An Affine Scaling Methodology for Best Basis Selection

Bhaskar D. Rao, *Senior Member, IEEE*, and Kenneth Kreutz-Delgado, *Senior Member, IEEE*

***Abstract*— A methodology is developed to derive algorithms for optimal basis selection by minimizing diversity measures proposed by Wickerhauser and Donoho. These measures include the p-norm-like $(\ell_{(p\leq 1)})$ diversity measures and the Gaussian and Shannon entropies. The algorithm development methodology uses a factored representation for the gradient and involves successive relaxation of the Lagrangian necessary condition. This yields algorithms that are intimately related to the Affine Scaling Transformation (AST) based methods commonly employed by the interior point approach to nonlinear optimization. The algorithms minimizing the $\ell_{(p\leq 1)}$ diversity measures are equivalent to a recently developed class of algorithms called FOCal Underdetermined System Solver (FOCUSS). The general nature of the methodology provides a systematic approach for deriving this class of algorithms and a natural mechanism for extending them. It also facilitates a better understanding of the convergence behavior and a strengthening of the convergence results. The Gaussian entropy minimization algorithm is shown to be equivalent to a well-behaved $p = 0$ norm-like optimization algorithm. Computer experiments demonstrate that the p-norm-like and the Gaussian entropy algorithms perform well, converging to sparse solutions. The Shannon entropy algorithm produces solutions that are concentrated but are shown to not converge to a fully sparse solution.**

I. Introduction

RECENTLY, there has been a great deal of interest in finding efficient representations of signals [1]–[6]. Of particular interest to us is the approach of using an overcomplete dictionary to represent a signal [7]–[11]. The motivation for such an approach is that a minimal spanning set of basis vectors is usually only adequate to efficiently represent a small class of signals while forming an overcomplete dictionary using a carefully chosen set of redundant basis vectors that can represent a larger class of signals compactly. Popular dictionaries used are the Wavelet and Gabor dictionaries, among others [7], [12]. The problem of basis selection, i.e., choosing a proper and succinct subset of vectors from the dictionary, naturally arises in this case, and developing algorithms for optimal basis selection is a subject of current research.

A sequential basis selection method called the matching pursuit method was developed in [7]. This method is computationally simple and quite effective. However, because the algorithm is greedy, there are situations where the basic algorithm does not result in effective sparse representations [7], [13], [14]. Another effective approach to basis selection was developed in [8] and [9] in the context of special dictionaries, wavelet packets, and cosine packets. An entropy-based measure of sparsity was used to choose the optimal basis, and an efficient algorithm was developed, exploiting the special structure in the dictionary vectors. The general problem of basis selection was addressed in [10] and [11], wherein an ℓ_1 norm measure was used as a measure of sparsity. Basis vectors were chosen that resulted in a representation with the smallest ℓ_1 norm, and the method was shown to be quite effective.

Interestingly, the problem of basis selection arises in many other applications, and researchers in other areas have also attempted to define diversity measures and to compute sparse/concentrated solutions based on minimizing them [15]–[19]. The use of the term "diversity" in this paper refers to a measure of antisparsity and is consistent with the terminology used in several research areas [33]. Note that minimizing diversity (antisparsity) is equivalent to maximizing concentration (sparsity). Our own interest in this problem was initially motivated by the biomagnetic imaging problem [20]. Basis selection has applications to linear inverse problems where the solution is known or required to be sparse, e.g., speech coding [21], bandlimited extrapolation and spectral estimation [22], [23], direction-of-arrival estimation [16], [24], functional approximation [25]–[27], failure diagnosis [28], and pattern recognition for medical diagnosis [19]. We can exploit the advances in these disparate areas to develop effective solutions to the best basis selection problem. It is clear that an effective solution to this problem has wide ranging consequences.

In this paper, we use the p-norm-like $(\ell_{(p\leq 1)})$ diversity measures and the Gaussian and Shannon entropy diversity measures proposed in [9] and [10] as the starting point for developing optimal basis selection methods. A novel methodology is developed and employed to minimize these sparsity measures and to develop algorithms for basis selection. An important outcome of this work is that it provides a deeper understanding of a class of algorithms called FOCal Underdetermined System Solver (FOCUSS), which was recently developed in [16] and [24]. An intuitive and informal approach was used in [16] and [24] to develop these algorithms, and their usefulness was then justified by applications and subsequent analysis. In this paper, we develop a formal and systematic framework for deriving and justifying them as p-norm-like diversity measure minimizers. In addition, our methodology provides a natural mechanism for deriving similar algorithms

Manuscript received June 7, 1997; revised July 7, 1998. This work was supported in part by the National Science Foundation under Grants MIP-922055 and IRI-9202581. The associate editor coordinating the review of this paper and approving it for publication was Dr. Henrique Malvar.

The authors are with the Electrical and Computer Engineering Department, University of California, San Diego, La Jolla, CA 92093-0407 USA (e-mail: brao@ece.ucsd.edu; kreutz@ece.ucsd.edu).

Publisher Item Identifier S 1053-587X(99)00148-8.

starting from other diversity measures. For example, we develop a variant of the FOCUSS algorithm to minimize the Shannon entropy diversity measure of [9]. Another important contribution of the paper is that the algorithms are shown to be equivalent to Affine Scaling Transformation (AST) algorithms, which have recently received attention in the literature on interior point optimization methods [29]–[31].

The outline of the paper is as follows. In Section II, we present the p-norm-like ($p \leq 1$, including p negative) diversity measures and the Gaussian and Shannon entropy measures of sparsity proposed in [9] and [10]. We then define the best basis selection problem as the problem of minimizing these measures subject to the constraint that the signal vector has a basis representation. In Section III, a methodology to derive an iterative algorithm that selects a sparse representation by minimizing the p-norm-like sparsity measures, excluding (temporarily) the case $p = 0$, is presented. The iterative algorithm is based on successive relaxation of the Lagrangian necessary conditions for a minimum. Additional insight into the resulting algorithm is obtained by interpreting the algorithm as solving a succession of constrained weighted minimum norm problems and as an AST-based gradient descent method. The AST interpretation shows that a natural scaling matrix is associated with a choice of value for p. In Section III-C and Appendix A, a detailed convergence analysis of the algorithm is performed, expanding the scope of the convergence results previously prescribed in [24] and [32]. In Section IV, we focus on the case $p = 0$. We show that the p-norm-like algorithm obtained by setting $p \to 0$ and the algorithm obtained from minimizing the Gaussian entropy are identical and argue that this algorithm effectively minimizes the numerosity measure described in [10]. In Section V, we develop an algorithm to minimize the Shannon entropy and discuss why this algorithm does not fully converge to a completely sparse solution. Section VI gives computer simulations comparing the performance of the sparse basis selection algorithms developed in the paper. Finally, conclusions are given in Section VII.

II. Problem Formulation

The problem of basis selection can be formulated as a problem of finding a sparse solution to an underdetermined system of equations [11], [14]. Let A be an $m \times n$ matrix formed using the basis vectors from the dictionary. Since we have an overcomplete dictionary, $m < n$, and it is assumed that $\text{rank}(A) = m$. Denoting the given signal to be represented by b, which is an $m \times 1$ vector representation of b, requires solving for x, which is an $n \times 1$ vector, such that

$$Ax = b. \tag{1}$$

The problem of basis selection and that of efficient representation of b requires that x be sparse, i.e., most of the entries of x are zero. Equation (1) ensures that x is a consistent representation of b, and the sparsity requirement ensures that the solution is concentrated and, hence, an efficient representation.

The representation problem has many solutions. Any solution can be expressed as

$$x = x_{mn} + v$$

where x_{mn} is the minimum 2-norm solution (i.e., solution with the smallest ℓ_2 norm[1] defined as $\|x\|_2^2 = \sum_{i=1}^{n} x[i]^2$) and is given by $x_{mn} = A^+ b$, where A^+ denotes the Moore–Penrose pseudoinverse. The vector v is any vector that lies in $\mathcal{N}(A)$, which is the null space of A. In this case, A has a nontrivial nullspace of dimension $(n-m)$. In many situations, a popular approach has been to set $v = 0$ and to select x_{mn} as the desired solution, e.g., the method of frames [4]. However, the minimum 2-norm criteria favors solutions with many small nonzero entries, which is a property that is contrary to the goal of sparsity/concentration [11], [16]. Consequently, there is a need to define other functionals that, when optimized, lead to sparse solutions.

The question of good diversity measures has been studied in the past, and a good discussion can be found in [9] and [10], and in the literature on linear inverse problems [15], [17], [18]. A popular diversity measure is $E^{(p)}(x)$, where

$$E^{(p)}(x) = \sum_{i=1}^{n} |x[i]|^p, \quad 0 \leq p \leq 1.$$

We extend this class to include negative values of p, leading to the general class of diversity measures

$$\begin{aligned} E^{(p)}(x) &= \text{sgn}(p) \sum_{i=1}^{n} |x[i]|^p, \quad p \leq 1 \\ &= \begin{cases} \sum_{i=1}^{n} |x[i]|^p, & 0 \leq p \leq 1 \\ -\sum_{i=1, x[i] \neq 0}^{n} |x[i]|^p, & p < 0 \end{cases} \end{aligned} \tag{2}$$

where $\text{sgn}(p) = \begin{cases} +1, & 0 \leq p \leq 1 \\ -1, & p < 0 \end{cases}$. The diversity measures $E^{(p)}(x)$ for $0 \leq p \leq 1$ are the general family of entropy-like measures defined in [9] and [10], as well as those discussed in [15] and [17] to compute sparse solutions. The motivation for these diversity measures is that their minimization subject to the constraint (1) results in sparse solutions. Due to the close connection to ℓ_p norms, we refer to these measures as "$\ell_{(p \leq 1)}$ diversity measures" and often, more simply, as the "p-norm-like diversity measures." It is well known that for $p < 1$, ℓ_p is not a true norm [15].

The diversity measure for $p = 0$, which is the *numerosity* discussed in [10], is of special interest because it is a *direct* measure of sparsity, providing a count of the number of nonzero elements of a vector x

$$E^{(0)}(x) = \#\{i : x[i] \neq 0\}.$$

Finding a global minimum to the numerosity measure requires an enumerative search and is NP hard [26]. Consequently, alternate diversity measures that are more amenable to optimization techniques are of interest. The $E^{(p)}(x)$ measures for $p \leq 1$, $p \neq 0$ are useful candidate measures in this context and are indirectly related to sparsity in that *when minimized*,

[1] For simplicity, by default, $\|\cdot\|$ will denote the 2-norm, and all other norms will be explicitly indicated.

they yield sparse solutions. However, these measures have the disadvantage that they can have many local minima, which can result in optimization algorithms converging to suboptimal solutions, i.e., solutions with more nonzero entries than absolutely necessary. This problem can be alleviated somewhat with the use of a good initial condition, which is likely to be available in engineering applications. For a more detailed discussion of these diversity measures for $0 \leq p \leq 1$, see [9], [10], and [15]. Additional discussion can be found in [33]. At this juncture, to avoid potential confusion, it is useful to note that minimization of $E^{(p)}(x)$ is considerably different from the standard ℓ_p optimization problem $\min_x ||Ax-b||_p^p$, $p \geq 1$ [34].

