SPARSE RECOVERY VIA DIFFERENTIAL INCLUSIONS

By Stanley Osher[*], Feng Ruan[†,‡], Jiechao Xiong[†], Yuan Yao[†] and Wotao Yin[*]

University of California, Los Angeles[*], Peking University[†], and Stanford University[‡]

In this paper, we recover sparse signals from their noisy linear measurements by solving nonlinear differential inclusions, which we call *Bregman ISS* and *Linearized Bregman ISS*. We show that under proper conditions, there exists a bias-free and sign-consistent point on their solution paths, which corresponds to a signal that is the unbiased estimate of the true signal and whose entries have the same signs as those of the true signs. Therefore, their solution paths are regularization paths better than the LASSO regularization path, since the points on the latter path are biased. We also show how to efficiently compute their solution paths in both continuous and discretized settings: the full solution paths can be exactly computed piece by piece, and a discretization leads to *Linearized Bregman iteration*, which is faster and easy to parallelize. Theoretical guarantees such as sign-consistency and minimax optimal l_2-error bounds are established in both continuous and discrete settings for specific points on the paths. Early-stopping rules for identifying these points are given. The key treatment relies on the development of differential inequalities for differential inclusions and their discretizations.

1. Introduction. We study two continuous time dynamics *Bregman ISS*[1] and *Linearized Bregman ISS*, as well as the forward-Euler discretization of the latter, for recovering a sparse unknown signal $\beta^* \in \mathbb{R}^p$ from its noisy linear measurements

$$y = X\beta^* + \epsilon. \tag{1.1}$$

Here, $y \in \mathbb{R}^n$ is a measurement vector, $X = [x_1, \ldots, x_p] \in \mathbb{R}^{n\times p}$ is a measurement matrix, and ϵ is unknown random noise. We allow $n < p$ and assume that β^* has $s \leq \min\{n, p\}$ nonzero components. For convenience, let $S = \mathrm{supp}(\beta^*)$ and T be its complement, i.e. $T = \{i : \beta_i^* = 0\}$.

The solution path $\{\rho_t, \beta_t\}_{t\geq 0}$ of Bregman ISS is given by the nonlinear

Keywords and phrases: Linearized Bregman, Differential Inclusion, Early Stopping Regularization, Statistical Consistency

[1]ISS abbreviates Inverse Scale Space, a name adopted from the imaging literature [BOXG05]. There, large-scale image features are recovered before small-scale ones.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
JUL 2014		**00-00-2014 to 00-00-2014**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Sparse Recovery via Differential Inclusions	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of California, Los Angeles,Department of Mathematics,Los Angeles,CA,90095**	8. PERFORMING ORGANIZATION REPORT NUMBER **CAM14-61**
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

In this paper, we recover sparse signals from their noisy linear measurements by solving nonlinear di erential inclusions, which we call Bregman ISS and Linearized Bregman ISS. We show that under proper conditions, there exists a bias-free and sign-consistent point on their solution paths, which corresponds to a signal that is the unbiased estimate of the true signal and whose entries have the same signs as those of the true signs. Therefore, their solution paths are regularization paths better than the LASSO regularization path, since the points on the latter path are biased. We also show how to e ciently compute their solution paths in both continuous and discretized settings the full solution paths can be exactly computed piece by piece and a discretization leads to Linearized Bregman iteration, which is faster and easy to parallelize. Theoretical guarantees such as signconsistency and minimax optimal l2-error bounds are established in both continuous and discrete settings for speci c points on the paths. Early-stopping rules for identifying these points are given. The key treatment relies on the development of di erential inequalities for di erential inclusions and their discretizations.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**39**	

differential inclusions:

$$\dot{\rho}_t = \frac{1}{n} X^T (y - X\beta_t), \tag{1.2a}$$

$$\rho_t \in \partial \|\beta_t\|_1, \tag{1.2b}$$

where $t \geq 0$ is time, $\rho_t \in \mathbb{R}^p$ is assumed to be right continuously differentiable in t, $\dot{\rho}_t$ is the right derivative of ρ_t, and β_t is assumed to be right continuous. The inclusion condition (1.2b) restricts ρ_t to a subgradient of ℓ_1-norm at β_t, $t \geq 0$. The initial conditions are, typically, $\rho_0 = 0$ and $\beta_0 = 0$. We will see that a solution to (1.2) exists and both ρ_t and $X\beta_t$, $t \geq 0$, are unique. In addition, ρ_t is piece-wise linear, and there exists a solution path β_t that is piece-wise constant. The entire path can be computed at finitely many break points.

Linearized Bregman ISS has its solution path $\{\rho_t, \beta_t\}_{t \geq 0}$ governed by the nonlinear differential inclusions:

$$\dot{\rho}_t + \frac{1}{\kappa}\dot{\beta}_t = \frac{1}{n} X^T (y - X\beta_t), \tag{1.3a}$$

$$\rho_t \in \partial \|\beta_t\|_1, \tag{1.3b}$$

where $\kappa > 0$ is a constant. Compared to (1.2a), equation (1.3a) has the additional term $\frac{1}{\kappa}\dot{\beta}$. As $\kappa \to \infty$, (1.3) is reduced to (1.2), and the solution path of (1.3) may converge to that of (1.2) exponentially fast as κ increases. We will see that (1.3) has a unique solution path ρ_t and β_t, $t \geq 0$, which are both continuous.

The discretizations of (1.2) and (1.3) are known as Bregman Iteration (equation (3.7) of [YODG08]) and Linearized Bregman Iteration (equations (5.19-20) of [YODG08]). They were introduced in the literature of variational imaging and compressive sensing before (1.2) and (1.3). Through a change of variable, Bregman Iteration becomes the iteration of the Augmented Lagrangian Method [Hes69, Pow67]. On the other hand, Linearized Bregman Iteration is a simple two-line iteration:

$$\rho_{k+1} + \frac{1}{\kappa}\beta_{k+1} = \rho_k + \frac{1}{\kappa}\beta_k + \frac{\alpha_k}{n} X^T (y - X\beta_k), \tag{1.4a}$$

$$\rho_k \in \partial \|\beta_k\|_1, \tag{1.4b}$$

which is evidently a forward Euler discretization to (1.3), where $\alpha_k > 0$ is a step size. Define $z_k = \rho_k + \frac{1}{\kappa}\beta_k$. Then (1.4) can be simplified to:

$$z_{k+1} = z_k + \frac{\alpha_k}{n} X^T (y - X\beta_k) \tag{1.5a}$$

$$\beta_{k+1} = \kappa \cdot \mathrm{shrink}(z_{k+1}, 1), \tag{1.5b}$$

where the mapping shrink is defined component-wise as

$$\text{shrink}(z,\lambda) := \text{sign}(z)\max\{|z|-\lambda,0\}, \quad z,\lambda\in\mathbb{R},\ \lambda\geq 0.$$

Note that $\text{shrink}(z,\lambda)$ is the unique solution to the convex program:

$$\min_{x\in\mathbb{R}} |x| + \frac{1}{2\lambda}(x-z)^2.$$

1.1. *Motivations and contributions.* Our exposition is motivated by the fact that solution path $\{\beta_t\}_{t\geq 0}$ of the differential inclusion (1.2) and the sequence $\{\beta_k\}_{k\geq 0}$ of (1.4) are *better* than the points on the LASSO regularization path. In particular, while LASSO regularization path is always biased, β_t can be unbiased when the correct set of variables is reached.

To see this, consider the LASSO problem [Tib96],

$$\min_{\beta} \lambda\|\beta\|_1 + \frac{1}{2n}\|y-X\beta\|_2^2, \tag{1.6}$$

where for the convenience of comparison we replace the regularization parameter λ by $t=1/\lambda$ in the following equivalent form

$$\min_{\beta} \|\beta\|_1 + \frac{t}{2n}\|y-X\beta\|_2^2. \tag{1.7}$$

Aside from the obvious relation $t=1/\lambda$, solution β is piece-wise linear in λ [EHJT04] though not so in t. Despite this, t will be convenient to our analysis by reflecting a nature of time evolution of the solution.

Since (1.7) is a convex program, $\hat{\beta}_t$ is a solution to (1.7) if and only if it obeys the first-order optimality conditions

$$\frac{\hat{\rho}_t}{t} = \frac{1}{n}X^T(y-X\hat{\beta}_t), \tag{1.8a}$$

$$\hat{\rho}_t \in \partial\|\hat{\beta}_t\|_1, \tag{1.8b}$$

which are obtained by taking the subdifferential of the objective in (1.7).

It is well-known that LASSO solution $\hat{\beta}_t$ is biased [FL01]. For example, considering the simple case that $n=p=1$, X is the identity and $y\geq 0$, then (1.8) yields

$$\hat{\beta}_t = \begin{cases} 0, & \text{if } t<1/y; \\ y-1/t, & \text{otherwise,} \end{cases} \tag{1.9}$$

while (1.2) has the solution

$$\beta_t = \begin{cases} 0, & \text{if } t<1/y; \\ y, & \text{otherwise,} \end{cases} \tag{1.10}$$

which is unbiased for $t \geq 1/y$ as $\mathbb{E}[\beta_t] = \mathbb{E}[y] = \beta^*$.

Moreover, the Linearized Bregman ISS (1.3) has the solution,

$$\beta_t = \begin{cases} 0, & \text{if } t < 1/y; \\ y(1 - e^{-\kappa(t-1/y)}), & \text{otherwise,} \end{cases} \tag{1.11}$$

which converges to the unbiased Bregman ISS estimator exponentially fast.

Let us discuss this phenomenon in the general setting. First, let the *oracle estimator* be the subset least-squares solution $\tilde{\beta}^*$ given the *true* set of variables S by an oracle, whose nonzero entries are given by

$$\tilde{\beta}^*_S = \left(\frac{1}{n} X_S^T X_S\right)^{-1} \frac{1}{n} X_S^T y = \beta^*_S + \left(\frac{1}{n} X_S^T X_S\right)^{-1} \frac{1}{n} X_S^T \epsilon. \tag{1.12}$$

Clearly $\tilde{\beta}^*_S \sim \mathcal{N}(\beta^*_S, \Sigma_n)$ where $\Sigma_n = \frac{\sigma^2}{n} \left(\frac{1}{n} X_S^T X_S\right)^{-1}$. Since in expectation with respect to noise, $\mathbb{E}[\tilde{\beta}^*] = \beta^*$, $\tilde{\beta}^*$ is an unbiased estimate of β^*.

In reality we are not given the support set S, so the following two properties are used to evaluate the performance of an estimator $\hat{\beta}$.

1. **Model selection consistency**: $\mathrm{supp}(\hat{\beta}) = S$;
2. **Asymptotic normality**: $\sqrt{n}(\hat{\beta} - \beta^*) \to \mathcal{N}(0, \Sigma^*)$, where

$$\Sigma^* = \lim_{n\to\infty} n\Sigma_n = \sigma^2 \left(\lim_{n\to\infty} \frac{1}{n} X_S^T X_S\right)^{-1}.$$

Since these properties hold for the oracle estimator, they are often referred to as the *oracle properties*.

A solution mapping $\hat{\beta}_t : [0, \infty) \to \mathbb{R}^p$ gives a *regularization path*. Model selection consistency, also known as path consistency, refers to the existence of a point $\hat{\beta}_\tau$ on this path that selects the correct variables, namely, $\mathrm{supp}(\hat{\beta}_\tau) = S$. Path consistency has been obtained for LASSO by establishing the stronger property of *sign consistency*, that is, $\mathrm{sign}(\hat{\beta}_\tau) = \mathrm{sign}(\beta^*)$, under certain conditions such as those in [ZY06, Zou06, YL07, Wai09]. Provided that path consistency is reached at τ, the LASSO estimate $\hat{\beta}_\tau$ is nonetheless biased since

$$\hat{\beta}_{\tau,S} = \left(\frac{1}{n} X_S^T X_S\right)^{-1} \frac{1}{n} X_S^T y - \left(\frac{1}{n} X_S^T X_S\right)^{-1} \frac{\hat{\rho}_\tau}{\tau}, \tag{1.13}$$

where $\rho_\tau = \mathrm{sign}(\hat{\beta}_\tau) \in \partial \|\hat{\beta}_\tau\|_1$. The first-term on the right-hand side equals the oracle estimator $\tilde{\beta}^*_S$, which is unbiased, whereas the second-term never vanishes and is the bias. Hence, the oracle properties are never completely met by LASSO.

The bias can be removed by a simple differentiation of LASSO solution. To see this, by multiplying t on both sides of (1.8a) and differentiating it with respect to t, any point on the LASSO path satisfies

$$\dot{\hat{\rho}}_t = \frac{1}{n} X^T (y - X(\hat{\beta}_t + t\dot{\hat{\beta}}_t)). \tag{1.14}$$

With path consistency assumed at time $t = \tau$, we have $\beta_{\tau,i} = 0,\ \forall\ i \notin S$, and from (1.14) we have

$$\dot{\hat{\rho}}_{\tau,S} = \frac{1}{n} X_S^T (y - X_S(\hat{\beta}_{\tau,S} + \tau\dot{\hat{\beta}}_{\tau,S})). \tag{1.15}$$

Generically, sign consistency occurs in a neighborhood and thus $\dot{\rho}_{\tau,S} = 0$. Therefore,

$$\hat{\beta}_{\tau,S} + \tau\dot{\hat{\beta}}_{\tau,S} = \left(\frac{1}{n} X_S^T X_S\right)^{-1} \frac{1}{n} X_S^T y = \tilde{\beta}_S^*,$$

which is the oracle estimator without bias! This motivates us to replace $(\hat{\beta}_t + t\dot{\hat{\beta}}_t)$ in (1.14) by just β_t, which gives the differential inclusions (1.2a) of Bregman ISS. Later we will show that the resulting β_t in (1.2) is indeed unbiased.

Therefore, in addition to giving the basic solution properties such as existence, uniqueness, and (dis)continuity, we also attempt to explain the good behaviors of the new solution paths and sequence by establishing their *path consistency* property. Basically we argue that

1. Under nearly the same conditions for LASSO [ZY06, Zou06, YL07, Wai09] that the covariates x_i are sufficiently uncorrelated and the signal β_S^* is strong enough, Bregman ISS (1.2) with a proper early stopping rule will return the *oracle estimator*;
2. Sign consistency and l_2-error bounds of minimax rates can be generalized to the Linearized Bregman iteration (1.4) and its limit dynamics (1.3), under similar conditions.

Some computational aspects are reviewed in the next subsection.

