Open access • Journal Article • DOI:10.1007/S10915-012-9616-5

Error Forgetting of Bregman Iteration — Source link

Wotao Yin, Stanley Osher

Institutions: Rice University, University of California, Los Angeles

Published on: 01 Feb 2013 - Journal of Scientific Computing (Springer US)

Topics: Bregman divergence

Noname manuscript No.
(will be inserted by the editor)

Error Forgetting of Bregman Iteration

Wotao Yin[§] · **Stanley Osher**[†]

the date of receipt and acceptance should be inserted later

Abstract This short article analyzes an interesting property of the Bregman iterative procedure, which is equivalent to the augmented Lagrangian method, for minimizing a convex piece-wise linear function $J(x)$ subject to linear constraints $Ax = b$. The procedure obtains its solution by solving a sequence of unconstrained subproblems of minimizing $J(x) + \frac{1}{2}\|Ax - b^k\|_2^2$, where b^k is iteratively updated. In practice, the subproblem at each iteration is solved at a relatively low accuracy. Let w^k denote the error introduced by early stopping a subproblem solver at iteration k. We show that if all w^k are sufficiently small so that Bregman iteration enters the optimal face, then while on the optimal face, Bregman iteration enjoys an interesting error-forgetting property: the distance between the current point $\bar{x}^k$ and the optimal solution set X^* is bounded by $\|w^{k+1} - w^k\|$, independent of the previous errors $w^{k-1}, w^{k-2}, \ldots, w^1$. This property partially explains why the Bregman iterative procedure works well for sparse optimization and, in particular, for ℓ_1-minimization. The error-forgetting property is unique to $J(x)$ that is a piece-wise linear function (also known as a polyhedral function), and the results of this article appear to be new to the literature of the augmented Lagrangian method.

Keywords Bregman iteration · error forgetting · sparse optimization · ℓ_1 minimization, piece-wise linear function, polyhedral function

1 Introduction

This article studies an interesting numerical property, which we call *error forgetting*, of the Bregman iterative procedure for solving the convex problem

$$\min_{x \in \mathbb{R}^n} \{J(x) : Ax = b\} \tag{1}$$

with a piece-wise linear function J, where matrix $A \in \mathbb{R}^{m \times n}$ and vector $b \in \mathbb{R}^m$ are given. A piece-wise linear function is also called a polyhedral function. Among the well-known examples of piece-wise linear functions is $J(x) = \mu\|x\|_1$, where μ is a constant scalar. The basis pursuit problem

$$\min\{\|x\|_1 : Ax = b\}, \tag{2}$$

tends to return a *sparse* solution of $Ax = b$. Problem (2) and its variants arise in compressive sensing, statistical learning, signal and image processing, etc. The Bregman iterative procedure [1,2] is based on solving subproblems of the form

$$\min_x F(x; d) := J(x) + \frac{1}{2}\|Ax - d\|_2^2, \tag{3}$$

The authors' work has been supported by NSF, ONR, DARPA, AFOSR, ARL and ARO.

Dedicated to the 70th birthday of Stanley Osher

§ Department of Computational and Applied Mathematics, Rice University, Houston, TX, USA. E-mail: wotao.yin@rice.edu
† Department of Mathematics, UCLA, Los Angeles, CA, USA. E-mail: sjo@math.ucla.edu

and iteratively performing

$$x^k \leftarrow \arg\min_x F(x; b^k) \tag{4a}$$

$$b^{k+1} \leftarrow b + (b^k - Ax^k) \tag{4b}$$

for $k = 1, \ldots$ starting with $b^1 = b$. Subproblem (3) can be efficiently solved for various different convex functions $J(\cdot)$. For $J(x) = \mu\|x\|_1$, there are first-order algorithms such as GPSR [3], FPC [4], SpaRSA [5], and so on.

In practice, the update (4a) is computed *inexactly*, giving rise to the inexact iteration

$$\bar{x}^k \leftarrow \arg\min_x F(x; \bar{b}^k) + w^k, \tag{5a}$$

$$\bar{b}^{k+1} \leftarrow b + (\bar{b}^k - A\bar{x}^k), \tag{5b}$$

Step (5a) introduces errors w^k, and they roll over to subsequent iterations through (5b). In general, these errors may accumulate over the iterations or can even cause the iteration to diverge. However, we show that as long as (5a) is solved with sufficient accuracy (which is not necessarily very high), the errors do not accumulate and can even cancel each other under certain situations. Consequently, the inexact iteration (5) can yield a highly accurate solution to (1) for a moderate amount of computation on every subproblem.

1.1 Background of Bregman Iteration

Bregman iteration (4) was introduced in [1] for total variation based image processing, extended to wavelet-based denoising in [6], nonlinear inverse scale space in [7, 8], MR imaging [9], matrix rank minimization [10], hyperspectral imaging [11], etc. Its usefulness for solving ℓ_1 and ℓ_1-related problems has been discussed in [2]. In this subsection, we briefly discuss another form of the iterative procedure.

The Bregman distance [12] induced by a convex function $J(\cdot)$ is defined as

$$D_J^p(u, v) = J(u) - J(v) - \langle p, u - v\rangle, \quad \text{where } p \in \partial J(v). \tag{6}$$

Because $D_J^p(u, v) \neq D_J^p(v, u)$, $D_J^p(u, v)$ is not a distance in the usual sense. Yet, 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 any point w being a convex combination of u and v.

