Operations Research and Fuzziology 运筹与模糊学, **2023, 13(6), 7333-7345**
Published Online December 2023 in Hans. https://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2023.136721

Hans汉斯

基于Huber损失和组Lasso惩罚问题的加速临近梯度算法

沈慧玲，彭定涛*，张 弦

贵州大学数学与统计学院, 贵州 贵阳

收稿日期：2023年10月19日；录用日期：2023年12月20日；发布日期：2023年12月28日

摘 要

在高维线性回归模型中，组稀疏恢复的做法是在原有线性模型的基础上增加一个组Lasso惩罚项，以诱导尽可能少的非零组，将问题转换为凸优化模型进行求解。本文研究Huber损失和组Lasoo组合问题，其中惩罚项包含了组稀疏惩罚，惩罚项的目的是用于保证组元素稀疏性结构。首先，由于惩罚项是一个凸但不光滑函数，为了刻画组Lasso模型的最优性条件，给出了其经典次微分。其次，利用Nesterov加速技术提出了加速临近梯度算法来求解我们的模型。最后证明了所提出算法的收敛性。

关键词

Huber损失，组Lasso，Nesterov加速，加速临近梯度算法

Accelerated Proximal Gradient Algorithm Based on Huber Loss and Group Lasso Penalty Problem

Huiling Shen, Dingtao Peng*, Xian Zhang

School of Mathematics and Statistics, Guizhou University, Guiyang Guizhou

* 通讯作者。

文章引用: 沈慧玲, 彭定涛, 张弦. 基于Huber损失和组Lasso惩罚问题的加速临近梯度算法[J]. 运筹与模糊学, 2023, 13(6): 7333-7345. DOI: 10.12677/orf.2023.136721

Received: Oct. 19th, 2023; accepted: Dec. 20th, 2023; published: Dec. 28th, 2023

Abstract

In the high-dimensional linear regression model, the method of group sparse recovery is to add a group Lasso penalty term to the original linear model so as to induce as few non-zero groups as possible, and then to transform the problem into a convex optimization model. This paper studies the group Lasso problem based on Huber loss, where the penalty term includes group sparsity penalty, and the purpose of the penalty term is to ensure the group sparsity structure of group element. Firstly, since the penalty term is a convex but not smooth function, in order to characterize the optimality conditions of the group Lasso model, its classical subdifferential is given. Secondly, an accelerated proximity gradient algorithm was proposed using Nesterov acceleration technology to solve our model. Finally, the convergence of the proposed algorithm was demonstrated.

Keywords

Huber Loss, Group Lasso, Nesterov Acceleration, Accelerated Proximal Gradient Algorithm

1. 引言

在多元线性回归模型中, 对于给定的预测变量$w_i \in \mathbb{R}^n$及响应变量$b_i \in \mathbb{R}$, 有:

$$b_i = w_i^\top x + \varepsilon_i.$$

其中ε_i是数据的随机误差且相互独立, $x \in \mathbb{R}^n$是需要估计的参数. 对该模型n次取样可得:

$$b_i = w_i^\top x + \varepsilon_i, i = 1, 2, \cdots, m, \tag{1}$$

将观测数据中的预测变量写成一个$m \times n$维的矩阵A, 将响应变量b_i和随机误差ε_i写成向量的形式, 则得到模型(1)的矩阵形式:

$$b = Ax + \varepsilon. \tag{2}$$

其中$A = (w_1^\top; w_2^\top; \cdots; w_m^\top) \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $\varepsilon \in \mathbb{R}^m$. 解模型(2)常用的方法是通过极小化残差平方和来估计回归系数, 即

$$\min_{x\in\mathbb{R}^n} \mathcal{L}(x) := \|Ax - b\|_2^2, \tag{3}$$

模型(3)的最小二乘估计: $x = (A^\top A)^{-1}(A^\top b)$. 然而当高维数据$m << n$时, $A^\top A$不可逆, 从而最优解很可能不止一个, 即无法得到唯一的最小估计. 为避免数据的多重共线性或可能出现的欠定现象, 许多研究者引入了如下ℓ_0正则稀疏优化模型:

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_0, \tag{4}$$

其中$\|x\|_0$表示向量x非零分量的个数, $\lambda > 0$是正则化参数, 用来平衡解的精确性和稀疏性.

