Advances in Applied Mathematics 应用数学进展, **2023, 12(8), 3665-3683**
Published Online August 2023 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2023.128364

Hans汉斯

稀疏逻辑回归问题的一个光滑化共轭梯度算法

李飘云，韦潇鹏*，唐敏笙

桂林电子科技大学数学与计算科学学院, 广西 桂林

收稿日期：2023年7月21日；录用日期：2023年8月13日；发布日期：2023年8月24日

摘　要

稀疏逻辑回归是一种具有稀疏约束的逻辑回归模型，它广泛应用于神经网络、机器学习和生物信息领域。本文基于近似l_1-范数的思想，采用六个光滑函数对稀疏逻辑回归模型中的l_1- 范数的每个分量进行近似，将问题转换为光滑化无约束最小化问题，然后设计共轭梯度法求解近似模型并给出收敛性分析。最后通过数值实验与已知求解稀疏逻辑回归模型的四个算法进行比较，得出共轭梯度法求解稀疏逻辑回归问题是有效的。

关键词

稀疏逻辑回归，共轭梯度法，光滑函数

A Smoothing Conjugate Gradient Algorithm for Sparse Logistic Regression Problems

Piaoyun Li, Xiaopeng Wei*, Minsheng Tang

School of Mathematics and Computational Science, Guilin University of Electronic Science and Technology, Guilin Guangxi

* 通讯作者。

文章引用: 李飘云, 韦潇鹏, 唐敏笙. 稀疏逻辑回归问题的一个光滑化共轭梯度算法[J]. 应用数学进展, 2023, 12(8): 3665-3683. DOI: 10.12677/aam.2023.128364

Received: Jul. 21st, 2023; accepted: Aug. 13th, 2023; published: Aug. 24th, 2023

Abstract

Sparse logistic regression is a kind of logistic regression model with sparse constraints, which is widely used in the fields of neural networks, machine learning, and bioinformatics. In this paper, based on the idea of approximating the l_1 norm, six smooth functions are used to approximate each component of the l_1 norm in the sparse logistic regression model, and the problem is transformed into a smoothed unconstrained minimization problem, then a conjugate gradient method is designed to solve the approximated model and the convergence analysis is given. Finally, numerical experiments are conducted to compare with four known algorithms for solving the sparse logistic regression model, and it is concluded that the conjugate gradient method is effective in solving the sparse logistic regression problem.

Keywords

Sparse Logistic Regression, Conjugate Gradient Method, Smoothing Function

1. 引言

逻辑回归是一种用于解决分类问题的特殊非线性回归模型, 在机器学习、数据挖掘、医学检测和统计学等领域有着广泛作用. 由于逻辑回归损失函数是严格凸函数, 因此在样本矩阵是全行秩的情况下有唯一的最小化解. 当样本数大于特征数时, 即m ≥ n 时, 最小化效果相对较好. 当m < n 的情况可能会导致过拟合现象. 一方面, m < n 的情况经常出现在许多实际应用中. 例如一份基因表达数据样本是由数千个基因组成的, 但是一般的医疗器械只能获得很有限的样本. 另一方面, 尽管这些数据中有许多特征, 但只有小部分特征是重要的. 随着时代的不断发展, 对于应用逻辑回归模型来解决实际问题的要求不断更新, 为了准确地描述实际问题的上述两个特征, 需要在原问题的基础上增加稀疏约束以减少特征数量. 这意味着在数据挖掘和机器学习应用中, 相对于输入特征而言一个稀

疏的决策规则能够更快地预测或更好地解释模型. 很多学者通过对逻辑回归模型进行正则化处理, 从而获得稀疏解, 进而提出了稀疏逻辑回归问题. 这种改进使得在处理大维度数据中占有很大优势.

当前求解稀疏逻辑回归问题的算法有很多, 例如快速迭代收缩阈值算法 [1]、邻近梯度算法 [2]等. 但求解逻辑回归的算法主要有梯度下降法和坐标下降法, 很多学者基于这两种算法给出了许多改进的算法, 如Kim [3]等人针对逻辑回归问题提出了一种内点法取得了较好的效果; Friedman [4]等人设计了一种在每次迭代开始时用二阶近似替换损失项，然后应用循环坐标下降法来最小化二次函数的快速算法, 该算法可以有效地处理稀疏特征; Yuan [5]等人通过考虑较少地计算损失函数的方法提出了一种改进的GLMNET 算法用来求解正则化逻辑回归问题; 另外Yuan [6]等人还提出了一种使用一维牛顿方向的坐标下降用于处理正则化问题; Bian [7]等人基于坐标下降牛顿法提出了一种不需要数据预处理的并行算法处理大规模l_1 逻辑回归问题, 该方法能够更好地利用并行性而且有很好的加速效果. Peng [8]等人基于随机逼近思想提出了求解惩罚逻辑回归的次线性收敛算法, 并将该算法应用在分布式存储系统. Hadoop [9]在海量数据集上实现显著加速. 此外, Yu [10]等人应用坐标下降法来求解逻辑回归的对偶问题, Balamurugan [11]等人提出了一种交替优化方法求解弹性网络正则化逻辑回归问题, 所得到的模型具有非常好的稀疏效果. 此外, 对于逻辑回归的l_1- 范数正则化问题的研究目前也有很多改进的算法, 例如梯度投影算法 [12]是最早求解l_1-范数正则化问题的基于梯度的算法之一, 它是将其重新表述为盒约束二次规划并使用梯度投影算法来求解. 近年来对l_1- 范数正则化问题的求解研究最广泛的一阶算法是迭代收缩/阈值(IST) 算法 [13]. 根据IST 算法, 在不同的优化方案和技术下提出了许多不同的算法. 例如Hale [14]等人提出了一种基于算子分割技术的IST 不动点延拓方法. 数值结果表明加速IST 方法(两种IST [15]和快速IST[1]) 具有较好的收敛性. 此外NESTA [16]给出了l_1- 范数的光滑函数, 然后使用Nesterov 的梯度法得到原问题的解; Tsuruoka [17]等人在逻辑回归问题的求解中使用随机梯度下降法, 比拟牛顿法数值效果更显著.