The diversity measures $E^{(p)}(x)$ for $p < 0$ are also good (indirect) concentration measures. For example consider $p = -1$,

$$E^{(-1)}(x) = -\sum_{i=1}^{n} \frac{1}{|x[i]|}.$$

$E^{(-1)}(x)$ will be minimized by making the entries of x small, thereby encouraging sparsity.

Many other diversity measures can be defined [33], [35]. We only examine here the Shannon entropy and Gaussian entropy, which are two other diversity measures described in [8]–[10]. The Shannon entropy diversity measure $H_S(x)$ is defined as

$$H_S(x) = -\sum_{i=1}^{n} \tilde{x}[i] \ln \tilde{x}[i], \quad \text{where } \tilde{x}[i] = \frac{|x[i]|^2}{||x||^2}. \tag{3}$$

The Gaussian entropy diversity measure $H_G(x)$ is defined as

$$H_G(x) = \sum_{i=1}^{n} \ln |x[i]|^2. \tag{4}$$

III. $\ell_{(p\leq 1)}$ Concentration Measures

In this section, we develop a novel methodology for deriving algorithms to minimize the $\ell_{(p\leq 1)}$ class of diversity measures defined by (2) subject to the linear constraint (1). For now, we exclude the case $p = 0$; the details pertaining to this special case are provided in Section IV. The algorithm is derived using successive relaxation of the Lagrangian necessary conditions in Section III-A. Interestingly, the approach turns out to be a systematic procedure for deriving a class of algorithms called FOCUSS developed in [24]. Additionally, the methodology employed provides a mechanism for generalizing and deriving FOCUSS-like algorithms to other situations. More insight into the methodology is obtained by interpreting the approach as a method of solving successive weighted constrained minimum norm problems. In Section III-B, we recast the algorithm as an affine scaling transformation (AST)-based gradient-descent optimization method. Sections III-A and B provide the foundation for a general affine scaling methodology for deriving best basis selection algorithms. In Section III-C and Appendix A, the convergence behavior of the algorithm is studied.

A. Algorithm Derivation

To minimize the $\ell_{(p\leq 1)}$ diversity measures subject to the equality constraints (1), we start with the standard method of Lagrange multipliers (see, e.g., [29], [36], and [37]). Define the Lagrangian $L(x, \lambda)$

$$L(x, \lambda) = E^{(p)}(x) + \lambda^T (Ax - b)$$

where λ is the $m \times 1$ vector of Lagrange multipliers. A necessary condition for a minimizing solution x_* to exist is that (x_*, λ_*) be stationary points of the Lagrangian function, i.e.,

$$\begin{aligned} \nabla_x L(x_*, \lambda_*) &= \nabla_x E^{(p)}(x_*) + A^T \lambda_* = 0 \\ \nabla_\lambda L(x_*, \lambda_*) &= Ax_* - b = 0. \end{aligned} \tag{5}$$

The gradient of the diversity measure $E^{(p)}(x)$ with respect to element $x[i]$ can be readily shown to be

$$\nabla_{x[i]} E^{(p)}(x) = |p||x[i]|^{p-2} x[i].$$

Substituting this in (5) results in a nonlinear equation in the variable x with no simple solution being evident. To remedy the situation, we suggest using a particular *factored representation* for the gradient vector of the diversity measure, i.e.,

$$\nabla_x E^{(p)}(x) = \alpha(x) \Pi(x) x \tag{6}$$

where $\alpha(x) = |p|$, and $\Pi(x) = \text{diag}(|x[i]|^{p-2})$.[2] From (5) and (6), the stationary points satisfy

$$\alpha(x_*) \Pi(x_*) x_* + A^T \lambda_* = 0 \quad \text{and} \quad Ax_* - b = 0. \tag{7}$$

Note that for $p \leq 1$, $\Pi^{-1}(x_*) = \text{diag}(|x[i]|^{2-p})$ exists for all x. From (7), we have

$$x_* = -\frac{1}{\alpha(x_*)} \Pi^{-1}(x_*) A^T \lambda_*. \tag{8}$$

Substituting for x_* in the second equation of (7) and solving for λ_* results in

$$\lambda_* = -\alpha(x_*)(A \Pi^{-1}(x_*) A^T)^{-1} b. \tag{9}$$

Substituting this expression for λ_* in (8) then results in

$$x_* = \Pi^{-1}(x_*) A^T (A \Pi^{-1}(x_*) A^T)^{-1} b. \tag{10}$$

Equation (10) is not in a convenient form for computation as the right side depends on x_*. However, it indicates the condition that the stationary point must satisfy and also suggests the iterative procedure for computing x_* as

$$x_{k+1} = \Pi^{-1}(x_k) A^T (A \Pi^{-1}(x_k) A^T)^{-1} b. \tag{11}$$

The computation of $\Pi^{-1}(x_k) = \text{diag}(|x_k[i]|^{(2-p)})$ for $p \leq 1$ does not pose any implementation problems, even as entries converge to zero (as is desired, the goal is a *sparse* stationary point x_*). Note that if any element $x[i]$ is zero, then the corresponding diagonal term in Π^{-1} is also zero.

[2] More generally, as will be seen in Section V, α is an explicit function of x.

1) Interpretation of Methodology: We now explore related optimization strategies that provide more insight into the approach. The need to solve the difficult nonlinear equation (5) is a standard occurrence in optimization problems. Several creative methods have been developed to address this problem in the nonlinear programming literature [29], [36]–[38]. Of particular interest to us in this context are the methods of finding iterative solutions by a relaxation approach where the Lagrangian is modified, and a related simpler problem is solved. In such iterative methods, one generates a sequence of estimates of the parameter vector x_k, along with estimates of the Lagrange multiplier vector λ_k, which is usually expressed in terms of the current iterate x_k [38]. The feasibility of the vector x_k is usually ensured at each step with the goal being that $x_k \rightarrow x_*$ and $\lambda_k \rightarrow \lambda_*$.

From this perspective, note that the iterative algorithm (11) can be viewed as arising from a successive relaxation of the Lagrangian necessary condition (7) at step $(k+1)$ to

$$\nabla_x L_{k+1}(x,\lambda) = \alpha(x_k)\Pi(x_k)x + A^T\lambda = 0.$$

Then, solving for the stationary point at step $(k+1)$ reduces to solving a linear system of equations of the form

$$\begin{bmatrix} \alpha(x_k)\Pi(x_k) & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} x \\ \lambda \end{bmatrix} = \begin{bmatrix} 0 \\ b \end{bmatrix}$$

which can be easily done in a manner similar to that used in computing (x_*, λ_*) [cf., (7)–(10)]. Note that at each step, x_{k+1} generated from (11) is feasible, i.e., $Ax_{k+1} = b$, as is easily demonstrated

$$\begin{aligned} Ax_{k+1} &= A\Pi^{-1}(x_k)A^T(A\Pi^{-1}(x_k)A^T)^{-1}b \\ &= (A\Pi^{-1}(x_k)A^T)(A\Pi^{-1}(x_k)A^T)^{-1}b \\ &= b. \end{aligned}$$

Given x_k, the sequence of Lagrange multiplier estimates naturally follow from (9) $\lambda_k = -\alpha(x_k)(A\Pi^{-1}(x_k)A^T)^{-1}b$. The convergence of the iterates $x_k \rightarrow x_*$ is proved in Section III-C and Appendix A.

Continuing this scrutiny further, we see that an appropriate approximate Lagrangian at each step $(k+1)$ is given by

$$L_{k+1}(x,\lambda) = \frac{\alpha(x_k)}{2}x^T\Pi(x_k)x + \lambda^T(Ax-b).$$

This approximation corresponds to replacing the problem of minimizing the original cost function $E^{(p)}(x)$ by a sequence of constrained weighted *minimum norm* problems. This is closely related to the algorithms developed in the context of the ℓ_p optimization problem of minimizing $||Ax-b||_p^p$, $p \geq 1$ [34]. In the ℓ_p optimization problem, there are no constraints, and it is customary to deal with an overdetermined system of equations; relaxation of the necessary condition for the minima leads to a sequence of weighted *least-squares* problems. The algorithm developed [cf., (11)] can be viewed as an extension of the methodology to the underdetermined problem.

For this procedure to be sound, it is necessary that the weighting matrix $\Pi(x_k)$ be positive definite at each iteration, which is true for the p-norm-like diversity measures. The contribution of the approach developed lies in the manner at which the weighting matrix is arrived. The weighting matrix $\Pi(x_k)$ is defined by the relationship (6). Due to the simple nature of $\Pi(x_k)$, it may be tempting to believe that an iterated weighted least squares formulation can be obtained directly from the cost function (rather than its gradient). Unfortunately, such an approach to defining the weighting matrix does not easily generalize to more complicated measures. For example, one might propose to use the direct approach to form

$$E^{(p)}(x) = \text{sgn}(p)\sum_{i=1}^{n}|x[i]|^p = \sum_{i=1}^{n}w[i]^2|x[i]|^2$$

where

$$w[i]^2 = \text{sgn}(p)|x[i]|^{p-2}$$

yielding $E^{(p)}(x) = x^T\Pi(x)x$ and a weighting matrix $\Pi(x) = \text{diag}(w[i]^2)$. However, this results in a positive definite weighting matrix for positive p but not for negative p. It is easy to generate other examples with similar negative conclusions. More general evidence that the relationship (6) is a natural and effective way to derive weighting matrices can be found in [33]. The algorithm developed in Section V for minimizing the Shannon entropy also uses the weighting matrix $\Pi(x_k)$ defined by the gradient relationship (6).

B. An AST Connection

The algorithm proposed in the previous subsection is closely related to the AST-based methods used by the interior point approach to solving linear and nonlinear programming problems [29]–[31]. A significant and interesting outcome is that the $\Pi(x)$ matrix defined from the gradient relationship (6) suggests a *natural* affine scaling matrix. To see this connection, let us define the symmetric scaling matrix W by

$$W^{-2}(x) \triangleq \Pi(x) = \text{diag}(|x[i]|^{p-2}) \tag{12}$$

with $W(x) = \text{diag}(|x[i]|^{1-\frac{p}{2}})$. For the $(k+1)$th iteration of the algorithm (11), let the W matrix, now denoted by W_{k+1}, be evaluated at the present solution x_k (i.e., $W_{k+1} = \text{diag}(|x_k[i]|^{1-\frac{p}{2}})$) and used to define a scaled variable q

$$q = W_{k+1}^{-1}x, \quad \text{equivalently } x = W_{k+1}q. \tag{13}$$

With this transformation, the optimization problem in x is transformed to an optimization problem in the scaled variable q, namely

$$\min_q E^{(p)}(W_{k+1}q) \quad \text{subject to } A_{k+1}q = b$$

where A_{k+1} is the rescaled A matrix defined by $A_{k+1} \triangleq AW_{k+1}$.

