High-dimensional Inference via Lipschitz Sparsity-Yielding Regularizers

Zheng Pan Changshui Zhang
Department of Automation, Tsinghua University, Beijing 100084, China
State Key Lab of Intelligent Technologies and Systems
Tsinghua National Laboratory for Information Science and Technology (TNList)

Abstract

Non-convex regularizers are more and more applied to high-dimensional inference with sparsity prior knowledge. In general, the non-convex regularizer is superior to the convex ones in inference but it suffers the difficulties brought by local optima and massive computation. A "good" regularizer should perform well in both inference and optimization. In this paper, we prove that some non-convex regularizers can be such "good" regularizers. They are a family of sparsity-yielding penalties with proper Lipschitz subgradients. These regularizers keep the superiority of non-convex regularizers in inference. Their estimation conditions based on sparse eigenvalues are weaker than the convex regularizers. Meanwhile, if properly tuned, they behave like convex regularizers since standard proximal methods guarantee to give stationary solutions. These stationary solutions, if sparse enough, are identical to the global solutions. If the solution sequence provided by proximal methods is along a sparse path, the convergence rate to the global optimum is on the order of $1/k$ where $k$ is the number of iterations.

1 INTRODUCTION

High-dimensional inference concerns the parameter estimation problems in which the dimensions of parameters are comparable to or larger than the sampling size. In general, high-dimensional inference is ill-posed and it needs additional prior knowledge about the structure of the parameters to obtain consistent estimations. In recent years, tremendous research works have demonstrated that the prior knowledge of sparsity can lead to good estimators, e.g., the well-known work of Compressed Sensing (see Candès and Plan, 2011, and reference therein) and its extension to general M-estimators (Negahban et al., 2009).

In methodology, sparsity is usually imposed as a sparsity-yielding (or sparsity-encouraging (Candès et al., 2008), sparsity-inducing (Bach, 2010)) regularizers for M-estimators. Many regularizers have been proposed to describe the prior of sparsity, e.g., $l_0$-norm, $l_1$-norm, $l_q$-norm with $0 < q < 1$, Smoothly Clipped Absolute Deviation (SCAD) penalty (Fan and Li, 2001), Log-Sum Penalty (LSP) (Candès et al., 2008), Minimax Concave Penalty (MCP) (Zhang, 2010a). Note that these sparsity-yielding regularizers are non-convex except $l_1$-norm.

Non-convex regularizers usually need less samples or, rather, weaker estimation conditions for high-dimensional inference than convex ones (Candès et al., 2008; Foucart and Lai, 2009; Davies and Gribonval, 2009; Saab and Yilmaz, 2010; Sun, 2012). However, their non-convexity makes the corresponding M-estimators difficult to solve. It can not be guaranteed to achieve a global optimum for arbitrary non-convex regularizer. Even worse, it is strongly NP-hard for $l_q$-regularized M-estimators with $l_2$ loss functions (Chen et al., 2011). Without any additional condition, the methods based on gradient descent or Iterative Reweighted $l_1$ (IRL1) methods (Candès et al., 2008; Foucart and Lai, 2009; Zhang, 2010b) only lead to local optimums. Zhang (2010b) analyzed the inference performance of IRL1, but the analysis needs the same estimation condition as $l_1$-norm. Besides, IRL1 suffers the same RIP based failure bound as $l_1$-norm for inference (Davies and Gribonval, 2009) and the computational burden is heavy since it needs to solve several $l_1$-regularized problems. For MCP, Zhang (2010a) pro-
posed the sparse Riesz condition for his path following algorithm, but this condition is also strong enough for the consistency of parameter estimation by LASSO.

In this paper, we consider what regularizers are able to hold good performance in both inference and optimization. We propose a family of sparsity-yielding non-convex regularizers with Lipschitz subgradients, named **Lipschitz Sparsity-Yielding (LSY)** regularizers. We prove that LSY regularizers hold good properties in both inference and optimization. Specifically, we make the following contributions:

- We provide estimation conditions for LSY regularizers under the frameworks of **Sparse Eigenvalue** (Section 3). The proposed estimation conditions are weaker than that of $\ell_1$-norm, which implies that LSY regularizers need less samples or has higher probabilities for consistent parameter estimation than the convex regularizers. With the estimation conditions, we establish an upper bound of the estimation errors, as well as an estimation to the sparseness of the global solutions.