共轭梯度法是一种有效地解决大规模非线性优化问题的算法. 它是一种在最速下降法和牛顿法之间的优化方法 [18]. 相比最速下降法求解优化问题的收敛速度得到了提高，同时也避免了牛顿法对于存储的要求以及在计算Hessian 矩阵和求逆矩阵的不足. 共轭梯度法的显著优点是其迭代步骤简便，只需求解出目标函数的一阶导数，对于所需的内存要求低. 随着大数据时代的快速发展，非线性共轭梯度法也得到了发展，并广泛应用于图像恢复、压缩感知 [19]等领域. 近年来针对共轭梯度法的研究主要集中在两个方面，一个是对共轭参数β_k 进行改进，另外一个是将不同的共轭梯度法进行混合，扬长避短达到改进算法的目的. 对共轭参数β_k 的直接修改是提高算法效率的有效手段，许多学者在这方面取得了较好的效果，例如Polak [20]等人与Polyak [21]提出了PRP共轭梯度法，该方法通过对FR 方法中共轭参数β_k进行修改从而提高了FR 方法的收敛速度; Hu [22]等人提出了一种新的共轭梯度法，这种新的算法通过扩展Polak-Ribiére-Polyak [21]提出的共轭参数, 使搜索方向满足独立于任何线搜索的充分下降条件，分别证明了在标准Wolfe线搜索条件以及标准Armijo 线搜索策略下的全局收敛性. 另外对于混合共轭梯度法研究的方法也有很多结果，Touati-Ahmed [23]等人首次将FR法与PRP 法结合, 引入了混合共轭梯度法, 这是对共轭梯度算法的一次有效尝试, 新算法同时具备两种方法的优点并克服了FR 法数值表现一般的不足; Yuan [24]等人设计了一种具有充分下降性质和信赖域性质的修正共轭梯度法, 该算法的搜索方向充分利用了最速下降算法与经典LS 共轭梯度方法的凸组合性质, 并在一定条件下建立了全局收敛性

质, 数值结果表明该修正方法对非线性方程组和图像复原问题是有效的.

本文通过采用六个光滑函数近似l_1-范数的每个分量, 将稀疏逻辑回归问题转化为求解光滑化后的无约束优化问题, 设计一个下降共轭梯度法对光滑后的l_1-范数逻辑回归模型进行求解, 并给出收敛性分析, 通过数值实验分析六个光滑函数近似稀疏逻辑回归问题的效果, 并将共轭梯度算法与已知的求解稀疏逻辑回归问题中的四个算法进行比较, 得出共轭梯度算法求解稀疏逻辑回归问题的有效性.

2. 预备知识

本节给出本文所涉及的相关定义和结论.

考虑独立同分布观测值$(a_i, b_i), i = 1, 2, \ldots, m$, 其中$a_i \in R^n$为预测因素, $b_i \in \{0, 1\}$为二进制响应值. 设矩阵$X = ({a_1}^T, {a_2}^T, \ldots, a_m^T)^T$, b为响应向量, 即$b = (b_1, b_2, \ldots, b_m)^T$ 且x∈R^n 为参数向量, 分类器标签y的条件概率形式如下 [25]:

$$\Pr ob(b_i|a_i) = \frac{\exp(b_i\eta_i(x))}{1 + \exp(\eta_i(x))}, i = 1, 2, \cdots m,$$

其中$\eta(x) = (\eta_1(x), \eta_2(x), \ldots, \eta_m(x))^T$，且$\eta_i(x) = {a_i}^T x$. 对数似然函数形式如下:

$$\log \prod_{i=1}^{m} \Pr ob(b_i|a_i) = \sum_{i=1}^{m} \{b_i\eta_i(x) - \log(1 + \exp(\eta_i(x)))\},$$

