Solving Sparse Linear Inverse Problems: Analysis of Reweighted l1 and l2 Methods

David Wipf, Srikantan Nagarajan

Solving Sparse Linear Inverse Problems: Analysis of Reweighted ℓ_1 and ℓ_2 Methods

David Wipf and Srikantan Nagarajan
Biomagnetic Imaging Laboratory
University of California, San Francisco CA 94143
Email: {dwipf,sri}@mrsc.ucsf.edu

***Abstract*—A variety of practical methods have recently been introduced for finding maximally sparse representations from overcomplete dictionaries, a central computational task in compressed sensing and source localization applications as well as numerous others. Many of the underlying algorithms rely on iterative reweighting schemes that produce more focal estimates as optimization progresses. Two such variants are iterative reweighted ℓ_1 and ℓ_2 minimization; however, some properties related to convergence and sparse estimation, as well as possible generalizations, are still not clearly understood or fully exploited. In this paper, we make the distinction between *separable* and *non-separable* iterative reweighting algorithms. The vast majority of existing methods are separable, meaning the weighting of a given coefficient at each iteration is only a function of that individual coefficient from the previous iteration (as opposed to dependency on all coefficients). We examine two such separable reweighting schemes: an ℓ_2 method from Chartand and Yin (2008) and an ℓ_1 approach from Candès et al. (2008), elaborating on convergence results and explicit connections between them. We then explore an interesting non-separable variant that can be implemented via either ℓ_2 or ℓ_1 reweighting and show several desirable properties relevant to sparse recovery. For the former, we show a direct connection with Chartrand and Yin's approach. For the latter, we demonstrate two desirable properties: (i) each iteration can only improve the sparsity and (ii), for any dictionary and sparsity profile, there will always exist cases where non-separable ℓ_1 reweighting improves over standard ℓ_1 minimization.**

I. Introduction

In recent years, there has been considerable interest in finding sparse signal representations from redundant dictionaries [7], [8], [11], [12], [19]. The canonical form of this problem is given by,

$$\min_{\boldsymbol{x}} \|\boldsymbol{x}\|_0, \qquad \text{s.t. } \boldsymbol{y} = \Phi\boldsymbol{x}, \tag{1}$$

where $\Phi \in \mathbb{R}^{n\times m}$ is a matrix whose columns $\boldsymbol{\phi}_i$ represent an overcomplete or redundant basis (i.e., rank$(\Phi) = n$ and $m > n$), $\boldsymbol{x} \in \mathbb{R}^m$ is a vector of unknown coefficients to be learned, and $\boldsymbol{y}$ is the signal vector. The cost function being minimized represents the ℓ_0 (quasi)-norm of $\boldsymbol{x}$ (i.e., a count of the nonzero elements in $\boldsymbol{x}$). If measurement noise or other complications are present, we instead solve the alternative problem

$$\min_{\boldsymbol{x}} \|\boldsymbol{y} - \Phi\boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{x}\|_0, \quad \lambda > 0, \tag{2}$$

noting that in the limit as $\lambda \to 0$, the two problems are equivalent (the limit must be taken outside of the minimization).

Unfortunately, an exhaustive search for the optimal representation requires the solution of up to $\binom{m}{n}$ linear systems of size $n\times n$, a prohibitively expensive procedure for even modest values of m and n. Consequently, in practical situations there is a need for approximate methods that efficiently solve (1) or (2) with high probability. Many recent sparse approximation algorithms rely on iterative reweighting schemes that produce more focal estimates as optimization progresses [3], [4], [5], [15], [16]. Two such variants are iterative reweighted ℓ_2 and ℓ_1 minimization. For the former, at the $(k+1)$-th iteration we must compute

$$\begin{aligned} \boldsymbol{x}^{(k+1)} &\rightarrow \arg\min_{\boldsymbol{x}} \|\boldsymbol{y} - \Phi\boldsymbol{x}\|_2^2 + \lambda \sum_i \frac{x_i^2}{w_i^{(k)}} \\ &= W^{(k)}\Phi^T\left(\lambda I + \Phi W^{(k)}\Phi^T\right)^{-1}\boldsymbol{y}, \end{aligned} \tag{3}$$

where $W^{(k)}$ is a diagonal weighting matrix[1] from the k-th iteration with i-th diagonal element $w_i^{(k)}$ that is potentially a function of all $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(k)}$. Similarly, the ℓ_1 reweighting variant solves

$$\boldsymbol{x}^{(k+1)} \rightarrow \arg\min_{\boldsymbol{x}} \|\boldsymbol{y} - \Phi\boldsymbol{x}\|_2^2 + \lambda \sum_i \frac{|x_i|}{w_i^{(k)}}, \tag{4}$$

although no analytic solution exists in this case and so a numerical program (e.g., interior point method) must be adopted. While this is typically an expensive computation, the number of iterations of (4) required is generally much less than (3) and, unlike (3), even a single iteration produces a very sparse solution. A second advantage of (4) is that it is often much easier to incorporate additional constraints, e.g., bounded activation or non-negativity of $\boldsymbol{x}$ [2]. This is because few iterations are needed and the per-iteration computational complexity need not change significantly with the inclusion of many useful constraints. In contrast, with (3) we lose the advantage of closed-form updates on $\boldsymbol{x}$ when such constraints are imposed.

