

# Adaptive $L_p$ ( $0 < p < 1$ ) Regularization: Oracle Property and Applications

Yunxiao Shi<sup>1(✉)</sup>, Xiangnan He<sup>1</sup>, Han Wu<sup>1</sup>, Zhong-Xiao Jin<sup>2</sup>, and Wenlian Lu<sup>1</sup>

<sup>1</sup> School of Mathematical Science, Fudan University, Shanghai, China  
kentsyx@gmail.com

<sup>2</sup> SAIC Motor Corporation Limited, No. 489, Wei Hai Road, Shanghai, China

**Abstract.** In this paper, we propose adaptive  $L_p$  ( $0 < p < 1$ ) estimators in sparse, high-dimensional, linear regression models when the number of covariates depends on the sample size. Other than the case of the number of covariates is smaller than the sample size, in this paper, we prove that under appropriate conditions, these adaptive  $L_p$  estimators possess the oracle property in the case that the number of covariates is much larger than the sample size. We present a series of experiments demonstrating the remarkable performance of this estimator with adaptive  $L_p$  regularization, in comparison with the  $L_1$  regularization, the adaptive  $L_1$  regularization, and non-adaptive  $L_p$  regularization with  $0 < p < 1$ , and its broad applicability in variable selection, signal recovery and shape reconstruction.

**Keywords:** Adaptive  $L_p$  regularization · Oracle property · Sparse regression · Variable selection · Compressed sensing

## 1 Introduction

High prediction accuracy and discovering relevant predictive variables are two fundamental problems in statistical learning. Variable selection is particularly important when the underlying model has a sparse representation, especially in high-dimensional and massive data analysis. It has been argued by [1]<sup>2</sup> that a good estimator should have oracle property, namely, the estimator

- correctly selects covariates with nonzero coefficients with probability converging to one, as the sample size goes to infinity, and
- has the same asymptotic distribution as if the zero coefficients were known in advance.

Consider the linear regression model  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ , where  $\mathbf{X} \in R^{n \times l_n}$  is a design matrix,  $\boldsymbol{\beta} \in R^{l_n}$  is the vector of unknown coefficients, and  $\boldsymbol{\epsilon} \in R^n$  is the vector of i.i.d. random variables with mean zero and finite variance  $\sigma^2$ . Note that  $l_n$ , the length of  $\boldsymbol{\beta}$  depends on the sample size  $n$  and may go to infinity as  $n \rightarrow \infty$ . Without loss of generality, we assume that the response vector  $\mathbf{y} \in R^n$

and the covariates are centered so that the intercept term can be excluded. In many situations we are to recover  $\beta$  from observation  $\mathbf{y}$  such that  $\beta$  is of the most sparse structure, that is,  $\beta$  has the fewest nonzero components. A direct approach is to formulate this problem as  $\min_{\beta \in R^n} \|\beta\|_0$  such that  $\mathbf{y} = \mathbf{X}\beta + \epsilon$ , which can be transformed into  $\min_{\beta \in R^n} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda \|\beta\|_0$ , which is called an  $L_0$  regularization problem, where  $\|\beta\|_0$  is the number of nonzero components of  $\beta$  and  $\lambda$  is the regularization parameter. Indeed this method can recover sparse solutions even in situations in which  $l_n \gg n$ , in fact, it can perfectly recover all the sparse  $\beta$  obeying  $\|\beta\|_0 \leq n/2$ . However this is of little practical use since generally solving an  $L_0$  regularization problem usually requires an intractable number of combinatorial searches. To conquer this difficulty, several approximations to the  $L_0$  problem have been proposed, such as  $L_1$  regularization [2–5], the adaptive Lasso [6], the  $L_p$  ( $0 < p < 1$ ) regularization [7, 8] and the adaptive  $L_p$  ( $0 < p < 1$ ) regularization [9].

Among the proposed techniques above,  $L_1$  regularization (or Lasso) overcame the huge computational cost for large problems of the  $L_0$  but may introduce inconsistent estimations [6] and extra bias [10]. The adaptive Lasso and  $L_p$  regularization solved the above problems and their oracle property were established in both low and high dimensional scenarios [6–8, 11]. Meanwhile it has been claimed that the  $L_p$  ( $0 < p < 1$ ) regularization yields more sparse solutions than both the Lasso and the adaptive Lasso [12, 13], but sometimes its sparsity would lead to unstable estimation [9], who therefore proposed the adaptive  $L_p$  ( $0 < p < 1$ ) regularization,  $\min_{\beta \in R^n} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda \sum_{j=1}^{l_n} \omega_j |\beta_j|^p$ , and proved its oracle property when the number of covariates is fixed.