因为ℓ_0是非凸、非光滑且不连续的正则, 因此求解该模型是NP难的 [1]. 斯坦福大学统计学教授Robert Tibshirani [2]于1996年首次利用ℓ_l 正则代替ℓ_0, 从而将问题转化为如下Lasso模型:

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_1, \tag{5}$$

其中$\|x\|_1 := \sum_{i=1}^{n} |x_i|$. Lasso模型是一种常用的凸罚模型, 基于互相关(Mutual Coherence, MC)性质和限制等距性质(Restricted Isometry Property, RIP), 许多研究者分析了模型(4)和(5)解的等价性和误差估计问题 [3]. Lasso问题是通过惩罚参数的ℓ_l范数来寻求欠定线性方程组的稀疏解. 然而在高维情形下, 通常自变量的大部分分组元素为零, 即决策自变量的分量之间往往具有一定的组结构, 比如基因表达数据可以按照生物学路径或者因子水平将指标进行分组. 而传统的Lasso模型, 仅考虑单个分量的稀疏性, 并不具备应对分组等复杂结构的能力, 因此很多数据本身重要的结构关系就会被忽视, 这会导致很多实际问题的求解失败. 基于未知变量具有组稀疏的先验信息, 解决上述问题通常采用具有组稀疏结构的正则化方法, 一般模型如下:

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_{2,0}. \tag{6}$$

$\|x\|_{2,0} := \sharp\{i : \|x_i\| \neq 0\}$, 它表示每个组的$\ell_2$范数非零的组的个数. 最近Jiao 和Jin [4]证明了$\ell_{2,0}$正则化问题的全局解的存在性, 并设计了有效的组原始对偶积极集算法来求解该问题. $\ell_{2,0}$ 是基于稀疏优化的ℓ_0正则, 它也是非凸、非光滑甚至是不连续的, 因此求解该问题依然具有很大的挑战. Yuan和Lin [5]提出了组Lasso 优化模型来松弛$\ell_{2,0}$ 模型:

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_{2,1}.$$

其中, $\|x\|_{2,1} = \sum_{l\in\Gamma} \|x_{G_l}\|_2$, $\Gamma = \{1, \cdots, L\}$， $x_{G_l} \in \mathbb{R}^{|G_l|}$, $\sum_{l\in\Gamma} |G_l| = n$, 且$|G_l|$表示为指标集$G_l$的

基数. 组Lasso是Lasso方法在组稀疏问题下的一种自然扩展, 因为组Lasso是凸的正则化函数, 因此它的局部极小点也是全局极小点. 相较与Lasso, 组Lasso具有许多令人满意的性质. 例如, 在恰当的组RIP条件下 [6], 它能够更好的恢复组稀疏信号, 且需要的样本量更小、受噪声影响更小、重构性能俱佳等优势. 组Lasso问题在机器学习、文本处理、生物信息学、信号解释、基因选择、变量选择等领域有着广泛的应用 [7–13], 研究者们已经发展了多种有效算法, 例如: 邻近梯度算法 [5,14], 块坐标算法 [15–17], 不动点迭代法 [18], 增广拉格朗日半光滑牛顿算法 [13] 等. 此外, 稀疏组Lasso估计量的统计性质也得到了充分的研究, 例如: Tony Cai等人 [7]研究了最优样本复杂度和误差估计的收敛速度，为稀疏组Lasso问题的样本复杂度与误差估计提供了理论保证; Candès等人 [19]研究了基于稀疏组Lasso估计量的多任务学习与分类问题; Benjamin [11]研究了自适应稀疏组Lasso估计量的渐近性质, 并证明了该估计量满足凸损失函数的oracle性质.

由于模型(6)的损失函数为最小二乘, 而最小二乘函数没有鲁棒性, 对离群点的容忍度不高 [20, 21], 基于此, 我们需要考虑抗异常损失函数的问题, 如最小一乘函数、Huber函数等. 而Huber函数结合了最小一乘和最小二乘的优良性质, 不仅具有鲁棒性, 而且是一个光滑函数. 因此, 在组合优化问题中, 将Huber函数作为损失函数具有很大的优势.

基于上述分析, 本文考虑如下基于Huber损失和组Lasso(GLasso)组合优化模型：

$$\min_{x\in\mathbb{R}^n}\sum_{i=1}^{m}H(A_i^Tx-b_i)+\lambda\|x\|_{2,1}, \tag{7}$$

其中

$$H(t)=\begin{cases}\frac{1}{2}t^2, & |t|\le\delta,\\ \delta\left(|t|-\frac{1}{2}\delta\right), & \text{其他},\end{cases}$$

是Huber函数, $A\in\mathbb{R}^{m\times n}$, $A_i^T\in\mathbb{R}^n$是A 的第i行, $b_i\in\mathbb{R}$, $i=1,\cdots,m$,

本文主要工作如下:

(i) 给出了问题(7)的最优性条件.

(ii) 计算了$\ell_{2,1}$范数的临近算子.

(iii) 基于问题(7)正则项的临近算子的结构, 提出了加速临近梯度算法并证明该算法的收敛性.

本文结构如下: 在第二节中, 我们介绍了凸函数的次微分并给出问题(7)的最优性条件; 第三节提出加速临近梯度算法且给出算法的收敛性; 第四节给出简要的结论.