Instead of solving (1) with the constraints, at each Bregman iteration, a problem in the form of

$$\min_x D_J^p(x, y) + \frac{1}{2}\|Ax - d\|,^2 \tag{7}$$

is solved, and the iteration is

$$x^k \leftarrow \text{solve (7) with } p := p^{k-1}, y := x^{k-1}, d := b, \tag{8a}$$

$$p^k \leftarrow p^{k-1} + A^\top(b - Ax^{k-1}). \tag{8b}$$

for $k = 1, \ldots$ starting with $x^0 = \mathbf{0}$ and $p^0 = \mathbf{0}$. Since $J(u)$ is not differentiable everywhere, there can be multiple choices for p in definition (6). Nevertheless p can be uniquely determined from the previous iteration using the optimality condition of (8a):

$$\mathbf{0} \in \partial J(x^k) - p^{k-1} + A^\top(Ax^k - b),$$

which gives update (8b). Assuming that (4a) and (8a) are computed exactly, one can verify that, for all iterations k, iteration (8a)–(8b) and (4a)–(4b) are equivalent through the identify $p^k = A^\top(b^k - Ax^k)$; see Theorem 3.1 of [2]. However, inexact computation of the subproblem in either iteration breaks down the equivalence.

We remind the reader not to confuse Bregman iteration with *linearized Bregman iteration* [2]:

$$x^k \leftarrow \text{solve (10) with } p = p^{k-1}, y = x^{k-1}, d = b, \tag{9a}$$

$$p^k \leftarrow p^{k-1} + A^\top(b - Ax^{k-1}) - \frac{1}{\alpha}(x^k - x^{k-1}), \tag{9b}$$

where

$$\min_x D_J^p(x, y) + \langle A^\top(Ay - b), x\rangle + \frac{1}{2\alpha}\|x - y\|_2^2 \tag{10}$$

is the prox-linear variant of (7). Not only do the two iterations solve different subproblems, they generate two sequences $\{x^k\}$ converging to different solutions. For example, if $J(x) = \|x\|_1$, the sequence generated by (9) converges to the solution of

$$\min\{\|x\|_1 + \frac{1}{2\alpha}\|x\|_2^2 : Ax = b\}, \tag{11}$$

which is a result first shown in [13] under some assumptions and then improved in [14, 15]. (Although with sufficiently large α, the solution of (11) becomes a solution to (2), according to analysis in [15–17].) Also see [18] for its application to matrix completion and equivalence to Uzawa's algorithm, [15] for its equivalence to dual gradient descent, and [17] for a global linear convergence proof.

Turning back to Bregman iteration, its equivalence to augmented Lagrangian iteration [19, 20] has been established in [2] and reviewed in [21]. We can introduce Lagrange multiplier $\lambda^k = b^k - b$; then from (4b) we obtain the update $\lambda^{k+1} \leftarrow \lambda^k + (b - Ax^k)$. Inexact augmented Lagrangian iteration has been studied in [22] and [23] where the first-order optimality conditions of the subproblem (4a) are satisfied up to tolerance ϵ_k that decreases to 0 in k. However, we study the case (5) where the error $\|w^k\|$ is sufficiently small but not diminishing in the limit and it is dominated by the error due to an early-terminated subproblem.

1.2 Organization

The main results are given in Section 2, in which subsection 2.1 introduces important notation, subsection 2.2 presents a numerical example to motivate the analysis in subsequent subsections. Subsection 2.3 reviews the existing convergence results. Subsections 2.4 and 2.5 study the error-forgetting and the strong error-cancellation property of Bregman iteration. Finally, Section 3 summarizes this article.

2 Main Results

2.1 Notation

The nature of iterative error analysis requires different sequences, which we now define:

- Exact sequences: $\{x^1, x^2, \ldots, x^k, \ldots\}$ and $\{b^1, b^2, \ldots, b^k, \ldots\}$ from *exactly computed* update (4).
- Inexact numerical sequences: $\{\bar{x}^1, \bar{x}^2, \ldots, \bar{x}^k, \ldots\}$ and $\{\bar{b}^1, \bar{b}^2, \ldots, \bar{b}^k, \ldots\}$ generated by (5), where w^k is the error at iteration k.
- Quasi-exact sequences: $\{\hat{x}^k\}$ and $\{\hat{b}^k\}$, which are artificial sequences defined for the convenience of analysis as follows. Each $\hat{x}^k = \arg\min_x F(x; \bar{b}^k)$ is the *exact* solution of (3) given the *inexact* input $\bar{b}^k$, and $\hat{b}^{k+1} = b + (\bar{b}^k - A\hat{x}^k)$.

In general, x^k, $\bar{x}^k$, and $\hat{x}^k$ are different from one another. Unless the word *inexact* is used, the computation is exact by default.

Note that w^k does *not* equal the collective error $\bar{x}^k - x^k$, which depends not only w^k but also $w^{k-1}, \ldots, w^1$. Instead, w^k is just the new error introduced at iteration k, and w^k obeys $w^k := \bar{x}^k - \hat{x}^k$.

X^* denotes the set of solutions of (1), and $x^* \in X^*$ denotes a solution of (1), which should be clear from the context.

2.2 A Motivating Example

In this simple test, the true solution $x^* \in \mathbb{R}^{500}$ had 25 nonzero entries that are sampled from i.i.d. Gaussian and located uniformly at random. Matrix $A \in \mathbb{R}^{250\times500}$ had i.i.d. Gaussian entries, and $b := Ax^*$. The exact solution of problem (2) was x^*.