- With properly tuned LSY regularizers, we optimize the corresponding M-estimator by **Proximal Methods** (Nesterov, 2007; Beck and Teboulle, 2009, 2010) and it guarantees to achieve stationary solutions. These stationary solutions are identical to the global solutions if they are enough sparse. Furthermore, the convergence rate will be sublinear* provided that the solutions of all iterations are along a sparse path.

- We give a simple method to tune the parameter of LSY regularizers (Section 6). The method prevents the troubles in parameter selection for non-convex regularized M-estimators. Our experiments show the effectiveness of the parameter selection method on synthetic data.

- Our LSY regularizers cover many existing non-convex regularizers, e.g., SCAD (Fan and Li, 2001), LSP (Candès et al., 2008; Trzasko and Manduca, 2009), MCP (Zhang, 2010a) and Ge-man Penalty (GP) (Trzasko and Manduca, 2009; Geman and Yang, 1995) (Section 2). The good properties in inference and optimization proposed in this paper give a theoretical explanation for the outperformance of these non-convex regularizers in the previous works.

---

*For a function $F(\theta)$ with the minimal function value $\hat{F}$ and a sequence $\{\theta_k\}$, if $F(\theta_k) - \hat{F} \leq C/k$ for some constant $C > 0$, we say the convergence rate is sublinear.

## 2 LSY PENALTY

In this paper, we focus on the component-decomposable regularizers, i.e., the regularizer $r(u)$ can be written as $r(u) = \sum_{i=1}^{n} R(|u_i|)$, where $R(\cdot)$ is called the basis functions. The basis function of LSY penalty, named **CLIF**, is defined in Definition 1.

**Definition 1 (CLIF)** The function $R : [0, +\infty) \mapsto [0, +\infty)$ is called a Concave Lipschitz Increasing Function (CLIF) if it holds the following three properties:

1. Concave over $[0, +\infty)$ and $R(u) = 0 \Leftrightarrow u = 0$;
2. Increasing (or non-decreasing) over $(0, +\infty)$;
3. Lipschitz continuity: differentiable over $[0, +\infty)$ and there exists $\nu_R > 0$ such that for any $u_1, u_2 \in (0, +\infty)$, $|R'(u_1) - R'(u_2)| \leq \nu_R |u_1 - u_2|$, where $R'(\cdot)$ is the derivative of $R(\cdot)$. $\nu_R$ is called the Lipschitz constant of $R'(\cdot)$.

**Definition 2 (LSY penalty)** Let $R(\cdot)$ be any CLIF. For any $u = (u_1, \cdots, u_n) \in \mathbb{R}^n$, we define the LSY Sparsity-Yielding (LSY) penalty derived from $R(\cdot)$ as $r(u) = \sum_{i=1}^{n} R(|u_i|)$.

Some special cases of LSY penalties have been widely used in sparsity related works, e.g., the penalties in Table 1. All of the basis functions in Table 1, except $\ell_1$-norm, can be written as

$$R(u) = \lambda^2 R_0(u/\lambda; \gamma),$$

where the parameter $\gamma$ determines the Lipschitz constants and controls the “degrees” of non-convexity and approximation to $\ell_0$-norm. Formally, we have

$$\lim_{\gamma \to 0^+} \frac{r_\gamma(\theta)}{R_\gamma(1)} = \|\theta\|_0, \quad \lim_{\gamma \to +\infty} \frac{r_\gamma(\theta)}{R_\gamma(1)} = \|\theta\|_1.$$

for any $\theta \in \mathbb{R}^n$. In addition, it should be noted that $\ell_p$-norm with $0 \leq p < 1$ is not an LSY penalty, since we can not find a finite Lipschitz constant to ensure its Lipschitz continuity.