符号标记: sign$(\cdot)$为示性函数, 即sign$(t)=1$, 若$t>0$; sign$(t)=0$, 若$t=0$; sign$(t)=-1$, 若$t<0$.

2. 最优性条件

本节给出后文所需的相关定义及最优性条件, 包括次微分和临近算子的定义.

2.1. 问题(7)的最优性条件

首先, 我们介绍问题(7)的最优性条件. 为了简便, 记$f(x)=\sum_{i=1}^{m}H(A_i^Tx-b_i)$, $g(x):=\lambda\|x\|_{2,1}$, 则问题(7)等价表示为如下问题:

$$\min_{x\in\mathbb{R}^n}\mathcal{F}(x):=f(x)+g(x). \tag{8}$$

定义 2.1 [22] 设$\mathcal{F}$是$\mathbb{R}$上的正常凸函数, $x\in\mathbb{R}^n$, 若存在$\xi\in\mathbb{R}^n$满足

$$\mathcal{F}(y)\geq\mathcal{F}(x)+\xi^\top(y-x),\forall y\in\mathbb{R}^n\},$$

则称ξ为$\mathcal{F}$在x处的次梯度. 函数$\mathcal{F}$在x处所有次梯度的集合称为$\mathcal{F}$在x处的次微分, 记为$\partial\mathcal{F}(x)$, 即

$$\partial\mathcal{F}(x)=\{\xi\in\mathbb{R}^n|\mathcal{F}(y)\geq\mathcal{F}(x)+\xi^\top(y-x),\forall y\in\mathbb{R}^n\}.$$

记$m(z):=\|z\|_2=(\sum_{j=1}^{|G_l|}z_j^2)^{\frac{1}{2}},\forall\, z\in\mathbb{R}^{|G_l|}$, $l\in\{1,\cdots,L\}$. 则当$i=1,\ldots,n$, $l=1,\ldots,L$时, $\|x\|_2$在$\widehat{x}$处的次微分表示为:

$$\partial m(\widehat{x}_{G_l})=\begin{cases}\{\frac{\widehat{x}_{G_l}}{\|\widehat{x}_{G_l}\|_2}\}, & \widehat{x}_{G_l}\neq 0,\\ \{v_{G_l}:\|v_{G_l}\|_2\leq 1\}, & \widehat{x}_{G_l}=0.\end{cases} \tag{9}$$

下面我们用次微分来刻画问题(7)的最优性条件. 此外, 问题(7)的最优性条件还可以使用方向稳定点刻画 [23].

定理 2.1 设$\widehat{x}$是问题(7)的最优值点, 则有

$$0\in\nabla f(\widehat{x})+\lambda\partial(\sum_{l=1}^{L}\|\widehat{x}_{G_l}\|_2),$$

其中, $\partial(\sum_{l=1}^{L}\|\widehat{x}_{G_l}\|_2)=\partial m(\widehat{x}_{G_1})\times\cdots\times\partial m(\widehat{x}_{G_L})$, $l\in\{1,\cdots,L\}$,

$$\nabla f(\widehat{x})=\begin{cases}\sum_{i=1}^{m}A_i(A_i^\top\widehat{x}-b_i), & |A_i^\top\widehat{x}-b_i|\leq\delta\\ \sum_{i=1}^{m}\delta A_i\,\mathrm{sign}(A_i^\top\widehat{x}-b_i), & |A_i^\top\widehat{x}-b_i|>\delta\end{cases}.$$

2.2. $g(\cdot)$的临近算子

定义 2.2 [18, 24] 令g是$\mathbb{R}^n$上的正常凸函数, 对$\forall\sigma>0$, $\sigma g(x)$的临近算子定义为:

$$\mathrm{prox}_{\sigma g}(x):=\arg\min_{y\in\mathbb{R}^n}\{g(y)+\frac{1}{2\sigma}\|y-x\|_2^2\},\ \forall x\in\mathbb{R}^n.$$

由于g是正常闭凸函数, 则对任意的$x\in\mathbb{R}^n$, $\mathrm{prox}_{\sigma g}(x)$的值存在且唯一, 说明临近算子是良定义的. 根据定义2.2, 不难计算出ℓ_2范数的邻近算子:

$$\mathrm{prox}_{t\lambda\|\cdot\|_2}(x)=\begin{cases}x-\frac{t\lambda x}{\|x\|_2}, & \|x\|_2\geq t\lambda,\\ 0, & \text{其他.}\end{cases}\tag{10}$$

其中, $t>0$, $\lambda>0$. 由于变量$x\in\mathbb{R}^n$可以分成L个组, 即$x=(x_{G_1},\cdots,x_{G_L})$, 且$g(x)=\lambda\|x\|_{2,1}=\lambda\sum\limits_{l=1}^{L}\|x_{G_l}\|_2$. 由邻近算子的定义和$g(x)$的组可分性, $tg(x)$的邻近算子如下所示:

$$\mathrm{prox}_{tg}(x)=(\mathrm{prox}_{tg}(x))_{G_1}\times\cdots\times(\mathrm{prox}_{tg}(x))_{G_L}\tag{11}$$

其中, $(\mathrm{prox}_{tg}(x))_{G_l}=\mathrm{prox}_{t\lambda\|\cdot\|_2}(x_{G_l})$ $(l=1,\cdots,L)$.