1.2. *Related work.*

1.2.1. *Regularization and other algorithms.* For general penalized least square problems, [FL01] has shown that no convex penalty functions can fully achieve the *oracle properties* and thus one has to resort to non-convex regularization, whose global minimizer is, however, algorithmically difficult to locate. Alternatively, one can apply LASSO for variable selection and then

remove the bias in LASSO by solving a subset least squares in the second stage. On the other hand, [OBG+05] noticed that Bregman iteration may reduce bias, also known as contrast loss, in the context of Total Variation image denoising. In this paper, we shall see that dynamics (1.2) can automatically remove bias without any non-convexity or second-stage subset least squares. It is a different kind of regularization via early stopping.

Early stopping regularization has been studied widely in linear inverse problems, e.g. [EHN96], and recently in Boosting, e.g. [Fri01, BY02, YRC07]. In fact, Linearized Bregman iterations can be viewed as an extension of Landweber iteration (also called L_2-Boost in statistics),

$$\beta_{k+1} = \beta_k + \frac{\alpha_k}{n} X^T (y - X\beta_k),$$

which follows the primal path β_t as a gradient descent method solving least square problem. To have solution sparsity, Linearized Bregman iterations (1.4) adds the dual path ρ_t in favor of sparse solutions.

Linearized Bregman iteration (1.5) is shown in [Yin10] equivalent to the gradient ascent iteration applied to the Lagrange dual of the problem

$$\min_{\beta} \|\beta\|_1 + \frac{1}{2\kappa}\|\beta\|_2^2 \quad \text{subject to } X\beta = y. \tag{1.16}$$

In particular, β_k converges to the unique solution of (1.16) at a linear rate (as long as $X \neq 0$ and $X\beta = y$ has a solution); see [LY13]. In addition, for sufficiently large κ, the solution to (1.16) is a solution to the basis pursuit model [CDS98], which is (1.16) without $\frac{1}{2\kappa}\|\beta\|_2^2$. In noisy settings, early stopping regularization is necessary for signal recovery. The results in this paper basically say that under nearly the same condition as LASSO, Bregman ISS with early stopping regularization may recover the signal without bias. We note that such dynamics can be easily extended to general settings with differentiable convex loss and non-differentiable convex penalty, e.g. Linearized Bregman iteration in matrix completion [CCS10].

One should not confuse Linearized Bregman iteration (1.5) with iterative soft-thresholding algorithm (ISTA), which has been widely used under different names in the literature (for example, see [DJ95, Don95, CDLL98, DD02, DDD04]),

$$\beta_{k+1} = \text{shrink}(\beta_k + \frac{\alpha_k}{n} X^T (y - X\beta_k), \lambda_k).$$

By moving the shrinkage operator to a different place in (1.5), Linearized Bregman iteration generates a sparse solution path with early stopping regularization, while ISTA exploits λ_k as the regularization parameter and its iterates converge to a LASSO solution.

1.2.2. *Parallel and distributed computing.* It is very easy to implement iteration (1.5) in parallel and distributed manners and apply it to very large-scale datasets. Suppose

$$X = [X_1,\ X_2,\ \ldots,\ X_L] \in \mathbb{R}^{n\times p},$$

where X_ℓ's are submatrices stored in a distributed manner (on a set of networked workstations). The sizes of X_ℓ's are flexible and can be chosen for good load balancing. Let each workstation ℓ hold data y and X_ℓ, and variables $z_{k,\ell}$ and $w_{k,\ell} := X_\ell\beta_{k,\ell}$, which are parts of z_k and summands of $w_k := X\beta_k$, respectively. The iteration (1.5) is carried out as

$$\text{for } \ell = 1,\ldots,L \text{ in parallel: } \begin{cases} z_{k+1,\ell} = z_{k,\ell} + \frac{\alpha_k}{n} X_\ell^T(y - w_k), \\ w_{k+1,\ell} = \frac{1}{\kappa} X_\ell \text{shrink}(z_{k+1,\ell}, 1), \end{cases}$$

$$\text{all-reduce summation: } w_{k+1} = \sum_{\ell=1}^{L} w_{k+1,\ell},$$

where the all-reduce step collects inputs from and then returns the sum to all the L workstations. It is the sum of L n-dimensional vectors, so no matter how the all-reduce step is implemented, the communication cost is independent of p. It is important to note that the algorithm is not changed at all. In particular, distributing the data into more computing units, i.e., increasing L, does *not* increase the number of iterations. Therefore, the parallel implementation is nearly embarrassingly parallel and truly scalable. In addition, it is also possible to develop implementations for data divided into blocks of rows of X or even smaller subblocks that split both rows and columns. Recently, (1.5) has also been extended in [YLYR13] to a *decentralized* setting where not only data and computation are distributed but communication is restricted to computing units with *direct* communication links so there is no data fusion center or long distance communication. The scheme fits sensor network or multi-party regression over the internet, where long-distance communication incurs long delays and high costs.

1.3. *Notation and assumptions.* We introduce the following notation and assumptions to β^*, X, and ϵ.

- Let the true support be denoted by $S = \text{supp}(\beta^*) = \{i : \beta_i^* \neq 0\}$, and $T = S^c$ be its complement. Clearly, $S \cup T = \{1,\ldots,p\}$.
- X_S denotes the submatrix of X formed by the columns of X in S, which are assumed to be *linearly independent*. Similarly define X_T so that $[X_S\ X_T] = X$.

- Assume $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$. It can generalize to sub-Gaussian without violating most of our results.

Define $\langle u, v\rangle = u^T v$ and $\langle u, v\rangle_n = \frac{1}{n} u^T v$ for $u, v \in \mathbb{R}^n$. Hence $\|u\|_n = \frac{1}{\sqrt{n}}\|u\|$. Let $X^* = \frac{1}{n} X^T$ be the adjoint operator of X with respect to inner product $\langle\cdot,\cdot\rangle_n$. Let the largest and the smallest nonzero magnitudes of β^* be $\beta^*_{\max} := \max(|\beta^*_i| : i \in S)$ and $\beta^*_{\min} := \min(|\beta^*_i| : i \in S)$, respectively. Similarly define $\tilde{\beta}^*_{\max}$ and $\tilde{\beta}^*_{\min}$ for the oracle estimator $\tilde{\beta}^*$ in (1.12). The dependence of ρ_t and β_t (or equivalently $\rho(t)$ and $\beta(t)$) on t is omitted where it is clear from the context. For the reason to be discussed in Section 2, we shall assume that ρ_t is right continuously differentiable and β_t is right continuous.

Throughout the paper, given two numbers a and b, let $a \vee b := \max(a, b)$.

1.4. *Outline.* In the rest of this paper, we establish basic solution properties of Bregman and Linearized Bregman ISS in Section 2. Section 3 and Section 4 describe statistical consistency properties of Bregman ISS and their generalizations to Linearized Bregman ISS/discretization, respectively. Section 5 is dedicated to the ideas of proofs. Section 6 provides some preliminary numerical results. Discussions and conclusions are summarized in Section 7.

2. Bregman and Linearized Bregman solution paths. The solution to Bregman ISS (1.2) is a piece-wise regularization path given iteratively as follows, starting with $k = 0$, $t_0 = 0$, and $\rho_0 = \beta_0 = 0$:

1. set $t_{k+1} := \sup\{t > t_k : \rho_{t_k} + \frac{t-t_k}{n} X^T(y - X\beta_{t_k}) \in \partial\|\beta_{t_k}\|_1\}$; if $t_{k+1} = \infty$, then *exit*;
2. set $\rho_{t_{k+1}} := \rho_{t_k} + \frac{t_{k+1}-t_k}{n} X^T(y - X\beta_{t_k})$;
3. set $S_{k+1} := \{i : |(\rho_{t_{k+1}})_i| = 1\}$ and $T_{k+1} = \{1, \dots, p\} \setminus S_{k+1}$;
4. set $\beta_{t_{k+1}}$ as any solution to

$$\begin{aligned} &\min_\beta && \|y - X\beta\|_2^2 && \\ &\text{subject to} && (\rho_{t_{k+1}})_i \beta_i \geq 0 && \forall\ i \in S_{k+1}, \\ & && \beta_j = 0 && \forall\ j \in T_{k+1}. \end{aligned} \tag{2.1}$$

5. set $k = k + 1$ and go to Step 1.

Paper [BMBO13] gives an algorithm but does not establish the uniqueness of solution path. One can show that the solution to (1.2) is piece-wise:

$$\begin{cases} \rho_t = \rho_{t_k} + \frac{t-t_k}{t_{k+1}-t_k}\rho_{t_{k+1}}, & \\ \beta_t = \beta_{t_k}, & \end{cases} \qquad t \in [t_k, t_{k+1}). \tag{2.2}$$

In other words, ρ_t is piece-wise linear and β_t is piece-wise constant. The following theorem presents some general conditions to ensure the existence and uniqueness of solution path.

THEOREM 2.1 (Solution existence and uniqueness for Bregman ISS). *Let ρ_t be right continuously differentiable and β_t be right continuous. Then, a solution to (1.2) is given by (β_t, ρ_t) generated by the above algorithm. Solution ρ_t and $X\beta_t$ are unique. In addition, if the columns x_i of X for $i \in \operatorname{supp}(\beta_t)$ are linearly independent for $t \geq 0$, then β_t is also unique.*

PROOF. The existence part follows from [BMBO13].

We show that the uniqueness part. Define $f(\beta) := \frac{1}{2n}\|y - X\beta\|^2$. Then, the differential inclusion (1.2) is equivalent to

$$\dot{\rho}_t = -\nabla f(\beta_t), \tag{2.3a}$$

$$\rho_t \in \partial\|\beta_t\|_1, \tag{2.3b}$$

Let $S_t^+ := \{i : (\rho_t)_i = 1\}$, $S_t^- := \{i : (\rho_t)_i = -1\}$, and $S_t = S_t^+ \cup S_t^-$. By (1.2b), in the case of $S_t = \emptyset$, we have $\beta_t = 0$, so $-\nabla f(\beta_t) = -\nabla f(0)$ is unique. In the case of $S_t \neq \emptyset$, we show below that $X\beta_t$ and $-\nabla f(\beta_t)$ are both unique. The uniqueness of ρ_t follows from these results and (2.3a).

In fact, (1.2a) and (1.2b) impose the following constraints on β_t:

$$\begin{cases} (\beta_t)_i \geq 0 \text{ and } (\nabla f(\beta_t))_i \geq 0, & \forall i \in S_t^+, \\ (\beta_t)_i \leq 0 \text{ and } (\nabla f(\beta_t))_i \leq 0, & \forall i \in S_t^-, \\ (\beta_t)_i = 0, & \forall i \notin S_t. \end{cases} \tag{2.4}$$

To see how $\nabla f(\beta_t)$ is involved, notice that $(\nabla f(\beta_t))_i \geq 0$ must hold for $\forall i \in S_t^+$ since $(\rho_t)_i \in [-1, 1]$ is already at its maximal value 1 and $\nabla f(\beta_t) < 0$ is forbidden as it would further increase $(\rho_t)_i$ to an impossible value. The same argument holds for $(\nabla f(\beta_t))_i \leq 0$ for $\forall i \in S_t^-$.

Furthermore, we will have $(\beta_t)_i \cdot (\nabla f(\beta_t))_i = 0$ for all i. To see this, assume $(\beta_t)_i \neq 0$. Then by the right continuity assumption, there exists an interval $[t, t + \epsilon)$ in which β_i remains nonzero with the same sign. By (2.3b), $(\rho_t)_i$ will remain either $+1$ or -1 in the same interval, so $(\nabla f(\beta_t))_i = 0$. On the other hand, assume $(\nabla f(\beta_t))_i \neq 0$. Then by (2.3a), ρ_i will change and thus it cannot stay either $+1$ or -1. By the right continuity of β, it must hold that $(\beta_t)_i = 0$. Therefore, we have the addition constraints

$$(\beta_t)_i \cdot (\nabla f(\beta_t))_i = 0. \tag{2.5}$$

Conditions (2.4) and (2.5) are precisely the KKT optimality conditions for

$$\min_{\beta} \ f(\beta)$$
$$\text{subject to } \begin{cases} \beta_i \geq 0, & \forall i \in S_t^+, \\ \beta_i \leq 0, & \forall i \in S_t^-, \\ \beta_i = 0, & \forall i \notin S_t, \end{cases} \tag{2.6}$$

which is identify to (2.1) except (2.1) specifies the time t_{k+1}. Let β_t be the solution to problem (2.6).

In general, if f is strictly convex, then the solution β_t is unique. In our case, f is not necessarily strictly convex, but $f = g(X\beta)$ for a strictly convex function g. Therefore, $X\beta_t$ is unique, and thus so is $\nabla f(\beta_t) = X^T \nabla g(X\beta_t)$. Lastly, β_t is unique if the columns of X corresponding to nonzero entries of β_t are linearly independent since $X\beta_t$ is unique. □

The existence and uniqueness of Linearized Bregman ISS is much simpler as shown in the following theorem.

THEOREM 2.2 (Solution existence and uniqueness for Linearized Bregman ISS). *Let ρ_t be right continuously differentiable and β_t be right continuous. Then (1.3) has a unique solution.*

PROOF. Let $z_t = f(\rho_t, \beta_t) = \rho_t + \frac{1}{\kappa}\beta_t$, then f is an injective function from the admissible set (ρ, β) to C^1 in variable t and $\beta_t = \kappa \text{shrink}(z_t, 1)$. Now differential inclusion (1.3) becomes the ODE

$$\dot{z}_t = \frac{1}{n} X^T (y - \kappa X \cdot \text{shrink}(z_t, 1)) =: g(z_t)$$

Obviously, $g(x)$ is Lipschitz continuous. Therefore, the Picard-Lindelöf Theorem implies that there exists a unique solution to this ODE, which leads to the solution of (1.3). □

We note that the solution of (1.3), though not piece-wise linear or constant, can still be computed in a piece-wise closed form where on each piece, the signs of β_t remain unchanged. This is left to the reader.

Provided that sign consistency is met by a point on the path at $t = \tau$, (2.1) returns an oracle solution as it is a least-squares problem subject to only sign constraints. Hence, natural questions are: *what conditions will guarantee sign consistency? And, how to determine τ*? In the sequel, we are going to provide an answer to this question. Throughout the remaining of this paper, we assume that ρ_t is right continuously differentiable and β_t is right continuous, so the existence and uniqueness of solution paths are guaranteed.

3. Consistency of Bregman ISS Dynamics. In this section necessary and sufficient conditions are established for noisy sparse signal recovery with Bregman ISS (1.2).