In this paper, we continue to investigate the adaptive  $L_p$  ( $0 < p < 1$ ) regularization when the number of covariates depends on the sample size and can go to infinity as the sample size goes to infinity. We prove that under a series of mild conditions, the adaptive  $L_p$  ( $0 < p < 1$ ) estimator enjoys the oracle property in high-dimensional settings even when  $l_n \gg n$ , and proposed algorithms that can efficiently solve the adaptive  $L_p$ . Finally we demonstrate the superior performance of the adaptive  $L_p$  in variable selection, signal recovery and image shape reconstruction by a series of numerical experiments, in comparison to the  $L_1$  estimator, the adaptive  $L_1$  estimator and the  $L_{1/2}$  estimator.

## 2 Preliminaries

The symbol  $\rightarrow$  stands for convergence in the common sense,  $\rightarrow_p$  for convergence in probability, and  $\rightarrow_d$  for convergence in distribution.  $\mathbb{P}(\cdot)$  stands for the probability.  $X_n = O_p(1)$  stands for some stochastically bounded sequence, and  $X_n = o_p(1)$  for  $X_n \rightarrow_p 0$  as  $n \rightarrow \infty$ . Meanwhile  $\beta_0 = [\beta_{01}^\top, \beta_{00}^\top]^\top \in R^{l_n}$ , where  $\beta_{01} \in R^{k_n}$  consists of the nonzero terms of  $\beta_0$  and  $\beta_{00} \in R^{m_n}$  are the zero ones, note that  $k_n + m_n = l_n$ . We center the response vector  $\mathbf{y} = [y_1, \dots, y_n]^\top$  and standardize the design matrix  $X = (x_{ij})_{n \times p_n}$  so that  $\sum_{i=1}^n y_i = 0$ ,  $\sum_{i=1}^n x_{ij} = 0$ ,  $\frac{1}{n} \sum_{i=1}^n x_{ij}^2 = 1$ ,  $j = 1, \dots, l_n$ .

Let  $J_{1n} = \{j \mid \beta_{0j} \neq 0\}$  and set  $\mathbf{X}_n^\top = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n] \in R^{l_n \times n}$ ,  $\mathbf{X}_{n1} = [\mathbf{x}_{11}, \mathbf{x}_{12}, \dots, \mathbf{x}_{1n}] \in R^{k_n \times n}$ , accordingly we define  $\Sigma_n = \frac{1}{n} \mathbf{X}_n^\top \mathbf{X}_n$ ,  $\Sigma_{n1} = \frac{1}{n} \mathbf{X}_{n1}^\top \mathbf{X}_{n1}$ . We denote  $\rho_{1n}$  and  $\tau_{1n}$  as the smallest eigenvalue of  $\Sigma_n$  and  $\Sigma_{n1}$  respectively, and  $\rho_{2n}$ ,  $\tau_{2n}$  the largest eigenvalue of  $\Sigma_n$  and  $\Sigma_{n1}$ . We consider the *oracle property* which was proposed by [1].

**Definition 2.1** (*Oracle Property*). Let  $\hat{\beta}_n^\top = [\hat{\beta}_{n1}^\top, \hat{\beta}_{n0}^\top]^\top$  be the estimator of the true parameter  $\beta_0 = [\beta_{01}^\top, \beta_{00}^\top]^\top$ . Then  $\hat{\beta}_n$  is said to possess oracle property if the two conditions below are satisfied: (1). (*Consistency*)  $\lim_{n \rightarrow \infty} \mathbb{P}(\beta_{n0} = \mathbf{0}) = 1$ ; (2). (*Asymptotic Normality*) Let  $s_n^2 = \sigma^2 \alpha_n^\top \Sigma_{n1}^{-1} \alpha_n$ , where  $\alpha_n$  is any  $k_n \times 1$  vector satisfying  $\|\alpha_n\|_2 \leq 1$  such that

$$n^{-\frac{1}{2}} s_n^{-1} \alpha_n^\top (\hat{\beta}_{n1} - \beta_{01}) = n^{-\frac{1}{2}} s_n^{-1} \sum_{i=1}^n \epsilon_i \alpha_n^\top \Sigma_{n1}^{-1} \mathbf{x}_{i1} + o_p(1) \rightarrow_d N(0, 1). \quad (2.1)$$

Let  $b_{1n} = \min_{j \in J_{1n}} \{|\beta_{0j}|\}$ ,  $b_{2n} = \max_{j \in J_{1n}} \{|\beta_{0j}|\}$ .

