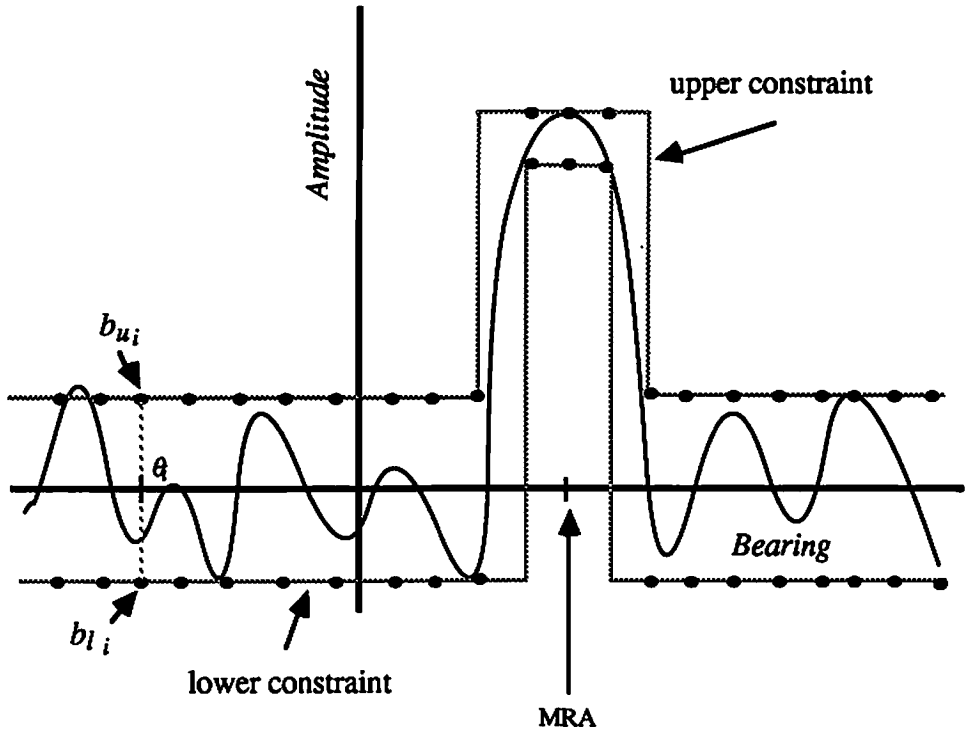


Figure 6.2. Beam Response Constraint Sampling Grid.
This constraint sampling technique is used in the $l_{1/q}$ search algorithms to specify the beam spatial response. For 2-D case $s_i = (sin\ \theta_i\ \ cos\ \theta_i, 0)$.

6.2. Related Work in Array Thinning, Placement, and 3-D Array Design

The need for thinned array designs arises in both hydrophone arrays for SONAR systems and in antenna arrays where the cost of array elements, or associated computational load, is a significant design factor. Thinning implies that an existing array design, typically with uniform element spacing, has some elements removed (i.e. corresponding γ_j=0) while weights are adjusted to maintain a response criterion [12,72]. Alternately one may use the nonuniformly spaced array approach [12,72], where placement of a fixed number

of elements in a continuous space (colinear, coplanar, etc.) is adjusted to optimize the desired response parameters. This corresponds to simultaneously solving (6.2) for γ_j and r_j, $1 \le j \le N$. The two problems can be viewed as essentially equivalent, if the original thinning array is spaced very densely.

A related problem is element placement in unusually shaped or "conformal" arrays, where elements are located on some nonplanar supporting structure (as the hull of a submarine) [73,74,75]. In this case, conventional uniform element spacing is not only difficult (impossible on some shapes), but can lead to wasteful oversampling when the local array surface is not perpendicular to the signal direction of interest. Thinning can become a trial and error proposition. Efficient element placement is thus a critical design issue for unusually shaped arrays unless there is no limit to available sensor elements and computational power. The available shading methods for conformal arrays, including maximizing the target signal to noise ratio with respect to a known noise field [77,78], linear programming methods, or using a pattern search algorithm [76,78], can yield useful shadings for these arbitrary arrays, but can give no information on how many elements are needed, or where they should be placed.

In the literature, the problems of thinning, nonuniform array shading, and element placement are generally solved only for specific array configurations (e.g. [72,76]). Maximally sparse optimization however, can be used in all these cases with the only constraint being that the element placement be symmetric about the origin. This uniform approach is possible since each case can be posed as an order minimization problem, and the algorithms of Chapter 3 are general in nature, requiring only that the problem be expressed as a set of linear inequality constraints. It should be noted that array thinning is usually most effective when very narrow mainlobe response is required [72].

Although little is known about the general properties of nonuniformly spaced arrays, a number of more or less effective design methods have been introduced for specific cases [12,72,88,89,90]. One classical method of thinning line arrays is equivalent to the design of "N^{th}-band" transversal filters [86,87]. Let h_i be the filter impulse response, with corresponding frequency response $H(e^{j\omega})$. If the frequency response is designed such that

$$\sum_{k=0}^{N-1} H(e^{\omega+\frac{2\pi k}{N}}) = Nh_0 \tag{6.4}$$

then it is easily shown that the zeros of the inverse transform of $H(e^{j\omega})$ fall exactly on multiples of N, i.e.

$$h_{Ni} = 0 \quad for\ i \neq 0. \tag{6.5}$$

For example, a lowpass filter design with frequency response symmetric about $\pi/2$ satisfies eqn (6.4) for $N = 2$, and has every even filter tap (except h_0) set to 0, thinning by nearly 50%. The equivalent beam response requirement for a $\lambda/2$ uniformly spaced line array is:

$$\sum_{k=0}^{N-1} R(e^{\frac{\omega}{c}sin(\theta)+\frac{2\omega k}{cN}}) = N\gamma_0 \tag{6.6}$$

$$\gamma_{Ni} = 0 \quad for\ i \neq 0.$$

where θ is the bearing response angle, and c the wave propagation speed.

Jarske, et al. have recently shown that for small symmetric linear arrays, a nonuniformly spaced design with element positions constrained to be placed at multiples of $\lambda/2$ spacing, can match the optimal response obtained with continuous nonuniform spacing [72]. They speculate that this property will follow for larger line arrays, and produce at least

locally optimum results. In their simple construction method, elements are positioned as symmetric pairs, and thinning is done by removing pairs of elements as follows:

1) Choose the number of free element pairs, M, to be included in the final design. Choose the maximum array length, $2D_{max}$, and form a filled array this long using $\lambda/2$ spacing and uniform coefficients.
2) Remove one pair of elements such that the resulting array has the smallest possible sidelobe level over the stopband interval.
3) Repeat step 2) until the number of elements is $2M+1$.
4) Calculate the optimal weights, χ, for the remaining elements using linear programming.

This is the approach used in the example comparisons of section 6.4. These examples show how the $l_{1/q}$ optimization algorithms can be applied directly without restriction on the array configuration, and how the maximally sparse optimization approach can improve on other thinning methods.

6.3. Formulation for $l_{1/q}$ Search Algorithms

In order to apply the algorithms of Chapter 3 to the beamformer of Figure 6.1, we will use the presteered beamformer approach, where $\gamma_j = a_j\phi_j$, with $\underline{\phi}$ the vector of precomputed unit magnitude phase shifts to steer the maximum response angle (MRA) to the direction of interest, and $\underline{a}$ the real valued weights, or shades, which are to be optimized. Additionally we require the elements to be placed symmetrically about the origin. We wish to solve for the maximally sparse $\underline{\chi}$ which meets the spatial response constraints. As in section 6.1, we take M samples of the desired upper and lower spatial magnitude

response bounds from a dense enough grid on an enclosing sphere to control sidelobe leakage. Let s_i be the vector of direction cosines to the point on the sphere where the upper and lower response constraints, b_{u_i} and b_{l_i} are sampled. Let s_0 be the vector direction cosines of the MRA, and r_j the position vector for the jth array element. Let a_j be the computed shade for the jth element, with $\underline{a} = \underline{a}^+ - \underline{a}^-$ where $a_j^+ , a_j^- \geq 0$ so we may obtain positive or negative shade values while using the positive only vectors $\underline{a}^+$ and $\underline{a}^-$ in the algorithm. We require symmetry about the origin to insure a real response value, y_k, i.e.:

$$r_j = - r_{N-j-1}, \quad \text{and} \quad a_j = a_{N-j-1} \tag{6.7}$$

where N is the number of array elements, therefore we need only solve for $N/2+1$ shades (coefficients) in $\underline{a}$. The real amplitude response at the constraint points is then given by a cosine transform, and is expressed in matrix form as:

$$\mathbf{H}_{ij} = \frac{2}{N} \cos[r_j \cdot (s_i - s_0)\ \omega/c] \ , \text{ for } i=1,\cdots M \ , \ j=1\cdots N/2+1 \tag{6.8}$$

We introduce slack vectors $\underline{z}^+$ and $\underline{z}^-$ of length M, and have the final form of the system:

$$H\underline{x} = \underline{b} : \begin{bmatrix} \mathbf{H} & -\mathbf{H} & \mathbf{I} & \mathbf{0} \\ \mathbf{H}^{\ddagger} & -\mathbf{H}^{\ddagger} & \mathbf{0} & -\mathbf{I}^{\ddagger} \\ \mathbf{1} & \mathbf{1} & \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \underline{a}^+ \\ \underline{a}^- \\ \underline{z}^+ \\ \underline{z}^- \end{bmatrix} = \begin{bmatrix} \underline{b}_u \\ \underline{b}_l^{\ddagger} \\ d \end{bmatrix} \tag{6.9}$$

where $\ddagger$ indicates rows may have been negated to force $\underline{b}_l^{\ddagger}$ to be non-negative. The bottom row of H is added to constrain the sum of shade absolute values. Simulations have shown d can be adjusted to improve beamformer stability and array gain relative to a noise field . Without the bottom constraint row, results often contain very large positive

and negative values. Although this does not affect the array spatial response, $R(\omega,s)$, the magnitude squared response to independent noise at the sensor elements is $\|\gamma\|^2\sigma^2$, where σ^2 is the noise variance. This can be significantly larger than the mainlobe response $R(\omega,s_0)$ when γ contains large negative values. Array stability, or lack of critical sensitivity to element gain or positions errors, is also better for designs with not large negative γ_j. By choosing d slightly larger than the expected mainlobe magnitude response, this problem can be avoided.

From eqn (6.9) we first use a phase one algorithm to find any basic solution and then optimize using the simplex or simulated annealing algorithms. With a low order solution a computed, the final complex element weight for the beamformer is

$$\gamma_j = a_j\phi_j = a_j e^{-j\frac{\omega}{c}(r_j\cdot s_0)} \tag{6.10}$$

6.4. Results

Consider the 60 element transparent concentric ring array of Figure 6.3a. We wish to form beams, steered horizontally, in the plane containing the array. This is similar to the configuration used by some "dipping" sonar systems which suspend a cylindrical ring array in the water from a helicopter and form horizontal search beams. Figure 6.3b shows the beam response for the full array using unity magnitude shading, with complex phase shifting at each element equal to the conjugate of the elemental propagation phase delay for a plane wave arriving from the maximum response angle (MRA) of zero degrees. A sinusoidal signal at 1 kHz is assumed, which gives an average element to element spacing of just over $\frac{\lambda}{4}$. We require the element positions to be symmetric about the origin.

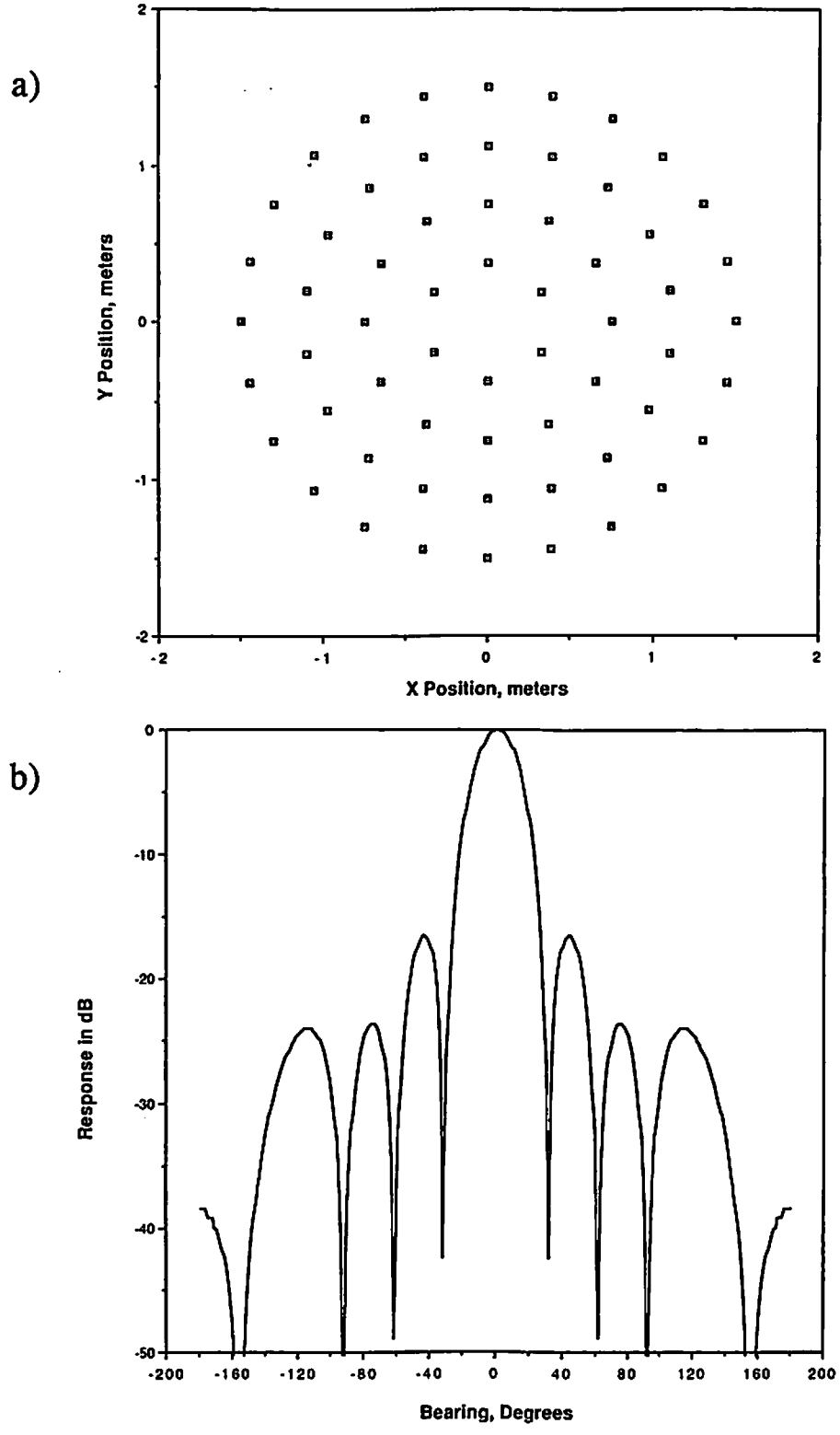


Figure 6.3. Original 60 Element Concentric Ring Array.
a) Element positions. Elements omnidirectional. b) Unity shaded beam response at 1 kHz in seawater.