因此，我们得到经典的Logistic回归模型:

$$\min_{x \in R^n} f(x) := \sum_{i=1}^{m} \{\log(1 + \exp(\eta_i(x))) - b_i\eta_i(x)\}. \tag{2-1}$$

平均逻辑损失函数定义为

$$\min_{x \in R^n} (1/m) f(x). \tag{2-2}$$

稀疏逻辑回归(SLR), 即$l1$-正则化逻辑回归, 是具有稀疏约束的逻辑回归模型, 广泛应用于分类和特征选择问题 [26]. 稀疏逻辑回归的数学模型定义如下：

$$\min_{x \in R^n} F(x) := f(x) + \lambda||x||_1. \tag{2-3}$$

其中：

$$f(x) = \frac{1}{m} \sum_{i=1}^{m} \log(1 + \exp(-b_i a_i^T x)). \tag{2-4}$$

是平均逻辑损失函数, $\lambda||x||_1$是l_1正则化函数, $\lambda > 0$. 问题的输入数据是一个训练数据点集$\{(a_i, b_i) \in R^n \times \{-1, 1\}, i = 1, 2, \ldots, m, m > 0\}$.

由于绝对值函数$|t|$在0处是不光滑的, 因此l_1-范数正则化极小化问题是一个不可微问题. 为了克服这一困难, 在本节中我们引入$|t|$ 的六个光滑函数尝试对l_1-范数的每个分量近似. 下面给出$|t|$光滑函数的定义和构造过程:

定义2.1 [19]: 如果同时满足以下两个条件:

(i)ψ在$(\mu,t)\in R_{++}\times R$连续可微;

(ii)对$\forall t\in R,\lim\limits_{\mu\to 0}\psi(\mu,t)=|t|$. 那么称函数$\psi$:$R_{++}\times R\to R$是$|t|$的光滑函数.

下面我们将介绍如何构造$|t|$的光滑函数. 首先可以观察到$|t|$可以分为两部分:

$$|t|=(t)_+-(t)_-=(t)_++(-t)_+,$$

其中$(t)_+=\max\{0,t\},(t_-)=\min\{0,t\}$. $(t)_+$ 的光滑函数可以通过引入一个密度（核）函数$d(t)$进行构造, 其满足$d(t)\geq 0$ 和$\int_{-\infty}^{+\infty}d(t)dt=1$. 定义

$$\widehat{s}(t,\mu):=\frac{1}{\mu}d(\frac{t}{\mu}),$$

其中μ是一个正参数. 如果下列条件成立

$$\int_{-\infty}^{+\infty}|t|d(t)dt<+\infty,$$

那么

$$\widehat{p}(t,\mu)=\int_{-\infty}^{+\infty}(t-s)_+\widehat{s}(s,\mu)ds=\int_{-\infty}^{t}(t-s)\widehat{s}(s,\mu)ds\approx(t)_+$$

是$(t)_+$的光滑逼近. 对于加函数$(t)_+=\max\{0,t\}$，目前已有一些著名的光滑函数 [27]，例如:

$$\widehat{\Psi_1}(\mu,t)=t+\mu\ln(1+e^{\frac{t}{\mu}}),\quad \widehat{\Psi_2}(\mu,t)=\begin{cases}t & if \quad t\geq\frac{\mu}{2},\\ \frac{1}{2\mu}(t+\frac{\mu}{2})^2 & if \quad -\frac{\mu}{2}<t<\frac{\mu}{2},\\ 0 & if \quad t\leq-\frac{\mu}{2},\end{cases}\tag{2-5}$$

$$\widehat{\psi_3}(\mu,t)=\frac{\sqrt{4\mu^2+t^2}+t}{2},\quad \widehat{\psi_4}(\mu,t)=\begin{cases}t-\frac{\mu}{2} & if \quad t>\mu,\\ \frac{t^2}{2\mu} & if \quad 0\leq t\leq\mu,\\ 0 & if \quad t<0.\end{cases}\tag{2-6}$$

它们对应的核函数分别如下

$$d_1(t)=\frac{e^{-x}}{(1+e^{-x})^2},\quad d_2(t)=\begin{cases}1 & if -\frac{1}{2}\leq x\leq\frac{1}{2},\\ 0 & otherwise,\end{cases}$$

$$d_3(t)=\frac{2}{(x^2+4)^{\frac{3}{2}}},\quad d_4(t)=\begin{cases}1 & if\ \ 0\le x\le 1,\\ 0 & otherwise.\end{cases}$$

本文中我们通过卷积构造$|t|$的光滑函数 [27,28], 如下所示:

$$\widehat{p}(|t|,\mu)=\widehat{p}(t,\mu)+\widehat{p}(-t,\mu)=\int_{-\infty}^{+\infty}|t-s|\widehat{s}(s,\mu)ds.$$

