Nonconvex Sparse Optimization and Applications

Xianchao Xiu

Department of Automation

上海大学
Shanghai University

Chinese Academy of Sciences, January 12, 2025

Joint work with Wanquan Liu (SYSU), Lingchen Kong (BJTU) and others

Outline

Introduction

Optimization

Applications

Future Work

Sparse Optimization

- Sparse optimization considers

$$\min_{x\in\mathbb{R}^n} \ f(x) + \lambda\|x\|_0$$

$$\Updownarrow$$

$$\min_{x\in\mathbb{R}^n} \ f(x) \quad \text{s.t.} \quad \|x\|_0 \le s$$

  - x can be extended to matrices and tensors
  - $f(x)$ may be nonsmooth even nonconvex
  - $\|x\|_0$ counts the number of nonzeros
  - λ and s are parameters
- Also called compressed sensing and variable selection
- Broad applications in machine learning, pattern recognition and engineering
- `https://github.com/xianchaoxiu/Sparse-Optimization`

Methods

- Convex methods
  - Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society Series B, 1996
  - Candès-Tao, Decoding by linear programming, IEEE TIT, 2005
  - Donoho, Compressed sensing, IEEE TIT, 2006
- Nonconvex methods
  - Fan-Li, Variable selection via nonconcave penalized likelihood and its oracle properties, Journal of the American Statistical Association, 2001
  - Chen-Xu-Ye, Lower bound theory of nonzero entries in solutions of $\ell_2 - \ell_p$ minimization, SIAM Journal on Scientific Computing, 2010
  - Xu-Chang-Xu-Zhang, $L_{1/2}$ regularization: A thresholding representation theory and a fast solver, IEEE TNNLS, 2012
- Direct methods
  - Blumensath-Davies, Iterative hard thresholding for compressed sensing, Applied and Computational Harmonic Analysis, 2009
  - Foucart, Hard thresholding pursuit: An algorithm for compressive sensing, SIAM Journal on Numerical Analysis, 2011
  - Yuan-Li-Zhang, Gradient hard thresholding pursuit, JMLR, 2018

More

- Bach-Jenatton-Mairal-Obozinski, Optimization with sparsity-inducing penalties, Foundations and Trends in Machine Learning, 2012
- Jain-Kar, Non-convex optimization for machine learning, Foundations and Trends in Machine Learning, 2017
- Hastie-Tibshirani-Wainwright, Statistical learning with sparsity: The Lasso and generalizations, CRC Press, 2015
- Zhao, Sparse optimization theory and methods, CRC Press, 2018
- Fan-Li-Zhang-Zou, Statistical foundations of data science, CRC Press, 2020
- Wright-Ma, High-dimensional data analysis with low-dimensional models: Principles, computation, and applications, Cambridge University Press, 2022
- Parhi-Nowak, Deep learning meets sparse regularization: A signal processing perspective, IEEE Signal Processing Magazine, 2023
- Tillmann-Bienstock-Lodi-Schwartz, Cardinality minimization, constraints, and regularization: A survey, SIAM Review, 2024

Outline

Introduction

Optimization

Applications

Future Work

$\ell_1 - \ell_p$ Minimization

- Xiu-Kong-Li-Qi, Computational Optimization and Applications, 2018

$$\min_{x \in \mathbb{R}^n} \quad \|Ax - b\|_1 + \lambda \|x\|_p^p \; (0 < p < 1) \tag{1}$$

- Consider the following ϵ-approximations

$$\min_{x \in \mathbb{R}^n} F_{\alpha,\epsilon}(x) = \|Ax - b\|_1 + \lambda \sum_{i=1}^{n} (|x_i|^\alpha + \epsilon_i)^{\frac{p}{\alpha}}$$

$$\Downarrow$$

$$\min_{x \in \mathbb{R}^n} F_\epsilon(x) = \|Ax - b\|_1 + \lambda \sum_{i=1}^{n} h_{u_\epsilon}(x_i) \tag{2}$$

where

$$h_{u_\epsilon}(x_i) = \min_{0 \le s \le u_\epsilon} p\left(|x_i| s - \frac{p-1}{p} s^{\frac{p}{p-1}}\right), \quad u_\epsilon = \left(\frac{\epsilon}{\lambda n}\right)^{\frac{p-1}{p}}$$