接下来的引理给出了g的邻近算子和次微分之间的关系.

引理 2.1 [25] 令g是$\mathbb{R}^n$上的正常凸函数, $y\in\mathbb{R}^n$, 则

$$x\in\partial g(y)\quad\text{当且仅当}\quad y=\mathrm{prox}_g(x+y).$$

接下来的定理给出Huber函数的梯度Lipschitz常数, 该定理保证了加速临近梯度算法的收敛结果.

定理 2.2 *Huber函数$f(x)=\sum_{i=1}^{m}H(A_i^Tx-b_i)$是连续可微凸函数, 且其梯度是Lipschitz 连续的, Lipschitz连续常数为$L=\|A\|_F^2$.*

证明. 对$\forall x,y\in\mathbb{R}^n$. (i) 当$|A_i^\top x-b_i|\leq\delta,|A_i^\top y-b_i|\leq\delta$时, 有

$$\begin{aligned}\|\nabla f(x)-\nabla f(y)\|_2&=\|\sum_{i=1}^{m}(A_i(A_i^\top x-b_i)-A_i(A_i^\top y-b_i))\|_2\\&\leq\sum_{i=1}^{m}\|A_iA_i^\top\|_2\cdot\|x-y\|_2=\sum_{i=1}^{m}\|A_i\|_2^2\cdot\|x-y\|_2\\&=\|A\|_F^2\cdot\|x-y\|_2.\end{aligned}$$

(ii) 当$|A_i^\top x-b_i|\leq\delta,|A_i^\top y-b_i|>\delta$时, 有

$$\begin{aligned}\|\nabla f(x)-\nabla f(y)\|_2&=\|\sum_{i=1}^{m}A_i(A_i^\top x-b_i)-\sum_{i=1}^{m}\delta A_i\,\mathrm{sign}(A_i^\top y-b_i)\|_2\\&\leq\sum_{i=1}^{m}\|A_i(A_i^\top x-b_i-\delta\,\mathrm{sign}(A_i^\top y-b_i))\|_2\\&=\sum_{i\in\{i|A_i^\top y-b_i<-\delta\}}\|A_i(A_i^\top x-b_i+\delta)\|_2+\sum_{i\in\{i|A_i^\top y-b_i>\delta\}}\|A_i(A_i^\top x-b_i-\delta)\|_2\\&\leq\sum_{i\in\{i|A_i^\top y-b_i<-\delta\}}\|A_i(A_i^\top x-A_i^\top y)\|_2+\sum_{i\in\{i|A_i^\top y-b_i>\delta\}}\|A_i(A_i^\top x-A_i^\top y)\|_2\\&\leq\sum_{i=1}^{m}\|A_iA_i^\top\|_2\cdot\|x-y\|_2=\|A\|_F^2\cdot\|x-y\|_2.\end{aligned}$$

(iii) 当$|A_i^\top x - b_i| > \delta, |A_i^\top y - b_i| \le \delta$ 时, 有

$$\begin{aligned}\|\nabla f(x) - \nabla f(y)\|_2 &= \|\sum_{i=1}^m \delta A_i \operatorname{sign}(A_i^\top x - b_i) - \sum_{i=1}^m A_i(A_i^\top y - b_i)\|_2 \\ &= \sum_{i\in\{i|A_i^\top x - b_i < -\delta\}} \|A_i(-\delta + b_i - A_i^\top y)\|_2 + \sum_{i\in\{i|A_i^\top x - b_i > \delta\}} \|A_i(\delta + b_i - A_i^\top y)\|_2 \\ &\le \sum_{i\in\{i|A_i^\top x - b_i < -\delta\}} \|A_i(A_i^\top x - A_i^\top y)\|_2 + \sum_{i\in\{i|A_i^\top x - b_i > \delta\}} \|A_i(A_i^\top x - A_i^\top y)\|_2 \\ &\le \sum_{i=1}^m \|A_i A_i^\top\|_2 \cdot \|x - y\|_2 = \|A\|_F^2 \cdot \|x - y\|_2.\end{aligned}$$