For the thinned array design, we use the same element phasing as in Figure 6.3, but let the algorithm adjust the real amplitude shading. The mainlobe width is constrained to be the same as Figure 6.3b, with sidelobes no larger than the first sidelobe of 6.3b. Allowing some of the secondary sidelobes to come up to level of the first allows some degree of freedom which is used by the algorithm in order minimization. Figure 6.4a shows the remaining elements of the array after thinning by the $l_{1/q}$ simplex search algorithm for q=15, and 6.4b shows the corresponding response pattern. Only 16 of the original elements are needed to maintain the original mainlobe shape and maximum sidelobe level, and the results agree with earlier observations that the outer elements of a ring array are the primary contributors to beam response. Note that the algorithm simultaneously selects the elements and computes the optimal shade weighting.

In [72], Jarske, et al. propose a simple thinning procedure for narrow beam arrays. In their example 4.1, a symmetric line array is designed with length constrained to be $\leq 50\lambda$, and a mainlobe width constraint of +/-3.6 degrees (wavenumber = .08 π/λ). The thinning procedure used in [72] requires the elements to be placed at multiples of $\lambda/2$ from the array center. Figure 6.5a shows the final element positions for the best solution in [72], which produced a maximum sidelobe level of .217 (-13.27 dB), using 25 elements. Figure 6.5b shows the element positions for the $l_{1/q}$ simplex search (q=15) solution to the same problem, but with the mainlobe further constrained to +/-2.06 degrees. 26 elements were required. Figure 6.6 shows the corresponding response pattern, The initial array used in the search was 251 elements long, with .2λ spacing, for a total length of 50λ. Using the stochastic search algorithm and truncating the sequence prior to reaching a temperature of zero, this solution was improved to that shown in Figures 6.5c

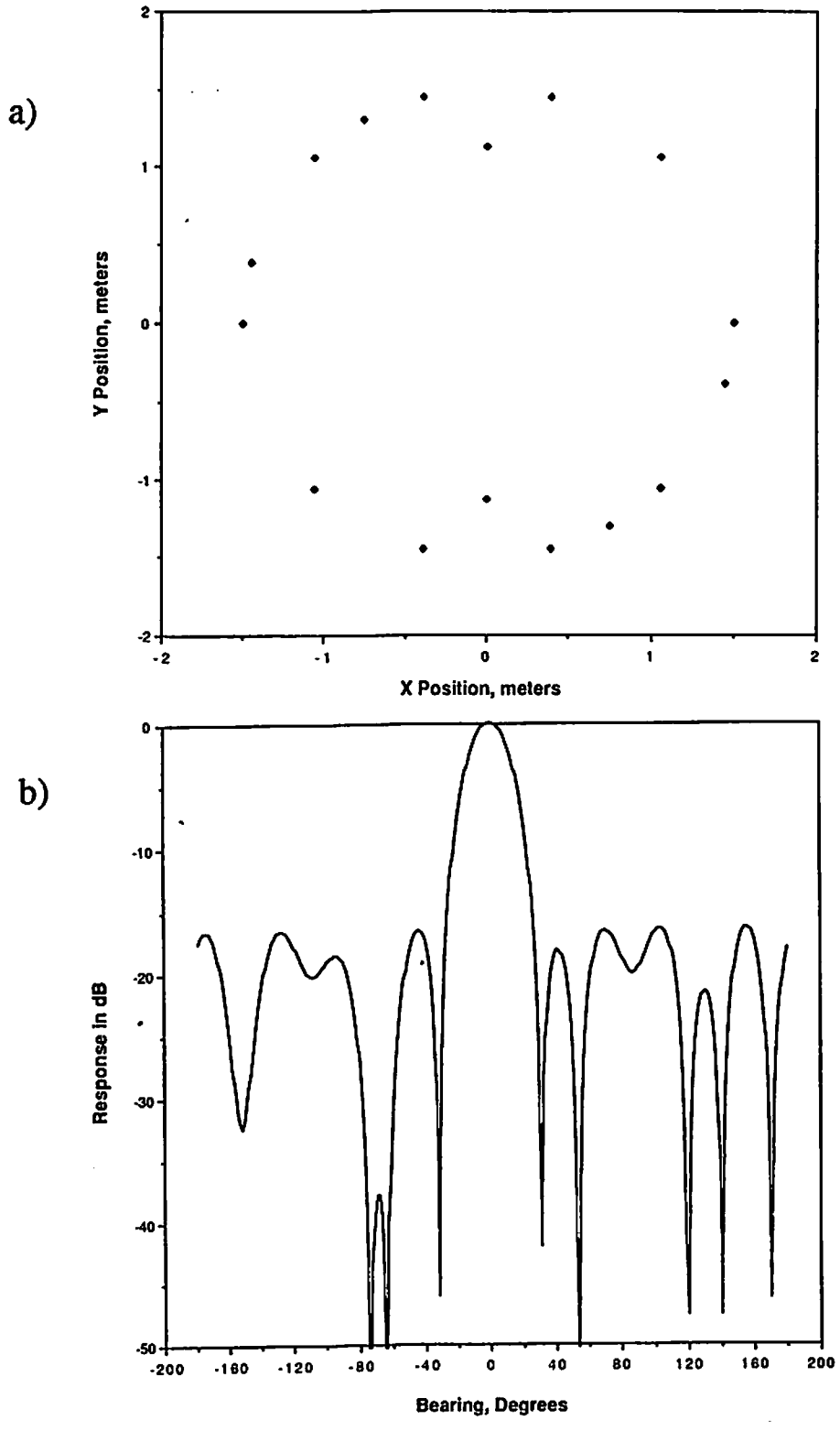


Figure 6.4. Thinned Array Results Using l_{11q} Simplex Search with Mainlobe and Maximum Sidelobe Constrained to Match Figure 6.3b.
a) Element positions. b) Thinned and optimally shaded beam response.

and 6.7. Note that here only 22 elements were needed, the aperture is slightly smaller, and the mainlobe narrower than example 4.1 of [72]. The difference between the $l_{1/q}$ and the stochastic search solutions is a example of termination at a local optimum which is overcome by the simulated annealing randomization of the search. Figure 6.5d shows the final element positions, and Figure 6.8 the corresponding beam response for the $l_{1/q}$ search with all constraints (including mainlobe width) identical to the example in [72]. Only 16 elements were needed and the aperture was reduced further.

Figure 6.9 shows the results of the $l_{1/q}$ simplex search (q=11) for a more complex array response specification, demonstrating the flexibility of the technique. The response constraints included:

1) MRA steered to +15 degrees
2) Mainlobe width +/- 4 degrees at -20 dB
3) -35 dB null in the range of 35 to 45 degrees
4) Response at less than -70 degrees and greater than +70 degrees unconstrained
5) Sidelobe level ≤ -20 dB.

As can be seen in in figure 6.9a, each of these requirements were met by the 30 element final design with element positions as shown in 6.9b. The starting array contained 200 elements, evenly spaced over 61 meters at approximately .2λ. The phase one linear programming solution contained 66 elements, which were then reduced to the final 30 in the $l_{1/q}$ search. This result is not a global optimum. The stochastic search could be used to further reduce the number of elements, and starting with the unnecessarily large 200 element array complicates the search process and makes less than optimum local solutions more likely.

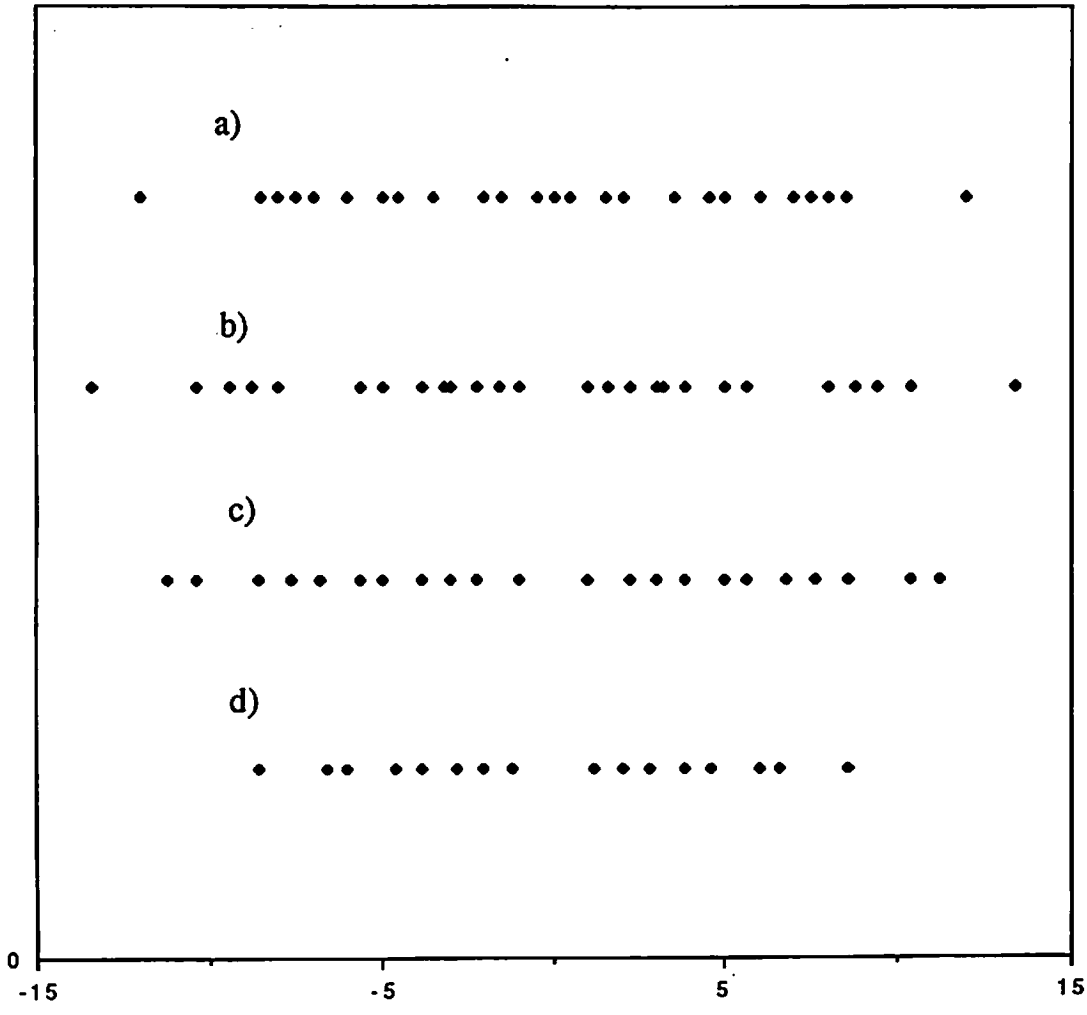


Figure 6.5. Element Positions of the Four Thinned Array Placement Examples. a) Jarske, et al. example 4.1 [72]. b) $l_{1/q}$ simplex search with more narrow mainlobe constraint than a. c) Stochastic search with narrow mainlobe. d) $l_{1/q}$ simplex search with same response constraints as a. above.