**Definition 2.2** (*Zero Consistency*). The estimator  $\tilde{\beta}_n$  is said to be zero consistent if it satisfies the two conditions: (1).  $\max_{j \in J_{0n}} |\tilde{\beta}_{nj}| = o_p(1)$ ; (2). There exists some constant  $c > 0$  such that for any  $\epsilon > 0$  when  $n$  is sufficiently large the following inequality holds

$$\mathbb{P}(\min_{j \in J_{1n}} |\tilde{\beta}_{nj}| \geq cb_{1n}) > 1 - \epsilon. \quad (2.2)$$

where  $\tilde{\beta}_{nj}$  is the marginal regression coefficient [11]. Furthermore, if for a certain constant  $C > 0$ , the following

$$\mathbb{P}(R_n \max_{j \in J_{0n}} |\tilde{\beta}_{nj}| > C) \rightarrow 0, \quad \text{as } n \rightarrow \infty \quad (2.3)$$

where  $\lim_{n \rightarrow \infty} R_n = \infty$ , then  $\tilde{\beta}_n$  is said to be zero consistent with rate  $R_n$ .

### 3 Methods

Now we present the conditions for the oracle property of the adaptive  $L_p$  regularization. Due to the limit of space, we omit the proofs of all theorems which will be seen in our future paper. We consider the estimator [9]

$$U_n(\beta) = \sum_{i=1}^n \sum_{j=1}^{l_n} (Y_i - x_{ij} \beta_j)^2 + \lambda_n \sum_{j=1}^{l_n} \omega_{nj} |\beta_j|^p, \quad (3.1)$$

where  $0 < p < 1$ . Let  $\bar{\beta}_n = \arg \min_{\beta} U_n(\beta) = [\bar{\beta}_{n1}^\top, \bar{\beta}_{n0}^\top]^\top$ , where  $\bar{\beta}_{n1}^\top$  and  $\bar{\beta}_{n0}^\top$  corresponds to the estimates of nonzero and zero coefficients of the true parameter  $\beta_0$  respectively. We give the following assumptions.

- (B1) (i)  $\{\epsilon_i\}_{i=1}^n$  is a sequence of i.i.d. random variables, with mean 0 and a finite variance  $\sigma^2$ ; (ii).  $\epsilon_i$  is sub-Gaussian, that is  $\mathbb{P}(|\epsilon_i| > x) \leq K \exp(-Cx^2)$ , for all  $i \in \mathbb{N}$ , some  $K > 0$  and  $C > 0$ .
- (B2)  $\bar{\beta}_n$  defined by Definition 2.4 is zero-consistent with rate  $R_n$ .
- (B3) (i) There exists a constant  $c_9 > 0$  such that  $\left| n^{-\frac{1}{2}} \sum_{i=1}^n x_{ij} x_{ik} \right| \leq c_9$  for all  $j \in J_{0n}$ , all  $k \in J_{1n}$  and sufficiently large  $n$ ; (ii). Let  $\xi_{nj} = n^{-1} \mathbb{E}(\sum_{i=1}^n Y_i x_{ij}) = n^{-1} \sum_{i=1}^n (\mathbf{x}_{i1}^\top \beta_{01} x_{ij})$ . There exists a  $\xi_0 > 0$ , such that  $\min_{j \in J_{1n}} |\xi_{nj}| > 2\xi_0 b_{1n} > 0$ .
- (B4) (i)  $\frac{\lambda_n}{n} \rightarrow 0$ ,  $\lambda_n n^{-\frac{p}{2}} R_n^\alpha k_n^{p-2} \rightarrow \infty$ ; (ii)  $\log(m_n) = o(1)(\lambda_n n^{-\frac{p}{2}} R_n^\alpha)^{\frac{2}{2-p}}$ .
- (B5) (i) There exists some constants  $0 < b_1 < b_2 < \infty$ , such that  $b_1 < b_{1n} < b_{2n} < b_2$ ; (ii)  $\lim_{n \rightarrow \infty} k_n \exp(-Cn) \rightarrow 0$ .

**Theorem 3.3.** *Suppose the conditions (B1)–(B5) hold. Then  $\bar{\beta}_n$  is consistent in variable selection, namely*

$$\mathbb{P}(\bar{\beta}_{n0} = \mathbf{0}) \rightarrow 1, \quad \mathbb{P}(\bar{\beta}_{n1j} \neq 0, j \in J_{1n}) \rightarrow 1, \quad \text{as } n \rightarrow \infty. \quad (3.2)$$

It can be seen from Theorem 3.3 that under appropriate conditions, the adaptive  $L_p$  ( $0 < p < 1$ ) correctly selects nonzero covariates with probability converging to one. Towards the oracle property, we denote the nonzero terms as  $\bar{\beta}_n$  and consider optimizing the following objective function

$$\tilde{U}_n(\beta_1) = \sum_{i=1}^n (Y_i - \mathbf{x}_{i1}^\top \beta_1)^2 + \lambda_n^* \sum_{i=1}^{k_n} \omega_{nj} |\beta_{1j}|^p, \quad (3.3)$$

where  $k_n$  is number of nonzero terms of  $\bar{\beta}_n$ .