We set $\mu = 0.01$ and ran inexact Bregman iteration (5). We repeated the same test five times, each time with one of the five different subproblem solvers: ISTA iteration (17), FPC–BB [4], GPSR [3], SpaRSA [5], and Mosek [24]. All the solvers started from 0 at each Bregman iteration (i.e., no warm start) and had the stopping

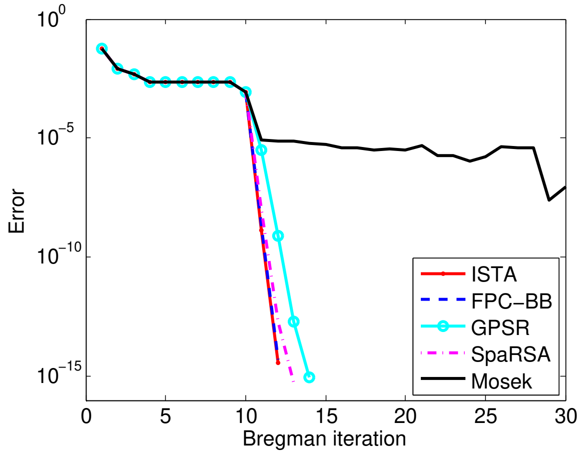

Fig. 1 The iterative errors, $\|\bar{x}^k - x^*\|_2/\|x^*\|_2$, of Bregman iteration using five different subproblem solvers. At each k, each of these solvers starts from 0 and stops at the ℓ_2 tolerance of 1e-6.

tolerance[1] of 10^{-6}. With this tolerance, the truncation error due to early termination dominated the round-off error. Hence, each solver at each Bregman iteration generated a point $\bar{x}^k$ containing an error on the order of 10^{-6}.

We recorded the normalized errors of $\bar{x}^k$ to x^*, $\|\bar{x}^k - x^*\|_2/\|x^*\|_2$, at different Bregman iterations. The Bregman iteration was stopped once this error fell below 10^{-14}, where round-off errors started to dominate the truncation error. These errors are depicted in Figure 1.

According to Figure 1, when the subproblems were solved by ISTA, FPC-BB, GPSR, or SpaSPA, the errors of $\bar{x}^k$ to x^* were not affected by the truncation errors $\approx 10^{-6}$ contained in $\bar{x}^1, \bar{x}^2, \ldots$. Not only did the truncation errors not accumulate, but they were canceling each other, letting $\bar{x}^{10}$ through $\bar{x}^{14}$ converge to $\mathbf{x}*$ at a geometric speed. When the subproblems were solved by Mosek, the truncation errors did not accumulate, yet the errors of $\bar{x}^k$ to x^* stagnated around 10^{-6}, the level of the truncation errors. We call the non-accumulating phenomenon *error forgetting* (see Subsection 2.4) and the cancellation phenomenon *error cancellation* (see Subsection 2.5). Roughly speaking, error cancellation does not occur with Mosek since it is based on the interior-point method, which consists of a series of linear and nonlinear operations, whereas the other four codes are based on operations either linear or piece-wise linear in the input (see Subsection 2.5).

The Matlab code to reproduce this test is available for download from the first author's webpage.

2.3 Related Existing Results

We first review some related existing results.

Assumption 1 *Let function $H(x)$ generalize the data fidelity term $\frac{1}{2}\|Ax - b\|_2^2$. $H(x)$ is convex and differentiable. Furthermore, every subproblem solution x^k exists.*

The following theorem summarizes the convergence of Bregman iteration.

Theorem 1 *([1]. Convergence) Under Assumption 1, the sequence $\{x^k\}$ of Bregman iteration (4) or (8) has the following properties:*

1. *Monotonic decrease in $H(\cdot)$: $H(x^{k+1}) \leq D_J^{p^k}(x^{k+1}) + H(x^{k+1}) \leq H(x^k)$.*
2. *Convergence to x^* in H with exact data: if x^* minimizes $H(\cdot)$ and satisfies $J(x^*) < \infty$, then $H(x^k) \leq H(x^*) + J(x^*)/k$.*

[1] The different solvers use different stopping criteria. ISTA and FPC-BB compute the ℓ_2-distance between 0 and the subdifferential of (7), scaled by μ, and compare it with the tolerance. GPSR and SpaRSA use the same ℓ_2-distance except it is not scaled by μ but normalized by the ℓ_2-norm of x instead, along with other minor differences. Underlying Mosek is an interior-point algorithm, which uses stopping criteria based on primal and dual feasibility violations, the duality gap, and the relative complementarity gap. For more details, the interested reader is referred to the companion code of this example.

3. *Convergence to x^* in $D(\cdot)$ with noisy data: Let $H(\cdot) = H(\cdot; b)$ (e.g., $H(x) = \frac{1}{2}\|Ax - b\|^2$) and suppose that $\tilde{x}$ obeys $H(\tilde{x}; b) \leq \delta$ and $H(\tilde{x}; b^0) = 0$ (b, b^0, $\tilde{x}$, and δ represent the noisy input, noiseless input, true signal, and noise level, respectively). Then, $D_J^{p^{k+1}}(\tilde{x}, x^{k+1}) < D_J^{p^k}(\tilde{x}, x^k)$ for k obeying $H(x^{k+1}; b) > \delta$.*

Remark 1 Point 2 indicates that if the input b is noiseless (i.e., $H(x^*; b) = \mathbf{0}$; e.g., $Ax^* = b$), then $H(u^k) \to H(x^*)$.