In both ℓ_2 and ℓ_1 reweighting schemes, different methods are distinguished by the choice of $W^{(k)}$, which ultimately determines the surrogate cost function for promoting sparsity that is being minimized (although not all choices lead to convergence or sparsity). In this paper we will explore several

[1] For notational convenience, our definition of the weights is the reciprocal of the weights used in several other papers.

different weighting selections, examining convergence issues, analytic properties related to sparsity, and connections between various algorithms. In particular, we will discuss a central dichotomy between what we will refer to as *separable* and *non-separable* choices for $W^{(k)}$. By far the most common, separable reweighting implies that each $w_i^{(k)}$ is only a function of $x_i^{(k)}$. This scenario is addressed in Section II where we examine the ℓ_2 method from Chartand and Yin (2008) [4] and an ℓ_1 approach from Candès et al. (2008) [3], elaborating on convergence results and explicit connections between them. In contrast, the non-separable case means that each $w_i^{(k)}$ is a function of all $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(k)}$, i.e., it is potentially dependent on all coefficients from all past iterations. Section III explores an interesting non-separable variant that can be implemented via either ℓ_2 or ℓ_1 reweighting and shows several desirable properties relevant to sparse recovery. For the former, we show a direct connection with Chartrand and Yin's approach. For the latter, we demonstrate two desirable properties: (i) each iteration can only improve the sparsity and (ii), for any dictionary and sparsity profile, there will always exist cases where non-separable ℓ_1 reweighting improves over standard ℓ_1 minimization. Simulations involving all of these methods are contained in Section IV.

II. Separable Reweighting Schemes

Separable reweighting methods have been applied to sparse recovery problems both in the context of the ℓ_2 norm [4], [10], [15], [16] and, very recently, the ℓ_1 norm [3]. All of these methods (at least locally) attempt to solve

$$\min_{\boldsymbol{x}} \|\boldsymbol{y} - \Phi\boldsymbol{x}\|_2^2 + \lambda \sum_i g(|x_i|), \tag{5}$$

where $g(\cdot)$ is a non-decreasing function. In general, (virtually) any square-concave[2] $g(\cdot)$ can be handled using reweighted ℓ_2 [15], while any concave $g(\cdot)$ can be used with reweighted ℓ_1 [9]. Many of these methods have been analyzed extensively in the past; consequently we will briefly address outstanding issues pertaining to two new approaches with substantial promise.

A. ℓ_2 Reweighting Method of Chartrand and Yin (2008) [4]

In [4] the ℓ_2 reweighting

$$w_i^{(k+1)} \rightarrow \left(x_i^{(k+1)}\right)^2 + \epsilon^{(k+1)} \tag{6}$$

is proposed (among others that are shown empirically to be less successful as discussed below), where $\epsilon^{(k+1)} \geq 0$ is a regularization factor that is reduced to zero as k becomes large. This procedure leads to state-of-the-art performance recovering sparse solutions in a series of empirical tests using a simple heuristic for updating $\epsilon^{(k+1)}$ and assuming $\lambda \rightarrow 0$ (noiseless case); however, some convergence issues are outstanding. Here we sketch a proof that this method will converge, for arbitrary sequences $\epsilon^{(k+1)} \rightarrow 0$, to a local minimum of a close surrogate cost function to (1) (similar ideas apply to the more general case where λ is nonzero).

To begin, we note that there is a one-to-one correspondence between minima of (1) and minima of

$$\min_{\boldsymbol{x}} \sum_i \log|x_i|, \qquad \text{s.t. } \boldsymbol{y} = \Phi\boldsymbol{x}, \tag{7}$$

which follows since $\|\boldsymbol{x}\|_0 \equiv \lim_{p\rightarrow 0} \sum_i |x_i|^p$ and $\lim_{p\rightarrow 0} \frac{1}{p}\sum_i (|x_i|^p - 1) = \sum_i \log|x_i|$. The penalty function in (7) can be upper-bounded using

$$\log|x_i| \;\leq\; \frac{1}{2}\log|x_i^2 + \epsilon| \;\leq\; \frac{x_i^2+\epsilon}{2\gamma_i} + \frac{1}{2}\log\gamma_i - \frac{1}{2}, \tag{8}$$

where $\epsilon, \gamma_i \geq 0$ are arbitrary. The second inequality, which follows directly from the concavity of the $\log$ function with respect to x_i^2, becomes an equality iff $\gamma_i = x_i^2 + \epsilon$. Now consider solving

$$\min_{\boldsymbol{x},\boldsymbol{\gamma},\epsilon} \sum_i \left(\frac{x_i^2+\epsilon}{\gamma_i} + \log\gamma_i\right), \;\; \text{s.t. } \boldsymbol{y} = \Phi\boldsymbol{x}, \epsilon \geq 0, \gamma_i \geq 0, \forall i, \tag{9}$$

