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 optimums 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 Candes 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_p-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-

Appearing in Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS) 2013, Scottsdale, AZ, USA. Volume XX of JMLR: W&CP XX. Copyright 2013 by the authors.

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 ℓ_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 parameters 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 Geman 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.

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 $R(\cdot)$. ν_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 Lipschitz 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 ℓ_1-norm, can be written as

$$R(u) = \lambda^2 R_0(u/\lambda; \gamma), \tag{1}$$

where the parameter γ determines the Lipschitz constants and controls the "degrees" of non-convexity and approximation to ℓ_0-norm. Formally, we have

$$\lim_{\gamma \to 0^+} \frac{r_\gamma(\theta)}{R_\gamma(1)} = \|\theta\|_0, \quad \lim_{\gamma \to +\infty} \frac{r_\gamma(\theta)}{R_\gamma(1)} = \|\theta\|_1. \tag{2}$$

for any $\theta \in \mathbb{R}^n$. In addition, it should be noted that ℓ_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 θ^* which is supported on $\mathcal{S}$ and satisfies

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

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

Table 1: Examples of LSY penalties.

Penalty	**CLIF**	**Lipschitz constant**
ℓ_1-norm	$R(u) = \lambda u$	$\nu_R > 0$
SCAD	$R(u) = \lambda \int_0^u \min\left\{1, \left(1 - \frac{x/\lambda - 1}{\gamma}\right)_+\right\} dx$	$\nu_R = 1/(2\gamma)$
MCP	$R(u) = \lambda \int_0^u \left(1 - \frac{x}{\lambda\gamma}\right)_+ dx$	$\nu_R = 1/(2\gamma)$
LSP	$R(u) = \lambda^2 \log\left(1 + \frac{u}{\lambda\gamma}\right)$	$\nu_R = 1/\gamma^2$
GP	$R(u) = \lambda^2 u/(\lambda\gamma + u)$	$\nu_R = 2/\gamma^2$

with a small noise $e \in \mathbb{R}^n$. In this paper, we assume that the noise satisfies

$$\|X^T e/m\|_\infty \le \epsilon$$

for some $\epsilon \ge 0$.

We focus on using the following regularized regression to recover θ^* from y.

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^p} \mathcal{L}(\theta) + r(\theta) \tag{3}$$

where

$$\mathcal{L}(\theta) = \frac{1}{2m}\|y - X\theta\|_2^2 \tag{4}$$

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

For the ℓ_2 loss in Eqn. (4), *Sparse Eigenvalues* (SE) is a widely used framework for estimation conditions.

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

$$\kappa_-(t) \le \|X\theta\|_2^2/(m\|\theta\|_2^2) \le \kappa_+(t)$$

holds for all θ with $\|\theta\|_0 \le t$.

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

$$\rho = \min\{\hat{\rho}, \rho^*\} \ge 0,$$

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

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

$$\max\{\|\hat{\theta}\|_\infty, \|\theta^*\|_\infty\} \le \rho_\infty.$$

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 $t - 1 \ge s$. If the SE satisfies

$$\kappa_+(2t)/\kappa_-(2t) < 4(\sqrt{2} - 1)G_R(\rho) + 1, \tag{5}$$

for some integer $t \ge s + 1$ and

$$(2t-1)\frac{2\rho_\infty}{R(2\rho_\infty)}\epsilon \le 1,$$

we have

$$\|\hat{\theta} - \theta^*\|_2 \le C_1(R'(0) + \epsilon), \tag{6}$$

where C_1 are positive constants independent of θ^* and ϵ and $G_R(\rho)$ is

$$G_R(\rho) = \begin{cases} \sqrt{\frac{s}{t}}\frac{R^{-1}(R(\rho)/s)}{R^{-1}(R(\rho)/(t-1))}, & \rho > 0 \\ (t-1)/\sqrt{st}, & \rho = 0. \end{cases} \tag{7}$$

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 ℓ_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 ℓ_1-norm, which means the non-convexity vanishes. Without zero gap, the non-convex regularizers with finite gradients at zero will behave like an ℓ_1-norm for the parameters with small magnitudes and can not have different performance from ℓ_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

$$\frac{\kappa_+(2t)}{\kappa_-(2t)} < 1 + 4(\sqrt{2} - 1)\sqrt{\frac{s}{t}}\frac{(1 + \rho/\gamma)^{1/s} - 1}{(1 + \rho/\gamma)^{1/(t-1)} - 1}.$$

With $t > s+1$, the upper bound for $\kappa_+(2t)/\kappa_-(2t)$ tends to infinity when $\gamma/\rho \to 0$. It means that if

$$\kappa_-(2t) := \inf_\theta \left\{ \frac{\|X\theta\|_2^2}{n\|\theta\|_2^2} : \|\theta\|_0 \le 2t \right\} > 0, \qquad (8)$$

there exists $\gamma > 0$ so that the SE condition in Eqn. (5) is satisfied for LSP. For ℓ_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 ℓ_1-norm. Note that $\gamma \to 0$ means that LSP tends to ℓ_0-norm and $\kappa_-(2t) > 0$ is the estimation condition for ℓ_0-norm (Foucart and Lai, 2009), which implies that our estimation conditions can be weakened "continuously" to that of ℓ_0-norm when LSP are more and more close to ℓ_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 $m_0 > 0$ such that*

$$\sqrt{\frac{2t\kappa_+(m_0)R(c_2(R'(0)+\epsilon))}{m_0}} + \frac{1}{m}\|X^T e\|_\infty < R'(l_0),$$

where c_2 is a positive constant [†]. *Then,*

$$|supp(\hat{\theta})\backslash\mathcal{S}| \le m_0 + \frac{tR(c_2(R'(0)+\epsilon))}{R(l_0)},$$

Corollary 1 *Suppose the basis function has the formulation in Eqn. (1) and the conditions of Theorem 2 hold with $t = 2s$, $m_0 = \beta_0 s$, $l_0 = \beta_1\lambda$ and $\lambda = \epsilon/\zeta$ for some β_0, β_1 and $\zeta > 0$. Let $C_3 = c_2(R_0'(0)+\zeta)$ where c_2 is the same as Theorem 2. If*

$$\frac{4\kappa_+(\beta_0 s)}{\beta_0} < \frac{(R_0'(\beta_1)-\zeta)^2}{R_0(C_3)}, \qquad (9)$$

then

$$|supp(\hat{\theta})\backslash\mathcal{S}| \le (\beta_0 + 2R_0(C_3)/R_0(\beta_1))s. \qquad (10)$$

Theorem 2 implies that the global solution $\hat{\theta}$ is sparse under appropriate conditions. Corollary 1 further shows that the sparseness of $\hat{\theta}$ can be $O(s)$:

$$\hat{s} := |\mathrm{supp}(\hat{\theta})| \le \left(1 + \beta_0 + \frac{2R_0(C_3)}{R_0(\beta_1)}\right)s. \qquad (11)$$

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 ℓ_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 $\nu_{\mathcal{L}}$, i.e., for all $\theta, \theta' \in \mathbb{R}^n$,

$$\|\nabla\mathcal{L}(\theta) - \nabla\mathcal{L}(\theta')\|_2^2 \le \nu_{\mathcal{L}}\|\theta - \theta'\|_2^2.$$

