Proximal Iteratively Reweighted Algorithm with Multiple Splitting for Nonconvex Sparsity Optimization

Canyi Lu[1], **Yunchao Wei**[2], **Zhouchen Lin**[3,*], **Shuicheng Yan**[1]
[1] Department of Electrical and Computer Engineering, National University of Singapore
[2] Institute of Information Science, Beijing Jiaotong University
[3] Key Laboratory of Machine Perception (MOE), School of EECS, Peking University
canyilu@gmail.com, wychao1987@gmail.com, zlin@pku.edu.cn, eleyans@nus.edu.sg

Abstract

This paper proposes the Proximal Iteratively REweighted (PIRE) algorithm for solving a general problem, which involves a large body of nonconvex sparse and structured sparse related problems. Comparing with previous iterative solvers for nonconvex sparse problem, PIRE is much more general and efficient. The computational cost of PIRE in each iteration is usually as low as the state-of-the-art convex solvers. We further propose the PIRE algorithm with Parallel Splitting (PIRE-PS) and PIRE algorithm with Alternative Updating (PIRE-AU) to handle the multi-variable problems. In theory, we prove that our proposed methods converge and any limit solution is a stationary point. Extensive experiments on both synthesis and real data sets demonstrate that our methods achieve comparative learning performance, but are much more efficient, by comparing with previous nonconvex solvers.

Introduction

This paper aims to solve the following general problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} F(\mathbf{x}) = \lambda f(\mathbf{g}(\mathbf{x})) + h(\mathbf{x}), \tag{1}$$

where $\lambda > 0$ is a parameter, and the functions in the above formulation satisfy the following conditions:

C1 $f(\mathbf{y})$ is nonnegative, concave and increasing.

C2 $\mathbf{g}(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}^d$ is a nonnegative multi-dimensional function, such that the following problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} \lambda\langle \mathbf{w}, \mathbf{g}(\mathbf{x})\rangle + \frac{1}{2}||\mathbf{x}-\mathbf{b}||_2^2, \tag{2}$$

is convex and can be cheaply solved for any given nonnegative $\mathbf{w} \in \mathbb{R}^d$.

C3 $h(\mathbf{x})$ is a smooth function of type $C^{1,1}$, i.e., continuously differentiable with the Lipschitz continuous gradient

$$||\nabla h(\mathbf{x}) - \nabla h(\mathbf{y})|| \le L(h)||\mathbf{x}-\mathbf{y}|| \ \text{ for any } \ \mathbf{x},\mathbf{y}\in\mathbb{R}^n, \tag{3}$$

$L(h) > 0$ is called the Lipschitz constant of ∇h.

C4 $\lambda f(\mathbf{g}(\mathbf{x})) + h(\mathbf{x}) \to \infty$ iff $||\mathbf{x}||_2 \to \infty$.

*Corresponding author.

Note that problem (1) can be convex or nonconvex. Though $f(\mathbf{y})$ is concave, $f(\mathbf{g}(\mathbf{x}))$ can be convex w.r.t $\mathbf{x}$. Also $f(\mathbf{y})$ and $\mathbf{g}(\mathbf{x})$ are not necessarily smooth, and $h(\mathbf{x})$ is not necessarily convex.

Based on different choices of f, $\mathbf{g}$, and h, the general problem (1) involves many sparse representation models, which have many important applications in machine learning and computer vision (Wright et al. 2009; Beck and Teboulle 2009; Jacob, Obozinski, and Vert 2009; Gong, Ye, and Zhang 2012b). For the choice of h, the least square and logistic loss functions are two most widely used ones which satisfy (**C3**):

$$h(\mathbf{x}) = \frac{1}{2}||\mathbf{A}\mathbf{x}-\mathbf{b}||_2^2, \text{or } \frac{1}{n}\sum_{i=1}^{n}\log(1+\exp(-b_i\mathbf{a}_i^T\mathbf{x})), \tag{4}$$

where $\mathbf{A} = [\mathbf{a}_1^T;\cdots;\mathbf{a}_n^T] \in \mathbb{R}^{n\times d}$, and $\mathbf{b} \in \mathbb{R}^n$. As for the choice of $\mathbf{g}(\mathbf{x})$, $|\mathbf{x}|$ (absolute value of $\mathbf{x}$ element-wise) and $\mathbf{x}^2$ (square of $\mathbf{x}$ element-wise) are widely used. One may also use $\mathbf{g}(\mathbf{X}) = ||\mathbf{x}_i||_2$ ($\mathbf{x}_i$ denotes the i-th column of $\mathbf{X}$) when pursuing column sparsity of a matrix $\mathbf{X}$. As for the choice of f, almost all the existing nonconvex surrogate functions of the ℓ_0-norm are concave on $(0,\infty)$. In element-wise, they include ℓ_p-norm y^p $(0 < p < 1)$ (Knight and Fu 2000), logarithm function $\log(y)$ (Candès, Wakin, and Boyd 2008), smoothly clipped absolute deviation (Fan and Li 2001), and minimax concave penalty (Zhang 2010).

The above nonconvex penalties can be further extended to define structured sparsity (Jacob, Obozinski, and Vert 2009). For example, let $\mathbf{x} = [\mathbf{x}_1;\cdots;\mathbf{x}_G]$. By taking $g(\mathbf{x}) = [||\mathbf{x}_1||_2;\cdots;||\mathbf{x}_G||_2]$ and $f(\mathbf{y}) = \sum_i f_i(y_i)$, with f_i being any of the above concave functions, then $f(\mathbf{g}(\mathbf{x}))$ is the nonconvex group Lasso $\sum_i f_i(||\mathbf{x}_i||_2)$. By taking $f(\mathbf{y}) = \sum_i y_i$, $f(\mathbf{g}(\mathbf{x})) = \sum_i ||\mathbf{x}_i||_2$ is the group Lasso.

Problem (1) contains only one variable. We will show that our proposed model can be naturally used for handling problem with several variables (which we mean a group of variables that can be updated simultaneously due to the separability structure of the problem). An example for multi-task learning can be found in (Gong, Ye, and Zhang 2012b).