where $\boldsymbol{\gamma} \triangleq [\gamma_1, \ldots, \gamma_m]^T$. Again, there is a one-to-one correspondence between local minima of the original problem (1) and local minima of (9). For fixed ϵ and $\boldsymbol{\gamma}$, the optimal $\boldsymbol{x}$ satisfies $\boldsymbol{x} = \Gamma\Phi^T\left(\Phi\Gamma\Phi^T\right)^{-1}\boldsymbol{y}$, with $\Gamma \triangleq \mathrm{diag}[\boldsymbol{\gamma}]$ and, from above, the minimizing $\boldsymbol{\gamma}$ for fixed $\boldsymbol{x}$ and ϵ is $\gamma_i = x_i^2 + \epsilon$, $\forall i$. By construction, coordinate descent over $\boldsymbol{x}$, $\boldsymbol{\gamma}$, and ϵ is guaranteed to reduce or leave unchanged (9) (the exact strategy for reducing ϵ is not crucial).

It can be shown that these updates, which are equivalent to Chartrand and Yin's ℓ_2 reweighting algorithm with $w_i^{(k)} = \gamma_i$, are guaranteed to converge monotonically to a local minimum (or saddle point) of (9) by satisfying the conditions of the *Global Convergence Theorem* (see for example [23]). At any such solution, $\epsilon = 0$ and we will also be at a local minimum (or saddle point) of (7). In the unlikely event that a saddle point is reached (such solutions to (7) are very rare [17]), a small perturbation leads to a local minimum.

Of course obviously we could set $\epsilon \rightarrow 0$ in the very first iteration, which reproduces the FOCUSS algorithm and its attendant quadratic rate of convergence near a minimum [17]; however, compelling evidence from [4] suggests that slow reduction of ϵ is far more effective in avoiding suboptimal local minima troubles.

The weight update (6) is part of a wider class given by

$$w_i^{(k+1)} \rightarrow \left[\left(x_i^{(k+1)}\right)^2 + \epsilon^{(k+1)}\right]^{1-\frac{p}{2}} \tag{10}$$

where $0 \leq p \leq 2$ is a used-defined parameter. With $p = 0$, we recover (6) and also obtain the best empirical performance solving (1) according to experiments in [4]; other values for p lead to alternative implicit cost functions and convergence properties. Additionally, as brought to our attention by a reviewer, for a carefully chosen $\epsilon^{(k+1)}$ update, interesting

[2]Square-concavity implies that $g(|x_i|) = f(x_i^2)$, where $f(\cdot)$ is concave.

and detailed convergence results are possible using (10), particularly for the special case where $p = 1$ which produces a robust means of finding a minimum ℓ_1 norm solution using reweighted ℓ_2 [6]. However, the selection for $\epsilon^{(k+1)}$ used to obtain these results may be suboptimal in certain situations relative to other prescriptions for choosing p and ϵ (see Section IV-A below).[3] Regardless, the underlying analysis from [6] provides useful insights into reweighted ℓ_2 algorithms.

B. ℓ_1 Reweighting Method of Candès et al. (2008) [3]

An interesting example of separable iterative ℓ_1 reweighting is presented in [3] where the selection

$$w_i^{(k+1)} \rightarrow \left| x_i^{(k+1)} \right| + \epsilon \tag{11}$$

is suggested. Here ϵ is generally chosen as a fixed, application-dependent constant. In the noiseless case, it is demonstrated based on [9] that this amounts to iteratively solving

$$\min_{\boldsymbol{x}} \sum_i \log\left(|x_i| + \epsilon\right), \qquad \text{s.t. } \boldsymbol{y} = \Phi \boldsymbol{x}. \tag{12}$$

The FOCUSS algorithm [17] can also be viewed as an iterative reweighted ℓ_2 method for locally solving (12) for the special case when $\epsilon = 0$; however, Candès et al. point out that just a few iterations of their method is far more effective in finding sparse solutions than FOCUSS. This occurs because, with $\epsilon = 0$, the cost function (12) has on the order of $\binom{m}{n}$ deep local minima and so convergence to a suboptimal one is highly likely. Here we present reweighted ℓ_2 updates for minimizing (12) for arbitrary ϵ. Using results from [15], it is relatively straightforward to show that

$$\begin{aligned} &\log\left(|x_i| + \epsilon\right) \quad \leq \qquad (13) \\ &\frac{x_i^2}{\gamma_i} + \log\left[\frac{\left(\epsilon^2 + 2\gamma_i\right)^{\frac{1}{2}} + \epsilon}{2}\right] - \frac{\left[\left(\epsilon^2 + 2\gamma_i\right)^{\frac{1}{2}} - \epsilon\right]^2}{4\gamma_i} \end{aligned}$$

for all $\epsilon, \gamma_i \geq 0$ with equality iff $\gamma_i = 2(x_i^2 + \epsilon|x_i|)$. Note that this is a very different surrogate cost function than (8) and utilizes the concavity of $\log\left(|x_i| + \epsilon\right)$, as opposed to $\log(x_i^2 + \epsilon)$, with respect to x_i^2. Using coordinate descent as before with $w_i^{(k)} = \gamma_i$ leads to the reweighted ℓ_2 iteration

