

ANALYSIS AND GENERALIZATIONS OF THE LINEARIZED BREGMAN METHOD*

WOTAO YIN[†]

Abstract. This paper reviews the Bregman methods, analyzes the linearized Bregman method, and proposes fast generalization of the latter for solving the basis pursuit and related problems. The analysis shows that the linearized Bregman method has the exact penalty property, namely, it converges to an exact solution of the basis pursuit problem if and only if its regularization parameter α is greater than a certain value. The analysis is based on showing that the linearized Bregman algorithm is equivalent to gradient descent applied to a certain dual formulation. This result motivates generalizations of the algorithm enabling the use of gradient-based optimization techniques such as line search, Barzilai-Borwein steps, L-BFGS, and nonlinear conjugate gradient steps. In addition, the paper discusses the selection and update of α . The analysis and discussions are limited to the ℓ_1 -norm but can be extended to other ℓ_1 -like functions.

Key words. Bregman, linearized Bregman, compressed sensing, ℓ_1 minimization, basis pursuit.

AMS subject classifications. 68U10, 65K10, 90C25, 90C51.

1. Introduction. Let $A \in \mathbb{R}^{m \times n}$ for $m < n$ (and sometimes, $m \ll n$ in compressed sensing), $b \in \mathbb{R}^m$, and $x \in \mathbb{R}^n$. The linearized Bregman method was introduced in [65] and extended or analyzed in [48, 11, 12] to approximately solve the basis pursuit problem

$$(1.1) \quad \min\{\|x\|_1 : Ax = b\},$$

which determines an ℓ_1 -minimal solution x_{opt} of the underdetermined linear system $Ax = b$. This problem arises in many applications, and in particular, in the recently emerging application of compressed sensing, which was brought to the forefront by Donoho [22] and Candes, Romberg, and Tao [14].

The *linearized* Bregman method is a variant of the *original* Bregman method introduced in [47, 65], and both Bregman methods can be applied to solve (1.1). They are reviewed in Subsection 1.2 below. Previous analysis in [11, 12] shows that the linearized Bregman method generates a sequence of points converging to x_α , the unique solution of

$$(1.2) \quad \min\{\|x\|_1 + \frac{1}{2\alpha}\|x\|^2 : Ax = b\},$$

where $\|x\| := \|x\|_2$ is the Euclidean of x . In this paper, we fully analyze (1.2), as well as the linearized Bregman method, and study how its solution x_α vary in terms of α .

1.1. Contributions. The first and main contribution of this report is to show that there exists a finite scalar $\alpha_\infty > 0$ such that the unique solution x_α of (1.2) is also a solution of (1.1) whenever $\alpha > \alpha_\infty$. Since (1.2) can be viewed as an quadratically regularized problem of (1.1), we call this result the exact regularization property.

*This research was supported in part by NSF CAREER Award DMS-07-48839, ONR Grant N00014-08-1-1101, and an Alfred P. Sloan Research Fellowship.

[†]Department of Computational and Applied Mathematics, Rice University, 6100 Main Street, MS-134, Houston, Texas, 77005, U.S.A. (wotao.yin@rice.edu).

The second contribution is to show that the linearized Bregman method is equivalent to the gradient descent method applied to the dual of (1.2)

$$(1.3) \quad \min_y -b^\top y + \frac{\alpha}{2} \left(\sum_{i=1}^n \max\{(A^\top y)_i - 1, 0\}^2 + \min\{(A^\top y)_i + 1, 0\}^2 \right).$$

which is a quadratic penalty problem for the Lagrange dual of (1.1):

$$(1.4) \quad \min\{-b^\top y : \|A^\top y\|_\infty \leq 1\}.$$

The constraint $\|A^\top y\|_\infty \leq 1$ can be rewritten as $(A^\top y)_i - 1 \leq 0$ and $(A^\top y)_i + 1 \geq 0$, for $i = 1, \dots, n$. The violation of each of these constraints is quadratically penalized in problem (1.3).

Following from the above result, we propose enhancements to the linearized Bregman method. Because problem (1.3) is unconstrained and its objective function is Lipschitz continuously differentiable, it appears easier to solve than (1.2) by standard optimization methods such as various line search methods, quasi-Newton methods, L-BFGS [42], and nonlinear conjugate gradient methods. Therefore, the linearized Bregman method can be improved by embracing more powerful optimization techniques.

Another contribution is to show it is trivial to get x_α from any dual solution y_α of (1.3) via

$$(1.5) \quad x_\alpha := \alpha(A^\top y_\alpha - \text{Proj}_{[-1,1]^n}(A^\top y_\alpha)).$$

This formulation was implicitly used in the linearized Bregman method without being explicitly stated.

In the review of the Bregman methods, we demonstrate the superb error tolerance of the original Bregman method and give a brief explanation.

As this paper focuses on numerical analysis rather than actual implementation of the linearized Bregman method, we propose practical procedures for updating α and identifying a sufficiently large α for solving problem (1.1). Numerical experiments for various instances of (1.1), including a set of compressed sensing problems, will be performed in future research.

The above results can be extended to more general problems of the form

$$(1.6) \quad \min\{J(x) : Ax = b\},$$

where J is certain ℓ_1 -like regularization functions. The author believes that similar results can be obtained for J being the nuclear norm of a matrix x , which is used in an optimization problem [10] similar to (1.6) for the matrix completion problem. The quadratic penalty function used in (1.3) is also extended to certain convex differentiable functions.

1.2. The Original and Linearized Bregman Methods.

1.2.1. Bregman Iterative Regularization. The original Bregman method (formally called the Bregman iterative regularization method) was introduced in [47], not for solving the basis pursuit problem, but in the context of total variation based image processing; it was then extended to wavelet-based denoising [61], nonlinear inverse scale space in [8, 9], and MR imaging [37]. Recently, its usefulness in compressed sensing, where ℓ_1 and ℓ_1 -related (such as total variation) minimization problems are solved, was pointed out in [65]. In this subsection, we briefly review the application of the method in imaging.

Let u represent an unknown greyscale image. The total variation of u is defined by

$$(1.7) \quad J(u) := \mu TV(u) = \mu \int \|\nabla u\|.$$

It is worth noting that the above definition is valid only if ∇u exists, and in this case, it equals the L^1 norm of $\|\nabla u(x)\|$ over the image domain. Formally the total variation of u is defined using a dual formulation (cf. [69]).

The authors of [47] extend the Rudin-Osher-Fatemi [53] model

$$(1.8) \quad \min_u \mu TV(u) + \frac{1}{2} \|u - b\|^2,$$

where b is typically an input noisy measurement of a clean image \bar{u} , and μ is a tuning parameter, into an iterative regularization model (1.10) below by replacing the term $\mu TV(u)$ with its Bregman distance. The Bregman distance [7] with respect to a convex functional $J(\cdot)$ between points u and v is defined as

$$(1.9) \quad D_J^p(u, v) = J(u) - J(v) - \langle p, u - v \rangle$$

where $p \in \partial J(v)$ is some subgradient of J at v . Because $D_J^p(u, v) \neq D_J^p(v, u)$ in general, $D_J^p(u, v)$ is not a distance in the usual sense. However, it measures the closeness between u and v in the sense that $D_J^p(u, v) \geq 0$, and $D_J^p(u, v) \geq D_J^p(w, v)$ for all points w on the line segment connecting u and v .

Instead of solving (1.8) once, the original Bregman method solves a sequence of convex problems

$$(1.10) \quad u^{k+1} \leftarrow \min_u D_J^{p^k}(u, u^k) + \frac{1}{2} \|u - b\|^2$$

for $k = 0, 1, \dots$ starting with $u^0 = \mathbf{0}$ and $p^0 = \mathbf{0}$ (hence, for $k = 0$, one indeed solves the original problem (1.8).) Since $\mu TV(u)$ is not differentiable everywhere, the subdifferential of $\mu TV(u)$ may contain more than one element, leading to many possible choices for p in (1.9). However, p can be uniquely determined from one iteration to the next. From the optimality of u^{k+1} in (1.10), it follows that $\mathbf{0} \in \partial J(u^{k+1}) - p^k + u^{k+1} - b$; hence, in [47] p^{k+1} is set according to

$$p^{k+1} := p^k + b - u^{k+1},$$

The difference between (1.8) and (1.10) is in the use of regularization. While (1.8) regularizes u by directly minimizing its total variation, (1.10) regularizes u by minimizing the total variation-based Bregman distance of u to a previous solution u^k .

In [47] two key results for the sequence $\{u^k\}$ were proved. First, $\|u^k - b\|$ converges to 0 monotonically; second, u^k also monotonically gets closer to \bar{u} , the *unknown* noiseless image, in terms of the Bregman distance $D_{TV}^{p^k}(\bar{u}, u^k)$, at least while $\|u^k - b\| \geq \|\bar{u} - b\|$. In fact, using the results in the paper [65], one can show that replacing $\|u - b\|$ by $\|Au - b\|$ in (1.10) yields a sequence of u^k that converges to a solution of

$$(1.11) \quad \min\{TV(u) : Au = b\},$$

in a finite number of steps. Since A is an identity matrix in the denoising problem, u^k simply converges to b in a finite number of steps. Recall that in [47], the original Bregman method is developed, not to solve (1.11), but to obtain a clean image that approximates the noisy image b . Therefore, based on the second key result

above, the iterations are stopped before u^k converges subject to the stopping rule $\|u^k - b\| < \sigma \approx \|\bar{u} - b\|$, where σ is an estimate of the noise level. Numerical results in [47] demonstrate that for this stopping rule, Bregman iterations yield an image that has remarkably better quality over the original model (1.8). As such, the original Bregman method is also called the Bregman iterative regularization method to emphasize the novelty of the method as a new regularization method.

One of the interesting ways to view the original Bregman method is as an “adding back the residual” method. Interestingly, not only for the first iteration $k = 0$, but for all k , the new problem (1.10) can be reduced to the original problem (1.8) with the input $b^{k+1} := b + (b^k - u^k)$ starting with $b^0 = u^0 = \mathbf{0}$, i.e., the iterations (1.10) are equivalent to

$$(1.12) \quad u^{k+1} \leftarrow \min_u J(u) + \frac{1}{2} \|u - b^{k+1}\|^2, \text{ where } b^{k+1} = b + (b^k - u^k),$$

and can be carried out using any existing algorithms for (1.8). The derivation is trivial.

The iterative procedure (1.12) has an intriguing interpretation: Let ω represent the noise in b , i.e., $b = \bar{u} + \omega$, and μ be large. At $k = 0$, $b^k - u^k = \mathbf{0}$, so (1.12) decomposes the input noisy image b into $u^1 + v^1$. Since μ is large, the resulting image u^1 is over smoothed (by total variation minimization) so it does not contain any noise. Consequently, u^1 can be considered to be a portion of the original clean image \bar{u} . The residual $v^1 = b - u^1 = (\bar{u} - u^1) + \omega$, hence, is the sum of the unrecovered “good” signal $(\bar{u} - u^1)$ and the “bad” noise ω . We wish to recover $(\bar{u} - u^1)$ from v^1 . Intuitively, one would next consider letting v^1 be the new input for (1.8) and solving it. However, the original Bregman method turns out to be both better and “nonintuitive”: it adds v^1 back to the original input b . The the new input of (1.12) in the 2nd iteration is

$$b + v^1 = (u^1 + v^1) + v^1 = u^1 + 2(\bar{u} - u^1) + 2\omega.$$

which, compared to the original input $b = u^1 + (\bar{u} - u^1) + \omega$, contains twice as much of both the unrecovered “good” signal $\bar{u} - u^1$ and the “bad” noise ω . What is remarkable is that the new decomposition u^2 is a better approximation to \bar{u} than u^1 (for μ large enough); one explanation is that u^2 not only inherits u^1 but also captures a part of $(\bar{u} - u^1)$, the previously un-captured “good” signal. Of course, as the convergence results indicate, u^k will eventually pick up the noise ω since $\{u^k\}$ converges to $b = \bar{u} + \omega$. However, a high quality image can be found among the sequence $\{u^k\}$: the image u^k that has $\|u^k - b\|$ closest to $\|\bar{u} - b\|$ is cleaner and has a higher contrast than the best image that one could possibly obtain from solving (1.8) one single time.