Related Works

If the condition (**C3**) holds and

$$\min_{\mathbf{x}} \lambda f(\mathbf{g}(\mathbf{x})) + \frac{1}{2}||\mathbf{x}-\mathbf{b}||_2^2, \tag{5}$$

can be cheaply computed, then problem (1) can be solved by iteratively solving a series of problem (5) (Gong et al. 2013). Such an updating procedure is the same as the ISTA algorithm (Beck and Teboulle 2009), which is originally for convex optimization. It can be proved that any accumulation point of $\{\mathbf{x}^k\}$ is a stationary point of problem (1). If $f(\mathbf{g}(\mathbf{x}))$ and $h(\mathbf{x})$ are convex, the Fast ISTA algorithm (FISTA) (Beck and Teboulle 2009) converges to the globally optimal solution with a convergence rate $O(1/T^2)$ (T is the iteration number). But for nonconvex optimization, it is usually very difficult to get the globally optimal solution to problem (5). Sometimes, it is also not easy even if $f(\mathbf{g}(\mathbf{x}))$ is convex.

The multi-stage algorithm in (Zhang 2008) solves problem (1) by solving a series of convex problem.

$$\min_{\mathbf{x}} \lambda\langle \mathbf{w}^k, \mathbf{g}(\mathbf{x})\rangle + h(\mathbf{x}). \tag{6}$$

However, solving such a convex problem requires other iterative solvers which is not efficient. It also fails when $h(\mathbf{x})$ is nonconvex.

More specially, the Iteratively Reweighted L1 (IRL1) (Chen and Zhou) and Iteratively Reweighted Least Squares (IRLS) (Lai, Xu, and Yin 2013) algorithms are special cases of the multi-stage algorithm. They aim to solve the following ℓ_p-regularization problem

$$\min_{\mathbf{x}} \lambda||\mathbf{x}||_p^p + \frac{1}{2}||\mathbf{A}\mathbf{x}-\mathbf{b}||_2^2. \tag{7}$$

The above problem is NP-hard. IRL1 instead considers the following relaxed problem

$$\min_{\mathbf{x}} \lambda\sum_{i=1}^{n}(|x_i|+\epsilon)^p + \frac{1}{2}||\mathbf{A}\mathbf{x}-\mathbf{b}||_2^2, \tag{8}$$

with $0 < \epsilon \ll 1$. IRL1 updates $\mathbf{x}^{k+1}$ by solving

$$\mathbf{x}^{k+1} = \arg\min_{\mathbf{x}} \lambda\sum_{i=1}^{n} w_i^k|x_i| + \frac{1}{2}||\mathbf{A}\mathbf{x}-\mathbf{b}||_2^2, \tag{9}$$

with $w_i^k = p/(|x_i^k|+\epsilon)^{1-p}$. Problem (8) is a special case of (1) by letting $f(\mathbf{y}) = \sum_i (y_i+\epsilon)^p$ ($0 < p < 1$) and $\mathbf{g}(\mathbf{x}) = |\mathbf{x}|$. However, IRL1 is not efficient since it has to solve a number of nonsmooth problem (9) by using some other convex optimization methods, e.g. FISTA.

The other method, IRLS, smooths problem (7) as

$$\min \lambda\sum_{i=1}^{n}(x_i^2+\epsilon)^{\frac{p}{2}} + \frac{1}{2}||\mathbf{A}\mathbf{x}-\mathbf{b}||_2^2, \tag{10}$$

and updates $\mathbf{x}^{k+1}$ by solving

$$\lambda \text{Diag}(\mathbf{w}^k)\mathbf{x} + \mathbf{A}^T(\mathbf{A}\mathbf{x}-\mathbf{b}) = 0, \tag{11}$$

with $w_i^k = p/\left((x_i^k)^2+\epsilon\right)^{1-\frac{p}{2}}$. Problem (10) is also a special case of (1) by taking $f(\mathbf{y}) = \sum_i (y_i+\epsilon)^{\frac{p}{2}}$ and $\mathbf{g}(\mathbf{x}) = \mathbf{x}^2$. However, the obtained solution by IRLS may not be naturally sparse, or it may require a lot of iterations to get a sparse solution. One may perform thresholding appropriately to achieve a sparse solution, but there is no theoretically sound rule to suggest a correct threshold.

Another related work is (Lu 2012) which aims to solve

$$\min_{\mathbf{x}} \lambda\sum_{i=1}^{n}(|x_i|+\epsilon)^p + h(\mathbf{x}). \tag{12}$$

In each iteration, $\mathbf{x}$ is efficiently updated by solving a series of problem

$$\min_{\mathbf{x}} \lambda\langle \mathbf{w}^k, |\mathbf{x}|\rangle + \frac{1}{2}||\mathbf{x}-\mathbf{b}||_2^2. \tag{13}$$

But their solver is only for problem (12) which is a special case of (1). The convergence proofs also depend on the special property of the ℓ_p-norm, thus is not general.

Furthermore, previous iterative algorithms can only solve the problem with only one variable. They cannot be naively generalized to solve multi-variable problems. However, there are many problems involving two or more variables, e.g. stable robust principle component analysis (Zhou et al. 2010) and robust multi-task feature learning (Gong, Ye, and Zhang 2012b). So it is desirable to extend the iteratively reweighted algorithms for the multi-variable case.

Contributions

In this work, we propose a novel method to solve the general problem (1), and address the scalablity and multi-variable issues. In each iteration we only need to solve problem (2), whose computational cost is usually the same as previous state-of-the-art first-order convex methods. This method is named as Proximal Iteratively REweighted (PIRE) algorithm. We further propose two multiple splitting versions of PIRE: PIRE with Parallel Splitting (PIRE-PS) and PIRE with Alternative Updating (PIRE-AU) to handle the multi-variable problem. Parallel splitting makes the algorithm highly parallelizable, making PIRE-PS suitable for distributed computing. This is important for large scale applications. PIRE-AU may converge faster than PIRE-PS. In theory, we prove that any sequences generated by PIRE, PIRE-PS and PIRE-AU are bounded and any accumulation point is a stationary point. To the best of our knowledge, PIRE-PS and PIRE-AU are the first two algorithms for problem (1) with multi-variables. If problem (1) is convex, the obtained solution is globally optimal.