Theorem 2 *([2]. Finite convergence.) Consider Bregman iteration (4) or (8) applied to the ℓ_1–minimization problem (2), where $J(x) = \mu\|x\|_1$ and $\mu > 0$. There exists $K > 0$ such that any x^k, $k > K$, is a solution of (2).*

Proof For completeness, we give a sketchy proof. Using vectors $p \in [-\mu, \mu]^n$, we define the subsets

$$U(p) := \{x \in \mathbb{R}^n : x_i \geq 0, \text{ if } p_i = \mu; x_i \leq 0, \text{ if } p_i = -\mu; x_i = 0, \text{ if } p_i \in (-\mu, \mu)\}.$$

There are finitely many distinct subsets $U(p)$, and their union is $\mathbb{R}^n$. Furthermore, we have $x^k \in U^k := U(p^k)$ (otherwise, $p^k \notin \partial(\mu\|x^k\|_1)$) and $D^{p^k}(x, x^k) = 0$ for any $x \in U^k$. Since x^* satisfies $Ax^* = b$ or $H(x^*) = 0$, the convergence $x^k \to x^*$ gives $H(x^k) \to H(x^*) = 0$. The finiteness of $\{U^k\}$ and the monotonicity of H in k dictate the existence of $K > 0$ such that $\min\{H(x) : x \in U^k\} = 0$ for all $k \geq K$. Therefor, for $k > K$, there exists $x \in U^k$ such that $D^{p^k}(x; x^k) + H(x) = 0$, so zero is the best possible value of (8a). Hence, $D^{p^k}(x^{k+1}; x^k) = 0$ and $H(x^{k+1}) = 0$. From Lemma 1 below, x^{k+1} is optimal. Since $p^k = p^{k+1} = \cdots$, (8a) remains the same problem; hence, all subsequent solutions are optimal. □

Lemma 1 *Suppose that in Bregman iteration (4) or (8) is applied to problem (1) with a convex function J. Once $Ax^k = b$ holds, x^k is a solution of (1).*

Proof The Bregman distance $D_J^p(\cdot, \cdot)$ induced by a convex function J is nonnegative, and according to iteration (8), $p^k \in \text{Range}(A^\top)$. Therefore, there exists vector c such that

$$\begin{aligned} J(x^k) &\leq J(x) - \langle p^k, x - x^k \rangle \\ &= J(x) - \langle A^\top c, x - x^k \rangle, \\ &= J(x) - \langle c, Ax - Ax^k \rangle \\ &= J(x) - \langle c, Ax - b \rangle. \end{aligned}$$

The above inequality shows that for any x satisfying $Ax = b$, $J(x^k) \leq J(x)$. Therefore, x^k is a solution of (1). □

It is not difficult to extend the result of Theorem 2 from $J(x) = \mu\|x\|_1$ to any piece-wise linear function $J(x)$ by defining subsets $U(p)$ based on $p \in \partial J$.

2.4 Error Forgetting

We start with an assumption below on the finite identification of the optimal face, namely, the exact and inexact sequences stay on the optimal face after a certain iteration K. While the faces of the ℓ_1 function can be conveniently characterized by the signs of the input, such characterization does not exist for a general convex function $J(x)$. However, observe that the Bregman distance (6) between two different points equals 0 only if between these two points J is linear, namely, the two points belong to the same face of J. Hence, we employ the Bregman distance in the assumption below for general piece-wise linear convex $J(x)$.

Assumption 2 *Let $\sigma > 0$ and $K > 0$ be such that as long as $\|w^k\| < \sigma$ holds uniformly for all k, we have (i) for $J(x) = \mu\|x\|_1$,*

$$\text{sign}(x^*) = \text{sign}(x^k) = \text{sign}(\bar{x}^k) = \text{sign}(\hat{x}^k) \tag{12}$$

holds for all $k \geq K$, and (ii) for general piece-wise linear convex $J(x)$,

$$D_J^p(x^*, x^k) = 0 \quad \textit{for any } p \in \partial J(x^k), \tag{13a}$$

$$D_J^p(x^*, \bar{x}^k) = 0 \quad \textit{for any } p \in \partial J(\bar{x}^k), \tag{13b}$$

$$D_J^p(x^*, \hat{x}^k) = 0 \quad \textit{for any } p \in \partial J(\hat{x}^k) \tag{13c}$$

hold for all $k \geq K$.

Remark 2 Under Assumption 2, all the sequences stay on the optimal face after a certain iteration K. There are results on the finite-identification property in the literature for the augmented Lagrangian iteration (e.g., [25, 26]), but nevertheless, they do not address inexact iterations with non-diminishing errors or give (12) or (13). Since error forgetting will be shown to occur after $k \geq K$ and yet establishing Assumption 2 (which is not our focus) is quite involved, we decide to leave it as an assumption with a non-rigorous explanation as follows. Theorem 2 addresses the finite property for the exact sequence x^k. Existing results such as [26, 27] on inexact iterations assert if the errors decay quickly enough, the inexact sequence $\bar{x}^k$ will converge to x^*. On the other hand, we can apply the arguments in the proof of Theorem 2 to $\bar{x}^k$: $\mathbb{R}^n$ is decomposed to finitely many subsets $U(p)$, among which the optimal U^k satisfies $\min\{H(x) : x \in U^k\} = 0$, and since $\bar{x}^k \to x^*$, after finitely many iterations, $\bar{x}^k$ will stay in the optimal U^k, which means $\bar{x}^k_i = 0$ if $x^*_i = 0$. Apparently, $\bar{x}^k \to x^*$ also gives, after finitely many iterations, $\bar{x}^k_i > 0$ if $x^*_i > 0$ and $\bar{x}^k_i < 0$ if $x^*_i < 0$. Furthermore, we shall be able to extend the property to $\hat{x}^k$ since it is no more inaccurate than $\bar{x}^k$.

