Sparse Recovery via Differential Inclusions

Yuan Yao

School of Mathematical Sciences
Peking University

September 2, 2014

with Stanley Osher (UCLA), Feng Ruan (PKU & Stanford), Jiechao Xiong (PKU), and Wotao Yin (UCLA), et al.

1. Inverse Scale Space (ISS) Dynamics
   - ISS
   - Dynamics of Bregman Inverse Scale Space
   - Discrete Algorithm: Linearized Bregman Iteration

2. Path Consistency Theory
   - Sign-consistency
   - l_2-consistency

3. Discussion

Background

Assume that $\beta^* \in \mathbb{R}^p$ is sparse and unknown. Consider recovering β^* from

$$y = X\beta^* + \epsilon,$$

where ϵ is **noise**.

Note

- $S := \mathrm{supp}(\beta^*)$ and T be its complement.
- X_S (X_T) be the columns of X with indices restricted on S (T)
- $\epsilon \sim \mathcal{N}(0, \sigma^2)$ (sub-Gaussian in general)
- X is n-by-p, with $p \gg n$.

Statistical Consistency of Algorithms

- Orthogonal Matching Pursuit (OMP, Mallat-Zhang'93)
  - noise-free: Tropp'04
  - noise: Cai-Wang'11
- LASSO (Tibshirani'96)
  - sign-consistency: Yuan-Lin'06, Zhao-Yu'06, Zou'07, Wainwright'09
  - l_2-consistency: Ritov-Bickel-Tsybakov'09 (also Dantzig)
  - related: BPDN (Chen-Donoho-Saunders'96), Dantzig Selector (Candes-Tao'07)
- Anything else do you wanna hear?

Optimization + Noise = H.D. Statistics?

- $p >> n$: impossible to be strongly convex

$$\min_{\beta} L(\beta) := \frac{1}{n}\sum_{i=1}^{n} \rho(y_i - x_i^T \beta), \quad \text{convex } \rho \text{ (Huber'73)}$$

- in presence of noise, not every optimizer $\arg\min L(\beta)$ is desired: mostly overfitting
- convex constraint/penalization: avoid overfiting, tractable but lead to bias $\Rightarrow$ non-convex? (hard to find global optimizer)
- dynamics: every algorithm is dynamics (Turing), not necessarily optimizing an objective function

Inverse Scale Space (ISS) Dynamics

- Bregman ISS

$$\dot{\rho}(t) = \frac{1}{n} X^T (y - X\beta(t)),$$
$$\rho(t) \in \partial \|\beta(t)\|_1.$$

Inverse Scale Space (ISS) Dynamics

- Bregman ISS

$$\dot{\rho}(t) = \frac{1}{n} X^T (y - X\beta(t)),$$
$$\rho(t) \in \partial \|\beta(t)\|_1.$$

Limit is solution to $\min_\beta \|\beta\|_1, \quad \text{s.t.} \quad X^T y = X^T X \beta.$

Inverse Scale Space (ISS) Dynamics

- Bregman ISS

$$\dot{\rho}(t) = \frac{1}{n} X^T (y - X\beta(t)),$$
$$\rho(t) \in \partial \|\beta(t)\|_1.$$

Limit is solution to $\min_\beta \|\beta\|_1, \quad \text{s.t.} \;\; X^T y = X^T X \beta.$

- Linearized Bregman ISS

$$\dot{\rho}(t) + \frac{1}{\kappa} \dot{\beta}(t) = \frac{1}{n} X^T (y - X\beta(t)),$$
$$\rho(t) \in \partial \|\beta(t)\|_1.$$

Inverse Scale Space (ISS) Dynamics

- Bregman ISS

$$\dot{\rho}(t) = \frac{1}{n} X^T (y - X\beta(t)),$$
$$\rho(t) \in \partial \|\beta(t)\|_1.$$

Limit is solution to $\min_\beta \|\beta\|_1, \quad \text{s.t.} \;\; X^T y = X^T X \beta.$

- Linearized Bregman ISS

$$\dot{\rho}(t) + \frac{1}{\kappa}\dot{\beta}(t) = \frac{1}{n} X^T (y - X\beta(t)),$$
$$\rho(t) \in \partial \|\beta(t)\|_1.$$

Limit is solution to $\min_\beta \|\beta\|_1 + \frac{1}{2\kappa}\|\beta\|_2^2, \quad \text{s.t.} \;\; X^T y = X^T X \beta.$

Algorithmic regularization

We claim that there exists points on their paths $(\beta(t), \rho(t))_{t\geq 0}$, which are