Proximal Iteratively Reweighted Algorithm

In this section, we show how to solve problem (1) by our Proximal Iteratively Reweighted (PIRE) algorithm. Instead of minimizing $F(\mathbf{x})$ in (1) directly, we update $\mathbf{x}^{k+1}$ by minimizing the sum of two surrogate functions, which correspond to two terms of $F(\mathbf{x})$, respectively.

First, note that $f(\mathbf{y})$ is concave, $-f(\mathbf{y})$ is convex. By the definition of subgradient of the convex function, we have

$$-f(\mathbf{g}(\mathbf{x})) \geq -f(\mathbf{g}(\mathbf{x}^k)) + \langle -\mathbf{w}^k, \mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{x}^k)\rangle, \tag{14}$$

where $-\mathbf{w}^k$ is the subgradient of $-f(\mathbf{y})$ at $\mathbf{y} = \mathbf{g}(\mathbf{x}^k)$, i.e.

$$-\mathbf{w}^k \in \partial\left(-f(\mathbf{g}(\mathbf{x}^k))\right) \text{ or } \mathbf{w}^k \in -\partial\left(-f(\mathbf{g}(\mathbf{x}^k))\right). \tag{15}$$

Eqn (14) is equivalent to

$$f(\mathbf{g}(\mathbf{x})) \leq f(\mathbf{g}(\mathbf{x}^k)) + \langle \mathbf{w}^k, \mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{x}^k)\rangle. \tag{16}$$

Algorithm 1 Solving problem (1) by PIRE

Input: $\mu > \frac{L(h)}{2}$, where $L(h)$ is the Lipschitz constant of $h(\mathbf{x})$.
Initialize: $k = 0$, $\mathbf{w}^k$.
Output: $\mathbf{x}^*$.
while not converge **do**

1. Update $\mathbf{x}^{k+1}$ by solving the following problem

$$\mathbf{x}^{k+1} = \arg\min_{\mathbf{x}} \lambda\langle \mathbf{w}^k, \mathbf{g}(\mathbf{x})\rangle + \frac{\mu}{2}\left\|\mathbf{x} - \left(\mathbf{x}^k - \frac{1}{\mu}\nabla h(\mathbf{x}^k)\right)\right\|^2.$$

2. Update the weight $\mathbf{w}^{k+1}$ by

$$\mathbf{w}^{k+1} \in -\partial\left(-f(\mathbf{g}(\mathbf{x}^{k+1}))\right).$$

end while

Then $f(\mathbf{g}(\mathbf{x}^k)) + \langle \mathbf{w}^k, \mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{x}^k)\rangle$ is used as a surrogate function of $f(\mathbf{g}(\mathbf{x}))$.

The loss function $h(\mathbf{x})$, which has Lipschitz continuous gradient, owns the following property (Bertsekas 1999)

$$h(\mathbf{x}) \le h(\mathbf{y}) + \langle \nabla h(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle + \frac{L(h)}{2}||\mathbf{x} - \mathbf{y}||_2^2. \quad (17)$$

Let $\mathbf{y} = \mathbf{x}^k$, $h(\mathbf{x}^k) + \langle \nabla h(\mathbf{x}^k), \mathbf{x} - \mathbf{x}^k\rangle + \frac{L(h)}{2}||\mathbf{x} - \mathbf{x}^k||_2^2$ is used as a surrogate function of $h(\mathbf{x})$.

Combining (16) and (17), we update $\mathbf{x}^{k+1}$ by minimizing the sum of these two surrogate functions

$$\begin{aligned}
&\mathbf{x}^{k+1}\\
&= \arg\min_{\mathbf{x}} f(\mathbf{g}(\mathbf{x}^k)) + \langle \mathbf{w}^k, \mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{x}^k)\rangle\\
&\quad + h(\mathbf{x}^k) + \langle \nabla h(\mathbf{x}^k), \mathbf{x} - \mathbf{x}^k\rangle + \frac{\mu}{2}||\mathbf{x} - \mathbf{x}^k||_2^2\\
&= \arg\min_{\mathbf{x}} \lambda\langle \mathbf{w}^k, \mathbf{g}(\mathbf{x})\rangle + \frac{\mu}{2}\left\|\mathbf{x} - \left(\mathbf{x}^k - \frac{1}{\mu}\nabla h(\mathbf{x}^k)\right)\right\|_2^2,
\end{aligned} \quad (18)$$

where $\mathbf{w}^k$ is also called the weight corresponding to $\mathbf{g}(\mathbf{x}^k)$.

For the choice of μ in (18), our theoretical analysis shows that $\mu > L(h)/2$ guarantees the convergence of the proposed algorithm. Note that f is concave and increasing, this guarantees that $\mathbf{w}^k$ in (15) is nonnegative. Usually problem (18) can be cheaply computed based on the condition (**C2**). For example, if $\mathbf{g}(\mathbf{x}) = |\mathbf{x}|$, solving problem (18) costs only $O(n)$. Such computational cost is the same as the state-of-the-art convex solvers for ℓ_1-minimization. This idea leads to the Proximal Iteratively REweighted (PIRE) algorithm, as shown in Algorithm 1. In the next section, we will prove that the sequence generated by PIRE is bounded and any accumulation point is a stationary point of problem (1).

Convergence Analysis of PIRE

Theorem 1. *Let $D = F(\mathbf{x}^1)$, and $\mu > \frac{L(h)}{2}$, where $L(h)$ is the Lipschitz constant of $h(\mathbf{x})$. The sequence $\{\mathbf{x}^k\}$ generated in Algorithm 1 satisfies the following properties:*

(1) $F(\mathbf{x}^k)$ is monotonically decreasing. Indeed,

$$F(\mathbf{x}^k) - F(\mathbf{x}^{k+1}) \ge \left(\mu - \frac{L(h)}{2}\right)||\mathbf{x}^k - \mathbf{x}^{k+1}||^2;$$

(2) The sequence $\{\mathbf{x}^k\}$ is bounded;

(3) $\sum_{k=1}^{\infty} ||\mathbf{x}^k - \mathbf{x}^{k+1}||_F^2 \le \frac{2D}{2\mu - L(h)}$. In particular, we have $\lim_{k\to\infty}(\mathbf{x}^k - \mathbf{x}^{k+1}) = \mathbf{0}$.