## 3 INFERENCE

Suppose we have $m$ samples

$$(y_1, a_1), (y_2, a_2), \cdots, (y_m, a_m),$$

where $y_i \in \mathbb{R}$ and $a_i \in \mathbb{R}^n$ for $i = 1, \cdots, m$. Let $X = (a_1, \cdots, a_m)^T$ and $y = (y_1, \cdots, y_m)$. We assume that there exists an underlying $s$-sparse parameter $\theta^*$ which is supported on $S$ and satisfies

$$y = X\theta^* + e$$

482
θ of sume all the magnitudes of the non-zero components To discriminate zero and non-zero components, we as-

\[ \hat{\theta} = \arg \min_{\theta \in \mathbb{R}^p} \mathcal{L}(\theta) + r(\theta) \]  

(3)

where

\[ \mathcal{L}(\theta) = \frac{1}{2m} \|y - X\theta\|_2^2 \]  

(4)

is the prediction error and \( r(\theta) \) is an LSY regularizer.

For the \( \ell_2 \) loss in Eqn. (4), Sparse Eigenvalues (SE) is a widely used framework for estimation conditions.

**Definition 3 (Sparse Eigenvalue)** For an integer \( t \geq 1 \), we say that \( \kappa_-(t) \) and \( \kappa_+(t) \) are the minimum and maximum sparse eigenvalue(SE) of a matrix \( X \) if

\[ \kappa_-(t) \leq \|X\theta\|_2^2/(m\|\theta\|_2^2) \leq \kappa_+(t) \]

holds for all \( \theta \) with \( \|\theta\|_0 \leq t \).

To discriminate zero and non-zero components, we assume all the magnitudes of the non-zero components of \( \theta^* \) and \( \hat{\theta} \) are not less than \( \rho \), i.e.,

\[ \rho = \min\{\hat{\rho}, \rho^*\} \geq 0, \]

where \( \hat{\rho} = \min\{|\hat{\theta}_i| : i \in \text{supp}({\hat{\theta}})\} \) and \( \rho^* = \min\{|\theta^*_i| : i \in \text{supp}(\theta^*)\} \). We call \( \hat{\rho} \) and \( \rho^* \) the zero gaps of \( \hat{\theta} \) and \( \theta^* \) respectively.

We observe that the solutions of Problem (3) are bounded. Thus, there exists \( \rho_\infty > 0 \) such that

\[ \max\{\|\hat{\theta}\|_\infty, \|\theta^*\|_\infty\} \leq \rho_\infty. \]

Based on SE, we establish the following parameter estimation theorem.

**Theorem 1** Suppose

\[ \frac{R^{-1}(u/s)}{R^{-1}(u/(t-1))} \]

is a non-decreasing function of \( u \) with \( t-1 \geq s \). If the SE satisfies

\[ \frac{\kappa_+(2t)}{\kappa_-(2t)} < 4\sqrt{2} - 1)G_R(\rho) + 1, \]

(5)

for some integer \( t \geq s + 1 \) and

\[ (2t-1) \frac{2\rho_\infty}{R(2\rho_\infty)} \leq 1, \]

we have

\[ \|\hat{\theta} - \theta^*\|_2 \leq C_1(R'(0) + \epsilon), \]

(6)

where \( C_1 \) are positive constants independent of \( \theta^* \) and \( \epsilon \) and \( G_R(\rho) \) is

\[ G_R(\rho) = \left\{ \begin{array}{ll} \sqrt{t} \frac{R^{-1}(R(\rho)/t)}{(t-1)\sqrt{s}}, & \rho > 0 \medskip \\ \frac{R^{-1}(R(\rho)/s)}{(t-1)}, & \rho = 0. \end{array} \right. \]

(7)

It can be verified that the LSY penalties in Table 1 satisfy the non-decreasing property of \( \frac{R^{-1}(u/s)}{R^{-1}(u/(t-1))} \).

If the zero gap \( \rho = 0 \), i.e., we can not distinguish the non-zero and zero component of the parameters, the estimation condition in Eqn. 5 becomes the same as the \( \ell_1 \)-norm case (Foucart and Lai, 2009). This is because when the non-zero components of parameters are scaled to be arbitrarily close to zero, the values of the regularizers at the scaled parameters can be approximated arbitrarily by a scaled \( \ell_1 \)-norm, which means the non-convexity vanishes. Without zero gap, the non-convex regularizers with finite gradients at zero will behave like an \( \ell_1 \)-norm for the parameters with small magnitudes and can not have different performance from \( \ell_1 \)-norm on sparse estimation in the meaning of the worst cases.