通过(2-5)-(2-6), 我们得到$|t|$的四个光滑函数

$$\psi_1(\mu,t)=\mu[\ln(1+e^{-\frac{t}{\mu}})+\ln(1+e^{\frac{t}{\mu}})],\quad \psi_2(\mu,t)=\begin{cases}t & if \quad t\ge\frac{\mu}{2},\\ \frac{t^2}{\mu}+\frac{\mu}{4} & if \quad -\frac{\mu}{2}<t<\frac{\mu}{2},\\ -t & if \quad t\le-\frac{\mu}{2},\end{cases}\tag{2-7}$$

$$\psi_3(\mu,t)=\sqrt{4\mu^2+t^2},\quad \psi_4(\mu,t)=\begin{cases}\frac{t^2}{2\mu} & if\ \ |t|\le\mu,\\ |t|-\frac{\mu}{2} & if\ \ |t|>\mu.\end{cases}\tag{2-8}$$

特别地，如果取伊帕涅奇尼科夫核函数

$$d(t)=\begin{cases}\frac{3}{4}(1-t^2) & if\ |t|\le 1,\\ 0 & otherwise.\end{cases}$$

那么$|t|$的光滑函数为

$$\psi_5(\mu,t)=\begin{cases}t & if \quad t>\mu,\\ -\frac{t^4}{8\mu^3}+\frac{3t^2}{4\mu}+\frac{3\mu}{8} & if \quad -\mu\le t\le\mu,\\ -t & if \quad t<-\mu.\end{cases}\tag{2-9}$$

此外，对于所有$t\in R$, 取高斯核函数$d(t)=\frac{1}{\sqrt{2\pi\mu^2}}e^{-\frac{1}{2}}$, 可得

$$\widehat{s}(t,\mu):=\frac{1}{\mu}d(\frac{t}{\mu})=\frac{1}{\sqrt{2\pi\mu^2}}e^{-\frac{t^2}{2\mu^2}},$$

这就得到了$|t|$的另一种光滑函数:

$$\psi_6(\mu,t)=terf(\frac{t}{\sqrt{2}\mu})+\sqrt{\frac{2}{\pi}}\mu e^{-\frac{t^2}{2\mu^2}}.\tag{2-10}$$

综上我们利用卷积得到了六个光滑函数, 下面我们将利用它们近似稀疏逻辑回归问题的l_1-范数的每个分量.

本文结合以上我们所构造的光滑函数以及近似后的稀疏逻辑回归模型, 尝试设计共轭梯度法对

光滑稀疏逻辑回归问题进行求解. 下面我们先简要介绍共轭梯度法.

考虑以下无约束的优化问题

$$\min f(x), x \in \mathrm{R}^{\mathrm{n}}$$

其中$f: R^n \to R$是一个连续可微的函数，用$g(x)$来表示$f(x)$ 的梯度向量, 即

$$g(x) = \nabla f(x).$$

共轭梯度法的迭代过程由以下公式给出

$$x_{k+1} = x_k + \alpha_k d_k, k = 0, 1 \ldots .$$

其中x_k是当前迭代点, x_0是给定的初始点，α_k 为沿着d_k方向上根据某种线搜索规则得到的步长，d_k为搜索方向且定义为

$$d_k = \begin{cases} -g_k, & k = 0, \\ -g_k + \beta_k d_{k-1}, & k > 0. \end{cases} \tag{2-11}$$

式(2-11)中的g_k表示$\nabla f(x_k)$, 即$f(x)$在点x_k 处的梯度方向. 共轭梯度法的关键在于构造迭代搜索方向，即式(2-11) 中参数β_k 的选择. 一般情况下β_k 的不同选择对应着不同类型的共轭梯度算法. 下列是几种著名的共轭梯度法参数β_k 的形式：

$$\beta_k^{FR} = \frac{\|g_k\|^2}{\|g_{k-1}\|^2}, \quad \beta_k^{PRP} = \frac{g_k^T(g_k - g_{k-1})}{\|g_{k-1}\|^2}, \tag{2-12}$$

$$\beta_k^{HS} = \frac{g_k^T(g_k - g_{k-1})}{d_{k-1}^T g_{k-1}}, \quad \beta_k^{CD} = \frac{\|g_k\|^2}{d_{k-1}^T g_{k-1}}, \quad \beta_k^{DY} = \frac{\|g_k\|^2}{d_{k-1}^T(g_k - g_{k-1})}, \tag{2-13}$$

其中$||\cdot||$表示向量的欧几里得范数. 上述几个表达式对于凸二次函数在精确线搜索下是相等的，但是当目标函数不是二次函数时算法的数值结果表现与收敛性会随着参数β_k 与所选择的线搜索的变化而不同, 但是数值结果表现好的PRP方法在某些线搜索中不能确保全局收敛性, 因此Hu [22]等人提出了一种新的共轭参数β_k 形式，使得改进后的共轭梯度算法在Wolfe 线搜索或者是Armijo 线搜索下都能够具有良好的收敛性，并且该共轭梯度算法(NPRP)生成的搜索方向满足充分下降条件，其参数β_k 如下：

$$\beta_k = \frac{\|g_k\|^2 - \frac{\|g_k\|}{\|g_{k-1}\|} g_k^T g_{k-1}}{\max\left\{\mu \|g_k\| \|d_{k-1}\|, \|g_{k-1}\|^2\right\}} = \frac{g_k^T(\|g_{k-1}\| g_k - \|g_k\| g_{k-1})}{\max\left\{\|g_{k-1}\|^3, \mu \|g_k\| \|g_{k-1}\| \|d_{k-1}\|\right\}}, \tag{2-14}$$

其中$\mu > 2$. 上述新的PRP共轭梯度法具有以下两个特点：

(1)NPRP方法满足独立于任意线搜索的充分下降条件；

(2)算法在没有目标函数的凸性假设下，采用标准Wolfe线搜索条件或标准Armijo 线搜索策略全局收敛.