Proof. Since $\mathbf{x}^{k+1}$ is the globally optimal solution to problem (18), the zero vector is contained in the subgradient with respect to $\mathbf{x}$. That is, there exists $\mathbf{v}^{k+1} \in \partial\langle \mathbf{w}^k, \mathbf{g}(\mathbf{x}^{k+1})\rangle$ such that

$$\lambda\mathbf{v}^{k+1} + \nabla h(\mathbf{x}^k) + \mu(\mathbf{x}^{k+1} - \mathbf{x}^k) = \mathbf{0}. \quad (19)$$

A dot-product with $\mathbf{x}^{k+1} - \mathbf{x}^k$ on both sides of (19) gives

$$\begin{aligned}
&\lambda\left\langle \mathbf{v}^{k+1}, \mathbf{x}^{k+1} - \mathbf{x}^k\right\rangle + \left\langle \nabla h(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x}^k\right\rangle\\
&+ \mu||\mathbf{x}^{k+1} - \mathbf{x}^k||^2 = 0.
\end{aligned} \quad (20)$$

Recalling the definition of the subgradient of the convex function, we have

$$\langle \mathbf{w}^k, \mathbf{g}(\mathbf{x}^k) - \mathbf{g}(\mathbf{x}^{k+1})\rangle \ge \left\langle \mathbf{v}^{k+1}, \mathbf{x}^k - \mathbf{x}^{k+1}\right\rangle. \quad (21)$$

Combining (20) and (21) gives

$$\begin{aligned}
&\lambda\langle \mathbf{w}^k, \mathbf{g}(\mathbf{x}^k) - \mathbf{g}(\mathbf{x}^{k+1})\rangle\\
&\ge -\left\langle \nabla h(\mathbf{x}^k), \mathbf{x}^k - \mathbf{x}^{k+1}\right\rangle + \mu||\mathbf{x}^{k+1} - \mathbf{x}^k||^2.
\end{aligned} \quad (22)$$

Since f is concave, similar to (16), we get

$$f(\mathbf{g}(\mathbf{x}^k)) - f(\mathbf{g}(\mathbf{x}^{k+1})) \ge \langle \mathbf{w}^k, \mathbf{g}(\mathbf{x}^k) - \mathbf{g}(\mathbf{x}^{k+1})\rangle. \quad (23)$$

By the condition (**C3**), we have

$$\begin{aligned}
&h(\mathbf{x}^k) - h(\mathbf{x}^{k+1})\\
&\ge \left\langle \nabla h(\mathbf{x}^k), \mathbf{x}^k - \mathbf{x}^{k+1}\right\rangle - \frac{L(h)}{2}||\mathbf{x}^{k+1} - \mathbf{x}^k||^2.
\end{aligned} \quad (24)$$

Now, combining (22)(23) and (24), we have

$$\begin{aligned}
&F(\mathbf{x}^k) - F(\mathbf{x}^{k+1})\\
&= \lambda f(\mathbf{g}(\mathbf{x}^k)) - \lambda f(\mathbf{g}(\mathbf{x}^{k+1})) + h(\mathbf{x}^k) - h(\mathbf{x}^{k+1})\\
&\ge \left(\mu - \frac{L(h)}{2}\right)||\mathbf{x}^{k+1} - \mathbf{x}^k||^2 \ge 0.
\end{aligned} \quad (25)$$

Hence $F(\mathbf{x}^k)$ is monotonically decreasing. Summing all the above inequalities for $k \ge 1$, it follows that

$$D = F(\mathbf{x}^1) \ge \left(\mu - \frac{L(h)}{2}\right)\sum_{k=1}^{\infty}||\mathbf{x}^{k+1} - \mathbf{x}^k||^2. \quad (26)$$

This implies that $\lim_{k\to\infty}(\mathbf{x}^{k+1} - \mathbf{x}^k) = \mathbf{0}$. Also $\{\mathbf{x}^k\}$ is bounded due to the condition (**C4**). ■

Theorem 2. *Let $\{\mathbf{x}^k\}$ be the sequence generated in Algorithm 1. Then any accumulation point of $\{\mathbf{x}^k\}$ is a stationary point $\mathbf{x}^*$ of problem (1). Furthermore, for every $n \ge 1$, we have*

$$\min_{1\le k\le n}||\mathbf{x}^{k+1} - \mathbf{x}^k||_2^2 \le \frac{F(\mathbf{x}^1) - F(\mathbf{x}^*)}{n\left(\mu - \frac{L(h)}{2}\right)}. \quad (27)$$

Please refer to the Supplementary Material for the proof.

We conclude this section with the following remarks:

(1) When proving the convergence of IRL1 for solving problem (8) or (12) in (Chen and Zhou ; Lu 2012), they use the Young's inequality which is a special property of the function y^p $(0 < p < 1)$

$$\sum_{i=1}^{n}(|x_i^k|+\epsilon)^p-(|x_i^{k+1}|+\epsilon)^p \geq \sum_{i=1}^{n} w_i^k\left(\left|x_i^k\right|-\left|x_i^{k+1}\right|\right), \tag{28}$$

where $w_i^k = p/(|x_i^k|+\epsilon)^{1-p}$. Eqn (28) is a special case of (23). But (23) is obtained by using the concavity of $f(\mathbf{y})$, which is much more general.

(2) In Eqn (27), $||\mathbf{x}^{k+1}-\mathbf{x}^k||_2$ is used to measure the convergence rate of the algorithm. The reason is that $||\mathbf{x}^{k+1}-\mathbf{x}^k||_2 \to 0$ is a necessary optimality condition as shown in the Theorem 1.

(3) PIRE requires that $\mu > L(h)/2$. But sometimes the Lipschitz constant $L(h)$ is not known, or it is not computable for large scale problems. One may use the backtracking rule to estimate μ in each iteration (Beck and Teboulle 2009). PIRE with multiple splitting shown in the next section also eases this problem.