$$w_i^{(k+1)} \rightarrow 2\left[\left(x_i^{(k+1)}\right)^2 + \epsilon \left| x_i^{(k+1)} \right|\right], \tag{14}$$

which will reduce (or leave unchanged) (12) for arbitrary $\epsilon \geq 0$. In preliminary empirical tests, this method is superior to regular FOCUSS and could be used as an alternative to reweighted ℓ_1 if computational resources for computing ℓ_1 solutions are limited. Additionally, the most direct comparison between reweighted ℓ_1 and ℓ_2 in this context would involve empirical tests using (14) versus the method from Candès et al., a subject addressed in [22]. Also, it would be worthwhile to compare (14) using a decreasing ϵ update with (6) since both are derived from different implicit bounds on $\log|x_i|$.

[3]Note that the convergence analysis we discuss above applies for any sequence of $\epsilon^{(k)} \rightarrow 0$ and can be extended to other values of $p \in [0, 1)$.

III. Non-Separable Reweighting Schemes

Non-separable selections for $W^{(k)}$ allow us to minimize cost functions based on general, non-separable sparsity penalties, meaning penalties that cannot be expressed as a summation over functions of the individual coefficients as in (5). Such penalties potentially have a number of desirable properties [20], [21]. One particularly useful non-separable penalty is given by

$$g_\alpha(|\boldsymbol{x}|) \triangleq \min_{\boldsymbol{\gamma} \geq 0} \; \boldsymbol{x}^T \Gamma^{-1} \boldsymbol{x} + \log|\alpha I + \Phi \Gamma \Phi^T|, \tag{15}$$

where $\alpha \geq 0$, $\Gamma \triangleq \text{diag}[\boldsymbol{\gamma}]$, and $|\boldsymbol{x}| \triangleq [|x_1|, \ldots, |x_m|]^T$. The motivation for (15) is derived from a dual-space view [20] of sparse Bayesian learning [18] and automatic relevance determination [14]. The analysis in [20] reveals that replacing $\|\boldsymbol{x}\|_0$ with $g_\alpha(|\boldsymbol{x}|)$ and $\alpha \rightarrow 0$ leaves the globally minimizing solution to (1) unchanged but drastically reduces the number of local minima (more so than *any possible* separable penalty function). It can be shown that $g_\alpha(|\boldsymbol{x}|)$ is a non-decreasing concave function of both $|\boldsymbol{x}|$ and $\boldsymbol{x}^2 \triangleq [x_1^2, \ldots, x_m^2]^T$ [21], therefore (perhaps not surprisingly) minimization can be accomplished using either iterative reweighted ℓ_2 or ℓ_1.

A. ℓ_2 Reweighting

There exist multiple ways to handle ℓ_2 reweighting in terms of the non-separable penalty function (15) for the purpose of solving

$$\min_{\boldsymbol{x}} \|\boldsymbol{y} - \Phi \boldsymbol{x}\|_2^2 + \lambda g_\alpha(|\boldsymbol{x}|). \tag{16}$$

Different sets of upper-bounding auxiliary functions (which are tight in different regions of $\boldsymbol{x}$ space) lead to different choices for $W^{(k)}$ with different properties. One useful variant that reveals a close connection with Chartand and Yin's method can be derived as follows. Using standard determinant identities we get

$$\begin{aligned} g_\alpha(|\boldsymbol{x}|) &\leq \boldsymbol{x}^T \Gamma^{-1} \boldsymbol{x} + \log|\alpha I + \Phi \Gamma \Phi^T| \\ &= \boldsymbol{x}^T \Gamma^{-1} \boldsymbol{x} + \log|\Gamma| + \log|\alpha^{-1} \Phi^T \Phi + \Gamma^{-1}| \\ &\quad + n \log \alpha \\ &\leq \boldsymbol{x}^T \Gamma^{-1} \boldsymbol{x} + \log|\Gamma| + \boldsymbol{z}^T \boldsymbol{\gamma}^{-1} - h^*(\boldsymbol{z}) + n \log \alpha \\ &= n \log \alpha - h^*(\boldsymbol{z}) + \sum_i \left(\frac{x_i^2 + z_i}{\gamma_i} + \log \gamma_i \right) \end{aligned} \tag{17}$$

where $\boldsymbol{\gamma}^{-1} \triangleq [\gamma_1^{-1}, \ldots, \gamma_m^{-1}]^T$ and $h^*(\boldsymbol{z})$ denotes the concave conjugate [4] of $\log|\alpha^{-1} \Phi^T \Phi + \Gamma^{-1}|$ with respect to Γ^{-1}. The bound holds for all non-negative vectors $\boldsymbol{z} \triangleq [z_1, \ldots, z_m]^T$. Coordinate descent over $\boldsymbol{x}$, $\boldsymbol{\gamma}$, and $\boldsymbol{z}$ (again with $w_i^{(k)} = \gamma_i$) then leads to the weight update