Formally the iterations of the original Bregman method applied to the problem

$$(1.13) \quad \min_x J(x) + H(x)$$

is given as Table 1.1 in which the Bregman distance $D_J^{b^k}(\cdot, \cdot)$ is defined by (1.9). Here we have switch to using x as the decision variable.

1.2.2. The Original Bregman Method. In [65], the original Bregman method is applied to solving the basis pursuit problem (1.1). Unlike its use for denoising, the iterations are carried out until they converges. It is shown in [65] that the method terminates in a finite number of iterations. That paper also points out the equivalence between the method and the augmented Lagrangian method, which is so called

TABLE 1.1
The original Bregman method for (1.1)

Input: $J(\cdot), H(\cdot)$

1. **Initialize:** $k = 0, x^0 = \mathbf{0}, p^0 = \mathbf{0}$.
 2. **while** stopping conditions not satisfied **do**
 3. $x^{k+1} \leftarrow \arg \min_x D_J^k(x, x^k) + H(x)$
 4. $p^{k+1} \leftarrow p^k - \nabla H(x^{k+1}) \in \partial J(x^{k+1})$
(If possible, replace Steps 3 and 4 by more stable updates, for example, (1.14) and (1.15) for solving (1.1). Correspondingly, set $b^0 = 0$ in Step 1.)
 5. $k \leftarrow k + 1$
 6. **end while**
-

the method of multipliers (cf. [39, 49, 51]). Below, we review the original Bregman method and explain the numerical differences between the two theoretically equivalent methods.

In the original Bregman method applied to (1.1), the regularizer $J(x)$ equals $\mu\|x\|_1$ and the iteration stops when Ax^k and b become almost equal, and ultimately from $Ax^k = b$ and convex duality, it follows that x^k is a minimizer of (1.1). At each iteration, Steps 4 and 5 in Table 1.1 can be equivalently replaced by

$$(1.14) \quad b^{k+1} \leftarrow b + (b^k - Ax^k)$$

$$(1.15) \quad x^{k+1} \leftarrow \min_x \mu\|x\|_1 + \frac{1}{2} \|Ax - b^{k+1}\|^2$$

under variable substitutions. The unconstrained problem in the form of (1.15) is the quadratic penalty problem associated with (1.1), called the basis pursuit denoising problem.

Several recent algorithms based on matrix-vector multiplications involving A and A^\top can efficiently solve (1.15) with large-scale data and sparse solutions. They include iterative soft thresholding (IST) algorithms [28, 46, 21, 3, 15, 23, 25, 20, 17, 35, 5, 60], GPSR [29], SPGL1 [56], ℓ_1 - ℓ_s [41], LASSO [55], FPC_AS [59], Nesterov's algorithm [44], and others. As a classical result in convex optimization, the constrained problem (1.1) can be approximated by

$$(1.16) \quad \min_x \mu\|x\|_1 + \frac{1}{2} \|Ax - b\|^2$$

in the sense that the solution of (1.16) converges to a solution of (1.1) as $\mu \rightarrow 0$; in other words, any method for (1.15) can give an approximate solution of (1.1) by solving (1.16) without carrying out the Bregman iteration. However, solving (1.16) often requires a very small μ , which slows down many algorithms (except for FPC_AS on problems with sparse solutions because of its suboptimization.) With the Bregman iteration, a relatively larger μ can be used to balance the subproblem difficulty and the total number of iterations. In the practice of solving the compressed sensing problems with sparse solutions, the authors of [65] suggest using a moderately large μ and report that only 2–6 Bregman iterations are needed on average.

Turning a constrained problem into a sequence of unconstrained problems is the characteristic of the well-known augmented Lagrangian method, so it is not surprising that the original Bregman method is

theoretically equivalent to the augmented Lagrangian method (see [65], and see [54, 26] in the context of total variation minimization and the split Bregman method). Through transforms, one can show that Step 4 or (1.14) is equivalent to the Lagrange multiplier update in the augmented Lagrangian method. However, the above equivalence holds only if the Bregman subproblem in Step 3 is *exactly* solved. Therefore, the two methods differ numerically. We quote from [65]:

It is interesting that Bregman iterations yield very accurate solutions even if the subproblems are *not* solved as accurately. ... The reason for this remains a subject of further study.

On the other hand, the convergence of the augmented Lagrangian method requires subproblems to be solved with increasing accuracies (cf. [4] and references therein). When x^k has nearly converged, a subproblem must be solved with a high accuracy. However, the subproblem (1.15) of the original Bregman method does not need to be solved highly accurately. We will this briefly in the next subsection.

1.2.3. High Accuracy and Error Tolerance. We begin with a small and randomly generated compressed sensing example. In MATLAB¹, we set the seeds of the random number generators `rand` and `randn` to 2009, and generated a random matrix A and a vector b using

```
K = 25; m = 250; n = 500;
A = randn(m,n);
x = zeros(n,1);
p = randperm(n);
x(p(1:K)) = randn(k,1);
b = A*x;
```

The matrix A has 250 rows and 500 columns, and its entries are Gaussian random. The vector x has 500 entries, out of which 25 elements were randomly picked and assigned with Gaussian random values while the rest were set to zero. The vector b was set to be the product of A and x . Then, we applied the original Bregman method with FPC_BB [36] for solving its subproblems. The optional stopping tolerance `gtol` of FPC_BB, which specifies the maximally allowed magnitude of the objective subgradients of (1.15), was set as

```
opts.gtol = 1e-6;
```

and μ was set to 0.01. The relative errors of x^k with respect to x , for iteration $k = 1, \dots, 6$, are given in the table below.

k	1	2	3	4	5
$\frac{\ x-x^k\ }{\ x\ }$	6.5e-2	2.3e-7	6.2e-14	7.9e-16	5.6e-16.

This example demonstrates that an extremely accurate solution can be obtained in a very small number of Bregman iterations, using a first-order (gradient-based) subproblem solver with a relatively loose stopping criterion. To show that the subproblems were in fact not accurately solved, we re-calculated the solutions of all the subproblems (1.15) with the same set of b^k to a higher accuracy of at least 10^{-12} . Let these accurate solutions be denoted by \bar{x}^k (but we did not use \bar{x}^k in the update (1.14).) The relative error of x^k with respect to \bar{x}^k was 3.8e-7 for $k = 1$ and remained 1.1e-7 for $k = 2, \dots, 5$. In other words, subproblem solutions were inaccurate but they turned into an extremely accurate final solution.

¹MATLAB version R2007b, the 32-bit Windows Edition.

Both the Bregman update (1.14) and FPC_BB played important roles. Assume that \bar{x}^k is an exact solution of the k -th subproblem and

$$x^k = \bar{x}^k + \omega^k,$$

where ω^k represents the error. If \bar{x}^k is used in the update, then one obtains

$$\bar{b}^{k+1} = b + (b^k - A\bar{x}^k) = b + (b^k - Ax^k) - A\omega^k = b^{k+1} - A\omega^k.$$

Hence, the update yields b^{k+1} that is the sum of the exact update \bar{b}^{k+1} and the error $A\omega^k$.

What happens in the next iteration is that under certain conditions, ω^k is almost equal to ω^{k+1} , which is the error of the FPC_BB solution of (1.15) at iteration $k+1$, and they almost cancel each other in the sense that x^{k+1} is almost equal to the exact solution of (1.15) obtained with the exact \bar{b}^{k+1} . In other words, the error of iteration k gets carried over to iteration $k+1$ and cancels with the new error of iteration $k+1$. In addition, such error cancelation can happen iteratively with increasing accuracies, for instance, at $k=3, 4, 5$ in the example above. To fully see this, many details of FPC_BB including continuation and certain finite convergence properties need to be very carefully examined, and will be studied in a forthcoming report. We shall note that FPC_BB is not the only solver that makes this happen.

1.2.4. Applications. The Bregman iterative regularization and original Bregman method have been extended and applied to various problems. In addition to compressed sensing based on the ℓ_1 -norm and TV, applications and extensions can be found in [38] for image blind deconvolution, [9, 8] for inverse scale space methods, [61] for wavelet-based image denoising, [32] for the split Bregman method (the ‘‘split’’ part is related to the alternating method in [57, 62, 63, 64]) and its applications in [31], [34] for denoising and partially parallel imaging, [66] and [43] for matrix rank minimization.

1.2.5. The Linearized Bregman Method. The linearized Bregman method [65] was obtained originally by linearizing the term $H(x)$ in Step 3 of the original Bregman method into $\langle \nabla H(x^k), x \rangle$ and adding the ℓ_2 -proximity term $\frac{1}{2\alpha}\|x - x^k\|^2$, yielding the new iteration:

$$(1.17) \quad x^{k+1} \leftarrow \arg \min_x D_J^{p^k}(x, x^k) + \langle \nabla H(x^k), x \rangle + \frac{1}{2\alpha}\|x - x^k\|^2.$$

The components of x are separable in the last two terms of (1.17). Hence, for componentwise separable regularization functions J such as $\mu\|x\|_1$, (1.17) is very simple to compute. The update formula of p can be derived from the optimality conditions of (1.17):

$$(1.18) \quad p^{k+1} \leftarrow p^k - \nabla H(x^k) - \frac{1}{\alpha}(x^{k+1} - x^k),$$

where $p^{k+1} \in \partial J(x^{k+1})$. The algorithm based on (1.17) and (1.18) is given in Table 1.2.

For $H(x) = \frac{1}{2}\|Ax - b\|^2$, Steps 3 and 4 in Table 1.2 can be significantly simplified. First, from (1.18) or Step 4, we get

$$p^{k+1} = p^k - A^\top(Ax^k - b) - \frac{1}{\alpha}(x^{k+1} - x^k) = \dots = \sum_{i=0}^k A^\top(b - Ax^i) - \frac{x^{k+1}}{\alpha}.$$

Introduce

$$(1.19) \quad v^k = p^{k+1} + \frac{x^{k+1}}{\alpha} = p^k - A^\top(Ax^k - b) + \frac{x^k}{\alpha} = \sum_{i=0}^k A^\top(b - Ax^i), \quad \forall k.$$

TABLE 1.2
The linearized Bregman method

Input: $J(\cdot)$, $H(\cdot)$, $\alpha > 0$; optional: x^0 and p^0

1. **Initialize:** $k = 0$, let $x^0 = \mathbf{0}$ and $p^0 = \mathbf{0}$ if not provided.
2. **while** stopping conditions not satisfied **do**
3. $x^{k+1} \leftarrow \arg \min_x D_J^{p^k}(x, x^k) + \langle \nabla H(x^k), x \rangle + \frac{1}{2\alpha} \|x - x^k\|^2$
4. $p^{k+1} \leftarrow p^k - \nabla H(x^k) - \frac{1}{\alpha}(x^{k+1} - x^k) \in \partial J(x^{k+1})$

(If possible, replace Steps 3 and 4 by simpler updates, for example, (1.21) and (1.22) for solving (1.1), respectively. In addition, set $v^0 = A^\top b$ in Step 1.)