(iv) 当$|A_i^\top x - b_i| > \delta, |A_i^\top y - b_i| > \delta$时, 有

$$\begin{aligned}\|\nabla f(x) - \nabla f(y)\|_2 &= \|\sum_{i=1}^m \delta A_i \operatorname{sign}(A_i^\top x - b_i) - \sum_{i=1}^m \delta A_i \operatorname{sign}(A_i^\top y - b_i)\|_2 \\ &= \delta\|\sum_{i=1}^m (A_i(\operatorname{sign}(A_i^\top x - b_i) - \operatorname{sign}(A_i^\top y - b_i)))\|_2 \\ &\le \delta \sum_{i=1}^m \|A_i\|_2 \frac{\|(\operatorname{sign}(A_i^\top x - b_i) - \operatorname{sign}(A_i^\top y - b_i))\|_2}{\|A_i^\top (x - y)\|_2} \cdot \|A_i^\top (x - y)\|_2 \\ &\le \delta \sum_{i=1}^m \|A_i\|_2^2 \frac{\|(\operatorname{sign}(A_i^\top x - b_i) - \operatorname{sign}(A_i^\top y - b_i))\|_2}{\|A_i^\top (x - y)\|_2} \cdot \|x - y\|_2 \\ &\le \delta \sum_{i=1}^m \|A_i\|_2^2 \frac{\|(\operatorname{sign}(A_i^\top x - b_i) - \operatorname{sign}(A_i^\top y - b_i))\|_2}{\|(A_i^\top x - b_i) - (A_i^\top y - b_i)\|_2} \cdot \|x - y\|_2 \\ &\le \sum_{i=1}^m \|A_i\|_2^2 \cdot \|x - y\|_2 = \|A\|_F^2 \cdot \|x - y\|_2.\end{aligned}$$

综上(i)-(iv), 对$\forall x, y \in \mathbb{R}^n$, 可得

$$\|\nabla f(x) - \nabla f(y)\|_2 \le \|A\|_F^2 \cdot \|x - y\|_2. \qquad \square$$

3. 加速邻近梯度算法

本节考虑优化问题(8)

$$\min_{x\in\mathbb{R}^n} \mathcal{F}(x) := f(x) + g(x)$$

的有效算法. 由定理2.2可知, $f(x)$是连续可微凸函数且梯度是Lipschitz连续的, 其梯度Lipschitz常数为$L = \|A\|_F^2 > 0$. 优化问题(8)由光滑部分$f(x)$和非光滑部分$g(x)$组成, 在(10)(11)中已经给出$g(x)$的临近算子, 因此可用临近梯度法来求解问题(8). 然而临近梯度法是一阶算法,

Beck和Teboulle [26]指出: 当步长为固定步长并小于或等于$\frac{1}{L}$时, 其收敛速度只有$O(\frac{1}{k})$, 而加速临近梯度算法在仅使用梯度信息的情况下,收敛速度能取到$O(\frac{1}{k^2})$. 因此, 本文采用如下加速临近梯度算法来求解问题(8).

首先对问题(8)引入如下辅助函数

$$Q_t(x,z)=f(z)+\langle x-z,\nabla f(z)\rangle+\frac{1}{2t}\|x-z\|_2^2+g(x), \tag{12}$$

其中$t>0$是一个常数. 如果第k步迭代已获得迭代点y^k, 我们算法的x^{k+1}由下式产生,

$$\begin{aligned}x^{k+1}&=\arg\min_{x\in\mathbb{R}^n}Q_\gamma(x,y^k)\\&=\arg\min_{x\in\mathbb{R}^n}\left\{f(y^k)+\langle x-y^k,\nabla f(y^k)\rangle+\frac{1}{2t_k}\|x-y^k\|_2^2+g(x)\right\}\\&=\arg\min_{x\in\mathbb{R}^n}\left\{g(x)+\frac{1}{2t_k}\|x-\left(y^k-t_k\nabla f(y^k)\right)\|_2^2\right\}\\&=\mathrm{prox}_{t_kg}\left(y^k-t_k\nabla f(y^k)\right).\end{aligned}$$

据此, 下面给出求解问题(8)的加速临近梯度算法的框架.

算法3.1 加速临近梯度算法

初始步: 给定$\lambda>0$, $\gamma_1=1$, $t_0>0$, $x^{-1}=x^0\in\mathbb{R}^n$, $\upsilon^0=x^0$, $\delta>0$, $\rho>1$, $k=1$.
步1: 令$\gamma_k=\frac{2}{1+\sqrt{1+\frac{4}{\gamma_{k-1}}}}$, $y^k=(1-\gamma_k)x^{k-1}+\gamma_k\upsilon^k$.
步2:

I. 计算z^k: $z^k=y^k-t_k\nabla f(y^k)$.

II. 将z^k, y^k按给定的方式分成L个组.

III. 对$l=1,\cdots,L$, 计算

$$x^k=\mathrm{prox}_{tg}(z^k)=(\mathrm{prox}_{tg}(z^k))_{G_1}\times\cdots\times(\mathrm{prox}_{tg}(z^k))_{G_L} \tag{13}$$

其中$(\mathrm{prox}_{tg}(z^k))_{G_l}$由(10)式定义.