Following the AST methodology [29]–[31], the gradient with respect to q is projected into the null space of A_{k+1} to obtain a feasible descent direction.[3] The gradient with respect to the scaled variable q is given by

$$\nabla_q E^{(p)}(W_{k+1}q) = W_{k+1}\nabla_x E^{(p)}(x) = \alpha(x)W_{k+1}\Pi(x)x.$$

[3]Projection of the gradient onto the nullspace of a constraint matrix to obtain a feasible descent direction is well-known as the gradient projection method of Rosen [37]. AST methods additionally rescale the constraint matrix at each step [29].

Evaluating this gradient at the current value of the iterate x_k (equivalently at $q'_k = W_{k+1}^{-1}x_k$) and projecting it into the null space of A_{k+1} results in the search direction l_k, where

$$\begin{aligned} l_k &= (I - A_{k+1}^{+}A_{k+1})\alpha(x_k)W_{k+1}\Pi(x_k)x_k \\ &= \alpha(x_k)(I - A_{k+1}^{+}A_{k+1})W_{k+1}W_{k+1}^{-2}W_{k+1}q'_k \\ &= \alpha(x_k)(I - A_{k+1}^{+}A_{k+1})q'_k. \end{aligned}$$

The new solutions q_{k+1} and x_{k+1} are computed as

$$q_{k+1} = q'_k - \mu_k l_k$$

and

$$x_{k+1} = W_{k+1}q_{k+1} = x_k - \mu_k W_{k+1}l_k$$

where μ_k is a positive step size. If μ_k is now chosen to equal $\frac{1}{\alpha(x_k)}$, then

$$q_{k+1} = A_{k+1}^{+}A_{k+1}q'_k = A_{k+1}^{+}AW_{k+1}W_{k+1}^{-1}x_k = A_{k+1}^{+}b \tag{14}$$

and

$$\begin{aligned} x_{k+1} &= W_{k+1}q_{k+1} = W_{k+1}A_{k+1}^{+}b \\ &= W_{k+1}A_{k+1}^{T}\left(A_{k+1}A_{k+1}^{T}\right)^{-1}b \\ &= W_{k+1}^{2}A^{T}\left(AW_{k+1}^{2}A^{T}\right)^{-1}b \end{aligned}$$

which is precisely the iterative procedure given by (11).

A closer examination of the scaling matrix $W(x) = \text{diag}(|x[i]|^{1-\frac{p}{2}})$ is worthwhile. In the standard AST methods, usually, the scaling matrix $W(x) = \text{diag}(|x[i]|) \triangleq X(x)$ is used[4] [29], [30]. For the algorithm suggested here, $\Pi(x)$ naturally defines a scaling matrix $W(x)$ via the relationship (12), which is more generally dependent on the choice of p. For $p = 1$, which corresponds to the ℓ_1 norm, (12) results in the scaling matrix $W(x) = \text{diag}(|x[i]|^{\frac{1}{2}})$, which differs from the commonly used weighting matrix $X(x)$; for $p = 0$, we obtain the standard affine scaling matrix $W(x) = \text{diag}(|x[i]|) = X(x)$, and for $p = -1$, we get the scaling matrix $W(x) = \text{diag}(|x[i]|^{\frac{3}{2}})$.

The AST derivation naturally leads to the following interesting algorithmic interpretation. Examination of (14) shows that at each step of the algorithm, we effectively solve for a minimum 2-norm solution with respect to q, i.e.,

$$\min ||q||_2^2 \quad \text{subject to} \quad A_{k+1}q = b.$$

Having found the minimum 2-norm solution, $q_{k+1} = A_{k+1}^{+}b$, x_{k+1} is then computed as $x_{k+1} = W_{k+1}q_{k+1}$.

The overall algorithm, which is equivalent to the FOCUSS algorithm originally proposed in [16] and [24], can be summarized as

$$\begin{aligned} W_{k+1} &= \text{diag}\left(|x_k[i]|^{1-\frac{p}{2}}\right) \\ q_{k+1} &= A_{k+1}^{+}b, \quad \text{where } A_{k+1} = AW_{k+1} \\ x_{k+1} &= W_{k+1}q_{k+1}. \end{aligned} \tag{15}$$

[4] Actually $X(x) = \text{diag}(x[i])$ is typically used, but this has the same effect as $W = \text{diag}(|x[i]|)$ since W always appears as W^2 in the computation of x.

We emphasize that algorithms (11) and (15) are entirely equivalent becuase they are related by the scaling transformation (13). The algorithms are initialized by a suitably chosen feasible x_0. As mentioned in Section II and experimentally demonstrated in Section VI, the choice of x_0 determines the sparse solution to which the iterations converge; therefore, care must be taken in making this choice. Often, the minimum 2-norm solution has been found to be a useful initial starting point [11], [24]. Note that unlike standard gradient descent algorithms, there is no need to compute a step size at each iteration of (15), which can significantly speed up the computation. For very large scale problems, a direct implementation of (15) can be onerous, and efficient implementations of the algorithm may become necessary. In particular, it is of interest to note that the algorithm (15) has an interpretation as an interior point optimization method [33]; this is a fact that can enable the use of recent breakthroughs in applying interior point methods to large scale problems similar to the $p = 1$ optimization problem described in [12]. Some additional details on initialization and computation can be found in [16], [24], and [39].

C. Convergence Analysis

Having proposed and motivated the $\ell_{(p\leq 1)}$-class of algorithms given by (11) [equivalently by (15)], we now turn to the issue of examining its convergence behavior. The special case of the numerosity measure, corresponding to $p = 0$, needs special attention and is deferred to the next section. A convergence analysis of the FOCUSS class of algorithms was earlier carried out in [32], [24]. However, the convergence analysis in this earlier investigation was limited, and in certain instances, more restrictive conditions were imposed than necessary. We follow the descent-function based analytical path proposed in [24] and [32] and improve on the results. The solution methodologies introduced here also enable a convergence analysis for the optimization of other sparsity measures, such as the Shannon entropy (3). As in [24], the convergence of the algorithm is established with the help of the global convergence theorem [36], [37]. The main result is as follows.

Theorem 1: Starting from a bounded feasible solution x_0, the algorithm (15) minimizes the $\ell_{(p\leq 1)}$ diversity measure and converges almost surely to a relative minimum, which for $p < 1$ is a basic or degenerate basic solution with at most m nonzero entries.

For the analysis, the assumption that we make about the A matrix is that its rank be m. No assumption of independence is required about any m columns selected from A. Here, we only describe some of the key aspects of the analysis, and the details are relegated to Appendix A. The main features of the analysis are as follows.

1) The analysis requires one to first determine a solution set Γ to which the algorithm converges, which in this case is defined as containing the stationary points (10). It is shown that the relevant and interesting stationary points of the algorithm is a subset of Γ, which is denoted by Γ_b, which contains the basic solutions, i.e.,

solutions with no more than m nonzero entries that are obtained by selecting m columns of A and solving the corresponding linear system of equations, if it exists. There are potentially $\binom{n}{m}$ such solutions. The systematic approach used in deriving the algorithm enables a simple and direct proof of the unstable nature of the sparse solutions with more than m entries.

2) The next important step required for establishing convergence is finding a descent function for the algorithm. The descent function for these algorithms is the diversity measure itself, and it is shown in the Appendix that

$$E^{(p)}(x_{k+1}) < E^{(p)}(x_k), \quad x_k \notin \Gamma.$$

In [24], to prove convergence of the FOCUSS class of algorithms, the $\ell_{(p\leq 1)}$ diversity measures were also used as descent functions. However, for $p \leq 1$ and $p \neq 0$, the decrease of the descent function was established under limited conditions [24, Th. 4]. In addition, the convergence result presented in [24, Th. 2] appears to be true only for $p = 0$. We show decrease in the descent function starting from any $x_0 \in R^n$.

3) Another requirement for the convergence analysis is to show that the sequence lies in a compact set. We establish this fact by providing a direct proof of the boundedness of the sequence x_k.

In addition to the convergence analysis, rate of convergence is another important revealing aspect of an algorithm. Results about the rate of convergence are available in [24], where it is shown that the order of convergence is $(2-p)$.

IV. Numerosity and Gaussian Entropy

We now pay special attention to the case where $p = 0$, which, as previously discussed, yields a numerosity measure that exactly counts the number of nonzero entries

$$E^{(0)}(x) = \#\{i : x[i] \neq 0\} = \sum_{i=1}^{n} \mathbf{1}(x[i])$$

where

$$\mathbf{1}(x[i]) = \begin{cases} 1, & x[i] \neq 0 \\ 0, & x[i] = 0 \end{cases}.$$

This is the measure we ideally would prefer to minimize as observed in [9], [10], and [15]. Unfortunately, this function is not directly suitable for minimization as the function is discontinuous in the regions of interest (when any $x[i]$ goes to zero) and has a gradient of zero everywhere else. However, the class of AST algorithms given by (15) [equivalently, by (11)] yields a well-behaved algorithm, even when $p = 0$. Indeed, letting $p = 0$ in (15) yields the basic FOCUSS algorithm of [16] and [24] and involves the use of a well-defined scaling matrix $W(x) = \text{diag}(|x[i]|)$. Although the algorithm (15) is well defined for $p = 0$, the convergence analysis differs somewhat from the $p \neq 0$ analysis described in Section III-C. In [24] and [32], a convergence analysis is given, and it is shown that the basic $(p = 0)$ FOCUSS algorithm minimizes the Gaussian entropy $H_G(x)$ defined by (4). Here, we show that there are even stronger connections, algorithmically and analytically, of the $p = 0$ algorithm to the Gaussian entropy.

Algorithmically, we can consider minimizing directly the Gaussian entropy or the monotonically related (and hence equivalent) cost function $\text{Exp}(H_G(x)) = \prod_{i=1}^{n} |x[i]|^2$. The latter one is preferable if we are interested in a function that is bounded from below. However, the Gaussian entropy is adequate for the discussion to follow.

An AST algorithm can be derived to minimize the Gaussian entropy following along the lines outlined in Section III-B; merely replace $E^{(p)}(x)$ with $H_G(x)$ in the analysis. The only new quantity required is the gradient of $H_G(x)$, which can be readily shown to be

$$\nabla_x H_G(x) = \alpha_G(x) \Pi_G(x) x$$

where $\alpha_G(x) = 2$ and $\Pi_G(x) = \text{diag}(\frac{1}{x[i]^2})$. The scalar factor $\alpha_G(x)$ does not affect the algorithm, and $\Pi_G(x)$ leads to an affine scaling algorithm with a scaling matrix given by $W(x) = \text{diag}(|x[i]|)$. Note that this is same scaling matrix as that obtained by setting $p = 0$ in algorithm (15), which was derived for the minimization of the $\ell_{(p\leq 1)}$ diversity measure assuming $p \neq 0$. A similar algorithmic conclusion is reached if we try to minimize $\text{Exp}(H_G(x)) = \prod_{i=1}^{n} |x[i]|^2$.