5. Optional: apply *kicking* if $x^{k+1} = x^k$
 6. $k \leftarrow k + 1$
 7. **end while**
-

Second, by $\nabla H(x^k) = A^\top(Ax^k - b)$, (1.17) or Step 3 can be simplified to

$$\begin{aligned}
 x^{k+1} &\leftarrow \arg \min_x J(x) - \langle p^k, x \rangle + \langle A^\top(Ax^k - b), x \rangle + \frac{1}{2\alpha} \|x - x^k\|^2 \\
 &\leftarrow \arg \min_x J(x) + \frac{1}{2\alpha} \left\| x - \alpha \left(p^k - A^\top(Ax^k - b) + \frac{x^k}{\alpha} \right) \right\|^2 \\
 (1.20) \quad &\leftarrow \arg \min_x J(x) + \frac{1}{2\alpha} \|x - \alpha v^k\|^2.
 \end{aligned}$$

The last problem (1.20) can be quickly solved for various choices of $J(x)$ such as $\mu\|x\|_1$, $\mu TV(x)$, $\mu\|\Phi x\|_1$ with a fast transform Φ , and more generally, for component-separable regularization terms in the form of $\sum_i \phi(x_i)$. After rearrangement we obtain the iterations

$$(1.21) \quad x^{k+1} \leftarrow \arg \min_x J(x) + \frac{1}{2\alpha} \|x - \alpha v^k\|^2,$$

$$(1.22) \quad v^{k+1} \leftarrow v^k + A^\top(b - Ax^{k+1}),$$

where (1.22) follows from (1.19). Problem (1.21) is simple to solve for many choices of $J(x)$ and also easy to program.

For $J(x)$ equal to $\mu\|x\|_1$ or several of its extensions, the iterations use only matrix-vector multiplications, vector additions and subtractions, as well as scalar soft thresholding. In the simplest case of $J(x) = \mu\|x\|_1$, the minimizer of (1.20) is given by

$$\text{shrink}(\alpha v^k, \alpha \mu) = \alpha \text{shrink}(v^k, \mu),$$

where $\text{shrink}(v^k, \mu)$ is known as component-wise soft thresholding:

$$(1.23) \quad (\text{shrink}(v^k, \mu))_i := \begin{cases} v_i^k - \mu, & v_i^k \in (\mu, +\infty), \\ 0, & v_i^k \in [-\mu, \mu], \\ v_i^k + \mu, & v_i^k \in (-\infty, \mu), \end{cases}$$

which can be computed as `sgn(vk) .* max(0, abs(vk))` in MATLAB. The form of the linearized Bregman method

$$(1.24) \quad x^{k+1} \leftarrow \alpha \text{shrink}(v^k, \mu),$$

$$(1.25) \quad v^{k+1} \leftarrow v^k + A^\top(b - Ax^{k+1}),$$

for compressed sensing problems was introduced in [65] and later accelerated by a technique called kicking in [48].

Kicking significantly reduces the total number of iterations during stagnation. According to (1.23) and (1.24), it can happen that over a sequence of consecutive iterations, the components v_i satisfying $|v_i| > \mu$ stay constant while the remaining components v_i , which satisfy $|v_i| \leq \mu$, are (slowly) updated. Until one of the latter components finally violates $|v_i| \leq \mu$, x remains unchanged. Kicking detects this stagnation by comparing x^k to x^{k+1} and breaks the stagnation by consolidating the remaining stagnated iterations.

It is proved in [11, 12] that the linearized Bregman method converges to the solution of

$$(1.26) \quad \min \left\{ \mu \|x\|_1 + \frac{1}{2\alpha} \|x\|^2 : Ax = b \right\}.$$

By scaling the objective function, (1.26) can be simplified to (1.2). The convergence was initially established for convex, continuously differentiable convex functions $J(x)$ in [11] (but neither ℓ_1 -norm nor total variation qualifies.) However, the same paper shows that if the convergence for $J(x) = \|x\|_1$ occurs, then the limit is the solution of (1.2). The convergence assumption was later removed in the authors' follow-up paper [12], which was drafted around the same time when the first version of this report was written. The proofs in [11, 12] are based on approximating $\|x\|_1$ by the Huber norm $\sum_{i=1}^n F_\epsilon(x_i)$, where

$$F_\epsilon(\xi) = \begin{cases} \frac{\xi}{2\epsilon}, & |\xi| \leq \epsilon \\ |\xi| - \frac{\epsilon}{2}, & |\xi| > \epsilon, \end{cases}$$

and ϵ is a smoothing parameter. The Huber norm is continuously differentiable, and in their proofs, $\epsilon \rightarrow 0$. In addition, it was shown that as $\alpha \rightarrow \infty$, the solution of (1.2) converges to one of (1.1).

In the next subsection, we show that the iterations in the linearized Bregman method are equivalent to a unit-step gradient descent iteration applied to a certain dual problem. Hence, the global convergence and the rate of convergence follow directly from results for gradient methods in the existing optimization literature. In addition, we show in Section 2 that requiring $\alpha \rightarrow \infty$ is not necessary; as long as α is larger than a threshold, the solution of (1.2) is an exact solution of (1.1).

Before ending this subsection, we list the applications of the linearized Bregman method that have appeared in the literature: compressed sensing [65, 48, 11], the matrix completion problem [10], and image processing [13]. Good numerical performance is reported in these papers.

1.3. Linearized Bregman as Dual Gradient Descent. First, the Lagrangian dual of

$$(1.27) \quad \min_x \{J(x) : Ax = b\}$$

is

$$\min_y -b^\top y + J^*(A^\top y),$$

where $J^*(z) := \max_x z^\top x - J(x)$ is the Fenchel dual (or Legendre transform) of $J(\cdot)$ (for simplicity, we assume this and other Fenchel duals exist). If we introduce

$$g_\alpha(x) := J(x) + \frac{1}{2\alpha} \|x\|^2,$$

and let $g_\alpha^*(\cdot)$ denote the Fenchel dual of $g_\alpha(\cdot)$, the Lagrangian dual of

$$\min_x \{g_\alpha(x) : Ax = b\}$$

is

$$(1.28) \quad \min_y -b^\top y + g_\alpha^*(A^\top y).$$

Since $g_\alpha(\cdot)$ is strictly convex, $g_\alpha^*(\cdot)$ is differentiable (cf. [50]).

THEOREM 1.1. *The linearized Bregman iterations (1.21)–(1.22) are equivalent to the gradient descent iteration applied to problem (1.28) with a unit step size.*

Proof. We shall relate the dual variable y^k to the variable v^k in (1.22), and then show that (1.22) is a gradient descent iteration. From (1.19), $v^k \in \mathcal{R}(A^\top)$ for all k , so we introduce y^k such that

$$(1.29) \quad v^k = A^\top y^k,$$

and (1.19) yields the iteration

$$(1.30) \quad y^{k+1} = y^k - (Ax^k - b).$$

Next, we show that $Ax^k - b$ is a gradient of the objective function of (1.3) at y^k assuming that p^k is a subgradient of $J(\cdot)$ at x^k :

$$\begin{aligned} p^k \in \partial_x J(x^k) &\iff A^\top y^k = p^k + \frac{1}{\alpha} x^k \in \partial_x g_\alpha(x^k) \\ &\iff x^k \in \nabla g_\alpha^*(A^\top y^k) \\ &\implies Ax^k = A \nabla g_\alpha^*(A^\top y^k) = \nabla_y g_\alpha^*(A^\top y^k) \\ &\iff Ax^k - b = \nabla_y (-b^\top y^k + g_\alpha^*(A^\top y^k)), \end{aligned}$$

where the second line is a well-known property of Fenchel duality (cf. [50]). \square

Dual gradient descent is equivalent to a multiplier method². Define the Lagrangian $L(x, y) := g_\alpha(x) + \langle y, b - Ax \rangle$. Then, the linearized Bregman updates (1.21) and (1.22) can be exactly obtained from the multiplier method

1. $x^{k+1} \leftarrow \min_x L(x, y^k)$,
2. $y^{k+1} \leftarrow y^k + \nabla_y L(x^{k+1}, y^k)$,

and let $v^k = A^\top y^k$.

Theorem 1.1 means that one can apply the general convergence results of gradient descent on the linearized Bregman method.

²The authors of [10] pointed out that the linearized Bregman method can be derived from Uzawa's method [1], which is a multiplier method motivated by economical equilibria.

Let us take $J(x) = \|x\|_1$ as an example (the derivation for $\mu\|x\|_1$ is similar), and show that iterations (1.24) and (1.25) generate a sequence $\{x^k\}$ that converges to the solution of (1.2) for $\alpha < 2/\|A\|^2$.

The Lagrangian dual problem of (1.1) is problem (1.4). According to the classic weak duality result, $\|x\|_1 \geq b^\top y$ for any x and y satisfying the constraints $Ax = b$ and $\|A^\top y\|_\infty \leq 1$. Because the feasible regions of both problems (1.1) and (1.4) have nonempty relative interiors (i.e., Slater's condition), there exist a pair of optimal primal and dual solutions (x^*, y^*) such that $\|x^*\|_1 = b^\top y^*$. Conversely, (x, y) is optimal if x and y satisfy $Ax = b$, $\|A^\top y\|_\infty \leq 1$, and $\|x\|_1 = b^\top y$. Strong duality is used in the proof of Theorem 2.8 below.

The Lagrangian dual problem of (1.2) is problem (1.3). In this example, g_α is

$$g_\alpha(x) = \|x\|_1 + \frac{1}{2\alpha}\|x\|^2,$$

and its Fenchel dual g_α^* can be straightforwardly computed as

$$g_\alpha^*(z) = \sum_{i=1}^n \frac{\alpha}{2}(z_i - \text{Proj}_{[-1,1]}(z_i))^2 = \frac{\alpha}{2}\|z - \text{Proj}_{[-1,1]^n}(z)\|_2^2,$$

where $\text{Proj}_{[-1,1]^n}(z)$ is the orthogonal projection of z to the hypercube $\{x \in \mathbb{R}^n : -1 \leq x_i \leq 1, \forall i\}$. Also, $\text{Proj}_{[-\mu, \mu]^n}(z) = z - \text{shrink}(z, \mu)$, so $z - \text{Proj}_{[-1,1]^n}(z) = \text{shrink}(z, 1)$. Let

$$(1.31) \quad F_\alpha(y) := -b^\top y + \frac{\alpha}{2}\|A^\top y - \text{Proj}_{[-1,1]^n}(A^\top y)\|^2,$$

which is equal to the objective function of (1.3). The problem of $\min_y F_\alpha(y)$ is precisely (1.3). Because

$$(1.32) \quad \nabla g_\alpha^*(z) = \alpha(z - \text{Proj}_{[-1,1]^n}(z)),$$

we have

$$(1.33) \quad \nabla F_\alpha(y) = -b + \alpha A \left(A^\top y - \text{Proj}_{[-1,1]^n}(A^\top y) \right).$$

Hence, the first-order optimality conditions of (1.3) are

$$(1.34) \quad \nabla F_\alpha(y) = 0$$

$$(1.35) \quad \text{or } \alpha A \left(A^\top y - \text{Proj}_{[-1,1]^n}(A^\top y) \right) = b.$$

It is easy to observe that

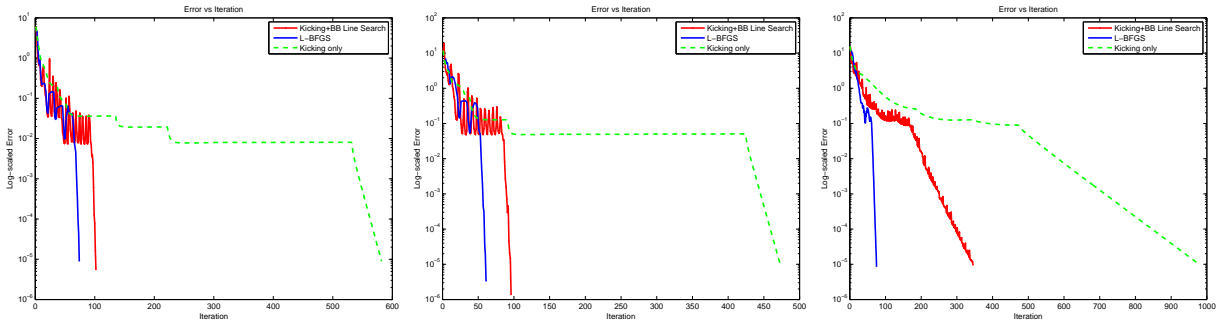
$$x = \alpha \left(A^\top y - \text{Proj}_{[-1,1]^n}(A^\top y) \right)$$

satisfies $Ax = b$, i.e., is a feasible solution of (1.1).