[4]The concave conjugate is a function that arises from convex analysis and duality theory [1]; in this instance it is computed via

$$h^*(\boldsymbol{z}) = \min_{\boldsymbol{\gamma} \geq 0} \boldsymbol{z}^T \boldsymbol{\gamma}^{-1} - \log|\alpha^{-1} \Phi^T \Phi + \Gamma^{-1}|.$$

$$
\begin{aligned}
w_i^{(k+1)} &\rightarrow \left(x_i^{(k+1)}\right)^2 + z_i^{(k+1)} \\
&= \left(x_i^{(k+1)}\right)^2 + w_i^{(k)} I \; - \\
&\qquad w_i^{(k)} \boldsymbol{\phi}_i^T \left(\alpha I + \Phi W^{(k)} \Phi^T\right)^{-1} \boldsymbol{\phi}_i w_i^{(k)},
\end{aligned} \tag{18}
$$

which can be computed in $O(n^2 m)$, the same expense as each solution to (3). Note that since each weight update is dependent on previous weight updates, it is implicitly dependent on previous values of $\boldsymbol{x}$, unlike in the separable case above.

The form of (18) is very similar to the one used by Chartrand and Yin. Basically, if we allow for a separate ϵ_i for each coefficient x_i, then the update (6) is equivalent to the selection

$$
\epsilon_i^{(k+1)} \rightarrow w_i^{(k)} I - w_i^{(k)} \boldsymbol{\phi}_i^T \left(\alpha I + \Phi W^{(k)} \Phi^T\right)^{-1} \boldsymbol{\phi}_i w_i^{(k)}. \tag{19}
$$

Moreover, the implicit auxiliary function from (8) being minimized by Chartrand and Yin's method has the exact same form as (17); with the latter, coefficients that are interrelated by a non-separable penalty term are effectively decoupled when conditioned on the auxiliary variables $\boldsymbol{z}$. And recall that one outstanding issue with Chartrand and Yin's approach is the optimal schedule for reducing $\epsilon^{(k)}$, which could be application dependent and potentially sensitive. So in this regard, (19) can be viewed as a principled way of selecting ϵ so as to avoid, where possible, convergence to local minima. In preliminary experiments, this method performs as well or better than the heuristic ϵ-selection strategy from [4] (see Section IV-A).

B. ℓ_1 Reweighting

Using similar methodology, a reweighted ℓ_1 implementation is easily derived with desirable convergence properties [20]. The requisite $W^{(k+1)}$ update is given by

$$
\begin{aligned}
w_i^{(k+1)} &\rightarrow \\
&\left[\boldsymbol{\phi}_i^T \left(\alpha I + \Phi W^{(k)} \mathrm{diag}\left[\left|\boldsymbol{x}^{(k+1)}\right|\right] \Phi^T\right)^{-1} \boldsymbol{\phi}_i\right]^{-\frac{1}{2}}
\end{aligned} \tag{20}
$$

for all i. This procedure is guaranteed to aid performance in the sense described by the following two results, which apply in the case where $\alpha \rightarrow 0, \lambda \rightarrow 0$ (zero noise limit):

Theorem 1: When applying iterative reweighted ℓ_1 using (20), the solution sparsity satisfies $\|\boldsymbol{x}^{(k+1)}\|_0 \leq \|\boldsymbol{x}^{(k)}\|_0$.

So continued iteration can never do worse. The proof is contained in [22]. Now define $\mathrm{spark}(\Phi)$ as the smallest number of linearly dependent columns in Φ. It follows then that $2 \leq \mathrm{spark}(\Phi) \leq n+1$.

Theorem 2: Assume that $\mathrm{spark}(\Phi) = n+1$. For any instance where standard ℓ_1 minimization fails to find some $\boldsymbol{x}^*$ drawn from support set $\mathcal{S}$ with cardinality $|\mathcal{S}| < \frac{(n+1)}{2}$, there exists a set of signals $\boldsymbol{y}$ (with non-zero measure) generated from $\mathcal{S}$ such that non-separable reweighted ℓ_1, with $W^{(k)}$ updated using (20), always succeeds but standard ℓ_1 always fails.

This result is proved in [22]. Note that Theorem 2 does not in any way indicate what is the best non-separable reweighting scheme in practice (for example, in our limited experience with empirical simulations, the selection $\alpha = 0$ is not necessarily always optimal). However, it does suggest that reweighting is potentially effective in a variety of situations.