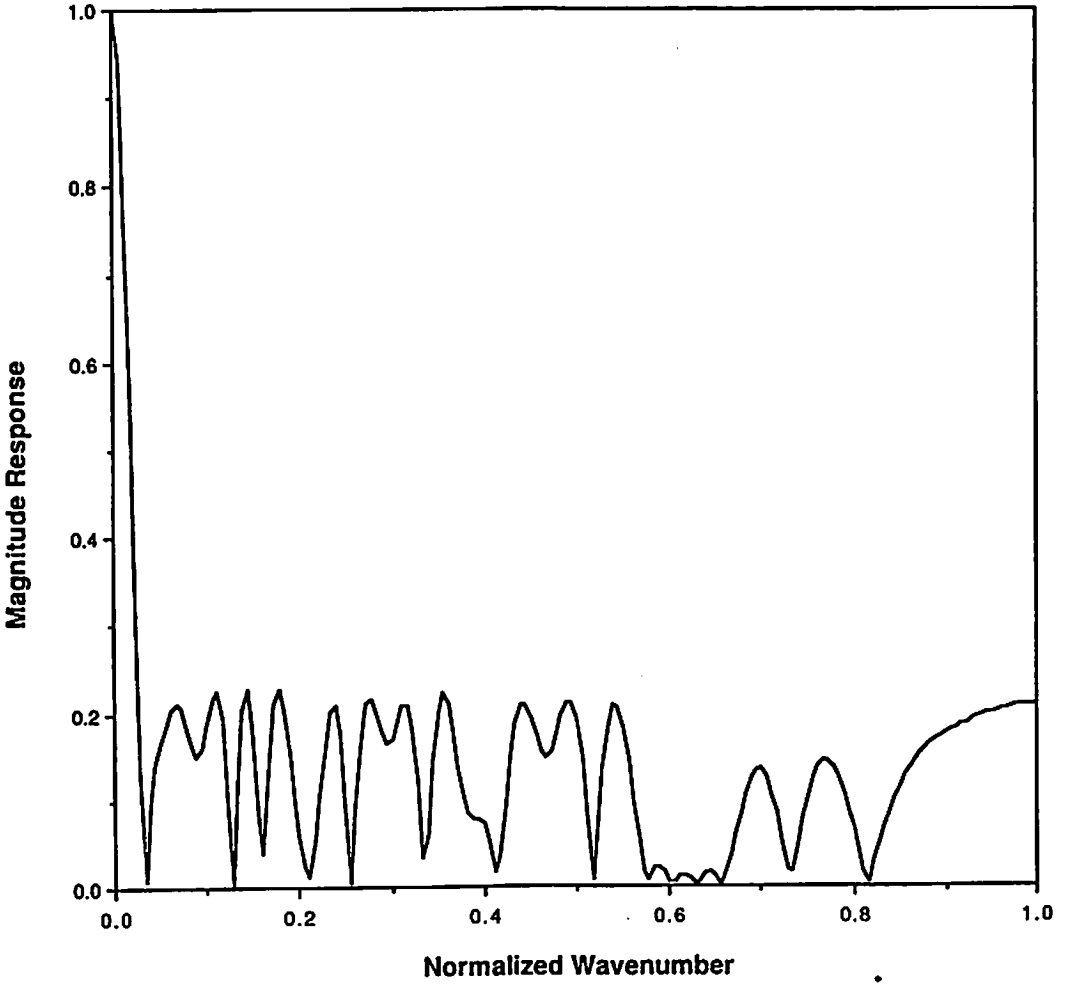


Figure 6.6. Narrow Mainlobe Beam Magnitude Response for $l_{1/q}$ Simplex Search Result, 26 Element Array of Figure 6.5b.

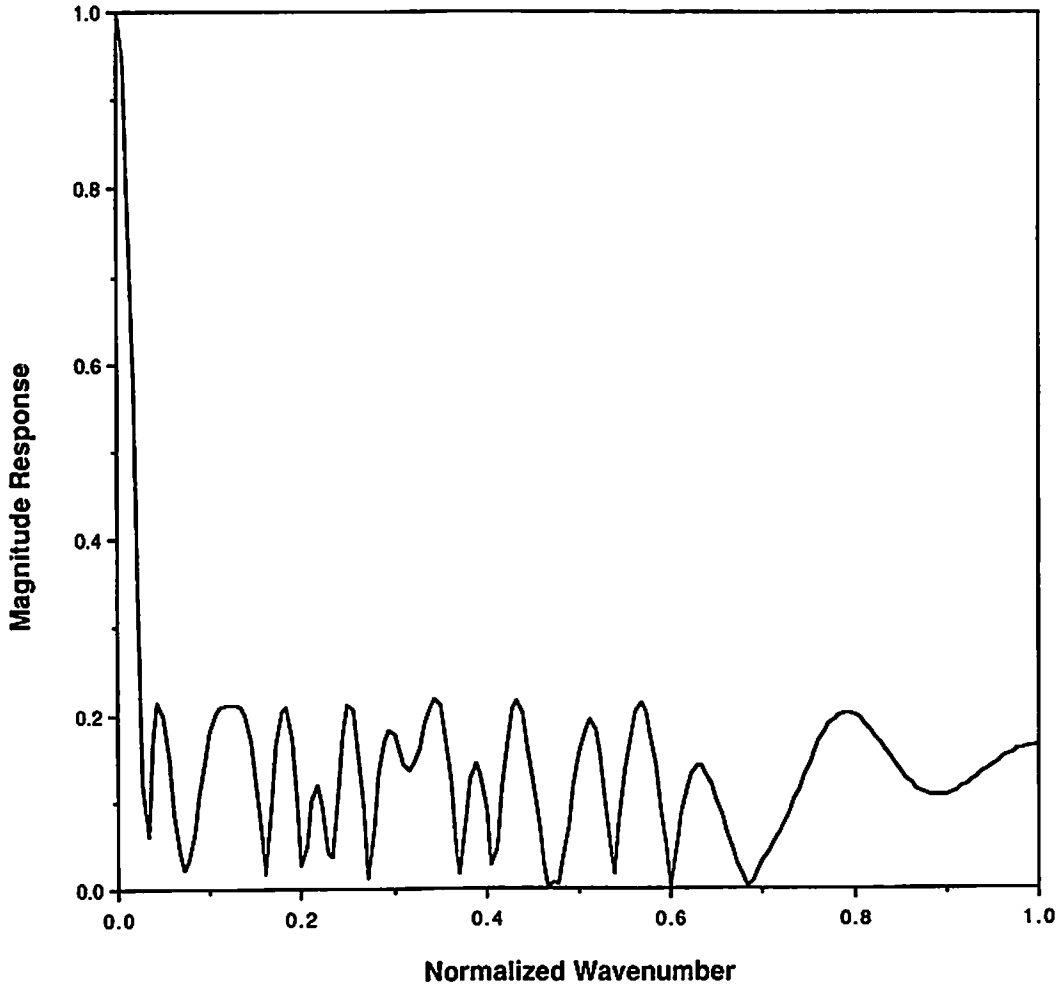


Figure 6.7. Narrow Mainlobe Beam Magnitude Response for Stochastic Search Result, 22 Element Array of Figure 6.5c.

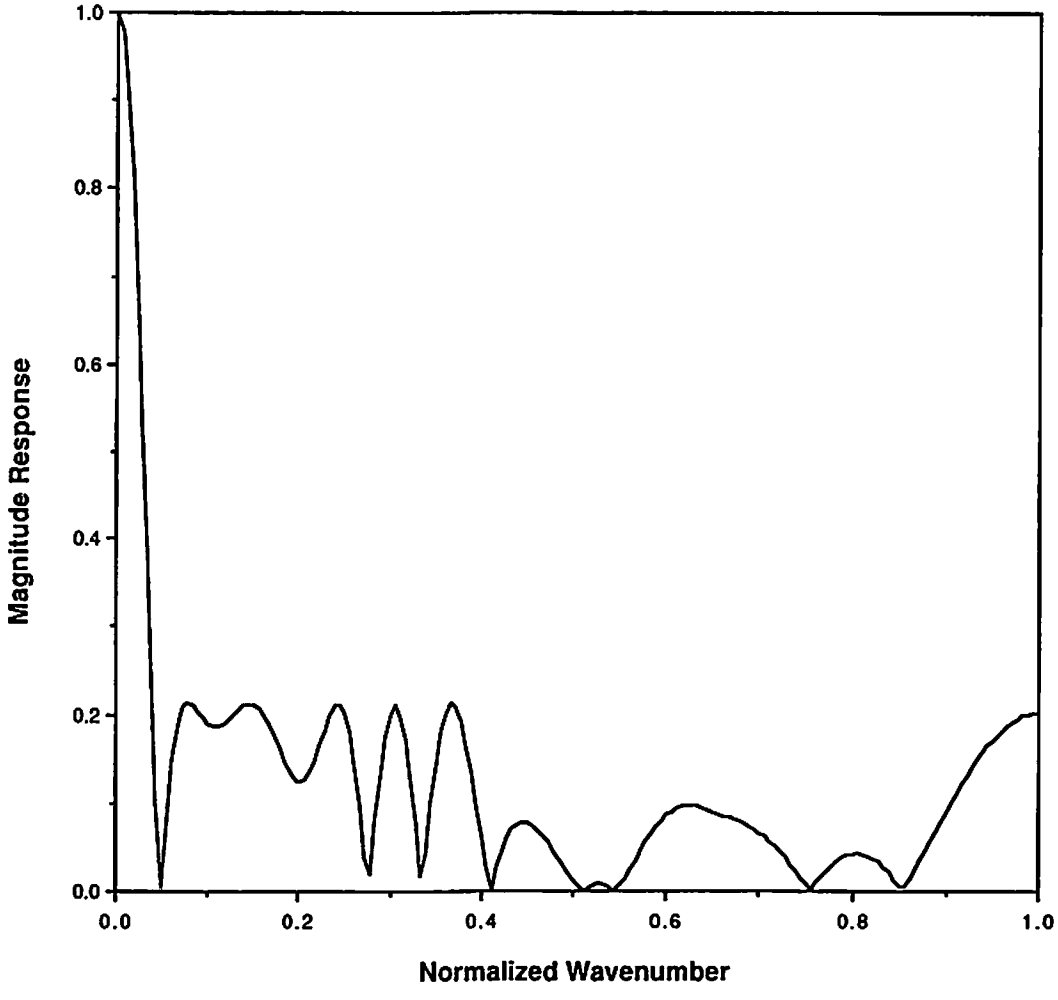
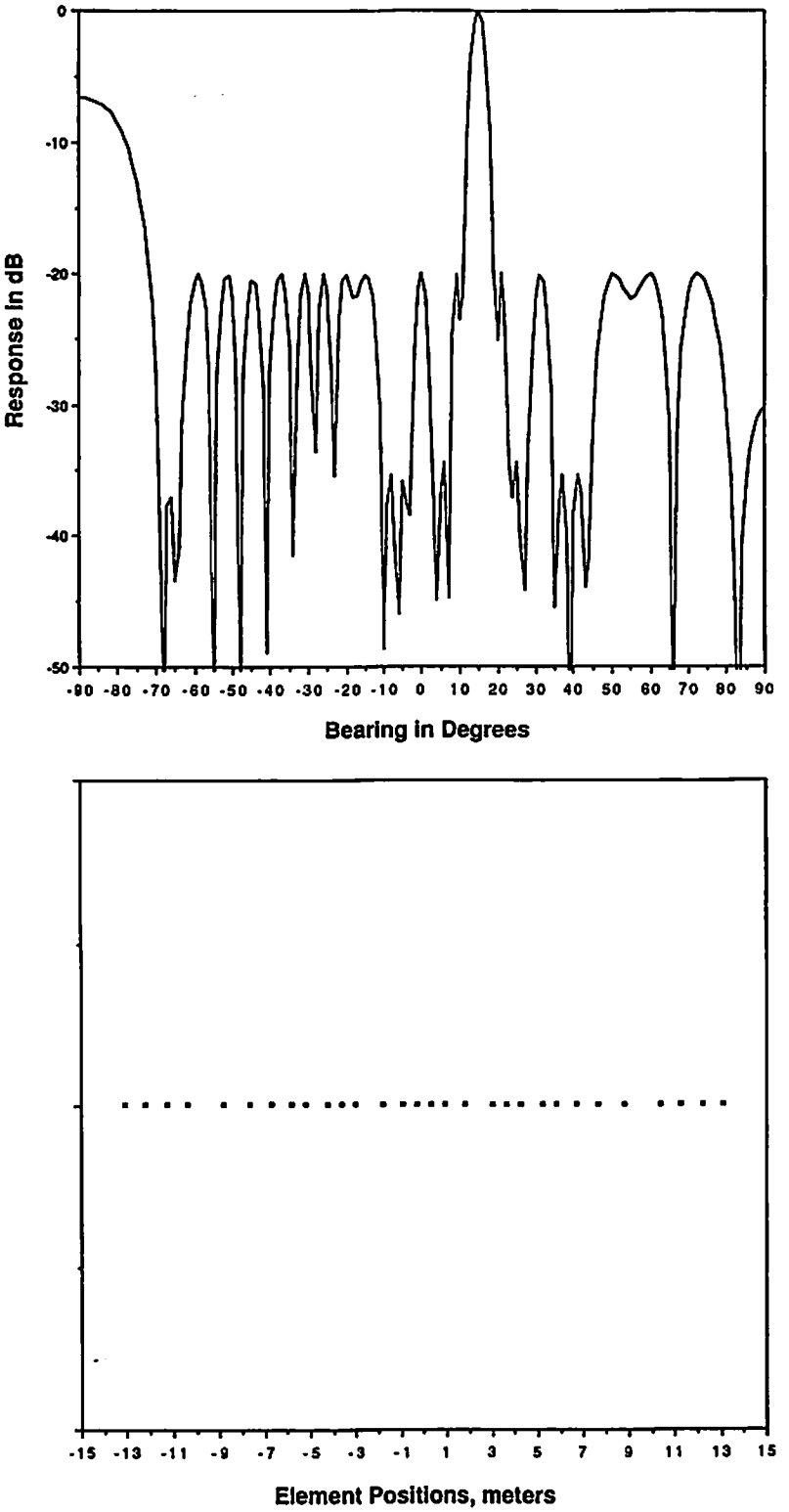


Figure 6.8. Beam Magnitude Response for $l_{1/q}$ Simplex Search Result, 16 Element Array of Figure 6.5d.



a)

b)

Figure 6.9. a) Beam Magnitude Response for $l_{1/q}$ Simplex Search Result, Arbitrary Response Specification. MRA = 15 deg., null from 35 to 45 deg., -20 dB sidelobes. b) Element positions.