To establish that $\{x^k\}$ in the iterations (1.24) and (1.25) converges to the solution of (1.2), all we need to show is that the $\psi(\cdot) := \nabla F_\alpha(\cdot)$ is Lipschitz continuous with the constant $L \leq \alpha\|A\|^2$. Then, the classical result of gradient descent yields the convergence given that the step size (which is 1) is no more than $2/L$, or equivalently, $0 < \alpha < 2/\|A\|^2$. To obtain Lipschitz continuity, we derive

$$\begin{aligned} \|\psi(y^1) - \psi(y^2)\| &= \|A(\nabla g(A^\top y^1) - \nabla g(A^\top y^2))\| \\ &\leq \|A\| \cdot \alpha \|(A^\top y^1 - \text{Proj}_{[-1,1]^n}(A^\top y^1)) - (A^\top y^2 - \text{Proj}_{[-1,1]^n}(A^\top y^2))\| \\ &\leq \alpha\|A\| \|A^\top(y^1 - y^2)\| \\ &\leq \alpha\|A\|^2 \|y^1 - y^2\|, \end{aligned}$$

where the second inequality holds because $|(s - \text{Proj}_{[-1,1]}s) - (t - \text{Proj}_{[-1,1]}t)| \leq |s - t|$ for any $s, t \in \mathbb{R}$.



Test	Signal Dim.	# Measurement	Signal Sparsity	Measurement Type	Signal Type	α
1	1024	512	102	Gaussian + QR	± 1	1
2	1024	307	31	Gaussian	Gaussian	5
3	1024	307	61	Bernoulli	Gaussian	5

FIG. 1.1. Plots of absolute errors in 2-norm versus iterations for Tests 1 through 3, from left to right.

1.4. Generalizations of the Linearized Bregman Method. Based upon the equivalence between the linearized Bregman method and the unit-step gradient descent algorithm applied to the dual problem (1.28). It is natural to consider enhancements such as line search, quasi-Newton methods, L-BFGS [42], Nesterov’s recent algorithm [44], and nonlinear conjugate gradient methods, all of which need only gradient computations.

Our purpose is not to detail the above enhancements one by one but to argue that with any of these enhancements, the main computation remains almost as simple as (1.21) and (1.22), or (1.24) and (1.25) for³ $J(x) = \|x\|_1$, because the gradient of the objective function in (1.28) is simple to compute. At $y = y^k$, the gradient is given by $Ax^k - b$, where x^k is further given by (1.21) in which $v^k = A^\top y^k$. For many choices of $J(x)$, computing gradients remains simple. For $J(x) = \|x\|_1$, we have shown that $x^k = \alpha \left(A^\top y^k - \text{Proj}_{[-1,1]^n}(A^\top y^k) \right)$. For $J(x) = \|\Phi x\|_1$ where Φ is a non-singular transform, one can introduce $\bar{x} := \Phi x$ and reduce problem (1.27) into $\min\{\|\bar{x}\|_1 : A\bar{x} = b\}$. Furthermore if Φ is orthonormal, then $\Phi^{-1} = \Phi^\top$ and thus $\bar{x}^k = \alpha \left(\Phi A^\top y^k - \text{Proj}_{[-1,1]^n}(\Phi A^\top y^k) \right)$. For $J(x) = TV(x)$, problem (1.21) is the ROF model, which can quickly solved by many algorithms including the latest graph-cut/max-flow algorithms [19, 16, 30]. The list of functions $J(x)$ permitting fast solutions is not short.

In Figure 1.1, we present comparison results of three different implementations of the linearized Bregman method applied to $J(x) = \|x\|_1$. It is important to note that from these very limited tests, no definitive conclusions on the three implementations should be drawn; however, the results demonstrate the possibility of accelerating the basic linearized Bregman method by applying standard optimization techniques. We added kicking, a technique combining kicking and the Barzilai-Borwein method accompanied by a non-monotone line search (written as kicking+BB_line_search), and L-BFGS⁴ to the algorithm in Table 1.2, and obtained the three corresponding implementations. We refer to [2, 18] for details on the Barzilai-Borwein method with line search. Recent uses of this method on ℓ_1 -minimization can be found in [29, 36, 59]. L-BFGS [45] is a well-known implementation of quasi-Newton optimization. It is based on the Broyden-Fletcher-Goldfarb-

³We do not work with $\mu\|x\|_1$ because it is redundant to scale both terms in the objective function of (1.2).

⁴Courtesy of Zaiwen Wen for a pure MATLAB implementation of L-BFGS

Shanno (BFGS) approximate Hessian update but does not explicitly store either the approximate Hessian or its inverse. Instead, it implicitly applies the approximate Hessian or its inverse that are generated from the last m updates of x and $\nabla f(x)$ on the fly, where m is generally as small as between 5 and 20. Hence, L-BFGS is particularly suited for large-scale optimization problems. To keep the discussion concise, we refer to the reader to the paper [59] for further details including parameter settings.

Figure 1.1 depicts three plots of absolute errors in 2-norm versus the number of iterations. In both implementations using `kicking` and `kicking+BB.line_search`, exactly two matrix–vector multiplications, one involving A and the other involving A^\top , are performed at each iteration; however, the L-BFGS implementation may perform more than one pair of such multiplications at each L-BFGS iteration. For fairness of comparison, we count each pair of A and A^\top -multiplications as one iteration in all of the three implementations. The comparison results clearly show that standard optimization enhancements can significantly accelerate the linearized Bregman method.

Above we have reviewed the Bregman iterative regularization method, the original Bregman method, as well as the linearized Bregman method. Along with these reviews, a few new observations and analysis are given. Next in Section 2 below, we present the main results of this paper. Conclusions and discussions are given in Section 3.

2. Main Results. The purpose of this section is to show that there exists a finite scalar $\alpha_\infty > 0$ such that the solution x_α of (1.2) is also a solution of the basis pursuit problem (1.1) whenever $\alpha > \alpha_\infty$. In other words, $\alpha > \alpha_\infty$ enables the linearized Bregman algorithm to compute a solution of (1.1). We note that the approach presented below in Subsections 2.1 and 2.2 is not concise as it could be, but the steps in the approach help us develop insights and ideas for checking $\alpha > \alpha_\infty$ and computing an upper bound of α_∞ , leading us to the results in Subsections 2.3 and 2.4.

2.1. Solutions of Problems (1.3) and (1.2). Let Y_α denote the set of solutions of (1.3) and $y_\alpha \in Y_\alpha$. For the convenience of subsequent analysis, we partition the index set $\{1, \dots, n\}$ into three subsets according to the values of $(A^\top y)_i$, $i = 1, \dots, n$. Define

$$q_i^1(y) := \begin{cases} 1, & (A^\top y)_i < -1, \\ 0, & \text{o.w.}, \end{cases} \quad q_i^2(y) := \begin{cases} 1, & -1 \leq (A^\top y)_i \leq 1, \\ 0, & \text{o.w.}, \end{cases} \quad q_i^3(y) := \begin{cases} 1, & (A^\top y)_i > 1, \\ 0, & \text{o.w.} \end{cases}$$

for $i = 1, \dots, n$. We also similarly define $\bar{q}_i^1(y)$, $\bar{q}_i^2(y)$, and $\bar{q}_i^3(y)$ by replacing all strict inequalities by non-strict ones (e.g., $<$ replaced by \leq) and vice versa. Let $Q^j(y) := \text{Diag}(q^j)$ for $j = 1, 2, 3$, which act as “selection” matrices. For example, $Q^1(y) \cdot (A^\top y - b) = 0$ means that $(A^\top y - b)_i = 0$ for all i satisfying $(A^\top y)_i < -1$. According to the definitions, for any i or y , exactly one of $Q_{ii}^1(y)$, $Q_{ii}^2(y)$, and $Q_{ii}^3(y)$ equals

1. The following example illustrates the above definitions:

$$A^\top y = \begin{bmatrix} 3 \\ 2 \\ -4 \\ -1 \\ \frac{1}{2} \end{bmatrix} \begin{array}{l} > 1 \\ > 1 \\ < -1 \\ \in [-1, 1] \\ \in [-1, 1] \end{array}$$

$$\implies Q^1(y) = \begin{bmatrix} 0 & & & & \\ & 0 & & & \\ & & 1 & & \\ & & & 0 & \\ & & & & 0 \end{bmatrix}, \quad Q^2(y) = \begin{bmatrix} 0 & & & & \\ & 0 & & & \\ & & 0 & & \\ & & & 1 & \\ & & & & 1 \end{bmatrix}, \quad Q^3(y) = \begin{bmatrix} 1 & & & & \\ & 1 & & & \\ & & 0 & & \\ & & & 0 & \\ & & & & 0 \end{bmatrix}.$$

Furthermore, we let $Q(y) := (Q^1(y), Q^2(y), Q^3(y))$. For given Q^1 , Q^2 , and Q^3 independent of y , we let $Q := (Q^1, Q^2, Q^3)$.

Depending only on Q , we divide A column-wise into three submatrices. For $j = 1, 2, 3$ each, let A^j be submatrix of A formed by the columns i of A that correspond to $(Q^j)_{ii} = 1$, and let $e^j = [1 \ 1 \ \dots \ 1]^\top$ with the dimension equal to the number of columns of A^j .

DEFINITION 2.1. *We say that y is consistent with a given Q if $Q(y) = Q$.*

The following theorem states that it is straightforward to obtain a primal solution from a dual solution.

THEOREM 2.2. *Let $\alpha > 0$. For any solution $y_\alpha \in Y_\alpha$ of (1.3), equation (1.5) gives the unique solution of (1.2). In particular, $Ax_\alpha = b$. This is a classical result of convex duality. For completeness, we provide a direct proof.*

Proof. According to (1.35), $Ax_\alpha = b$, i.e., x_α is feasible. As we can apply the result of classic convex duality, it remains to show that the duality gap vanishes, namely, $b^\top y_\alpha - g_\alpha^*(A^\top y_\alpha) = \|x_\alpha\|_1 + \frac{1}{2\alpha}\|x_\alpha\|_2^2$. Let $p := \text{Proj}_{[-1,1]^n}(A^\top y_\alpha)$. Since $(x_\alpha)_i$ is strictly positive (strictly negative) if and only if p_i equals 1 (-1 , respectively), we have $p_i \cdot (x_\alpha)_i = |x_i|$ and thus $p^\top x_\alpha = \|x_\alpha\|_1$; hence,

$$\begin{aligned} b^\top y_\alpha - g_\alpha^*(A^\top y_\alpha) &= (Ax_\alpha)^\top y_\alpha - \frac{\alpha}{2} \|A^\top y_\alpha - p\|_2^2 \\ &= (A^\top y_\alpha)^\top x_\alpha - \frac{1}{2\alpha} \|\alpha(A^\top y_\alpha - p)\|_2^2 \\ &= \frac{1}{\alpha} (\alpha(A^\top y_\alpha - p))^\top x_\alpha + p^\top x_\alpha - \frac{1}{2\alpha} \|\alpha(A^\top y_\alpha - p)\|_2^2 \\ &= \frac{1}{\alpha} \|x_\alpha\|_2^2 + \|x_\alpha\|_1 - \frac{1}{2\alpha} \|x_\alpha\|_2^2 \\ &= \|x_\alpha\|_1 + \frac{1}{2\alpha} \|x_\alpha\|_2^2. \end{aligned}$$

Finally, because problem (1.2) has a strictly convex objective function, its solution x_α is unique. \square

This theorem lets one recover the primal solution x_α from any dual solution $y_\alpha \in Y_\alpha$. In case that y_α is not an exact but *approximate* solution, x_α is not an exact solution of (1.2) either, and the primal feasibility measure $\|Ax_\alpha - b\|_2$ is equal to the first-order dual optimality measure $\|\nabla F_\alpha(y_\alpha)\|$ while, on the other hand, the duality gap given by this pair of x_α and y_α is always zero.