Interestingly, the monotonically related functional $\text{Exp}(H_G(x)) = \prod_{i=1}^{n} |x[i]|^2$ provides an analytic connection to the $\ell_{(p\leq 1)}$ diversity measures via the arithmetic-geometric mean inequality [40]

$$\left(\prod_{i=1}^{n} |x[i]|^p \right)^{\frac{1}{n}} \leq \frac{1}{n} \sum_{i=1}^{n} |x[i]|^p.$$

This implies that for all p and $|x[i]| > 0$

$$\left(-\frac{1}{n} E^{(p^-)}(x) \right)^{\frac{1}{p^-}} \leq [\text{Exp}(H_G(x))]^{\frac{1}{2n}} \leq \left(\frac{1}{n} E^{(p^+)}(x) \right)^{\frac{1}{p^+}}$$

where $p^+ \geq 0$ and $p^- \leq 0$. We have equality in the limit $p \to 0$, establishing a connection between the Gaussian entropy and the $\ell_{(p\leq 1)}$ diversity measures, i.e.,

$$e^{\frac{1}{2n} H_G(x)} = \lim_{p \to 0} \left(\frac{1}{n} E^{(p)}(x) \right)^{\frac{1}{p}}. \tag{16}$$

We can also relate the Gaussian entropy to the $\ell_{(p\leq 1)}$ diversity measures via a Taylor series expansion. This follows along the lines of the argument used in [10] to link the numerosity measure to the Shannon entropy. The diversity measures $E^{(p)}$ are continuous and differentiable with respect to p. It can be shown that

$$\frac{dE^{(p)}(x)}{dp} = \sum_{i=1}^{n} |x[i]|^p \ln |x[i]|.$$

Taking the limit as $p \to 0$, we get

$$\left. \frac{dE^{(p)}(x)}{dp} \right|_{p \to 0} = \frac{1}{2} H_G(x).$$

Performing a Taylor series expansion about $p = 0$, we get

$$E^{(p)}(x) \approx E^{(0)}(x) + \frac{p}{2} H_G(x). \tag{17}$$

As p gets small, the diversity measure $E^{(p)}(x)$ begins to behave like the Gaussian entropy (except at sparsity points where the otherwise constant numerosity measure $E^{(0)}(x)$ jumps discontinuously), establishing another point of contact between Gaussian entropy and $E^{(p)}(x)|_{p\to 0}$ minimization.

The convergence of the algorithm (11) for the case $p = 0$ can be proved using the global convergence theorem. In fact, most of the results discussed in Section III-C and derived in Appendix A hold true, except that we need to identify a proper descent function (the almost everywhere constant, otherwise discontinuous, function $E^{(0)}(x)$ being unsuitable). In [24], it is shown that the Gaussian entropy function $H_G(x)$ is a suitable descent function, which is a fact now further supported by the relationships (16) and (17). The fact that $H_G(x)$ is a valid descent function can be shown from (15) using the observation that $||q_k||^2 \leq n$ and the inequality $\log t \leq t-1,\ t>0$. See [24] for details.

V. Shannon Entropy

In this section, we develop an algorithm for minimizing the Shannon entropy diversity measure $H_S(x)$ defined by (3) and discussed in [8]–[10]. The approach follows the steps employed in Section III-A to minimize the $\ell_{(p\leq 1)}$ diversity norms. This necessitates taking the gradient of the diversity measure, which in this case can be shown to equal

$$\nabla_x H_S(x) = \alpha_S(x) \Pi_S(x) x$$

where $\alpha_S(x) = \frac{2}{||x||_2^2}$ and

$$\Pi_S(x) = -\text{diag}(H_S(x) + \ln \tilde{x}[i]) \quad \text{where } \tilde{x}[i] = \frac{|x[i]|^2}{||x||^2}.$$

Retracing the argument given in Section III-A through (11) suggests that we focus on the iteration

$$x_{k+1}^r = \Pi_S^{-1}(x_k) A^T \left(A \Pi_S^{-1}(x_k) A^T\right)^{-1} b.$$

The superscript r is used here because, unlike the p-norm-like case where $\Pi(x)$ is positive definite, $\Pi_S(x)$ is indefinite, calling for some modifications in order to develop an algorithm that provably converges. In the next subsection, we develop some preliminary results for this purpose. Note that we assume the invertibility of $\Pi_S(x)$, which is a generic property.

A. Preparatory Results

In order to derive the modified algorithm, we need the following identities.

Identity 1:

$$x_k^T \Pi_S(x_k) x_k = 0.$$

Proof:

$$\begin{aligned} x_k^T \Pi_S(x_k) x_k &= -x_k^T x_k H_S(x_k) - \sum_{i=1}^n x_k^2[i] \ln \frac{|x_k[i]|^2}{||x_k||^2} \\ &= -||x_k||^2 \left(H_S(x_k) + \sum_{i=1}^n \frac{x_k^2[i]}{||x_k||^2} \ln \frac{|x_k[i]|^2}{||x_k||^2} \right) \\ &= -||x_k||^2 (H_S(x_k) - H_S(x_k)) = 0. \end{aligned}$$

□

Identity 2: Let x_k be feasible, i.e., $Ax_k = b$; then

$$\begin{aligned} \left(x_{k+1}^r\right)^T \Pi_S(x_k) x_k &= x_k^T \Pi_S(x_k) x_{k+1}^r \\ &= \left(x_{k+1}^r\right)^T \Pi_S(x_k) x_{k+1}^r \\ &= b^T \left(A \Pi_S^{-1}(x_k) A^T\right)^{-1} b. \end{aligned}$$

Proof:

$$\begin{aligned} & x_k^T \Pi_S(x_k) x_{k+1}^r \\ &= x_k^T \Pi_S(x_k) \Pi_S^{-1}(x_k) A^T \left(A \Pi_S^{-1}(x_k) A^T\right)^{-1} b \\ &= x_k^T A^T \left(A \Pi_S^{-1}(x_k) A^T\right)^{-1} b \\ &= b^T \left(A \Pi_S^{-1}(x_k) A^T\right)^{-1} b. \end{aligned}$$

The other expressions can be shown in a similar manner. □

Lemma 1:

$$H_S(y) - H_S(x) \leq \frac{y^T \Pi_S(x) y}{||y||^2}.$$

Proof:

$$\begin{aligned} y^T \Pi_S(x) y &= -||y||^2 H_S(x) - ||y||^2 \sum_{i=1}^n \frac{y^2[i]}{||y||^2} \ln \frac{|x[i]|^2}{||x||^2} \\ &\geq ||y||^2 (H_S(y) - H_S(x)). \end{aligned}$$

The last inequality follows from the fact that $\sum_k p_k \log p_k - \sum_k p_k \log q_k \geq 0$ with equality if and only if for all k, we have $p_k = q_k$, where here, p and q are probabilities [9], [41]. □

B. Modified Algorithm

Using the above results, we consider the following form of an algorithm for minimizing the Shannon entropy (3)

$$x_{k+1} = x_k + \mu_k \left(x_k - x_{k+1}^r\right).$$

If x_k is feasible, then it can be readily shown that x_{k+1} is feasible. The increment $x_k - x_{k+1}^r$ then provides a feasible direction of descent, and by proper choice of the step size μ_k, which we explore next, we can ensure that $H_S(x)$ is minimized at each step. By Lemma 1, it is sufficient to select μ_k such that $x_{k+1}^T \Pi_S(x_k) x_{k+1} \leq 0$ as this ensures that $H_S(x_{k+1}) \leq H_S(x_k)$. Examining $x_{k+1}^T \Pi_S(x_k) x_{k+1}$ and simplifying it using the above identities, it can be shown that

$$x_{k+1}^T \Pi_S(x_k) x_{k+1} = -\mu_k(\mu_k + 2) \left(x_{k+1}^r\right)^T \Pi_S(x_k) x_{k+1}^r.$$

If $(x_{k+1}^r)^T \Pi_S(x_k) x_{k+1}^r \leq 0$, then we need to select a value of μ_k such that $\mu_k(\mu_k+2)$ is negative. To obtain a potentially optimum step size, we can choose μ_k to equal the value where $\mu_k(\mu_k+2)$ attains its minimum. This choice leads to the selection of $\mu_k = -1$, resulting in $x_{k+1} = x_{k+1}^r$ and yielding an iteration step equivalent to (11). We should mention that in practice, we have found the term $(x_{k+1}^r)^T \Pi_S(x_k) x_{k+1}^r$ to be always negative. Some reasoning as to why this should be true is available, but a detailed explanation would take us too far astray, and we refrain from doing so. However, lacking a rigorous proof that this term is indeed always negative, for completeness, it is necessary to check and to deal with the case when $(x_{k+1}^r)^T \Pi_S(x_k) x_{k+1}^r > 0$, which we do next.

If $(x_{k+1}^r)^T \Pi_S(x_k) x_{k+1}^r > 0$, then any positive value of μ_k is acceptable. We can choose μ_k optimally by trying to ensure a large decrease in the Shannon entropy $H_S(x)$. For simplicity, we suggest the choice $\mu_k = 1$. The overall algorithm then is

$$x_{k+1} = \begin{cases} x_{k+1}^r, & \left(x_{k+1}^r\right)^T \Pi_S(x_k) x_{k+1}^r \le 0 \\ 2x_k - x_{k+1}^r, & \left(x_{k+1}^r\right)^T \Pi_S(x_k) x_{k+1}^r > 0. \end{cases} \quad (18)$$

C. Stationary Points

Algorithm (18) is guaranteed to decrease the Shannon entropy $H_S(x)$. The stationary points of the algorithm are given by (10) and satisfy the necessary first order condition for a local minimum, viz., that the gradient at x_* is in the range space of A^T. With the gradient proportional to $\Pi(x)x$, an equivalent statement is

$$y_* \triangleq \Pi_S(x_*) x_* = A^T \left(A \Pi_S^{-1}(x_*) A\right)^{-1} b = A^T \beta_*. \quad (19)$$

The nature of the stationary solutions can be ascertained by examining this condition more closely. Examining the ith entry $y_*[i]$ of y_*, we have

$$y_*[i] = -x_*[i] H_S(x_*) - x_*[i] \ln \tilde{x}_*[i].$$

If the solution is sparse, i.e., many of the entries $x_*[i]$ are zero, then the corresponding entries $y_*[i]$ must be zero. However, (19) shows this can only be true if by using linear combination of m columns of A^T, we can find a vector with a large number of zeros. This is, in general, not possible, and therefore, the stationary points of the minimum Shannon entropy solution cannot generally be completely sparse, as our simulations in Section VI demonstrate. However, consistent with the discussion given in [9], they do tend to have a large number of entries with very small (albeit nonzero) amplitudes. An explanation as to why complete sparsity is not attained for the Shannon entropy $H_S(x)$ and possible ways to rectify this situation is given in [33].