$\ell_1 - \ell_p$ Minimization

- (Definition) We say that $x^* \in \mathbb{R}^n$ is a generalized first-order stationary point of (1) if

$$0 \in (A^\top \mathrm{sgn}(Ax^* - b))_i x_i^* + \lambda p |x_i^*|^p, \quad i = 1, 2, \cdots, n$$

Furthermore, the following statement holds

$$|x_i^*| \ \geq \ (\frac{\lambda p}{\|A_i\|_1})^{\frac{1}{1-p}}, \quad \forall i \in T \tag{3}$$

- (Lower Bound) Let ϵ be a constant such that

$$0 < \epsilon < \lambda n \left(\frac{\|A_i\|_1}{\lambda p} \right)^{\frac{p}{p-1}} \tag{4}$$

Suppose that x^* is a generalized first-order stationary point of (2). Then, x^* is also a generalized first-order stationary point of (1). Moreover, the nonzero entries of x^* satisfy the lower bound property (3).

$\ell_1 - \ell_p$ Minimization

- (Convergent Theorem) Assume that ϵ satisfies (4) and set q as $\frac{1}{p} + \frac{1}{q} = 1$. Suppose that x^* is an accumulation point of $\{x^k\}$. Then x^* is a generalized first-order stationary point of (1). Moreover, the nonzero entries of x^* satisfy the lower bound (3).

Choose an arbitrary $x^0 \in \mathbb{R}^n$ and ϵ such that (4) holds. Set $k = 0$

1) Solve the weighted ℓ_1 minimization problem

$$x^{k+1} \in \mathrm{argmin}_x \left\{\|Ax - b\|_1 + \lambda p \sum_{i=1}^n s_i^k |x_i|\right\}$$

where $s_i^k = \min\left\{\left(\frac{\epsilon}{\lambda n}\right)^{\frac{1}{q}}, |x_i^k|^{\frac{1}{q-1}}\right\}$ for all i

2) Set $k \leftarrow k + 1$ and go to step 1)

End

Distributed Optimization

- Qu-Chen-Xiu-Liu, Neurocomputing, 2024

$$\min_{Y \in \mathbb{R}^{n\times p}} \quad \sum_{i=1}^{d} f_i(Y) \qquad (5)$$
$$\text{s.t.} \quad \|Y\|_{2,0} \leq s, \;\; Y^\top Y = I_p$$

$$\Downarrow$$

$$\min_{Y \in \mathbb{R}^{n\times p}} \quad \sum_{i=1}^{d} f_i(Y) + \frac{\mu}{4}\|Y^\top Y - I_p\|_F^2$$
$$\text{s.t.} \quad \|Y\|_{2,0} \leq s$$

$$\Downarrow$$

$$\min_{Y,\{X_i\} \in \mathbb{R}^{n\times p}} \quad \sum_{i=1}^{d} f_i(X_i) + \frac{\mu}{4}\|Y^\top Y - I_p\|_F^2 \qquad (6)$$
$$\text{s.t.} \quad X_i = Y, \; \forall i \in [d], \;\; \|Y\|_{2,0} \leq s$$

Distributed Optimization

- (Lemma) Let $(\widetilde{Y}^*, \{\widetilde{X}_i^*\})$ be the (local) minimizer of (6). Then there exists $\mu_\epsilon > 0$ such that $\widetilde{Y}^*$ is an ϵ-(local) minimizer of (5) for any $\mu \geq \mu_\epsilon$.
- (Definition) We say $(Y^*, \{X_i^*\}, \{\Lambda_i^*\})$ is a KKT point of (6) if it satisfies

$$\begin{cases} 0 \in \nabla g(Y^*) + \sum_{i=1}^{d} \Lambda_i^* + \mathcal{N}_{\mathcal{S}}(Y^*) \\ 0 = \nabla f_i(X_i^*) - \Lambda_i^*, \ \forall i \in [d] \\ 0 = X_i^* - Y^*, \ \forall i \in [d] \end{cases}$$

- (Definition) We say $(Y^*, \{X_i^*\}, \{\Lambda_i^*\})$ is a stationary point of (6) if there exists $\alpha > 0$ such that

$$\begin{cases} Y^* = \mathcal{P}_{\mathcal{S}}(Y^* - \alpha(\nabla g(Y^*) + \sum_{i=1}^{d} \Lambda_i^*)) \\ 0 = \nabla f_i(X_i^*) - \Lambda_i^*, \ \forall i \in [d] \\ 0 = X_i^* - Y^*, \ \forall i \in [d] \end{cases}$$

- (Optimal Conditions) Suppose that $(Y^*, \{X_i^*\})$ is a local minimizer of (6). Then, there exists Λ_i^* $(i \in [d])$ such that $(Y^*, \{X_i^*\}, \{\Lambda_i^*\})$ is a KKT point of (6).