Let  $\hat{\beta}_{n1}$  be the nonzero terms of  $\hat{\beta}_0 = \arg \min_{\beta} \tilde{U}_n(\beta)$ . We further give the following conditions.

- (B6) (i) There exists constants  $0 < \tau_1 < \tau_2 < \infty$ , such that  $0 < \tau_1 < \tau_{1n} < \tau_{2n} < \tau_2$ ; (ii)  $n^{-\frac{1}{2}} \max_{1 \leq i \leq n} \mathbf{x}_{i1}^\top \mathbf{x}_{i1}$ .
- (B7) (i)  $k_n(1 + \lambda_n^*)/n \rightarrow 0$ , (ii)  $\lambda_n^*(k_n/l_n \sqrt{n})^{\frac{1}{2}} \rightarrow 0$ .

Therefore we have

**Theorem 3.4.**  *$\hat{\beta}_{n1}$  is the estimate of the true non-zero parameter  $\hat{\beta}_{01}$ . Suppose condition (B1)–(B7) hold, then*

$$n^{\frac{1}{2}} s_n^{-1} \alpha_n^\top (\hat{\beta}_{n1} - \hat{\beta}_{01}) = n^{\frac{1}{2}} s_n^{-1} \sum_{i=1}^n \epsilon_i \alpha_n^\top \Sigma_{n1}^{-1} \mathbf{x}_{i1} + o_p(1) \rightarrow_d N(0, 1).$$

where  $s_n^2 = \sigma^2 \alpha_n^\top \Sigma_{n1}^{-1} \alpha_n$  and  $\alpha_n$  is an arbitrary  $k_n \times 1$  vector with  $\|\alpha_n\|_2 \leq 1$ .

The assumption that  $\bar{\beta}_n$  is zero-consistent with rate  $R_n$  is critical in establishing the oracle property of the adaptive  $L_p$  ( $0 < p < 1$ ) regularizer. [11] points out that when  $l_n$  is fixed or of the order  $o(\sqrt{n})$ , the OLS estimator

$\tilde{\beta}_{ols} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$  is feasible as  $\tilde{\beta}_n$ . But when  $l_n > O(\sqrt{n})$ , the OLS estimator is no longer zero-consistent. Here we follow the work of [11] but with necessary modification, i.e.,  $\exists b_1 > 0$  such that  $b_{1n} > b_1 > 0$  to present initial estimator. Refer to Sect. 3 of [11] for the rest of the details of discussions. For the case  $l_n < n$ , refer to [9] for details.

However, for the case  $l_n > n$ , the OLS is no longer feasible as an initial estimator. By [11] and Theorem 3.3, we perform variable selection first to obtain the nonzero  $\beta_j$ , which induces the following Algorithm 1.

---

**Algorithm 1.** adaptive  $L_p$  algorithm when  $l_n > n$ .

---

**input** : Predictor matrix  $\mathbf{X}$ , observation vector  $\mathbf{y}$

**output**: The adaptive  $L_p$  estimator  $\hat{\beta}$

**begin**

Let  $\tilde{\beta}_{nj} = \sum_{i=1}^n Y_i x_{ij} / \sum_{i=1}^n x_{ij}^2$

Let  $\omega_{nj} = |\tilde{\beta}_{nj}|^{-\gamma}$

**while**  $l_n > n$  **do**

$\lambda \leftarrow \lambda_n \omega_{nj} / \sum_{i=1}^n x_{ij}^2$

$a \leftarrow \sum_{i=1}^n y_i x_{ij} / \sum_{i=1}^n x_{ij}^2$

**for**  $j = 1$  **to**  $l_n$  **do**

**if**  $\lambda \geq c_p |a|^{2-p}$  **then**

$\beta_j$  is zero

**else**

$\beta_j$  is nonzero

**end if**

**end for**

$I$  is the index of all nonzero  $\beta_j$

$\mathbf{X} \leftarrow \mathbf{X}_I$ ,  $\beta \leftarrow \beta_I$ ,  $\mathbf{y} \leftarrow \mathbf{X}_I \beta_I$

$l_n \leftarrow$  column number of  $\mathbf{X}$

$n \leftarrow$  row number of  $\mathbf{X}$

**end while**

**end**

Use Algorithm 3.1 in [9] with latest  $\mathbf{X}$ ,  $\mathbf{y}$  as input.

Output adaptive  $L_p$  estimator  $\hat{\beta}$ .