- sparse
- sign-consistent (the same sparsity pattern of nonzeros as true signal)
- unbiased (or less bias) than LASSO

Oracle Estimator

If S is disclosed by an oracle, the *oracle estimator* is the subset least square solution with $\tilde{\beta}^*_T = 0$ and for $\Sigma_n = \frac{1}{n} X_S^T X_S \to \Sigma_S$,

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

"Oracle properties"

- **Model selection consistency**: $\mathrm{supp}(\tilde{\beta}^*) = S$;
- **Normality**: $\tilde{\beta}^*_S \sim \mathcal{N}(\beta^*, \frac{\sigma^2}{n} \Sigma_n^{-1})$.

So $\tilde{\beta}^*$ is unbiased, i.e. $E[\tilde{\beta}^*] = \beta^*$.

Recall LASSO

LASSO:

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

optimality condition:

$$\frac{\rho_t}{t} = \frac{1}{n} X^T (y - X\beta_t), \tag{2a}$$

$$\rho_t \in \partial \|\beta_t\|_1, \tag{2b}$$

where $\lambda = 1/t$ is often used in literature.

- Tibshirani'1996 (LASSO)
- Chen-Donoho-Saunders'1996 (BPDN)

The Bias of LASSO

- Path consistency: $\exists \tau_n \in (0,\infty)$, $\mathrm{supp}(\hat{\beta}_{\tau_n}) = S$ (e.g. , Zhao-Yu'06, Zou'06, Yuan-Lin'07, Wainwright'09)
- LASSO is biased

$$(\hat{\beta}_{\tau_n})_S = \tilde{\beta}^*_S - \frac{1}{\tau_n}\Sigma_n^{-1}\rho_{\tau_n}, \quad \tau_n > 0$$

  - e.g. $X = Id$, $n = p = 1$,

$$\hat{\beta}_\tau = \begin{cases} 0, & \text{if } \tau < 1/y; \\ y - 1/\tau, & \text{otherwise,} \end{cases}$$

  - (Fan-Li'2001) non-convex penalty is necessary (SCAD, Zhang's PLUS, Zou's Adaptive LASSO, etc.)
  - Any other simple scheme?

Differentiation of LASSO's KKT Equation

Taking derivative (assuming differentiability) w.r.t. t

$$\rho_t = \frac{1}{n} X^T (y - X\beta_t) t$$

$$\Rightarrow \dot{\rho}_t = \frac{1}{n} X^T (y - X(\dot{\beta}_t t + \beta_t)), \quad \rho_t \in \partial \|\beta_t\|_1$$

- Debias: sign-consistency $(\mathrm{sign}(\beta_\tau) = \mathrm{sign}(\beta^*)) \Rightarrow$ oracle estimator $\beta'_\tau := \dot{\beta}_\tau \tau + \beta_\tau = \tilde{\beta}^*$
- e.g. $X = Id$, $n = p = 1$,

$$\beta'_t = \begin{cases} 0, & \text{if } t < 1/y; \\ y, & \text{otherwise,} \end{cases}$$

Inverse scale space (ISS)

Nonlinear ODE (differential inclusion)

$$\dot{\rho}_t = \frac{1}{n} X^T (y - X\beta_t), \tag{3a}$$

$$\rho_t \in \partial \|\beta_t\|_1. \tag{3b}$$

starting at $t = 0$ and $\rho(0) = \beta(0) = \mathbf{0}$.

- Replace ρ/t in LASSO by $\mathrm{d}\rho/\mathrm{d}t$
- Burger-Gilboa-Osher-Xu'06 (image recovery and recovers the objects in an image in an inverse-scale order as t increases (larger objects appear in β_t first))

Solution Path

- β_t is piece-wise constant in t:

$$\begin{aligned} \beta_{t_{k+1}} = \arg\min_\beta \quad & \|y - X\beta\|_2^2 & \\ \text{subject to} \quad & (\rho_{t_{k+1}})_i \beta_i \geq 0 & \forall\ i \in S_{k+1}, \\ & \beta_j = 0 & \forall\ j \in T_{k+1}. \end{aligned} \tag{4}$$

- $t_{k+1} = \sup\{t > t_k : \rho_{t_k} + \frac{t - t_k}{n} X^T(y - X\beta_{t_k}) \in \partial\|\beta_{t_k}\|_1\}$
- ρ_t is piece-wise linear in t,

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

- Sign consistency $\rho_t = \mathrm{sign}(\beta^*) \Rightarrow \beta_t = \tilde{\beta}^*$