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*

$$\tau_\alpha(\omega) = \arg\min_\theta \left\{ \alpha r(\theta) + \frac{1}{2}\|\theta - \omega\|_2^2 \right\} \qquad (12)$$

for $\omega \in \mathbb{R}^n$. Since the LSY penalty $r(\theta)$ is component-decomposable, we can calculate each component of $\tau_\alpha(\omega)$ separately. We can even obtain the closed solutions of $\tau_\alpha(\omega)$ for the five LSY penalties in Table 1. For ℓ_1-norm, the shrinkage function is the well known *soft-thresholding* (Donoho, 1995)

$$\tau_\alpha(\omega) = \mathrm{sign}(\omega) \cdot \max\{\omega - \alpha, 0\},$$

where all the computation is component-wise. For LSP, the shrinkage function can be written as

$$\tau_\alpha(\omega) = \begin{cases} \frac{1}{2}\mathrm{sign}(\omega)\cdot(|\omega| - \gamma & \\ \qquad + \sqrt{(|\omega|+\gamma)^2 - 4\alpha}\Big), & |\omega| \ge \lambda\alpha/\gamma, \\ 0, & |\omega| < \lambda\alpha/\gamma. \end{cases}$$

In general, $\tau_\alpha(\omega)$ shrinkages ω to 0 if $|\omega| \le \alpha R'(0)$, 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 $2\alpha\nu_R \le 1$.*

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

[†] c_2 is defined in Eqn. (25) of the supplementary material

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

$$p_{\bar{\nu}}(\theta_{k-1}) = \tau_{1/\bar{\nu}}(\theta_{k-1} - \nabla\mathcal{L}(\theta_{k-1})/\bar{\nu}),$$

and

$$\mathcal{D}_{\mathcal{L}}(\theta;\theta') = \mathcal{L}(\theta) - \mathcal{L}(\theta') - \langle\nabla\mathcal{L}(\theta'), \theta - \theta'\rangle.$$