VI. Computer Simulations

In order to gain insight into the behavior of the algorithms discussed in this paper, we perform a simulation study of their behavior on a synthetic test case. A random $m \times n$ matrix A is created whose entries are each Gaussian random variables with mean zero and variance 1. The columns of A are then normalized to have a 2-norm of 1. A sparse solution x_s with a specified number of nonzero entries r is then created; the indices, i.e., location, of these r entries is random, and their amplitudes are random. The vector b is then computed as $b = Ax_s$. For convenience in interpreting the results, b and x_s are then suitably rescaled such that $\|b\|_2 = 1$. With a known sparse solution, x_s, which now is at hand to provide a benchmark, the algorithms are run to select the optimal basis vectors (columns of A). The number of vectors chosen are compared with the actual number r used to generate the data. The experiment is repeated 100 times, with algorithm initialized each time by the minimum 2-norm solution $x_0 = A^+ b$. The histogram of the *redundancy index*, which is defined as the ratio of the number of distinct columns chosen by the method to the number of columns actually used to generate the data, is computed. Algorithms with a redundancy index histogram concentrated around 1 indicate a good procedure.

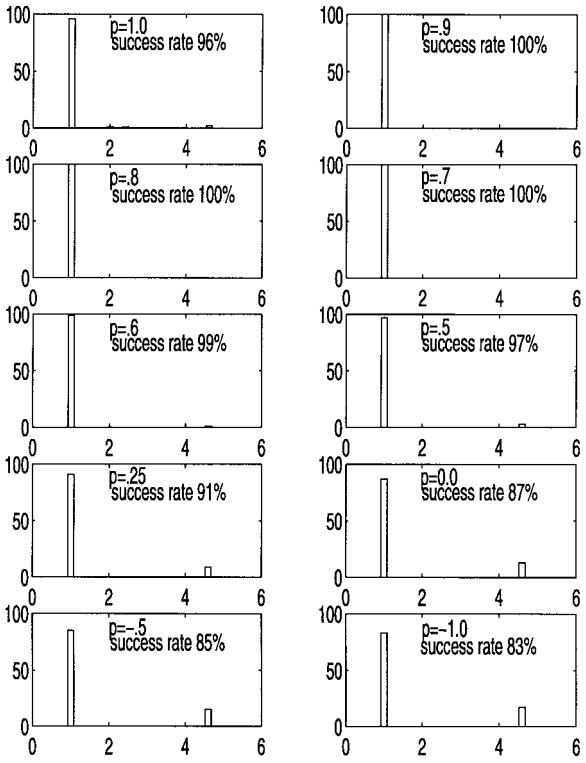


Fig. 1. Performance (histogram of 100 trials) of the $\ell_{(p \le 1)}$ measures are shown for $p = 1, 0.9, 0.8, 0.7, 0.6, 0.5, 0.25, 0, -0.5$, and -1. The parameters used in the simulation are $m = 20$, $n = 30$, sparsity $r = 4$, and the algorithms are initialized by the minimum 2-norm solution $x_0 = A^+ b$. The $p = 0$ algorithm is equivalent to the Gaussian entropy algorithm.

Experiment 1: In this experiment, A is chosen to be a 20×30 matrix, i.e., $m = 20$ and $n = 30$. Figs. 1 and 2 detail the results for sparsity $r = 4$ and $r = 7$, respectively. The p-norm-like diversity measures are optimized for p-values $p = 1, 0.9, 0.8, 0.7, 0.6, 0.5, 0.25, 0, -0.5$ and -1. Recall that the $p = 0$ and the Gaussian Entropy algorithms are equivalent. The histogram results shown are obtained by thresholding the solution obtained at the end of 50 iterations. A threshold of 10^{-8} was used in these simulations, and components of the solution with magnitude less than the threshold are taken to be zero. The number of iterations is chosen to ensure the convergence of the slowest variant of the $l_{(p \le 1)}$ algorithms, which in this case is the ℓ_1 variant. Success rate is defined as the percentage of trials in which the redundancy index was 1. From the simulations, it can be seen that the results are superior when a value of p close to 1 is used. However, they have slower convergence compared with lower values of p and, for $p = 1$, may not be able to reduce the entries sufficiently quickly. In this case, external monitoring procedures may be necessary to null out small entries. The rate of convergence analysis given in [24] indicates that lower values of p have provably faster convergence rate. It may be possible to develop faster algorithms with high reliability by trying to combine the faster convergence behavior of small values of p with the superior basis selection ability of the larger values of p. The

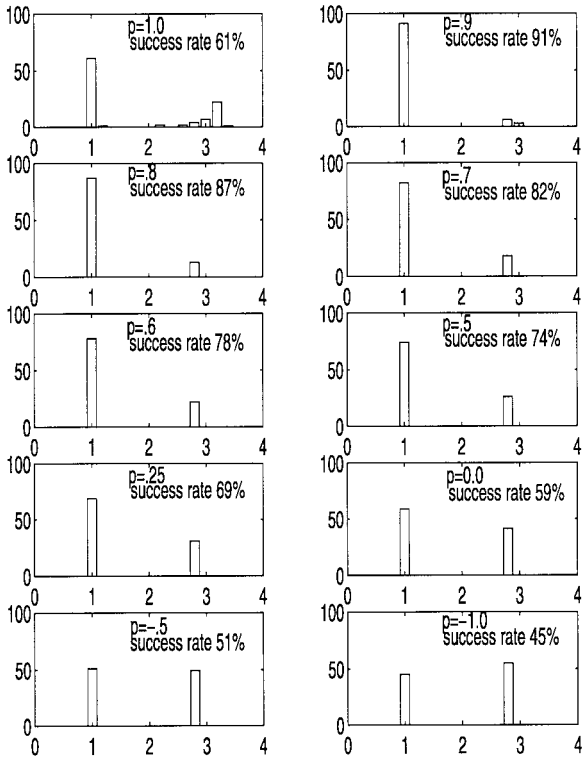


Fig. 2. Performance (histogram of 100 trials) of the $\ell_{(p\leq 1)}$ measures are shown for $p = 1,\ 0.9,\ 0.8,\ 0.7,\ 0.6,\ 0.5,\ 0.25,\ 0,\ -0.5$, and -1. The parameters used in the simulation are $m = 20$, $n = 30$, sparsity $r = 7$, and the algorithms are initialized by the minimum 2-norm solution $x_0 = A^+b$. The $p = 0$ algorithm is equivalent to the Gaussian entropy algorithm.

TABLE I
CORRELATION BETWEEN THE A MATRIX COLUMNS SELECTED BY THE ALGORITHM AND THE COLUMNS PRESENT IN THE TRUE SPARSE SOLUTION ARE TABULATED. THE PARAMETERS ARE $m = 20$, $n = 30$ AND SPARSITY $r = 7$. SUCCESS CORRESPONDS TO THE SEVEN COLUMNS OF THE TRUE SPARSE SOLUTION BEING INCLUDED IN THE SET OF COLUMNS CHOSEN BY THE ALGORITHM

p	Successes	Failures		
		Match 6	Match 5	Match 4
1	91	8	1	0
.9	91	6	3	0
.8	87	9	4	0
.7	82	13	5	0
.6	78	15	7	0
.5	74	19	7	0
.25	69	22	8	1
0	59	31	9	1
-.5	51	35	13	1
-1.0	45	43	12	0

methods employed for l_1 norm minimization in [11], [29], and [30] can be viewed as employing such an approach and are potentially extensible for minimizing the other diversity measures.

To get a better understanding of the performance of the method, we checked the correlation between the columns selected by the algorithms and the actual columns used to generate the true sparse solution. These results are tabulated in Table I for the sparsity $r = 7$ case. The entries under column

TABLE II
CORRELATION OF THE SUCCESSFUL TRIALS BETWEEN VARIOUS $\ell_{(p\leq 1)}$ ALGORITHMS, I.E., COMPUTES THE NUMBER OF TRIALS IN WHICH BOTH ALGORITHMS WERE SUCCESSFUL

p	1	.9	.8	.7	.6	.5	.25	0	-.5	-1
1	91	89	85	81	75	72	67	57	49	44
.9	89	91	87	82	77	74	68	58	50	44
.8	85	87	87	81	76	73	67	57	49	43
.7	81	82	81	82	76	73	67	57	49	44
.6	75	77	76	76	78	73	68	58	51	45
.5	72	74	73	73	73	74	68	58	50	44
.25	67	68	67	67	68	68	69	58	50	44
0	57	58	57	57	58	58	58	59	49	43
-.5	49	50	49	49	51	50	50	49	51	44
-1.0	44	44	43	44	45	44	44	43	44	45

labeled "Match 6" indicate the number of times only six columns of the true solutions were included in the computed solution. At no time was there a match of less than 4 for any of the variants. Here, success is defined as a trial where the solution obtained selected all the seven columns in the true sparse solution, even though the redundancy index may be greater than 1. The correlation of the $l_{(p\leq 1)}$ variants is also examined by comparing the number of trials in which two different diversity measures both lead to the correct choice of including all the seven desired columns. These results are tabulated in Table II. Although there is a strong correlation between the methods, and the diversity measures with p closer to 1 exhibit superior (more reliable) performance, there are trials where diversity measures with lower values of p identified the correct solution, whereas the measure with a larger value of p did not.