引理2.1 [22] 设搜索方向d_k由NPRP方法生成，则

$$g_k^T d_k \leq -(1-\frac{2}{\mu})\|g_k\|^2,$$

其中$\mu > 2$.

3. 算法描述及其收敛性分析

基于上一节所给出的六个光滑函数, 我们可以得到如下稀疏逻辑回归$F(x)$的近似模型.

令$\varphi_i(\mu, x)$为l_1-范数$\|x\|_1$ 的光滑函数，$\varphi_i(\mu, x)$可表示如下：

$$\varphi_i(\mu, x) := \sum_{j=1}^{n} \psi_i(\mu, x_j), \quad i = 1, 2, 3, 4, 5, 6. \tag{3-1}$$

则有

$$F_{\mu_k}(x) := \lambda\varphi_i(\mu, x) + f(x), x \in R^n. \tag{3-2}$$

其中

$$f(x) = \frac{1}{m}\sum_{i=1}^{m} \ln(1 + \exp(-b_i a_i^T x)).$$

因此，求解问题(2-3)转化为求解以下光滑无约束最小化问题：

$$\min_{x \in R^n} F_{\mu_k}(x). \tag{3-3}$$

在本文中我们将尝试使用共轭梯度算法求解问题(3-3). 下面给出如下光滑化共轭梯度算法.

算法3.1(光滑化共轭梯度法)

Step 1. 给定初始点$x_0 \in R^n$，$\varepsilon > 0$，$\mu > 2$，$\mu_0 > 0$，选择参数$\rho, \delta, \gamma, \overline{\gamma} \in (0, 1)$，令$k = 0$;

Step 2. 如果满足终止条件$\|g_k\| \leq \varepsilon$，则停止；否则转到Step 3;

Step 3. 如果$k = 0$，则$d_0 = -g_0$，否则利用公式(2-11)计算搜索方向d_k，公式(2-14)计算β_k^*;

Step 4. 确定$\alpha_k = \max\{\rho^j, j = 0, 1, 2, \ldots\}$，满足

$$F_{\mu_k}(x_k + \alpha_k d_k) \leq F_{\mu_k}(x_k) - 2\delta(1-\gamma)\alpha_k(\nabla F_{\mu_k}(x_k))^T d_k.$$

Step 5. 令$x_{k+1} = x_k + \alpha_k d_k$，且$\mu_{k+1} = \overline{\gamma}\mu_k$，$k = k + 1$，返回Step 2.

注：上面给出的算法与文献 [19]和文献 [22]的区别在于文献 [19]采用的是另外一种共轭梯度算法求解稀疏优化问题，而算法3.1 采用的是文献 [22]中共轭梯度法，但是它所使用的线搜索与文献 [22]不同.

为了分析上述共轭梯度法的收敛性质，我们首先给出三个引理.

引理3.1 若$f(x)=\frac{1}{m}\sum_{i=1}^{m}\log(1+\exp(-b_ia_i^Tx))$, 则$\|\nabla f(x)-\nabla f(y)\|\leq\frac{1}{m}\|A\|^2\|x-y\|$

证明：因为$\nabla f(x)=-\frac{1}{m}\sum_{i=1}^{m}(1-p_i(x))b_ia_i$，其中$p_i(x)=\frac{1}{1+\exp(-b_ia_i^Tx)}$.
又因为$\nabla^2 f(x)=\frac{1}{m}\sum_{i=1}^{m}(1-p_i(x))p_i(x)a_ia_i^T$ 从而$\nabla f(x)-\nabla f(y)=\int_0^1\nabla^2 f(y+t(x-y))(x-y)dt$.
记$A=[a_1,a_2,\cdots,a_m]^T$, $W(x)$是由$\{(1-p_i(x))p_i(x)\}_{i=1}^m$ 生成的对角阵，则有

$$\begin{aligned}\|\nabla f(x)-\nabla f(y)\|&\leq\int_0^1\|\nabla^2 f(y+t(x-y))(x-y)\|dt\\&\leq\int_0^1\left\|\tfrac{1}{m}A^TW(y+t(x-y))A\right\|\|x-y\|dt\\&\leq\int_0^1\tfrac{1}{m}\|A\|^2\|W(y+t(x-y))\|\|x-y\|dt\\&\leq\tfrac{1}{m}\|A\|^2\|x-y\|,\end{aligned}$$

其中$\|W(y+t(x-y))\|\leq 1$.