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:** $\theta_0 = \arg\min \mathcal{L}(\theta) + R'(0)\|\theta\|_1$.
Step $k$ ($k \geq 1$):
    Find the smallest integer $i_k \geq 0$ such that
        $\mathcal{D}(p_{\bar{\nu}}(\theta_{k-1}), \theta_{k-1}) \leq \frac{\bar{\nu}}{2}\|p_{\bar{\nu}}(\theta_{k-1}) - \theta_{k-1}\|_2^2$    (13)
    with $\bar{\nu} = \eta^{i_k}\nu_{k-1}$.
    Set $\nu_k = \eta^{i_k}\nu_{k-1}$ and compute
        $\theta_k = \tau_{1/\nu_k}\left(\theta_{k-1} - \frac{1}{\nu_k}\nabla\mathcal{L}(\theta_{k-1})\right).$
```

The parameter ν_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 γ and regularization factor λ can be selected simply by the parameter selection method in Section 5. Sparse initialization usually accelerates the algorithm, hence we use the solution of

$$\min_{\theta} \mathcal{L}(\theta) + R'(0)\|\theta\|_1$$

as θ_0, which can be solved efficiently by standard solvers for convex regularized problems, e.g., FISTA (Beck and Teboulle, 2009). The computation complexity of ν_k and the shrinkage functions are similar to ℓ_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 $\tilde{\theta}$ is a φ-approximate stationary (φ-AS) solution of $\mathcal{F}(\theta)$ if the directional derivative of $\mathcal{F}$ at $\tilde{\theta}$ in any direction $d \in \mathbb{R}^n$ is no less than $-\varphi$, i.e.,*

$$\mathcal{F}'(\theta; d) \geq -\varphi.$$

Lemma 1 *$\{\mathcal{F}(\theta_k)\}$ is a decreasing sequence and converges; For any $\varphi > 0$, $\{\theta_k\}$ gives a φ-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 (Candès and Tao, 2005; Candes and Plan, 2011) and the general loss with RSC conditions (Negahban *et al.*, 2009). With a positive minimum SE, the ℓ_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 $\mathcal{L}(\theta)$ holds the property of (s, s')-Sparse Strong Convexity (SSC) if there exists a constant $\kappa_-(s, s') > 0$ such that*

$$\begin{aligned}&\mathcal{L}(a\theta + (1-a)\theta')\\ \leq& a\mathcal{L}(\theta) + (1-a)\mathcal{L}(\theta') - a(1-a)\kappa_-(s,s')\|\theta - \theta'\|_2^2\end{aligned}$$

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

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

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

The more sparse θ and θ' are, the more probably SSC holds.

Theorem 4 *Suppose $\tilde{\theta}$ is φ-AS and q-sparse.*

1. *$\mathcal{F}(\tilde{\theta}) \leq \mathcal{F}(\theta^*) + \varphi$ if $\nu_R \leq \kappa_-(q, s)$;*
2. *$\mathcal{F}(\tilde{\theta}) \leq \mathcal{F}(\hat{\theta}) + \varphi$ if $\nu_R \leq \kappa_-(q, \hat{s})$, where $\hat{s} = \mathrm{supp}(\hat{\theta})$.*

With Lemma 1, the φ-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 θ^*, 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 $\mathcal{F}(\hat{\theta})$. 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* *integers k_0 and k with $k > k_0$. Let $\hat{s} = \text{supp}(\hat{\theta})$. If $\kappa_-(q_{k_0,k}, \hat{s}) > 0$ and $\nu_R \leq \kappa_-(q_{k_0,k}, \hat{s})$, we have*

$$\|\theta_{k+1} - \hat{\theta}\|_2 \leq \|\theta_k - \hat{\theta}\|_2$$

and

$$\mathcal{F}(\theta_k) - \mathcal{F}(\hat{\theta}) \leq \frac{(\eta\nu_{\mathcal{L}} - 2\nu_R)\|\hat{\theta} - \theta_{k_0}\|_2^2}{2(k - k_0)}$$

for any $k > k_0$.

Theorem 5 requires the solution sequence $\{\theta_j\}_{j=k_0}^k$ is along a sparse path, i.e., every solution in the sequence is sparse. Along a sparse path, the solution sequence holds a sublinear convergence rate to the global optimums.

In practice, the shrinkage effect of ISTA usually drives θ_k to become sparse rapidly and keep sparse stably, especially for non-convex regularizers which are close to ℓ_0-norm. The experiments in Section 5 also demonstrate that empirically. With this experience as an assumption, Theorem 5 shows that the convergence rate of ISTA with non-convex regularizers are the same as the rate with convex regularizers (Beck and Teboulle, 2009), which explains the efficiency of non-convex regularized ISTA in many practical experience.

5 EXPERIMENT

In this section, we provide some brief experiments to confirm the proposed theory. More complicated experiments and applications of LSY regularizers have been demonstrated in some previous work, e.g., Candès *et al.* (2008); Fan and Li (2001); Trzasko and Manduca (2009); Zhang (2010a). The model is introduced in Section 3, where θ^* is s-sparse and the designed matrix X has i.i.d. elements drawn from $N(0, 1)$. The dimensions of the model are set as $n = 10,000$, $s = \log^3 n$ and $m = \beta s \log n$ where $\beta > 0$ is varied to control the sampling size. The observation noise e is drawn from $N(0, \sigma^2 I)$ and it follows that

$$\frac{1}{m}\|X^T e\|_\infty \leq \epsilon = 2\sigma\sqrt{\frac{\log n}{m}}$$

with high probability. The original sparse vectors are generated with random ±1 for non-zero components.

5.1 Parameter Selection Method

The conditions of Theorem 1, 4 and 5 put a constraint on the choice of the parameters λ and γ:

$$\frac{\kappa_-}{2\nu_R} \geq 1 \geq (2t - 1)\frac{2\rho_\infty}{R(2\rho_\infty)}\epsilon, \tag{14}$$