Experiment 2: An important feature of choosing p less than 1 is that the diversity measures potentially then have multiple local minima with each minimum having a basin of attraction. The choice of initial condition then decides the minimum attained. The use of different initial conditions to obtain different sparse solutions can be a valuable attribute when sparsity alone is not of paramount importance [24]. For computing sparse solutions, performance can then be improved by multiple reinitialization. The results of reinitialization for sparsity 7 are shown in Fig. 3 for $p = 0$. The histogram of the redundancy index for the procedure initialized by the minimum 2-norm solution has been shown in Fig. 2. Fig. 3(a) shows the redundancy index for $p = 0$ with a random initial condition, and Fig. 3(b)–(d) show the results of repeated initialization. Note that for $p = 0$, it is possible to achieve 100% success by repeated reinitialization. In this experiment, the minimum 2-norm solution was first used, and then, after detecting a failure, random initialization was used. A histogram of the number of initializations required to obtain a successful outcome and the histogram of the total complexity are also shown. Total complexity is measured by taking all the initializations, as well as the number of iterations per initialization, into account. The number of iterations is controlled by examining the q vector in (15). The entries of the q vector in the $p = 0$ case converge to either 0 or 1, enabling easy identification of convergence

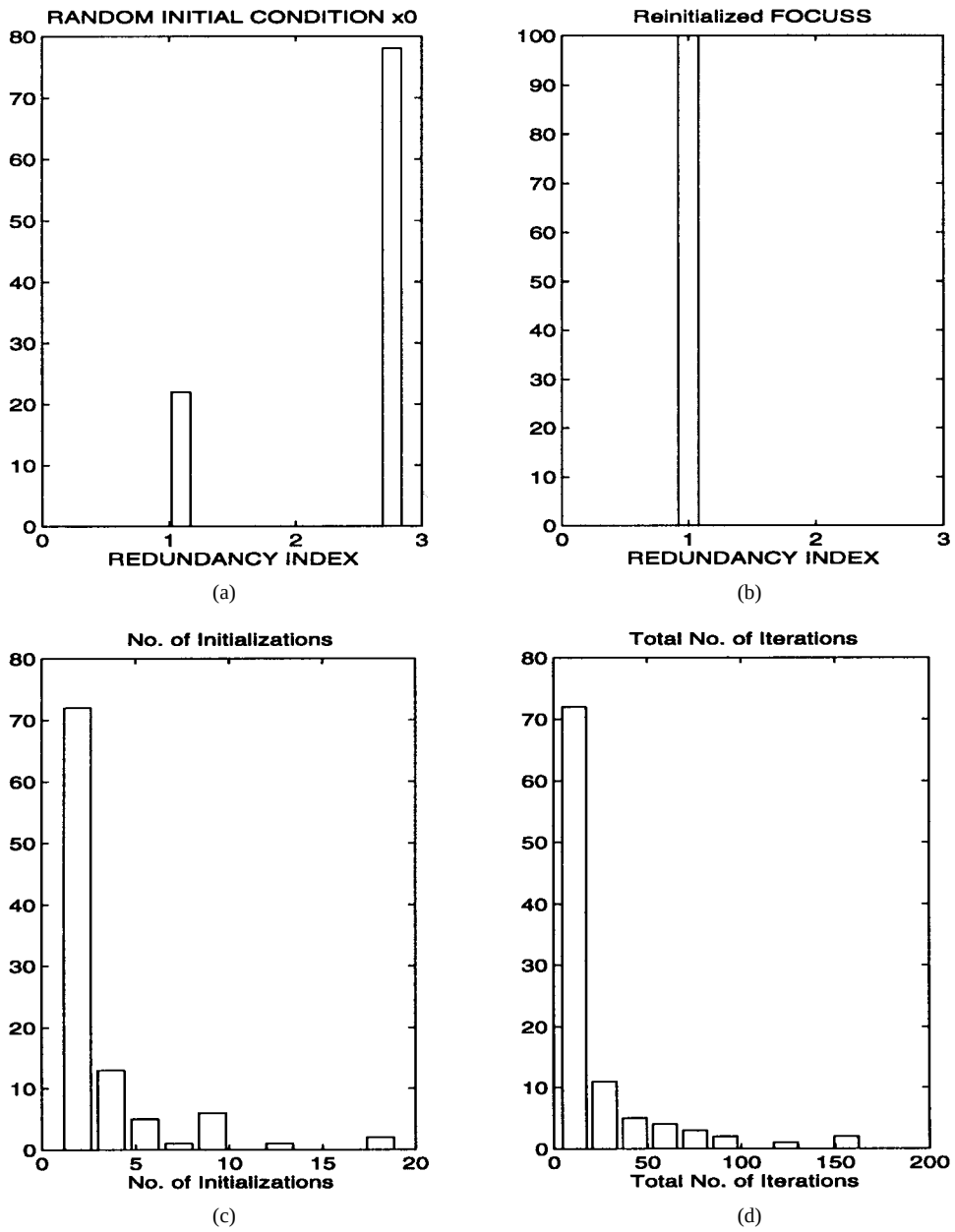


Fig. 3. Experiments with initial conditions and the $p = 0$, p-norm-like algorithm, which is equivalent to the basic FOCUSS algorithm of [24] and the Gaussian entropy algorithm. Parameters are $m = 20$, $n = 30$, and $r = 7$. (a) Success rate with a random initial condition. This can be compared with the $p = 0$ result in Fig. 2, where the minimum 2-norm solution was used for initialization. (b) Success rate after repeated initialization. If the minimum 2-norm initialization failed, thereafter, repeated random initialization was used. (c) Histogram of the number of initializations needed to ensure success. (d) Histogram of the total number of iterations (summing over all reinitializations) needed for eventual success.

[39]. As can be seen, the number of reinitializations needed is usually less than 10, and the total number of iterations is less than 100. A closer examination of the iterations suggest that successful initializations require fewer iterations than the failed ones.

Experiment 3: A similar simulation study of algorithm (18) was conducted for the algorithm developed for minimizing the Shannon entropy. As suggested by the discussion in Section V-C, the solution does not converge to a true sparse solution but does result in concentration. A typical solution obtained by the algorithm, for sparsity $r = 4$ and $r = 7$, and the corresponding true sparse solution are shown in Fig. 4. A base 10 logarithm scale is used in these plots. The nonzero entries of the true sparse solution are denoted by "$*$" and the minimum Shannon entropy solution by dotted lines. As is evident from Fig. 4, the algorithm does produce concentration but not truly sparse solutions, i.e., has many very small (albeit nonzero) amplitudes. To test that the algorithm indeed minimized the entropy measure, we compared the entropy of the converged solution to that of the entropy of the known true sparse solution. Out of 100 trials, the converged solution had lower entropy in 96 trials. In the remaining four trials, the algorithm had converged to a local minima of the entropy function with a value larger than that of the true sparse solution.

In summary, the simulations provide interesting insight into the algorithms and provide support to the theoretical analysis.

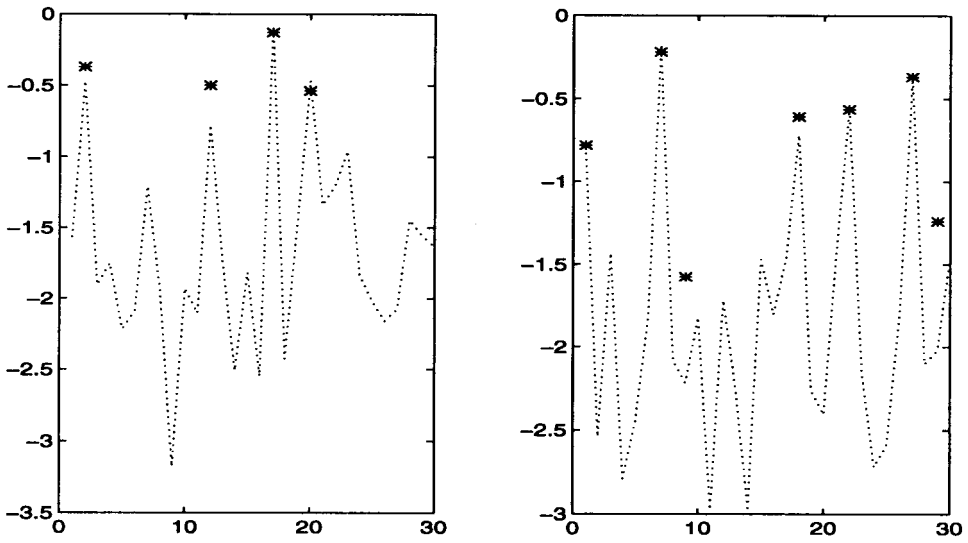


Fig. 4. Typical results (in $\log 10$ scale) obtained using the Shannon entropy algorithm, which is denoted by the dotted line, are shown for (a) sparsity $r = 4$ and (b) for sparsity $r = 7$. The nonzero entries of the true sparse solution are denoted by "$*$."

More extensive application-based study is still necessary to understand the methods more fully. Some results in this context are already available [11], [24]. In particular, details of the application of FOCUSS to the biomagnetic imaging problem can be found in [20] and [24]. We expect to conduct additional application-based evaluations in the near future and are optimistic that this work will stimulate other researchers to also conduct such experiments.

VII. Conclusion

We have developed a novel methodology to develop algorithms for best basis selection. The procedure yields algorithms that are intimately related to the affine scaling transformation (AST)-based methods commonly employed by the interior point approach to nonlinear optimization. The methodology is quite general and is used to develop effective algorithms to minimize several well-known diversity measures, e.g., the p-norm-like diversity measures and the Gaussian and Shannon entropies proposed in [9] and [10]. A detailed convergence analysis of the algorithm for minimizing $\ell_{(p\leq 1)}$ diversity measures, which are equivalent to the FOCUSS-class of algorithms, is conducted, showing them to be provably convergent. Both the theoretical evidence and the computer simulations show the algorithms developed to be quite effective and promising for optimal basis selection. Generalizations of the algorithms and results presented here can be found in [33].

Appendix A

In this Appendix, we examine the convergence behavior of the algorithm (11) [equivalently (15)] developed to minimize the $\ell_{(p\leq 1)}$ diversity measures. These algorithms are equivalent to the FOCUSS-class of algorithms [24], [32]. As noted in Section III-C, a convergence analysis of FOCUSS can be found in [24] and [32]. However, the convergence analysis in this earlier investigation was limited, and in certain instances, more restrictive conditions were imposed than necessary. We improve on the results and, for brevity, concentrate on the generalizations/simplifications that are facilitated by the new systematic framework employed in this paper. The convergence analysis is based on the global convergence theorem (GCT) discussed in [36] and [37]. We first state the GCT for completeness before conducting the analysis.

Theorem 2 (Global Convergence Theorem) [36], [37]: Let $\mathcal{A}$ be an algorithm on a set X, and suppose that, given x_0, the sequence $\{x_k\}_{k=0}^{\infty}$ is generated, satisfying

$$x_{k+1} = \mathcal{A}(x_k).$$

Let a solution set $\Gamma \subset X$ be given, and suppose the following.

1) All points x_k are contained in a compact set $S \subset X$.
2) There is a continuous function (the descent function) Z on X such that
   a) if $x \notin \Gamma$, then $Z(y) < Z(x),\ \forall y \in \mathcal{A}(x)$;
   b) if $x \in \Gamma$, then $Z(y) \leq Z(x),\ \forall y \in \mathcal{A}(x)$.
3) The mapping $\mathcal{A}$ is closed at points outside Γ.

Then, the limit of any convergent subsequence of x_k is a solution, and $Z(x_k) \to Z(x_*)$ for some $x_* \in \Gamma$.

The convergence analysis is now carried out by showing that the sequence generated by the algorithms for $p \leq 1$, $p \neq 0$ satisfies all the conditions required by the theorem. For ease of exposition, the convergence analysis is subdivided into four stages. They include the following: 1) Defining the solution set Γ, 2) identifying the descent function, 3) refining the solution set, and 4) establishing the boundedness of the sequences. We start with the definition of the solution set.