6.5 Extension to Broadband Beamforming Designs

The development and examples above assumed that beamforming was to be performed only over a very narrow band of interest, so that a single complex beamforming coefficient could be used for each array element. In this section, the broadband case is considered, and beamformer architectures and application approaches are proposed which could potentially yield sparse array designs. The following discussion addresses theoretical aspects and implementation issues, while experimental evaluation is left for future research.

6.5.1. Sidelobe Control for Small Percentage Bandwidth Beamformers

The beamformer architecture of Figure 6.1, though designed for a single frequency, can often be operated over a reasonably broad bandwidth under the following conditions:

1) The band of interest is a small percentage of the center frequency:
$\frac{\omega_{+1} - \omega_{-1}}{\omega_0} << 1$, where ω_0 = band center frequency, ω_{+1} = band upper edge, and ω_{-1} = band lower edge. A percentage bandwidth of 10% is often used.

2) Inter-element spacing is no greater than $\lambda/2$ at the maximum frequency of interest.

3) The complex beamformer coefficients and element placement are of non-critical design. This condition will not be defined explicitly, but acceptable examples include line array designs using eqn (6.10) with a set of a_j which are rectangular window, Hamming window, Kaiser-Bessel Window, or Chebyshev shaded. A rectangular windowed, phase shift shading beamformer is an example of an unacceptable critical design.

With these conditions met, the performance of most uniformly spaced beamformers performance is still degraded at off center frequencies by a mis-steering of the beam's MRA, proportional to sin^{-1}(MRA), and variation in the mainlobe width. Sidelobe levels are not affected for most shading designs until frequencies are high enough to cause spatial undersampling, at which point grating lobes appear due to spatial aliasing. The thinned array designs of Section 6.4 do not generally meet criteria 2) or 3). As shown in Figure 6.10, the thinned line array of figure 6.8 shows no degradation at half the design frequency other than mainlobe widening, but at 1200 Hz, severe grating lobes appear. This occurs at a lower frequency for thinned arrays than for $\lambda/2$ spaced uniform arrays since the degrees of freedom in the "invisible response region" are utilized in the thinning process [72]. Figure 6.11 shows the effect of frequency changes on the response of the thinned circular array of Figure 6.4. Figure 6.11a shows how the sidelobe level is dramatically increased as operating frequency drops from 1000 to 700 Hz. For comparison, Figure 6.11b shows how the 700 Hz response for the full unity shaded array has no change in sidelobe level, but has widened mainlobe response with respect to Figure 6.3b. These examples demonstrate the need for a different approach if thinned arrays are to be used over more than just a narrow band of temporal frequencies.

The following approach is proposed to overcome this problem. As shown in Figure 6.12, additional constraints are specified for operation at an arbitrary number of frequencies spanning the band of interest. Their corresponding MRA directions are adjusted to account for frequency dependent mis-steering. For example, we may specify constraints at $\omega_{-\Delta} = \frac{2}{3}(\omega_0 - \omega_{-1})$, $\omega_{+\Delta} = \frac{2}{3}(\omega_{+1} - \omega_0)$, and ω_0; with $MRA_{-\Delta}=sin^{-1}\left(\frac{\omega_0}{\omega_\Delta} sin\ (MRA_0)\right)$ and $MRA_{+\Delta}$ computed likewise.

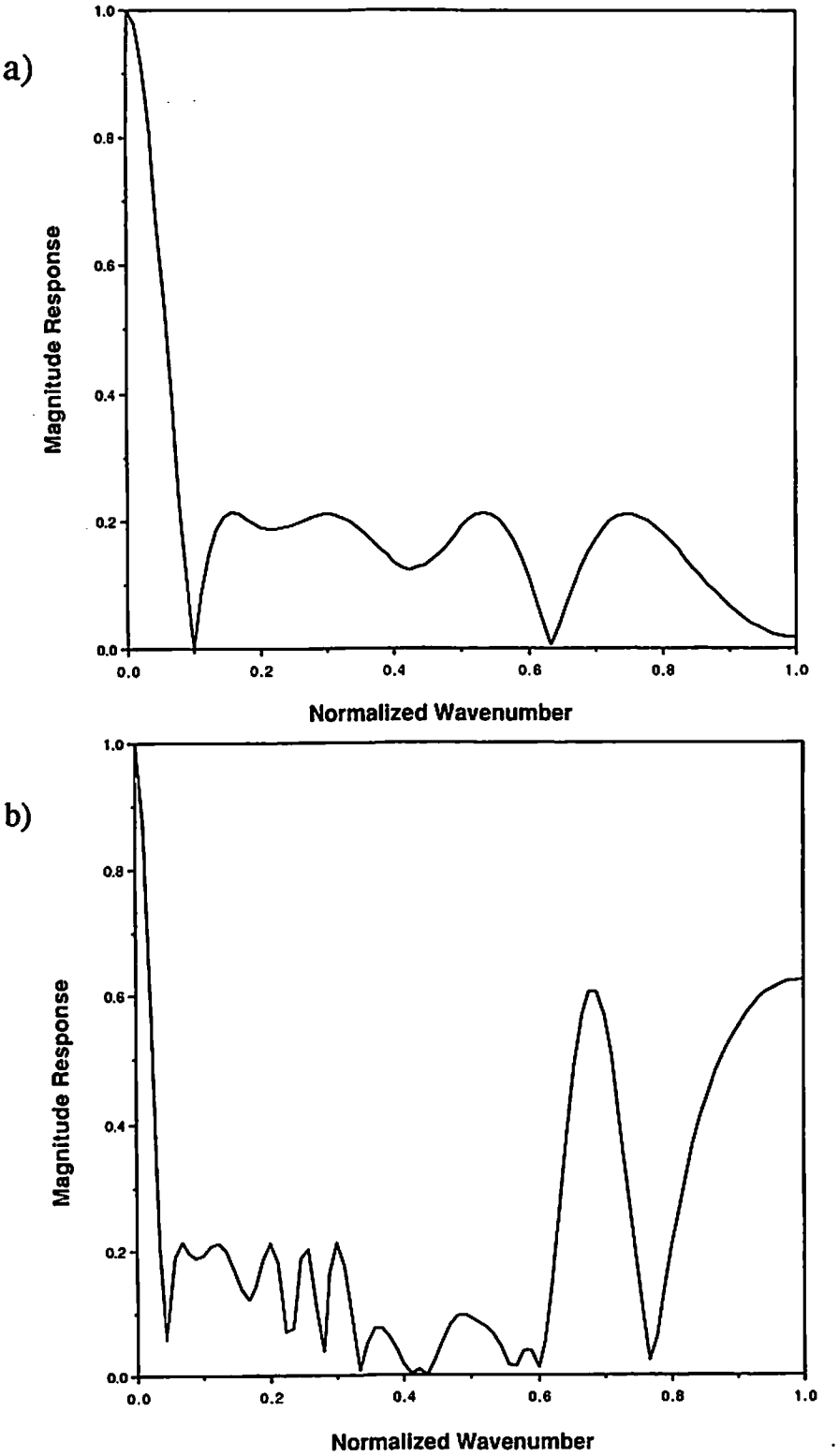


Figure 6.10. Out of Band Performance for the Thinned Line Array of Figure 6.8. a) 500 Hz response. b) 1200 Hz response.

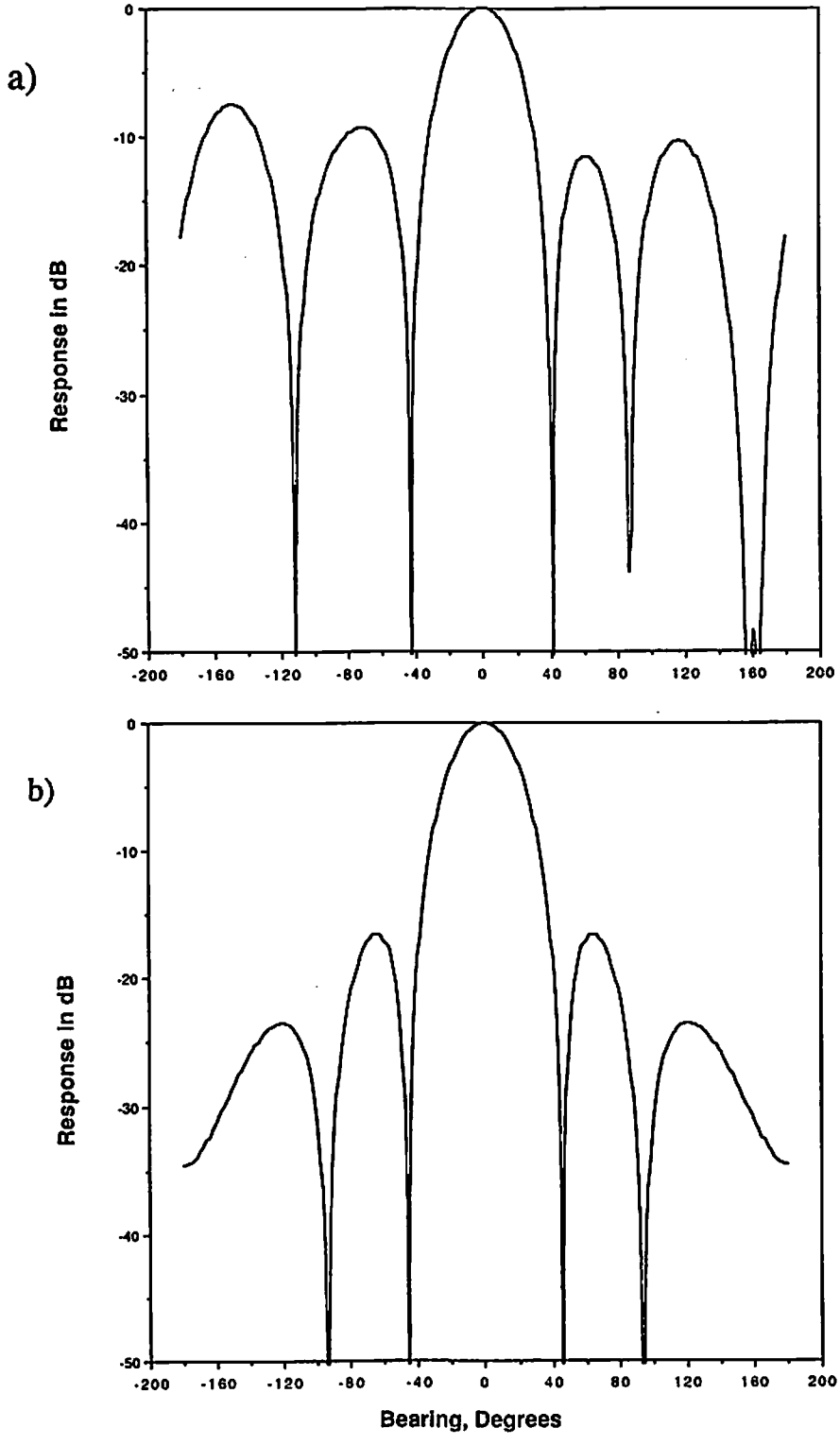


Figure 6.11. Out of Band Performance for the Thinned Circular Array of Figure 6.4. a) 700 Hz response. b) 700 Hz response of the unthinned full array given for comparison.

The additional constraints are incorporated into eqn (6.9) by letting

$$\mathbf{H} = \begin{bmatrix} H^{0} \\ H^{-\Delta} \\ H^{+\Delta} \end{bmatrix} \tag{6.11}$$

where H^{0}, $H^{-\Delta}$, and $H^{+\Delta}$ are computed using ω_0, $\omega_{-\Delta}$ and $\omega_{+\Delta}$ respectively in eqn (6.8). $\underline{b}_u$ and $\underline{b}_l$ are likewise extended by concatenating the corresponding constraint sample points. It is expected than this method will not yield designs as sparse as for the single frequency case, but significant thinning should be possible while controlling sidelobe leakage.

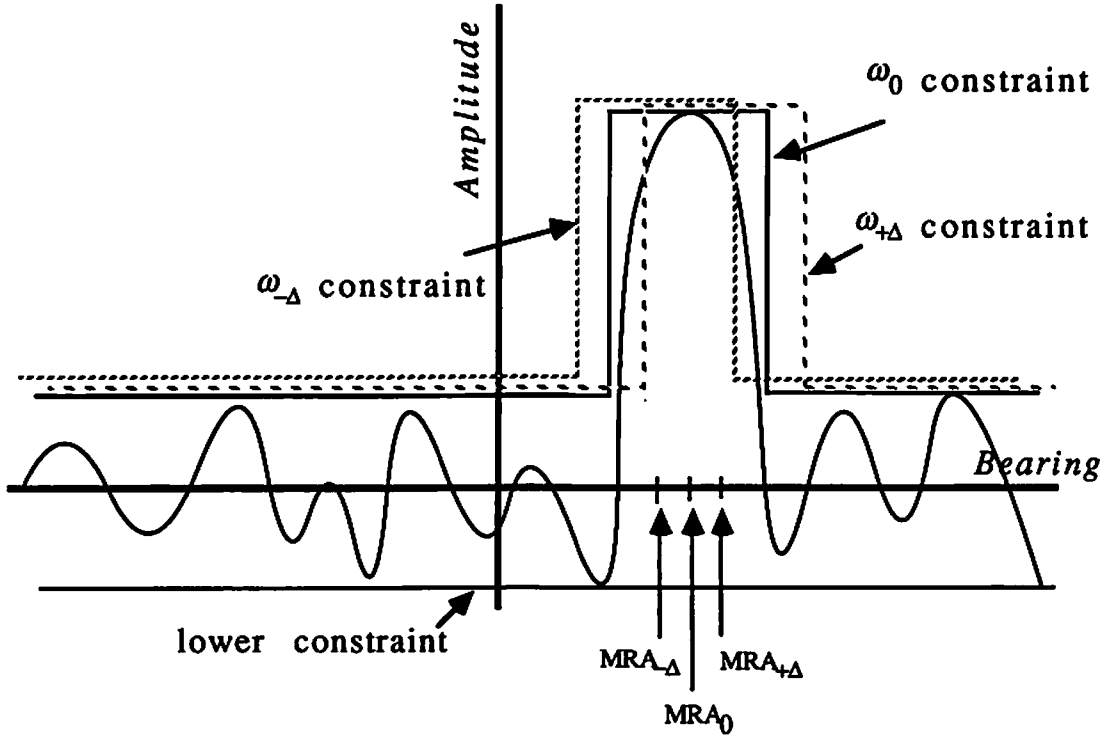


Figure 6.12. Beam Response Constraints at Three Frequencies for a Small Percentage Bandwidth Beamformer.