where $\kappa_- = \kappa_-(q + s)$, $\kappa_-(q + \hat{s})$ or $\kappa_-(q_{k_0,k} + \hat{s})$. Hence, we use it to select the parameters of LSY regularizers.

In the following experiments, we consider three LSY penalties: MCP, GP and LSP. Their Lipschitz constants are listed in Table 1. We set $\rho_\infty = 1$, $t = s + 1$ and $\kappa_- = 0.5$ in our experiment[‡]. Hence, we use

$$\frac{1}{4\nu_R} = 1 = \frac{4(2s + 1)\sigma}{\lambda^2 R_0(2/\lambda; \gamma)}\sqrt{\frac{\log n}{m}}$$

to select λ and γ. If the noise is close to zero, λ can be small enough to give a good approximation to ℓ_0-norm[§]. Hence, our parameter selection method is particularly suitable for the cases with low noise level ($\sigma_m = \sigma/\sqrt{m} = 10^{-6}$ in our experiment) and small sampling size (small β).

5.2 Optimization

The sublinear convergence in Theorem 5 assumes that the sequence $\{\theta_k\}$ becomes sparse rapidly and $\|\theta_k\|_0$ keeps small stably. Figure 1 illustrates the decrease of of non-zero numbers and function values. In Figure 1, the non-zero numbers may increase at first iteration, but after 4 ∼ 5 iterations, the sparsity becomes the same as that of the original vector ($\log^3 n$) and the function values converge to be almost the same. The even columns show the corresponding results of IRL1 methods, which usually need 2 iterations for the convergence of sparsity and function value. It should be noted that an iteration of ISTA only means a shrinkage function while an iteration of IRL1 means a solver for ℓ_1-regularized problems, e.g., FISTA. Hence, ISTA is more efficient for LSY regularizers than IRL1.

5.3 High-dimensional Inference

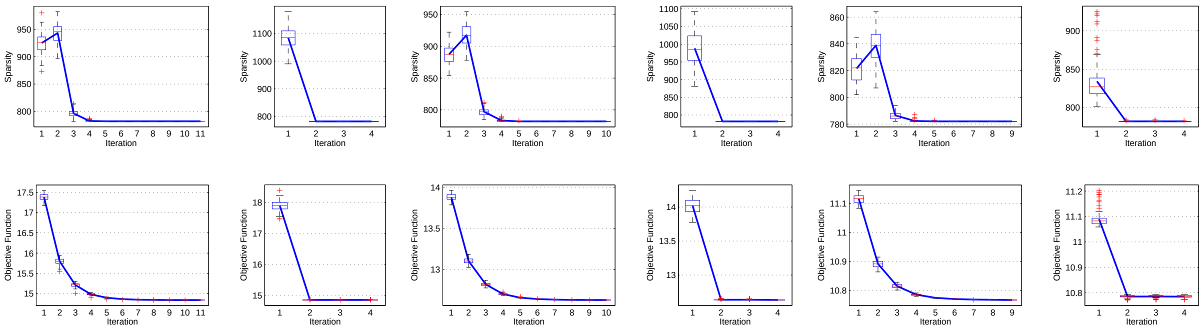
We compare the performance of sparsity inference between the three LSY regularizers and popular sparse regularizers (ℓ_1, ℓ_0, ℓ_p with $p = 0.1, 0.2, 0.5$). We minimize ℓ_1, ℓ_0 and ℓ_p-regularized problems with FISTA ($\lambda = 0.03$) (Beck and Teboulle, 2009), OMP (Tropp and Gilbert, 2007) and IRL1 ($\lambda = \mu\phi^{1-p}/p\|X^T y\|_\infty$ where $\mu = 0.1, \phi = 0.01$) [¶] methods respectively.

First, we concern the problems of support recovery, i.e., whether the recovered vector $\hat{\theta}$ has the same support as θ^*. Denote $\mathcal{T}$ and $\mathcal{S}$ as the supports of $\hat{\theta}$ and

[‡] For Gaussian matrix, $\kappa_- = 0.5$ holds with high probability when θ and θ' in Definition 5 are sparse.

[§] MCP, GP and LSY tend to ℓ_0-norm also when $\lambda \to 0$.

[¶] IRL1 uses $(|\theta| + \phi)^p$ to approximate ℓ_p-norm with small ϕ. As far as we know, there is still no widely accepted rule to choose λ for ℓ_p-regularized problems. We varied μ and find $\mu = 0.1$ has a good balance between avoiding zero solution and yielding enough sparse solutions.



(a) (b) (c) (d) (e) (f)

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.

θ^*. We use *Support Recovery Ratio (SRR)*

$$r = |\mathcal{S} \cap \mathcal{T}| / |\mathcal{S} \cup \mathcal{T}|$$

to indicate the performance of support recovery. The supports of $\hat{\theta}$ and θ^* 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)*

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

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 ℓ_p-norms. However, the LSY regularizers can give smaller RREs than ℓ_p-norms.

The smaller RREs, as well as more stable SRRs, show that LSY regularizers are more robust to noise than ℓ_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).

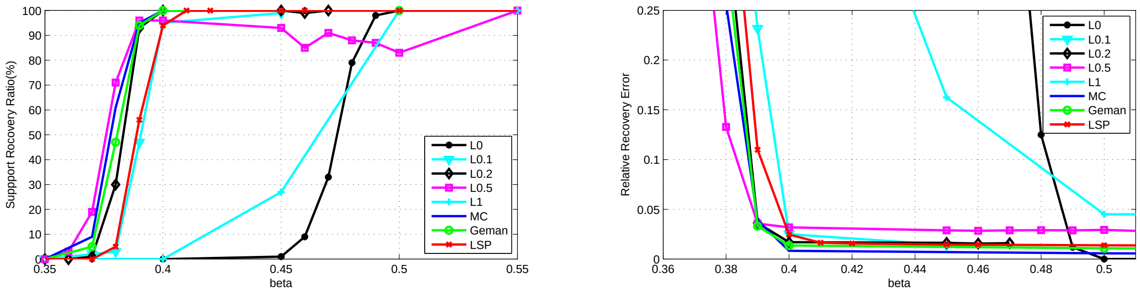
References

Agarwal, A., Negahban, S., and Wainwright, M. (2011). Fast global convergence of gradient methods for high-dimensional statistical recovery. *Arxiv preprint arXiv:1104.4824*.

Bach, F. (2010). Structured sparsity-inducing norms through submodular functions. *NIPS 2010*.

Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, **2**(1), 183–202.

Beck, A. and Teboulle, M. (2010). Gradient-based algorithms with applications to signal-recovery problems. *Convex Optimization in Signal Processing and Communications*, pages 42–88.



(a) SRR

(b) RRE

Figure 2: The comparison of SRR (left) and RRE (right) between ℓ_1(FISTA), ℓ_0(OMP), ℓ_p (IRL1, $p = 0.1, 0.2, 0.5$), LSP, MC and Geman.

Candes, E. and Plan, Y. (2011). A probabilistic and ripless theory of compressed sensing. *Information Theory, IEEE Transactions on*, **57**(11), 7235 –7254.

Candès, E. and Tao, T. (2005). Decoding by linear programming. *IEEE Transactions on Information Theory*, **51**(12), 4203–4215.

Candès, E., Wakin, M., and Boyd, S. (2008). Enhancing sparsity by reweighted ℓ_1 minimization. *Journal of Fourier Analysis and Applications*, **14**(5), 877–905.

Chen, X., Ge, D., Wang, Z., and Ye, Y. (2011). Complexity of unconstrained l2-lp minimization. *Arxiv preprint arXiv:1105.0638*.

Davies, M. and Gribonval, R. (2009). Restricted isometry constants where ℓ^p sparse recovery can fail for $0 < p \leq 1$. *Information Theory, IEEE Transactions on*, **55**(5), 2203 –2214.

Donoho, D. (1995). De-noising by soft-thresholding. *Information Theory, IEEE Transactions on*, **41**(3), 613–627.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**(456), 1348–1360.