引理3.2 设函数$F_\mu(x)$如(3-2)定义, 则存在一个常数$L>0$, 对所有的$x,y\in R^n$, 有$\|\nabla F_\mu(x)-\nabla F_\mu(y)\|\leq L\|x-y\|$成立.

证明：对任意固定的$\mu>0$，为了证明$\nabla F_\mu(x)$的Lipschitz 性质，需要验证$\psi_i'(i=1,2,3,4,5,6)$的Lipschitz条件. 因此，我们分以下两种情况讨论.

情况(1): 当$i=1,3,5,6$时，对$\forall t_1,t_2\in R$, 令$t_1<t_2$，由拉格朗日中值定理可得

$$\left|\psi_i'(\mu,t_1)-\psi_i'(\mu,t_2)\right|=\left|\psi_i''(\mu,\xi)\right||t_1-t_2|,\xi\in(t_1,t_2).$$

为了后续的分析需要，对于每个$i=1,3,5,6$，我们需要估计$\left|\psi_i''(\mu,\xi)\right|$.

对于$i=1$, 我们知道

$$\left|\psi_1''(\mu,\xi)\right|=\frac{1}{\mu}\left[\frac{e^{\frac{\xi}{\mu}}}{\left(1+e^{\frac{\xi}{\mu}}\right)^2}+\frac{e^{\frac{-\xi}{\mu}}}{\left(1+e^{\frac{-\xi}{\mu}}\right)^2}\right]<\frac{2}{\mu}.$$

对于$i=3$, 很明显我们可以得到

$$\left|\psi_3'(\mu,\xi)\right|=\frac{4\mu^2}{\left(4\mu^2+\xi^2\right)^{3/2}}<\frac{1}{2\mu}.$$

对于$i=5$, 我们有

$$\psi_5'(\mu,t)=\begin{cases}1 & if\ t>\mu,\\ -\frac{t^3}{2\mu^3}+\frac{3t}{2\mu} & if\ -\mu\leq t\leq\mu,\\ -1 & if\ t<-\mu.\end{cases}\qquad\psi_5''(\mu,t)=\begin{cases}0 & if\ t>\mu,\\ -\frac{3t^2}{2\mu^3}+\frac{3}{2\mu} & if\ -\mu\leq t\leq\mu,\\ 0 & if\ t<-\mu.\end{cases}$$

对于$i=6$, 我们计算

$$\psi_6{}'(\mu,t)=\frac{2}{\sqrt{\pi}}\int_0^{\frac{t}{\sqrt{2}\mu}}e^{-u^2}du,\quad \psi_6{}''(\mu,t)=\frac{\sqrt{2}}{\sqrt{\pi}\mu}e^{-\frac{-t^2}{2\mu^2}},$$

由此可得出

$$\left|\psi_6{}''(\mu,t)\right|<\frac{\sqrt{2}}{\sqrt{\pi}\mu}.$$

根据上述对$\left|\psi_i{}''(\mu,\xi)\right|$ 估计的结果表明

$$\left|\psi_i{}'(\mu,t_1)-\psi_i{}'(\mu,t_2)\right|\le\frac{3}{\mu}\left|t_1-t_2\right|,i=1,3,5,6,\forall t_1,t_2\in R. \tag{3-4}$$

情况(2): 当$i=2,4$时, 类似情况(1), 同理将$\psi_2{}'$和$\psi_4{}'$ 进行估计.

对于$i=2$，我们有

$$\psi_2{}'(\mu,t)=\begin{cases}1 & if\ t\ge\frac{\mu}{2},\\ \frac{2t}{\mu} & if\ -\frac{\mu}{2}<t<\frac{\mu}{2},\\ -1 & if\ t\le-\frac{\mu}{2}.\end{cases}$$

由于t在每个区间对应不同的$\psi_2{}'(\mu,t)$表达式, 因此我们对t 的区间范围进行讨论:
(a)如果$t_1\ge\frac{\mu}{2}$，$t_2\ge\frac{\mu}{2}$,$t_1\le-\frac{\mu}{2}$，$t_2\le-\frac{\mu}{2}$或者$t_1,t_2\in(-\frac{\mu}{2},\frac{\mu}{2})$，那么

$$\left|\psi_2{}'(\mu,t_1)-\psi_2{}'(\mu,t_2)\right|\le\frac{2}{\mu}\left|t_1-t_2\right|.$$

如果$t_1\ge\frac{\mu}{2}$，$t_2\le-\frac{\mu}{2}$，则有

$$\left|\psi_2{}'(\mu,t_1)-\psi_2{}'(\mu,t_2)\right|=2=\mu\frac{2}{\mu}\le\frac{2}{\mu}\left|t_1-t_2\right|.$$