Discretized Algorithm

Damped Dynamics: continuous solution path

$$\dot{\rho}_t + \frac{1}{\kappa}\dot{\beta}_t = \frac{1}{n}X^T(y - X\beta_t), \quad \rho_t \in \partial\|\beta_t\|_1. \tag{5}$$

Linearized Bregman Iteration as forward Euler discretization (Osher-Burger-Goldfarb-Xu-Yin'05, Yin-Osher-Goldfarb-Darbon'08): for $\rho_k \in \partial\|\beta_k\|_1$,

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

- Damping factor: $\kappa > 0$
- Step size: α_k

Comparisons

Linearized Bregman Iteration:

$$z_{t+1} = z_t - \alpha_t X^T (X \kappa Shrink(z_t, 1) - y)$$

- This is not **ISTA**:

$$z_{t+1} = Shrink(z_t - \alpha_t X^T (X z_t - y), \lambda)$$

  - ISTA solves **LASSO** for fixed λ
- This is not **OMP** which only adds in variables.
- This is not Donoho-Maleki-Montanari's AMP

AUC of ISS often beats LASSO

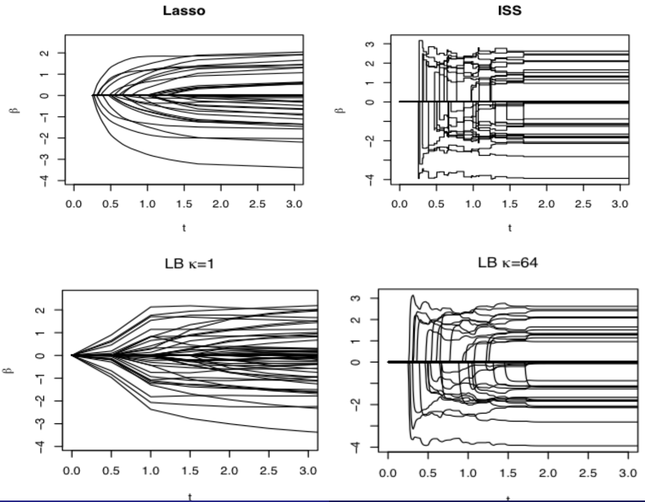
$n = 200$, $p = 100$, $S = \{1, \ldots, 30\}$, $x_i \sim N(0, \Sigma_p)$ ($\sigma_{ij} = 1/(3p)$ for $i \neq j$ and 1 otherwise)

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.

But regularization paths are different.



Path Consistency Theory

We are going to present a consistency theory where

- Under what conditions one can achieve
  - sign consistency (model selection consistency)
  - l_2-consistency ($\|\beta(t) - \tilde{\beta}^*\|_2 \leq O(\sqrt{s \log p/n})$)
- When sign-consistency holds, Bregman ISS path returns the oracle estimator without bias
- Early stopping regularization against overfitting noise

Assumptions

(A1) Restricted Strongly Convex: $\exists \gamma \in (0,1]$,

$$\frac{1}{n} X_S^T X_S \geq \gamma I$$

(A2) Incoherence/Irrepresentable Condition: $\exists \eta \in (0,1)$,

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

- The incoherence condition is used independently in Tropp'04, Yuan-Lin'05, Zhao-Yu'06, and Zou'06, Wainwright'09,etc.

Path Consistency

Theorem (Path Consistency of Bregman ISS)

Assume (A1) and (A2). Define

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

and the smallest magnitude $\beta^*_{\min} = \min(|\beta^*_i| : i \in S)$*. Then*

- **(No-false-positive)** *for all* $t \leq \overline{\tau}$*, the path has no-false-positive with high probability,* $\mathrm{supp}(\beta(t)) \subseteq S$*;*

Path Consistency, continued

Theorem (continued)

- **(Sign consistency for path)** *instead if the signal is strong enough such that*

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

then there is $\tau \leq \bar{\tau}$ *such that solution path* $\beta(t)$ *reaches sign consistency for every* $t \in [\tau, \bar{\tau}]$.