Distributed Optimization

- (Nonincreasing Lemma) Let $\{(Y^k, \{X_i^k\}, \{\Lambda_i^k\})\}$ be the generated sequence and $\beta \geq \sqrt{2}r$. Then the generated augmented Lagrangian sequence is nonincreasing, i.e.,

$$\mathcal{L}_{\beta}(Y^{k+1}, \{X_i^{k+1}\}; \{\Lambda_i^{k+1}\}) \leq \mathcal{L}_{\beta}(Y^k, \{X_i^k\}; \{\Lambda_i^k\})$$

- (Bounded Lemma) Suppose that $\beta \geq 2r$ holds. Then the sequence $\{(Y^k, \{X_i^k\}, \{\Lambda_i^k\})\}$ is bounded. Moreover, it satisfies

$$\begin{cases} \lim\limits_{k\to\infty} \|Y^{k+1} - Y^k\|_F = 0 \\ \lim\limits_{k\to\infty} \|X_i^{k+1} - X_i^k\|_F = 0, \ \forall i \in [d] \\ \lim\limits_{k\to\infty} \|\Lambda_i^{k+1} - \Lambda_i^k\|_F = 0, \ \forall i \in [d] \end{cases}$$

- (Convergent Theorem) Let $\{(Y^k, \{X_i^k\}, \{\Lambda_i^k\})\}$ be the generated sequence and $\beta \geq 2r$. Then, any accumulation point $(Y^*, \{X_i^*\}, \{\Lambda_i^*\})$ is a stationary point of (6).

$L_{1/2}$ Reglarization

- Fan-Yan-Xiu-Liu, under review

$$\min_{x \in \mathbb{H}^p} F(x) := f(x) + \lambda \|x\|_{1/2}^{1/2} \; (\mathbb{H} = \mathbb{R} \text{ or } \mathbb{C}) \tag{7}$$

where $f(x) := \frac{1}{n} \sum_{i=1}^{n} h_\alpha(|\langle a_i, x\rangle|^2 - b_i)$ and

$$h_\alpha(u) = \begin{cases} \frac{1}{2}u^2, & \text{if } |u| \leq \alpha \\ \alpha|u| - \frac{1}{2}\alpha^2, & \text{if } |u| > \alpha \end{cases}$$

- For ease of expression, define

$$g(x) := \frac{1}{n} \sum_{i=1}^{n} h'_\alpha(|\langle a_i, x\rangle|^2 - b_i)\langle a_i, \rangle \bar{a}_i$$

which implies that $\nabla f(x) = 2g(x)$ for $\mathbb{H} = \mathbb{R}$ and $\nabla_x f(x) = g(x)$ for $\mathbb{H} = \mathbb{C}$

- Wirtinger derivative

$L_{1/2}$ Reglarization

- (Optimal Conditions) There exists a global minimizer $\hat{x}$ which lies in the level set $S = \{x \in \mathbb{H}^p : F(x) \leq F(x^0)\}$, and further satisfies the fixed point inclusion

$$\hat{x} \in \mathcal{H}_{\lambda\tau}(\hat{x} - 2\tau g(\hat{x}))$$

Initialize spectral point x^0, let $k = 0,\ j = 0,\ \tau_0 = \beta$

1) Do

$$x^{k+1} = \mathcal{H}_{\lambda\tau_k}(x^k - 2\tau_k \nabla f(x^k))$$

where $\tau_k = \gamma\beta^{j_k}$ and j_k is the smallest nonnegative integer such that

$$F(x^k) - F(x^{k+1}) \geq \delta\|x^{k+1} - x^k\|^2$$

2) Check convergence: if

$$\|x^{k+1} - x^k\| \leq \epsilon \max\{1, \|x^k\|\}$$

otherwise, set $k \leftarrow k+1$, and go back to Step 1)