步3: 令$\upsilon^k=x^{k-1}+\frac{1}{\gamma_k}(x^k-x^{k-1})$.
步4: 令$k:=k+1$, 转入**步1**.
输出: x^k.

在算法3.1中,

$$\nabla f(y^k)=\begin{cases}\sum_{i=1}^m A_i(A_i^\top y^k-b_i), & |A_i^\top y^k-b_i|\le\delta\\ \sum_{i=1}^m \delta A_i\,\mathrm{sign}(A_i^\top y^k-b_i), & |A_i^\top y^k-b_i|>\delta\end{cases}.$$

且γ_k满足以下条件:

$$\gamma_1=1,\ \frac{(1-\gamma_k)t_k}{\gamma_k^2}\le\frac{t_{k-1}}{\gamma_{k-1}^2},\ k>1, \tag{14}$$

$$\frac{\gamma_k^2}{t_k}=O(\frac{1}{k^2}). \tag{15}$$

定理 3.1 *设$\{x^k\}$是算法3.1产生的序列, 取步长$t_k=\frac{1}{L}$, x^*为极小解, 则*

$$\mathcal{F}(x^k)-\mathcal{F}(x^*)\leqslant\frac{2L}{(k+1)^2}\|x^0-x^*\|_2^2. \tag{16}$$

证明. 由引理2.1可知, $-x^k+y^k-t_k\nabla f(y^k)\in t_k\partial g(x^k)$. 因为惩罚函数$g$ 为凸函数, 则对$\forall x\in\mathbb{R}^n$, 有

$$t_kg(x)\geqslant t_kg(x^k)+\langle-x^k+y^k-t_k\nabla f(y^k),x-x^k\rangle, \tag{17}$$

又因为可微函数f是凸的且梯度$L-$Lipschitz连续的, 则由f的二次上界性质和$t_k=\frac{1}{L}$ 可知

$$f(x^k)\leqslant f(y^k)+\langle\nabla f(y^k),x^k-y^k\rangle+\frac{1}{2t_k}\|x^k-y^k\|_2^2. \tag{18}$$

结合(17)和(18), 可得

$$\begin{aligned}\mathcal{F}(x^k)=&f(x^k)+g(x^k)\\ \leqslant&g(x)+f(y^k)+\langle\nabla f(y^k),x-y^k\rangle+\frac{1}{t_k}\langle x^k-y^k,x-x^k\rangle+\frac{1}{2t_k}\|x^k-y^k\|_2^2\\ \leqslant&g(x)+f(x)+\frac{1}{t_k}\langle x^k-y^k,x-x^k\rangle+\frac{1}{2t_k}\|x^k-y^k\|_2^2\\ =&\mathcal{F}(x)+\frac{1}{t_k}\langle x^k-y^k,x-x^k\rangle+\frac{1}{2t_k}\|x^k-y^k\|_2^2.\end{aligned}$$

在上式中分别取$x=x^{k-1}$, $x=x^*$, 再分别乘$1-\gamma_k$和γ_k并相加得到:

$$\begin{aligned}&\mathcal{F}(x^k)-\mathcal{F}(x^*)-(1-\gamma_k)(\mathcal{F}(x^{k-1})-\mathcal{F}(x^*))\\ \leqslant&\frac{1}{t_k}\langle x^k-y^k,(1-\gamma_k)x^{k-1}+\gamma_kx^*-x^k\rangle+\frac{1}{2t_k}\|x^k-y^k\|_2^2.\end{aligned} \tag{19}$$

结合$y^k=(1-\gamma_k)x^{k-1}+\gamma_k v^{k-1}$和$v^k=x^{k-1}+\frac{1}{\gamma_k}(x^k-x^{k-1})$, 则(19)可表示为

$$\begin{aligned}&\mathcal{F}(x^k)-\mathcal{F}(x^*)-(1-\gamma_k)(\mathcal{F}(x^{k-1})-\mathcal{F}(x^*))\\ \leqslant&\frac{1}{2t_k}(\|y^k-(1-\gamma_k)x^{k-1}-\gamma_k x^*\|_2^2-\|x^k-(1-\gamma_k)x^{k-1}-\gamma_k x^*\|_2^2)\\ =&\frac{\gamma_k^2}{2t_k}(\|v^{k-1}-x^*\|_2^2-\|v^k-x^*\|_2^2).\end{aligned}\tag{20}$$