3.1. *Assumptions.*

(A1) **Restricted Strong Convexity**: there is a $\gamma \in (0, 1]$,

$$X_S^* X_S \geq \gamma I.$$

(A2) **Irrepresentable Condition**: there is a $\eta \in (0, 1)$,

$$\left\| X_T^* X_S^\dagger \right\|_\infty = \left\| \frac{1}{n} X_T^T X_S \left(\frac{1}{n} X_S^T X_S \right)^{-1} \right\|_\infty \leq 1 - \eta$$

where $X_S^\dagger := X_S \left(\frac{1}{n} X_S^T X_S \right)^{-1}$.

Condition A1 says that the Hessian matrix of the empirical risk $\frac{1}{2n}\|y - X\beta\|_2^2$ restricted on the index set $S \times S$ is strictly positive definitive, so the empirical risk is strongly convex when restricted on the support set S. Such a condition is necessary in the sense that once it fails, X_S will be linearly dependent and no unique representation is possible under the basis X_S.

Condition A2 says that the absolute row sums of matrix $X_T^* X_S^\dagger$ are all less than one. It has been proposed independently under a variety of names, e.g. Exact Recovery Condition [Tro04], Irrepresentable Condition [ZY06], among [YL07, Zou06]. Here we adopt the name in [ZY06] as it refers to the fact that the regression coefficients of X_S for response X_j ($j \in T$) all have ℓ_1-norm less than one, i.e.

$$\beta_j' = \arg\min_{\beta \in \mathbb{R}^s} \frac{1}{2n} \|X_j - X_S \beta\|^2 \implies \|\beta_j'\|_1 < 1,$$

so in this sense one cannot represent the irrelevant covariates X_T by the relevant ones X_S effectively.

Neither A1 nor A2 can be checked when the support set S of signal is not known. Alternatively we can use a more strict but checkable condition proposed in [DH01].

(A3) **Mutual Incoherence Condition**:

$$\mu := \max_{i,j} \left| \frac{1}{n} \langle X_i, X_j \rangle \right| < \frac{1}{(2s - 1)}, \quad s = |S|.$$

It can be shown [Tro04, CW11] that once A3 holds, then A1 and A2 simultaneously hold with

$$\gamma = 1 - \mu(s-1)$$

since $(1-\mu(s-1))I_S \leq X_S^* X_S \leq (1+\mu(s-1))I_S$, and

$$\eta = \frac{1-\mu(2s-1)}{1-\mu(s-1)}.$$

We note that condition A3 is shown to be sharp in the noisy case in [CWX10]. With these one can translate all the theoretical results with condition A1 and A2 into condition A3.

3.2. *Mean Bregman ISS Path versus LASSO Path.* As we have seen in Section 1.1 near equation (1.14), Bregman ISS (1.2) can be derived by differentiating LASSO's KKT conditions. Such a relation can be seen precisely by considering the consistency conditions of LASSO on the following temporal *mean path* of Bregman ISS:

$$\bar{\beta}(t) := \frac{1}{t}\int_0^t \beta(s)ds. \tag{3.1}$$

According to Theorem 2.1 and Condition A1, Bregman ISS path β_t is unique and thus $\bar{\beta}(t)$ is well defined as long as $\mathrm{supp}(\beta(s)) \subseteq S$, $s \in [0,t)$, where S is the true support.

A connection between Bregman ISS and LASSO lies in the same condition under which their paths from start to time t are supported within the true support S. In addition, the Bregman ISS mean path $\bar{\beta}(t)$ is identical to the LASSO path if the Bregman ISS path is incremental with only adding variables, but without dropping. In general, the two paths are distinct.

THEOREM 3.1. *Let (β_t, ρ_t) be either the Bregman ISS path (1.2) or the LASSO path (1.8) with $\rho(t) \in \partial\|\beta_t\|_1$. Assume that for all $t \leq \tau$,*

$$\|X_T^* X_S^\dagger \rho_S(t) + t X_T^* P_T \epsilon\|_\infty < 1, \tag{3.2}$$

where $P_{S^\perp} = I - P_S = I - X_S^\dagger X_S^*$ is the projection matrix onto $\mathrm{im}(X_S)^\perp$. Then for all $t \leq \tau$,

A. the Bregman ISS path, its mean path, and the LASSO path all have supports in S;
B. the mean Bregman ISS path $\bar{\beta}(1/\lambda)$ is piecewise linear with $\lambda = 1/t$;

C. *if the Bregman ISS path is incremental in the sense that* $S_t = \operatorname{supp}(\beta_t)$ *satisfies* $S_t \subseteq S_{t'} \subseteq S$ *for all* $t \leq t' \leq \tau$, *then the mean Bregman ISS path is identical to the LASSO path; but they are distinct in general.*

REMARK 3.1. *In particular in noiseless setting,* $\epsilon = 0$, *(3.2) becomes*

$$\|X_T^* X_S^\dagger \rho_S(t)\|_\infty < 1$$

or dropping $\rho_S(t)$ *by*

$$\|X_T^* X_S^\dagger\|_\infty = \|X_T^* X_S (X_S^* X_S)^{-1}\|_\infty < 1$$

which is sufficient and necessary to guarantee that both Bregman ISS, LASSO, and OMP [Tro04] recovers the sparse signal in noiseless setting; once it is violated there is some S*-sparse signal for which these methods fail.*

PROOF OF THEOREM 3.1. Assume there exists a $\tau \geq 0$, such that for all $t \leq \tau$, solution path $\beta(t)$ satisfies $\operatorname{supp}(\beta(t)) \subseteq S$. Then Bregman ISS (1.2) splits into

$$\dot{\rho}_S = -X_S^* X_S(\beta_S - \beta_S^*) + X_S^* \epsilon, \tag{3.3a}$$

$$\dot{\rho}_T = -X_T^* X_S(\beta_S - \beta_S^*) + X_T^* \epsilon. \tag{3.3b}$$

From (3.3a) one gets the Bregman ISS solution

$$\beta_S(t) = \beta_S^* - (X_S^* X_S)^{-1} \dot{\rho}_S + (X_S^* X_S)^{-1} X_S^* \epsilon, \tag{3.4}$$

which leads to the following equation by plugging into (3.3b)

$$\dot{\rho}_T = X_T^* X_S^\dagger \dot{\rho}_S + X_T^* P_T \epsilon. \tag{3.5}$$

Integration on both sides of this equation and setting

$$\|\rho_T(t)\|_\infty = \|X_T^* X_S^\dagger \rho_S(t) + t X_T^* P_T \epsilon\|_\infty < 1 \tag{3.6}$$

which ensures that $\beta_T(t) = 0$. So is the mean path.

On the other hand, LASSO starts from the KKT condition (1.8) which splits into

$$\hat{\rho}_S/t = -X_S^* X_S(\hat{\beta}_S - \beta_S^*) + X_S^* \epsilon \tag{3.7a}$$

$$\hat{\rho}_T/t = -X_T^* X_S(\hat{\beta}_S - \beta_S^*) + X_T^* \epsilon \tag{3.7b}$$

Following the same trick above one can see the same condition (3.6) is met for LASSO to ensure $\hat{\beta}_T(t) = 0$. This finishes the proof of part A.

As to part B, for $t \leq \tau$, the mean path is obtained by integration on (3.3a)

$$\bar{\beta}_S(t) = \frac{1}{t}\int_0^t \beta_S(s)ds = \beta^*_S - \frac{1}{t}(X^*_S X_S)^{-1}\rho_S(t) + (X^*_S X_S)^{-1}X^*_S\epsilon. \tag{3.8}$$

Equation (2.2) implies that $\frac{1}{t}\rho_t = \frac{1}{t}\rho_{t_k} + \frac{1-t_k/t}{t_{k+1}-t_k}\rho_{t_{k+1}}$, which is piecewise linear with respect to $\lambda = 1/t$.

To see part C, let $S_t = \mathrm{supp}(\beta_t)$ for Bregman ISS. If for all $s \leq t \leq \tau$, $S_s \subseteq S_t \subseteq S$, then similar reasoning as above implies that the Bregman ISS path satisfies

$$\bar{\beta}_{S_t}(t) = \beta^*_{S_t} - \frac{1}{t}(X^*_{S_t} X_{S_t})^{-1}\rho_{S_t}(t) + (X^*_{S_t} X_{S_t})^{-1}X^*_{S_t}\epsilon. \tag{3.9}$$

For such incremental processes, $\rho_{S_t}(t) = \mathrm{sign}(\beta_{S_t}(t)) = \mathrm{sign}(\bar{\beta}_{S_t}(t))$ which meets the LASSO path equation

$$\hat{\beta}_{\hat{S}_t}(t) = \beta^*_{\hat{S}_t} - \frac{1}{t}(X^*_{\hat{S}_t} X_{\hat{S}_t})^{-1}\hat{\rho}_{\hat{S}_t}(t) + (X^*_{\hat{S}_t} X_{\hat{S}_t})^{-1}X^*_{\hat{S}_t}\epsilon, \tag{3.10}$$

where $\hat{S}_t = \mathrm{supp}(\hat{\beta}_t)$ for LASSO. But such a relation is lost when variable dropping happens. □

Despite of the difference to the LASSO path, the mean Bregman ISS path may reach statistical model-selection consistency under the same conditions as LASSO.

THEOREM 3.2 (Sign Consistency of Mean Path). *Let*

$$\overline{\tau} := \frac{\eta}{2\sigma}\sqrt{\frac{n}{\log p}}\left(\max_{j\in T}\|X_j\|_n\right)^{-1}.$$

Assume that both (A.1) and (A.2) hold. Then the following holds.

A. ***(No-false-positive)*** *the mean path has no-false-positive before time $\overline{\tau}$, i.e., $\forall t \leq \overline{\tau}$ $\mathrm{supp}(\bar{\beta}_t) \subseteq S$, with probability at least $1 - \frac{2}{p\sqrt{\pi \log p}}$;*

B. ***(No-false-negative for Mean Path)*** *moreover if the signal is strong enough such that $\beta^*_{\min} > c_1/\overline{\tau}$,*

$$c_1 = \left(\frac{\eta}{\sqrt{\gamma}\max_{j\in T}\|X_j\|_n} + \|(X^*_S X_S)^{-1}\|_\infty\right),$$

then with probability at least $1 - \frac{2}{p\sqrt{\pi \log p}}$, the mean path $\bar{\beta}_{\overline{\tau}}$ has no false-negative, i.e. $\mathrm{sign}(\bar{\beta}_{\overline{\tau}}) = \mathrm{sign}(\beta^*)$.

REMARK 3.2. *Under the same conditions as LASSO with* $\lambda^* = 1/\bar{\tau}$ *[Wai09], the mean path* $\bar{\beta}$ *of Bregman ISS reaches sign-consistency. These conditions are sufficient and necessary in the sense that once violated, there exists an instance such that the probability of failure will be larger than* $1/2$ *due to noise. In this sense, the mean path estimator* $\bar{\beta}(\bar{\tau})$ *is "statistically equivalent" to the LASSO estimator.*

The mean Bregman ISS path geometrically sheds light on why LASSO incurs bias while Bregman ISS can avoid it. The LASSO path, likes the mean Bregman ISS path, involves some kind of averaging that ensures the path continuity but causes bias. The Bregman ISS path is piecewise constant, allows it to be bias-free.

Now we need to answer the following question: *what are conditions to ensure the sign consistency of the Bregman ISS path*?

3.3. *Consistency of Bregman ISS.* The following theorem tells us that under the irrepresentable (incoherence) condition, the Bregman ISS dynamics always evolves in the support of true signals in the early stage; furthermore if the signal is strong enough then the dynamics will pick up all the true variables before selecting any incorrect ones. When such a sign consistency is reached, Bregman ISS returns the oracle estimator which is unbiased.

THEOREM 3.3 (Sign Consistency of Bregman ISS). *Let*

$$\bar{\tau} := \frac{\eta}{2\sigma}\sqrt{\frac{n}{\log p}}\left(\max_{j\in T}\|X_j\|_n\right)^{-1}.$$

Assume that both (A.1) and (A.2) hold. Then Bregman ISS (1.2) has paths satisfying:

A. ***(No-false-positive)*** *the path has no-false-positive before time* $\bar{\tau}$*, i.e.* $\forall t \le \bar{\tau}\ \mathrm{supp}(\beta_t) \subseteq S$*, with probability at least* $1 - \frac{2}{p\sqrt{\pi \log p}}$*;*

B. ***(Sign-consistency)*** *moreover if the signal is strong enough such that*

$$\beta^*_{\min} \ge \left(\frac{4\sigma}{\gamma^{1/2}} \vee \frac{8\sigma(2+\log s)\left(\max_{j\in T}\|X_j\|_n\right)}{\gamma\eta}\right)\sqrt{\frac{\log p}{n}} \tag{3.11}$$

Then with probability at least $1 - \frac{2}{p\sqrt{\pi \log p}}$*,* $\mathrm{sign}(\beta_{\bar{\tau}}) = \mathrm{sign}(\beta^*)$*.*

REMARK 3.3. *Once the sign consistency holds,* $\beta(t)$ *meets the oracle estimator* $\tilde{\beta}^*$ *which is unbiased and has a* l_2*-error rate* $\|\beta(t) - \beta^*\|_2 \le O(\sigma\sqrt{s\log s/n})$*, even better than the* l_2*-error rate* $O(\sigma\sqrt{s\log p/n})$ *for the LASSO estimator which is* ***minimax optimal*** *up to a logarithmic factor [RWY11].*

To have sign consistency, Theorem 3.3 makes a *strong signal* condition with a lower bound on $\beta^*_{\min}$. However even without such a strong signal assumption, the minimax optimal l_2-error rates can be achieved disregarding sign consistency.

THEOREM 3.4 (Minimax Optimal l_2-Error Bound). *Assume that both (A1) and (A2) hold. There is a $\tau \in [0, \overline{\tau}]$ such that with probability at least $1 - \frac{2}{p\sqrt{\pi \log p}}$,*

$$\|\beta_\tau - \beta^*\|_2 \leq \frac{2\sigma}{\eta\gamma}\left(4 \max_{j\in T} \|X_j\|_n + \eta\sqrt{\gamma}\right)\sqrt{\frac{s \log p}{n}}.$$

The existence of such τ does not ensure us to find it easily. However one can use $\bar{\tau}$ at a cost of enlarging the constants by a square root of condition number of $\Sigma_S = X_S^* X_S$.

COROLLARY 3.1. *Under the same condition of Theorem 3.4 and assuming an upper eigenvalue bound $X_S^* X_S \leq \gamma_{\max} I_S$, then the following holds for all $t \in [\tau, \bar{\tau}]$ with probability at least $1 - \frac{2}{p\sqrt{\pi \log p}}$*

$$\|\beta_t - \beta^*\|_2 \leq \frac{2\sigma\sqrt{\mathcal{K}(X_S^* X_S)}}{\eta\gamma}\left(4 \max_{j\in T} \|X_j\|_n + \eta\sqrt{\gamma}\right)\sqrt{\frac{s \log p}{n}}$$

where $\mathcal{K}(X_S^* X_S) = \gamma_{\max}/\gamma$ is the condition number of $X_S^* X_S$.

