Appl. Comput. Harmon. Anal. 41 (2016) 436–469

Contents lists available at ScienceDirect

Applied and Computational Harmonic Analysis

www.elsevier.com/locate/acha

Sparse recovery via differential inclusions

Stanley Osher [a], Feng Ruan [b,c], Jiechao Xiong [b], Yuan Yao [b,*], Wotao Yin [a]

[a] Department of Mathematics, University of California, Los Angeles, CA 90095, United States
[b] School of Mathematical Sciences, Peking University, Beijing, 100871, China
[c] Department of Statistics, Stanford University, Stanford, CA 94305, United States

ARTICLE INFO

Article history:
Received 19 September 2014
Received in revised form 26 December 2015
Accepted 9 January 2016
Available online 14 January 2016
Communicated by Spec. Issue Guest Editor

Keywords:
Linearized Bregman
Differential inclusion
Early stopping regularization
Statistical consistency

ABSTRACT

In this paper, we recover sparse signals from their noisy linear measurements by solving nonlinear differential inclusions, which is based on the notion of inverse scale space (ISS) developed in applied mathematics. Our goal here is to bring this idea to address a challenging problem in statistics, *i.e.* finding the oracle estimator which is unbiased and sign consistent using dynamics. We call our dynamics *Bregman ISS* and *Linearized Bregman ISS*. A well-known shortcoming of LASSO and any convex regularization approaches lies in the bias of estimators. However, we show that under proper conditions, there exists a bias-free and sign-consistent point on the solution paths of such dynamics, 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, *i.e.* the oracle estimator. Therefore, their solution paths are regularization paths better than the LASSO regularization path, since the points on the latter path are biased when sign-consistency is reached. 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 a simple iterative thresholding rule 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, which extends the previous results and leads to exponentially fast recovering of sparse signals before selecting wrong ones.

1. Introduction

We study a dynamic approach to recover a sparse signal $\beta^* \in \mathbb{R}^p$ from its noisy linear measurements

* Corresponding author. Fax: +1 310 206 6673 (c/o S. Osher).
E-mail addresses: sjo@math.ucla.edu (S. Osher), fengruan@stanford.edu (F. Ruan), xiongjiechao@pku.edu.cn (J. Xiong), yuany@math.pku.edu.cn (Y. Yao), wotaoyin@math.ucla.edu (W. Yin).
URL: http://www.math.pku.edu.cn/teachers/yaoy/ (Y. Yao).

http://dx.doi.org/10.1016/j.acha.2016.01.002

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

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 $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ is Gaussian 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. $T = \{i : \beta^*_i = 0\}$. 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 $[X_S\ X_T] = X$.

Such a problem has been widely studied in applied mathematics [12], engineering, and statistics [36], see for example surveys in [22,13]. In these works, convex regularization or relaxation approach has been exploited to overcome the combinatorial explosion of searching the best sparse signals using subset least squares. However, it has been known since [25] that all convex regularization approaches lead to biased estimators whose expectation does not meet the true signal, which motivates the exploration of using non-convex regularization yet it may suffer from a computational hurdle of locating the global optima [27,29].

To address this dilemma between statistical accuracy and computational hurdle, in this paper we introduce some dynamics from the *Inverse Scale Space* (ISS) method, which first appeared in the image restoration literature in [6,3,7,2,8] and analyzed and implemented carefully in [5]. The name refers to the observation there that large-scale (image) features are recovered before small-scale ones. Our goal here is to show that such dynamics provides a surprisingly simple way to statistically accurate (unbiased and sign-consistent) estimator if equipped with a new type regularization – early stopping. Our results also extend those early error analysis on ISS to statistical consistency, establishing model selection consistency as well as minimax optimal l_2 error bounds under comparable conditions to LASSO, *etc.*

The first one, called *Bregman ISS* here, is given by the nonlinear 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 $\beta_0 = 0$. As it evolves, the component which reaches $|\rho_t(i)| = 1$ enters into our selection $\beta_t(i) \neq 0$. Hence roughly speaking, the larger magnitude $X_i^T(y - X\beta_t)$ has, the faster the component is selected. In the ideal case, we hope the signals in S are selected faster than non-signals in T, whose conditions will be our main concern in this paper. Under general conditions, 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.

A damping version of the first one, called *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 for all $\kappa > 0$. Alternatively, (1.3) can be obtained as a differential inclusion replacing the l_1-norm in (1.2b) by the Elastic Net [44] penalty $\|\beta_t\|_1 + \frac{1}{\kappa}\|\beta_t\|_2^2$ which will be discussed later.