6.5.2. Application to General Broadband Beamformer Structures

Two frequently used, and roughly equivalent, broadband beamformer architectures are shown in Figures 6.13 and 6.14 [78]. Figure 6.13 is a time domain implementation, while Figure 6.14 operates in the frequency domain. They both offer many degrees of freedom, with multiple tap weights per array element, so that virtually any combined spatial and spectral-temporal response, $R(s,\omega)$, can be achieved (within the limitations of element placement and filter or FFT length). These structures are particularly useful for broadband statistically optimal design where the spatial and spectral-temporal characteristics of the noise and interference field are incorporated in the design optimization [79].

These structures do however provide some difficult challenges for sparse array design. One would think it possible to specify the desired spatial response at enough discrete frequencies to fully specify the design. The constraint equation becomes $|\mathbf{H}\underline{w} - \underline{b}| \leq \underline{\varepsilon}$, with $\underline{w}$ the extended vector of all w_{ij}, $\underline{b}$ the vector of upper and lower error bounds for deviation from $R(s_k,\omega_l)$ for all combinations of constraint sample position vectors, s_k, and frequencies ω_l, and with $\mathbf{H}$ the corresponding spectral - spatial response system matrix. Using this form directly in eqn (1.2) will certainly minimize the number of nonzero weights, but will not thin the array since if for any j, $w_{ij} \neq 0$, then array element i cannot be eliminated. A modified $l_{1/q}$ cost function which would force the desired thinned array optimization is:

$$g(\underline{w}) = \sum_i \left(\sum_j |w_{ij}| \right)^{1/q} \tag{6.12}$$

Minimizing $g(\underline{w})$ has the effect of forcing the l_1 norms, separately computed for the weights of each channel, to be maximally sparse. This attempts to set all the weights of a channel to zero simultaneously, so that an array element may be eliminated. $g(\underline{w})$ is

shown in Appendix C to be concave, and therefore the fundamental theorem (section 2.2) can also be applied to eqn (6.12). This implies that the simplex search based algorithms described above could be used. This needs more study to determine if the algorithms can be used efficiently. Additionally, this approach requires a very large system matrix, **H**, in order to accommodate sufficient discrete frequencies. These very high dimensionality systems can create implementation and performance problems for the simplex search programs.

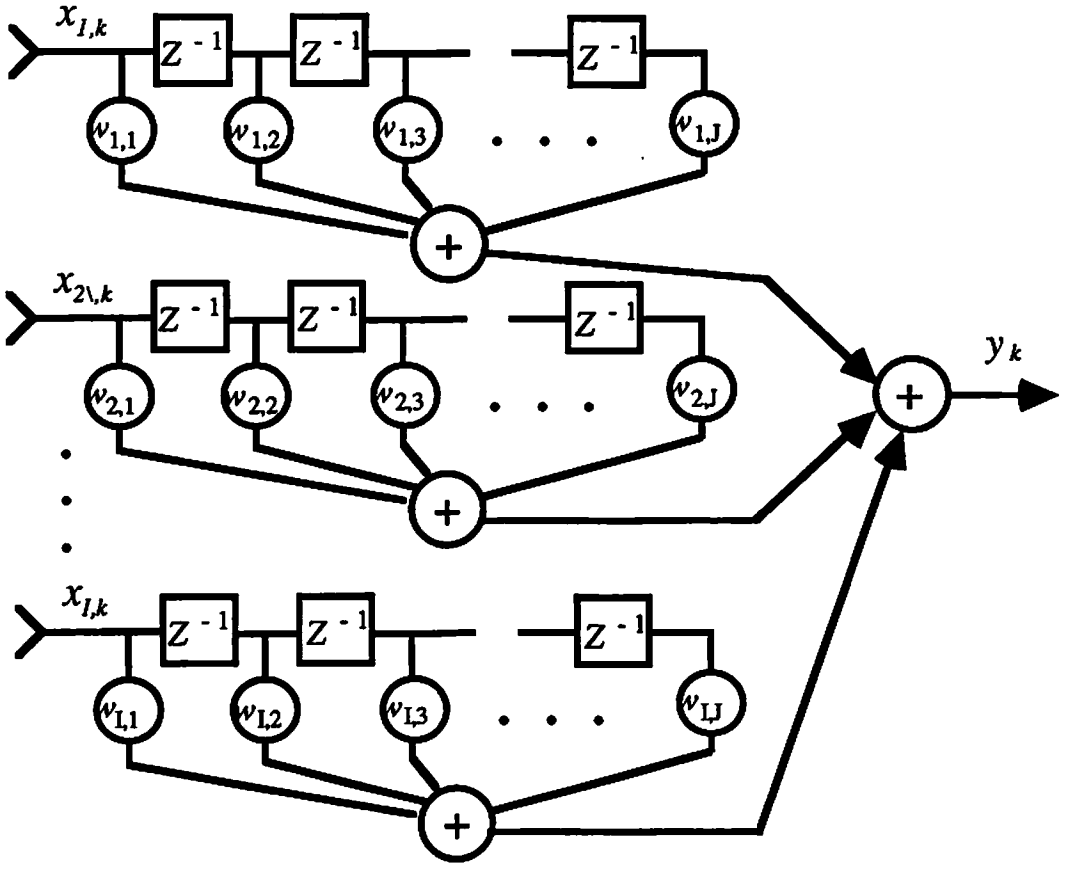


Figure 6.13. Broadband Beamforming Structure Using Temporal FIR Filters.

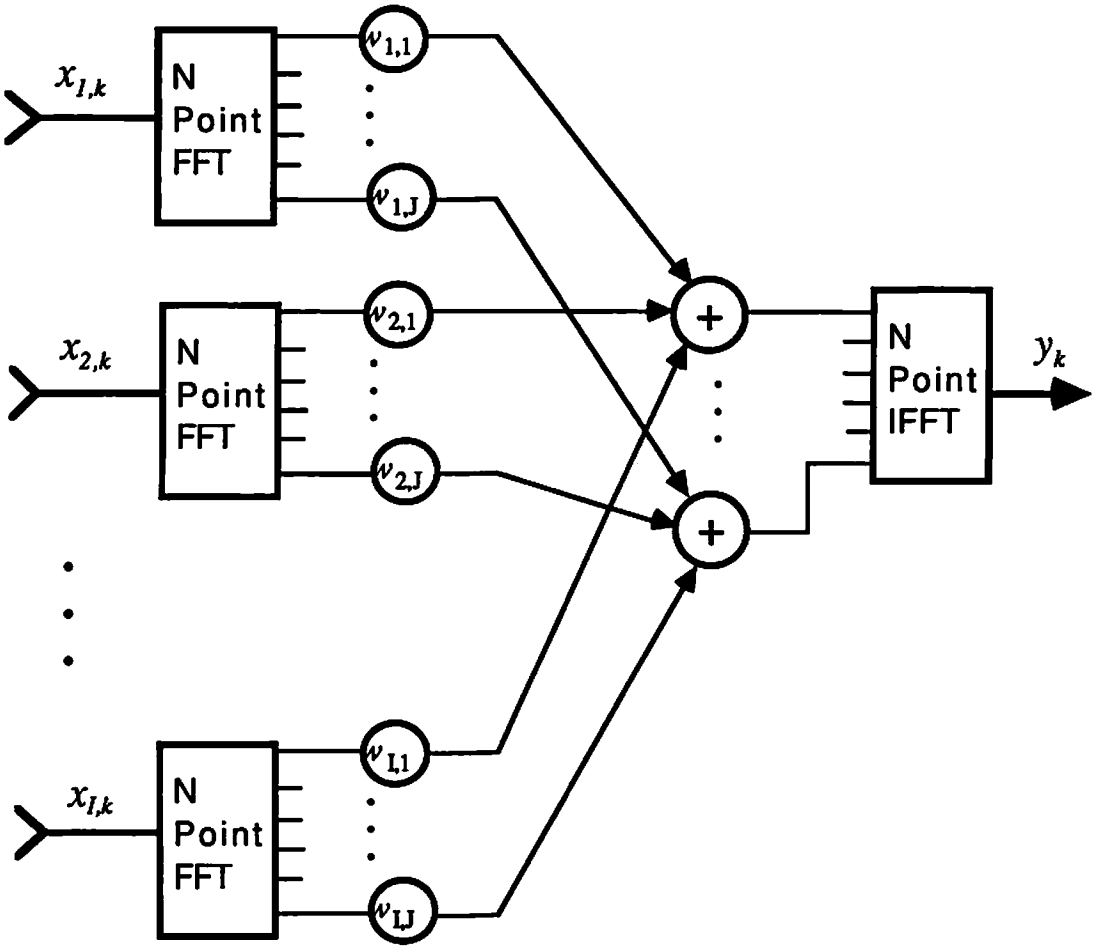


Figure 6.14. Broadband Beamforming Structure for Frequency Domain Weighting.

6.5.3. Application to Separable Broadband Structures

An adaptation of the general broadband structures of section 6.5.2. will allow a tractable maximally sparse formulation with only a small penalty in system design flexibility. If the spatial and temporal-spectral response requirements can be specified in a separable form, i.e. $R(s_k,\omega_l) = R(s_k)R(\omega_l)$, we may use the architecture shown in Figure 6.15.

This is the same as Figure 6.13 or 6.14, except that an additional series weight, γ_i has been added to each array element, and we require that each spectral processing block be identical, such that $w_{ij} = w_j$ for all i. The spectral blocks must satisfy $R(\omega_l)$ to within allowable tolerance, and all spatial response control is provided by the γ_i weights.

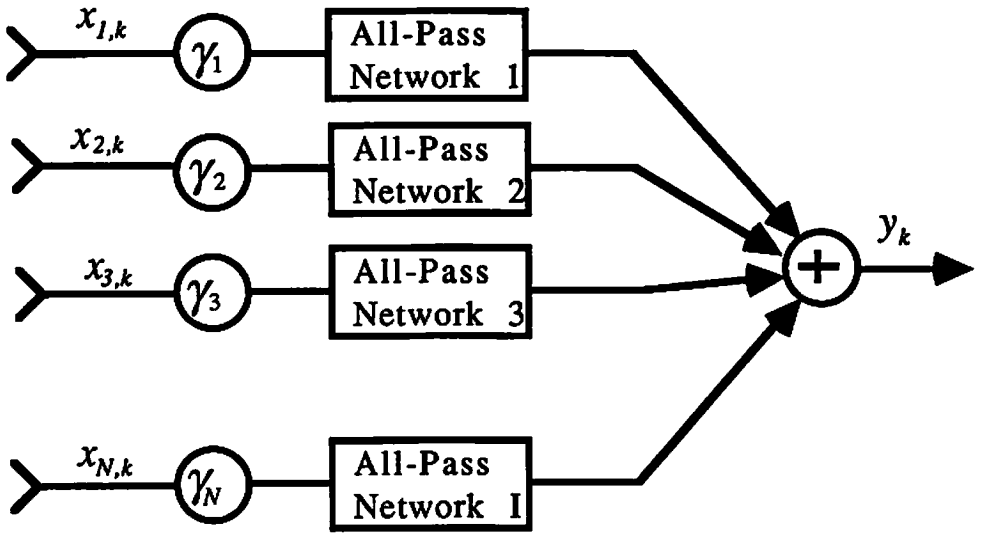


Figure 6.15. Separable Broadband Architecture.

The following design procedure may be used:

1) Using any acceptable filter design method, compute all w_j to satisfy $R(\omega_l)$. Either the FIR filter architecture of Figure 6.13, or the FFT based structure of 6.14 can be used.

2) set up $\underline{b}$ as the spatial constraint vector derived from $R(s_k)$ at a few discrete frequencies across the band of interest. $R(\omega_l)$ is not used in this step, and we need only enough sample frequencies to control sidelobe leakage. H is computed from

eqn (6.8) as in section 6.5.1. above, using each discrete spatial constraint frequency.

3) Solve $min\ g(\underline{x})\ s.t.\ |\mathbf{H}\underline{x} - \underline{b}| \leq \underline{\varepsilon}$ using any of the algorithms described in Chapter 3.

This method insures a thinned broadband array result and keeps the system dimensionality low enough to make the algorithms practical.

A special case, often used in industry, of the separable broadband beamformer is the all-pass filter structure. This can be used when the spatial temporal-spectral requirements are simply to maintain a single specified spatial response across a broad band of frequencies. The spectral filtering is designed as an all-pass network which does not affect amplitude, but introduces a frequency dependent phase shift to compensate for phase differences, relative to band center, at each element. With this phase compensation, equally spaced line array shading can be computed as if for a narrow band beamformer, sidelobe levels are not frequency dependent, and the MRA does not shift with frequency. The phase correction is computed such that for a plane wave signal arriving at the MRA, the corrected output of each element has zero relative phase across the band. $R_i(\omega)$, the spectral response for element i:, is defined as:

$$R_i(\omega) = e^{-j\frac{\omega-\omega_0}{c}(r_i \cdot s_0)} \tag{6.13}$$

where r_i is the position vector for element i, s_0 is the MRA direction cosine vector, and ω_0 is the band center frequency. Maximally sparse optimization then proceeds as described above.