$L_{1/2}$ Reglarization

- (Subsequence Convergence) Assume that $\{x^k\}$ is the generated sequence. Then the following conclusions hold
  (a) $\lim_{k\to\infty} \|x^{k+1} - x^k\| = 0$.
  (b) Every accumulation point of $\{x^k\}$ satisfies the following fixed point equation
  $$x = \mathcal{H}_{\lambda\tau}(x - 2\tau g(x)) \tag{8}$$
  for $\tau \in [\gamma\beta^{\bar{j}}, \gamma]$ when $\gamma \leq \lambda^2/(64\ell^3)$ with $\ell = \alpha \sum_{i=1}^{n} \|a_i\|^2 \sup_{F(x)\leq F(x^0)} \|x\|/n$.
  (c) $\{F(x^k)\}$ decreasingly converges to $F(x^*)$, where x^* is any accumulation point of $\{x^k\}$.
- (Whole Sequence Convergence) Assume that $\{x^k\}$ is the generated sequence. Then the whole sequence $\{x^k\}$ is convergent.
- (Convergence Rate) Under mild conditions, the whole sequence $\{x^k\}$ is convergent and converges at least sublinearly to a vector x^* satisfying (8).

Outline

Data Analysis

- Xiu-Liu-Li-Kong, Computational Statistics & Data Analysis, 2019

$$\min_{\beta} \quad \frac{1}{2}\|y - X\beta\|^2 + \Phi_{\tau_1}(\beta) + \sum_{i=1}^{p} \Phi_{\tau_2}(\beta_{i+1} - \beta_i)$$

  - Φ_{τ_1} and Φ_{τ_2} can be the same or different
  - Nonconvex penalty functions: ℓ_p, SCAD, MCP, capped ℓ_1
- For notational simplicity, define

$$\min_{\beta} \quad \frac{1}{2}\|y - X\beta\|^2 + \Phi_{\tau_1}(\beta) + \Phi_{\tau_2}(D\beta)$$

with

$$D = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & -1 & 1 \end{pmatrix} \in \mathbb{R}^{(p-1)\times p}$$

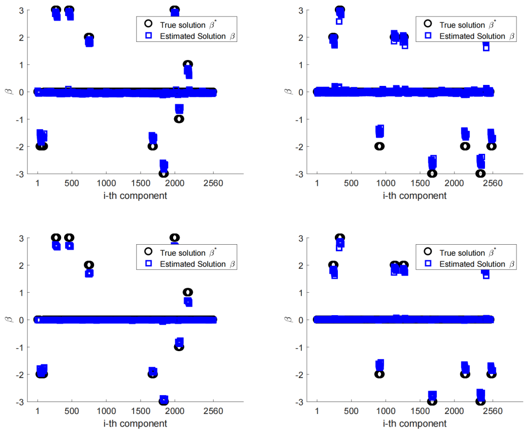
Data Analysis

- Alternating direction method of multipliers (ADMM)

$$\begin{aligned} \min_{\alpha,\gamma,\beta} \quad & \frac{1}{2}\|y - X\beta\|^2 + \Phi_{\tau_1}(\alpha) + \Phi_{\tau_2}(\gamma) \\ \text{s.t.} \quad & \alpha = \beta \\ & \gamma = D\beta \end{aligned}$$

- (Convergent Theorem) Suppose that $\{(\alpha^k, \gamma^k, \beta^k, w_1^k, w_2^k)\}$ is a generated sequence. Then the sequence converges to a stationary point.

- Recovery results



Signal Processing

- Li-Xiu-Liu-Miao, IEEE Signal Processing Letters, 2022

$$\begin{aligned} \min_{U,P_v} \quad & \sum_{v=1}^{M} \|U - X_v P_v\|_F^2 \\ \text{s.t.} \quad & U^\top U = I_d, \; \|P_v\|_{2,0} \le s_v \end{aligned}$$

- Alternating minimization algorithm (AMA): Update U, then update P_v
- Denote $f(P_v) := \|U^{k+1} - X_v P_v\|_F^2$. Then

$$\nabla f(P_v) = 2X_v^\top (X_v P_v - U^{k+1}) \in \mathbb{R}^{d_v \times d}$$

$$\nabla^2 f(P_v) = 2I_d \otimes X_v^\top X_v \in \mathbb{R}^{d_v d \times d_v d}$$