---

## 4 Results

In this section, we give three application examples of variable selection, signal recovery and image shape reconstruction respectively. We note that by Algorithm 1 solving the adaptive  $L_p$  ( $0 < p < 1$ ) is equivalent to solving a series of adaptive  $L_1$  which is very quick on a modern computer. We take  $p = 1/2$  in the experiments, which was recommended by [12]. To show the performance of the present algorithm, we compare with the  $L_1$  (Lasso), adaptive  $L_1$  and (nonadaptive)  $L_{1/2}$  regularized estimators. In practice, we use the algorithm proposed in [6] to compute the adaptive Lasso and apply the iterative  $L_1$  algorithm in [12] to the  $L_{1/2}$  regularization. The parameter  $\lambda_n$  in  $L_{1/2}$  is selected by using the generalized cross-validation(GCV) method described in [1, 16].

For comparison with the adaptive  $L_1$  regularizer and the adaptive  $L_{1/2}$  regularizer, two-dimensional cross-validation is used and selected  $\gamma$  from  $\{0.5, 1.0, 1.5\}$ . For fixed  $\gamma$  and  $\lambda$ , we apply Algorithm 1 to obtain a numerical solution  $\hat{\beta} = \hat{\beta}_{\gamma, \lambda}$ . Note that  $\hat{\beta}_{\gamma, \lambda}$  is the minimum of  $L_{1/2}$  regularizer [12], namely  $\hat{\beta}_{\gamma, \lambda} = (X^\top X + \lambda_n W D^*)^{-1} X^\top y$ , where  $W$  is a diagonal matrix with elements  $\omega_{nj}$  and  $D = \text{diag}\{|\hat{\beta}_j|^{3/2}\}$ . Here  $D^*$  of  $D$ . Meanwhile the number of nonzero components of  $\hat{\beta}_{\gamma, \lambda}$  can be approximated by (refer to [16, 17])  $P_{\gamma, \lambda} = \text{tr}(X(X^\top X + \lambda W D^*)^{-1} X^\top)$ . Thus the generalized cross-validation statistic is given by  $\text{GCV}_{\gamma, \lambda} = \frac{1}{l_n} \frac{\text{RSS}_{\gamma, \lambda}}{(1 - P_{\gamma, \lambda}/l_n)^2}$ , where RSS stands for the residual sum of squared errors:  $\sum_{i=1}^n (y_i - x_i^\top \beta)^2$ . Therefore we obtain the solution of adaptive  $L_{1/2}$  regularizer by the minimization problem  $\{\hat{\gamma}, \hat{\lambda}\} = \arg \min_{\gamma, \lambda} \text{GCV}_{\gamma, \lambda}$ , that is,  $\hat{\beta} = \hat{\beta}_{\hat{\gamma}, \hat{\lambda}}$ . All the simulation codes are written in Python with using the package SPAMS [15].

#### 4.1 Variable Selection

Consider the following linear regression model mentioned in [1, 12, 16]  $y = X\beta^* + \sigma\epsilon$ , where the true values of  $\beta^*$  are  $[3, 1.5, 0, 0, 2, 0, 0, 0]^\top$ ,  $\epsilon$  is the i.i.d. noise following certain distribution, and  $\sigma$  is the strength of the noise. We first take  $\sigma = 1$  and second  $\sigma = 3$  with  $\epsilon$  following the standard normal distribution. Finally take  $\sigma = 1$  but  $\epsilon$  follows the linear mixture of 30% standard Cauchy distribution and 70% standard normal distribution (denoted as MIXTURE in the result table). The correlation between each  $x_i$  and  $x_j$  is equal to  $(1/2)^{|i-j|}$ .

For each type of noise, we simulate 100 datasets of  $n = 100$  observations respectively, and the relative model error in each dataset is defined as  $\|\hat{y} - X\hat{\beta}^*\|_2 / \|y - X\beta^*\|_2$ , where  $\hat{y} = X\hat{\beta}$ ,  $\|\hat{y} - X\hat{\beta}^*\|_2$  is the model error and  $\|y - X\beta^*\|_2$  is the inherent prediction error due to the noise. The results shown in Tables 1 and 2 illustrated that adaptive  $L_{1/2}$  is more accurate and sparse than the other three regularizers.