All the results in this subsection follow from the more general results on Linearized Bregman ISS (1.3) in the next section by taking $\kappa \to 0$. Therefore we omit the proofs.

4. Generalizations to Linearized Bregman ISS and Its Discretization. In this section, we state a general consistency result for Linearized Bregman ISS (1.3) and Linearized Bregman Iterations (1.4) whose proof will be given in the next section.

4.1. *Consistency of Linearized Bregman ISS.* The following theorem establishes general conditions for statistical consistency of Linearized Bregman ISS (LBISS) (1.3).

THEOREM 4.1 (Consistency of LBISS). *Let*

$$\overline{\tau} := \frac{(1 - B/(\kappa\eta))\eta}{2\sigma}\sqrt{\frac{n}{\log p}}\left(\max_{j\in T} \|X_j\|_n\right)^{-1}.$$

Assume κ is big enough to satisfy

$$\beta^*_{\max} + 2\sigma\sqrt{\frac{\log p}{\gamma n}} + \frac{\|X\beta^*\|_2 + 2\sigma\sqrt{s\log n}}{n\sqrt{\gamma}} \triangleq B \le \kappa\eta.$$

Then (1.3) has paths satisfying

A. ***(No-false-positive)*** *the path has no-false-positive before time $\overline{\tau}$,i.e., $\forall t \le \overline{\tau}$ $\mathrm{supp}(\beta_t) \subseteq S$, with probability at least $1 - \frac{2}{p\sqrt{\pi\log p}} - \frac{1}{n\sqrt{\pi\log n}}$;*

B. ***(No-false-negative for Mean Path)*** *moreover if the signal is strong enough such that $\beta^*_{\min} > c_1/\bar{\tau}$,*

$$c_1 = \left(\frac{(1 - B/(\kappa\eta))\eta}{\sqrt{\gamma}\max_{j\in T}\|X_j\|_n} + (1 + B/\kappa\eta)\|(X_S^*X_S)^{-1}\|_\infty\right),$$

then with probability at least $1 - \frac{2}{p\sqrt{\pi\log p}} - \frac{1}{n\sqrt{\pi\log n}}$, the mean path $\bar{\beta}(t)$ satisfies $\mathrm{sign}(\bar{\beta}_{\bar{\tau}}) = \mathrm{sign}(\beta^*)$;

C. ***(Sign-consistency for LBISS)*** *Moreover if the smallest magnitude $\beta^*_{\min}$ is strong enough and κ big enough such that*

$$\beta^*_{\min} \ge \frac{4\sigma}{\gamma^{1/2}}\sqrt{\frac{\log p}{n}},$$

$$\frac{8 + 4\log s}{\beta^*_{\min}} + \frac{1}{\kappa}\log(\frac{3\|\beta^*\|_2}{\beta^*_{\min}}) \le \overline{\tau},$$

then with probability at least $1 - \frac{2}{p\sqrt{\pi\log p}} - \frac{1}{n\sqrt{\pi\log n}}$, $\mathrm{sign}(\beta_{\bar{\tau}}) = \mathrm{sign}(\beta^*)$;

D. ***(l_2-bound)*** *For some constant C and κ large enough to satisfy*

$$\frac{4}{C\gamma}\sqrt{\frac{n}{\log p}} + \frac{1}{2\kappa\gamma}(1 + \log\frac{n\|\beta^*\|_2^2 + 4\sigma^2 s\log p/\gamma}{C^2 s\log p}) \le \overline{\tau},$$

there is a $\tau \in [0, \overline{\tau}]$ such that $\|\beta_\tau - \beta^*\|_2 \le (C + \frac{2\sigma}{\gamma^{1/2}})\sqrt{\frac{s\log p}{n}}$ with probability at least $1 - \frac{2}{p\sqrt{\pi\log p}} - \frac{1}{n\sqrt{\pi\log n}}$.

REMARK 4.1. *A. For sign-consistency of LBISS,*

$$\beta^*_{\min} \ge \left(\frac{4\sigma}{\gamma^{1/2}} \vee \frac{8\sigma(2 + \log s)\,(\max_{j\in T}\|X_j\|)}{\gamma\eta}\right)\sqrt{\frac{\log p}{n}}$$

is enough to guarantee the existence of κ.

B. For l_2-consistency

$$C \geq \frac{8\sigma\left(\max_{j\in T}\|X_j\|_n\right)}{\eta\gamma}$$

is enough to guarantee the existence of κ.

C. Taking $\kappa = \infty$, we get the Theorem 3.3 for Bregman ISS.

D. An l_2-error bound of the same rate for estimator $\beta(\bar{\tau})$ can be established using the monotonicity of $\|X_S(\tilde{\beta}_S^* - \beta_S(t))\|_2$ (see Appendix) for $t \leq \bar{\tau}$,

$$\begin{aligned}\|\beta(\bar{\tau}) - \tilde{\beta}\|_2 &\leq \frac{\|X_S(\beta_S(\bar{\tau}) - \tilde{\beta}_S^*)\|_2}{\sqrt{n\gamma}} \leq \frac{\|X_S(\beta_S(\tau) - \tilde{\beta}_S^*)\|_2}{\sqrt{n\gamma}}, \quad \tau \leq \bar{\tau} \\ &\leq \sqrt{\mathcal{K}(X_S^* X_S)}\left(C + \frac{2\sigma}{\sqrt{\gamma}}\right)\sqrt{\frac{s\log p}{n}},\end{aligned}$$

where $\mathcal{K}(X_S^* X_S)$ is the condition number of $X_S^* X_S$.

4.2. *Consistency of Linearized Bregman iterations.* The following theorem establishes statistical consistency conditions for Linearized Bregman Iteration (1.4).

THEOREM 4.2 (Consistency of Linearized Bregman Iterations). *Let $t_n = \sum_{k=0}^{n-1}\alpha_k$ and*

$$\bar{\tau} := \frac{(1 - B/(\kappa\eta))\eta}{2\sigma}\sqrt{\frac{n}{\log p}}\left(\max_{j\in T}\|X_j\|_n\right)^{-1}.$$

Assume that κ is big enough to satisfy

$$\beta_{\max}^* + 2\sigma\sqrt{\frac{\log p}{\gamma n}} + \frac{\|X\beta^*\|_2 + 2\sqrt{s\log n}}{n\sqrt{\gamma}} \triangleq B \leq \kappa\eta,$$

and step size α is small such that $\kappa\alpha\|X_S X_S^*\| < 2$. Then any solution path of (1.3) satisfies

*A. **(No-false-positive)** for all n s.t. $t_n \leq \bar{\tau}$, the path has no-false-positive with probability at least $1 - \frac{2}{p\sqrt{\pi\log p}} - \frac{1}{n\sqrt{\pi\log n}}$, $\operatorname{supp}(\beta_k) \subseteq S$;*

*B. **(Sign-consistency)** moreover if the smallest magnitude $\beta_{\min}^*$ is strong enough and κ is big enough to ensure*

$$\beta_{\min}^* \geq \frac{4\sigma}{\gamma^{1/2}}\sqrt{\frac{\log p}{n}},$$

$$\frac{8+4\log s}{\tilde{\gamma}\beta^*_{\min}}+\frac{1}{\kappa\tilde{\gamma}}\log(\frac{3\|\beta^*\|_2}{\tilde{\beta}_{\min}})+3\alpha\le\bar{\tau},$$

where $\tilde{\gamma}=\gamma(1-\kappa\alpha\|X_SX_S^*\|/2)$, *then with probability at least* $1-\frac{2}{p\sqrt{\pi\log p}}-\frac{1}{n\sqrt{\pi\log n}}$, $\operatorname{sign}(\beta_{k^*})=\operatorname{sign}(\beta^*)$ *for* $k^*=\max\{k:t_k\le\bar{\tau}\}$.

C. ***(l_2-bound)*** *for some large enough constants* κ *and* C *such that*

$$\frac{4}{C\tilde{\gamma}}\sqrt{\frac{n}{\log p}}+\frac{1}{2\kappa\tilde{\gamma}}(1+\log\frac{n\|\beta^*\|_2^2+4\sigma^2 s\log p/\gamma}{C^2 s\log p})+2\alpha\le\bar{\tau},$$

with probability at least $1-\frac{2}{p\sqrt{\pi\log p}}-\frac{1}{n\sqrt{\pi\log n}}$, *there is a* k^*, $t_{k^*}\le\bar{\tau}$, *such that* $\|\beta_{k^*}-\beta^*\|_2\le(C+\frac{2\sigma}{\gamma^{1/2}})\sqrt{\frac{s\log p}{n}}$.

REMARK 4.2. A. *Taking* $\alpha\to 0$, *we have* $\tilde{\gamma}=\gamma$, *and Theorem 4.1 for Linearized Bregman ISS follows.*

B. *The condition* $\kappa\alpha\|X_SX_S^*\|<2$ *is necessary to ensure the convergence of LB algorithm in the noiseless case. This condition also guarantees the monotonic descent of* $\|X_S(\beta_{S,k}-\tilde{\beta}_S)\|$ *before* $\bar{\tau}$.

5. Analysis of ISS Dynamics. The general idea to analyze differential inclusions in (1.2) and (1.3) is to associate these dynamics with some potential or Lyapunov functions, which control a fast convergence of solutions to the oracle estimator. The restricted strongly convex condition A1 suggests us that when the solution path $\beta(t)$ evolves in the support set S, a suitable choice of potential functions should be expected with exponentially fast decay, which enables us to estimate the stopping time of reaching sign consistency and small l_2-error.

The difficulty lies in that ISS dynamics are differential inclusions, hence we exploit *differential inequalities* of such a potential function to derive the bounds.

5.1. *Potential function.* One would like to study the dynamics of the following differential inclusion

$$\dot{\rho}_t+\frac{1}{\kappa}\dot{\beta}_t=-X^*X(\beta_t-\tilde{\beta}^*) \tag{5.1a}$$

$$\rho_t\in\partial\|\beta_t\|_1, \tag{5.1b}$$

where $\tilde{\beta}^*$ is the oracle estimator. Assuming the right continuity of solutions and multiplying both sides above by $\beta(t)-\tilde{\beta}^*$, one obtains a *potential* or *Lyapunov* function $\Psi:\mathbb{R}^p\to\mathbb{R}_0^+$ associated with the dynamics

$$\frac{d}{dt}(\Psi(\beta_t))=-\frac{1}{n}\|X(\beta_t-\tilde{\beta}^*)\|_2^2,$$

where

$$\Psi(\beta) = D(\tilde{\beta}, \beta) + \frac{\|\beta - \tilde{\beta}\|^2}{2\kappa} \tag{5.2}$$

and $D(\tilde{\beta}, \beta)$ is the Bregman distance

$$D_V(\tilde{\beta}, \beta) := V(\tilde{\beta}) - V(\beta) - \langle \partial V(\beta), \tilde{\beta} - \beta \rangle \tag{5.3}$$

induced by the particular convex function $V(\beta) = \|\beta\|_1$. As $n \ll p$, matrix X has a large null-space, and to ensure the stationary point of the dynamics being the oracle solution, one must restrict the dynamics evolving outside the subspace $\ker(X)$.

5.2. *Differential inequality with restricted exponential decay of potential.* Define the following *Oracle Dynamics* as if an oracle discloses the true variable set S such that we restrict our attention on a subspace defined by S,

$$\dot{\rho}'_S + \frac{1}{\kappa}\dot{\beta}'_S = -X_S^* X_S(\beta'_S - \tilde{\beta}^*_S), \quad \rho'_S(t) \in \partial\|\beta'_S(t)\|_1. \tag{5.4}$$

Here $X_S^* X_S$ is a $s \times s$ symmetric matrix satisfying the strong convexity $X_S^* X_S \geq \gamma I_s$, which will lead to exponentially fast decay of potential function.

To reach this goal, our key treatment here is a differential inequality associated differential inclusion in Oracle Dynamics which is tight enough to ensure the exponential decay of potential function. This is a Bihari's type [Bih56] nonlinear differential inequality, which generalizes the linear cases of Grönwall-Bellman inequalities [Gro19, Bel43]. In our treatment, a piecewise continuous bound is given which leads to the tight rates in this paper.

LEMMA 5.1 (Generalized Bihari's Inequality). *The potential Ψ of the Oracle Dynamics above satisfies the following differential inequality*

$$\frac{d}{dt}(\Psi(\beta'_S)) \leq -\gamma F^{-1}(\Psi(\beta'_S)),$$

where F^{-1} is the right-continuous inverse of the following strictly increasing function

$$F(x) = \frac{x}{2\kappa} + \begin{cases} 0 & 0 \leq x < \tilde{\beta}^2_{min} \\ 2x/\tilde{\beta}_{min} & \tilde{\beta}^2_{min} \leq x \leq s\tilde{\beta}^2_{min} \\ 2\sqrt{xs} & x \geq s\tilde{\beta}^2_{min}. \end{cases} \tag{5.5}$$

Such an inequality ensures a decrease of the potential function at a fast enough speed which leads to the following tight estimates on stopping time.

We are concerned with the following stopping time reaching sign-consistency and l_2-consistency of Oracle Dynamics, respectively. Define

$$\tilde{\tau}_1 := \inf\{t > 0 : \operatorname{sign}(\beta'_S) = \operatorname{sign}(\tilde{\beta}^*_S)\}, \tag{5.6}$$

$$\tilde{\tau}_2(C) := \inf\left\{t > 0 : ||\beta'_S - \tilde{\beta}^*_S||_2 \le C\sqrt{\frac{s\log p}{n}}\right\}. \tag{5.7}$$

Equipped with the generalized Bihari's inequality, one can build up the following bounds for stopping time on sign-consistency and l_2-consistency, respectively.