However, with positive zero gap \( \rho > 0 \), much weaker estimation conditions are available. For example, the SE based estimation condition for LSP becomes

\[ \frac{\kappa_+(2t)}{\kappa_-(2t)} < 1 + 4(\sqrt{2} - 1)\sqrt{\frac{s}{t} (1 + \rho/\gamma)^{1/s} - 1}. \]
With \( t > s + 1 \), the upper bound for \( \kappa_+(2t)/\kappa_-(2t) \) tends to infinity when \( \gamma/\rho \to 0 \). It means that if
\[
\kappa_-(2t) := \inf_{\theta} \left\{ \frac{\|X\theta\|_2^2}{\gamma(t \theta_0 - 2t)} : \|\theta\|_0 \leq 2t \right\} > 0,
\] there exists \( \gamma > 0 \) so that the SE condition in Eqn. (5) is satisfied for LSP. For \( \ell_1 \)-norm, the estimation conditions need \( \kappa_+(2t)/\kappa_-(2t) \) to be upper bounded by a fixed constant, e.g., \( \kappa_+(2t)/\kappa_-(2t) < 1 + 4(\sqrt{2} - 1)(t/s)^{1/2} \) (Foucart and Lai, 2009) and \( \kappa_+(2t)/\kappa_-(2t) < 4t/s - 1 \) (Zhang, 2010a; Zhang and Huang, 2008). Hence, the proposed estimation condition of LSP is much weaker than that of \( \ell_1 \)-norm. Note that \( \gamma \to 0 \) means that LSP tends to \( \ell_0 \)-norm and \( \kappa_-(2t) > 0 \) is the estimation condition for \( \ell_0 \)-norm (Foucart and Lai, 2009), which implies that our estimation conditions can be weakened "continuously" to that of \( \ell_0 \)-norm when LSP are more and more close to \( \ell_0 \)-norm.

In addition to the error bound, we usually hope that the M-estimators yield enough sparse solutions. Thus, estimating the sparseness of the global solutions is also an important problem. We extend the result from Zhang and Zhang (2012) and show that the global solutions of Problem (3) are sparse under appropriate conditions.

**Theorem 2** Suppose the conditions of Theorem 1 hold. Consider \( l_0 > 0 \) and integer \( m > 0 \) such that
\[
\sqrt{\frac{2\kappa_+(m_0)R(c_2(R'(0) + \epsilon))}{m_0}} + \frac{1}{m} \|X^T e\|_\infty < R'(l_0),
\]
where \( c_2 \) is a positive constant \( \dagger \). Then,
\[
|\supp(\hat{\theta}) \setminus \mathcal{S}| \leq m_0 + \frac{tR(c_2(R'(0) + \epsilon))}{R(l_0)},
\]
where \( R'(0) \) is the first derivative of \( R(e) \) at \( e = 0 \). Moreover, Problem (12) is convex and differentiable and its gradient is Lipschitz with Lipschitz constant \( \nu_L \), i.e., for all \( \theta, \theta' \in \mathbb{R}^n \),
\[
\|\nabla \mathcal{L}(\theta) - \nabla \mathcal{L}(\theta')\|_2^2 \leq \nu_L \|\theta - \theta'\|_2^2.
\]

**Corollary 1** Suppose the basis function has the formulation in Eqn. (1) and the conditions of Theorem 2 hold with \( l = 2s, m_0 = \beta_0 s, l_0 = \beta_1 \lambda \) and \( \lambda = \epsilon/\zeta \) for some \( \beta_0, \beta_1 \) and \( \zeta > 0 \). Let \( C_3 = c_2(R'(0) + \epsilon) \) where \( c_2 \) is the same as Theorem 2. If
\[
\frac{4\kappa_+(\beta_0 s)}{\beta_0} < \left( \frac{R'(0)}{R'_{\alpha_0}(C_3)} \right)^2 - \frac{\zeta^2}{R'(\beta_1)},
\]
then
\[
|\supp(\hat{\theta}) \setminus \mathcal{S}| \leq (\beta_0 + 2R_0(C_3)/R_0(\beta_1)) s.
\]

It is important that the global solutions are sparse, since it is one of the motivations of sparse learning. The sparseness of global solutions is also important for finding the global solutions, which will be stated in the next section.