Before proceeding to empirical comparisons, it is worth relating Theorem 2 with results from Davies et al. (2008) [5], where it is shown that for any sparsity level, there will always exist cases (albeit of measure zero) where, if standard ℓ_1 minimization fails, any *admissible* ℓ_1 reweighting strategy will also fail. In this context a reweighting scheme is said to be admissible if: (i) $w_i^{(1)} = 1$ for all i and, (ii) there exists a $w_{min}^{(k)} > 0$ such that for all k and i, $w_i^{(k)} \geq w_{min}^{(k)}$ and if $x_i^{(k)} = 0$, then $w_i^{(k)} = w_{min}^{(k)}$.[5]

Interestingly, the non-separable reweighting from (20) does *not* satisfy this definition despite its effectiveness in practice (see Section IV-B below); however, in [22] we suggest slightly modified versions of admissibility that accommodate a wider class of useful algorithms. While details are deferred, more general reweighting strategies can sometimes be advantageous for avoiding local minima. Regardless, we stress the contrasting nature of Davies et al.'s result versus our Theorem 2. The former demonstrates that on a set of measure zero in $\boldsymbol{x}$ space, a large class of reweighting schemes will not improve upon basic ℓ_1 minimization, while the latter specifies that on a different set of nonzero measure on $\mathcal{S}$, some non-separable reweighting will always do better.

IV. Empirical Results

This section contains a few brief experiments involving the various reweighting schemes discussed herein. First we include a comparison of ℓ_2 approaches followed by an ℓ_1 example involving non-negative sparse coding.

A. Reweighted ℓ_2 Example

Monte-Carlo simulations were conducted similar to those performed in [4], [6] allowing us to compare the separable method of Chartrand and Yin with the non-separable update (18) using $\alpha = 0$. As discussed above, these two methods differ only in the effective choice of the ϵ parameter. We also include results from the related method in Daubechies et al. using $p = 1$, which gives us the basis pursuit (minimum ℓ_1 norm) solution, and $p = 0.6$ which works well in conjunction with the proscribed ϵ update based on the simulations from [6]. Note that the optimal value of p and ϵ for sparse recovery purposes can be interdependent and [6] reports poor results with p much smaller than 0.6 when using their ϵ update.

[5] Note that we have modified this definition slightly to account for the fact that we define our weights as the reciprocal of those used in [5].

Additionally, there is an additional parameter K associated with Daubechies et al.'s algorithm that must be set; we used the heuristic taken from the authors' Matlab code.[6]

The experimental particulars are as follows: First, a random, overcomplete 50×250 dictionary Φ is created with iid unit Gaussian elements and ℓ_2 normalized columns. Next, sparse weight vectors $\boldsymbol{x}^*$ are randomly generated with the number of nonzero entries varied to create different test conditions. Nonzero amplitudes are drawn iid from one of two experiment-dependent distributions. Signals are then computed as $\boldsymbol{y} = \Phi\boldsymbol{x}^*$. Each algorithm is presented with $\boldsymbol{y}$ and Φ and attempts to estimate $\boldsymbol{x}^*$ using an initial weighting of $w_i^{(1)} = 1, \forall i$. In all cases, we ran 1000 independent trials and compared the number of times each algorithm failed to recover $\boldsymbol{x}^*$. Under the specified conditions for the generation of Φ and $\boldsymbol{y}$, all other feasible solutions $\boldsymbol{x}$ almost surely have a sparsity less than $\boldsymbol{x}^*$, so our synthetically generated coefficients must be maximally sparse.

Figure 1 displays results where the nonzero elements in $\boldsymbol{x}^*$ were drawn with unit magnitudes. The performance of four algorithms is shown: the three separable methods discussed above and the non-separable update given by (18). For algorithms with non-convex underlying sparsity penalties, unit magnitude coefficients can be much more troublesome than other distributions because local minima may become more pronounced or numerous [21]. In contrast, the performance will be independent of the nonzero coefficient magnitudes when minimizing the ℓ_1 norm (i.e., the $p = 1$ case) [13], so we expect this situation to be most advantageous to the ℓ_1 norm solution relative to the others. Nevertheless, from the figure we observe that the non-separable reweighting still performs best; out of the remaining separable examples, the $p = 1$ case is only slightly superior.

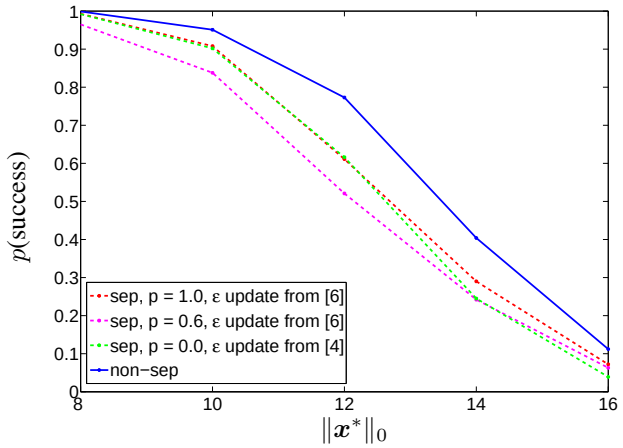

Fig. 1. Iterative reweighted ℓ_2 example using *unit magnitude* nonzero coefficients. Probability of success recovering sparse coefficients for different sparsity values, i.e., $\|\boldsymbol{x}^*\|_0$.