LEMMA 5.2. *The following bounds hold for the Oracle Dynamics (5.4)*

$$\tilde{\tau}_1 \le \frac{4 + 2\log s}{\gamma\tilde{\beta}^*_{min}} + \frac{1}{\kappa\gamma}\log(\frac{\|\tilde{\beta}^*\|_2}{\tilde{\beta}^*_{min}}),$$

$$\tilde{\tau}_2(C) \le \frac{4}{C\gamma}\sqrt{\frac{n}{\log p}} + \frac{1}{2\kappa\gamma}(1 + \log\frac{n\|\tilde{\beta}^*\|_2^2}{C^2 s\log p}).$$

REMARK 5.1. *A. $\tilde{\tau}_1 \le O(\log s/\beta^*_{\min})$ says that $\beta(t)$ will reach sign-consistency after $t \ge O(\log s/\tilde{\beta}_{\min})$. The factor $\log s$ is due to the potential method above which converts a multidimensional dynamics into a one-dimensional differential inequality, and dropping potential exponentially from at least $\|\tilde{\beta}^*_S\|_1 \ge s\tilde{\beta}^*_{\min}$ to 0 requires necessarily the $O(\log s)$ time.*

B. $\tilde{\tau}_2(C) \le O(\frac{1}{C}\sqrt{\frac{n}{p}})$ says that l_2-consistency can be reached before $\overline{\tau} = O(\sqrt{\frac{n}{p}})$ as long as C is a sufficiently large constant.

5.3. *Sign-consistency and l_2-error bound.* Now we are ready to reach the sign-consistency and l_2-error bound for $\beta(t)$ by setting $\tilde{\tau}_1 \le \bar{\tau}$ and $\tilde{\tau}_2(C) \le \bar{\tau}$, respectively. In these cases, Oracle Dyanmics (5.4) $\beta'_S(t)$ meets the original path $\beta_S(t)$ when restricted on S. The complete proofs of Theorem 4.1 and its discrete version of Theorem 4.2 will be found in Appendix A, together with their supporting lemmas.

6. Data-dependent Stopping Rules for Bregman ISS. All the previous results enable us to select $\bar{\tau}$ as a stopping time which however depends on unknown parameters γ, η, and noise level σ, hence is not a data-dependent stopping rule. In this section we present two preliminary results with early stopping rules comparable to [CW11], which only depend on the noise level σ and thus can be estimated from data. We leave it our future work to explore fully adaptive stopping rules.

In the following, define the residue $r(t) := y - X\beta(t)$. The first theorem adopts the stopping rule based on $\|r(t)\|_2$ and the second theorem is based on $\|Xr(t)\|_\infty$.

THEOREM 6.1. *Suppose*

$$\beta^*_{\min} \geq \left(\frac{4\sigma}{\gamma^{1/2}} \vee \frac{8\sigma(2+\log s)\left(\max_{j\in T}\|X_j\|_n\right)}{\gamma\eta} \right) \sqrt{\frac{\log p}{n}},$$

and

$$\beta^*_{\min} \geq \frac{2\sigma}{\sqrt{\gamma}} \left(\sqrt{1+2\sqrt{\frac{\log n}{n}}} + \sqrt{\frac{\log s}{n}} \right).$$

Then Bregman ISS with the stopping rule $\|r(t)\|_2 \leq \sigma\sqrt{n+2\sqrt{n\log n}}$ selects the true subset S with probability at least $1 - O(1/n)$.

REMARK 6.1.

- *This result is comparable to Theorem 7 in [CW11].*
- *The first condition on the minimum of magnitude of signals ensures the model selection consistency of the Bregman ISS path and thus indicates that one can find some t along the path so that the residual term satisfies $\|r(t)\|_2 \leq \sigma\sqrt{n+2\sqrt{n\log n}}$. Once the path achieves sign consistency, the Bregman ISS must stop.*
- *The second condition $\beta^*_{\min} \geq \frac{2\sigma}{\sqrt{\gamma}} \left(\sqrt{1+2\sqrt{\frac{\log n}{n}}} + \sqrt{\frac{\log s}{n}} \right)$ guarantees that one can not stop earlier before Bregman ISS achieves a full recovery. Note that as $n \to \infty$, one needs $\beta^*_{\min} \geq 2\sigma/\sqrt{\gamma}$ which is a constant.*

THEOREM 6.2. *In addition to (3.11), suppose*

$$\beta^*_{\min} \geq \frac{2\sigma \max_i \|X_i\|_n \sqrt{2(1+c)s\log p}}{\sqrt{n}\gamma} + 2\sigma\sqrt{\frac{\log s}{n\gamma}}.$$

Then Bregman ISS with the stopping rule $\|X^T r(t)\|_\infty \leq 2\sigma\sqrt{\max_i \|X_i\| \log p}$ ($\delta > 0$) selects the true subset S with probability at least $1 - O(1/p + 1/n)$.

REMARK 6.2. *This result is comparable to Theorem 8 in [CW11], though the lower bound* $\beta^*_{\min} \geq O(\sigma\sqrt{s\log p/n})$ *loses a factor* $\sqrt{s}$ *here. As* $n \to \infty$, *the lower bound can be arbitrarily small.*

The remaining of this section presents the proofs of the theorems above.

PROOF OF THEOREM 6.1. Lemma 3 in [CW11] or Lemma 5.2 in [CXZ09] shows that with probability at least $1-1/n$, ϵ is essentially l_2 upper bounded

$$\|\epsilon\|_2 \leq \sigma\sqrt{n + 2\sqrt{n\log n}}.$$

Hence with the same probability,

$$\|r(\tau^*)\| = \|(I - X_S(X_S^*X_S)^{-1}X_S)\epsilon\|_2 \leq \|\epsilon\|_2 \leq \sigma\sqrt{n + 2\sqrt{n\log n}}$$

We have now shown that the Bregman ISS stops once the path acheives sign consistency.

Next we are going to show that the algorithm will not stop whenever there is some $i \in S$ such that $\beta_i(t) = 0$. By Lemma A.5,

$$\begin{aligned} \|r_t\| &\geq \|X_S(\tilde{\beta}^*_S - \beta_S(t))\| \\ &\geq \sqrt{n\gamma}\|\tilde{\beta}^*_S - \beta_S(t)\| \\ &\geq \sqrt{n\gamma}\tilde{\beta}^*_{\min} \\ &\geq 2\sigma\sqrt{n + 2\sqrt{n\log n}} \end{aligned}$$

provided that $\tilde{\beta}^*_{\min} \geq 2\sigma\sqrt{\frac{1+2\sqrt{\log n/n}}{\gamma}}$. Note that

$$\|(X_S^*X_S)^{-1}X_S^*\varepsilon\|_\infty \leq 2\sigma\sqrt{\frac{\log s}{n\gamma}}, \quad \text{w. p. at least } 1 - 2n^{-1},$$

so it suffices to have $\beta^*_{\min} \geq \frac{2\sigma(\sqrt{n+2\sqrt{n\log n}}+\sqrt{\log s})}{\sqrt{n\gamma}}$. □

PROOF OF THEOREM 6.2. By assumptions

$$\beta^*_{\min} \geq \left(\frac{4\sigma}{\gamma^{1/2}} \vee \frac{8\sigma(2+\log s)\,(\max_{j\in T}\|X_j\|_n)}{\gamma\eta}\right)\sqrt{\frac{\log p}{n}}.$$

Hence, according to Theorem 4.2, the Bregman ISS achieves the sign consistency with high probability. Assume that at time τ^*, $\beta(\tau^*)$ has the same sign as the underlying sparse signal β. For each t,

$$r_t = (I - X_{S(t)}(X^*_{S(t)}X_{S(t)})^{-1}X^*_{S(t)})(X_S\beta_S + \epsilon) = s_t + n_t,$$

where $s_t = (I - P_{S(t)})X_S\beta_S$ is the signal part of the residual and $n_t = (I - P_{S(t)})\epsilon$ is the noise part of the residual. Then $r_{\tau^*} = n_{\tau^*}$. Let $b_\infty = \sigma\sqrt{2(1+c)\max_i \|X_i\| \log p}$.

$$\begin{aligned}\text{Prob}(\|X^T n_t\|_\infty = \|X^T(I-P_t)\epsilon\|_\infty \geq b_\infty) &\leq \sum_i \text{Prob}(|X_i^T(I-P_t)\epsilon| \geq b_\infty) \\ &\leq \sum_i \text{Prob}(|X_i^T\epsilon| \geq b_\infty) \\ &\leq \frac{2}{p^c\sqrt{2\log p}},\end{aligned}$$

which means the algorithm stops at τ^*.

Next we are going to show that the algorithm will not stop whenever there is some $i \in A_t \subseteq S$ such that $\beta_i(t) = 0$. By Lemma A.5,

$$\begin{aligned}\|X^T r_t\|_\infty &= \|X^T[X_S(\tilde{\beta}_S^* - \beta_S(t)) + (I-P_S)\epsilon]\|_\infty, \\ &\geq \|X_S^T[X_S(\tilde{\beta}_S^* - \beta_S(t)) + (I-P_S)\epsilon]\|_\infty, \\ &= \|X_S^T X_S(\tilde{\beta}_S^* - \beta_S(t))\|_\infty, \quad X_S^T(I-P_S)\epsilon = 0, \\ &\geq \frac{1}{\sqrt{s}}\|X_S^T X_S(\tilde{\beta}_S^* - \beta_S(t))\|_2, \\ &\geq \frac{n\gamma}{\sqrt{s}}\|\tilde{\beta}_S^* - \beta_S(t)\|_2, \\ &\geq \frac{n\gamma}{\sqrt{s}}\tilde{\beta}_{\min}^* \geq b_\infty,\end{aligned}$$

provided that $\tilde{\beta}_{\min}^* \geq \frac{\sqrt{s}b_\infty}{n\gamma}$. Note that

$$\|(X_S^* X_S)^{-1}X_S^*\varepsilon\|_\infty \leq 2\sigma\sqrt{\frac{\log s}{n\gamma}}, \quad \text{w. p. at least } 1 - 2n^{-1},$$

so it suffices to have

$$\beta_{\min}^* \geq b_\infty + 2\sigma\sqrt{\frac{\log s}{n\gamma}} = \frac{\sigma(\max_i \frac{\|X_i\|}{\sqrt{n}}\sqrt{2(1+c)s\log p/\gamma} + \sqrt{\log s})}{\sqrt{n\gamma}}$$

with probability at least $1 - O(p^{-1} + n^{-1})$. □

7. Experiments. In this section we provide some experimental results to illustrate the relations among LASSO, Bregman ISS (ISS) and Linearized Bregman iteration (LB).

In this experiment we choose $n = 200$, $p = 100$ and only the first $s = 30$ elements of β are nonzero ($\beta_j = r_j + \text{sign}(r_j)$, where $r_j \sim \mathcal{N}(0,1)$, $j = 1, \ldots, 30$). Each sample x_i is drawn from the distribution $\mathcal{N}(0, \Sigma_p)$. We choose $\Sigma_p = (\sigma_{ij})$, where $\sigma_{ij} = 1$ if $i = j$, and $\sigma_{ij} = 1/(3p)$ otherwise. In such a setting, the Irrepresentable (Incoherence) Condition holds with high probability, since Σ_p is nearly identity matrix. We choose noise level $\sigma = 1$ here, considering the choice that the magnitude of β_i is $O(1)$.

Figure 1 is an example of regularization path of three methods. As κ goes bigger, the LB path becomes closer to that of ISS. For LB we choose $\kappa\alpha = 1/2$ such that the step size of gradient decent is $1/2$, to satisfy the convergence condition. Note that if α is too big, the solution is oscillating.

To compare the performance of three methods quantitatively, we choose the AUC of ROC curve, to measure the goodness of three regularization paths. ROC (receiver-operating-characteristic) curve is plotted by thresholding the regularization parameter λ in LASSO, t in ISS, or k in LB at different levels which create different true positive rates (TPR) and false positive rates (FPR):

$$TPR = \frac{\#\{Selected\ True\ Variables\}}{\#\{True\ Variables\}}, FPR = \frac{\#\{Selected\ False\ Variables\}}{\#\{False\ Variables\}}.$$

ROC is a curve from $(0,0)$ to $(1,1)$. AUC (Area Under the Curve) means the area under the ROC curve. Large AUC values indicate that the signals are picked out earlier than noise on regularization paths. Repeating the experiments for 100 times, in Table 1 we report the mean AUC with standard deviations for the three methods at different noise levels. It shows that all the three methods work reasonably well in this example, while Bregman ISS performs slightly better than LASSO. As κ becomes bigger, the performance of LB gets closer to that of Bregman ISS. Notice that as noise level σ gets larger, all the methods have their performance decay since signal and noise get confused.

σ	LB($\kappa = 4$)	LB($\kappa = 64$)	LB($\kappa = 1024$)	ISS	LASSO
1	0.9771(0.0124)	0.994(0.0069)	0.9947(0.0065)	0.9948(0.0064)	0.9945(0.0068)
3	0.9604(0.0169)	0.9867(0.009)	0.9882(0.0083)	0.9884(0.0082)	0.9879(0.0086)
5	0.9393(0.0226)	0.9659(0.0188)	0.9673(0.0188)	0.9676(0.0187)	0.9671(0.0187)

TABLE 1

Mean AUC (standard deviation) for three methods at different noise levels (σ): ISS has a slightly better performance than LASSO in terms of AUC and as κ increases, the performance of LB approaches that of ISS. As noise level σ increases, the performance of all the methods drops.