Solution Set: The solution set Γ is obtained by collecting all the stationary solutions of the algorithm. More precisely

$$\Gamma = \{x_* : Ax_* = b, \quad \text{and} \quad x_* = W_*(AW_*)^{+}b$$

where

$$W_* = \operatorname{diag}\left(|x_*[i]|^{1-\frac{p}{2}}\right)\}.$$

This set contains precisely those x_k that satisfy (10). It can be shown to contain the set Γ_b, which contains all the basic and degenerate basic solutions, i.e., solutions with no more than m nonzero entries that are obtained by selecting m columns of A and solving the corresponding linear system of equations whenever it exists [30]. There are potentially $\binom{n}{m}$ such solutions. Later, we will examine Γ more closely and show that Γ_b is the more relevant set.

Descent Function: The next important step is the determination of the descent function. It is shown that the descent function for the algorithm is the diversity measure $E^{(p)}(x)$ itself, i.e., for x_k generated by the algorithm (15), we have

$$E^{(p)}(x_{k+1}) < E^{(p)}(x_k), \quad x_{k+1} \neq x_k. \tag{20}$$

To show the validity of (20), we make use of the Hölder inequality. Since we use a more general form of the Hölder inequality than is commonly presented in textbooks, we state it here for completeness.

Theorem 3 (Generalized Hölder Inequality) [40]: If $x_i, y_i \geq 0, r > 1, \frac{1}{r} + \frac{1}{s} = 1$, then

$$\sum_{i=1}^{n} x_i y_i \leq \left(\sum_{i=1} x_i^r\right)^{\frac{1}{r}} \left(\sum_{i=1} y_i^s\right)^{\frac{1}{s}}.$$

The inequality is reversed for $r < 1$ $(r \neq 0)$, assuming that $x_i, y_i > 0$ (strict positivity). In each case, equality holds if and only if the sets (x^r) and (y^s) are proportional.

We first consider the case $0 < p \leq 1$. Recall from Section III-A that x_k generated by (11) is feasible (i.e., $Ax_k = b$) for all k, and note from (15) that q_{k+1} is obtained as the optimal minimum 2-norm solution to the problem $AW_{k+1}q = b$, where $W_{k+1} = \text{diag}(|x_k[i]|^{\frac{2-p}{2}})$. In addition, note that feasibility of x_k implies that $\bar{q}$ defined by $\bar{q}[i] = \text{sgn}(x_k[i])|x_k[i]|^{\frac{p}{2}}$ is a feasible (but nonoptimal) solution to $AW_{k+1}q = b$, assuming $x_{k+1} \neq x_k$ (nonconvergence of (11)). Therefore

$$\|q_{k+1}\|_2^2 < \|\bar{q}\|_2^2 = \sum_{i=1}^{n} |x_k[i]|^p. \tag{21}$$

The entries of $x_{k+1} = W_{k+1}q_{k+1}$ can be written as $x_{k+1}[i] = |x_k[i]|^{\frac{2-p}{2}} q_{k+1}[i]$. Hence

$$\begin{aligned} E^{(p)}(x_{k+1}) &= \sum_{i=1}^{n} |x_{k+1}[i]|^p \\ &= \sum_{i=1}^{n} |x_k[i]|^{\frac{(2-p)p}{2}} |q_{k+1}[i]|^p. \end{aligned} \tag{22}$$

Let $r = \frac{2}{p}$, and define s by $\frac{1}{r} + \frac{1}{s} = 1$. Therefore, we have $s = \frac{2}{2-p}$. Applying the Hölder inequality to (22), we have

$$E^{(p)}(x_{k+1}) \leq \left(\sum_{i=1}^{n} |x_k[i]|^{\frac{(2-p)ps}{2}}\right)^{\frac{1}{s}} \left(\sum_{i=1}^{n} |q_{k+1}[i]|^{pr}\right)^{\frac{1}{r}} \tag{23}$$

$$= \left(\sum_{i=1}^{n} |x_k[i]|^p\right)^{\frac{2-p}{2}} \left(\sum_{i=1}^{n} |q_{k+1}[i]|^2\right)^{\frac{p}{2}} \tag{24}$$

$$< \left(\sum_{i=1}^{n} |x_k[i]|^p\right)^{\frac{2-p}{2}} \left(\sum_{i=1}^{n} |x_k[i]|^p\right)^{\frac{p}{2}} \tag{25}$$

$$= \sum_{i=1}^{n} |x_k[i]|^p = E^{(p)}(x_k) \tag{26}$$

where the second strict inequality follows from (21). Thus, we have shown that the diversity measure $E^{(p)}(x_k)$ is reduced in each iteration for $0 < p \leq 1$.

The proof can similarly be shown for $p < 0$. Note that this time, the Hölder inequality is reversed but is compensated for by the negative sign arising from the factor $\text{sgn}(p)$ included in the diversity measure (2), making (23) still valid, except that the term on the right-hand side is negated. To successfully carry through the proof, it is necessary to make use of the fact that p is negative in (25).

Another useful observation to make is that the above descent function is actually valid for $p < 2$, $p \neq 0$ and not just $p \leq 1$, $p \neq 0$. Therefore, the FOCUSS algorithm can actually also be used to minimize the l_p norm of x for $1 < p < 2$. We have not emphasized this range of p because of their inability to generate truly concentrated solutions.

Refinement of Solution Set: Here, we show that the points to which the algorithm converges almost surely lies in the set Γ_b, which contains solutions with a maximum of m nonzero entries. A useful observation in this context is to note that in the FOCUSS algorithm, once an entry becomes zero, it remains zero in the rest of the iterations. Therefore, we concentrate only on the nonzero entries. Let us suppose that x_* has r nonzero entries. Let $A^{(r)}$ be the $m \times r$ matrix formed by collecting the selected basis vectors, and let $x_*^{(r)}, W_*^{(r)}$, and $\Pi(x_*^{(r)})$ be the corresponding quantities extracted from x_*, W_*, and $\Pi(x_*)$, respectively. Then, we can redefine the solution set Γ and Γ_b as

$$\begin{aligned} \Gamma = \big\{ & x_* : Ax_*^{(r)} = b, \quad \text{and} \quad x_*^{(r)} = W_*^{(r)} \big(A^{(r)} W_*^{(r)}\big)^{+} b \\ & 1 \leq r \leq n \big\} \\ \Gamma_b = \big\{ & x_* : Ax_*^{(r)} = b, \quad \text{and} \quad x_*^{(r)} = W_*^{(r)} \big(A^{(r)} W_*^{(r)}\big)^{+} b \\ & 1 \leq r \leq m \big\}. \end{aligned}$$

Γ_b contains solutions with a maximum of m nonzero entries and is the solution set of interest. They will be shown to be stable fixed points and the remaining stationary points in Γ to be saddle points or unstable fixed points. Such conclusions were also reached in [24]. The insights gained by the approach used in Section III to derive the algorithm provides an alternate direct approach to showing these results. For brevity, we only show that stationary points with the number of nonzero entries between m and n are saddle points, i.e., for $m < r < n$. Note that the stationary points satisfy

$$A^{(r)} x_*^{(r)} = b \quad \text{and} \quad \Pi\big(x_*^{(r)}\big) x_*^{(r)} \in \mathcal{R}\big(A^{(r)T}\big).$$

For $m < r < n$, such points are indeed rare. This can be seen by examining closely what is required of the stationary points. For example, if we consider the case $p = 0$, then $x_*^{(r)}$, whichis a vector in R^r, lies in a linear variety of dimension $(r - m) = \dim(\mathcal{N}(A^{(r)}))$,, and it is simultaneously required

that the r-dimensional vector $\Pi(x_*^{(r)})x_*^{(r)}$, which is obtained by inverting element wise the entries of $x_*^{(r)}$, lie in $\mathcal{R}(A^{(r)T})$, which is an m-dimensional subspace. Generic solutions to $A^{(r)}x_*^{(r)} = b$ will not satisfy this condition. Not only are these solutions rare, but they are saddle points. This can be seen by performing a simple Taylor series expansion of $E^{(p)}(x^{(r)})$ about the feasible solution $x_*^{(r)}$. Let $x^{(r)} = x_*^{(r)} + \mu d$, where d is any vector in the null space of $A^{(r)}$. Note that we are looking at perturbations that do not change the sparsity of $x_*^{(r)}$. Since $r > m$, a nontrivial null space for $A^{(r)}$ exists. Then

$$\begin{aligned} E^{(p)}\big(x^{(r)}\big) &= E^{(p)}\big(x_*^{(r)}\big) + \mu|p|x_*^{(r)T}\Pi\big(x_*^{(r)}\big)d \\ &\quad + \mu^2|p|(p-1)d^T\operatorname{diag}\big(|x_*^{(r)}[i]|^{p-2}\big)d + O(\mu^3) \qquad (27) \\ &= E^{(p)}\big(x_*^{(r)}\big) + \mu^2|p|(p-1)d^T\operatorname{diag}\big(|x_*^{(r)}[i]|^{p-2}\big)d \\ &\quad + O(\mu^3). \qquad (28) \end{aligned}$$

The simplification of (27) to (28) is possible because d and $\Pi(x_*^{(r)})x_*^{(r)}$ are orthogonal as $d \in \mathcal{N}(A^{(r)})$ and $\Pi(x_*^{(r)})x_*^{(r)} \in \mathcal{R}(A^{(r)T})$. For $p < 1$, $E^{(p)}(x^{(r)}) < E^{(p)}(x_*^{(r)})$ in an arbitrarily small neighborhood of $x_*^{(r)}$. Thus, $x_*^{(r)}$ must be a local maximum along directions $d \in \mathcal{N}(A^{(r)})$. On the other hand, perturbing $x_*^{(r)}$ such that the number of nonzero entries increases can be shown to increase $E^{(p)}(x)$. Therefore, $x_*^{(r)}, m < r < n$ are saddle points. They are not a source of much concern as they are hard to get to, and a small perturbation can nudge the algorithm away from these points. Furthermore, they are easy to identify from the fact that $r > m$ for these points. More generally rank$(A^{(r)}) \neq r$.

Boundedness of the Sequence x_k: We now prove that the sequence generated by (11) [equivalently, (15)] is contained in a compact set by showing that the sequence $\|x_k\|$ is bounded. This is fairly easy to show if we restrict p to be positive. Based on the descent function analysis, for $0 < p \leq 1$, we have $|x_k[i]| \leq \big(E^{(p)}(x_0)\big)^{\frac{1}{p}}$.