(b)如果$t_1\ge\frac{\mu}{2}$, $t_2\in(-\frac{\mu}{2},\frac{\mu}{2})$, 那么

$$\left|\psi_2{}'(\mu,t_1)-\psi_2{}'(\mu,t_2)\right|=1-\frac{2t_2}{\mu}<\frac{2t_1}{\mu}-\frac{2t_2}{\mu}=\frac{2}{\mu}\left|t_1-t_2\right|.$$

(c)如果$t_1\le-\frac{\mu}{2}$，$t_2\in(-\frac{\mu}{2},\frac{\mu}{2})$，那么

$$\left|\psi_2{}'(\mu,t_1)-\psi_2{}'(\mu,t_2)\right|=1+\frac{2t_2}{\mu}<-\frac{2t_1}{\mu}+\frac{2t_2}{\mu}=\frac{2}{\mu}\left|t_1-t_2\right|.$$

根据上述讨论我们可以得出以下结论

$$\left|\psi_2{}'(\mu,t_1)-\psi_2{}'(\mu,t_2)\right|\le\frac{2}{\mu}\left|t_1-t_2\right|,\ \forall t_1,t_2\in R.$$

对于$i=4$, 运用$i=2$情况下相同的讨论方式, 同样可以得出

$$\left|\psi_4'(\mu,t_1)-\psi_4'(\mu,t_2)\right| \leq \frac{1}{\mu}\left|t_1-t_2\right|,\ \forall t_1,t_2\in R.$$

以上我们对$i=2,4$进行了分析讨论，得出以下结果

$$\left|\psi_i'(\mu,t_1)-\psi_i'(\mu,t_2)\right| \leq \frac{2}{\mu}\left|t_1-t_2\right|,\ i=2,4,\forall t_1,t_2\in R. \tag{3-5}$$

根据以上结论, 对$\forall x,y\in R^n$ 应用(3-4)和(3-5), 同时结合引理3.1, 我们有

$$\begin{aligned}
&\|\nabla F_\mu(x)-\nabla F_\mu(y)\|\\
&=\|\lambda\left[\nabla\varphi_i(\mu,x)-\nabla\varphi_i(\mu,x)\right]+\nabla f(x)-\nabla f(y)\|\\
&\leq \tfrac{3}{\mu}\lambda n\|x-y\|+\tfrac{\|A\|^2}{m}\|x-y\|\\
&=L\|x-y\|,
\end{aligned}$$

其中$L=\frac{3}{\mu}\lambda n+\frac{\|A\|^2}{m}$.

引理3.3 设$\mu>0$，则水平集$L(x_0)=\{x\in R^n|F_\mu(x)\leq F_\mu(x_0)\}$是有界的.

证明：反证法. 假设集合是$L(x_0)$无界的. 那么存在一个指标集K_1，使得$\left\|x^k\right\|\to\infty, k\to\infty, k\in K_1$. 由$F_\mu(x)$ 的定义，我们有$F_\mu(x_k)\to\infty, k\to\infty, k\in K_1$. 这与$F_\mu(x_k)\leq F_\mu(x_0)$ 相矛盾，因此，水平集$L(x_0)$ 是有界的.

下面我们根据上述三个引理给出算法3.1的全局收敛性分析.

定理3.1 设$\{x_k\}$为算法生成的序列, 则

$$\liminf_{k\to\infty}\|\nabla F_\mu(x_k)\|=0. \tag{3-6}$$

证明：首先假设$\liminf\limits_{k\to\infty}\|\nabla F_\mu(x_k)\|=0$不成立, 那么存在一个常数$\varepsilon_0>0$，使得

$$\|\nabla F_\mu(x_k)\|\geq\varepsilon_0,\ \forall k. \tag{3-7}$$

由引理2.1可得$(\nabla F_\mu(x_k))^Td_k\leq-(1-\frac{2}{\mu})\|\nabla F_\mu(x_k)\|^2$. 因此存在$\alpha_k>0$，使得

$$F_{\mu_k}(x_k+\alpha_kd_k)\leq F_{\mu_k}(x_k)-2\delta(1-\gamma)\alpha_k(\nabla F_\mu(x_k))^Td_k. \tag{3-8}$$

这意味着$\{F_\mu(x_k)\}$是递减且有界的, 可得$x_k\in L(x_0)$ 和$\{F_\mu(x_k)\}$ 是收敛的. 令$\lim\limits_{k\to\infty}F_\mu(x_k)=F_*$. 由引理2.1 和(3-8) 可得$\lim\limits_{k\to\infty}\alpha_k\|\nabla F_\mu(x_k)\|=0$. 显然$\|\nabla F_*\|>0$, 因此有

$$\lim_{k\to\infty}\alpha_k=0. \tag{3-9}$$

由$F_\mu(x)$连续可微性知, 存在常数$\bar{r}>0$，使得

$$\|\nabla F_\mu(x_k)\| \leq \bar{r},\ \ \forall k.$$

下面利用反证法证明$\{\|d_k\|\}$ 是有界的. 假设$\{\|d_k\|\}$ 无界，那么存在K_2使得$\|d_k\| \to \infty, k \to \infty, k \in K_2$, 设$\theta_k$为$-\nabla F_\mu(x_k)$与$d_k$之间的夹角, 则有

$$\cos\theta_k = \frac{-(\nabla F_\mu(x_k))^T d_k}{\|\nabla F_\mu(x_k)\|\,\|d_k\|} = \frac{\|\nabla F_\mu(x_k)\|}{\|d_k\|}. \tag{3-10}$$