PIRE with Multiple Splitting

In this section, we will show that PIRE can also solve multi-variable problem as follows

$$\min_{\mathbf{x}_1,\cdots,\mathbf{x}_S} F(\mathbf{x}) = \lambda\sum_{s=1}^{S} f_s(\mathbf{g}_s(\mathbf{x}_s)) + h(\mathbf{x}_1,\cdots,\mathbf{x}_S), \tag{29}$$

where f_s and $\mathbf{g}_s$ holds the same assumptions as f and $\mathbf{g}$ in problem (1), respectively. Problem (29) is similar to problem (1), but splits $\mathbf{x}$ into $\mathbf{x} = [\mathbf{x}_1;\cdots;\mathbf{x}_S] \in \mathbb{R}^n$, where $\mathbf{x}_s \in \mathbb{R}^{n_s}$, and $\sum_{i=1}^{S} n_s = n$.

Based on different assumptions of $h(\mathbf{x}_1,\cdots,\mathbf{x}_S)$, we have two splitting versions of the PIRE algorithm. They use different updating orders of the variables.

PIRE with Parallel Splitting

If we still assume that (**C3**) holds, i.e. $h(\mathbf{x}_1,\cdots,\mathbf{x}_S)$ has a Lipschitz continuous gradient, with Lipschitz constant $L(h)$, PIRE is naturally parallelizable. In each iteration, we parallelly update $\mathbf{x}_s^{k+1}$ by

$$\begin{aligned}\mathbf{x}_s^{k+1} = &\arg\min_{\mathbf{x_s}} \lambda\langle\mathbf{w}_s^k, \mathbf{g}_s(\mathbf{x}_s)\rangle \\ &+\frac{\mu}{2}\left\|\mathbf{x}_s-\left(\mathbf{x}_s^k-\frac{1}{\mu}\nabla_s h\left(\mathbf{x}_1^k,\cdots,\mathbf{x}_S^k\right)\right)\right\|^2,\end{aligned} \tag{30}$$

where the notion $\nabla_s h\left(\mathbf{x}_1,\cdots,\mathbf{x}_S\right)$ denotes the gradient w.r.t $\mathbf{x}_s$, $\mu > L(h)/2$, and $\mathbf{w}_s^k$ is the weight vector corresponding to $\mathbf{g}(\mathbf{x}_s^k)$, which can be computed by

$$\mathbf{w}_s^k \in -\partial\left(-f_s(\mathbf{g}_s(\mathbf{x}_s^k))\right),\ s=1,\cdots,S. \tag{31}$$

When updating $\mathbf{x}_s$ in the $(k+1)$-th iteration, only the variables in the k-th iteration are used. Thus the variables $\mathbf{x}_s^{k+1}$, $s = 1,\cdots,S$, can be updated in parallel. This is known as Jacobi iteration in numerical algebra (Liu, Lin, and Su 2013). This algorithm is named as PIRE with Parallel Splitting (PIRE-PS). Actually the updating rule of PIRE-PS is the same as PIRE, but in parallel. It is easy to check that the proofs in Theorem 1 and 2 also hold for PIRE-PS.

For some special cases of $h(\mathbf{x}_1,\cdots,\mathbf{x}_S)$, we can use different μ_s, usually smaller than μ, for updating $\mathbf{x}_s^{k+1}$. This may lead to faster convergence (Zuo and Lin 2011). If $h(\mathbf{x}_1,\cdots,\mathbf{x}_S) = \frac{1}{2}\left\|\sum_{s=1}^{S}\mathbf{A}_s\mathbf{x}_s-\mathbf{b}\right\|_2^2$, we can update $\mathbf{x}_s^{k+1}$ by

$$\begin{aligned}\mathbf{x}_s^{k+1} = &\arg\min_{\mathbf{x_s}} \lambda\langle\mathbf{w}_s^k, \mathbf{g}_s(\mathbf{x}_s)\rangle \\ &+\frac{\mu_s}{2}\left\|\mathbf{x}_s-\left(\mathbf{x}_s^k-\frac{1}{\mu_s}\mathbf{A}_s^T(\mathbf{A}\mathbf{x}^k-\mathbf{b})\right)\right\|_2^2,\end{aligned} \tag{32}$$

where $\mu_s > L_s(h)/2$ and $L_s(h) = ||\mathbf{A}_s||_2^2$ is the Lipschitz constant of $\nabla_s h(\mathbf{x}_1,\cdots,\mathbf{x}_S)$. If the size of $\mathbf{A}$ is very large, $L(h) = ||\mathbf{A}||_2^2$ may not be computable. We can split it to $\mathbf{A} = [\mathbf{A}_1,\cdots,\mathbf{A}_S]$, and compute each $L_s(h) = ||\mathbf{A}_s||_2^2$ instead. Similar convergence results in Theorem 1 and 2 also hold by updating $\mathbf{x}_s^{k+1}$ in (32). For detailed proofs, please refer to the Supplementary Material. A main difference of the convergence poof is that we use the Pythagoras relation

$$||\mathbf{a}-\mathbf{c}||_2^2-||\mathbf{b}-\mathbf{c}||_2^2 = ||\mathbf{a}-\mathbf{b}||_2^2+2\langle\mathbf{a}-\mathbf{b},\mathbf{b}-\mathbf{c}\rangle, \tag{33}$$

for the squared loss $h(\mathbf{x}_1,\cdots,\mathbf{x}_S)$. This property is much tighter than the property (17) of function with Lipschitz continuous gradient.

The result that using the squared loss leads to smaller Lipschitz constants by PIRE-PS is very interesting and useful. Intuitively, it results to minimize a tighter upper bounded surrogate function. Our experiments show that this will lead to a faster convergence of the PIRE-PS algorithm.

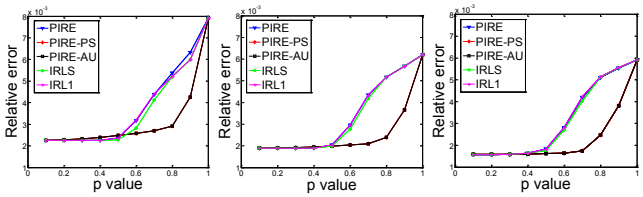
PIRE with Alternative Updating

In this section, we propose another splitting method to solve problem (29) based on the assumption that each $\nabla_s h(\mathbf{x}_1,\cdots,\mathbf{x}_S)$ is Lipschitz continuous with constant $L_s(h)$. Different from PIRE-PS, which updates each $\mathbf{x}_s^{k+1}$ based on $\mathbf{x}_s^k$, $s = 1,\cdots,S$, we instead update $\mathbf{x}_s^{k+1}$ based on all the latest $\mathbf{x}_s$. This is the known Gauss-Sidel iteration in numerical algebra. We name this method as PIRE with Alternative Updating (PIRE-AU).