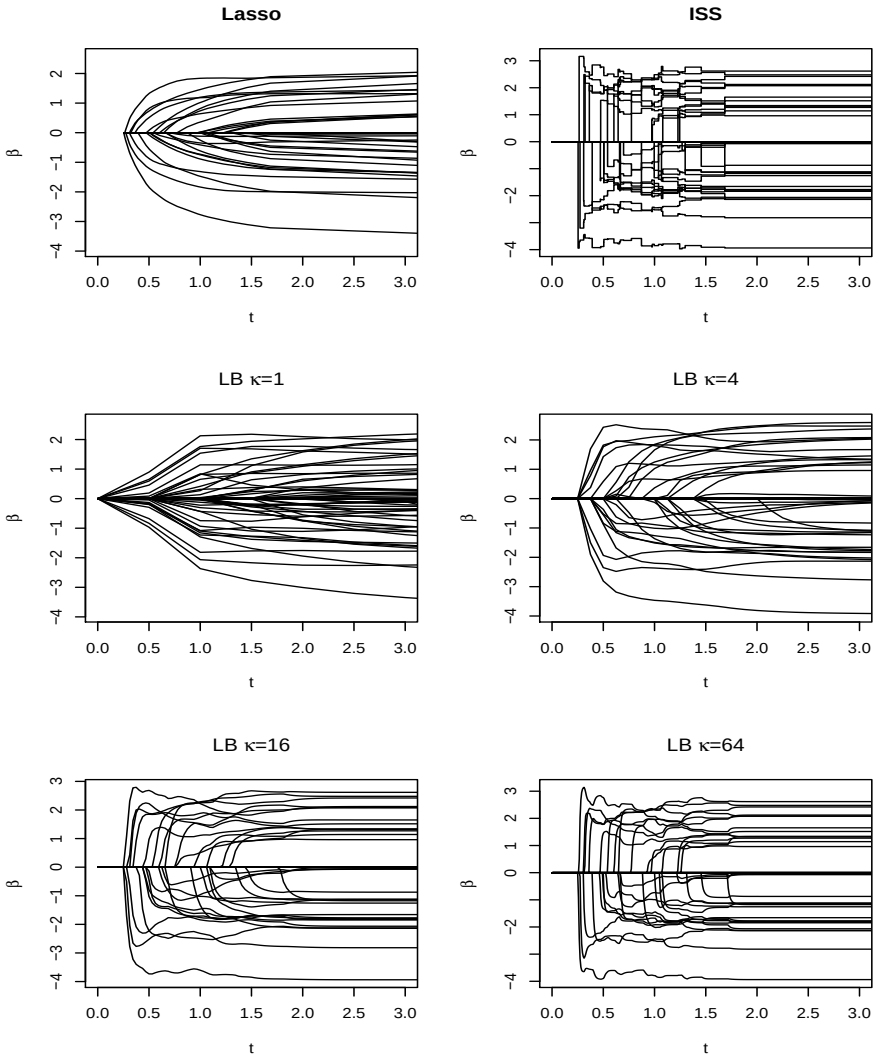

FIG 1. *Regularization path of LASSO, Bregman ISS, and Linearized Bregman Iterations with different choices of* κ ($\kappa\alpha = 1/2$). *As* κ *grows, the paths of Linearized Bregman iterations approach that of Bregman ISS.*

8. Discussion and Conclusion. In this paper, noisy sparse signal recovery is approached via two continuous dynamics, called Bregman ISS and Linearized Bregman ISS, where a discretization of the latter leads to the widely used Linearized Bregman Iteration algorithm. Equipped with an early stopping regularization, Bregman ISS can simultaneously achieve

model selection consistency and unbiased estimation. As a discretization of Linearized Bregman ISS paths, model selection consistency and minimax optimal l_2-error bounds for Linearized Bregman Iteration are also established. Some data-dependent stopping rules are given for Bregman ISS solution paths.

Future directions of our study include fully data-dependent stopping rules and generalization of our results in nonlinear settings.

APPENDIX A: PROOFS

A.1. Proof of Consistency of LBISS.

LEMMA A.1. *Assume that X_S has full column rank.*

(A) For all $t \le \tau$, solution of (1.3) $\beta(t)$ contains no false positive if

$$\|X_T^* X_S^\dagger(\rho_S + \beta_S/\kappa) + tX_T^* P_T \epsilon\|_\infty < 1, \quad \forall t \le \tau,$$

where $P_T = I - X_S^\dagger X_S^*$ is the projection operator onto the column space of X_T.

(B) Mean path $\bar{\beta}(\tau)$ is sign-consistent if

$$\operatorname{sign}(\bar{\beta}_S(\tau)) = \operatorname{sign}(\beta_S^* + \Phi_S^{-1} X_S^* \epsilon - \frac{1}{\tau}\Phi_S^{-1}(\rho_S(\tau) + \frac{1}{\kappa}\beta_S(\tau)) = \operatorname{sign}(\beta_S^*)$$

where $\Phi_S = X_S^* X_S = \frac{1}{n} X_S^T X_S$.

No-false-positivity and the sign-consistency for mean path in Theorem 4.1, directly follow this lemma.

PROOF OF LEMMA A.1. Consider the differential inclusion (1.3)

$$\dot{\rho} + \frac{1}{\kappa}\dot{\beta} = -\frac{1}{n} X^T(X\beta - y) = -X^* X(\beta - \beta^*) + X^*\epsilon$$

Assume there exists a $\tau \ge 0$, such that for all $t \le \tau$, solution path $\beta(t)$ contains no false-positive, i.e. $\operatorname{supp}(\beta(t)) \subseteq S$. Then for all $t \le \tau$,

$$\dot{\rho}_S + \dot{\beta}_S/\kappa = -X_S^* X_S(\beta_S - \beta_S^*) + X_S^*\epsilon, \tag{A.1}$$

and

$$\dot{\rho}_T + \dot{\beta}_T/\kappa = -X_T^* X_S(\beta_S - \beta_S^*) + X_T^*\epsilon. \tag{A.2}$$

From (A.1) one gets $-(\beta_S - \beta_S^*) = (X_S^* X_S)^{-1}(\dot{\rho}_S + \dot{\beta}_S/\kappa) - (X_S^* X_S)^{-1} X_S^*\epsilon$, which leads to the following equation by plugging into (A.2)

$$\dot{\rho}_T + \dot{\beta}_T/\kappa = X_T^* X_S^\dagger(\dot{\rho}_S + \dot{\beta}_S/\kappa) + X_T^* P_T \epsilon$$

where $P_T = I - P_S = I - X_S^\dagger X_S^*$ is the projection matrix onto $\mathrm{im}(X_T)$.

Integration on both sides and setting

$$\|\rho_T(t) + \beta_T(t)/\kappa\|_\infty = \|X_T^* X_S^\dagger(\rho_S(t) + \beta_S(t)/\kappa) + tX_T^* P_T \epsilon\|_\infty < 1$$

the first part follows from $\beta_T(t) = \kappa \cdot \mathrm{shrink}(\rho_T(t) + \beta_T(t)/\kappa, 1)$.

The second part is obtained by integration on (A.1)

$$\bar{\beta}_S(\tau) = \frac{1}{\tau}\int_0^\tau \beta_S(t)dt = \beta_S^* - \frac{1}{\tau}\Phi_S^{-1}(\rho_S(\tau) + \frac{1}{\kappa}\beta_S(\tau)) + \Phi_S^{-1}X_S^*\epsilon$$

followed by taking $\mathrm{sign}(\bar{\beta}_S(\tau)) = \mathrm{sign}(\beta_S^*)$. □

LEMMA A.2. *Suppose $\epsilon \sim N(0, \sigma^2 I_n)$, and $X \in R^{n\times p}$*

$$\text{(A.3)}\quad \mathrm{Prob}(\|X^T\epsilon\|_\infty > \sigma\sqrt{2(1+\mu)\log p}\max_j \|X_j\|) \leq \frac{1}{p^\mu\sqrt{\pi\log p}};$$

$$\text{(A.4)}\quad \mathrm{Prob}(\|X^T\epsilon\|_2 > \sigma\sqrt{2(1+\mu)\mathrm{tr}(X^TX)\log p}) \leq \frac{1}{p^\mu\sqrt{\pi\log p}}.$$

PROOF OF LEMMA A.2. From the Gaussian tail probability bound,

$$\begin{aligned}\mathrm{Prob}(|X_j^T\epsilon| > \sigma\sqrt{2(1+\mu)\log p}\|X_j\|) &\leq 2\frac{1}{\sqrt{2(1+\mu)\log p}\sqrt{2\pi}}e^{-\frac{2(1+\mu)\log p}{2}}\\ &\leq \frac{1}{p^{1+\mu}\sqrt{\pi\log p}}.\end{aligned}$$

The first inequality is directly the union bound of index j. The second inequality is obtained by the fact

$$\{\epsilon : \|X^T\epsilon\|_2 > \sigma\sqrt{2(1+\mu)\mathrm{tr}(X^TX)\log p}\} \in \bigcup_j \{\epsilon : |X_j^T\epsilon| > \sigma\sqrt{2(1+\mu)\log p}\|X_j\|\},$$

which ends the proof. □

PROOF OF LEMMA 5.1. Denote

$$A_t = \{i \in S | \mathrm{sign}(\tilde{\beta}_i^*) \neq \mathrm{sign}(\beta_i')\} \subseteq S.$$

Noticed that

$$\begin{aligned}\|\tilde{\beta}_S^* - \beta_S'\|_2^2 &\geq \sum_{i\in A_t}\tilde{\beta}_i^{*2}\\ &\geq \max\{\tilde{\beta}_{min}^*\sum_{i\in A_t}|\tilde{\beta}_i^*|, (\sum_{i\in A_t}|\tilde{\beta}_i^*|)^2/s\}\\ &\geq \max\{\tilde{\beta}_{min}^* D(\tilde{\beta}_S^*, \beta_S')/2, D(\tilde{\beta}_S^*, \beta_S')^2/4s\}\end{aligned}$$

and

$$\|\tilde{\beta}_S^* - \beta_S'\|_2 < \tilde{\beta}_{min}^* \Rightarrow A_t = \emptyset \Rightarrow \tilde{\beta}_S^* = \beta_S' \Rightarrow D(\tilde{\beta}_S^*, \beta_S') = 0$$

then according to the definition of Ψ and F, we have

$$\Psi(\beta_S') = \frac{\|\tilde{\beta}_S^* - \beta_S'\|_2^2}{2\kappa} + D(\tilde{\beta}_S^*, \beta_S') \leq F(\|\tilde{\beta}_S^* - \beta_S'\|_2^2)$$

which implies

$$F^{-1}(\Psi(\beta_S')) \leq \|\tilde{\beta}_S^* - \beta_S'\|_2^2$$

.

Combining the following result from right continuous differentiability

$$\left\langle \frac{d\rho_S'}{dt}, \beta_S' \right\rangle = 0$$

and the strong convexity conditions of $X_s^* X_s$, we have

$$\frac{d}{dt}(\Psi(\beta_S')) = -\left\langle \beta_S' - \tilde{\beta}_S^*, X_s^* X_s(\beta_S' - \tilde{\beta}_S^*) \right\rangle \leq -\gamma \|\tilde{\beta}_S^* - \beta_S'\|_2^2 \leq -\gamma F^{-1}(\Psi(\beta_S')),$$

as desired. □

Proof of Lemma 5.2. From the generalized Bihari's inequality

$$\tilde{\tau}_1 \leq -\int_0^{\tilde{\tau}_1} \frac{\frac{d}{dt}(\Psi(\beta_S'))}{\gamma F^{-1}(\Psi(\beta_S'))} dt = \frac{1}{\gamma} \int_{\Psi(\tilde{t}_\infty)}^{\Psi(0)} \frac{dx}{F^{-1}(x)}$$

Note that $\Psi(0) = \|\tilde{\beta}_S^*\|_1 + \frac{\|\tilde{\beta}_S^*\|^2}{2\kappa}$. so $F^{-1}(\Psi(0)) \leq \|\tilde{\beta}_S^*\|_2^2$. By continuity and monotonicity of $F(x)$ on $(\tilde{\beta}_{min}^2, +\infty)$ and and $\Psi(\tilde{t}_1) \geq \frac{\tilde{\beta}_{min}^2}{2\kappa}$

$$\begin{aligned}
\gamma\tilde{\tau}_1 &\leq \int_{\frac{\tilde{\beta}_{min}^2}{2\kappa}}^{\frac{\tilde{\beta}_{min}^2}{2\kappa}+2\tilde{\beta}_{min}} \frac{dx}{F^{-1}(x)} + \int_{\tilde{\beta}_{min}^2}^{s\tilde{\beta}_{min}^2} \frac{dF}{x} + \int_{s\tilde{\beta}_{min}^2}^{\|\tilde{\beta}\|_2^2} \frac{dF}{x} \\
&\leq \int_{\frac{\tilde{\beta}_{min}^2}{2\kappa}}^{\frac{\tilde{\beta}_{min}^2}{2\kappa}+2\tilde{\beta}_{min}} \frac{dx}{\tilde{\beta}_{min}^2} + \int_{\tilde{\beta}_{min}^2}^{s\tilde{\beta}_{min}^2} (\frac{1}{2\kappa x} + \frac{2}{\tilde{\beta}_{min} x}) dx + \int_{s\tilde{\beta}_{min}^2}^{\|\tilde{\beta}\|_2^2} (\frac{1}{2\kappa x} + \frac{\sqrt{s}}{x^{\frac{3}{2}}}) dx \\
&\leq \frac{4 + 2\log s}{\tilde{\beta}_{min}} + \frac{1}{\kappa} \log(\frac{\|\tilde{\beta}\|_2}{\tilde{\beta}_{min}}).
\end{aligned}$$

Proof of $\tilde{\tau}_2$ is straightforward now. For $t < \tilde{\tau}_2$,

$$\frac{d\Psi}{dt} \leq -\gamma |\tilde{\beta} - \beta|_2^2 \leq -\gamma \frac{C^2 s \log d}{n}$$

Let

$$\tilde{F}(x) = \frac{x}{2\kappa} + 2\sqrt{xs} \geq F(x) \quad \forall x > 0$$

Let $\tilde{F}^{-1}$ be the right-continuous inverse. Then $\|\beta(t) - \tilde{\beta}\|_2^2 \geq F^{-1}(\Psi(\beta)) \geq \tilde{F}^{-1}(\Psi(\beta))$. By generalized Bihari's inequality

$$-\frac{1}{\gamma}\frac{d\Psi}{dt} \geq |\tilde{\beta} - \beta|_2^2 \geq \tilde{F}^{-1}(\Psi)$$

Therefore

$$-\frac{1}{\gamma}\frac{d\Psi}{dt} \geq \max\{\tilde{F}^{-1}(\Psi), \frac{C^2 s \log d}{n}\}$$

Again, we have:

$$\tilde{\tau}_2 \leq \frac{1}{\gamma}\int_{\Psi(\tilde{t}_2)}^{\Psi(0)} \frac{dx}{\max\{\tilde{F}^{-1}(x), \frac{C^2 s \log d}{n}\}}$$

Noticed that

$$\tilde{F}^{-1}(\Psi(0)) \leq F^{-1}(\Psi(0)) \leq |\tilde{\beta}|_2^2$$

Therefore,

$$\begin{aligned}
\gamma\tilde{\tau}_2 &\leq \int_0^{\tilde{F}(C^2 s\log d/n)} \frac{dx}{\frac{C^2 s\log d}{n}} + \int_{\tilde{F}(C^2 s\log d/n)}^{\Psi(0)} \frac{dx}{\tilde{F}^{-1}(x)} \\
&\leq \int_0^{(C^2 s\log d/n)/2\kappa + 2Cs\sqrt{\log d/n}} \frac{dx}{\frac{C^2 s\log d}{n}} + \int_{C^2 s\log d/n}^{|\tilde{\beta}|_2^2} \frac{d\tilde{F}}{x} \\
&\leq \frac{1}{2\kappa} + \frac{2}{C}\sqrt{\frac{n}{\log d}} + \int_{C^2 s\log d/n}^{|\tilde{\beta}|_2^2} (\frac{1}{2\kappa x} + \frac{\sqrt{s}}{x^{3/2}})dx \\
&\leq \frac{4}{C}\sqrt{\frac{n}{\log d}} + \frac{1}{2\kappa}(1 + \log\frac{n|\tilde{\beta}|_2^2}{C^2 s\log d})
\end{aligned}$$

which gives the bounds. □

Proof of Theorem 4.1.