Foucart, S. and Lai, M. (2009). Sparsest solutions of underdetermined linear systems via lq-minimization for $0 < q \leq 1$. *Applied and Computational Harmonic Analysis*, **26**(3), 395–407.

Geman, D. and Yang, C. (1995). Nonlinear image recovery with half-quadratic regularization. *Image Processing, IEEE Transactions on*, **4**(7), 932–946.

Negahban, S., Ravikumar, P., Wainwright, M., and Yu, B. (2009). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *NIPS 2009*.

Nesterov, Y. (2007). *Gradient methods for minimizing composite objective function*. CORE.

Saab, R. and Yilmaz, O. (2010). Sparse recovery by non-convex optimization - instance optimality. *Applied and Computational Harmonic Analysis*, **29**(1), 30 – 48.

Sun, Q. (2012). Recovery of sparsest signals via ℓ_q-minimization. *Applied and Computational Harmonic Analysis*, **32**(3), 329 – 341.

Tropp, J. and Gilbert, A. (2007). Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory*, **53**(12), 4655–4666.

Trzasko, J. and Manduca, A. (2009). Relaxed conditions for sparse signal recovery with general concave priors. *IEEE Transactions on Signal Processing*, **57**(11), 4347–4354.

Zhang, C. (2010a). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, **38**(2), 894–942.

Zhang, C. and Huang, J. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics*, **36**(4), 1567–1594.

Zhang, C. and Zhang, T. (2012). A general theory of concave regularization for high dimensional sparse estimation problems. *Statistical Science*.

Zhang, T. (2010b). Analysis of multi-stage convex relaxation for sparse regularization. *The Journal of Machine Learning Research*, **11**, 1081–1107.

7 APPENDIX

In this appendix, we use $\theta[i]$ to denote the i-th component of a vector θ. Let $\mathcal{S} = \text{supp}(\theta^*)$, $\Delta = \hat{\theta} - \theta^*$ and $\mathcal{T}$ be any index set with $|\mathcal{T}| \leq s$. We use $\bar{\mathcal{T}}$ to denote the complement of $\mathcal{T}$. Let $i_1, i_2, \cdots$ be a sequence of decreasing indices for $\bar{\mathcal{S}}$, i.e., $i_k \in \bar{\mathcal{S}}$ for $k \geq 1$ and

$$|\Delta[i_1]| \geq |\Delta[i_2]| \geq |\Delta[i_3]| \geq \cdots,$$

Given an integer $t \geq s$, we partition $\bar{\mathcal{S}}$ as $\bar{\mathcal{S}} = \cup_{i \geq 1} \mathcal{S}_i$ such that $\mathcal{S}_1 = \{i_1, \cdots, i_t\}$, $\mathcal{S}_2 = \{i_{t+1}, \cdots, i_{2t}\}$, $\cdots$. Define $\Sigma = \sum_{i \geq 2} \|\Delta_{\mathcal{S}_i}\|_2$.

7.1 Proof of Theorem 1

First, we introduce the following three lemmas.

Lemma 2 *For any* $\alpha > 1$ *and* $\frac{\alpha+1}{\alpha-1} \cdot \frac{2\rho_\infty}{R(2\rho_\infty)} \|\nabla \mathcal{L}(\theta^*)\|_\infty \leq 1$*, it holds that*

$$r(\Delta_{\bar{\mathcal{S}}}) + \frac{1+\alpha}{4m} \|X\Delta\|_2^2 \leq \alpha r(\Delta_{\mathcal{S}}). \tag{15}$$

Proof. Since $\hat{\theta} = \theta^* + \Delta$ minimizes $\mathcal{L}(\theta) + r(\theta)$, we have

$$0 \geq \mathcal{L}(\theta^* + \Delta) - \mathcal{L}(\theta^*) + r(\theta^* + \Delta) - r(\theta^*) \tag{16}$$

Since $\mathcal{L}(\theta) = \|y - X\theta\|_2^2/(2m)$, $\mathcal{L}(\theta^* + \Delta) - \mathcal{L}(\theta^*) = \langle \nabla \mathcal{L}(\theta^*), \Delta \rangle + \|X\Delta\|_2^2/(2m)$. Substituting it to (16), we have

$$0 \geq \frac{1}{2m}\|X\Delta\|_2^2 + \langle \nabla \mathcal{L}(\theta^*), \Delta \rangle + r(\theta^* + \Delta) - r(\theta^*). \tag{17}$$

For the first term of (17)'s right hand, we have

$$\begin{aligned} \langle \nabla \mathcal{L}(\theta^*), \Delta \rangle &\geq -\|\nabla \mathcal{L}(\theta^*)\|_\infty \|\Delta\|_1 \\ &\geq -\|\nabla \mathcal{L}(\theta^*)\|_\infty \sum_{i=1}^{n} \frac{|\Delta[i]|}{R(|\Delta[i]|)} \cdot R(|\Delta[i]|) \\ &\geq -\frac{2\rho_\infty}{R(2\rho_\infty)} \|\nabla \mathcal{L}(\theta^*)\|_\infty r(\Delta) \\ &\geq -\frac{\alpha - 1}{\alpha + 1} r(\Delta). \end{aligned} \tag{18}$$

Note that

$$\begin{aligned} r(\theta^* + \Delta) - r(\theta^*) &= r(\theta^*_{\mathcal{S}} + \Delta_{\mathcal{S}}) + r(\Delta_{\bar{\mathcal{S}}}) - r(\theta^*) \\ &\geq r(\Delta_{\bar{\mathcal{S}}}) - r(\Delta_{\mathcal{S}}). \end{aligned} \tag{19}$$

Substituting (18) and (19) to (17), we have

$$\begin{aligned} 0 &\geq -\frac{\alpha-1}{\alpha+1} r(\Delta) + r(\Delta_{\bar{\mathcal{S}}}) - r(\Delta_{\mathcal{S}}) + \frac{1}{2m}\|X\Delta\|_2^2 \\ &= \frac{2}{\alpha+1} r(\Delta_{\bar{\mathcal{S}}}) - \frac{2\alpha}{\alpha+1} r(\Delta_{\mathcal{S}}) + \frac{1}{2m}\|X\Delta\|_2^2. \ \square \end{aligned} \tag{20}$$

Lemma 3

$$R(\Sigma/\sqrt{t}) \leq r(\Delta_{\bar{\mathcal{T}}})/t.$$

Proof. For any $i \in \mathcal{T}_k$ and $j \in \mathcal{T}_{k-1}$ $(k \geq 2)$, we have $|\Delta[i]| \leq |\Delta[j]|$. Thus, $R(|\Delta[i]|) \leq r(\Delta_{\mathcal{T}_{k-1}})/t$, i.e.,

$$|\Delta[i]|^2 \leq (R^{-1}(r(\Delta_{\mathcal{T}_{k-1}})/t))^2.$$

We can derive that

$$\|\Delta_{\mathcal{T}_k}\|_2 \leq \sqrt{t} R^{-1}(r(\Delta_{\mathcal{T}_{k-1}})/t).$$

Hence,

$$\begin{aligned} R(\Sigma/\sqrt{t}) &\leq \textstyle\sum_{k \geq 1} R(\|\Delta_{\mathcal{T}_{k+1}}\|_2/\sqrt{t}) \\ &\leq \textstyle\sum_{k \geq 1} r(\Delta_{\mathcal{T}_k})/t \\ &= r(\Delta_{\bar{\mathcal{T}}})/t. \ \square \end{aligned}$$