- Newton hard thresholding pursuit (NHTP)

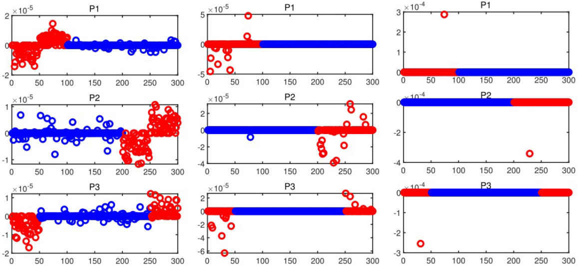
$$F(P_v; T_v) := \begin{pmatrix} \nabla_{T_v} f(P_v) \\ (P_v)_{\overline{T}_v} \end{pmatrix} = 0$$

Signal Processing

- Runtime comparison

Problem Scale	GCCA	SGCCA	SCGCCA
(1,000;300;300;300)	0.04	0.04	**0.01**
(5,000;300;300;300)	0.23	0.28	**0.03**
(10,000;300;300;300)	0.40	0.41	**0.07**
(50,000;300;300;300)	2.32	2.27	**0.34**
(100,000;300;300;300)	4.58	4.35	**0.66**
(1,000;1,500;1,500;1,500)	0.42	0.40	**0.02**
(5,000;1,500;1,500;1,500)	1.35	1.16	**0.12**
(10,000;1,500;1,500;1,500)	2.63	2.24	**0.24**
(50,000;1,500;1,500;1,500)	13.21	10.56	**1.18**
(100,000;1,500;1,500;1,500)	26.60	22.53	**2.35**
(1,000;3,000;3,000;3,000)	1.53	1.58	**0.17**
(5,000;3,000;3,000;3,000)	3.92	3.49	**0.23**
(10,000;3,000;3,000;3,000)	6.87	5.65	**0.45**
(50,000;3,000;3,000;3,000)	32.02	23.18	**2.29**
(100,000;3,000;3,000;3,000)	667.69	629.54	**4.91**

- Extracted feature comparison



Pattern Recognition

- Liu-Feng-Xiu-Liu, Pattern Recognition, 2024

$$\begin{aligned} \min_{Q} \quad & \mathrm{Tr}\left(Q^\top S Q\right) + \lambda \|Q\|_{2,1} \\ \text{s.t.} \quad & Q^\top Q = I \end{aligned}$$

$$\Downarrow$$

$$\begin{aligned} \min_{P,Q,E} \quad & \mathrm{Tr}\left(Q^\top S Q\right) + \lambda_1 \|Q\|_{2,1} + \lambda_2 \|E\|_1 \\ \text{s.t.} \quad & X = PQ^\top X + E, \;\; P^\top P = I \end{aligned}$$

$$\Downarrow$$

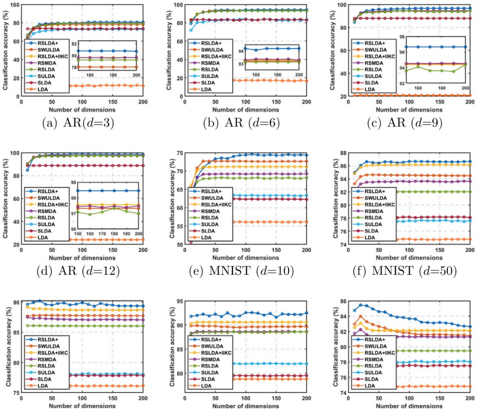
$$\begin{aligned} \min_{P,Q,E} \quad & \mathrm{Tr}\left(Q^\top S Q\right) + \lambda_1 \|Q\|_{2,0} + \lambda_2 \|E\|_0 \\ \text{s.t.} \quad & X = PQ^\top X + E, \;\; P^\top P = I \end{aligned}$$

$$\Downarrow$$

$$\begin{aligned} \min_{P,Q,E} \quad & \mathrm{Tr}\left(Q^\top S Q\right) + \lambda_1 \|Q\|_{2,0} + \lambda_2 \|Q\|_0 + \lambda_3 \|E\|_0 \\ \text{s.t.} \quad & X = PQ^\top X + E, \;\; P^\top P = I \end{aligned}$$