Whether Y_α is a singleton or not, x_α is unique. Therefore, we have

COROLLARY 2.3. *Let $\alpha > 0$. Then, both $A^\top y_\alpha - \text{Proj}_{[-1,1]^n}(A^\top y_\alpha)$ and $Q(y_\alpha)$ are constant over all $y_\alpha \in Y_\alpha$. Hence, $Q(y_\alpha)$ only depends on α , so we introduce the y -independent notation*

$$Q_\alpha^j := Q^j(y_\alpha), \quad j = 1, 2, 3, \quad y_\alpha \in Y_\alpha.$$

We similarly define A_α^j , $j = 1, 2, 3$, as the partitions of A that depend on α . Then, it is easy to see

$$(2.1) \quad \begin{aligned} \min_y F_\alpha(y) = \\ \min_y -b^\top y + \frac{\alpha}{2} \|(A_\alpha^1)^\top y + e_\alpha^1\|_2^2 + \frac{\alpha}{2} \|(A_\alpha^3)^\top y - e_\alpha^3\|_2^2, \end{aligned}$$

where e_α^1 and e_α^3 are vectors of all ones of appropriate dimensions. It is important to note that the above equation only means the two problems have the same optimal objective value, but problem (2.1) can have a solution that is not a solution of $\min_y F_\alpha(y)$, which is problem (1.3). Problem (2.1) can have multiple solutions some of which are *not* consistent with Q_α , thus not solutions of (1.3). On the other hand, any solution of (1.3) satisfies the optimality conditions of (2.1) and thus is a solution of (2.1). Let $y_\alpha \in Y_\alpha$. The set of solutions of (2.1) is $\{y_\alpha\} + \text{Null}(A_\alpha^1(A_\alpha^1)^\top + A_\alpha^3(A_\alpha^3)^\top)$, which can be larger than Y_α . From Corollary 2.3 and noticing $A_\alpha^1(A_\alpha^1)^\top + A_\alpha^3(A_\alpha^3)^\top = A(Q_\alpha^1 + Q_\alpha^3)A^\top$, it is easy to show

COROLLARY 2.4. *Let $y_\alpha \in Y_\alpha$. Then,*

$$Y_\alpha = (\{y_\alpha\} + \text{Null}(A(Q_\alpha^1 + Q_\alpha^3)A^\top)) \cap \{y : Q(y) = Q_\alpha\}.$$

In general, Y_α can contain more than one element. Given a solution $y_\alpha \in Y_\alpha$, it is trivial to get $\text{Proj}_{[-1,1]^n}(A^\top y_\alpha) = \text{sign}(x_\alpha)$ and thus Q_α . However, similar to the result of classic degenerate complementary slackness, finding a solution y_α in Y_α from the corresponding x_α is generally more difficult because y_α may be subject to additional inequality constraints implied by $Q(y) = Q_\alpha$. In other words, not all solution to the normal equations of (2.1) satisfies $Q(y) = Q_\alpha$. An exception arises when $A(Q_\alpha^1 + Q_\alpha^3)A^\top$ has a full rank because then, the normal equations of (2.1) have a unique solution y_α , which must lie in Y_α and thus satisfy $Q(y) = Q_\alpha$ automatically. This is unfortunate news for compressed sensing problems. Because they tend to have sparse solutions, the exceptional case would not normally arise. On the other hand, solution sparsity provides other means to check optimality; see the discussions in Subsection 2.4.

2.2. Exact Penalty. In this subsection, through analyzing the point set

$$(2.2) \quad I(\alpha) = \{\beta > 0 : Q_\beta = Q_\alpha\} \subset \mathbb{R},$$

we show that for α sufficiently large, the solution x_α of (1.2) is also an solution of (1.1).

LEMMA 2.5. *Let $\alpha > 0$. $I(\alpha)$ is nonempty and connected, so it is either a singleton or an interval (possibly unbounded).*

Proof. Since $\alpha \in I(\alpha)$, $I(\alpha)$ is nonempty. It remains to show that for $\alpha_1, \alpha_2 \in I(\alpha)$ and $\gamma \in (0, 1)$, $\beta := \gamma\alpha^1 + (1 - \gamma)\alpha^2 \in I(\alpha)$. From $\alpha_1, \alpha_2 \in I(\alpha)$, it follows that $A_\alpha^1 = A_{\alpha_1}^1 = A_{\alpha_2}^1$, $A_\alpha^3 = A_{\alpha_1}^3 = A_{\alpha_2}^3$, and there exist $y_{\alpha_1} \in Y_{\alpha_1}$ and $y_{\alpha_2} \in Y_{\alpha_2}$, which satisfy the optimality conditions of (2.1) in the following form: for $\nu = \alpha_1, \alpha_2$ each

$$(2.3) \quad (A_\alpha^1(A_\alpha^1)^\top + A_\alpha^3(A_\alpha^3)^\top) y_\nu = (A_\alpha^1 e_\alpha^1 - A_\alpha^3 e_\alpha^3) + \nu^{-1} b.$$

Define $\Delta y := (y_{\alpha_1} - y_{\alpha_2})/(\alpha_1^{-1} - \alpha_2^{-1})$ and $y_\beta := y_{\alpha_2} + (\beta^{-1} - \alpha_2^{-1})\Delta y$. Since y_β satisfies (2.3) for $\nu = \beta$ and is on the line segment connecting y_{α_1} and y_{α_2} , we have $Q(y_\beta) = Q_\alpha$. Since $Q(y_\beta) = Q_\beta$, we get $Q_\beta = Q_\alpha$ and thus, $\beta \in I(\alpha)$. \square Let $\mathcal{I} = \{I(\alpha) : \alpha > 0\}$. Since there are finitely many distinct Q_α 's, \mathcal{I} is a finite set. Since $I \cap I' = \emptyset$ for any two *distinct* $I, I' \in \mathcal{I}$, Lemma 2.5 lets us order the elements of \mathcal{I} by their corresponding α -values in the following way:

$$\mathcal{I} = \{I_1, I_2, \dots, I_J\}, \quad J := |\mathcal{I}| < \infty$$

such that

$$\alpha < \alpha', \quad \forall \alpha \in I_i, \forall \alpha' \in I_{i+1}, \quad i = 1, \dots, J-1.$$

Since Q_α does not depend on the choice of $\alpha \in I_j$, we introduce the α -independent notation

$$Q_j := Q_\alpha, \quad \alpha \in I_j, \quad j = 1, \dots, J.$$

Similarly, we define $A_j^1 := A_\alpha^1$, $A_j^2 := A_\alpha^2$, and $A_j^3 := A_\alpha^3$, where $\alpha \in I_j$, for $j = 1, \dots, J$.

Because $\cup \mathcal{I} = \{\alpha : \alpha > 0\}$ and \mathcal{I} is a finite set, we have

LEMMA 2.6. *J is finite and $\sup I_j = +\infty$.* To proceed we need the following assumption, which leads to the boundedness of $\cup_{\beta \geq \alpha_0} Y_\beta$ for any $\alpha_0 > 0$.

ASSUMPTION 1. *A has full row rank.* When this assumption does not hold, there exists at least one redundant constraint in the system $Ax = b$ given its consistency.

Next, we show $x_\alpha, \forall \alpha \in I_j$, solves problem (1.1) by finding a corresponding solution of the dual problem (1.4). The following lemma proves that a set of key equations have joint a solution.

LEMMA 2.7. *Let $\alpha \in I_j$ and $y_\alpha \in Y_\alpha$. Under Assumption 1, the following system has a solution Δy :*

$$(2.4) \quad (A_\alpha^1(A_\alpha^1)^\top + A_\alpha^3(A_\alpha^3)^\top) \Delta y = b,$$

$$(2.5) \quad \|A^\top(y_\alpha - \alpha^{-1}\Delta y)\|_\infty \leq 1,$$

$$(2.6) \quad (A_\alpha^1)^\top(y_\alpha - \alpha^{-1}\Delta y) = -e_\alpha^1,$$

$$(2.7) \quad (A_\alpha^3)^\top(y_\alpha - \alpha^{-1}\Delta y) = e_\alpha^3.$$

The purpose is to obtain a vector Δy such that

$$(2.8) \quad y_\infty := y_\alpha - \alpha^{-1}\Delta y$$

is a solution of (1.4). Equation (2.4) is a result of (2.3) after varying ν within I_j . If Δy satisfies (2.4) then

$$y_\beta := y_\alpha + (\beta^{-1} - \alpha^{-1})\Delta y$$

solves (2.3) for $\nu = \beta > \alpha$ and thus, is a solution of (2.1) with β replacing α . However, (2.4) alone is not enough because not all such Δy yields a feasible solution y_∞ of (1.4). Equation (2.5) poses the feasibility condition. The remaining equations (2.6) and (2.7) follow from (2.4) and (2.5) when $\alpha \in I_j$ and $y_\alpha \in Y_\alpha$, as shown in Theorem 2.9 below. However, (2.6) and (2.7) are explicitly given because when $\alpha \in I_j$ and $j \neq J$, they sometimes still hold while (2.5) does not; we can show that for $\alpha \in I_j$, if (2.4), (2.6), and (2.7) hold, then x_α is constant over I_j .

Proof. [Proof of Lemma 2.7] Let $\alpha' > \alpha$ and $y_{\alpha'} \in Y_{\alpha'}$. Clearly, $\alpha' \in I_J$; hence, there exists $y_{\alpha'} \in Y_{\alpha'}$ satisfying (2.3) with $\nu = \alpha'$. So does y_α with $\nu = \alpha$. By taking the differences between two copies of (2.3) with $\nu = \alpha$ and $\nu = \alpha'$, we get

$$(2.9) \quad \Delta y_{\alpha'} := \frac{y_{\alpha'} - y_\alpha}{\alpha'^{-1} - \alpha^{-1}},$$

which satisfies (2.4). Hence, the equations in (2.4) are consistent. Because $y_{\alpha'} = y_\alpha + (\alpha'^{-1} - \alpha^{-1})\Delta y_{\alpha'}$ and $Q(y_{\alpha'}) = Q_J$, the set

$$S_{\alpha'} := \{\Delta y : \Delta y \text{ satisfies (2.4)}\} \cap \{\Delta y : Q(y_\alpha + (\alpha'^{-1} - \alpha^{-1})\Delta y) = Q_J\}$$

contains $y_{\alpha'}$ and thus is nonempty. As in the proof of Lemma 2.5, from $\Delta y \in S_{\alpha'}$ one can get $\Delta y \in S_\beta$ for any $\beta \in [\alpha, \alpha']$, which means $S_{\alpha'}$ is monotonically non-increasing in α' . From Assumption 1, $T_\alpha := \cup_{\beta \geq \alpha} Y_\beta$ is bounded, so $S_{\alpha'} \subset \{u - v : u, v \in T_\alpha\}$ is bounded. From the theorem of nested sets, there exists a $\Delta y \in \cap_{\alpha' > \alpha} \text{cl}(S_{\alpha'})$ satisfying (2.4) and using this Δy , we have

$$(2.10) \quad y_\beta := y_\alpha + (\beta^{-1} - \alpha^{-1})\Delta y \in Y_\beta, \quad \forall \beta \geq \alpha.$$

It is a classical result of the quadratic penalty method that in (1.3), the penalized terms vanish as the penalty parameter goes to infinity, i.e.,

$$(2.11) \quad \lim_{\beta \rightarrow \infty} \|(A_\alpha^1)^\top y_\beta + e_\alpha^1\| = 0, \quad \lim_{\beta \rightarrow \infty} \|(A_\alpha^3)^\top y_\beta - e_\alpha^3\| = 0, \quad \forall y_\beta \in Y_\beta.$$

Combining (2.10) and (2.11) and letting $\beta \rightarrow \infty$, we get (2.6) and (2.7).