Remark 3 It is generally difficult to predict K or to detect $k > K$ unless more computation is done. Methods such as the adaptive inverse scale space method [28] are able to detect $k > K$ by solving nonnegative least-squares problems. In fact, that method finds the solution of (2) by solving a sequence of nonnegative least-squares problems, each of which either identities the solution or provides information to update the current point. It is related to Bregman iteration, so the error forgetting property might also apply.

Theorem 3 (Two exact steps on the optimal face) *Under Assumptions 1 and 2, consider* inexact *Bregman iteration (5) applied to problem (1) with a piece-wise linear $J(x)$ through iteration $k-1$, where $k > K$. Then, two* exact *steps of (4a) at iterations k and $k+1$ yield an exact solution to (1).*

Proof As an exact minimizer of $F(x; \bar{b}^k)$, $\hat{x}^k$ satisfies the first order optimality condition:

$$\mathbf{0} = \hat{p}^k + A^\top(A\hat{x}^k - \bar{b}^k), \quad \text{where } \hat{p}^k \in \partial J(\hat{x}^k).$$

Hence, $A^\top \hat{b}^{k+1} = A^\top b + A^\top(\bar{b}^k - A\hat{x}^k) = A^\top b + \hat{p}^k$. Hence, $\breve{x}$ is the exact solution to

$$\min_x J(x) + \frac{1}{2}\left\|Ax - \hat{b}^{k+1}\right\|_2^2, \tag{14}$$

or after re-organizing the terms,

$$\min_x D_J^{\hat{p}^k}(x, \hat{x}^k) + \frac{1}{2}\|Ax - b\|_2^2.$$

According to Assumption 2 and $k > K$, the subdifferential $\hat{p}^k \in \partial J(\hat{x}^k)$ satisfies $D_J^{\hat{p}^k}(x^*, \hat{x}^k) = 0$ for $x^* \in X^*$. Hence, considering the facts that $D_J^{\hat{p}^k}(x, \hat{x}^k) + \frac{1}{2}\|Ax - b\|_2^2 \geq 0$, $\forall x$, and $D_J^{\hat{p}^k}(x^*, \hat{x}^k) + \frac{1}{2}\|Ax^* - b\|_2^2 = 0$, we conclude that the solution $\breve{x}$ obeys $D_J^{\hat{p}^k}(\breve{x}, \hat{x}^k) + \frac{1}{2}\|A\breve{x} - b\|_2^2 = 0$ and thus $A\breve{x} = b$. Although Lemma 1 only addresses exact Bregman iteration, its proof applies to $\breve{x}$; in particular, the first-order optimality conditions of (14) are $\mathbf{0} = \breve{p} + A^\top(A\breve{x} - \hat{b}^{k+1})$, where $\breve{p} \in \partial J(\breve{x})$, and thus $\breve{p} \in \text{Range}(A^\top)$. Therefore, $\breve{x} \in X^*$. □

Theorem 3 means that once Bregman iteration (exact or not) identifies the optimal support, another two *exact* steps will yield the exact solution to problem (1). The first exact step corrects the subdifferential $\hat{p}^k$ (or the Lagrange multiplier, in terms of augmented Lagrangian iteration), which allows the second step to find the exact solution to problem (1).

Since it is difficult to know when the optimal support has surely been identified, it is hard for one to decide when to take the two exact steps. So, Theorem 3 does not provide a powerful computational rule. Nevertheless, Theorem 3 introduces $\breve{x}$, which is used in the analysis below.

Based on Theorem 3, we can treat $\breve{x}$ as a solution to problem (1), just the same as x^*. Although in case X^* has multiple elements, different $\breve{x}$ may be obtained by two exact steps starting at different $k > K$, this is not a problem in our analysis below, so we do not associate $\breve{x}$ with k.

Theorem 4 (Error forgetting) *Consider* inexact *Bregman iteration (5) applied to problem (1) with a piece-wise linear $J(x)$. Define $\breve{x} \in X^*$ as in Theorem 3, i.e., $\breve{x}$ is the solution of (1) that one would get after applying two exact Bregman steps on the optimal face. Under Assumption 1 and the additional assumption $-A^\top(A\breve{x} - \hat{b}^{k+1}) \in \partial J(\breve{x} - w^k)$ (see Remark 4 below), we have*

$$\bar{x}^{k+1} - \breve{x} = w^{k+1} - w^k. \tag{15}$$

Proof By definition of $\breve{x}$ in Theorem 3, $\breve{x}$ is an exact minimizer of $J(x) + \frac{1}{2}\|A\mathbf{x} - \hat{b}^{k+1}\|^2$, so we have $-A^\top(A\breve{x} - \hat{b}^{k+1}) \in \partial J(\breve{x})$. The assumption $-A^\top(A\breve{x} - \hat{b}^{k+1}) \in \partial J(\breve{x} - w^k)$ means that $\breve{x}$ is also an exact minimizer of $J(x - w^k) + \frac{1}{2}\|A\mathbf{x} - \hat{b}^{k+1}\|^2$, a fact we shall use below.