由(20)可得

$$\frac{t_k}{\gamma_k^2}(\mathcal{F}(x^k)-\mathcal{F}(x^*))+\frac{1}{2}\|v^k-x^*\|_2^2\leq\frac{t_k}{\gamma_k^2}(1-\gamma_k)(\mathcal{F}(x^{k-1})-\mathcal{F}(x^*))+\frac{1}{2}\|v^{k-1}-x^*\|_2^2\tag{21}$$

结合(14)和(21), 得

$$\frac{t_k}{\gamma_k^2}(\mathcal{F}(x^k)-\mathcal{F}(x^*))+\frac{1}{2}\|v^k-x^*\|_2^2\leqslant\frac{t_{k-1}}{\gamma_{k-1}^2}(\mathcal{F}(x^{k-1})-\mathcal{F}(x^*))+\frac{1}{2}\|v^{k-1}-x^*\|_2^2.\tag{22}$$

反复使用(22)可得

$$\frac{t_k}{\gamma_k^2}(\mathcal{F}(x^k)-\mathcal{F}(x^*))+\frac{1}{2}\|v^k-x^*\|_2^2\leqslant\frac{t_1}{\gamma_1^2}(\mathcal{F}(x^1)-\mathcal{F}(x^*))+\frac{1}{2}\|v^1-x^*\|_2^2.\tag{23}$$

由于$\gamma_1=1$, $v^0=x^0$, 根据不等式(20), 则不等式(23)的右边可转化为

$$\begin{aligned}&\frac{t_1}{\gamma_1^2}(\mathcal{F}(x^1)-\mathcal{F}(x^*))+\frac{1}{2}\|v^1-x^*\|_2^2\\ \leqslant&\frac{(1-\gamma_1)t_1}{\gamma_1^2}(\mathcal{F}(x^0)-\mathcal{F}(x^*))+\frac{1}{2}\|v^0-x^*\|_2^2=\frac{1}{2}\|x^0-x^*\|_2^2.\end{aligned}\tag{24}$$

□

故结合(14), (23) 和(24)可得

$$\mathcal{F}(x^k)-\mathcal{F}(x^*)\leqslant\frac{2L}{(k+1)^2}\|x^0-x^*\|_2^2.$$

4. 总结

本文研究了基于Huber损失函数和组Lasso的组合优化问题, 首先我们使用经典次微分给出问题的最优性条件. 其次, 针对该组合优化问题, 设计了加速临近梯度算法并分析了算法的收敛性, 同时该算法的收敛速度达到$O(\frac{1}{k^2})$. 本文为求解组Lasso优化问题提供了理论和方法基础. 将进一步通过数值实验验证算法的效果.

基金项目

国家自然科学基金项目(12261020)、贵州省科技计划项目(ZK[2021]009, [2018]5781)、贵州省

高层次留学人才创新创业择优资助重点项目([2018]03)和贵州省青年科技人才成长项目([2018]121).

参考文献

[1] Natarajan, B. (1995) Sparse Approximate Solutions to Linear Systems. *SIAM Journal on Computing*, **24**, 227-234. https://doi.org/10.1137/S0097539792240406

[2] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society: Series B*, **58**, 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[3] Jiao, Y., Jin, B. and Lu, X. (2015) A Primal Dual Active Set with Continuation Algorithm for the ℓ_0 Regularized Optimization Problem. *Applied and Computational Harmonic Analysis*, **39**, 400-426. https://doi.org/10.1016/j.acha.2014.10.001

[4] Jiao, Y., Jin, B. and Lu, X. (2016) Group Sparse Recovery via the $\ell_0 - \ell_2$ Penalty: Theory and Algorithm. *IEEE Transactions on Signal Processing*, **65**, 998-1012. https://doi.org/10.1109/TSP.2016.2630028

[5] Yuan, M. and Lin, Y. (2006) Model Selection and Estimation in Regression with Grouped Variables. *Journal of the Royal Statistical Society: Series B*, **68**, 49-67. https://doi.org/10.1111/j.1467-9868.2005.00532.x

[6] Eren Ahsen, M. and Vidyasagar, M. (2017) Error Bounds for Compressed Sensing Algorithms with Group Sparsity: A Unified Approach. *Applied and Computational Harmonic Analysis*, **43**, 212-232. https://doi.org/10.1016/j.acha.2015.11.006

[7] Cai, T.T., Zhang, A.R. and Zhou, Y.C. (2022) Sparse Group Lasso: Optimal Sample Complexity, Convergence Rate, and Statistical Inference. *IEEE Transactions on Information Theory*, **68**, 5975-6002. https://doi.org/10.1109/TIT.2022.3175455