### 4 Optimization Method

We provide optimization methods for Problem (3). In this section, the loss function \( \mathcal{L}(\theta) \) is not restricted to the \( \ell_2 \) loss in Eqn. (4). We assume the loss function \( \mathcal{L}(\theta) \) is convex and differentiable and its gradient is Lipschitz with Lipschitz constant \( \nu_L \), i.e., for all \( \theta, \theta' \in \mathbb{R}^n \),
\[
\|\nabla \mathcal{L}(\theta) - \nabla \mathcal{L}(\theta')\|_2 \leq \nu_L \|\theta - \theta'\|_2.
\]

#### 4.1 Proximal Methods

We use proximal methods to solve Problem (3). For convex regularizers, proximal methods have been proved to be efficient first order methods both theoretically and empirically (Beck and Teboulle, 2009; Agarwal et al., 2011). To apply it to non-convex regularizers, we need to tackle the following problem, named shrinkage function
\[
\tau_\omega(\theta) = \arg \min_{\theta} \left\{ \alpha r(\theta) + \frac{1}{2} \|\theta - \omega\|_2^2 \right\}
\]
for \( \omega \in \mathbb{R}^n \). Since the LSY penalty \( r(\theta) \) is component-decomposable, we can calculate each component of \( \tau_\omega(\theta) \) separately. We can even obtain the closed solutions of \( \tau_\omega(\theta) \) for the five LSY penalties in Table 1. For \( \ell_1 \)-norm, the shrinkage function is the well known soft-thresholding (Donoho, 1995)
\[
\tau_\omega(\theta) = \text{sign}(\omega) \cdot \max\{\omega - \alpha, 0\},
\]
where all the computation is component-wise. For LSP, the shrinkage function can be written as
\[
\tau_\omega(\theta) = \begin{cases} \frac{1}{2} \text{sign}(\omega) \cdot (|\omega| - \gamma) & \text{if } |\omega| > \lambda \alpha/\gamma, \\ \frac{1}{4} \sqrt{(|\omega| + \gamma)^2 - 4\alpha} & \text{if } |\omega| \leq \lambda \alpha/\gamma. \end{cases}
\]

In general, \( \tau_\omega(\theta) \) shrinks \( \omega \) to 0 if \( |\omega| \leq \alpha R'(0) \), which we call shrinkage effect. Moreover, Problem (12) is convex for suitable LSY penalties.

**Theorem 3** For any \( \omega \in \mathbb{R}^n \), Problem (12) is convex under the condition that \( 2\nu_L \) is less than 1.

The convexity of Problem (12) is useful when LSY penalty is too complicated to give a closed or simple solution for \( \tau_\omega(\theta) \). In this case, we can still obtain \( \tau_\omega(\theta) \) fast by one-dimensional convex optimization.
4.2 Global Optimality

We concern the convergence property of the standard proximal methods ISTA (Beck and Teboulle, 2009). ISTA is illustrated in Algorithm 1, where

\[ p_\nu(\theta_{k-1}) = \tau_1/\nu(\theta_{k-1} - \nabla L(\theta_{k-1}))/\nu, \]

and

\[ D_L(\theta; \theta') = L(\theta) - L(\theta') - \langle \nabla L(\theta'), \theta - \theta' \rangle. \]

**Algorithm 1 Iterative Shrinkage-Thresholding Algorithm (ISTA)**

**Input:** \( \nu_0 > 0, > 2\nu R \) if needed and \( \eta > 1, \gamma \) and \( \lambda \) are selected according to (14).

**Initialization:** \( \theta_0 = \arg \min L(\theta) + R'(0)\|\theta\|_1. \)

Step \( k \geq 1): \)

- Find the smallest integer \( i_k \geq 0 \) such that
  \[ \nabla p_\nu(\theta_{k-1}) \leq \frac{\nu}{2}\|p_\nu(\theta_{k-1}) - \theta_{k-1}\|_2^2 \] (13)
  with \( \nu = \eta i_k \nu k - 1 \).