$$\begin{aligned}
\mathcal{A} &= \{\epsilon : \|X_S(X_S^*X_S)^{-1}X_S^*\epsilon\|_2 > 2\sigma\sqrt{s\log n}\} \\
\mathcal{B} &= \{\epsilon : \|(X_S^*X_S)^{-1}X_S^*\epsilon\|_\infty > 2\sigma\sqrt{\frac{\log p}{n\gamma}}\} \\
\mathcal{C} &= \{\epsilon : \|X_T^*P_T\epsilon\|_\infty > 2\sigma\sqrt{\frac{\log p}{n}}\max_{j\in T}\|X_j\|_n\}
\end{aligned}$$

Note $tr(X_S(X_S^*X_S)^{-1}X_S^*) = s, (X_S^*X_S)^{-1}X_S^* \cdot X_S(X_S^*X_S)^{-1} = (X_S^*X_S)^{-1} \preceq 1/\gamma$, and $X_T^*P_T \cdot P_TX_T \preceq X_T^*X_T$, using Lemma A.2, we have

$$\text{Prob}(A) \leq \frac{1}{n\sqrt{\pi \log n}}, \quad \text{Prob}(B) \leq \frac{1}{p\sqrt{\pi \log p}}, \quad \text{Prob}(C) \leq \frac{1}{p\sqrt{\pi \log p}}.$$

(1) (no-false-positivity for $\beta(t)$ up to τ) First consider the LB-ISS

$$\frac{d\rho_S}{dt} + \frac{1}{\kappa}\frac{d\beta_S}{dt} = -X_S^*X_S(\beta_S - \tilde{\beta}_S) \tag{A.5}$$

where $\tilde{\beta}_S = \beta_S^* + (X_S^*X_S)^{-1}X_S^*\epsilon$. It is easy to conclude $\|X_S(\tilde{\beta}_S - \beta_S)\|_2$ is monotonically decreasing based on the following observation

$$\frac{d}{dt}\frac{\|X_S(\tilde{\beta}_S - \beta_S)\|_2^2}{2n} = -\left\langle \frac{d\rho_S}{dt}, \frac{d\beta_S}{dt} \right\rangle - \frac{1}{\kappa}\left\| \frac{d\beta_S}{dt} \right\|_2^2 = -\frac{1}{\kappa}\left\| \frac{d\beta_S}{dt} \right\|_2^2 \leq 0$$

using $\langle d\rho_S(t)/dt, d\beta_S(t)/dt \rangle = 0$ from the assumption of Bregman ISS paths. On the set $\mathcal{A}^c \bigcup \mathcal{B}^c$,

$$\begin{aligned}
\|\beta_S\|_\infty &\leq \|\tilde{\beta}_S\|_\infty + \|\tilde{\beta}_S - \beta_S(t)\|_2 \\
&\leq \tilde{\beta}_{max} + \frac{\|X_S(\tilde{\beta}_S - \beta_S(t))\|_2}{\sqrt{n\gamma}} \\
&\leq \tilde{\beta}_{max} + \frac{\|X_S\tilde{\beta}_S\|_2}{\sqrt{n\gamma}} \\
&\leq \beta_{max}^* + 2\sigma\sqrt{\frac{\log p}{\gamma n}} + \frac{\|X_S\beta^*\|_2 + 2\sigma\sqrt{s\log n}}{\sqrt{n\gamma}}
\end{aligned}$$

Denote this upper bound as B. Returning to the original problem, by Lemma A.1, it suffices to have for all $t \leq \tau$,

$$1 > \|X_T^*X_S^\dagger(\rho_S + \beta_S/\kappa) + tX_T^*P_T\epsilon\|_\infty$$

The first part

$$\|X_T^*X_S^\dagger(\rho_S + \beta_S/\kappa)\|_\infty \leq (1-\eta)(1 + \|\beta_S\|_\infty/\kappa) \leq 1 - (1 - B/\kappa\eta)\eta$$

$$t \leq \tau := \frac{1 - B/\kappa\eta}{2}\eta\sigma^{-1}\sqrt{n/\log p}\left(\max_{j\in T}\|A_j\|\right)^{-1} = O(\eta\sigma^{-1}\sqrt{n/\log p}).$$

On the set $\mathcal{C}^c$, we have $t\|X_T^*P_T\epsilon\|_\infty < (1 - B/\kappa\eta)\eta$.

(2) (no-false-negativity for the mean path) it suffices to ensure

$$\beta^*_{\min} > \|\Phi_S^{-1} X_S^* \epsilon\|_\infty + \|\frac{1}{\tau}\Phi_S^{-1}(\rho_S + \beta_S/\kappa)\|_\infty.$$

Where $\Phi_S = X_S^* X_S$. The second part on the right hand side is $\|\frac{1}{\tau}\Phi_S^{-1}(\rho_S + \beta_S/\kappa)\|_\infty \le \frac{1}{\tau}\|\Phi_S^{-1}\|_\infty(1 + B/\kappa)$. The first part is bounded on the set $\mathcal{B}|^c$

(3) (l_2-error bound) Lemma 5.2 implies if $C > \frac{8\sigma\left(\max_{j\in T}\|X_j\|_n\right)}{\eta\gamma}$, when κ is big enough, we have

$$\begin{aligned}
\tilde{t}_2(C) &\le \frac{4}{C\gamma}\sqrt{\frac{n}{\log p}} + \frac{1}{2\kappa\gamma}(1 + \log\frac{n|\tilde{\beta}|_2^2}{C^2 s^2 \log p}) \\
&\le \frac{4}{C\gamma}\sqrt{\frac{n}{\log p}} + \frac{1}{2\kappa\gamma}(1 + \log\frac{n\|\beta^*\|_2^2 + 4\sigma^2 s\log p/\gamma}{C^2 s\log p}) \\
&\le \overline{\tau}
\end{aligned}$$

Thus $\exists \tau \in [0, \overline{\tau}]$

$$||\beta_S(\tau) - \tilde{\beta}_S||_2 \le C\sqrt{s\log(p)/n}$$

Note that with high probability

$$||\beta_S^* - \tilde{\beta}_S||_2 \le 2\sigma\sqrt{s\log(p)/n}\gamma^{-1/2}$$

(4) (Sign Consistency for β_t) The condition

$$\beta^*_{min} \ge \frac{4\sigma}{\gamma^{1/2}}\sqrt{\frac{\log p}{n}}$$

implies that $\tilde{\beta}$ has the same sign as β^* as well as $1/2|\beta_i^*| \le |\tilde{\beta}_i| \le 3/2|\beta_i^*|$ for each component i. Thus sign consistency is reached when $\tilde{t}_\infty \le \overline{\tau}$, or

$$\begin{aligned}
\frac{4 + 2\log s}{\gamma\tilde{\beta}_{min}} + \frac{1}{\kappa\gamma}\log(\frac{\|\tilde{\beta}\|_2}{\tilde{\beta}_{min}}) &\le \frac{8 + 4\log s}{\beta^*_{min}\gamma} + \frac{1}{\kappa\gamma}\log(\frac{3\|\beta^*\|_2}{\beta^*_{min}}) \\
&\le \overline{\tau}
\end{aligned}$$

which is ensured by κ big enough and

$$\beta^*_{\min} \ge 2\tilde{\beta}_{\min} \ge \left(\frac{4\sigma}{\gamma^{1/2}} \vee \frac{8\sigma(2 + \log s)\left(\max_{j\in T}\|X_j\|_n\right)}{\gamma\eta}\right)\sqrt{\frac{\log p}{n}}.$$

This completes the proof. □

A.2. Proof of Consistency of Linearized Bregman Iterations. First of all, we give a discrete version of generalized Bihari's inequality which is useful for Linearized Bregman iterations (1.4).

LEMMA A.3 (Discrete Generalized Bihari's inequality). *Consider the LB*

$$(\rho_{k+1}-\rho_k)+(\beta_{k+1}-\beta_k)/\kappa=-\alpha_k X_S^*X_S(\beta_k-\tilde\beta)$$

where $X_S^*X_S\geq\gamma I$. *Let the potential (or Lyapunov) function be*

$$\Psi_k=D(\tilde\beta,\beta_k)+\frac{||\beta_k-\tilde\beta||^2}{2\kappa}$$

Then the following difference inequality holds

$$\Psi_{k+1}-\Psi_k\leq-\alpha_k\gamma(1-\kappa\alpha_k\|X_SX_S^*\|/2)F^{-1}(\Psi_k)$$

where F *is defined by (5.5).*

PROOF OF LEMMA A.3. Similar to continue case, we have

$$\|\beta_k-\tilde\beta\|_2^2\geq F^{-1}(\Psi_k).$$

Since ℓ_1-norm is homogeneous of degree 1, its subgradient $\rho\in\partial\|\beta\|_1$ satisfies $\langle\rho,\beta\rangle=\|\beta\|_1$. Multiplying $\beta_k-\tilde\beta$ on the both sides of iteration equation, it leads to

$$\Psi_{k+1}-\Psi_k+(\rho_{k+1}-\rho_k)\beta_k-\|\beta_{k+1}-\beta_k\|^2/2\kappa=-\alpha_k\left\langle\beta_k-\tilde\beta,X_S^*X_S(\beta_k-\tilde\beta)\right\rangle$$

Note that for $i\in S$, $(\rho_{k+1}^{(i)}-\rho_k^{(i)})\beta_{k+1}^{(i)}=|\beta_{k+1}^{(i)}|-\rho_k^{(i)}\beta_{k+1}^{(i)}\geq 0$

$$\begin{aligned}
&\|\beta_{k+1}-\beta_k\|^2/\kappa-2(\rho_{k+1}-\rho_k)\beta_k\\
\leq\ &\|\beta_{k+1}-\beta_k\|^2/\kappa+2(\rho_{k+1}-\rho_k)(\beta_{k+1}-\beta_k)\\
\leq\ &\|\beta_{k+1}-\beta_k\|^2/\kappa+2(\rho_{k+1}-\rho_k)(\beta_{k+1}-\beta_k)+\|\rho_{k+1}-\rho_k\|^2\\
\leq\ &\kappa\|\rho_{k+1}-\rho_k+(\beta_{k+1}-\beta_k)/\kappa\|^2\\
=\ &\kappa\alpha_k^2\|X_S^*X_S(\beta_k-\tilde\beta)\|^2
\end{aligned}$$

$$\begin{aligned}
\Psi_{k+1}-\Psi_k\ &\leq\ -\frac{\alpha_k}{n}\left\langle X_S(\beta_k-\tilde\beta),X_S(\beta_k-\tilde\beta)\right\rangle+\frac{\alpha_k^2\kappa}{2n^2}\left\langle X_S^TX_S(\beta_k-\tilde\beta),X_S^TX_S(\beta_k-\tilde\beta)\right\rangle\\
&=\ -\frac{\alpha_k}{n}\left\langle X_S(\beta_k-\tilde\beta),(I-\kappa\alpha_kX_SX_S^*/2)X_S(\beta_k-\tilde\beta)\right\rangle\\
&\leq\ -\frac{\alpha_k}{n}(1-\kappa\alpha_k\|X_SX_S^*\|/2))\|X_S(\beta_k-\tilde\beta)\|^2\\
&\leq\ -\alpha_k\gamma(1-\kappa\alpha_k\|X_SX_S^*\|/2))\|\beta_k-\tilde\beta\|^2\\
&\leq\ -\alpha_k\gamma(1-\kappa\alpha_k\|X_SX_S^*\|/2))F^{-1}(\Psi_k)
\end{aligned}$$

which gives the result. □

Next we present a discrete stopping time bound from the inequality above.

LEMMA A.4 (Discrete Stopping Time Bounds). *Consider the LB*

$$(\rho_{k+1}-\rho_k)+(\beta_{k+1}-\beta_k)/\kappa=-\alpha_k X_S^* X_S(\beta_k-\tilde{\beta})$$

where $X_S^* X_S \geq \gamma I$ *and* $\alpha_k \leq \alpha, \quad \forall k$.

Define

$$\tilde{\tau}_1 := \inf\left\{\sum_{t=0}^{k-1}\alpha_t : \operatorname{sign}(\beta_k)=\operatorname{sign}(\tilde{\beta})\right\}$$

and

$$\tilde{\tau}_2(C) := \inf\left\{\sum_{t=0}^{k-1}\alpha_t : ||\beta_k-\tilde{\beta}||_2 \leq C\sqrt{\frac{s\log p}{n}}\right\}$$

Then the following bounds hold,

$$\tilde{\tau}_\infty \leq \frac{4+2\log s}{\tilde{\gamma}\tilde{\beta}_{min}}+\frac{1}{\kappa\tilde{\gamma}}\log(\frac{\|\tilde{\beta}\|_2}{\tilde{\beta}_{min}})+3\alpha$$

$$\tilde{\tau}_2(C) \leq \frac{4}{C\tilde{\gamma}}\sqrt{\frac{n}{\log p}}+\frac{1}{2\kappa\tilde{\gamma}}(1+\log\frac{n|\tilde{\beta}|_2^2}{C^2 s\log p})+2\alpha$$

where $\tilde{\gamma}=\gamma(1-\kappa\alpha\|X_S X_S^*\|/2)$

REMARK A.1. *Taking* $\alpha \to 0$, *it recovers the stopping time bounds in continuous case, Lemma 5.2.*

PROOF OF LEMMA A.4. Consider

$$\Psi_k = D(\tilde{\beta},\beta_k)+\frac{||\beta_k-\tilde{\beta}||^2}{2\kappa}$$

For a uniform upper bound on step sizes $\alpha_t \leq \alpha$, by the discrete Bihari's inequality in Lemma A.3

$$\Psi_{k+1}-\Psi_k \leq -\alpha_k\tilde{\gamma}F^{-1}(\Psi_k) \leq -\alpha_k\tilde{\gamma}\tilde{F}^{-1}(\Psi_k)$$

where $\tilde{\gamma}=\gamma(1-\kappa\alpha\|XX^*\|/2)$ and $\tilde{F}(x)=\frac{x}{2\kappa}+2\sqrt{xs} \geq F(x)$, $\forall x>0$.