7. CONCLUSIONS

7.1. Concluding Remarks

The significance of the preceding work lies both in its contribution to solution of a difficult class of optimization problems, and demonstration of their usefulness in practical engineering applications. Both the optimization theoretic developments and the demonstrated solutions of real-world engineering problems are novel. The area of maximally sparse optimization has not been sufficiently exploited in the past, either because of the lack of practical algorithms for global solutions, or because problems which could benefit from the sparseness criterion were not recognized. This dissertation has made progress in addressing both of these deficiencies.

Three new algorithms have been presented for determining minimum order solutions to problems with linear or quadratic inequality constraints. These algorithms are based on minimization of the $l_{1/q}$ quasi-norm, $q>1$, which for a sufficiently large value of q is equivalent to the minimum order criterion, as shown in Theorem 2. In the case of linear constraints, it was shown in Theorem 1 that the solution lies at a vertex of the convex polytope formed by the constraints, and can therefore be found using a simplex algorithm to search the vertices for either a locally or globally optimal solution. The stochastic search algorithm uses a simulated annealing approach to guide the simplex search and converge to a global minimum order result. This algorithm is the only currently available method for truly optimal thinning of arbitrary beamforming arrays, and should enable similarly successful solution of related sparse problems. In the case of quadratic constraints, a convex transformation was described which permitted use of gradient descent techniques for locally optimum solutions.

An estimation theory interpretation of the problem showed that the deterministic $l_{1/q}$ based sparse optimization also gives the maximum a-posteriori estimate for sparse generalized p-Gaussian distributed parameters. This provides the justification for use of the technique in the presence of noise when the gpG model is applicable, and proves the superiority of the $l_{1/q}$ norm as a sparseness measure.

All three algorithms were demonstrated to achieve good, sparse solutions when applied to neuromagnetic image reconstruction, beamforming array design, and seismic deconvolution. These applications demonstrate significant practical engineering use for maximally sparse optimization. The majority of research goals stated in Chapter 1 have been accomplished.

7.2. Future Research

Future related work should include application of the technique to additional signal and image processing problems. Restoration of sparse degraded images is currently an area of continued research interest for the author. Blurred and noisy images from astronomical star photographs, or written text, should be well suited for sparse optimization and merit continued study. It is also felt that with some further investigation a large number of additional DSP problems could be found which could be solved with the algorithms presented here.

For each of the applications presented in this dissertation there are extensions which could be included in future research. For NMI all of the work to date has been with simulated magnetic data. A major experiment needs to be mounted to collect clean evoked response data and demonstrate minimum dipole reconstructions with human source data. The model and algorithms also need to be extended to handle full degrees of freedom for

dipole orientation as well as position. The use of magnetic resonance brain maps to give local cortex normal orientation for dipole orientation constraint needs to be demonstrated. For the deconvolution problem of Chapter 5 it would be instructive to investigate more examples, including 2-D and 3-D cases. For beamforming array design the broadband array merits considerable future attention. Also, methods of modifying the formulation to eliminate the element symmetry requirement would be worthy of study.

A more detailed understanding of the simplex search behavior is needed to explain the algorithm's success. Perhaps bounds on the non-optimality of solutions can be inferred directly from **H** and $\underline{b}$. General algorithm enhancements should include efficient system reinversion procedures to eliminate cumulative pivoting error, particularly in the stochastic search. Additional research into deterministic methods for finding global optima is also anticipated, including investigation into adaptations of some global concave minimization algorithms [20] for signal processing application.

8. BIBLIOGRAPHY

[1] B. Jeffs, R. Leahy and M. Singh, "An Evaluation of Methods for Neuromagnetic Image Reconstruction", *IEEE Trans. Biomed. Eng.*, Vol. BME-34, pp. 713-723, 1987.

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

[3] U. Schwarz, "Mathematical statistical description of the iterative beam removing technique,[method CLEAN]", *Astr. Astrophy.*, Vol. 65, pp. 345-356, 1978.

[4] I. Barrodale and F. Roberts, "Application of mathematical programming to l_p approximation", in *Nonlinear Programming*, J. Rosen et al, Eds., Academic Press, 1970.

[5] A. Pietsch, "Approximation spaces", *J. Approx. Theory*, Vol. 32, pp. 115-134, 1981.

[6] P. Zwart, "Global maximization of a convex function with linear inequality constraints", *Oper. Res.*, Vol. 22, pp. 602-609, 1974.

[7] N. Thoai and H. Tuy, "Convergent algorithms for minimizing a concave function", *Math. Oper. Res.*, Vol. 5, pp. 556-566, 1980.

[8] W. Gray, "Variable norm deconvolution", Ph. D. Thesis, Stanford University, 1979.

[9] R. Wiggins, "Entropy guided deconvolution", *Geophys.*, Vol. 50, pp. 2720-2726, 1985.

[10] D. Donoho, "On minimum entropy deconvolution", in *Applied Time Series Analysis, II*, Academic Press, 1981.

[11] R. Mammone, "Image restoration using linear programming", in *Image Recovery: Theory and Applications*, Ed H. Stark, Academic Press, 1987.

[12] A. Ishimaru and Y-S Chen, "Thining and broadbanding antenna arrays by unequal spacings", *IEEE Trans. Antennas Propagat.*, Vol. AP-13, pp. 34-42, 1965.

[13] R. Leahy and B. Jeffs, "Maximally Sparse Optimization for Array Beamforming and Other Applications," USC-SIPI Report 131, University of Southern California, Los Angeles, CA, submitted to *IEEE Trans. Acoust., Speech, Signal Processing*, 1988.

[14] B. Jeffs, R. Leahy, and M. Singh, "Analysis of Reconstruction Algorithms for Neuromagnetic Imaging Using SQUID-detectors," *Proc. of IEEE ASSP 1986 Digital Signal Processing Workshop*, Chatham, MA, Oct. 20-22, 1986.

[15] R. Leahy, B. Jeffs, M. Singh, and R. Brechner, "Evaluation of Algorithms for a SQUID Detector Neuromagnetic Imaging System," *Proc. of the SPIE conf. on Medical Imaging*, Newport Beach, CA, Feb. 1987.

[16] R. Leahy and B. Jeffs, "Optimal Element Placement in Conformal Beamforming," *Proc. 22nd Asilomar Conf. Signals, Syst. Comp.*, Nov. 1988.

[17] L. Scales, *Introduction to non-linear optimization,* Springer-Verlag, New York, 1985.

[18] D.G. Luenberger, *Linear and Nonlinear Programming*, second edition, Addison-Wesley, Reading, Mass, 1984.

[19] G. Strang, *Linear algebra and its applications*, 2nd Ed., Academic Poress, 1980.

[20] J. Falk and K. Hoffman,"Concave minimization via collapsing polytopes", *Oper. Res.*, Vol. 34, pp. 919-929, 1986.

[21] S. Kirkpatrick, C.D. Gelatt, Jr., M.P. Vecchi, "Optimization by Simulated Annealing," *Science*, Vol. 220, pp. 671-680, 1983.

[22] M.T. Subbotin, "On the Law of Frequency of Errors," *Matematicheskii Sbornik,* Vol. 31, no. 1, pp. 296-301, 1923.

[23] J.H. Miller and J.B. Thomas, "Detector for Discrete-Time Signals in Non-Gaussian Noise," *IEEE Trans. Information Theory*, Vol. IT-18, no. 2, pp. 241-250, Mar. 1972.

[24] T.T. Pham, R.J.P. deFigueiredo, "Maximum Likelihood Estimation of a Class of Non-Gaussian Densities with Application to l_p Deconvolution," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-37, no. 1, pp. 73-82, Jan. 1989.

[25] H. Kruse, *Degeneracy graphs and the neighbourhood problem,* Lecture notes in Econ. Math. Systems, Vol. 260, Springer-Verlag, 1986.

[26] N. Karmarkar, "A New Polynomial-Time Algorithm for Linear Programming," *Combinatorica*, Vol. 4 no. 4, pp. 373-395, 1984

[27] A. Schrijver, *Theory of linear and integer programming*, J. Wiley, 1986.

[28] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images", *IEEE Trans. Patt. Anal. Mach. Int.*, Vol. PAMI-6, pp. 721-741, 1984.

[29] D. Mitra, F. Romeo and A. Sangiovanni-Vincentelli, "Convergence and finite-time behaviour of simulated annealing," *Adv. Appl. Prob.*, Vol. 18, pp. 747-771, 1986.

[30] J. Marroquin, *Probabilistic solutions of inverse problems*, Ph.D. Thesis, MIT, 1985.

[31] R. Duffin, "Linearizing geometric programs", *SIAM Review*, Vol. 12, pp. 211-227, 1970.

[32] R. Duffin, E. Peterson and C. Zener, *Geometric programming*, John Wiley, New York, 1967.

[33] K. Schittkowksi, "NLPQL: A FORTRAN subroutine solving constrained nonlinear programming problems", *Anals Oper. Research*, Vol. 5, pp. 485-500, 1980.

[34] M.J.D. Powell, "A fast algorithm for nonlinearly constrainded optimization calculations," in *Numerical Analysis Proceedings, Dundee 1977, Lecture Notes in Mathematics,* (edited by G.A. Watson, Vol 630, Springer-Verlab, Berlin, Germany, 1978.

[35] S.J. Williamson, G.L. Ramani, L. Kaufman and I. Modena (Eds.), *Biomagnetism: An Interdisciplinary Approach..* Plenum Press, New York, 1982.

[36] D. Brenner, S. J. Williamson, and L. Kaufman, "Visually Evoked Magnetic Fields of the Human Brain," *Science*, Vol. 190, pp. 480-482, 1975.

[37] D. Brenner, J. Lipton, L. Kaufman, and S. J. Williamson, "Somatically Evoked Magnetic Fields of the Human Brain," *Science*, Vol. 199, pp. 81-83, 1978.

[38] M. Singh, D. Doria, V.W. Henderson, G.C. Huth and J. Beatty, "Reconstruction of Images From Neuromagnetic Fields," *IEEE Trans. on Nuclear Science*, Vol. NS-31, pp. 585-589, 1984.

[39] O.D. Kellogg, *Foundation of Potential Theory* , Dover, New York, p 221, 1953.

[40] G.M. Baule and R. McFee, *American Heart Journal*, pp. 66-95, 1963.

[41] D. Cohen, E. Edelsack, and J.E. Zemmerman, *Applied Physics Letters*, Vol. 16, p 278, 1970.

[42] D. Cohen, *Science*, Vol. 175, p 664, 1972.

[43] S.J. Williamson and L. Kaufman, "Biomagnetism," *J. Magnetism and Magnetic Materials*, Vol. 22, pp. 129-201, 1981.

[44] D. Cohen, "Magnetoencephalography: Evidence of Magnetic Fields Produced by Alpha Rhythm Currents," *Science*, Vol. 161, pp. 784-786, 1968.

[45] P. Carelli, I. Modena, G.B. Ricci and G.L. Romani, "Magnetoencephalography," in *Biomagnetism: An Interdisciplinary Approach*, S.J. Williamson, G.L. Ramani, L. Kaufman and I. Modena (Eds.), Plenum Press, New York, pp. 469-482, 1982.

[46] G.L. Romani and R.R. Fenici, "Study of the Human Heart Conduction System by the Biomagnetic Method," U.S.A. - Italy Joint Symposium on Methods of Non-Invasive Diagnosis in Cardiovascular Disease, Bethesda, 12-13 Nov., 1981.

[47] D.S. Barth, W. Sutherling, J. Engel Jr., and J. Beatty, "Neuromagnetic localization of Epileptiform Spike Activity in the Human Brain," *Science*, Vol. 218, pp. 891-894, 1982.

[48] M. Kutas and S.A. Hillyard, "The Lateral Distribution of Event-Related Potentials During Sentence Processing," *Neuropsychologia*, Vol. 5, pp. 579-590, 1982.

[49] L. Kaufman, "Perception and Event-Related Potentials and Fields," in *Biomagnetism: An Interdisciplinary Approach*, S.J. Williamson, G.L. Ramani, L. Kaufman and I. Modena (Eds.), Plenum Press, New York, pp. 385-398, 1982.

[50] P.L. Nunez, *Electric Fields of the Brain*, Oxford University Press, New York, 1981.

[51] Y. Okada, "Neurogenesis of Evoked Magnetic Fields," in *Biomagnetism: An Interdisciplinary Approach*, S.J. Williamson, G.L. Ramani, L. Kaufman and I. Modena (Eds.), Plenum Press, New York, pp. 399-408, 1982.

[52] B.N. Cuffin, "A Comparison of Moving Dipole Inverse Solutions Using EEG's and MEG's," *IEEE Trans. Biomed. Eng.*, Vol. BME-32, pp. 905-910, 1985.

[53] B.N. Cuffin, "Effects of Measurement Errors and Noise on MEG Moving Dipole Inverse Solutions," *IEEE Trans. Biomed. Eng.*, Vol. BME-33, pp. 854-861, 1986.

[54] Y. Okada, "Discrimination of Localized and Distributed Current Dipole Sources and Localized Single and Multiple Sources," in *Biomagnetism: Applications and Theory*, H. Weinberg, G. Stroink, and T. Katila (Eds.), Pergamon, New York, pp. 266-272, 1985.

[55] H. Weinberg, P. Brickett, F. Coolsma, and M. Baff, "Topography of Simulated MEG and EEG generated by Multiple Intracranial Dipoles" in *Biomagnetism: Applications and Theory*, H. Weinberg, G. Stroink, and T. Katila (Eds.), Pergamon, New York, pp. 273-277, 1985.