Since $\nabla_s h(\mathbf{x}_1,\cdots,\mathbf{x}_S)$ is Lipschitz continuous, similar to (17), we have

$$\begin{aligned}&h(\mathbf{x}_1^{k+1},\cdots,\mathbf{x}_{s-1}^{k+1},\mathbf{x}_s,\mathbf{x}_{s+1}^k,\cdots,\mathbf{x}_S^k)\\ \leq &h(\mathbf{x}_1^{k+1},\cdots,\mathbf{x}_{s-1}^{k+1},\mathbf{x}_s^k,\cdots,\mathbf{x}_S^k)+\\ &\langle\nabla_s h(\mathbf{x}_1^{k+1},\cdots,\mathbf{x}_{s-1}^{k+1},\mathbf{x}_s^k,\cdots,\mathbf{x}_S^k),\mathbf{x}_s-\mathbf{x}_s^k\rangle\\ &+\frac{L_s(h)}{2}||\mathbf{x}_s-\mathbf{x}_s^k||_2^2.\end{aligned} \tag{34}$$

The hand right part of (34) is used as a surrogate function of $h(\mathbf{x}_1^{k+1},\cdots,\mathbf{x}_{s-1}^{k+1},\mathbf{x}_s,\mathbf{x}_{s+1}^k,\cdots,\mathbf{x}_S^k)$, which is tighter



(a) (b) (c)

Figure 1: Recovery performance comparison with a different number of measurement $\mathbf{A} \in \mathbb{R}^{m\times 1000}$: (a) $m = 200$; (b) $m = 300$; and (c) $m = 400$.

than (17) in PIRE. Then we update $\mathbf{x}_s^{k+1}$ by

$$\mathbf{x}_s^{k+1} = \arg\min_{\mathbf{x}_s} \lambda\langle \mathbf{w}_s^k, \mathbf{g}_s(\mathbf{x}_s)\rangle + \frac{\mu_s}{2}||\mathbf{x}_s - \mathbf{x}_s^k||_2^2 + \langle \nabla_s h(\mathbf{x}_1^{k+1}, \cdots, \mathbf{x}_{s-1}^{k+1}, \mathbf{x}_s^k, \cdots, \mathbf{x}_S^k), \mathbf{x}_s - \mathbf{x}_s^k\rangle, \quad (35)$$

where $\mu_s > L_s(h)/2$ and $\mathbf{w}_s^k$ is defined in (31).

The updating rule in PIRE-AU by (35) and (31) also leads to converge. Any accumulation point of $\{\mathbf{x}^k\}$ is a stationary point. See the detailed proofs in the Supplementary Material.

Both PIRE-PS and PIRE-AU can solve the multi-variable problems. The advantage of PIRE-PS is that it is naturally parallelizable, while PIRE-AU may converge with less iterations due to smaller Lipschitz constants. If the squared loss function is used, PIRE-PS use the same small Lipschitz constants as PIRE-AU.

Experiments

We present several numerical experiments to demonstrate the effectiveness of the proposed PIRE algorithm and its splitting versions. All the algorithms are implemented by Matlab, and are tested on a PC with 8 GB of RAM and Intel Core 2 Quad CPU Q9550.

ℓ_p-Minimization

We compare our proposed PIRE, PIRE-PS and PIRE-AU algorithms with IRLS and IRL1 for solving the ℓ_p-minimization problem (7). For fair comparison, we try to use the same settings of all the completed algorithms. We use the solution to the ℓ_1-minimization problem as the initialization. We find that this will accelerate the convergence of the iteratively reweighted algorithms, and also enhance the recovery performance. The choice of ϵ in (8) and (10) plays an important role for sparse signal recovery, but theoretical support has not been carried out so far. Several different decreasing rules have been tested before (Candès, Wakin, and Boyd 2008; Mohan and Fazel 2012; Lai, Xu, and Yin 2013), but none of them dominates others. Since the sparsity of sparse signal is usually unknown, we empirically set $\epsilon^{k+1} = \epsilon^k/\rho$, with $\epsilon^0 = 0.01$, and $\rho = 1.1$ (Mohan and Fazel 2012). The algorithms are stopped when $||\mathbf{x}^k - \mathbf{x}^{k+1}||_2/||\mathbf{x}^k||_2 \leq 10^{-6}$.

IRL1 requires solving (9) as inner loop. FISTA is employed to solve (9) with warm start, i.e. using $\mathbf{x}^k$ as initialization to obtain $\mathbf{x}^{k+1}$. This trick greatly reduces the inner loop iteration, which is the main cost for IRL1. For PIRE-PS and PIRE-AU algorithms, we solve problem (29) by setting $S = 20$.

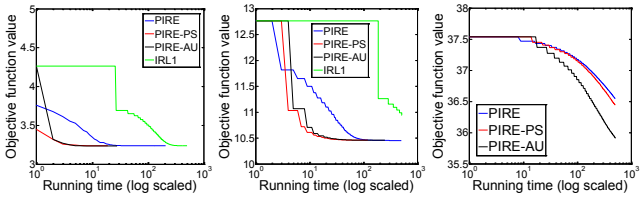
Table 1: Comparison of iteration number, running time (in seconds), objective function value and relative recovery error of different iterative reweighted methods.