Lemma 4

$$\begin{aligned} &\max\{\|\Delta_{\mathcal{T}}\|_2, \|\Delta_{\mathcal{T}_1}\|_2\} \\ \leq &\frac{1+\sqrt{2}}{2\kappa_-(2t)} \left[\frac{\kappa_+(2t) - \kappa_-(2t)}{2} \Sigma + \sqrt{t}(R'(0) + \epsilon) \right]. \end{aligned} \tag{21}$$

Proof. Since zero vector is a subgradient of the objective function in Problem (3) at $\theta = \hat{\theta}$, we have

$$\|X^T(X\hat{\theta} - y)/m\|_\infty \leq R'(0)$$

Then it follows from the triangle inequality that

$$\begin{aligned} \|X^T X \Delta/m\|_\infty &\leq \|X^T(X\hat{\theta} - y)/m\|_\infty + \|X^T e/m\|_\infty \\ &\leq R'(0) + \epsilon. \end{aligned}$$

We modify Eqn. (12) of Foucart and Lai (2009) to that

$$\begin{aligned} & \langle X\Delta, X(\Delta_{\mathcal{T}} + \Delta_{\mathcal{T}_1}) \rangle \\ \leq \ & (\|\Delta_{\mathcal{T}}\|_1 + \|\Delta_{\mathcal{T}_1}\|_1) \|X^T X \Delta\|_\infty \\ \leq \ & \sqrt{t}(R'(0) + \epsilon)(\|\Delta_{\mathcal{T}}\|_2 + \|\Delta_{\mathcal{T}_1}\|_2). \end{aligned}$$

Following the proof of Theorem 3.1 in Foucart and Lai (2009), we conclude (21). □

Next, we turn to the proof the theorem. Let $\kappa_- = \kappa_-(2t)$ and $\kappa_+ = \kappa_+(2t)$. There are two cases according to the difference of supports of $\hat{\theta}$ and θ^*.

Case 1: the supports of $\hat{\theta}$ and θ^* are the same.

We have $\Delta_i = 0$ for $i \in \bar{\mathcal{S}}$ and $\Sigma = 0$, with which and Lemma 4, we can obtain that $\|\Delta\|_2 = \|\Delta_{\mathcal{S}}\|_2 \leq c_1(R'(0) + \epsilon)$ where $c_1 = (1 + \sqrt{2})\sqrt{t}/(2\kappa_-)$.

Case 2: the support of $\hat{\theta}$ and θ^* are not the same.

There exists j satisfying $|\Delta_j| \geq \rho$. Let $\mathcal{T}$ be the indices of first s largest components of Δ in the sense of magnitudes. Obviously, $r(\Delta_{\mathcal{T}}) \geq R(\rho)$.

By Lemma 4, we have

$$r(\Delta_{\mathcal{T}}) \leq sR\left(\frac{\|\Delta_{\mathcal{T}}\|_2}{\sqrt{s}}\right)$$
$$\leq sR\left(\frac{1+\sqrt{2}}{2\sqrt{s}\kappa_-}\left(\frac{\kappa_+-\kappa_-}{2}\Sigma + \sqrt{t}(R'(0)+\epsilon)\right)\right), \tag{22}$$

Combining with Lemma 2 with $\alpha = t/(t-1)$ and 3, it follows that

$$R^{-1}\left(\tfrac{r(\Delta_{\mathcal{S}})}{s}\right) - \tfrac{(1+\sqrt{2})(\kappa_+-\kappa_-)}{4\kappa_-}\sqrt{\tfrac{t}{s}}R^{-1}\left(\tfrac{r(\Delta_{\mathcal{S}})}{t-1}\right)$$
$$\leq \tfrac{(1+\sqrt{2})}{2\kappa_-}\sqrt{\tfrac{t}{s}}(R'(0)+\epsilon).$$

Note that $r(\Delta_{\mathcal{S}}) \geq R(\rho)$. Since $\frac{R^{-1}(u/s))}{R^{-1}(u/(t-1))}$ is an non-decreasing function of u, we have that

$$\frac{R^{-1}(r(\Delta_{\mathcal{S}})/s)}{R^{-1}(r(\Delta_{\mathcal{S}})/(t-1))} \geq \frac{R^{-1}(R(\rho)/s)}{R^{-1}(R(\rho)/(t-1)))}.$$

for $\rho > 0$. If $\rho = 0$, the left hand of the above is lower bounded by $(t-1)/s$. Define $G_R(\rho)$ as Eqn. (7). Under the condition that

$$G_R(\rho) - \frac{(1+\sqrt{2})(\kappa_+-\kappa_-)}{4\kappa_-} > 0, \tag{23}$$

it holds that

$$R^{-1}\left(r(\Delta_{\mathcal{S}})/(t-1)\right) \leq c_2(R'(0)+\epsilon), \tag{24}$$

where

$$c_2 = \left(G_R(\rho,\alpha,s,t) - \frac{(1+\sqrt{2})}{4}\left(\frac{\kappa_+}{\kappa_-}-1\right)\right)^{-1}\frac{1+\sqrt{2}}{2\kappa_-}. \tag{25}$$

Hence, we have

$$\Sigma \leq \sqrt{t}c_2(R'(0)+\epsilon) \tag{26}$$

and

$$\|\Delta\|_2 \leq \|\Delta_{\mathcal{T}}\|_2 + \|\Delta_{\mathcal{T}_1}\|_2 + \Sigma$$
$$\leq \frac{1+\sqrt{2}}{\kappa_-}\left(\frac{\kappa_+-\kappa_-}{2}\Sigma + \sqrt{t}(R'(0)+\epsilon)\right) + \Sigma$$
$$\leq C_1(R'(0)+\epsilon), \tag{27}$$

where

$$C_1 = \frac{(1+\sqrt{2})}{2}\left[\left(\frac{\kappa_+}{\kappa_-}+2\sqrt{2}-3\right)c_2 + \frac{2}{\kappa_-}\right]\sqrt{t} \tag{28}$$
$$\geq c_1. \ \square$$

7.2 Proof of Theorem 2

The proof is similar to Theorem 2 in Zhang and Zhang (2012) except that we bound $r(\Delta_{\mathcal{S}})$ and $\|X\Delta\|_2^2/(2m)$ as follows. By Eqn. (24), we have

$$r(\Delta_{\mathcal{S}}) \leq (t-1)R(c_2(R'(0)+\epsilon))$$

and

$$\frac{1}{2m}\|X\Delta\|_2^2 \leq \frac{2\alpha}{1+\alpha}r(\Delta_{\mathcal{S}}) \leq tR(c_2(R'(0)+\epsilon)).$$

7.3 Proof of Theorem 3

For $w \in \mathbb{R}$, consider the function $h(u) = \alpha R(|u|) + \frac{1}{2}(u-w)^2$ for $u \in \mathbb{R}$. For any $u_1,\ u_2 \in \mathbb{R}$ and $\alpha_1,\ \alpha_2 \geq 0$ $(\alpha_1+\alpha_2 = 1)$, we have

$$R(|\alpha_1 u_1 + \alpha_2 u_2|)$$
$$\leq R(\alpha_1|u_1| + \alpha_2|u_2|)$$
$$\leq R(|u_1|) + R'(|u_1|)(\alpha_1|u_1| + \alpha_2|u_2| - |u_1|)$$
$$\leq R(|u_1|) + \alpha_2 R'(|u_1|)(|u_2| - |u_1|). \tag{29}$$

Similarly,

$$R(|\alpha_1 u_1 + \alpha_2 u_2|) \leq R(|u_2|) + \alpha_1 R'(|u_2|)(|u_1| - |u_2|). \tag{30}$$

From (39)×α_1+(40)×α_2, we have

$$R(|\alpha_1 u_1 + \alpha_2 u_2|)$$
$$\leq \alpha_1 R(|u_1|) + \alpha_2 R(|u_2|)$$
$$+ \alpha_1\alpha_2(R'(|u_1|) - R'(|u_2|))(|u_2| - |u_1|)$$
$$\leq \alpha_1 R(|u_1|) + \alpha_2 R(|u_2|)$$
$$+ \alpha_1\alpha_2|R'(|u_1|) - R'(|u_2|)|\cdot|u_2 - u_1|$$
$$\leq \alpha_1 R(|u_1|) + \alpha_2 R(|u_2|) + \alpha_1\alpha_2\nu_R(u_1-u_2)^2 \tag{31}$$

and

$$(\alpha_1 u_1 + \alpha_2 u_2 - w)^2$$
$$= \alpha_1(u_1-w)^2 + \alpha_2(u_2-w)^2 - \alpha_1\alpha_2(u_1-u_2)^2 \tag{32}$$