**Table 1.** Median value of relative model error ( $n = 100$ ) (with min/max)

	$\epsilon \sim N(0, 1)$		$\epsilon \sim \text{MIXTURE}$
	$\sigma = 1$	$\sigma = 3$	$\sigma = 1$
Lasso	.2871(.1692/.4239)	.3315(.2341/.4673)	.2437(.1266/.3319)
Adaptive Lasso	.2179(.0936/.3937)	.3067(.2226/.4183)	.1635(.0455/.3243)
$L_{1/2}$	.2367(.1441/.4173)	.3184(.2304/.4540)	.1982(.0602/.3299)
Adaptive $L_{1/2}$	.1896(.0823/.3947)	.2941(.2215/.4191)	.1527(.0454/.3240)

**Table 2.** Average number of zero coefficients ( $n = 100$ ) (with standard deviation)

	$\epsilon \sim N(0, 1)$		$\epsilon \sim \text{MIXTURE}$
	$\sigma = 1$	$\sigma = 3$	$\sigma = 1$
Lasso	2.91 <sub>(.92)</sub>	1.81 <sub>(1.21)</sub>	2.55 <sub>(1.03)</sub>
Adaptive Lasso	3.64 <sub>(1.06)</sub>	2.29 <sub>(1.18)</sub>	3.09 <sub>(1.23)</sub>
$L_{1/2}$	4.27 <sub>(1.12)</sub>	2.72 <sub>(1.31)</sub>	3.68 <sub>(.90)</sub>
Adaptive $L_{1/2}$	4.64 <sub>(.86)</sub>	3.27 <sub>(1.18)</sub>	4.09 <sub>(1.19)</sub>

## 4.2 Signal Recovery

In this and the next experiment, we show the application of adaptive  $L_p$  regularization in compressed sensing [2, 19, 20]. Consider a real-valued and finite length signal  $\mathbf{x} \in R^N$ , which is represented by an orthonormal basis  $\{\psi_i\}_{i=1}^N$  of  $R^N$ . Let  $\Psi = [\psi_1, \dots, \psi_N]$ . There exists  $s \in R^N$  such that  $\mathbf{x} = \Psi s = \sum_{i=1}^N \psi_i s_i$ .

Consider  $y = \Phi x + \epsilon$ , where  $\epsilon$  is a noise term which is either stochastic or deterministic and  $\Phi$  is the “sensing matrix”. It was shown by [19, 21] that reconstruction of  $\mathbf{x}$  can be formulated to minimize the following  $L_0$  problem  $\min_{\mathbf{x} \in R^N} \sum_{i=1}^N I_{x_i \neq 0}$  such that  $\|y - \Phi x\|_2 \leq \delta$ , where the parameter  $\delta$  is adjustable so that the true signal  $\mathbf{x}$  can be feasible. According to the work of [2], if  $\mathbf{x}$  is sufficiently sparse and  $\Phi$  satisfies the Restricted Isometry Property [23], this  $L_0$  problem is equivalent to  $\min_{\mathbf{x} \in R^N} \sum_{i=1}^N |x_i|$  such that  $\|y - \Phi x\|_2 \leq \delta$ , an  $L_1$  regularization. Now we apply the adaptive  $L_{1/2}$  regularizer to solve the original problem, that is, we consider the following minimization problem  $\min_{\mathbf{x} \in R^N} \sum_{i=1}^N \omega_i |x_i|^{1/2}$  such that  $\|y - \Phi x\|_2 \leq \delta$ , to reconstruct the signal.

As a numerical experiment, take  $\mathbf{x} = \sin(2\pi f_1 t) + \cos(2\pi f_2 t) + \epsilon$  with a fix signal length  $N = 512$ ,  $t \in [0, 0.1]$  with fixed-length step,  $f_1 = 16$  and  $f_2 = 384$ . We consider the noise  $\epsilon$  follows the standard normal distribution with  $\mu = 0$  and  $\sigma = 0.01$ . Discrete cosine transform (DCT) is used to obtain the sparse representation of  $\mathbf{x}$  (denoted as  $\hat{\mathbf{x}}$ ). Then we set  $M = 128$  and sample a random  $M \times N$  matrix  $\Phi$  with i.i.d. Gaussian entries. We first apply the  $L_1$ , adaptive  $L_1$ ,  $L_{1/2}$  and our adaptive  $L_{1/2}$  regularization to recover  $\hat{\mathbf{x}}$  and then employ the inverse discrete cosine transform (idct) to obtain the reconstructed  $\mathbf{x}$  respectively. We run each estimator for 100 trials.

We compare the recovery performance of both  $\hat{\mathbf{x}}$  and  $\mathbf{x}$  (denoted as  $\hat{\mathbf{x}}_{\text{re}}$  and  $\mathbf{x}_{\text{re}}$  respectively). For  $\hat{\mathbf{x}}_{\text{re}}$ , we measure the performance in terms of “sparseness”, by the ratio of the number of nonzero coefficients in  $\hat{\mathbf{x}}_{\text{re}}$  to signal length. and for  $\mathbf{x}_{\text{re}}$  we consider the relative error  $\|\mathbf{x}_{\text{re}} - \mathbf{x}\|_2 / \|\mathbf{x}\|_2$ . Tables 3 and 4 shows us that though in terms of accuracy (relative error) our adaptive  $L_{1/2}$  performs slightly worse than adaptive  $L_1$  but better than the others and yields the most sparse solution.