- Set \( \nu_k = \eta i_k \nu k - 1 \) and compute
  \[ \theta_k = \tau_1/\nu_k \left( \theta_{k-1} - \frac{1}{\nu_k} \nabla L(\theta_{k-1}) \right). \]

The parameter \( \nu_0 \) can be tuned to ensure the convexity of Problem (12) if needed, i.e., \( \nu_0 \geq 2\nu R \). The parameters of LSY penalty \( \gamma \) and regularization factor \( \lambda \) can be selected simply by the parameter selection method in Section 5. Sparse initialization usually accelerates the algorithm, hence we use the solution of

\[ \min_{\theta} L(\theta) + R'(0)\|\theta\|_1 \]

as \( \theta_0 \), which can be solved efficiently by standard solvers for convex regularized problems, e.g., FISTA (Beck and Teboulle, 2009). The computation complexity of \( \nu_k \) and the shrinkage functions are similar to \( \ell_1 \)-regularized ISTA.

Before analyzing the convergence performance of LSY regularized ISTA, we first define the concept of approximate stationary solutions.

**Definition 4** Give \( \varphi \geq 0 \), we say that \( \bar{\theta} \) is a \( \varphi \)-approximate stationary \((\varphi-AS)\) solution of \( F(\theta) \) if the directional derivative of \( F \) at \( \bar{\theta} \) in any direction \( d \in \mathbb{R}^n \) is no less than \( -\varphi \), i.e.,

\[ F'(\bar{\theta}; d) \geq -\varphi. \]

**Lemma 1** \( \{ F(\theta_k) \} \) is a decreasing sequence and converges; For any \( \varphi > 0, \{ \theta_k \} \) gives a \( \varphi \)-AS solution of Problem (3) within finite steps.

Considering the prior of sparsity, if the loss function is restricted to operate on sparse vectors, it behaves like a strongly convex function, e.g., the works on compressed sensing with RIP conditions (Candes and Tao, 2005; Candes and Plan, 2011) and the general loss with RSC conditions (Negahban et al., 2009). With a positive minimum SE, the \( \ell_2 \) loss in Eqn. (4) also behaves like a strongly convex function. Hence, we propose SSC conditions which extend SE to general loss functions.

**Definition 5 (SSC)** We say \( L(\theta) \) holds the property of \((s, s')\)-Sparse Strong Convexity (SSC) if there exists a constant \( \kappa_-(s, s') > 0 \) such that

\[ L(a\theta + (1 - a)\theta') \leq a L(\theta) + (1 - a) L(\theta') - a(1 - a) \kappa_-(s, s') \| \theta - \theta' \|_2^2 \]

holds for any \( a \in (0, 1) \), any \( s \)-sparse vector \( \theta \) and any \( s' \)-sparse vector \( \theta' \).

For the case of \( L(\theta) = \frac{1}{m} \| y - X\theta \|_2^2 \), SSC becomes the same as the minimum SE, i.e.,

\[ 2\kappa_-(s, s') = \kappa_-(s + s'). \]

The more sparse \( \theta \) and \( \theta' \) are, the more probably SSC holds.

**Theorem 4** Suppose \( \bar{\theta} \) is \( \varphi \)-AS and \( q \)-sparse.

1. \( \mathcal{F}(\bar{\theta}) \leq \mathcal{F}(\theta^*) + \varphi \) if \( \nu R \leq \kappa_-(q, s); \)
2. \( \mathcal{F}(\bar{\theta}) \leq \mathcal{F}(\theta) + \varphi \) if \( \nu R \leq \kappa_-(q, \hat{s}), \) where \( \hat{s} = \text{supp}(\bar{\theta}). \)

With Lemma 1, the \( \varphi \)-AS solutions required in Theorem 4 can be obtained by ISTA. Theorem 4.1 implies that the stationary solutions \((\varphi = 0)\) achieve better function values of \( \mathcal{F}(\theta) \) than \( \theta^* \), which is enough to let Theorem 1 be applicable for these stationary solutions, i.e., under the conditions of Theorem 1, the parameter estimation by these stationary solutions also hold the error bound in Eqn. (6).

Theorem 2 and Corollary 1 show the global solutions are sparse under appropriate conditions. With the guarantees of sparse global solutions, the sparse stationary solutions are also global solutions by Theorem 4.2.