From the definition $w^k = \bar{x}^k - \hat{x}^k$ and the updates $\hat{b}^{k+1} = b + (\bar{b}^k - A\hat{x}^k)$ and $\bar{b}^{k+1} = b + (\bar{b}^k - A\bar{x}^k)$, we have $\bar{b}^{k+1} = \hat{b}^{k+1} - A(\bar{x}^k - \hat{x}^k) = \hat{b}^{k+1} - Aw^k$.

The exact solution of $\min J(x) + \frac{1}{2}\|Ax - \bar{b}^{k+1}\|^2$ is $\hat{x}^{k+1}$ by definition. Plugging $\bar{b}^{k+1} = \hat{b}^{k+1} - Aw^k$ into the objective function, we obtain $J(x) + \frac{1}{2}\|A(x + w^k) - \hat{b}^{k+1}\|^2$, which after variable transforming $y := x + w^k$, becomes $J(y - w^k) + \frac{1}{2}\|Ay - \hat{b}^{k+1}\|^2$, for which we have argued $\breve{x}$ is an exact minimizer. Hence, $\breve{x} - w^k$ exactly minimizes $J(x) + \frac{1}{2}\|A(x + w^k) - \hat{b}^{k+1}\|^2$, so $\hat{x}^{k+1} = \breve{x} - w^k$.

Finally, from the definition $w^{k+1} = \bar{x}^{k+1} - \hat{x}^{k+1}$, we get $\bar{x}^{k+1} = \hat{x}^{k+1} + w^{k+1} = \breve{x} + w^{k+1} - w^k$. □

Corollary 1 *Under Assumption 2 and the assumptions of Theorem 4, we have*

$$\mathrm{dist}(\bar{x}^{k+1}, X^*) \leq \|w^{k+1} - w^k\|_2, \ \forall k > K. \tag{16}$$

Proof For $k > K$ and under Assumption 2, Theorem 3 gives $\breve{x} \in X^*$, and Theorem 4 gives $\bar{x}^{k+1} - \breve{x} = w^{k+1} - w^k$. Hence, $\mathrm{dist}(\bar{x}^{k+1}, X^*) \leq \|\bar{x}^{k+1} - \breve{x}\|_2 = \|w^{k+1} - w^k\|_2$. □

Remark 4 Since $-A^\top(A\breve{x} - \hat{b}^{k+1}) \in J(\breve{x})$ by the definition of $\breve{x}$, the assumption $-A^\top(A\breve{x} - \hat{b}^{k+1}) \in J(\breve{x} - w^k)$ in Theorem 4 holds if $\breve{x}$ and $\breve{x} - w^k$ are both in the relative interior of the same face of the epigraph of J, or simply speaking, the error w^k is small enough so that $\breve{x} - w^k$ stays within the same face with $\breve{x}$. Consider $J(\cdot) = \|\cdot\|_1$ for example. If x and w satisfy $\mathrm{sign}(x) = \mathrm{sign}(x - w)$, then any $p \in J(x)$ also satisfies $p \in J(x - w)$.

Remark 5 Corollary 1 shows that the absolute error $\mathrm{dist}(\bar{x}^{k+1}, X^*)$ depends only on w^k and w^{k+1}, which are the numerical errors introduced at iterations k and $k+1$ and are independent of the early errors $w^{k-1}, w^{k-2}, \ldots, w^1$. That means that they do not accumulate to large absolute errors. This result holds no matter how each step (5a) is computed as long as the assumptions hold, and it explains why none of the curve in Figure 1 increases above 10^{-6} even though $\|w^k\| \approx 10^{-6}$ at each k. While it explains the solid curve, it does not explain the other four curves, which not only stay below 10^{-6} but also decrease geometrically to 10^{-15}. In general, the two vectors w^k and w^{k+1} can be anything of size 10^{-6} or less. So, $\|w^{k+1} - w^k\|$ is not necessarily small. The additional analysis in the next subsection shall explain why $\|w^{k+1} - w^k\|$ decrease geometrically corresponding to those four curves.

Remark 6 Although iterations (4) and (8) are equivalent on paper, the error-forgetting property holds *only* for iteration (4)'s inexact version (5). When the subproblems of (4) and (8) are solved inexactly, iterations (4) and (8) do not generate the same sequences. Iteration (8) is less stable since p^k generated by the *inexact version* of (8b) does not obey $p^k \in \partial J(x^k)$.

2.5 Error Cancellation

In this section, we focus on $J(\cdot) = \mu\|\cdot\|_1$ and explain that in the inexact iteration (5), when every subproblem (3) is solved by first-order (i.e., gradient based) iterations, the right-hand side of (16) — $\|w^{k+1} - w^k\|_2$ — can decrease geometrically in k. As one will soon see, this property requires "perfect synchronization" of the subproblem iterations at iterations k, $k+1$, $k+2$, …, where $k > K$. While it does happen, it is less likely to happen than error-forgetting. Instead of rigorously proving the property, we merely explain why it can happen.