Size (m, n, t)	Methods	Iter.	Time (second).	Obj. $(\times 10^{-2})$	Recovery error $(\times 10^{-3})$
(100,500,50)	PIRE	116	0.70	**5.238**	2.529
	PIRE-PS	58	**0.48**	5.239	2.632
	PIRE-AU	**56**	0.63	5.239	2.632
	IRLS	168	81.82	5.506	**2.393**
	IRL1	**56**	3.43	5.239	2.546
(200,800,100)	PIRE	119	1.48	16.923	2.246
	PIRE-PS	37	**0.82**	**16.919**	2.192
	PIRE-AU	**36**	0.88	**16.919**	2.192
	IRLS	169	474.19	17.784	**2.142**
	IRL1	81	13.53	16.924	2.248
(300,1000,200)	PIRE	151	4.63	42.840	2.118
	PIRE-PS	29	1.38	**42.815**	1.978
	PIRE-AU	**28**	**1.34**	**42.815**	**1.977**
	IRLS	171	1298.70	44.937	2.015
	IRL1	79	35.59	42.844	2.124
(500,1500,200)	PIRE	159	8.88	64.769	2.010
	PIRE-PS	26	2.27	**64.718**	**1.814**
	PIRE-AU	**25**	**2.20**	**64.718**	**1.814**
	IRLS	171	3451.79	67.996	1.923
	IRL1	89	80.89	64.772	2.013
(800,2000,200)	PIRE	140	14.99	87.616	1.894
	PIRE-PS	33	5.15	**87.533**	**1.648**
	PIRE-AU	**32**	**4.97**	**87.533**	**1.648**
	IRLS	177	7211.2	91.251	1.851
	IRL1	112	173.26	87.617	1.895

Sparse Signal Recovery The first experiment is to examine the recovery performance of sparse signals by using the proposed methods. The setup for each trial is as follows. The dictionary $\mathbf{A} \in \mathbb{R}^{m\times n}$ is a Gaussian random matrix generated by Matlab command `randn`, with the sizes $m = 200, 300, 400$, and $n = 1000$. The sparse signal $\mathbf{x}$ is randomly generated with sparsity $||\mathbf{x}||_0 = 20$. The response $\mathbf{b} = \mathbf{A}\mathbf{x} + 0.01\mathbf{e}$, where $\mathbf{e}$ is Gaussian random vector. Given $\mathbf{A}$ and $\mathbf{b}$, we can recover $\hat{\mathbf{x}}$ by solving the ℓ_p-minimization problem by different methods. The parameter is set to $\lambda = 10^{-4}$. We use the relative recovery error $||\hat{\mathbf{x}}-\mathbf{x}||_2/||\mathbf{x}||_2$ to measure the recovery performance. Based on the above settings and generated data, we find that the recovery performances are stable. We run 20 trials and report the mean relative error for comparison.

Figure 1 plots the relative recovery errors v.s. different p values ($p = 0.1, \cdots, 0.9, 1$) on three data sets with different numbers of measurements. The result for $p = 1$ is obtained by FISTA for ℓ_1-minimization. We can see that all the iteratively reweighted algorithms achieve better recovery performance with $p < 1$ than ℓ_1-minimization. Also a smaller value of p leads to better recovery performance, though the ℓ_p-minimization problem is nonconvex and a globally optimal solution is not available. In most cases, PIRE is comparative with IRLS and IRL1. A surprising result is that PIRE-PS and PIRE-AU outperform the other methods when $0.5 < p < 1$. They use a smaller Lipschitz constant than PIRE, and thus may converge faster. But none of these iteratively reweighted methods is guaranteed to be optimal.

Running Time Comparison The second experiment is to show the advantage in running time of the proposed methods. We implement all the completed methods in matrix



(a) (b) (c)

Figure 2: Running time v.s. objective function value on three synthesis data sets with size (m, n, t): (a) (1000,3000,500); (b) (1000,5000,1000); (c) (1000,10000,1000).

form for solving the following ℓ_p-minimization problem

$$\min_{\mathbf{X}\in\mathbb{R}^{n\times t}} \lambda||\mathbf{X}||_p^p + \frac{1}{2}||\mathbf{AX}-\mathbf{B}||_F^2, \tag{36}$$

where $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{B} \in \mathbb{R}^{m\times t}$, $||\mathbf{X}||_p^p = \sum_{ij}|X_{ij}|^p$, and p is set to 0.5 in this test. $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{X}$ are generated by the same procedure as the above section, and the same settings of algorithm parameters are followed. Each column of $\mathbf{X}$ is with sparsity $n \times 2\%$. We test on several different sizes of data sets, parameterized as (m, n, t). The iteration number, running time, objective function value and the relative recovery error are tabulated in Table 1. It can be seen that the proposed methods are much more efficient than IRLS and IRL1. PIRE-PS and PIRE-AU converge with less iteration and less running time. In our test, IRL1 is more efficient than IRLS. The reasons lie in: (1) initialization as a sparse solution to ℓ_1-minimization is a good choice for IRL1, but not for IRLS; (2) For each iteration in IRLS, solving t equations (11) in a loop by Matlab is not efficient; (3) IRL1 converges with less inner loop iterations due to warm start.

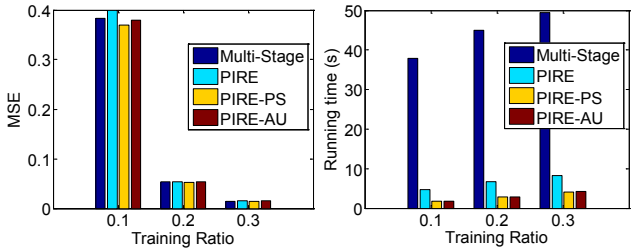
We also plot the running time v.s. objective function value on three larger data sets in Figure 2. The algorithms are stopped within 500 seconds in this test. IRLS costs much more time, and thus it is not plotted. IRL1 is not plotted for the case $n = 10,000$. It can be seen that PIRE-PS and PIRE-AU decreases the objective function value faster than PIRE.