Figure 2 reproduces the same experiment with nonzero elements in $\boldsymbol{x}^*$ now drawn from a unit Gaussian distribution. The performance of the $p = 1$ separable algorithm is unchanged as expected; however, the others all improve significantly, especially the non-separable update and Chartrand and Yin's method. Note that we have also included a second non-separable variant labeled 'non-sep+' in which we added an additional decaying regularization term to (18). Although we adopted a simple heuristic for this task (like Chartrand and Yin's method), based on results in the figure, this demonstrates that better performance is definitely possible and that (18) is by no means optimal. Of course presumably all of these methods could be improved through additional modifications and tuning (e.g., a simple hybrid scheme is suggested in [6] that involves reducing p after a 'burn-in' period that improves recovery probabilities marginally); however, we save thorough evaluation of such extensions to future exploration.

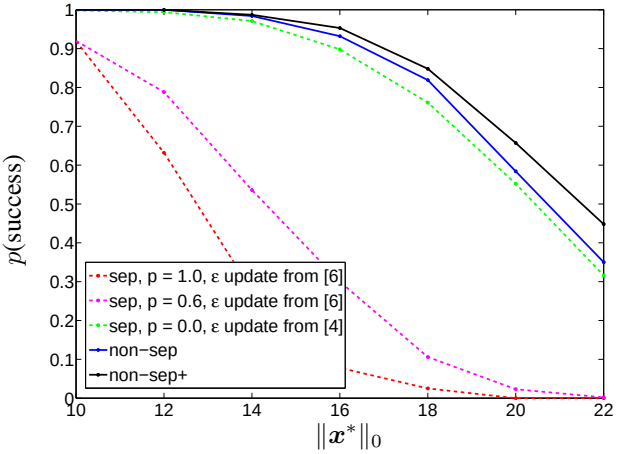

Fig. 2. Iterative reweighted ℓ_2 example using *Gaussian distributed* nonzero coefficients. Probability of success recovering sparse coefficients for different sparsity values, i.e., $\|\boldsymbol{x}^*\|_0$.

B. Non-Negative Sparse Coding Example via Reweighted ℓ_1 Minimization

As an example of non-separable ℓ_1 reweighting, we performed a simple experimental test similar to the previous section. Each trial consisted of generating a 50×100 dictionary Φ with iid Gaussian entries and a sparse vector $\boldsymbol{x}^*$ with 30 nonzero, non-negative coefficients. A signal is computed using $\boldsymbol{y} = \Phi\boldsymbol{x}^*$ as above. We then attempted to recover $\boldsymbol{x}^*$ by applying non-separable ℓ_1 reweighting to $\boldsymbol{y}$ and Φ, with a non-negativity constraint applied to (4) at each iteration and using $\alpha = 1$. This selection was found to work somewhat better than $\alpha = 0$ (as was used for reweighted ℓ_2), although we have not fully explored what the optimal value might be. Since we are working with a noise-free signal, we also assume $\lambda \to 0$ and so the requisite coefficient update becomes

$$\boldsymbol{x}^{(k+1)} \to \arg\min_{\boldsymbol{x}} \sum_i \frac{x_i}{w_i^{(k)}}, \quad \text{s.t. } \boldsymbol{y} = \Phi\boldsymbol{x}, \;\; x_i \geq 0, \forall i, \tag{21}$$

which can be solved using a straightforward linear program. Assuming $W^{(1)} = I$, then the first iteration amounts to the

[6]http://www.ricam.oeaw.ac.at/people/page/fornasier/menu3.html

non-negative minimum ℓ_1 solution. Results from 1000 random trials are displayed in Figure 3, where standard non-negative ℓ_1 succeeds less than 40% of the time; however, with only four reweighted iterations using (20), success improved to almost 90%. This demonstrates both the efficacy of reweighting and the ability to handle constraints on $\boldsymbol{x}$.

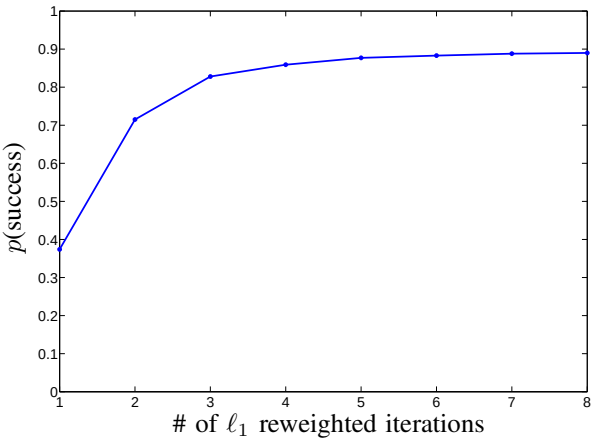

Fig. 3. Non-negative sparse coding example. Probability of success recovering sparse, non-negative coefficients as the number of reweighted ℓ_1 iterations is increased.