From (2.10), we also have $(A^\top y_\beta)_i \in [-1, 1]$ for all i such that $(Q_J^1)_{ii} = (Q_J^3)_{ii} = 0$. Therefore, (2.5) follows from (2.6) and (2.7). \square It is worth noting that (2.4) can have multiple solutions, not all satisfying (2.5)–(2.7). The whole system of (2.4)–(2.7) can have multiple solutions as well. However, it can be referred from the above Lemma that if (2.4) has a unique solution Δy , Δy must satisfy (2.5)–(2.7). Next, we prove the main result of this section.

THEOREM 2.8. x_α is constant over $\alpha \in I_J$ and is an optimal solution of problem (1.1).

Proof. Let $\alpha \in I_J$, $\beta \geq \alpha$, $y_\alpha \in Y_\alpha$, Δy be a solution of (2.4)–(2.7). Define y_∞ and y_β as in (2.8) and (2.10), respectively. We have $y_\infty = y_\beta - \beta^{-1}\Delta y$ and, from (2.6) and (2.7), $Q_J^1(A^\top y_\infty + e) = 0$ and $Q_J^3(A^\top y_\infty - e) = 0$. Therefore,

$$\begin{aligned} x_\beta &= \beta(A^\top y_\beta - \text{Proj}_{[-1,1]^n}(A^\top y_\beta)) \\ &= \beta Q_J^1(A^\top y_\beta + e) + \beta Q_J^3(A^\top y_\beta - e) \\ &= \beta Q_J^1(A^\top y_\infty + e) + \beta Q_J^3(A^\top y_\infty - e) + (Q_J^1 + Q_J^3)A^\top \Delta y \\ &= (Q_J^1 + Q_J^3)A^\top \Delta y. \end{aligned}$$

Since both $\alpha \in I_J$ and $\beta \geq \alpha$ are arbitrary and x_β is independent of β , the first half of the theorem is proved.

The second half follows from the strong duality theorem and the facts that x_α is primal feasible (i.e., $Ax_\alpha = b$), y_∞ is dual feasible from (2.5), and the duality gap vanishes as

$$\|x_\alpha\|_1 = x_\alpha^\top Q_J^3 e - x_\alpha^\top Q_J^1 e = x_\alpha^\top Q_J^3 (A^\top y_\infty) + x_\alpha^\top Q_J^1 (A^\top y_\infty) = x_\alpha^\top (Q_J^1 + Q_J^3) A^\top y_\infty = x_\alpha^\top A^\top y_\infty = b^\top y_\infty.$$

\square

2.3. An Pathological Example. For a given α , it is generally tricky to test $\alpha \in I_J$ based only on a primal-dual solution pair x_α and $y_\alpha \in Y_\alpha$. One needs to solve (2.4)–(2.7) (in fact, only (2.4) and (2.5) will suffice as is shown below), which include inequality constraints implicitly in (2.5). Is there a simple alternative to avoid the inequalities?

A property of $\alpha \in I_J$ is that x_α is constant, so one may wonder this property is sufficient, namely, if $x_\alpha = x_\beta$ for $\alpha \neq \beta$, then $\alpha \in J$. This does not hold in the following example.

Let

$$A = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 0 & -2 \end{bmatrix}, \quad b = \begin{bmatrix} 4 \\ 3 \end{bmatrix}.$$

Then, for $\alpha = 1, 2, 3, 4, 8$, as well as ∞ (for problem (1.1)), the primal and dual solutions x_α and y_α of (1.2) are given in the following table:

α	x_α	y_α	$\ x_\alpha\ _1$
1	$[3 \ 1 \ 0]^\top$	$[2 \ 2]^\top$	4
2	$[3 \ 1 \ 0]^\top$	$[\frac{3}{2} \ 1]^\top$	4
3	$[\frac{65}{21} \ \frac{17}{21} \ \frac{1}{21}]^\top$	$[\frac{80}{63} \ \frac{16}{21}]^\top$	$\frac{83}{21}$
4	$[\frac{67}{21} \ \frac{13}{21} \ \frac{2}{21}]^\top$	$[\frac{97}{84} \ \frac{9}{14}]^\top$	$\frac{82}{21}$
8	$[\frac{7}{2} \ 0 \ \frac{1}{4}]^\top$	$[\frac{125}{128} \ \frac{59}{128}]^\top$	$\frac{15}{4}$
∞	$[\frac{7}{2} \ 0 \ \frac{1}{4}]^\top$	$[\frac{3}{4} \ \frac{1}{4}]^\top$	$\frac{15}{4}$

It is interesting to observe that for $\alpha = 1, 2$, the primal solution x_α of (1.2) remain the same but is not optimal to (1.1). Therefore, one cannot conclude the optimality of x_α from it being over different α . Suppose $\alpha, \alpha' \in I_j \in \mathcal{I}$ and $\alpha \neq \alpha'$. From the proof of Theorem 2.8, it is easy to see when equations (2.4), (2.6) and (2.7) hold for $\Delta y := \frac{y_{\alpha'} - y_\alpha}{\alpha' - 1 - \alpha - 1}$ for $y_\alpha \in Y_\alpha$ and $y_{\alpha'} \in Y_{\alpha'}$, then $x_\alpha = x_{\alpha'}$ and it is unique over $\alpha \in I_j$. Therefore, condition (2.5) is indispensable for identifying $\alpha \in I_J$.

The above example also demonstrates that x_α can vary over α lying in the same interval I_j . For $\alpha = 3$ and $\alpha = 4$, all entries of x_α have the same sign. Hence, $\alpha = 3$ and 4 belong to the same interval I_j . However, $x_3 \neq x_4$.

Since minimizing $\|x\|_1$ and $\|x\|_2$ tend to yield sparse and non-sparse solutions, respectively, it is natural to conjecture that the solution x_α of (1.2) becomes monotonically more sparse as α increases. However, in the above example x_α has more nonzero entries for $\alpha = 4$ than $\alpha = 1$ or 2, so the number of nonzero entries in x_α generally may not be monotonic in α .

Finally, exact penalty holds for $\alpha = 8$ since $x_8 = x_\infty$.

2.4. Recognize $\alpha \in I_J$. To verify $\alpha \in I_J$, one needs to solve (2.4) and (2.5) for Δy , but not (2.6) and (2.7).

THEOREM 2.9. $\alpha \in I_J$ if and only if (2.4) and (2.5) have a solution.

Proof. The “only if” part is shown in Lemma 2.7.

We show the “if” part by contradiction. Let $\alpha \in I_j$. Suppose (2.4) and (2.5) hold for Δy but $j \neq J$.

First, we show that some equation in (2.6) or (2.7) must be violated by contradiction. Suppose that all equations in (2.6) and (2.7) hold for Δy . Then, we know $y_\beta = y_\alpha + (\beta^{-1} - \alpha^{-1})\Delta y \in Y_\beta$ from (2.10) in the proof of Lemma 2.7. From (2.6), (2.7), and the fact $(A_\alpha^1)^\top y_\alpha < -e_\alpha^1$ and $(A_\alpha^3)^\top y_\alpha > e_\alpha^3$, we have

$$(2.12) \quad (A_\alpha^1)^\top (y_\alpha + (\beta^{-1} - \alpha^{-1})\Delta y) < -e_\alpha^1 \text{ and } (A_\alpha^3)^\top (y_\alpha + (\beta^{-1} - \alpha^{-1})\Delta y) > e_\alpha^3, \quad \forall \beta \geq \alpha.$$

Recalling the definition of A_α^2 , we have $-e_\alpha^2 \leq (A_\alpha^2)^\top y_\alpha \leq e_\alpha^2$, and from (2.5), we also have

$$(2.13) \quad -e_\alpha^2 \leq (A_\alpha^2)^\top (y_\alpha + (\beta^{-1} - \alpha^{-1})\Delta y) \leq e_\alpha^2, \quad \forall \beta \geq \alpha.$$

From (2.12) and (2.13), we conclude that $y_\beta = y_\alpha + (\beta^{-1} - \alpha^{-1})\Delta y$ is compatible with Q_α for all $\beta > \alpha$. From this and $\alpha \in I_j$, it follows that $I_j = I_J$, which contradicts $j \neq J$.

Let V^1 and V^3 denote the index sets of violated equations in (2.6) and (2.7), respectively. Therefore, we have shown $V^1 \cup V^3 \neq \emptyset$. Together with (2.5), we have

$$(2.14) \quad -1 \leq A_i^\top (y_\alpha - \alpha^{-1}\Delta y) < 1, i \in V^1,$$

$$(2.15) \quad -1 < A_i^\top (y_\alpha - \alpha^{-1}\Delta y) \leq 1, i \in V^3.$$

Second, we show that there exists a vector $z \neq \mathbf{0}$ such that $Az = \mathbf{0}$ and

$$(2.16) \quad z_i \begin{cases} < 0, & i \in V^1, \\ = 0, & i \in (V^1 \cup V^3)^C, \\ > 0, & i \in V^3. \end{cases}$$

From (2.3) and (2.4), it is easy to derive

$$(A_\alpha^1 (A_\alpha^1)^\top + A_\alpha^3 (A_\alpha^3)^\top) (y_\alpha - \alpha^{-1}\Delta y) = A_\alpha^3 e_\alpha^3 - A_\alpha^1 e_\alpha^1,$$

or using the convention $A = [A_\alpha^1 \ A_\alpha^2 \ A_\alpha^3]$,

$$A \begin{bmatrix} -e_\alpha^1 - (A_\alpha^1)^\top (y_\alpha - \alpha^{-1}\Delta y) \\ \mathbf{0} \\ e_\alpha^3 - (A_\alpha^3)^\top (y_\alpha - \alpha^{-1}\Delta y) \end{bmatrix} = \mathbf{0}.$$

Let z be the vector in the brackets. We know that the entries in $(A_\alpha^1)^\top (y_\alpha - \alpha^{-1}\Delta y)$ equal -1 except those in V^1 and the entries in $(A_\alpha^3)^\top (y_\alpha - \alpha^{-1}\Delta y)$ equal 1 except those in V^3 . From (2.14) and (2.15), we obtain (2.16).

Finally, we show that x_α defined in Theorem 2.2, which assumed to be optimal to (1.2) as $\alpha \in J$, is however not optimal. According to the definition of V^1 and V^3 , we have $(x_\alpha)_i < 0$, $i \in V^1$, and $(x_\alpha)_i > 0$, $i \in V^3$. From (2.16), there exists a small scalar $\rho > 0$ such that $x_\alpha - \rho z$ yields a strictly smaller objective of (1.2). Moreover, since $Ax_\alpha = b$ and $Az = 0$, we have $A(x_\alpha - \rho z) = b$. Hence, $x_\alpha - \rho z$ is a better solution than x_α , meaning that x_α is not optimal. This contradicts to the optimality of x_α . \square

COROLLARY 2.10. *Equations (2.6) and (2.7) are implied by (2.4) and (2.5).*

Next we study the special cases in which $\alpha \in I_J$ or $\alpha \notin I_J$ can be determined without fully solving (2.4) or (2.5) for Δy .