Add (41)×α and (32)×1/2, we have

$$h(\alpha_1 u_1 + \alpha_2 u_2)$$
$$\leq \alpha_1 h(u_1) + \alpha_2 h(u_2) + (\alpha\nu_R - 1/2)\alpha_1\alpha_2(u_1-u_2)^2$$
$$\leq \alpha_1 h(u_1) + \alpha_2 h(u_2). \tag{33}$$

Hence, $h(\cdot)$ is a convex function given $2\alpha\nu_R \leq 1$.

Problem (12) can be recast as n one dimensional problems $\min_{\theta[i]\in\mathbb{R}} \alpha R(|\theta[i]|) + \frac{1}{2}(\theta[i] - w[i])^2$ for $i = 1, \cdots, n$. Hence, Problem (12) is convex when $2\alpha\nu_R \leq 1$. $\square$

7.4 Proof of Lemma 1

The proof of Lemma 1 needs the following lemma.

Lemma 5 *Let $\theta' \in \mathbb{R}^n$, $\nu > 0$. If $\mathcal{D}(p_\nu(\theta'), \theta') \le \frac{\nu}{2}\|p_\nu(\theta') - \theta'\|_2^2$, it holds that for any $\theta \in \mathbb{R}^n$,*

$$\begin{aligned}\mathcal{F}(\theta) - \mathcal{F}(p_\nu(\theta')) \ge \frac{\nu - 2\nu_R}{2}\|p_\nu(\theta') - \theta\|_2^2 \\ -\frac{\nu}{2}\|\theta' - \theta\|_2^2 + \mathcal{D}_{\mathcal{L}}(\theta; \theta'). \qquad (34)\end{aligned}$$

Proof. First, for any $u, v \in \mathbb{R}^n$, we define

$$\mathcal{Q}_\nu(u, v) = \mathcal{L}(v) + \langle \nabla\mathcal{L}(v), u - v\rangle + \frac{\nu}{2}\|u - v\|_2^2 + r(u).$$

Thus, $\mathcal{Q}_\nu(p_\nu(\theta'), \theta')$ is an upper bound to $\mathcal{F}(p_\nu(\theta'))$ under the condition that $\mathcal{D}(p_\nu(\theta'), \theta') \le \frac{\nu}{2}\|p_\nu(\theta') - \theta'\|_2^2$, i.e. $\mathcal{F}(p_\nu(\theta')) \le \mathcal{Q}_\nu(p_\nu(\theta'), \theta')$.

For $\mathcal{L}(\theta)$, we have

$$\mathcal{L}(\theta) = \mathcal{L}(\theta') + \langle \theta - \theta', \nabla\mathcal{L}(\theta')\rangle + \mathcal{D}_{\mathcal{L}}(\theta; \theta'). \qquad (35)$$

For the CLIF $R(\cdot)$ and any $\theta, \theta' \ge 0$, we have

$$\begin{aligned}R(\theta) &\ge R(\theta') + R'(\theta)(\theta - \theta') \\ &= R(\theta') + R'(\theta')(\theta - \theta') + (R'(\theta) - R'(\theta'))(\theta - \theta') \\ &\ge R(\theta') + R'(\theta')(\theta - \theta') - \nu_R(\theta - \theta')^2. \qquad (36)\end{aligned}$$

Note that we define $R'(0) = \lim_{\theta \to 0^+} R'(\theta)$.

Since $p_\nu(\theta')$ the minimizer of $Q_\nu(\theta, \theta')$ w.r.t. θ, there exists $\Gamma(\theta') \in \partial r(z)$ such that such that $p_\nu(\theta')$ satisfies

$$\Gamma(\theta') + \nu(z - \theta') + \nabla\mathcal{L}(\theta') = 0,$$

where $z = p_\nu(\theta')$. For any $i = 1, \cdots, n$, we have the following three cases:

1. if $p_\nu(\theta')[i] = 0$, there exists $\phi_i \in [-1, 1]$ such that $\Gamma(\theta')[i] = \phi_i R'(0)$ and $R'(|p_\nu(\theta')[i]|)(|\theta[i]| - |p_\nu(\theta')[i]|) = R'(0)|\theta[i]| \ge \phi_i R'(0)\theta[i] = (\Gamma(\theta')[i])(\theta[i] - p_\nu(\theta')[i])$;

2. if $p_\nu(\theta')[i] > 0$, we have $\Gamma(\theta')[i] = R'(p_\nu(\theta')[i]) > 0$. Hence, $R'(|p_\nu(\theta')[i]|)(|\theta[i]| - |p_\nu(\theta')[i]|) = \Gamma(\theta')[i](\theta[i] - p_\nu(\theta')[i])$;

3. if $p_\nu(\theta')[i] < 0$, we have $\Gamma(\theta')[i] = -R'(p_\nu(\theta')[i]) < 0$. Hence, $R'(|p_\nu(\theta')[i]|)(|\theta[i]| - |p_\nu(\theta')[i]|) = \Gamma(\theta')[i](-|\theta[i]| - p_\nu(\theta')[i]) \ge \Gamma(\theta')[i](\theta[i] - p_\nu(\theta')[i])$.

Hence, it holds that

$$R'(|p_\nu(\theta')[i]|)(|\theta[i]| - |p_\nu(\theta')[i]|) \ge \Gamma(\theta')[i](\theta[i] - p_\nu(\theta')[i]). \qquad (37)$$

With (36) and (37), we have

$$\begin{aligned}& r(\theta) \\ \ge & \sum_i [R(|p_\nu(\theta')[i]|) + (|\theta[i]| \\ & -|p_\nu(\theta')[i]|)R'(|p_\nu(\theta')[i]|) - \nu_R(|\theta[i]| - |p_\nu(\theta')[i]|)^2] \\ \ge & r(p_\nu(\theta')) + \langle R'(|p_\nu(\theta')|), |\theta| - |p_\nu(\theta')|\rangle \\ & - \nu_R\||\theta| - |p_\nu(\theta')|\|_2^2 \\ \ge & r(p_\nu(\theta')) + \langle \Gamma(\theta'), \theta - p_\nu(\theta')\rangle - \nu_R\|\theta - p_\nu(\theta')\|_2^2 \qquad (38)\end{aligned}$$

With (35), (38) and $\mathcal{D}(p_\nu(\theta'), \theta') \le \frac{\nu}{2}\|p_\nu(\theta') - \theta'\|_2^2$, Eqn. (34) follows with some algebra. □

Invoking Lemma 5 with $\theta = \theta' = \theta_k$ and $\nu = \nu_k$, we have $\mathcal{F}(\theta_k) \ge \mathcal{F}(\theta_{k+1}) \ge 0$. Hence, $\{\mathcal{F}(\theta^{(k)})\}$ is a decreasing sequence and converges. Furthermore, $\|\theta_k - \theta_{k-1}\|_2 \to 0$ as $k \to \infty$.

For any $u_1, u_2 \in \mathbb{R}^n$, we define

$$\mathcal{Q}_\nu(u_1, u_2) = \mathcal{L}(u_2) + \langle \nabla\mathcal{L}(u_2), u_1 - u_2\rangle + \frac{\nu}{2}\|u_1 - u_2\|_2^2 + r(u_1).$$

θ_k is a minimizer of $\min_\theta Q_{\nu_k}(\theta, \theta_k)$ in fact. Thus, the directional derivative along any $d \in \mathcal{R}^n$ is non-negative, i.e.,

$$r'(\theta_k; d) + d^T\nabla\mathcal{L}(\theta_{k-1}) + \nu_k d^T(\theta_k - \theta_{k-1}) \ge 0.$$

Then,

$$\begin{aligned}& \mathcal{F}'(\theta_k; d) \\ = & d^T\nabla\mathcal{L}(\theta_k) + r'(\theta_k; d) \\ \ge & d^T(\nabla\mathcal{L}(\theta_k) - \nabla\mathcal{L}(\theta_{k-1}) - \nu_k(\theta_k - \theta_{k-1})) \\ \ge & -\|d\|_2(\|\nabla\mathcal{L}(\theta_k) - \nabla\mathcal{L}(\theta_{k-1})\|_2 + \nu_k\|\theta_k - \theta_{k-1}\|_2) \\ \ge & -(1 + \eta)\nu_{\mathcal{L}}\|d\|_2\|\theta_k - \theta_{k-1}\|_2.\end{aligned}$$