**Table 3.** Averaged sparseness of recovered  $\hat{x}$  (with standard deviation(SD))

	Number of nonzero coefficients	Sparseness
Lasso	122.3 <sub>(2.1)</sub>	.238 <sub>(.004)</sub>
Adaptive $L_1$	63.4 <sub>(5.6)</sub>	.124 <sub>(.011)</sub>
$L_{1/2}$	48.4 <sub>(4.2)</sub>	.095 <sub>(.008)</sub>
Adaptive $L_{1/2}$	43.2 <sub>(3.4)</sub>	.084 <sub>(.006)</sub>

**Table 4.** Averaged relative error under 2-norm of recovered  $x$  (with SD)

	Relative error under 2-norm
Lasso	.185 <sub>(.009)</sub>
Adaptive $L_1$	.145 <sub>(.007)</sub>
$L_{1/2}$	.175 <sub>(.008)</sub>
Adaptive $L_{1/2}$	.168 <sub>(.007)</sub>

### 4.3 Shape Reconstruction

We use the example proposed by [24] to reconstruct an image from a set of parallel projections, acquired along different angles. Similar patterns are commonly seen in computed tomography (CT) data. Without prior knowledge on the sample, the number of projections that are required to reconstruct the image is of order  $O(N)$  (in pixels). Here, we consider the case of the sparse image with the objects that are basic shapes where only the boundary of objects have non-zero value. These images are artificially generated but still correspond to real-life applications including monitoring cellular material.

The sparse image we use here is of size  $128 \times 128$  and we added Gaussian noise with standard variance  $\sigma = 0.2$  (shown in Fig. 1(a)). In reconstruction, we stretch it into a  $128 \times 128 = 16384$  dimensional vector. The reconstruction results using the Lasso and our adaptive  $L_{1/2}$  with  $N/7$  pixels and  $N/10$  sampled are shown in Fig. 1(a).

Both estimators recovered the original image with highly visible accuracy with  $N/7$  pixels sampled, while the adaptive  $L_{1/2}$  regularizer has a better performance numerically which is demonstrated in Table 5. But when the sampling ratio drops to  $N/10$ , the Lasso starts to fail (notice that the shapes break down), while our adaptive  $L_{1/2}$  still gives reconstruction result with high accuracy (shown in Fig. 1(b)).

We use the Structural Similarity Image Metric (SSIM) [25, 26] to compare the reconstruction performance among the four estimators. The SSIM index can be viewed as a quality measure of one of the images being compared, provided the other image is regarded as of perfect quality and improve consistence with human visual perception, in comparison to the traditional indices such as peak signal-to-noise ratio (PSNR) and mean squared error (MSE) [27].



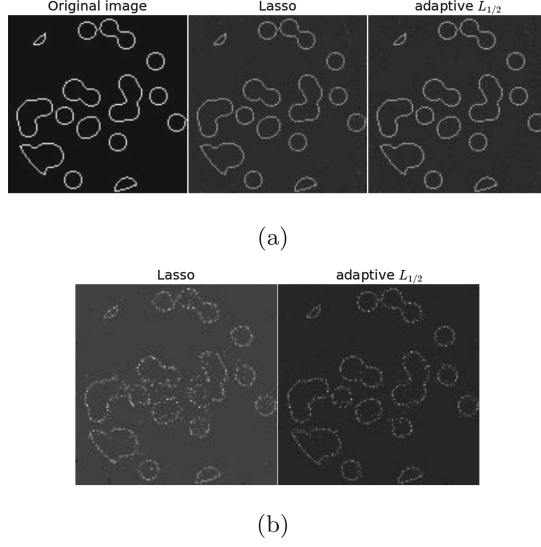
**Fig. 1.**

Table 5 shows that the adaptive  $L_{1/2}$  regularizer has the best performance of reconstruction in both cases that the sampling ratio is  $N/7$  or  $N/10$ .

**Table 5.** Performance comparison of reconstruction results measured by SSIM among the Lasso, adaptive  $L_1$ ,  $L_{1/2}$ , adaptive  $L_{1/2}$ .

	SSIM ( $N/7$ pixels sampled)	SSIM ( $N/10$ pixels sampled)
Lasso	0.217	0.090
Adaptive $L_1$	0.197	0.116
$L_{1/2}$	0.224	0.113
Adaptive $L_{1/2}$	0.233	0.120

## 5 Concluding Remarks

We have conducted a study of a specific framework of the adaptive  $L_p$  ( $0 < p < 1$ ) regularization, towards better performance for the estimation of sparsity problems. We have shown that the adaptive  $L_p$  regularized estimators possess the oracle property when  $l_n \gg n$ . We also proposed a fast and efficient algorithm to solve the adaptive  $L_p$  regularization problem. Our results offer new insights into the  $L_p$  ( $0 < p < 1$ ) related methods and reveals its potential application in diverse fields of compressed sensing.