By perfection synchronization, we refer to the phenomenon that when an iterative procedure consisted of linear and *piece-wise linear* operators starts from input $\bar{b}^{k+1}$ and then starts again from input $\bar{b}^k$, the input difference $\|\bar{b}^{k+1} - \bar{b}^k\|$ is so small that the two sequences of operations take place on the same linear piece at all iterations, i.e, $\bar{b}^{k+1}$ and $\bar{b}^k$ are processed by two identical sequences of linear operations. The first-order algorithms FPC, GPSR, and SpaRSA are based on the ISTA iteration

$$x_{(i+1)} \leftarrow \mathrm{shrink}(x_{(i)} - \tau A^\top(Ax_{(i)} - f), \mu\tau), \tag{17}$$

where the operator $\mathrm{shrink}(x, \alpha) := x - \mathrm{Proj}_{[-\alpha,\alpha]^n}(x)$ is piece-wise linear. While on the optimal face, as $\bar{x}^k$ gets close to $\breve{x}$, the difference between $\bar{b}^{k+1}$ and $\bar{b}^k$ is also very small, so small that perfect synchronization occurs and leading to good consequences as we shall continue explain below.

Now let us explain how $w^{k+1} - w^k$ geometrically decreases in k. According to the definitions of $\{w^k\}$, $\{\bar{x}^k\}$, and $\{\hat{x}^k\}$, they obey

$$w^{k+1} - w^k = (\bar{x}^{k+1} - \bar{x}^k) - (\hat{x}^{k+1} - \hat{x}^k). \tag{18}$$

We first argue $(\hat{x}^{k+1} - \hat{x}^k) = (w^{k-1} - w^k)$. We assume that

$$\eta := w^{k-1} - w^k \tag{19}$$

is sufficiently small so that $L(w) := J(x) - J(x+w)$ is well defined for points x and w specified below. From $\bar{b}^{k+1} - \bar{b}^k = b - A\bar{x}^k = A(\breve{x} - A\bar{x}^k) = A\eta$, we have $\bar{b}^{k+1} = \bar{b}^k + A\eta$. Hence, the problem

$$\min\{J(x) + \frac{1}{2}\|Ax - \bar{b}^{k+1}\|^2\}, \tag{20}$$

is equivalent to $\min\{J(x) + \frac{1}{2}\|A(x-\eta) - \bar{b}^k\|^2\}$, or with the change of variable $y := x - \eta$, is further equivalent to $\min\{J(y+\eta) + \frac{1}{2}\|A(y) - \bar{b}^k\|^2\}$. Since $J(y+\eta) = J(y) - L(\eta)$ near $y = \hat{x}^k$, and $\hat{x}^k$ is the minimizer of $J(y) + \frac{1}{2}\|A(y) - \bar{b}^k\|^2$, we can let $\hat{x}^k$ be the minimizer y^* and thus, the solution $\hat{x}^{k+1}$ to problem (20) obeys $\hat{x}^{k+1} = y^* + \eta = \hat{x}^k + \eta$, or

$$(\hat{x}^{k+1} - \hat{x}^k) = \eta. \tag{21}$$

Next, we study $(\bar{x}^{k+1} - \bar{x}^k)$ in (18). Since $\bar{x}^{k+1}$ and $\bar{x}^k$ are generated from the inputs $f := \bar{b}^{k+1}$ and $f := \bar{b}^k$, respectively, we shall first get

$$\bar{b}^{k+1} - \bar{b}^k = b - A\bar{x}^k = A(\breve{x} - \bar{x}^k) = A(w^{k-1} - w^k) = A\eta,$$

where the first equality follows from (5b), the second from Theorem 3, the third from Theorem 4, and the last from (19). Since η is assumed to be sufficiently small, we can assume the same for $(\bar{b}^{k+1} - \bar{b}^k)$.

Suppose from the inputs $f := \bar{b}^{k+1}$ and $f := \bar{b}^k$, (17) generates sequences $\{x^{k+1}_{(0)}, x^{k+1}_{(1)}, \ldots, x^{k+1}_{(i)}, \ldots\}$ and $\{x^k_{(0)}, x^k_{(1)}, \ldots, x^k_{(i)}, \ldots\}$, respectively, where $x^{k+1}_{(0)} = x^k_{(0)} = \mathbf{0}$. At iteration $i = 0$, the shrinkage operation is applied to $x^{k+1}_{(0)} - \tau A^\top(Ax^{k+1}_{(0)} - \bar{b}^{k+1})$ and $x^k_{(0)} - \tau A^\top(Ax^k_{(0)} - \bar{b}^k)$, respectively, which are very close since $\bar{b}^{k+1}$ and $\bar{b}^k$ are very close. Consequently, the same linear piece of the shrinkage operation is applied to both, yielding $x^{k+1}_{(1)}$ and $x^k_{(1)}$ that are again very close. This, in turn, lets the same linear piece of the shrinkage operator applied at iteration $i = 1$, and by induction, also at iterations $i = 2, 3, \ldots$ until the iterations are terminated, typically after the same number of iterations when the same stopping is used. Hence, $\{x^{k+1}_{(1)}, x^{k+1}_{(2)}, \ldots, x^{k+1}_{(i)}, \ldots\}$ and $\{x^k_{(1)}, x^k_{(2)}, \ldots, x^k_{(i)}, \ldots\}$ are generated by the same sequence of linear operations from slightly different inputs $f := \bar{b}^{k+1}$ and $f := \bar{b}^k$. The difference $(\bar{b}^{k+1} - \bar{b}^k) = A\eta$ causes the difference $(x^{k+1}_{(i)} - x^k_{(i)})$. Without going into further details, we claim that they give $\bar{x}^{k+1} := x^{k+1}_{(i_{\text{stop}})}$ and $\bar{x}^k := x^k_{(i_{\text{stop}})}$ obeying

$$\bar{x}^{k+1} - \bar{x}^k = (I + M)\eta, \tag{22}$$

where M depends on both the stopping iteration i_{stop} and matrix A, but not on η, $\bar{b}^{k+1}$, or $\bar{b}^k$. Furthermore, M obeys $\|M\| < 1$ provided that $\tau\|A^\top A\|_2 < 2$. Therefore, from (18), (21), and (21), we finally obtain

$$w^{k+1} - w^k = M\eta = M(w^k - w^{k-1}). \tag{23}$$

According to Corollary 1, $\|w^{k+1} - w^k\|$ upper bounds the absolute error of $\bar{x}^{k+1}$. From (23) and $\|M\| < 1$, it follows that the absolute error decays geometrically.