Case 1. If (2.4) has a unique solution or, equivalent, the matrix $[A_\alpha^1 \ A_\alpha^3]$ has full row rank, then one can simply test the solution against (2.5). If it satisfies (2.5), then $\alpha \in I_J$ and x_α is optimal to (1.1); otherwise, $\alpha \notin I_J$ and x_α is not optimal.

Case 2. In the application of (1.1) for finding a sparsest linear representation of b using the columns of A , there are results stating that there exists a number M such that any x satisfying $Ax = b$ and $\|x\|_0 \leq M$ is the sparsest representation. Similar results exist for compressed sensing problems with sparse solutions. See papers [24, 27, 52, 33, 67, 68, 40] and references therein. For compressed sensing, cross validation [6, 58] can also be applied.

Alternatively, given a sparse x_α , one can check the following sufficient optimality condition: let $S := \text{supp}(x_\alpha)$, the set of nonzero components of x_α ; construct $\tilde{y} = A_S(A_S^\top A_S)^{-1} \text{sign}(x_\alpha)$, where $\text{sign}(x_\alpha)$ is the vector with ± 1 components indicating the signs of x_α ; if $\|A\tilde{y}\|_\infty \leq 1$ then x_α is an optimal solution of (1.1). Here, \tilde{y} is the least-squares solution of $A_S^\top y = \text{sign}(x_\alpha)$, and it satisfies $b^\top \tilde{y} = \|x_\alpha\|_1$. Hence, the dual feasibility condition $\|A\tilde{y}\|_\infty \leq 1$ is sufficient for (x_α, \tilde{y}) to form an optimal primal-dual solution pair for (1.1). This condition is not necessary because x_α can be optimal but $\|A\tilde{y}\|_\infty > 1$. However, when x_α is optimal, \tilde{y} as the least-squares solution tends to satisfy $\|A\tilde{y}\|_\infty \leq 1$ so the above sufficient condition is indeed practical.

Case 3. If two solutions x_α and $x_{\alpha'}$, $\alpha \neq \alpha'$, have the same signs but different values, then we can conclude neither α nor α' is in I_J .

Case 4. When (1.1) has a unique solution, the solution has no more than m nonzeros. Consequently, if x_α has more than m nonzeros, then $\alpha \notin I_J$.

In a compressed sensing problem where the matrix A is drawn from a random distribution, formed by a set of randomly chosen rows of a known matrix, or constructed in other ways involving randomization, the expected solution is almost always either highly sparse or has exactly m nonzero elements. In the former situation, Case 3 applies. In the latter situation, $[A_\alpha^1 \ A_\alpha^3]$ often has full rank so Case 1 applies. Therefore, it is often straightforward to test the optimality for a compressed sensing problem.

In general, if none of the first three cases applies at a solution $y_\alpha \in Y_\alpha$, the solution is degenerate. To guarantee that the corresponding x_α is an optimal solution of (1.1), one ultimately needs to obtain y_∞ that is optimal to (1.4) and thus forms an optimal primal-dual pair with x_α . However, (2.5) depends on both y_α and Δy , and neither y_α nor the solution Δy of (2.4) is necessarily unique. It is possible that (2.5) is satisfied by certain $y_\alpha \in Y_\alpha$ and Δy that satisfies (2.4) but not by a different pair. Our solution to this difficulty is: among the solutions of (2.4), obtain Δy that is likely to satisfy (2.5) though not guaranteed; if this fails, then increase α and warmly start the linearized Bregman method. The detail of this procedure is given in the algorithm in next subsection.

2.5. The Enhanced Linearized Bregman Method. The enhanced linearized Bregman method given in Table 2.1 is designed for problem (1.1) with general (non-sparse) solutions. It is *not* best for the compressed sensing problem with a sparse solution as α can be smartly chosen and updated thanks for the solution sparsity. In this subsection, we focus on the strategies of solving for Δy on Line 7 and updating α on Lines 11 and 13, leaving out the implementation of Line 3 to the linearized Bregman algorithm.

In Line 7, (2.4) is solved for Δy (recall that a solution always exists.) If the left-hand side of (2.4) has full rank, it becomes a traditional linear system for which many efficient solvers exist. Otherwise, we suggest

Input: A, b

Output: solution x^* of (1.1)

1. **Initialize:** $l = 0$
 2. **for** $l = 0, \dots, l_{\max}$ **do**
 3. $(x_\alpha, y_\alpha) \leftarrow$ solve (1.3), warm start if $l \geq 1$;
 4. **if** x_α passes sparsity tests (Cases 2 and 3 in Subsection 2.4)
 5. **return** $x^* = x_\alpha$;
 6. **otherwise**
 7. $\Delta y \leftarrow$ solve (2.4);
 8. **if** Δy satisfies (2.5) (thus, both (2.6) and (2.7))
 9. **return** $x^* = x_\alpha$;
 10. **else if** Δy satisfies both (2.6) and (2.7) but not (2.5)
 11. increase α (see (2.17));
 12. **otherwise**
 13. increase α (see second to last paragraph of subsection 2.5);
 14. **end if**
 15. **end if**
 16. **end for**
-

one obtain Δy by solving

$$\min \|\Delta y\|, \text{ subject to (2.4).}$$

Given $\alpha \in I_J$, the minimizer of this problem has a better chance to satisfy (2.5) among the solutions of (2.4).

It is necessary to increase α if x_α cannot be guaranteed optimal. It is important to note that excessively large α will make (1.3) difficult to solve. Therefore, it is not practical to simply choose a large α at the beginning or in an update. On the other hand, an increase in α should be sufficiently large so that there is a good chance to have the corresponding x_α optimal. Bearing this in mind, we treat Lines 11 and 13 differently.

When Line 11 is reached, Δy satisfies (2.4), (2.6), and (2.7). From the discussions in 2.2, x_α remains constant in α locally; it can be either optimal or not. When Line 13 is reached, on the other hand, x_α is guaranteed non-optimal. Therefore, different α -updating strategies should be applied. We suggest a conservative update for Line 11 and a bold one for Line 13.

For Line 11, let β be the largest value such that $y_\beta := y_\alpha + (\beta^{-1} - \alpha^{-1})\Delta y$ satisfies (2.5). Then, set

$$(2.17) \quad \alpha_{\text{new}} \leftarrow 1.05 \times \beta,$$

and warm start the solver of (1.3) with $y_\alpha + (\alpha_{\text{new}} - \alpha^{-1})\Delta y$. Notice that it is possible that $Q_{\alpha_{\text{new}}}$ does not change in the next iteration. If so, Δy does not need to be resolved.

For Line 13, we set α_{new} by counting, as β increases, the total number of components of $(A_\alpha^1)^\top(y_\alpha + (\beta^{-1} - \alpha^{-1})\Delta y)$ become greater than -1 and those of $(A_\alpha^3)^\top(y_\alpha + (\beta^{-1} - \alpha^{-1})\Delta y)$ become less than 1 , called the number of crossings. It can be shown that for $\beta = \infty$, the number of crossings is strictly positive. Let this number be $\gamma > 0$. Therefore, one set α_{new} be smallest β such that the number of crossings is $\lceil \gamma/2 \rceil$, for instance.

In fact, one can show that the above updating strategies guarantee finding an optimal solution after a finite number of updates on α . However, we do not investigate this further because efficient strategies depend on the type of data and solutions.

3. Discussions and Conclusions. One of the main results of this paper is the exact penalty property, which implies that to solve the basis pursuit problem (1.1), one can choose to solve the simpler unconstrained problem (1.3) with an appropriate parameter α using, for example, the linearized Bregman algorithm or its faster generalizations (see Subsection 1.3). On the other hand, it is tricky to choose α : problem (1.3) tends to be numerically easier with a smaller α but α must exceed the exact penalty cutoff. This leaves us the following numerical questions: how to choose α , how to check exact penalty, and how to update α .

In papers [48, 11, 12, 10], the authors have demonstrated good numerical results with relatively small α values in their tested compressed sensing and matrix completion problems. Their empirical choices of α worked fine. For problems with sparse solutions (or low-rank solutions in the matrix completion problem), simple posterior optimality tests are available (see Subsection 2.4). For non-degenerate, non-sparse solutions, solving the linear system (2.4) seems unavoidable. In the worst case with degenerate yet non-sparse solutions, both (2.4) and (2.5) need to be solved.

It is easy to see that the exact penalty cutoff of α depends on the magnitudes of the components of the solution, so no α value will suffice in all problems or just all the problems where solutions are known to be sparse. Therefore, a proper initialization and runtime update of α are needed. This was not studied in [48, 11, 12, 10]. An update strategy for general problems is proposed in Subsection 2.5. For compressed sensing problems with sparse or compressible solutions, we have tested a few strategies. So far, the following strategy seems to have the best performance: one sets an upper bound for the number of nonzero (or dominating) components in the solution; since the linearized Bregman algorithm (including the original and generalizations) tends to generate a sequence of solutions with growing numbers of nonzero (or dominating) components, one increases α whenever the number violates the upper bound. We leave details to our future numerical report.

It is yet to see whether the algorithms and methods studied in this paper can be applied to problems other than those minimizing ℓ_1 , total variation, or the matrix nuclear norm. We are optimistic, especially given the generalizations of the linearize Bregman algorithm that can significantly increase speed but not programming complexity (see Subsections 1.3 and 1.4).

Acknowledgements. The author wants to thank Donald Goldfarb and Zaiwen Wen from Columbia University and Wenye Ma from UCLA for proofreading this paper.