Since $\|\theta_k - \theta_{k-1}\|_2 \to 0$, there exists $N > 0$ so that $\mathcal{F}'(\theta_k; d) \ge -\varphi$ holds for any $k > N$. □

7.5 Proof of Theorem 4

Proof. For any $u_1,\ u_2 \in \mathbb{R}$ and $a_1,\ a_2 \ge 0$ ($a_1 + a_2 = 1$), we have

$$\begin{aligned}& R(|a_1u_1 + a_2u_2|) \\ \le & R(a_1|u_1| + a_2|u_2|) \\ \le & R(|u_1|) + R'(|u_1|)(a_1|u_1| + a_2|u_2| - |u_1|) \\ \le & R(|u_1|) + a_2R'(|u_1|)(|u_2| - |u_1|). \qquad (39)\end{aligned}$$

Similarly,

$$R(|a_1u_1 + a_2u_2|) \le R(|u_2|) + a_1R'(|u_2|)(|u_1| - |u_2|). \qquad (40)$$

From (39)$\times a_1$+(40)$\times a_2$, we have

$$\begin{aligned}
&R(|a_1u_1+a_2u_2|)\\
\le& a_1R(|u_1|)+a_2R(|u_2|)\\
&+a_1a_2(R'(|u_1|)-R'(|u_2|))(|u_2|-|u_1|)\\
\le& a_1R(|u_1|)+a_2R(|u_2|)\\
&+a_1a_2|R'(|u_1|)-R'(|u_2|)|\cdot|u_2-u_1|\\
\le& a_1R(|u_1|)+a_2R(|u_2|)+a_1a_2\nu_R(u_1-u_2)^2 \qquad (41)
\end{aligned}$$

Thus,

$$\begin{aligned}
&\mathcal{F}(a_1\tilde{\theta}+a_2\hat{\theta})\\
=&a_1\mathcal{L}(\tilde{\theta})+a_2\mathcal{L}(\hat{\theta})-\frac{a_2a_1}{2m}\|X(\tilde{\theta}-\hat{\theta})\|_2^2\\
&+\sum_{i=1}^{n}R(|a_1\tilde{\theta}+a_2\hat{\theta}|)\\
\le&a_1\mathcal{L}(\tilde{\theta})+a_2\mathcal{L}(\hat{\theta})+a_1r(\tilde{\theta})+a_2r(\hat{\theta})\\
&a_1a_2(\nu_R\|\tilde{\theta}-\hat{\theta}\|_2^2-\|X(\tilde{\theta}-\hat{\theta})\|_2^2/(2m))\\
\le&a_1\mathcal{F}(\tilde{\theta})+a_2\mathcal{F}(\hat{\theta})+a_1a_2(\nu_R-\kappa_-(q,\hat{s}))\\
\le&a_1\mathcal{F}(\tilde{\theta})+a_2\mathcal{F}(\hat{\theta}). \qquad (42)
\end{aligned}$$

The directional derivative along $\hat{\theta}-\tilde{\theta}$ satisfies that

$$\begin{aligned}
-\varphi\le\mathcal{F}'(\tilde{\theta};\hat{\theta}-\tilde{\theta})&=\liminf_{a\downarrow 0}\frac{\mathcal{F}(\tilde{\theta}+a(\hat{\theta}-\tilde{\theta}))-\mathcal{F}(\tilde{\theta})}{a}\\
&\le\liminf_{a\downarrow 0}\mathcal{F}(\hat{\theta})-\mathcal{F}(\tilde{\theta}). \qquad (43)
\end{aligned}$$

Thus, Theorem 4.1 follows. With the same analysis, Theorem 4.2 follows. □

7.6 Proof of Theorem 5

Remark 3.2 of Beck and Teboulle (2009) provides that

$$\nu_0\le\nu_j\le\eta\nu_{\mathcal{L}}, \qquad (44)$$

for $j=0,1,2,\cdots$. With Lemma 5 and (44), the proof of this theorem is similar with the proof of Theorem 3.1 in Beck and Teboulle (2009):

Let $\kappa_-=\kappa_-(q_{k_0,k},\hat{s})$. From Lemma 5, we obtain that

$$\begin{aligned}
&\frac{2}{\nu_{j+1}}(\mathcal{F}(\hat{\theta})-\mathcal{F}(\theta_{j+1}))\\
\ge&\left(1-\frac{2\nu_R}{\nu_{j+1}}\right)\|\theta_{j+1}-\hat{\theta}\|_2^2-\|\theta_j-\hat{\theta}\|_2^2+\frac{2\kappa_-}{\nu_{j+1}}\|\theta_j-\hat{\theta}\|_2^2\\
\ge&\left(1-\frac{2\nu_R}{\nu_{j+1}}\right)\left(\|\theta_{j+1}-\hat{\theta}\|_2^2-\|\theta_j-\hat{\theta}\|_2^2\right)
\end{aligned}$$

holds for $j=k_0,\cdots,k-1$. Combined with (44) and the fact that $\mathcal{F}(\hat{\theta})-\mathcal{F}(\theta_{j+1})\le 0$ yields

$$\|\theta_{j+1}-\hat{\theta}\|_2\le\|\theta_j-\hat{\theta}\|_2$$

and

$$\begin{aligned}
&\frac{2}{\eta\nu_{\mathcal{L}}}(\mathcal{F}(\hat{\theta})-\mathcal{F}(\theta_{j+1}))\\
\ge&\left(1-\frac{2\nu_R}{\eta\nu_{\mathcal{L}}}\right)(\|\theta_{j+1}-\hat{\theta}\|_2^2-\|\theta_j-\hat{\theta}\|_2^2).
\end{aligned}$$

Summing the above inequality over $j=k_0,\cdots,k-1$ gives

$$\begin{aligned}
&\frac{2}{\eta\nu_{\mathcal{L}}}\left((k-k_0)\mathcal{F}(\hat{\theta})-\sum_{j=k_0}^{k-1}\mathcal{F}(\theta_{j+1})\right)\\
\ge&\left(1-\frac{2\nu_R}{\eta\nu_{\mathcal{L}}}\right)(\|\hat{\theta}-\theta_k\|_2^2-\|\hat{\theta}-\theta_{k_0}\|_2^2). \qquad (45)
\end{aligned}$$

Invoking Lemma 5 again with $\theta=\theta'=\theta_j$ and $\nu=\nu_{j+1}$ yields

$$\frac{2}{\nu_{j+1}}(\mathcal{F}(\theta_j)-\mathcal{F}(\theta_{j+1}))\ge\left(1-\frac{2\nu_R}{\nu_{j+1}}\right)\|\theta_j-\theta_{j+1}\|_2^2.$$

Since $\nu_{j+1}\ge\nu_0$ and $\mathcal{F}(\theta_j)-\mathcal{F}(\theta_{j+1})\ge 0$, it follows that

$$\frac{2}{\nu_0}(\mathcal{F}(\theta_j)-\mathcal{F}(\theta_{j+1}))\ge\left(1-\frac{2\nu_R}{\nu_0}\right)\|\theta_j-\theta_{j+1}\|_2^2.$$

Multiplying the last inequality by $j-k_0$ and summing over $j=k_0,\cdots,k-1$, we obtain

$$\begin{aligned}
&\frac{2}{\nu_0}\sum_{j=k_0}^{k-1}((j-k_0)\mathcal{F}(\theta_j)-(j-k_0+1)\mathcal{F}(\theta_{j+1})+\mathcal{F}(\theta_{j+1}))\\
\ge&\left(1-\frac{2\nu_R}{\nu_0}\right)\sum_{j=k_0}^{k-1}(j-k_0)\|\theta_j-\theta_{j+1}\|_2^2,
\end{aligned}$$

which simplifies to

$$\begin{aligned}
&\frac{2}{\nu_0}\left(-(k-k_0)\mathcal{F}(\theta_k)+\sum_{j=k_0}^{k-1}\mathcal{F}(\theta_{j+1})\right)\\
\ge&\left(1-\frac{2\nu_R}{\nu_0}\right)\sum_{j=k_0}^{k-1}(j-k_0)\|\theta_j-\theta_{j+1}\|_2^2. \qquad (46)
\end{aligned}$$

Adding (45) and (46) $\times\frac{\nu_0}{\eta\nu_{\mathcal{L}}}$, we get

$$\begin{aligned}
&\frac{2(k-k_0)}{\eta\nu_{\mathcal{L}}}(\mathcal{F}(\hat{\theta})-\mathcal{F}(\theta_k))\\
\ge&\left(1-\frac{2\nu_R}{\eta\nu_{\mathcal{L}}}\right)(\|\hat{\theta}-\theta_k\|_2^2-\|\hat{\theta}-\theta_{k_0}\|_2^2)\\
&+\frac{\nu_0}{\eta\nu_{\mathcal{L}}}\left(1-\frac{2\nu_R}{\nu_0}\right)\sum_{j=k_0}^{k-1}(j-k_0)\|\theta_j-\theta_{j+1}\|_2^2.
\end{aligned}$$

Theorem 5 follows. □