[8] Chatterjee, S. and Steinhaeuser, K. (2012) Sparse Group Lasso: Consistency and Climate Applications. In: Ghosh, J., Liu, H., Davidson, I., Domeniconi, C. and Kamath, C., Eds., *Proceedings of the SIAM International Conference on Data Mining*, SIAM, 47-58. https://doi.org/10.1137/1.9781611972825.5

[9] Li, Y.M., Nan, B. and Zhu, J. (2015) Multivariate Sparse Group Lasso for the Multivariate Multiple Linear Regression with an Arbitrary Group Structure. *Biometrics*, **71**, 354-363. https://doi.org/10.1111/biom.12292

[10] Li, X., Sun, D.F. and Toh, K.C. (2018) On Efficiently Solving the Subproblems of a Level-set Method for Fused Lasso Problems. *SIAM Journal on Optimization*, **28**, 1842-1862. https://doi.org/10.1137/17M1136390

[11] Poignard, B. (2020) Asymptotic Theory of the Adaptive Sparse Group Lasso. *Annals of the Institute of Statistical Mathematics*, **72**, 297-328. https://doi.org/10.1007/s10463-018-0692-7

[12] Simon, N., Friedman, J. and Hastie, T. (2013) A Sparse-Group Lasso. *Journal of Computational and Graphical Statistics*, **22**, 231-245. https://doi.org/10.1080/10618600.2012.681250

[13] Zhang, Y.J., Zhang, N. and Sun, D.F. (2020) An Efficient Hessian Based Algorithm for Solving Large-Scale Sparse Group Lasso Problems. *Mathematical Programming*, **17**, 223-263. https://doi.org/10.1007/s10107-018-1329-6

[14] Zhang, X. and Peng, D.T. (2022) Solving Constrained Nonsmooth Group Sparse Optimization via Group Capped-ℓ_1 Relaxation and Group Smoothing Proximal Gradient Algorithm. *Computational Optimization and Applications*, **83**, 801-844. https://doi.org/10.1007/s10589-022-00419-2

[15] Friedman, J., Hastie, T. and Tibshirani, R. (2010) A Note on the Group Lasso and a Sparse Group Lasso. https://arxiv.org/pdf/1001.0736

[16] Qin, Z., Scheinberg, K. and Goldfarb, D. (2013) Efficient Block-Coordinate Descent Algorithms for the Group Lasso. *Mathematical Programming Computation*, **5**, 143-169. https://doi.org/10.1007/s12532-013-0051-x

[17] Richtárik, P. and Takáč, M. (2014) Iteration Complexity of Randomized Block-Coordinate Descent Methods for Minimizing a Composite Function. *Mathematical Programming*, **144**, 1-38. https://doi.org/10.1007/s10107-012-0614-z

[18] Argyriou, A., Micchelli, C.A. and Pontil, M. (2011) Efficient First Order Methods for Linear Composite Reularizers. https://doi.org/10.48550/arXiv.1104.1436

[19] Candés, E.J. and Plan, Y. (2011) A Probabilistic and Ripless Theory of Compressed Sensing. *IEEE Transactions on Information Theory*, **57**, 7235-7254. https://doi.org/10.1109/TIT.2011.2161794

[20] Candés, E.J., Romberg, J. and Tao, T. (2004) Robust Uncertainty Principles: Exact Signal Reconstruction from Highly Incomplete Frequency Information. *IEEE Transactions on Information Theory*, **52**, 489-509. https://doi.org/10.1109/TIT.2005.862083

[21] Fan, J.Q. and Li, R.Z. (2001) Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association*, **96**, 1348-1360. https://doi.org/10.1198/016214501753382273

[22] Bauer, F. (2016) Proximal Newton-Type Methods for Large-Scale Non-Smooth Optimization. Master's Thesis, Technische Universität München, München.

[23] Peng, D.T. and Chen, X. (2020) Computation of Second-Order Directional Stationary Points for Group Sparse Optimization. *Optimization Methods and Software*, **35**, 348-376. https://doi.org/10.1080/10556788.2019.1684492

[24] Bian, W. and Chen, X.J. (2020) A Smoothing Proximal Gradient Algorithm for Nonsmooth Convex Regression with Cardinality Penalty. *SIAM Journal on Numerical Analysis*, **58**, 858-883. https://doi.org/10.1137/18M1186009

[25] Micchelli, C.A., Shen, L. and Xu, Y. (2011) Proximity Algorithms for Image Models: Denoising. *Inverse Problems*, **27**, Article 045009. https://doi.org/10.1088/0266-5611/27/4/045009

[26] Beck, A. and Teboulle, M. (2009) Gradient-Based Algorithms with Applications to Signal-Recovery Problems. In: Palomar, D.P. and Eldar, Y.C., Eds., *Convex Optimization in Signal Processing and Communications*, Cambridge University Press, Cambridge, 42-88. https://doi.org/10.1017/CBO9780511804458.003