REFERENCES

- [1] K. J. ARROW, L. HURWICZ, AND H. UZAWA, *Studies in Nonlinear Programming*, Stanford University Press, Stanford, CA, 1958.
- [2] J. BARZILAI AND J. BORWEIN, *Two point step size gradient methods*, IMA Journal of Numerical Analysis, 8 (1988), pp. 141–148.
- [3] J. BECT, L. BLANC-FERAUD, G. AUBERT, AND A. CHAMBOLLE, *A ℓ_1 -unified variational framework for image restoration*, European Conference on Computer Vision, Prague, Lecture Notes in Computer Sciences 3024, (2004), pp. 1–13.
- [4] D. P. BERTSEKAS, *Constrained Optimization and Lagrange Multiplier Methods (Paperback)*, Athena Scientific, 1996.
- [5] J.M. BIOUCAS-DIAS AND M. FIGUEIREDO, *A new TwIST: two-step iterative shrinkage/thresholding algorithms for image restoration*, IEEE Transactions on Image Processing, 16 (2007), pp. 2992–3004.
- [6] P. BOUFONOUS, M. DUARTE, AND R. BARANIUK, *Sparse signal reconstruction from noisy compressive measurements using cross validation*, IEEE Workshop on Statistical Signal Processing, (2007), pp. 299–303.
- [7] L. BREGMAN, *The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming*, USSR Computational Mathematics and Mathematical Physics, 7 (1967), pp. 200–217.
- [8] M. BURGER, G. GILBOA, S. OSHER, AND J. XU, *Nonlinear inverse scale space methods*, Communications in Mathematical Sciences, 4 (2006), pp. 175–208.
- [9] M. BURGER, S. OSHER, J. XU, AND G. GILBOA, *Nonlinear inverse scale space methods for image restoration*, Variational, Geometric, and Level Set Methods in Computer Vision, Lecture Notes in Computer Science 3752, (2005), pp. 25–36.
- [10] J.-F. CAI, E CANDÈS, AND Z. SHEN, *A singular value thresholding algorithm for matrix completion*, arXiv:0810.3286, (2008).
- [11] J.-F. CAI, S. OSHER, AND Z. SHEN, *Linearized Bregman iterations for compressed sensing*, To appear in Mathematics of Computation, (2008).
- [12] ———, *Convergence of the linearized Bregman iteration for ℓ_1 -norm minimization*, To appear in Mathematics of Computation, (2009).
- [13] ———, *Linearized Bregman iterations for frame-based image deblurring*, SIAM Journal on Imaging Sciences, 2 (2009), pp. 226–252.
- [14] E. CANDÈS, J. ROMBERG, AND T. TAO, *Stable signal recovery from incomplete and inaccurate information*, Communications on Pure and Applied Mathematics, 2005 (2005), pp. 1207–1233.
- [15] A. CHAMBOLLE, *An algorithm for total variation minimization and applications*, Journal of Mathematical Imaging and Vision, 20 (2004), pp. 89–97.
- [16] ———, *Total variation minimization and a class of binary MRF models*, Tech. Report UMR CNRS 7641, Ecole Polytechnique, 2005.
- [17] P. L. COMBETTES AND J.-C. PESQUET, *Proximal thresholding algorithm for minimization over orthonormal bases*, To appear in SIAM Journal on Optimization, (2007).
- [18] Y.H. DAI AND R. FLETCHER, *Projected barzilai-borwein methods for large-scale box-constrained quadratic programming*, Numerische Mathematik, 100 (2005), pp. 21–47.
- [19] J. DARBON AND M. SIGELLE, *Image restoration with discrete constrained total variation, Part I: fast and exact optimization*, Journal of Mathematical Imaging and Vision, 26 (2006), pp. 261–276.
- [20] I. DAUBECHIES, M. DEFRISE, AND C. DE MOL, *An iterative thresholding algorithm for linear inverse problems with a sparsity constraint*, Communications in Pure and Applied Mathematics, 57 (2004), pp. 1413–1457.
- [21] C. DE MOL AND M. DEFRISE, *A note on wavelet-based inversion algorithms*, Contemporary Mathematics, 313 (2002), pp. 85–96.
- [22] D. DONOHO, *Compressed sensing*, IEEE Transactions on Information Theory, 52 (2006), pp. 1289–1306.
- [23] M. ELAD, *Why simple shrinkage is still relevant for redundant representations?*, IEEE Transactions on Information Theory, 52 (2006), pp. 5559–5569.
- [24] M. ELAD AND A. BRUCKSTEIN, *A generalized uncertainty principle and sparse representations in pairs of bases*, IEEE Transactions on Information Theory, 48 (2002), pp. 2558–2567.
- [25] M. ELAD, B. MATALON, J. SHTOK, AND M. ZIBULEVSKY, *A wide-angle view at iterated shrinkage algorithms*, SPIE (Wavelet XII), San-Diego CA, August 26-29, 2007., (2007).
- [26] ERNIE ESSER, *Applications of Lagrangian-based alternating direction methods and connections to split Bregman*, UCLA

- CAM Report 09-31, (2009).
- [27] A. FEUER AND A. NEMIROVSKI, *On sparse representation in pairs of bases*, IEEE Transactions on Information Theory, 49 (2003), pp. 1579–1581.
 - [28] M. FIGUEIREDO AND R. NOWAK, *An EM algorithm for wavelet-based image restoration*, IEEE Transactions on Image Processing, 12 (2003), pp. 906–916.
 - [29] M. FIGUEIREDO, R. NOWAK, AND S. J. WRIGHT, *GPSR: Gradient projection for sparse reconstruction*, <http://www.lx.it.pt/~mtf/gpsr/>, (2007).
 - [30] D. GOLDFARB AND W. YIN, *Parametric maximum flow algorithms for fast total variation minimization*, Rice University CAAM Technical Report TR07-09, (2007).
 - [31] T. GOLDSTEIN, X. BRESSON, AND S. OSHER, *Geometric applications of the split Bregman method: Segmentation and surface reconstruction*, UCLA CAM Report 09-06, (2009).
 - [32] T. GOLDSTEIN AND S. OSHER, *The split Bregman algorithm for L1 regularized problems*, UCLA CAM Report 08-29, (2008).
 - [33] R. GRIBONVAL, R. M. FIGUERAS I VENTURA, AND P. VANDERGHEYNST, *A simple test to check the optimality of a sparse signal approximation*, EURASIP Signal Processing, special issue on Sparse Approximations in Signal and Image Processing, 86 (2006), pp. 496–510.
 - [34] W. GUO AND F. HUANG, *A local mutual information guided denoising technique and its application to self-calibrated partially parallel imaging*, Proceedings of Medical Image Computing and Computer Assisted Intervention, Part II, Lecture notes on Computer Science 5242, (2008), pp. 937–947.
 - [35] E. HALE, W. YIN, AND Y. ZHANG, *A fixed-point continuation method for ℓ_1 -regularization with application to compressed sensing*, Rice University CAAM Technical Report TR07-07, (2007).
 - [36] E. T. HALE, W. YIN, AND Y. ZHANG, *Fixed-point continuation for ℓ_1 -minimization: methodology and convergence*, Submitted to SIAM Journal on Optimization, (2007).
 - [37] L. HE, T.-C. CHANG, S. OSHER, T. FANG, AND P. SPEIER, *MR image reconstruction by using the iterative refinement method and nonlinear inverse scale space methods*, UCLA CAM Report 06-35, (2006).
 - [38] L. HE, A. MARQUINA, AND S. OSHER, *Blind deconvolution using TV regularization and Bregman iteration*, International Journal of Imaging Systems and Technology, 5 (2005), pp. 74–83.
 - [39] M. R. HESTENES, *Multiplier and gradient methods*, Journal of Optimization Theory and Applications, 4 (1969), pp. 303–320.
 - [40] A. JUDITSKY AND A. NEMIROVSKI, *On verifiable sufficient conditions for sparse signal recovery via ℓ_1 minimization*, preprint, <http://hal.archives-ouvertes.fr/docs/00/32/17/75/PDF/CSNote-Submitted.pdf>, (2008).
 - [41] K. KOH, S.-J. KIM, AND S. BOYD, *ℓ_1 - ℓ_s : A simple MATLAB solver for ℓ_1 -regularized least squares problems*, http://www.stanford.edu/~boyd/l1_ls/, (2007).
 - [42] D. C. LIU AND J. NOCEDAL, *On the limited memory method for large scale optimization*, Mathematical Programming, Series B, 45 (1989), pp. 503–528.
 - [43] S. MA, D. GOLDFARB, AND L. CHEN, *Fixed point and bregman iterative methods for matrix rank minimization*, Submitted to Mathematical Programming, (2009).
 - [44] Y. NESTEROV, *Gradient methods for minimizing composite objective function*, www.optimization-online.org, CORE Discussion Paper 2007/76, (2007).
 - [45] J. NOCEDAL AND S. J. WRIGHT, *Numerical Optimization*, Springer-Verlag, New York, 1999.
 - [46] R. NOWAK AND M. FIGUEIREDO, *Fast wavelet-based image deconvolution using the EM algorithm*, Proceedings of the 35th Asilomar Conference on Signals, Systems, and Computers, Monterey, CA, (2001).
 - [47] S. OSHER, M. BURGER, D. GOLDFARB, J. XU, AND W. YIN, *An iterative regularization method for total variation-based image restoration*, SIAM Journal on Multiscale Modeling and Simulation, 4 (2005), pp. 460–489.
 - [48] S. OSHER, Y. MAO, B. DONG, AND W. YIN, *Fast linearized Bregman iteration for compressive sensing and sparse denoising*, To appear in Communications in Mathematical Sciences. Rice University CAAM Technical Report TR08-07., (2008).
 - [49] M. J. D. POWELL, *A method for nonlinear constraints in minimization problems*, in Optimization, R. Fletcher, ed., Academic Press, New York, 1972, pp. 283–298.
 - [50] R.T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, 1970.

- [51] R. T. ROCKAFELLAR, *A dual approach to solving nonlinear programming problems by unconstrained optimization*, Mathematical Programming, 5 (1973), pp. 354–373.
- [52] M. RUDELSON AND R. VERSHYNIN, *Geometric approach to error correcting codes and reconstruction of signals*, International Mathematical Research Notices, 64 (2005), pp. 4019–4041.
- [53] L. RUDIN, S. OSHER, AND E. FATEMI, *Nonlinear total variation based noise removal algorithms*, Physica D, 60 (1992), pp. 259–268.
- [54] X. C. TAI AND CHUNLIN WU, *Augmented Lagrangian method, dual methods and split Bregman iteration for ROF model*, UCLA CAM Report 09-05, (2009).
- [55] R. TIBSHIRANI, *Regression shrinkage and selection via the lasso*, Journal Royal Statistical Society B, 58 (1996), pp. 267–288.
- [56] E. VAN DEN BERG AND M. P. FRIEDLANDER, *SPGL1: A MATLAB solver for large-scale sparse reconstruction*, <http://www.cs.ubc.ca/labs/scl/index.php/main/spgl1>, (2007).
- [57] Y. WANG, J. YANG, W. YIN, AND Y. ZHANG, *A new alternating minimization algorithm for total variation image reconstruction*, To appear in SIAM Journal on Imaging Sciences, (2007).
- [58] R. WARD, *Compressed sensing with cross validation*, arXiv:0803.1845v2, (2008).
- [59] Z. WEN, W. YIN, D. GOLDFARB, AND Y. ZHANG, *A fast algorithm for sparse reconstruction based on shrinkage, subspace optimization and continuation*, Submitted to SIAM Journal on Scientific Computing. Rice University CAAM Technical Report TR09-01, (2009).
- [60] S. WRIGHT, R. NOWAK, AND M. FIGUEIREDO, *Sparse reconstruction by separable approximation*, Submitted, (2008).
- [61] J. XU AND S. OSHER, *Iterative regularization and nonlinear inverse scale space applied to wavelet-based denoising*, IEEE Transactions on Image Processing, 16 (2006), pp. 534–544.
- [62] J. YANG, W. YIN, Y. ZHANG, AND Y. WANG, *A fast algorithm for edge-preserving variational multichannel image restoration*, Rice University CAAM Technical Report TR08-09, (2008).
- [63] J. YANG, Y. ZHANG, AND W. YIN, *An efficient TVL1 algorithm for deblurring multichannel images corrupted by impulsive noise*, Rice University CAAM Technical Report TR08-12, (2008).
- [64] ———, *A fast TVL1-L2 algorithm for image reconstruction from partial fourier data*, Submitted to IEEE Journal of Selected Topics in Signal Processing Special Issue on Compressed Sensing. Rice University CAAM Technical Report TR08-27, (2008).
- [65] W. YIN, S. OSHER, D. GOLDFARB, AND J. DARBON, *Bregman iterative algorithms for ℓ_1 -minimization with applications to compressed sensing*, SIAM Journal on Imaging Sciences, 1 (2008), pp. 143–168.
- [66] X. ZHANG, M. BURGER, X. BRESSON, AND S. OSHER, *Bregmanized nonlocal regularization for deconvolution and sparse reconstruction*, UCLA CAM Report 09-03, (2009).
- [67] Y. ZHANG, *A simple proof for recoverability of ℓ_1 -minimization: go over or under?*, Rice University CAAM Technical Report TR05-09, (2005).
- [68] ———, *When is missing data recoverable?*, Rice University CAAM Technical Report TR06-15, (2006).
- [69] W. P. ZIEMER, *Weakly Differentiable Functions: Sobolev Spaces and Functions of Bounded Variation*, Graduate Texts in Mathematics, Springer, 1989.