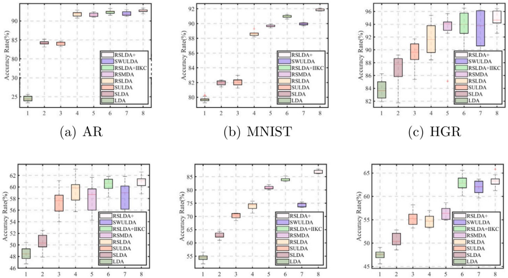
Pattern Recognition

▶ Classification accuracy



(a) AR(d=3)
(b) AR (d=6)
(c) AR (d=9)
(d) AR (d=12)
(e) MNIST (d=10)
(f) MNIST (d=50)
(g) MNIST (d=100)
(h) MNIST (d=200)
(i) HGR (d=4)

▶ Model stability

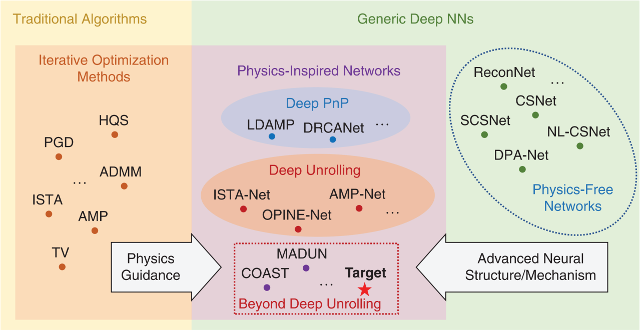


(a) AR
(b) MNIST
(c) HGR
(d) NEU
(e) Caltech
(f) CAR

Outline

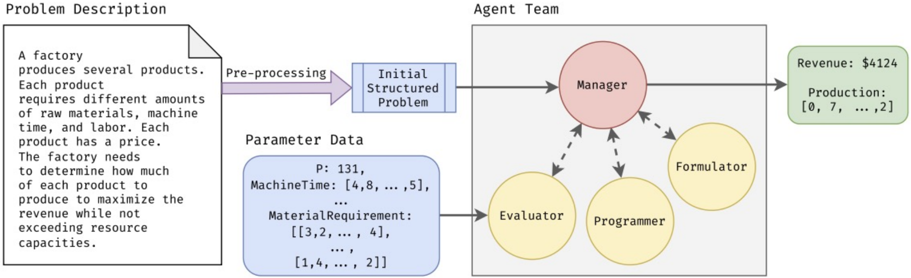
Deep Unfolding Networks

- Gregor-LeCun, Learning fast approximations of sparse coding, ICML, 2010
- Zhang-Ghanem, ISTA-Net: Interpretable optimization-inspired deep network for image compressive sensing, CVPR, 2018
- Chen-Liu-Yin, Learning to optimize: A tutorial for continuous and mixed-integer optimization, Science China Mathematics, 2024



Large Language Models

- Yang-Wang-Lu et al., Large language models as optimizers, ICLR, 2024
- AhmadiTeshnizi-Gao-Udell, OptiMUS: Scalable optimization modeling with (MI) LP solvers and large language models, ICML, 2024
- Romera-Barekatain-Novikov et al., Mathematical discoveries from program search with large language models, Nature, 2024



References

- Fan-Yan-Xiu-Liu, Robust sparse phase retrieval: Model, theoretical guarantee and efficient algorithm, under review
- Liu-Feng-Xiu-Liu, Towards robust and sparse linear discriminant analysis for image classification, Pattern Recognition, 2024
- Qu-Chen-Xiu-Liu, Distributed sparsity constrained optimization over the Stiefel manifold, Neurocomputing, 2024
- Li-Xiu-Liu-Miao, An efficient Newton-based method for sparse generalized canonical correlation analysis, IEEE Signal Processing Letters, 2022
- Xiu-Liu-Li-Kong, Alternating direction method of multipliers for nonconvex fused regression problems, Computational Statistics & Data Analysis, 2019
- Xiu-Kong-Li-Qi, Iterative reweighted methods for $\ell_1 - \ell_p$ minimization, Computational Optimization and Applications, 2018