Now, we concentrate on the case $p \leq 0$. The proof is somewhat more involved and is by contradiction. Suppose that the sequence x_k is unbounded. This implies that at least one element $x_k[l_1] \to \infty$. Then, since $Ax_k = b$, rearranging the equations, we have

$$A'x'_k = b - a_{l_1}x_k[l_1] \qquad (29)$$

where A' is the matrix A with the l_1th column a_{l_1} removed, and x'_k is the x_k vector with the l_1th entry $x_k[l_1]$ deleted. If $x_k[l_1] \to \infty$, then by (29), certain elements of x'_k must also tend to infinity. Let those elements be $x_k[l_r], r = 2, \ldots, s-1$. If we assume that any m columns of A are linearly independent,[5] then generically, $s \geq m$. Note that the vectors $a_{l_r}, r = 1, \ldots, m$ form a basis set, and let x_b be the basic solution corresponding to this set, which solves

$$\sum_{r=1}^{m} a_{l_r}x_b[l_r] = b.$$

[5]This assumption is not really necessary and is made to simplify the proof.

The vector x_b is bounded. If $x_k \to \infty$, then there exists an iteration index k_1 such that

$$|x_{k_1+1}[l_r]| > |x_b[l_r]|, \quad r = 1, \ldots, m. \qquad (30)$$

Note that $x_{k_1+1} = W_{k_1+1}q_{k_1+1}$ and that $\|q_{k_1+1}\|^2$ is smallest of all solutions to $AW_{k_1+1}q = b$. Define q' as

$$q'[i] = q_{k_1+1}[i]R[i]$$

where

$$R[i] = \begin{cases} \frac{x_b[l_r]}{x_{k_1+1}[l_r]}, & r = 1, \ldots, m \\ 0, & \text{otherwise} \end{cases}. \qquad (31)$$

Note that q' is a feasible solution, i.e., satisfies $AW_{k_1+1}q' = b$. By (30) and (31), it can be concluded that $|R[i]| < 1$, and hence, $\|q'\|^2 < \|q_{k_1+1}\|^2$. This contradicts the fact that $\|q_{k_1+1}\|$ is the smallest.

REFERENCES

[1] M. Vetterli and J. Kovacevic, *Wavelets and Subband Coding*. Englewood Cliffs, NJ: Prentice-Hall, 1995.

[2] G. Strang and T. Nguyen, *Wavelets and Filter Banks*. Wellesley, MA: Wellesley-Cambridge, 1996.

[3] A. N. Akansu and R. A. Haddad, *Multiresolution Signal Decomposition: Transforms, Subbands and Wavelets*. New York: Academic, 1992.

[4] I. Daubechies, *Ten Lectures on Wavelets*. Philadelphia, PA: SIAM, 1992.

[5] N. Hess-Nielsen and M. V. Wickerhauser, "Wavelets and time-frequency analysis," *Proc. IEEE*, vol. 84, pp. 523–540, Apr. 1996.

[6] K. Ramachandran, M. Vetterli, and C. Herley, "Wavelets, subband coding and best bases," *Proc. IEEE*, vol. 84, pp. 541–560, Apr. 1996.

[7] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 41, pp. 3397–3415, Dec. 1993.

[8] R. R. Coifman and M. V. Wickerhauser, "Entropy-based algorithms for best basis selection," *IEEE Trans. Inform. Theory*, vol. IT-38, pp. 713–718, Mar. 1992.

[9] M. V. Wickerhauser, *Adapted Wavelet Analysis from Theory to Software*. Wellesley, MA: A. K. Peters, 1994.

[10] D. Donoho, "On minimum entropy segmentation," in *Wavelets: Theory, Algorithms, and Applications*, C. K. Chui, L. Montefusco, and L. Puccio, Eds. New York: Academic, 1994, pp. 233–269.

[11] S. Chen and D. Donoho, "Basis pursuit," in *Proc. Twenty-Eighth Asilomar Conf. Signals, Syst., Comput.*, Monterey, CA, Nov. 1994, vol. I, pp. 41–44.

[12] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," Tech. Rep. 479, May 1995.

[13] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proc. 27th Asilomar Conf. Signals, Syst., Comput.*, Nov. 1993, pp. 40–44.

[14] J. M. Adler, B. D. Rao, and K. Kreutz-Delgado, "Comparison of basis selection methods," in *Proc. 30th Asilomar Conf. Signals, Syst., Comput.*, Monterey, CA, Nov. 1996, vol. 1, pp. 252–257.

[15] B. Jeffs and M. Gunsay, "Restoration of blurred star field images by maximally sparse optimization," *IEEE Trans. Image Processing*, vol. 2, pp. 202–211, Apr. 1993.

[16] I. F. Gorodnitsky and B. D. Rao, "A recursive weighted minimum-norm algorithm: Analysis and applications," in *Proc. ICASSP*, Minneapolis, MN, Apr. 1993, vol. III, pp. 456–459.

[17] G. Harikumar and Y. Bresler, "A new algorithm for computing sparse solutions to linear inverse problems," in *Proc. ICASSP*, Atlanta, GA, May 1996, vol. III, pp. 1331–1334.

[18] I. Santamaria-Caballero, C. J. Pantaleon-Prieto, and A. Artes-Rodriguez, "Sparse deconvolution using adaptive mixed-Gaussian models," *Signal Process.*, vol. 54, pp. 161–172, 1996.

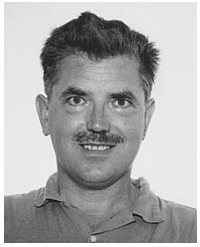
[19] P. S. Bradley and O. L. Mangasarian, "Feature selection via mathematical programming," Univ. Wisconsin Math. Program. Tech. Rep. 95-21. Available ftp at ftp://ftp.cs.wisc.edu/math-prog/tech-reports/95-21.ps.Z.

[20] I. F. Gorodnitsky, J. S. George, and B. D. Rao, "Neuromagnetic source imaging with FOCUSS: A recursive weighted minimum norm algorithm," *J. Electroencephalograph. Clinical Neurophysiol.*, vol. 95, no. 4, pp. 231–251, Oct. 1995.
[21] A. M. Kondoz, *Digital Speech: Low Bit Rate Coding for Communication Systems*. New York; Wiley, 1996.
[22] H. Lee, D. P. Sullivan, and T. H. Huang, "Improvement of discrete band-limited signal extrapolation by iterative subspace modification," in *Proc. ICASSP*, Dallas, TX, Apr. 1987, vol. 3, pp. 1569–1572.
[23] S. D. Cabrera and T. W. Parks, "Extrapolation and spectral estimation with iterative weighted norm modification," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 39, pp. 842–851, Apr. 1991.
[24] I. F. Gorodnitsky and B. D. Rao, "Sparse signal reconstructions from limited data using FOCUSS: A re-weighted minimum norm algorithm," *IEEE Trans. Signal Processing*, vol. 45, pp. 600–616, Mar. 1997.
[25] S. Chen and J. Wigger, "Fast orthogonal least squares algorithm for efficient subset model selection," *IEEE Trans. Signal Processing*, vol. 43, pp. 1713–1715, July 1995.
[26] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM J. Comput.*, vol. 24, pp. 227–234, Apr. 1995.
[27] R. E. Carlson and B. K. Natarajan, "Sparse approximate multiquadric interpolation," *Comput. Math Applicat.*, vol. 27, pp. 99–108, 1994.
[28] P. Duhamel and J. C. Rault, "Automatic test generation techniques for analog circuits and systems: A review," *IEEE Trans. Circuits Syst.*, vol. CAS-26, pp. 411–440, July 1979.
[29] S. G. Nash and A. Sofer, *Linear and Nonlinear Programming*. New York: McGraw-Hill, 1996.
[30] S.-C. Fang and S. Putenpura, *Linear Optimization and Extensions*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[31] D. den Hertog, *Interior Point Approach to Linear, Quadratic, and Convex Programming: Algorithms and Complexity*. Boston, MA: Kluwer, 1994.
[32] I. F. Gorodnitsky and B. D. Rao, "Convergence analysis of a class of adaptive weighted norm extrapolation algorithms," in *Proc. Twenty-Seventh Asilomar Conf. Signals, Syst., Comput.*, Monterey, CA, Nov. 1993, vol. I, pp. 339–343.
[33] K. Kreutz-Delgado and B. D. Rao, "A general approach to sparse basis selection: Majorization, concavity, and affine scaling," Tech. Rep. UCSD-CIE-97-7-1, Univ. Calif, San Diego, July 1997.
[34] Y. Li, "A globally convergent method for ℓ_p problems," *SIAM J. Optimiz.*, vol. 3, no. 3, pp. 609–629, Aug. 1993.
[35] C. Taswell, "Satisficing search algorithms for selecting near-best bases in adaptive tree-structured wavelet transforms," *IEEE Trans. Signal Processing*, vol. 44, pp. 2423–2438, Oct. 1996.
[36] D. Luenberger, *Linear and Nonlinear Programming*. Reading, MA: Addison-Wesley, 1989.
[37] M. S. Bazaraa and C. M. Shetty, *Nonlinear Programming: Theory and Algorithms*. New York: Wiley, 1979.
[38] P. E. Gill, W. Murray, and M. W. Wright, *Practical Optimization*. New York: Academic, 1981.
[39] B. D. Rao, "Analysis and extensions of the FOCUSS algorithm," in *Proc. 30th Asilomar Conf. Signals, Syst., Comput.*, Monterey, CA, Nov. 1996, vol. 2, pp. 1218–1223.
[40] E. F. Beckenbach and R. Bellman, *Inequalities*. New York: Springer-Verlag, 1970.
[41] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

Bhaskar D. Rao (SM'91) received the B.Tech. degree in electronics and electrical communication engineering from the Indian Institute of Technology, Kharagpur, in 1979 and the M.S. and Ph.D. degrees from the University of Southern California, Los Angeles, in 1981 and 1983 respectively.

Since 1983, he has been with the University of California, San Diego, where he is currently a Professor in the Electrical and Computer Engineering Department. His interests are in the areas of digital signal processing, estimation theory, and optimization theory with applications to speech, communications, and human-computer interactions.

Dr. Rao has been a member of the Statistical Signal and Array Processing Technical Committee of the IEEE Signal Procesing Society.



Kenneth Kreutz-Delgado (S'79–M'84–SM'93) received the B.A. and M.S. degrees in physics and the Ph.D. degree in engineering systems science, all from the University of California, San Diego (UCSD).

Currently, he is an Associate Professor in the UCSD Electrical and Computer Engineering Department and is affiliated with the Center for Information Engineering and the Institute for Neural Computation. Before joining the faculty at UCSD, he was a researcher at the NASA Jet Propulsion Laboratory, California Institute of Technology, Pasadena, where he worked on the development of intelligent telerobotic systems for use in space/planetary exploration and satellite servicing and repair. His interest in sensor-based intelligent systems that can function in unstructured and nonstationary environments is the basis for his research activities in sensing and signal processing in natural environments; real-time pattern recognition/classification from multiple time-series data sets; adaptive sensory-motor control; multibody systems theory (dynamics and kinematics) applied to vision and control; and the development of emergent intelligent behavior from distributed complex systems.

Dr. Kreutz-Delgado is a member of the AAAS and the IEEE Signal Processing, Robotics, SMC, Control, and Computer societies.