Deriving Algorithms for Computing Sparse Solutions to Linear Inverse Problems

B.D. Rao* and K. Kreutz-Delgado
Electrical and Computer Engineering Department
University of California, San Diego
La Jolla, California 92093-0407

Abstract

*A novel methodology is employed to develop algorithms for computing sparse solutions to linear inverse problems, starting from suitably defined diversity measures whose minimization promotes sparsity. These measures include p-norm-like ($\ell_{(p\leq 1)}$) diversity measures, and the Gaussian and Shannon Entropies. The algorithm development methodology uses a factored representation of the gradient, and involves successive relaxation of the Lagrangian necessary condition. The general nature of the methodology provides a systematic approach for deriving a recently developed class of algorithms called FOCUSS (***FOC****al* ***U****nderdetermined* ***S****ystem* ***S****olver), and a natural mechanism for extending them.*

1 Introduction

The need to compute sparse solutions to linear inverse problems arises in many applications, e.g. biomagnetic imaging, signal representation, speech coding, bandlimited extrapolation and spectral estimation, direction of arrival estimation, functional approximation, failure diagnosis, and pattern recognition for medical diagnosis [1, 2, 3, 4, 5, 6, 7, 8]. From the large number of potential applications, it is clear that an effective solution to this problem has wide ranging consequences.

This work is motivated by the **FOC**al **U**nderdetermined **S**ystem **S**olver (FOCUSS) algorithm presented and analyzed in [5]. We present a systematic derivation of the FOCUSS class of algorithms by applying optimization principles to minimize suitable diversity measures. The benefits of the approach are twofold. Firstly, it provides a strong theoretical foundation for the FOCUSS algorithm which results in new insights, and facilitates a more complete analysis. Secondly, it exposes the general principles underlying the algorithm providing a natural mechanism for extending them.

The outline of the paper is as follows. In Section 2 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 [6] and [2], and define the problem as that of minimizing these measures subject to the constraint that the measurement vector be a feasible solution. In Section 3, we employ a novel methodology to derive an iterative algorithm that selects a sparse representation by minimizing the p-norm-like sparsity measures, excluding (temporarily) the case $p = 0$. The iterative algorithm is derived using a factored representation of the gradient, and by successive relaxation of the Lagrangian necessary conditions for a minimum. In Section 3.2 convergence analysis of the algorithm is performed, expanding the scope of the convergence results previously prescribed in [5]. In Section 4, 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 by [2]. In Section 5, the methodology is used to derive an algorithm to minimize the Shannon entropy.

2 PROBLEM FORMULATION

The problem of computing sparse solutions to linear inverse problems can be formulated as the problem of finding a sparse solution to an underdetermined system of equations [7, 8]. Let A be an $m \times n$ matrix formed using the vectors derived from the forward model. We have an underdetermined system where $m < n$ and it is assumed that $\text{rank}(A) = m$. Denoting the given measurement vector by b, a $m \times 1$ vector, the inverse

*Corresponding author; email: brao@ece.ucsd.edu

1058-6393/98 $10.00 © 1998 IEEE

problem consists of solving for x, a $n \times 1$ vector, such that

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

The problem of computing a sparse solution requires that x be sparse, i.e. most of the entries of x be zero. Eq. (1) ensures that x is a consistent representation of b, and the sparsity requirement ensures that the solution is concentrated.

The inverse 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 $\mathcal{N}(A)$, 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. However, the minimum 2-norm criteria favors solutions with many small nonzero entries, a property that is contrary to the goal of sparsity/concentration [5, 7]. Consequently there is a need to define other functionals, referred to here as diversity measures, which when minimized lead to sparse/concentrated solutions.

The question of good diversity measures has been studied in the past and a good discussion can be found in [6, 2], and in the literature on linear inverse problems [1]. A popular diversity measure is $E^{(p)}(x)$, where

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

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

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

where $\mathrm{sgn}(p) = \begin{cases} +1, & 0 \le p \le 1 \\ -1, & p < 0 \end{cases}$. The diversity measures $E^{(p)}(x)$ for $0 \le p \le 1$ are the general family of entropy-like measures defined in [6, 2], and also discussed in [1], for computing 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 "$\ell_{(p\le 1)}$ 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 [1].