Multi-Task Feature Learning

In this experiment, we use our methods to solve the multi-task learning problem. Assume we are given m learning tasks associated with $\{(\mathbf{X}_1, \mathbf{y}_1), \cdots, (\mathbf{X}_m, \mathbf{y}_m)\}$, where $\mathbf{X}_i \in \mathbb{R}^{n_i\times d}$ is the data matrix of the i-th task with each row a sample, $\mathbf{y}_i \in \mathbb{R}^{n_i}$ is the label of the i-th task, n_i is the number of samples for the i-th task, and d is the data dimension. Our goal is to find a matrix $\mathbf{Z} = [\mathbf{z}_1, \cdots, \mathbf{z}_m] \in \mathbb{R}^{d\times m}$ such that $\mathbf{y}_i \approx \mathbf{X}_i\mathbf{z}_i$. The capped-$\ell_1$ norm is used to regularize $\mathbf{Z}$ (Gong, Ye, and Zhang 2012a)

$$\min_{\mathbf{Z}} \lambda \sum_{j=1}^{d} \min(||\mathbf{z}^j||_1, \theta) + h(\mathbf{Z}), \tag{37}$$

where $h(\mathbf{Z}) = \sum_{i=1}^{m} ||\mathbf{X}_i\mathbf{z}_i - \mathbf{y}_i||_2^2/mn_i$ is the loss function, $\theta > 0$ is the thresholding parameter, and $\mathbf{z}^j$ is the j-th row of $\mathbf{Z}$. The above problem can be solved by our proposed PIRE, PIRE-PS and PIRE-AU algorithms, by letting $f(\mathbf{y}) = \sum_{j=1}^{d} \min(y_i, \theta)$, and $g(\mathbf{Z}) = [||\mathbf{z}^1||_1; \cdots; ||\mathbf{z}^m||_1]$.



(a) (b)

Figure 3: Comparison of (a) mean squared error (MSE) and running time on the Isolet data set for multi-task feature learning.

The Isolet (Bache and Lichman 2013) data set is used in our test. 150 subjects spoke the name of each letter of the alphabet twice. Hence, we have 52 training examples from each speaker. The speakers are grouped into 5 subsets of 30 speakers each. Thus, we have 5 tasks with each task corresponding to a subset. There are 1560, 1560, 1560, 1558, and 1559 samples of 5 tasks, respectively. The data dimension is 617, and the response is the English letter label (1-26). We randomly select the training samples from each task with different training ratios (0.1, 0.2 and 0.3) and use the rest of samples to form the test set. We compare our PIRE, PIRE-PS and PIRE-AU (we set $S = m = 5$ in PIRE-PS and PIRE-AU) with the Multi-Stage algorithm (Zhang 2008). We report the Mean Squared Error (MSE) on the test set and the running time for solving (37) on the training set. The results are averaged over 10 random splittings. As shown in Figure 3, it can be seen that all these methods achieve comparative performance, but our PIRE, PIRE-PS and PIRE-AU are much more efficient than the Multi-Stage algorithm.

Conclusions

This paper proposes the PIRE algorithm for solving the general problem (1). PIRE solves a series of problem (2), whose computational cost is usually very cheap. We further propose two splitting versions of PIRE to handle the multi-variable problems. In theory, we prove that PIRE (also its splitting versions) converges and any limit point is a stationary point. We test our methods to solve the ℓ_p-minimization problem and multi-task feature learning problem. Experimental results on both synthesis and real data sets show that our methods are with comparative learning performance, but much more efficient, by comparing with IRLS and IRL1 or multi-stage algorithms. It would be interesting to apply PIRE for structured sparsity optimization, and also the non-convex low rank regularized minimization problems (Lu et al. 2014).

Acknowledgements

This research is supported by the Singapore National Research Foundation under its International Research Centre @Singapore Funding Initiative and administered by the IDM Programme Office. Z. Lin is supported by NSF of China (Grant nos. 61272341, 61231002, and 61121002) and MSRA.

References

Bache, K., and Lichman, M. 2013. UCI machine learning repository. http://archive.ics.uci.edu/ml.

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.

Bertsekas, D. P. 1999. *Nonlinear programming*. Athena Scientific (Belmont, Mass.), 2nd edition.

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., and Zhou, W. Convergence of the reweighted ℓ_1 minimization algorithm for $\ell_2 - \ell_p$ minimization. *to appear in Comp. Optim. Appl*.

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.

Gong, P.; Zhang, C.; Lu, Z.; Huang, J.; and Ye, J. 2013. A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. *ICML*.

Gong, P.; Ye, J.; and Zhang, C. 2012a. Multi-stage multi-task feature learning. In *NIPS*.

Gong, P.; Ye, J.; and Zhang, C. 2012b. Robust multi-task feature learning. In *ACM SIGKDD*, 895–903. ACM.

Jacob, L.; Obozinski, G.; and Vert, J.-P. 2009. Group Lasso with overlap and graph Lasso. In *ICML*, 433–440. ACM.

Knight, K., and Fu, W. 2000. Asymptotics for Lasso-type estimators. *Annals of Statistics* 1356–1378.

Lai, M.-J.; Xu, Y.; and Yin, W. 2013. Improved iteratively reweighted least squares for unconstrained smoothed ℓ_q minimization. *SIAM Journal on Numerical Analysis* 51(2):927–957.

Liu, R.; Lin, Z.; and Su, Z. 2013. Linearized alternating direction method with parallel splitting and adaptive penalty for separable convex programs in machine learning. In *ACML*.

Lu, C.; Tang, J.; Lin, Z.; and Yan, S. 2014. Generalized nonconvex nonsmooth low-rank minimization. In *CVPR*.

Lu, Z. 2012. Iterative reweighted minimization methods for ℓ_p regularized unconstrained nonlinear programming. *Mathematical Programming*.

Mohan, K., and Fazel, M. 2012. Iterative reweighted algorithms for matrix rank minimization. In *JMLR*, volume 13, 3441–3473.

Wright, J.; Yang, A. Y.; Ganesh, A.; Sastry, S. S.; and Ma, Y. 2009. Robust face recognition via sparse representation. *TPAMI* 31(2):210–227.

Zhang, T. 2008. Multi-stage convex relaxation for learning with sparse regularization. In *NIPS*, 1929–1936.

Zhang, C. 2010. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*.

Zhou, Z.; Li, X.; Wright, J.; Candès, E.; and Ma, Y. 2010. Stable principal component pursuit. In *IEEE International Symposium on Information Theory Proceedings*, 1518–1522. IEEE.

Zuo, W., and Lin, Z. 2011. A generalized accelerated proximal gradient approach for total-variation-based image restoration. *TIP* 20(10):2748–2759.