Path Consistency, continued

Theorem (continued)

- **(l_2-consistency)** *Under (A1) and (A2), there is an early stopping $\tau_n \in [0, \overline{\tau}]$, such that with high probability $\|\beta(\tau_n) - \beta^*\|_2 \leq C_0 \sqrt{\frac{s \log d}{n}}$, where*

$$C_0 = \frac{2\sigma}{\gamma^{1/2}} + \frac{8\sigma\left(\max_{j \in T} \|X_j\|\right)}{\eta\gamma}$$

Note: for $\bar{\gamma} I_s \geq \frac{1}{n} X_S^T X_S \geq \underline{\gamma} I_s$,

$$\|\beta(\overline{\tau}) - \beta^*\|_2 \leq \sqrt{\frac{\bar{\gamma}}{\underline{\gamma}}}\left(C_0 + \frac{2\sigma}{\sqrt{\underline{\gamma}}}\right)\sqrt{\frac{s \log p}{n}}$$

Remark

- Similar results for LASSO are established in Wainwright'09 with $\lambda^* = 1/\bar{\tau}$, where the lasso path are sign-consistent
- $\beta(\bar{\tau})$ is unbiased, while LASSO estimator is biased
- The l_2-error bound is of minimax optimal rates
- The temporal mean path

$$\bar{\beta}(\tau) := \frac{1}{\tau}\int_0^\tau \beta(s)ds \qquad (6)$$

is sign-consistent under precisely the same condition as LASSO, though they are different!

Generalization To Discrete Setting

Theorem (Linearized Bregman Iterations)

Assume that κ is large enough and α is small enough, with $\kappa\alpha\|X_S^* X_S\| < 2$,

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

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

then all the results can be extended to discrete algorithm setting (Linearized Bregman Iterations).

Understanding the Dynamics

Bregman ISS as gradient descent in dual space:

$$\dot{\rho}_t = -\nabla L(\beta_t) = \frac{1}{n} X^T (y - X(\dot{\beta}_t t + \beta_t)), \quad \rho_t \in \partial \|\beta_t\|_1$$

- incoherence condition and strong signals ensure it firstly evolves on index set S to reduce the loss
- strongly convex in subspace restricted on index set $S \Rightarrow$ fast decay in loss
- early stopping after all strong signals are detected, before picking up the noise

Idea of Proof: I

1. No-false-positive condition is the same as LASSO
2. For $t \leq \bar{\tau}$ consider *Oracle dynamcs*

$$\frac{d\rho'_S}{dt} = -\frac{1}{n} X_S^T X_S(\beta'_S - \tilde{\beta}^*_S), \quad \rho'_S(t) \in \partial \|\beta'_S(t)\|_1, \qquad (7)$$

where $\frac{1}{n} X_S^T X_S \geq \gamma I_s$.

- a generalized Grönwall-Bellman-Bihari inequality:

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

where F is a piecewise polynomial and D is the Bregman distance associated to $\| \cdot \|_1$.

Idea of Proof: II

3 Sign-consistency and l_2-consistency are reached by setting these stopping time $\tilde{\tau}_i \leq \bar{\tau}$ where oracle dynamics meets Bregman ISS

$$\tilde{\tau}_1 := \inf\{t > 0 : \mathrm{sign}(\beta'_S) = \mathrm{sign}(\tilde{\beta}^*_S)\} \leq O(\log s/\beta^*_{\min})$$

$$\tilde{\tau}_2(C) := \inf\left\{t > 0 : ||\beta'_S - \tilde{\beta}^*_S||_2 \leq C\sqrt{\frac{s\log p}{n}}\right\} \leq O(\frac{1}{C}\sqrt{\frac{n}{p}})$$

Discussion

These results can be extended to discrete algorithm, the simple 1-line Linearized Bregman iteration:

- achieve mean path sign-consistency, **equivalent to LASSO**
- and path sign-consistency with less bias, **better than LASSO**
- LB iteration is as simple as ISTA, but more powerful
  - cost: two free-parameters, κ and step-size α_k
  - tips: $\alpha_k \kappa \|\Sigma_n\| < 2$, large κ to remove Elastic-net effect
- A simple dynamics acts as if nonconvex optimization...

Reference

- Osher, Ruan, Xiong, Yao, and Yin, *Sparse Recovery via Differential Equations*, arXiv:1406.7728
- Xu, Xiong, Huang, and Yao, *Robust Statistical Ranking: Theory and Algorithms*, arXiv:1408.3467

Acknowledgement

- Theory
  - *Stanley Osher*, *Wotao Yin*, UCLA
  - *Feng Ruan*, *Jiechao Xiong*, PKU
- Applications: *Ming Yan* UCLA; *Qianqian Xu*, *Chendi Huang* PKU
- Discussions: *Ming Yuan* U Wisconsin, *Lie Wang* MIT, *Peter Bickel* and *Bin Yu*, UCB
- Grants:
  - IPAM, National Basic Research Program of China (973 Program), NSFC