The diversity measure for $p = 0$, the *numerosity* discussed in [2], 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 measures requires an enumerative search and is NP hard [3]. Consequently, alternate diversity measures that are more amenable to optimization techniques are of interest. The $E^{(p)}(x)$ measures for $p \le 1, p \neq 0$ are useful candidate measures in this context, and are indirectly related to sparsity in that *when minimized* 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 [5]. For a more detailed discussion of these diversity measures for $0 \le p \le 1$ the reader is referred to [6, 2, 1]. Additional discussion can be found in [9]. To avoid potential confusion, it should noted that minimization of $E^{(p)}(x)$ is considerably different from the standard ℓ_p optimization problem $\min_x \|Ax - b\|_p^p$, $p \ge 1$ [10].

The diversity measures $E^{(p)}(x)$ for $p < 0$ are also good (indirect) diversity 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 [9]. We only examine here the Shannon Entropy and Gaussian Entropy, two other diversity measures described in [6, 2]. The Shannon Entropy diversity measure $H_S(x)$ is defined as

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

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

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

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

3 $\ell_{(p\leq 1)}$ DIVERSITY 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 4. The algorithm is derived in Section 3.1 using a factored representation for the gradient, and by successive relaxation of the Lagrangian necessary conditions. Interestingly, the approach turns out to be a systematic procedure for deriving a class of algorithms called FOCUSS developed in [5]. Additionally the methodology employed provides a mechanism for generalizing and deriving FOCUSS-like algorithms to other situations.

3.1 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. 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.

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

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 respresentation* for the gradient vector of the diversity measure, i.e.

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

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

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

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

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

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

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

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

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

Eq. (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 following iterative procedure for computing x_*,

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

The computation of $\Pi^{-1}(x_k) = \mathrm{diag}(|x_k[i]|^{(2-p)})$ for $p \leq 1$ does not pose any implementation problems, even as elements $x_k[i]$ converge to zero (as is desired, the goal being 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.

More insight into the methodology is obtained by interpreting the approach as a method of solving successive constrained weighted *minimum norm* problems. Note that each iteration of (11) corresponds to computing a weighted minimum-norm solution to (1), i.e.

$$x_{k+1} = \arg\min \left[x^T\Pi^{-1}(x_k)x \text{ subject to } Ax = b\right]$$

Defining the symmetric scaling matrix W by $W^{-2}(x) \triangleq \Pi(x) = \mathrm{diag}(|x[i]|^{p-2})$, with $W(x) = \mathrm{diag}(|x[i]|^{1-\frac{p}{2}})$, a computational alternative to the algorithm can be obtained which has the form

$$\begin{array}{rcl} W_{k+1} & = & \mathrm{diag}(|x_k[i]|^{1-\frac{p}{2}}) \\ q_{k+1} & = & A_{k+1}^{+}b, \text{ where } A_{k+1} = AW_{k+1} \\ x_{k+1} & = & W_{k+1}q_{k+1}. \end{array} \quad (12)$$

The algorithm (12) has similarities to Affine Scaling Methods, a class of interior point optimization methods, where the weighting matrix $W(x)$ plays the role of the Affine Scaling Transformation (AST) matrix. More details on this connection can be found in [11], and because of this connection we often refer to the algorithm as an affine scaling algorithm or an AST algorithm.

The methodology employed in arriving at the iterative algorithm (11) from (6) and (7) is novel. It

is closely related to the algorithms developed in the context of the ℓ_p optimization problem of minimizing $\|Ax - b\|_p^p$, $p \geq 1$ [10]. 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* problem. The algorithm developed (c.f. (11)) can be viewed as an extension of the methodology to the underdetermined problem.

3.2 Convergence Analysis

Having proposed and motivated the $\ell_{(p\leq 1)}$-class of algorithms given by (11) (equivalently, by (12)), 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 [5]. The insights provided by the systematic derivation presented in section 3.1 enables strengthening of the result shown in [5]. The main result is as follows.

Theorem 1 *Starting from a bounded feasible solution x_0, the algorithm (12) 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 non-zero entries.*

The details of the proof can be found in [11]. Here we highlight the main improvement obtained. To establish convergence, a descent function is required and it is shown that the descent function for these algorithms is the diversity measure itself, i.e.

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

where Γ is the solution set. In [5], 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 more restrictive conditions; by either restricting the $x[i]$ to be positive, or, more generally, by requiring that the $x[i]$ all lie in the same quadrant. We remove this restriction and show decrease in the descent function starting from any $x_0 \in R^n$.