For k such that $\Psi_k \geq 2\tilde{\beta}_{min}+\tilde{\beta}_{min}^2/2\kappa$, denote $L_k=F^{-1}(\Psi_k)$, which is non-increasing. Define $t_m=\sum_{t=0}^{m-1}\alpha_t$. Let $n_1=\sup\{n : L_n > s\tilde{\beta}_{min}^2\}$, then

$$\tilde{\gamma}\alpha_k \leq \frac{F(L_k)-F(L_{k+1})}{L_k}$$

then for $0 \le k \le n_1 - 1$,

$$\frac{F(L_k) - F(L_{k+1})}{L_k} \le (\frac{\log L_k}{2\kappa} - 2\sqrt{\frac{s}{L_k}}) - (\frac{\log L_{k+1}}{2\kappa} - 2\sqrt{\frac{s}{L_{k+1}}})$$

This is because of

$$\frac{L_k - L_{k+1}}{L_k} \le \log(\frac{L_k}{L_{k+1}})$$

using $1 - x \le -\log x$ for $x \le 1$, and

$$\frac{\sqrt{L_k} - \sqrt{L_{k+1}}}{L_k} \le \frac{\sqrt{L_k} - \sqrt{L_{k+1}}}{\sqrt{L_k}\sqrt{L_{k+1}}} = \frac{1}{\sqrt{L_{k+1}}} - \frac{1}{\sqrt{L_k}}$$

$$\begin{aligned}
\tilde{\gamma} t_{n_1} &\le (\frac{\log L_0}{2\kappa} - 2\sqrt{\frac{s}{L_0}}) - (\frac{\log L_{n_1}}{2\kappa} - 2\sqrt{\frac{s}{L_{n_1}}}) \\
&\le (\frac{\log \|\tilde{\beta}\|^2}{2\kappa} - 2\sqrt{\frac{s}{\|\tilde{\beta}\|^2}}) - (\frac{\log s\tilde{\beta}_{min}^2}{2\kappa} - 2\sqrt{\frac{s}{s\tilde{\beta}_{min}^2}})
\end{aligned}$$

Let $n_2 = \sup\{n : L_n > \tilde{\beta}_{min}^2\}$,

$$\tilde{\gamma}\alpha_k \le \frac{F(L_k) - F(L_{k+1})}{L_k}$$

Then similarly, we have

$$\begin{aligned}
\tilde{\gamma}(t_{n_2} - t_{n_1+1}) &\le (\frac{1}{2\kappa} + \frac{2}{\tilde{\beta}_{min}})(\log L_{n_1+1} - \log L_{n_2}) \\
&\le (\frac{1}{2\kappa} + \frac{2}{\tilde{\beta}_{min}})(\log s\tilde{\beta}_{min}^2 - \log \tilde{\beta}_{min}^2)
\end{aligned}$$

Let $n_3 = \sup\{n : \Psi_n > \tilde{\beta}_{min}^2/2\kappa\}$

$$\begin{aligned}
\tilde{\gamma}(t_{n_3} - t_{n_2+1}) &\le \sum_{k=n_2+1}^{n_3-1} \frac{\Psi_k - \Psi_{k+1}}{\tilde{\beta}_{min}^2} \\
&\le \frac{\frac{\tilde{\beta}_{min}^2}{2\kappa} + 2\tilde{\beta}_{min} - \frac{\tilde{\beta}_{min}^2}{2\kappa}}{\tilde{\beta}_{min}^2} \\
&= \frac{2}{\tilde{\beta}_{min}}
\end{aligned}$$

To sum up, we have

$$\tilde{\tau}_1 \leq t_{n_3+1} \leq \frac{4+2\log s}{\tilde{\gamma}\tilde{\beta}_{min}} + \frac{1}{\kappa\tilde{\gamma}}\log(\frac{\|\tilde{\beta}\|_2}{\tilde{\beta}_{min}}) + 3\alpha$$

Similarly, we have

$$\tilde{\tau}_2(C) \leq \frac{4}{C\tilde{\gamma}}\sqrt{\frac{n}{\log d}} + \frac{1}{2\kappa\tilde{\gamma}}(1+\log\frac{n|\tilde{\beta}|_2^2}{C^2 s^2 \log d}) + 2\alpha,$$

which ends the proof. □

Proof of Theorem 4.2. The proof is the same to the continue case. The only difference is the decreasing of $\|X(\beta_k - \tilde{\beta})\|_2$ needs the condition $\kappa\alpha\|X_S X_S^*\| < 2$.

Consider the LB

$$(\rho_{k+1} - \rho_k) + (\beta_{k+1} - \beta_k)/\kappa = -\alpha_k X_S^* X_S(\beta_k - \tilde{\beta})$$

where $X_S^* X_S \geq \gamma I$.

$$\begin{aligned}
& \|X_S(\beta_{k+1} - \tilde{\beta})\|^2 - \|X_S(\beta_k - \tilde{\beta})\|^2 \\
= \;& \|X_S(\beta_{k+1} - \beta_k)\|^2 + 2(\beta_{k+1} - \beta_k)^T X_S^T X_S(\beta_k - \tilde{\beta}) \\
= \;& \|X_S(\beta_{k+1} - \beta_k)\|^2 - 2n/\alpha_k(\beta_{k+1} - \beta_k)^T[(\rho_{k+1} - \rho_k) + (\beta_{k+1} - \beta_k)/\kappa] \\
\leq \;& \|X_S(\beta_{k+1} - \beta_k)\|^2 - 2n/\alpha_k(\beta_{k+1} - \beta_k)^T(\beta_{k+1} - \beta_k)/\kappa \\
= \;& n(\beta_{k+1} - \beta_k)^T(X_S^* X_S - 2/\alpha_K\kappa)(\beta_{k+1} - \beta_k) \\
\leq \;& 0,
\end{aligned}$$

where we have used $\|X_S X_S^*\| = \|X_S^* X_S\|$. Hence $\|X_S(\tilde{\beta}_S - \beta_k)\|_2$ is monotonically nonincreasing. □

Note that this implies that $\|r_t\| := \|y - X\beta_t\|$ is monotonically nonincreasing for all $t \in (0, \bar{\tau})$. The following lemma makes it precise.

Lemma A.5. *For $t \in [0, \bar{\tau}]$, the residue admits an orthogonal decomposition*

$$\|r_t\|^2 = \|y - X\beta_t\|^2 = \|X_S(\tilde{\beta}_S^* - \beta_S(t))\|^2 + \|P_T\varepsilon\|^2$$

and is monotonically nonincreasing.

PROOF. By Pythagorean Theorem,

$$\begin{aligned}\|r_t\|^2 &= \|X_S(\beta^* - \beta_t) + \varepsilon\|^2 = \|P_S X_S(\beta^* - \beta_t) + P_S\varepsilon\|^2 + \|(I - P_S)\varepsilon\|^2 \\ &= \|X_S(\beta^* - \beta_t) + X_S(X_S^* X_S)^{-1} X_S^*\varepsilon\|^2 + C_{\epsilon,S} \\ &= \|X_S(\tilde{\beta}_S^* - \beta_S(t))\|^2 + C_{\epsilon,S}\end{aligned}$$

and the conclusion follows from that $\|X_S(\tilde{\beta}_S^* - \beta_S(t))\|$ is monotonically nonincreasing. □

ACKNOWLEDGEMENTS

The research of Stanley Osher was supported in part by NSF grant 1118971 and ONR grant N000141210838. The research of Yuan Yao was supported in part by National Basic Research Program of China under grant 2012CB825501 and NSFC grant 61071157. The research of Wotao Yin was supported in part by NSF grants DMS-1349855 and DMS-1317602 and ARO MURI grant W911NF-09-1-0383.

SUPPLEMENTARY MATERIAL

Supplement A: Matlab Linearized Bregman codes (http://www.math.ucla.edu/ wotaoyin/software.html).

Supplement B: R Package of Linearized Bregman algorithms (http://www.math.pku.edu.cn/teachers/yaoy/reference/Libra_1.1.tar.gz).

REFERENCES

[Bel43] Richard Bellman, *The stability of solutions of linear differential equations*, Duke Math. J. **10** (1943), no. 4, 643–647.

[Bih56] Imre Bihari, *A generalization of a lemma of bellman and its application to uniqueness problems of differential equations*, Acta Mathematica Hungarica **7** (1956), no. 1, 81–94.

[BMBO13] Martin Burger, Michael Möller, Martin Benning, and Stanley Osher, *An adaptive inverse scale space method for compressed sensing*, Mathematics of Computation **82** (2013), no. 281, 269–299.

[BOXG05] Martin Burger, Stanley Osher, Jinjun Xu, and Guy Gilboa, *Nonlinear inverse scale space methods for image restoration*, Variational, Geometric, and Level Set Methods in Computer Vision, Springer, 2005, pp. 25–36.

[BY02] Peter Bühlmann and Bin Yu, *Boosting with the l_2-loss: Regression and classification*, Journal of American Statistical Association **98** (2002), 324–340.

[CCS10] Jian-Feng Cai, Emmanuel J. Candès, and Zuowei Shen, *A Singular Value Thresholding Algorithm for Matrix Completion*, SIAM Journal on Optimization **20** (2010), no. 4, 1956–1982 (en).

[CDLL98] Antonin Chambolle, Ronald A. DeVore, Nam-Yong Lee, and Bradley J. Lucier, *Nonlinear wavelet image processing: variational problems, compression, and noise removal through wavelet shrinkage*, IEEE Transactions on Image Processing **7** (1998), no. 3, 319–335.

[CDS98] Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders, *Atomic decomposition by basis pursuit*, SIAM Journal on Scientific Computing **20** (1998), 33–61.

[CW11] Tony Cai and Lie Wang, *Orthogonal matching pursuit for sparse signal recovery*, IEEE Transactions on Information Theory **57** (2011), no. 7, 4680–4688.

[CWX10] Tony Cai, Lie Wang, and Guangwu Xu, *Stable recovery of sparse signals and an oracle inequality*, IEEE Transactions on Information Theory **56** (2010), no. 7, 3516–3522.

[CXZ09] Tony Cai, Guangwu Xu, and Jun Zhang, *On recovery of sparse signals via l_1 minimization*, IEEE Transactions on Information Theory **55** (2009), no. 7, 3588–3397.

[DD02] Christine De Mol and Michael Defrise, *A note on wavelet-based inversion algorithms*, Contemporary Mathematics **313** (2002), 8596.

[DDD04] Ingrid Daubechies, Michel Defrise, and Christine De Mol, *An iterative thresholding algorithm for linear inverse problems with a sparsity constraint*, Comm. Pure Appl. Math. **57** (2004), no. 11, 1413–1457.

[DH01] David L. Donoho and Xiaoming Huo, *Uncertainty principles and ideal atomic decomposition*, IEEE Transactions on Information Theory **47** (2001), no. 7, 2845–2862.

[DJ95] David L. Donoho and Iain M. Johnstone, *Adapting to unknown smoothness via wavelet shrinkage*, J. Amer. Statist. Assoc. **90** (1995), 1200–1224.

[Don95] David Donoho, *De-noising by soft-thresholding*, IEEE Transactions on Information Theory **41** (1995), no. 3, 613–627.

[EHJT04] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani, *Least angle regression*, Annals of Statistics **32** (2004), no. 2, 407–499.

[EHN96] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of inverse problems*, Kluwer Academic Publishers, 1996.

[FL01] Jianqing Fan and Runze Li, *Variable selection via nonconcave penalized likelihood and its oracle properties*, Journal of American Statistical Association (2001), 1348–1360.

[Fri01] J. H. Friedman, *Greedy function approximation: A gradient boosting machine*, The Annals of Statistics **29** (2001), 1189–1232.

[Gro19] Thomas Hakon Gronwall, *Note on the derivatives with respect to a parameter of the solutions of a system of differential equations*, Annals of Mathematics **20** (1919), no. 2, 292–296.

[Hes69] Magnus R Hestenes, *Multiplier and gradient methods*, Journal of optimization theory and applications **4** (1969), no. 5, 303–320.

[LY13] Ming-Jun Lai and Wotao Yin, *Augmented ℓ_1 and Nuclear-Norm Models with a Globally Linearly Convergent Algorithm*, SIAM Journal on Imaging Sciences **6** (2013), no. 2, 1059–1091 (en).

[OBG+05] Stanley Osher, Martin Burger, Donald Goldfarb, Jinjun Xu, and Wotao Yin, *An iterative regularization method for total variation-based image restoration*, SIAM Journal on Multiscale Modeling and Simulation **4** (2005), no. 2, 460–489.

[Pow67] Michael JD Powell, *A method for non-linear constraints in minimization problems*, UKAEA, 1967.

[RWY11] Garvesh Raskutti, Martin J. Wainwright, and Bin Yu, *Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls*, IEEE Transactions on Information Theory **57** (2011), no. 10, 6976–6994.

[Tib96] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. of the Royal

Statistical Society, Series B **58** (1996), no. 1, 267–288.

[Tro04] Joel A. Tropp, *Greed is good: Algorithmic results for sparse approximation*, IEEE Trans. Inform. Theory **50** (2004), no. 10, 2231–2242.

[Wai09] Martin J. Wainwright, *Sharp thresholds for high-dimensional and noisy sparsity recovery using l_1-constrained quadratic programming (lasso)*, IEEE Transactions on Information Theory **55** (2009), no. 5, 2183–2202.

[Yin10] Wotao Yin, *Analysis and Generalizations of the Linearized Bregman Method*, SIAM Journal on Imaging Sciences **3** (2010), no. 4, 856–877 (en).

[YL07] Ming Yuan and Yi Lin, *On the nonnegative garrote estimator*, Journal of the Royal Statistical Society, Series B **69** (2007), no. 2, 143–161.

[YLYR13] Kun Yuan, Qing Ling, Wotao Yin, and Alejandro Ribeiro, *A Linearized Bregman Algorithm for Decentralized Basis Pursuit*, EUSIPCO (2013).

[YODG08] Wotao Yin, Stanley Osher, Jerome Darbon, and Donald Goldfarb, *Bregman iterative algorithms for compressed sensing and related problems*, SIAM Journal on Imaging Sciences **1** (2008), no. 1, 143–168.

[YRC07] Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto, *On early stopping in gradient descent learning*, Constructive Approximation **26** (2007), no. 2, 289–315.

[Zou06] Hui Zou, *The adaptive lasso and its oracle properties*, Journal of the American Statistical Association **101** (2006), no. 476, 1418–1429.

[ZY06] Peng Zhao and Bin Yu, *On model selection consistency of lasso*, J. Machine Learning Research **7** (2006), 2541–2567.

School of Mathematical Sciences
Peking University
Beijing, China 100871
E-mail: fengruan@stanford.edu
xiongjiechao@pku.edu.cn
yuany@math.pku.edu.cn
URL: http://www.math.pku.edu.cn/teachers/yaoy/

Department of Mathematics
University of California
Los Angels, CA 90095
E-mail: sjo@math.ucla.edu
wotaoyin@math.ucla.edu