Since this analysis is not thorough, we prepared a Matlab demo for download from the first author's webpage. We suggest that the interested reader play with the demo and verify error cancellation, and we also hope that it will motivate the discoveries of more interesting properties of minimizing a piece-wise linear function.

3 Conclusions

When minimizing a piece-wise linear function subject to linear constraints, Bregman (or augmented Lagrangian) iteration is more tolerant to numerical errors due to inaccurately solved subproblems. While the standard practice requires diminishing errors over the steps, Bregman iteration only requires sufficiently small errors so that the optimal face can be identified. Once on the optimal face, errors from two or more steps back will be "forgotten." Furthermore, the errors of last two steps of subproblem solvers based on linear and piece-wise linear operators may nearly cancel each other, causing geometric convergence of the intermediate solutions to the exact solution. Hence, a highly accurate solution can be obtained by inaccurate Bregman steps.

Acknowledgements The authors thank Professor Yin Zhang and two anonymous reviewers for their valuable comments.

References

1. 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(2):460–489, 2005.
2. 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(1):143–168, 2008.
3. M. Figueiredo, R. Nowak, and S. J. Wright. Gradient projection for sparse reconstruction: application to compressed sensing and other inverse problems. *IEEE Journal of Selected Topics in Signal Processing: Special Issue on Convex Optimization Methods for Signal Processing*, 4(1):586–597, 2007.
4. E. T. Hale, W. Yin, and Y. Zhang. Fixed-point continuation for ℓ_1-minimization: methodology and convergence. *SIAM Journal on Optimization*, 19(3):1107–1130, 2008.
5. S.J. Wright, R.D. Nowak, and M.A.T. Figueiredo. Sparse reconstruction by separable approximation. *Signal Processing, IEEE Transactions on*, 57(7):2479–2493, 2009.
6. J. Xu and S. Osher. Iterative regularization and nonlinear inverse scale space applied to wavelet-based denoising. *IEEE Transactions on Image Processing*, 16(2):534–544, 2006.
7. M. Burger, G. Gilboa, S. Osher, and J. Xu. Nonlinear inverse scale space methods. *Communications in Mathematical Sciences*, 4(1):175–208, 2006.
8. 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*, pages 25–36, 2005.
9. 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.
10. S. Ma, D. Goldfarb, and L. Chen. Fixed point and bregman iterative methods for matrix rank minimization. *Mathematical Programming*, page 133, 2009.
11. Z. Guo, T. Wittman, and S. Osher. L1 unmixing and its application to hyperspectral image enhancement. *Proceedings of SPIE Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XV*, 7334:73341M, 2009.
12. 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*, 7:200–217, 1967.
13. J.-F. Cai, S. Osher, and Z. Shen. Linearized Bregman iterations for compressed sensing. *Mathematics of Computation*, 78(267):1515–1536, 2008.
14. J.-F. Cai, S. Osher, and Z. Shen. Convergence of the linearized Bregman iteration for ℓ_1-norm minimization. *Mathematics of Computation*, 78(268):2127–2136, 2009.
15. W. Yin. Analysis and generalizations of the linearized Bregman method. *SIAM Journal on Imaging Sciences*, 3(4):856–877, 2010.
16. M. Friedlander and P. Tseng. Exact regularization of convex programs. *SIAM Journal on Optimization*, 18(4), 2007.
17. M.-J. Lai and W. Yin. Augmented ℓ_1 and nuclear-norm models with a globally linearly convergent algorithm. *Rice University CAAM Technical Report TR12-02*, 2012.
18. J.-F. Cai, E. Candes, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2008.
19. M. R. Hestenes. Multiplier and gradient methods. *Journal of Optimization Theory and Applications*, 4:303–320, 1969.
20. M. J. D. Powell. A method for nonlinear constraints in minimization problems. In R. Fletcher, editor, *Optimization*, pages 283–298. Academic Press, New York, 1972.
21. Ernie Esser. Applications of Lagrangian-based alternating direction methods and connections to split Bregman. *UCLA CAM Report 09-31*, 2009.
22. D.P. Bertsekas. Combined primal-dual and penalty methods for constrained minimization. *SIAM Journal on Control*, 13(3):521–544, 1975.
23. D. P. Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods.* Athena Scientific, 1996.
24. Mosek ApS Inc. The Mosek optimization tools, ver 4., 2006.
25. D.P. Bertsekas. Necessary and sufficient conditions for a penalty method to be exact. *Mathematical Programming*, 9(1):87–99, 1975.
26. R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14:877–898, 1976.
27. J. Eckstein and P.J.S. Silva. A practical relative error criterion for augmented lagrangians. *Mathematical Programming*, pages 1–30, 2010.
28. M. Burger, M. Moller, M. Benning, and S. Osher. An adaptive inverse scale space method for compressed sensing. *UCLA CAM Report 11-08*, 2011.