4 NUMEROSITY AND GAUSSIAN ENTROPY

We now pay special attention to the case where $p = 0$, which, as previously discussed, yields a numerosity measure which 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 one ideally would prefer to minimize as observed in [1, 6, 2]. 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 (12) (equivalently, by (11)) yields a well-behaved algorithm even when $p = 0$. Indeed, letting $p = 0$ in (12) yields the basic FOCUSS algorithm of [5] and involves the use of a well-defined scaling matrix $W(x) = \text{diag}(|x[i]|)$. Although the algorithm (12) is well-defined for $p = 0$, the convergence analysis differs somewhat from the $p \neq 0$ analysis described in section 3.2. In [5], 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, one can consider minimizing directly the Gaussian Entropy, or the monotonically related (and hence equivalent) cost function $\text{Exp}(H_G(x)) = \Pi_{i=1}^{n} |x[i]|^2$. The latter one is preferable if one is 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 3.1; merely replace $E^{(p)}(x)$ by $H_G(x)$ in the analysis. The only new quantity required is the gradient of $H_G(x)$, which can be readily shown to have the following factored representation:

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

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 (12), 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 one tries 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 [12]

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

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. [12]

$$e^{\frac{1}{2n}H_G(x)} = \lim_{p\to 0^+}\left(\frac{1}{n}E^{(p)}(x)\right)^{\frac{1}{p}}. \qquad (13)$$

One can also relate the Gaussian Entropy to the $\ell_{(p\leq 1)}$ diversity measures via a Taylor series expansion [11]. A related discussion along these lines can be found in [2].

5 SHANNON ENTROPY

An algorithm for minimizing the Shannon Entropy diversity measure $H_S(x)$ defined by (3) and discussed in [6, 2] can also be developed using the factored representation of the gradient, and the relaxation of the Lagrangian neccesary condition approach developed in section 3.1. This necessitates taking the gradient of the diversity measure, which has the following factored representation:

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

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

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

Retracing the argument given in Section 3.1 through equation (11) suggests that we focus on the iteration

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

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. It is shown in [11] that a suitably modified algorithm to minimize the Shannon entropy is

$$x_{k+1} = \begin{cases} x_{k+1}^r, & (x_{k+1}^r)^T\Pi_S(x_k)x_{k+1}^r \leq 0 \\ 2x_k - x_{k+1}^r, & (x_{k+1}^r)^T\Pi_S(x_k)x_{k+1}^r > 0\,. \end{cases} \qquad (14)$$

Though the algorithm converges, it does not converge to a truly sparse solution. However, the diversity measure does promote concentration in that the final solution does tend to have a large number of entries with very small (albeit nonzero) amplitudes. More details can be found in [11]. Extensions of the diversity measures along with a general convergence framework can be found in [9].

References

[1] B. Jeffs and M. Gunsay. "Restoration of Blurred Star Field Images by Maximally Sparse Optimization". *IEEE Trans. on Image Processing.*, 1993.

[2] D. Donoho. "On Minimum Entropy Segmentation". In *Wavelets: Theory, Algorithms, and Applications, edited by C. K. Chui, et al*, 1994.

[3] B. K. Natarajan. "Sparse Approximate Solutions to Linear Systems". *SIAM Journal on Computing*, 24(2):227–234, April 1995.

[4] I.F. Gorodnitsky, J.S. George, and B.D. Rao. "Neuromagnetic source imaging with FOCUSS: a recursive weighted minimum norm algorithm". *Journal of Electroencephalography and Clinical Neurophysiology*, 95(4):231–251, Oct. 1995.

[5] I.F. Gorodnitsky and B.D. Rao. "Sparse Signal Reconstructions from Limited Data using FOCUSS: A Reweighted Minimum Norm Algorithm". *IEEE Trans. on Signal Processing*, 45:600–616, 1997.

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

[7] S. Chen and D. Donoho. "Basis Pursuit". In *Twenty-Eighth Asilomar Conference on Signals, Systems and Computers, Vol. I*, pages 41–44, CA, Nov. 1994.

[8] J. M. Adler, B. D. Rao, and K. Kreutz-Delgado. "Comparison of Basis Selection Methods". In *Proc. of the 30th Asilomar Conference on Signals, Systems and Computers*, volume 1, pages 252–257, Nov. 1996.

[9] K. Kreutz-Delgado and B.D. Rao. "A General Approach to Sparse Basis Selection: Majorization, Concavity, and Affine Scaling". *Submitted to the IEEE Trans. on Signal Processing*, September 1997.

[10] Y. Li. "A Globally Convergent Method for ℓ_p Problems". *SIAM J. Optim.*, Aug. 1993.

[11] B. D. Rao and K. Kreutz-Delgado. "An Affine Scaling Methodology for Best Basis Selection". *submitted to IEEE Trans. on Signal Processing*, Jan. 1997.

[12] E. F. Beckenbach and R. Bellman. *Inequalities*. Springer-Verlag, 1970.