[56] D.S. Barth, W. Sutherling, J. Broffman, and J. Beatty, "Magnetic Localization of a Dipolar Current Source Implanted in a Sphere and a Human Cranium," *Electroenceph. Clin. Neurophysiol.*, Vol. 63, pp. 260-273, 1986.

[57] Y. Okada, "Somatic Evoked Field," in *Biomagnetism: An Interdisciplinary Approach*, S.J. Williamson, G.L. Ramani, L. Kaufman and I. Modena (Eds.), Plenum Press, New York, pp. 409-421, 1982.

[58] J.H. Tripp, "Physical Concepts and Mathematical Models," in *Biomagnetism: An Interdisciplinary Approach*, S.J. Williamson, G.L. Ramani, L. Kaufman and I. Modena (Eds.), Plenum Press, New York, pp. 409-421, 1982.

[59] D.L. Arthur, E.R. Flynn and S.J. Williamson, "Source Localization of Long Latency Auditory Evoked Magnetic Fields in Human Temporal Cortex," submitted for publication 1986.

[60] T.B. Smith, "Best-fit Multipole Expansions for Fields from Static Currents," *Inverse Problems*, Vol. 1, pp. 173-179, 1985.

[61] R.J. Ilmoniemi, M.S. Hamalainen and J. Knuutila, "The Forward and Inverse Problems in the Spherical Model," *Biomagnetism: Applications and Theory*, H. Weinberg, G. Stroink, and T. Katila (Eds.), Pergamon, New York, pp. 278-282, 1985.

[62] D.B. Geselowitz and W.T. Miller, III, "Extracorporeal Magnetic Fields Generated by Internal Bioelectric Sources," *IEEE Trans. Magnetics*, Vol. MAG-9, no. 3, pp. 392-398, 1973.

[63] Y. Censor, "Finite Series-Expansion Reconstruction Methods," *Proc. of the IEEE*, Vol. 71, pp. 409-418, 1983.

[64] W.J. Dallas, "Fourier Space Solution to the Magnetostatic Imaging Problem," *Applied Optics*, Vol. 24, pp. 4543-4546, 1985.

[65] D. Cohen and H. Hosaka, "Magnetic Field Produced by a Current Dipole," *J. Electrocardiol.*, Vol. 9, pp. 409-417, 1976.

[66] Y.S. Shim and Z.H. Cho, "SVD Pseudoinversion Image Reconstruction," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-29, no. 4, pp. 904-909, 1981.

[67] S.F. Burch, S.F. Gull, and J. Skilling, "Image Restoration by a Powerful Maximum Entropy Method," *Computer Vision, Graphics, and Image Processing*, Vol. 23, pp. 113-128, 1983.

[68] B.R. Frieden, "Restoring with Maximum Likelihood and Maximum Entropy," *J. Optical Soc. of America*, Vol. 62, pp. 511-518, 1972.

[69] T. Elfving, "On Some Methods for Entropy Maximization and Matrix Scaling," *Linear Algebra and its Applications*, Vol. 34, pp. 321-339, 1980.

[70] J. Mendel, *Optimal seismic deconvolution*, Acad. Press, 1983.

[71] R.A. Wiggins, "Minimum Entropy Deconvolution," *Geoexploration*, vol. 16, pp. 21-35, 1978.

[72] P. Jarske, T. Saramaki, S. Mitra, U Neuvo, "On Properties and Design of Nonuniformly Spaced Linear Arrays", *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-36, pp. 372-380, 1988.

[73] T. Frank, J. Kesner and H. Gruen, "Conformal array beam patterns and directivity indices", *J. Acoust. Soc. Am.*, Vol. 63, pp. 841-847, 1978.

[74] R. Streit, "Optimization of discrete arrays of arbitrary geometry", *J. Acoust., Soc. Am.*, Vol. 69, pp. 199-212, 1981.

[75] E. Sullivan, "Side-lobe behaviour of conformal arrays", *J. Acoust. Soc. Am.*, Vol. 71, pp. 402-404, 1982.

[76] S. Prasad and R. Charan, "On the constrained synthesis of array patterns with applications to circular and arc arrays", *IEEE Trans. Antennas Propagat.*, Vol. AP-32, pp. 725-730, 1984.

[77] L. Grffiths and C. Jim, "An alternative approach to linearly constrained beamforming", *IEEE Trans. Antennas Propagat.*, Vol. AP-30, pp. 27-34, 1982.

[78] B.D. Van Veen and K.M. Buckley, "Beamforming: A Versatile Approach to Spatial Filtering," *IEEE ASSP Magazine*, pp. 4-24, April 1988

[79] L.J. Griffiths and K.M. Buckley, "Quiescent Pattern Control in Linearly Constrained Adaptive Arrays," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-35, no. 7, 1987.

[80] R. Hooke and T. Jeeves, "Direct search solution of numerical and statistical problems", *J. ACM.*, Vol. 8, pp. 212-229, 1961.

[81] B. Jeffs, "Neuromagnetic Image Reconstruction," proposal for Ph.D Dissertation, University of Southern California, Electrical Engineering Department, July, 1987.

[82] B. Kalantari and J.B. Rosen, "An Algorithm for Global Minimization of Linearly Constrained Concave Quadratic Functions," Institute of Management Sciences, Operations Research Society of America, 1987.

[83] H. Tuy, "Concave Programming under Linear Constraints," *Dokl. Akad. Nauk SSSR.*, Vol. 159, pp. 32-35, 1964, Translated: *Soviet Math.*, Vol. 5, pp. 1437-1440, 1964.

[84] J.H. McClellan, T.W. Parks, and L.R. Rabiner, "A computer program for designing optimum FIR linear phase digital filters," *IEEE Trans. Audio Electroacoust.*, Vol. AU-21, pp. 506-536, Dec. 1973.

[85] S.W. Autrey, "Approximate Synthesis of Nonseparable Design Responses for Rectangular Arrays," *IEEE Trans. Antennas Propagat.*, Vol. AP-35, no. 8, pp. 907-912, Aug. 1987.

[86] F. Mintzer, "On half-band, third-band, and *N*th-band FIR filters and their design," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-30, pp. 734-738, Oct. 1982.

[87] R.E. Crochiere and L.R. Rabiner, *Multirate Digital Signal Processing*, Prentice-Hall, Englewood Cliffs, New Jersey, 1983.

[88] R.W. Redlich, "Iterative least-squares synthesis of nonuniformly spaced linear arrays," *IEEE Trans. Antennas Propagat.*, Vol. AP-21, pp. 106-108, Jan. 1973

[89] N. Maeda, "Transversal filters with nonuniform tap spacings," *IEEE Trans. Circuits Syst.*, Vol. CAS-27, pp. 1-11, Jan. 1980.

[90] N.M. Mitrou, "Results on nonrecursive digital filters with nonequidistant taps," *IEEE Trans. Acoust., Speech, Signal Processing.*, Vol. ASSP-33, pp. 1621-1624, Dec. 1985.

9. APPENDICES

9.1. Appendix A, Proof of the Fundamental Theorem of $l_{1/q}$ Programming

Theorem 1A

Given a problem of the form (2.2a) or (2.2b), if a solution exists, then a basic solution exists.

Theorem 1A, Proof:

The existence of a solution, or a basic solution, is dependent only on the constraint equation $\mathbf{H}\underline{x}=\underline{b}$, not on the cost functional. Therefore, this portion of the theorem is equivalent to the linear programming case, for which a proof is available in many texts [18].

Theorem 1B

If a global optimal solution to eqn (2.2a) exists, it is a basic solution, and is globally optimal to eqn (1.2) for $\underline{\varepsilon} = \underline{0}$

Theorem 1B, Proof:

1) From eqn (2.2a) above, we have the form

$$\min_{\underline{x}} \; g(\underline{x}) = \sum_{i=1}^{N} |x_i|^{1/q} \; such \; that \; \mathbf{H}\underline{x} = \underline{b} \; , \; q>1 \tag{9.1}$$

We then construct the system

$$\min_{\underline{y}} \; h(\underline{y}) = \sum_{i=1}^{2N} (y_i)^{1/q} \;\; such \;\; that \;\; C\underline{y}=\underline{b} \, , \, \underline{y} \geq \underline{0}, \;\; q>1 \qquad (9.2)$$

where

$$\underline{y} = \begin{bmatrix} \underline{x}^+ \\ \underline{x}^- \end{bmatrix}, C = \begin{bmatrix} \mathbf{H} \mid -\mathbf{H} \end{bmatrix}, \underline{x} = (\underline{x}^+ - \underline{x}^-)$$

2) By construction, any feasible $\underline{y}$ for (9.2) implies an $\underline{x}$ feasible for (9.1). Similarly, $\underline{y}$ basic for (9.2) implies an $\underline{x}$ basic for (9.1).

3) Over $\underline{y} \geq 0$ under, $\sum_{i=1}^{2N}(y_i)^{1/q}$ is strictly concave, and $C\underline{y}=\underline{b}$ defines a convex set.

Any strictly concave functional achieves its global minimum at an extreme point of a convex constraint set. Since a solution is an extreme point iff it is basic [18], then eqn (9.2) is minimized at some basic solution, $\underline{y}_{opt}$ with cost $h(\underline{y}_{opt})$.

4) Given $\underline{y}_{opt} = [\underline{x}^+_{opt}, \underline{x}^-_{opt}]^T$ let $\hat{\underline{x}} = \underline{x}^+_{opt} - \underline{x}^-_{opt}$. We prove by contradiction $\underline{y}_{opt}$ is properly basic. Assume that for a pair of terms in $\underline{y}_{opt}$, $(x^+_{i_{opt}}) \cdot (x^-_{i_{opt}}) \neq 0$, and thus $\underline{y}_{opt}$ is not properly basic. We may form a new vector, $\underline{y}'$ by replacing only these terms with:

$$x'^{+}_{i} = x^+_{i_{opt}} - min(x^+_{i_{opt}}, x^-_{i_{opt}}) \text{ and } x'^{-}_{i} = -min(x^+_{i_{opt}}, x^-_{i_{opt}}) \qquad (9.3)$$

By inspection $\underline{y}'$ is a basic feasible solution to (9.2) and $h(\underline{y}')<h(\underline{y}_{opt})$ since $\underline{y}'$ differs from $\underline{y}_{opt}$ in only two terms, which are both smaller than the corresponding terms in $\underline{y}_{opt}$. This contradicts the optimality of $\underline{y}_{opt}$, so we conclude $(\underline{x}^+_{opt})^T (\underline{x}^-_{opt})=0$.

5) $g(\hat{\underline{x}})=h(\underline{y}_{opt})$ for any optimum solution to (9.2) since due to 4)

$$h(\underline{y}_{opt})=\sum_{i=1}^{2N}(y_{i_{opt}})^{1/q}=\sum_{i=1}^{N}(x^{+}_{i_{opt}})^{1/q}+\sum_{i=1}^{N}(x^{-}_{i_{opt}})^{1/q}=\sum_{i=1}^{N}|\hat{x}_i|^{1/q}=g(\hat{\underline{x}}) \tag{9.4}$$

6) We prove by contradiction that $\hat{\underline{x}}$ is optimum for (9.1). Assume $\hat{\underline{x}}$ is not optimum for eqn (9.1), but some $\ddot{\underline{x}}$ is, with $g(\ddot{\underline{x}})<g(\hat{\underline{x}})$. We could then construct a feasible solution

$\ddot{\underline{y}}=(\ddot{\underline{x}}^{+},\ddot{\underline{x}}^{-})^{T}$, with $\ddot{\underline{x}}^{+}$ containing the positive terms of $\ddot{\underline{x}}$ and $\ddot{\underline{x}}^{-}$ the negative terms. $(\ddot{\underline{x}}^{+})^{T}(\ddot{\underline{x}}^{-})=0$, so as in 5) $g(\ddot{\underline{x}})=h(\ddot{\underline{y}})<h(\underline{y}_{opt})=g(\hat{\underline{x}})$ which contradicts the assumption $\underline{y}_{opt}$ is optimal for (9.2). Therefore, $\hat{\underline{x}}$ is optimal for (9.1).

7) Thus, given $\underline{y}_{opt}$ optimum for (9.2), by 3) it must be basic, 2) implies an $\hat{\underline{x}}$ which is also basic for (9.1), which by 6) is also optimum. **Q.E.D.**

Theorem 1C

If a globally optimal solution to (2.2b) exists, then a properly basic globally optimal solution exists, furthermore, this solution implies $\underline{x}=\underline{x}^{+}-\underline{x}^{-}$ is a globally optimal solution to eqn (1.2) for $\underline{\varepsilon}>\underline{0}$

Theorem 1C, Proof:

1) $\tilde{H}\tilde{\underline{x}}-\tilde{\underline{b}}=\underline{0}, \tilde{\underline{x}}\geq\underline{0}$ defines a convex solution set over which $g(\tilde{\underline{x}})$ is concave. If an optimal solution $\tilde{\underline{x}}^{0}$ exists, then, since $g(\tilde{\underline{x}})$ is (not strictly) concave, there exists a solution $\tilde{\underline{x}}_{opt}$ which is an extreme point, therefore basic, with $g(\tilde{\underline{x}}^{0})=g(\tilde{\underline{x}}_{opt})$.

2) We may use an argument similar to step 4) of Theorem 2B above to prove $(\underline{x}^{+}_{opt})^{T}(\underline{x}^{-}_{opt})=0$ and thus $\tilde{\underline{x}}$ is properly basic.