V. Discussion

In this paper we have briefly explored various iterative reweighting schemes for solving sparse linear inverse problems, elaborating on a distinction between separable and non-separable weighting functions. Although a large number of separable algorithms have been proposed in the literature, the non-separable case is relatively uncommon and, on the surface, may appear much more difficult to work with. However, iterative reweighted ℓ_1 and ℓ_2 approaches provide a convenient means of decoupling coefficients via auxiliary variables leading to efficient updates that can potentially be related to existing separable schemes. In general, a variety of different algorithms are possible by forming different upper-bounding auxiliary functions. While the non-separable algorithms we have discussed show considerable promise, we envision that superior strategies and interesting extensions are very possible.

One direction of future research is the application and analysis of various reweighting schemes to the simultaneous sparse approximation problem, which is a special case of covariance component estimation [20]. In this scenario we are presented with multiple signals $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots$ that we assume were produced by coefficient vectors $\boldsymbol{x}_1^*, \boldsymbol{x}_2^*, \ldots$ characterized by the same sparsity profile or support. All of the algorithms discussed herein can naturally be expanded to this domain by applying an ℓ_2 norm penalty to the aligned elements of each coefficient vector. With reweighted ℓ_2, this transition is very straightforward and easy to implement. In contrast, reweighted ℓ_1 requires the solution of a second-order-cone program at each iteration. Preliminary results indicate that non-separable methods can significantly widen their performance advantage over their separable counterparts in this domain.

Acknowledgment

This research was supported by NIH grants R01DC04855 and R01DC006435. Also, thanks to the reviewers for suggesting reference [6] and other useful comments.

References

[1] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[2] A.M. Bruckstein, M. Elad, and M. Zibulevsky, "A non-negative and sparse enough solution of an underdetermined linear system of equations is unique," *IEEE Trans. Information Theory*, vol. 54, no. 11, pp. 4813–4820, Nov. 2008.

[3] E. Candès, M. Wakin, and S. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *To appear in J. Fourier Anal. Appl.*, 2008.

[4] R. Chartrand and W. Yin, "Iteratively reweighted algorithms for compressive sensing," *Proc. Int. Conf. Accoustics, Speech, and Signal Proc.*, 2008.

[5] M. Davies and R. Gribonval, "Restricted isometry constants where ℓ_p sparse recovery can fail for $0 < p \leq 1$," *Technical Report IRISA*, no. 1899, July 2008.

[6] I. Daubechies, R. DeVore, M. Fornasier, and S. Gunturk, "Iteratively re-weighted least squares minimization for sparse recovery," to appear in *Commun. Pure Appl. Math.*, 2009.

[7] D. Donoho, "For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution," *Stanford University Technical Report*, September 2004.

[8] D. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization," *Proc. National Academy of Sciences*, vol. 100, no. 5, pp. 2197–2202, March 2003.

[9] M. Fazel, H. Hindi, and S. Boyd, "Log-Det heuristic for matrix rank minimization with applications to Hankel and Euclidean distance matrices," *Proc. American Control Conf.*, vol. 3, pp. 2156–2162, June 2003.

[10] M. Figueiredo, "Adaptive sparseness using Jeffreys prior," *Advances in Neural Information Processing Systems 14*, pp. 697–704, 2002.

[11] J. Fuchs, "On sparse representations in arbitrary redundant bases," *IEEE Transactions on Information Theory*, vol. 50, no. 6, pp. 1341–1344, June 2004.

[12] R. Gribonval and M. Nielsen, "Sparse representations in unions of bases," *IEEE Transactions on Information Theory*, vol. 49, pp. 3320–3325, Dec. 2003.

[13] D. Malioutov, M. Çetin, and A. Willsky, "Optimal sparse representations in general overcomplete bases," *IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 2, pp. II–793–796, May 2004.

[14] R. Neal, *Bayesian Learning for Neural Networks*. New York: Springer-Verlag, 1996.

[15] J. Palmer, D. Wipf, K. Kreutz-Delgado, and B. Rao, "Variational EM algorithms for non-Gaussian latent variable models," *Advances in Neural Information Processing Systems 18*, pp. 1059–1066, 2006.

[16] B. Rao, K. Engan, S. F. Cotter, J. Palmer, and K. Kreutz-Delgado, "Subset selection in noise based on diversity measure minimization," *IEEE Trans. Signal Processing*, vol. 51, no. 3, pp. 760–770, March 2003.

[17] B. Rao and K. Kreutz-Delgado, "An affine scaling methodology for best basis selection," *IEEE Transactions on Signal Processing*, vol. 47, no. 1, pp. 187–200, January 1999.

[18] M. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Machine Learning Research*, vol. 1, pp. 211–244, 2001.

[19] J. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Transactions on Information Theory*, vol. 50, no. 10, pp. 2231–2242, October 2004.

[20] D. Wipf and S. Nagarajan, "A new view of automatic relevance determination," *Advancies in Neural Information Processing Systems 20*, 2008.

[21] ——, "Latent variable Bayesian models for promoting sparsity," *UC San Francisco Technical Report*, 2008.

[22] ——, "Iterative Reweighted ℓ_1 and ℓ_2 Methods for Finding Sparse Solutions," *UC San Francisco Technical Report*, 2008.

[23] W. Zangwill, *Nonlinear Programming: A Unified Approach*. New Jersey: Prentice Hall, 1969.