Theorem 4 analyzes the the performance of a single sparse solution on approximating the global optimum \( \mathcal{F}(\bar{\theta}) \). We also concern the performance of a sparse solution sequence \( \{ \theta_k \}_{k=0}^{\infty} \) given by ISTA.

**Theorem 5** Suppose \( \{ \theta_k \}_{k=0}^{\infty} \) is generated by ISTA. Let \( q_{k_0, k} = \max \{ \| \theta_i \|_0, k_0 \leq i \leq k \} \) for any positive
This page contains a discussion on parameter selection for sparse recovery problems. It mentions the importance of choosing appropriate parameters \( \lambda \) and \( \gamma \) for techniques like ISTA and LSY regularizers. The page includes mathematical formulations and theorems, such as Theorems 5, which require the solution sequence to be stable and sparse. The text discusses the convergence rate of ISTA for different regularizers and the significance of parameters in ensuring that the solution sequence behaves as expected. It also highlights the need for good approximation to the original vector's \( \ell_0 \)-norm under certain conditions. The optimization section outlines methods for selecting these parameters, emphasizing the importance of empirical validation.
Objective Function

Sparsity

780
15
10.8
10.8
800
15.5
10.9
16
10.9
850
820
13.5
1000
900
17
16.5
13.5
900
1100
840
11.1
18
950
17.5
1100
11.1

Figure 1: Boxplots of the sparsity (top) and objective function (bottom) for MC (a)(b), Geman (c)(d) and LSP (e)(f). The columns of (a)(c)(e) and (b)(d)(f) are for ISTA and IRL1 respectively. We set $\beta = 0.5$ and run 100 trials for each regularizer.

$\theta^*$. We use Support Recovery Ratio (SRR)

$$r = \frac{|S \cap T|}{|S \cup T|}$$

to indicate the performance of support recovery. The supports of $\hat{\theta}$ and $\theta^*$ are the same iif SRR $r = 1$.

We run 100 trials for each regularizer and each sampling size. Figure 2(a) illustrates the frequencies of successful support recovery (SRR $r = 1$). MCP, GP and LSP have almost the same SRR as $\ell_{0.5}$, $\ell_{0.2}$ and $\ell_{0.1}$ respectively for small sampling sizes ($\beta \leq 0.4$). Note that all the three LSY regularizers can stabilize SRRs at 1 when $\beta \geq 0.41$. However, the SRRs of $\ell_{0.5}$-regularizer, minimized by IRL1, can not give exact support recovery stably, since the solutions given by IRL1 methods usually have small noise for some components that should have been zero.

In Figure 2(b), we compare the average Relative Recovery Error (RRE)

$$\text{RRE} = \frac{\|\hat{\theta} - \theta^*\|_2}{\|\theta^*\|_2}$$

for different regularizers and different sampling sizes. The three LSY regularizers can reduce RRE to a low level ($< 5\%$) with similar sampling sizes to the three $\ell_p$-norms. However, the LSY regularizers can give smaller RREs than $\ell_p$-norms.

The smaller RREs, as well as more stable SRRs, show that LSY regularizers are more robust to noise than $\ell_p$-norms.

6 CONCLUSION

We have presented the theoretical analysis for LSY regularizers in high-dimensional inference with the prior knowledge of sparsity. The proposed SE based estimation conditions are weaker than that of convex regularizers. Proximal methods provide stationary solutions of LSY regularized regression, which are identical to the global optimums with sparseness assumptions. With the parameter selection method in Section 6, we can avoid the exhausted search for good parameters. The experiments on synthetic examples demonstrate a numerical confirmation for the outperformance of LSY regularizers in optimization and inference.

This paper provides a general theory to non-convex high-dimensional inference with sparsity priors and can serve as a guideline for selecting regularizers and developing algorithms for non-convex regularized regression.

Acknowledgements

Supported by 973 Program (2013CB329503) and NSFC (Grant No. 9112030161075004, 61021063).

References


Figure 2: The comparison of SRR (left) and RRE (right) between $\ell_1$ (FISTA), $\ell_0$ (OMP), $\ell_p$ (IRL1, $p = 0.1, 0.2, 0.5$), LSP, MC and Geman.