**Acknowledgments.** This work is jointly supported by Natural Science Foundation of China (NSFC) under Grant No. 61673119 and the Shanghai Committee of Science and Technology, China under Grant No. 14DZ1118700.

## References

1. Fan, J.Q., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **96**, 1348–1360 (2001)
2. Donoho, D.L.: Compressed sensing. *IEEE Trans. Inf. Theory* **52**, 1289–1306 (2006)
3. Candes, E.J., Romberg, J., Tao, T.: Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure. Appl. Math.* **59**, 1207–1223 (2006)
4. Donoho, D.L.: Neighbourly polytypes and the sparse solution of under-determined systems of linear equations. *IEEE Trans. Inf. Theory* (2005, to appear)
5. Donoho, D.L.: High-dimensional centrally symmetric polytypes with neighbour proportional to dimension. *Discrete Comput. Geom.* **35**, 617–652 (2006)
6. Zou, H.: The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.* **101**, 1418–1429 (2006)
7. Knight, K., Fu, W.: Asymptotics for Lasso-type estimators. *Ann. Stat.* **28**, 1356–1378 (2000)
8. Huang, J., Horowitz, J., Ma, S.: Asymptotic properties for bridge estimators in sparse high-dimensional regression models. *Ann. Stat.* **36**, 587–613 (2008)
9. He, X.N., Lu, W.L., Chen, T.P.: A note on adaptive  $L_p$  regularization. In: *The 2012 International Joint Conference on Neural Networks* (2012)
10. Meinshausen, N., Yu, B.: Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Stat.* **37**, 246–270 (2009)
11. Huang, J., Ma, S., Zhang, C.: Adaptive lasso for sparse high dimensional regression models. *Stat. Sinica* **18**, 1603–1618 (2008)
12. Xu, Z.B., et al.:  $L_{1/2}$  regularization. *SCIENCE CHINA-Inf. Sci.* **53**, 1159–1169 (2010)
13. Xu, Z.B., et al.:  $L_{1/2}$  regularization: a thresholding representation theory and a fast solver. *IEEE Trans. Neural Netw. Learn. Syst.* **23**, 1013–1027 (2012)
14. Vaart, A.W., Wellner, J.A.: *Weak Convergence and Empirical Processes*, pp. 16–28. Springer, New York (1996)
15. Marial, J.: SPARse Modeling Software: an optimization toolbox for solving various sparse estimation problems. <http://spams-devel.gforge.inria.fr/>
16. Tibshirani, R.: Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Series B (Methodological)* **58**, 267–288 (1996)
17. Peter, C., Grace, W.: Smoothing noisy data with spline functions. *Numer. Math.* **31**, 377–403 (1978)
18. Chen, S., Donoho, D.L., Saunders, M.: Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.* **20**, 33–61 (1998)
19. Candes, E.J., Romberg, J., Tao, T.: Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theory* **52**, 489–509 (2006)
20. Candes, E.J., Tao, T.: Near-optimal signal recovery from random projections: universal encoding strategies? *IEEE Trans. Inf. Theory* **52**, 5406–5425 (2006)
21. Candes, E.J.: The restricted isometry property and its application for compressed sensing. *C.R. Math.* **346**, 589–592 (2006)
22. Candes, E.J., Wakin, M.B., Boyd, S.P.: Enhancing sparsity by reweighted  $\ell_1$  minimization. *J. Fourier Anal. Appl.* **14**, 877–905 (2008)
23. Candes, E.J., Tao, T.: Decoding by linear programming. *IEEE Trans. Inf. Theory* **51**, 4203–4215 (2005)
24. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011)

25. Wang, Z., et al.: Image quality assesment: From error visibility to structral similarity. IEEE Trans. Image Process. **13**, 600–612 (2004)
26. Wang, Z., Simoncelli, E.P.: Multiscale structural similarity for image quality assessment. In: Conference Record of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers, pp. 1398–1402 (2004)
27. Wang, Z., Bovik, A.C.: Mean squared error: love it or leave it? - A new look at signal fidelity measures. IEEE Signal Process. Mag. **26**, 98–117 (2009)

Neural Information Processing

24th International Conference, ICONIP 2017,

Guangzhou, China, November 14-18, 2017,

Proceedings, Part I

Liu, D.; Xie, S.; Li, Y.; Zhao, D.; El-Alfy, E.-S.M. (Eds.)

2017, XVIII, 936 p. 263 illus., Softcover

ISBN: 978-3-319-70086-1