3) Since $\underline{x}^+$ and $\underline{x}^-$ are not included in the cost computation $g(\tilde{x})$, as in 5) and 6) above $g(\tilde{x}_{opt}) = g(\underline{x})$, $\underline{x} = \underline{x}^+_{opt} - \underline{x}^-_{opt}$, and hence if $\tilde{x}_{opt}$ is an optimal solution to (2.2b), $\underline{x}$ is an optimal solution to (1.2). **Q.E.D.**

Alternate Proof, Fundamental Theorem of $l_{1/q}$ Programming

Proof:

1) Let $\underline{x}$ be a non-basic feasible solution with k non-zero terms, $k > M$. Without loss of generality, reorder $\underline{x}$, **H**, and $\underline{b}$ such that the first k terms are the non-zero terms, then with $\underline{h}_k$ the k th column of **H**

$$x_1\underline{h}_1 + x_2\underline{h}_2 + \cdots x_k\underline{h}_k = \underline{b} \tag{9.5}$$

$$g(\underline{x}) = \sum_{i=1}^{k} (x_i)^{1/q} \quad , \quad x_{k+1} \cdots x_N = 0$$

2) Since Rank (H) $\leq$ M, there are at most M independent vectors in the set $\underline{h}_1, \underline{h}_2, \cdots \underline{h}_k$. Since $k > M$ there exists a non-trivial linear combination of these vectors equal to zero, i.e. there exists an $\underline{\alpha}$ such that $\alpha_1 \underline{h}_1 + \alpha_2 \underline{h}_2 + \cdots \alpha_k \underline{h}_k = 0$ where at least one $\alpha_i > 0$ and at least one $\alpha_j \neq 0$ with $i \neq j$. There are two cases which must be evaluated: where at least one $\alpha_j < 0$ exists, or where all non-zero $\alpha_j > 0$. We set $\alpha_k+1 \cdots \alpha_N = 0$ so $\mathbf{H}\underline{\alpha} = \underline{0}$.

Case 1: at least 1 $\alpha_j < 0$

3) Form two new feasible solutions $\underline{x}_{\varepsilon_+}$ and $\underline{x}_{\varepsilon_-}$ as follows: Any $\underline{x}_\varepsilon$ of the form $\underline{x}_\varepsilon = (\underline{x} - \varepsilon\underline{\alpha})$ is a feasible solution for $|\varepsilon|$ sufficiently small that for all i,

$x_{i\varepsilon} \geq 0$. As we increase ε from zero, at least one term $x_{i\varepsilon}$ will go to zero when $x_i = \varepsilon\alpha_i$ since by 2) at least one $\alpha_i > 0$ and all $x_i \geq 0$. $\underline{x}_{\varepsilon_+}$ is chosen to be the solution where the first term goes to zero and ε_+ is the corresponding value of ε. $\underline{x}_{\varepsilon_-}$ is formed similarly by allowing ε to decrease until a term $x_{j\varepsilon}$ goes to zero at $\varepsilon = \varepsilon_-$, i.e. ε_+ and ε_- are chosen as the smallest positive and negative values respectively such that at least one of the elements $x_{i\varepsilon_+}$ and $x_{j\varepsilon_-}$ are zero for $1 \leq i, j \leq k$. Both $\underline{x}_{\varepsilon_+}$ and $\underline{x}_{\varepsilon_-}$ have k - 1 non zero terms.

4) The cost for any value of ε can be computed as:

$$g(\underline{x}_\varepsilon) = \sum_{i=1}^{k} (x_i - \varepsilon\alpha_i)^{1/q} \tag{9.6}$$

Note that the sum is taken only over the terms that were non-zero in $\underline{x}$. Treat this as a scalar valued function of ε, $C(\varepsilon) = g(\underline{x}_\varepsilon)$ defined over the interval $[\varepsilon_-, \varepsilon_+]$.

5) We may compute the second derivative of $C(\varepsilon)$ over the open interval $(\varepsilon_-, \varepsilon_+)$,

$$\frac{d^2C(\varepsilon)}{d\varepsilon^2} = \frac{1-q}{q^2} \sum_{i=1}^{k} \alpha_i^2 (x_i - \varepsilon\alpha_i)^{1/q - 2} \tag{9.7}$$

Also over this interval $(x_i - \varepsilon\alpha_i)$ is strictly positive for $1 < i < k$ so

$(x_i - \varepsilon\alpha_i)^{1/q - 2} > 0$

and since

$$\alpha_i^2 > 0, \text{ for some } i, 1 \leq i \leq k, \quad \text{and } \frac{1-q}{q^2} < 0 \text{ for } q > 1$$

it follows that

$$\frac{d^2C(\varepsilon)}{d\varepsilon^2} < 0 \text{ for } \varepsilon_- < \varepsilon < \varepsilon_+$$

6) From 5) we see that $C(\varepsilon)$ has no inflection points and is concave everywhere over $(\varepsilon_-, \varepsilon_+)$. Since $C(\varepsilon)$ is continuous over the closed interval $[\varepsilon_-, \varepsilon_+]$ and due to the concavity over the open interval, min $C(\varepsilon)$ occurs in the limit as ε approaches

either ε_+ or ε_- . But at these values, $\underline{x}_\varepsilon$ has only k - 1 non-zero terms. Thus for any case 1 feasible solution with $k > M$ non-zero terms and cost $C(\underline{x})$, there exists a solution at $\underline{x}_{\varepsilon_+}$ or $\underline{x}_{\varepsilon_-}$ with k - 1 non-zero terms and cost $C(\underline{x}_{\varepsilon_{+,-}}) < C(\underline{x})$.

Case 2: all non-zero $\alpha_j > 0$

7) Again we form $\underline{x}_{\varepsilon_+}$ and find ε_+ as in step 3) but we note that for negative values of ε, with all elements of $\underline{\alpha} \geq 0$, there is no ε_- which will drive an element of $\underline{x}$ to zero. We define an open interval, $(-\infty, \varepsilon_+)$, over which $(x_i - \varepsilon \alpha_i) > 0$

8) Differentiating with respect to ε,

$$\frac{dC(\varepsilon)}{d\varepsilon} = \frac{-1}{q} \sum_{i=1}^{k} \alpha_i (x_i - \varepsilon \alpha_i)^{1/q - 1} \qquad (9.8)$$

and since in case 2, at least two elements α_i, $\alpha_j > 0$ exist, with no negative elements, we conclude that $\frac{dC(\varepsilon)}{d\varepsilon}$ is strictly negative over $(-\infty, \varepsilon_+)$.

9) From 8) we see that $C(\varepsilon)$ is a strictly decreasing function over the open interval, so due to continuity we conclude that for case 2, the minimum value is reached in the limit at ε_+, i.e. at $\underline{x}_{\varepsilon_+}$ which has k - 1 non-zero elements.

10) Assume $\underline{x}$ is a non-basic optimum solution. By repeated application of 6) or 9) above, we can find a basic solution which has lower cost. This contradiction implies the optimum solution cannot be non-basic. Q. E. D.

9.2. Appendix B, Proof of Equivalence Theorem for $l_{1/q}$ Optimization

Theorem 2: Let S denote the set of all basic feasible solutions to $\mathbf{H}\underline{x} = \underline{b}$ s.t. $\mathbf{H} \in R^{M \times N}$. If the solutions in S are bounded, then $\Omega = \max_{\underline{x} \in S} [l_\infty(\underline{x})]$ is finite.

Let $\varepsilon = \min_{x_i \in S, 1 \le j \le N} \left\{ |x_{ij}| \; s.t. \; x_{ij} \neq 0 \right\}$. i.e. ε is the smallest non-zero magnitude of any element of any vectors in S.

Given $\varepsilon > 0$ and $\Omega < \infty$, if V is the set of all globally optimal solutions to

$$\min_{\underline{x}} f(\underline{x}) = \sum_{i=1}^{N} \mathbf{I}(x_i) \quad \text{such that} \quad \mathbf{H}\underline{x} = \underline{b} \tag{2.3}$$

with $r = f(\underline{x})$ for any $\underline{x} \in V$ (i.e. r is the optimal solution order), and U is the set of all globally optimal solutions to

$$\min_{\underline{x}} g(\underline{x}) = \sum_{i=1}^{N} |x_i|^{1/q} \quad \text{such that} \quad \mathbf{H}\underline{x} = \underline{b}, \; q > 1 \tag{2.4}$$

then if $q \ge q_1$ U is a subset of V, where

$$q_1 = \frac{\log\left(\frac{\Omega}{\varepsilon}\right)}{\log\left(\frac{r+1}{r}\right)} \tag{2.5}$$

Proof:

1) By Theorem 1A, and since any non-basic solution to eqn (2.3) is of higher order than a basic solution, $V \subset$ S. Also, by Theorem 1B, $U \subset$ S.

2) Since $|x_j| \le \Omega$, $1 \le j \le N$ for all $\underline{x} \in$ S, $g(\underline{x}) \le r\,\Omega^{1/q}$ for all $\underline{x} \in V$

3) Since the order of any basic solution, $\underline{x}'$, not in V is $\geq r+1$, and $|x'_j| \geq \varepsilon$, $1 \leq j \leq N$, for $x'_j \neq 0$, $g(\underline{x}') \geq (r+1)\ \varepsilon^{1/q}$ for all $\underline{x}' \in S \cap \overline{V}$ (where $\overline{V}$ is the complementary set of V).

4) If for some q, $r\Omega^{1/q} < (r+1)\ \varepsilon^{1/q}$ then by 2) and 3), $g(\underline{x}) < g(\underline{x}')$ for all $\underline{x} \in V$ and all $\underline{x}' \in S \cap \overline{V}$, then by 1), solutions minimizing $g(\underline{x})$ must be contained in V, i.e. $U \subset V$.

5) Solving for q in 4):

$$\frac{r+1}{r}\left(\frac{\varepsilon}{\Omega}\right)^{1/q} > 1 \tag{9.9}$$

$$log\left(\frac{r+1}{r}\right) + \frac{1}{q} log\left(\frac{\varepsilon}{\Omega}\right) > 0$$

$$q > \frac{log\left(\frac{\Omega}{\varepsilon}\right)}{log\left(\frac{r+1}{r}\right)} = q_1$$

and since $0 < r,\ \varepsilon, \Omega < \infty$, q_1 is finite. **Q.E.D.**

Note that this is a sufficient (but not necessary) condition which gives a conservative upper bound on q and that in practice a much smaller value may be used. Even though the equations may contain unbounded or very large finite solutions, the application itself may suggest a more realistic value for Ω. This smaller Ω will yield a smaller q which will minimize the objective, $g(\underline{x})$, for the lowest order realistic solution, but will return a higher cost for lower order, yet unrealistic solutions. For example, in the beamforming

problem, we know that for a stable shading design with a normalized MRA response of 1.0, any single shade cannot be significantly larger than N. Mathematically feasible, yet practically unrealistic solutions of low order can be rejected by using the "common sense" value for $\Omega \sim N$. Likewise, a realistic value for ε can be inferred from the measurement and quantization noise in our problem.

9.3. Appendix C, Proof of Convexity for Modified $l_{1/q}$ Cost

Consider the cost functional defined by eqn (6.12). It is always possible, as shown in Appendix B, to transform the associated bipolar linear system to an equivalent system with twice as many positive only variables, so (6.12) becomes

$$g'(\underline{w}) = \sum_i \left(\sum_j w_{ij} \right)^{1/q} \qquad q>1, \quad w_{ij} \geq 0 \tag{9.9}$$

where the indices i and j of the column vector $\underline{w}$ correspond to the array element and filter tap indices respectively. We wish to know if this is a concave function in order to invoke the arguments of Theorem 2. We will show that the Hessian matrix, $\boldsymbol{H}$, is negative semi-definite, which implies $g'(\underline{w})$ is concave.

$$H_{ij,i'j'} = \frac{\partial^2 g(\underline{w})}{\partial w_{ij} \partial w_{i'j'}} = \begin{cases} \frac{1-q}{q^2} \left(\sum_j w_{ij} \right)^{1/q\,-2} & for\ i=i' \\ 0 & for\ i \neq i' \end{cases} \tag{9.10}$$

Now, $\frac{1-q}{q^2} < 0$ for $q > 1$, and $\left(\sum_j w_{ij}\right)^{1/q\,-1} \geq 0$ for all i, so $H_{ij,i'j'} \leq 0$. If we order, without loss of generality, the elements of $\underline{w}$ such that all w_{ij} with a common i value are contiguous, then $\boldsymbol{H}$ takes a block diagonal form, with each block consisting of a single repeated negative value:

$$H = \begin{bmatrix} a \cdots a & & & \\ \vdots \quad \vdots & & 0 & \\ a \cdots a & & & \\ & b \cdots b & & \\ & \vdots \quad \vdots & & \\ & b \cdots b & & \\ & & \ddots & \\ & & c \cdots c & \\ 0 & & \vdots \quad \vdots & \\ & & c \cdots c & \end{bmatrix} \tag{9.11}$$

$$\text{with} \quad a = \frac{1-q}{q^2}\left(\sum_j w_{1j}\right)^{1/q\,-2} \le 0, \quad b = \frac{1-q}{q^2}\left(\sum_j w_{2j}\right)^{1/q\,-2} \le 0, \quad etc.$$

Each submatrix is negative semidefinite since $\underline{y}^T \begin{bmatrix} a \cdots a \\ \vdots \quad \vdots \\ a \cdots a \end{bmatrix} \underline{y} = a\,(\sum_i y_i)^2 \le 0$ for all real $\underline{y}$. Due to the block diagonal structure of H, it follows that $\underline{z}^T H \underline{z} \le 0$ for all real $\underline{z}$ since there is no cross-coupling between elements in $\underline{z}$ which multiply one submatrix, and those that multiply another. Therefore H is negative semi-definite, and $g'(\underline{w})$ is concave over $\underline{w} \ge \underline{0}$. **q.e.d.**