由上述关系式, 结合$\varepsilon_0 \leq \|\nabla F_\mu(x_k)\| \leq \bar{r}$可以得出$\cos\theta_k \to 0, k \in K_2, k \to \infty$, 即$\theta_k \to \frac{\pi}{2}, k \in K_2, k \to \infty$ 由此可以得出$(\nabla F_\mu(x_k))^T d_k \to 0, k \in K_2, k \to \infty$, 即有如下关系

$$(1-\frac{2}{\mu})\|\nabla F_\mu(x_k)\|^2 \leq -(\nabla F_\mu(x_k))^T d_k \to 0, k \in K_2, k \to \infty,$$

这与(3-7)相矛盾. 因此$\{\|d_k\|\}$ 是有界的, 即存在一个常数$M^*>0$使得

$$\|d_k\| \leq M^*,\ \ \forall k. \tag{3-11}$$

由此将(3-9)与(3-11)结合起来，可以得到

$$\lim_{k\to\infty} \alpha_k \|d_k\| = 0. \tag{3-12}$$

结合以上三个式子, 可得$\cos\theta_k \geq \frac{\varepsilon_0}{M^*}$，进一步得到

$$-(\nabla F_\mu(x_k))^T d_k = \|\nabla F_\mu(x_k)\|\,\|d_k\|\cos\theta_k \geq \frac{{\varepsilon_0}^2}{M^*}\|d_k\|.$$

由中值定理可得

$$\begin{aligned}\nabla F_\mu(x_k+\alpha_k d_k) &= F_\mu(x_k) + \alpha_k(\nabla F_\mu(\xi_k))^T d_k \\ &= F_\mu(x_k) + \alpha_k(\nabla F_\mu(x_k))^T d_k + \alpha_k(\nabla F_\mu(\xi_k)-\nabla F_\mu(x_k))^T d_k \\ &\leq F_\mu(x_k) + \alpha_k\|d_k\|\,(\tfrac{(\nabla F_\mu(x_k))^T d_k}{\|d_k\|} + \|\nabla F_\mu(\xi_k)-\nabla F_\mu(x_k)\|),\end{aligned} \tag{3-13}$$

其中$x_k \leq \xi_k \leq x_k+\alpha_k d_k$. 结合引理3.2和水平集$L(x_0)$ 的紧性，可以得出$\nabla F_\mu(x)$在$L(x_0)$上是一致连续的. 这与(3-12)一起意味着存在一个常数$\widehat{\alpha}>0$且有

$$\|\nabla F_\mu(\xi_k)-\nabla F_\mu(x_k)\| < \frac{1}{2}\frac{{\varepsilon_0}^2}{M^*}. \tag{3-14}$$

当k足够大时, 式(3-13)有

$$\begin{aligned}\nabla F_\mu(x_k+\alpha_k d_k) &\leq F_\mu(x_k) + \widehat{\alpha}(-\tfrac{{\varepsilon_0}^2}{M^*} + \tfrac{1}{2}\tfrac{{\varepsilon_0}^2}{M^*}) \\ &= F_\mu(x_k) - \tfrac{\widehat{\alpha}{\varepsilon_0}^2}{2M^*},\end{aligned} \tag{3-15}$$

这与$F_\mu(x_{k+1}) - F_\mu(x_k) \to 0, k \to \infty$ 矛盾. 于是有下列式子成立

$$\lim_{k\to\infty} \inf \|\nabla F_\mu(x_k)\| = 0.$$

4. 数值实验

为检验本文所提出光滑化共轭梯度算法的有效性, 本小节我们将利用六个光滑函数对l_1- 范数的每个分量近似, 得出六种不同的光滑化稀疏逻辑回归模型, 利用光滑化共轭梯度算法对光滑化后的稀疏逻辑回归模型进行求解, 并将光滑化共轭梯度算法（NCG）的数值结果与邻近梯度法(Proximal Gradient Method) [1]、线性邻近梯度法(LS- Proximal Gradient Method) [1]、加速邻近梯度法(Fast Proximal Gradient Method) [1]、线性加速邻近梯度法(LS-Fast Proximal Gradient Method) [1]的数值结果进行对比分析.

Table 1. Comparison of numerical results of six smoothing functions with 260 iterations

表 1. 迭代260次的六种光滑函数的数值比较结果

光滑函数	目标函数值	误差	CPU时间(单位：s)
ψ_1	Inf	0.003531	8.46
ψ_2	0.801205	0.000002	8.42
ψ_3	0.794706	0.000001	8.44
ψ_4	6.706516	0.000086	50.23
ψ_5	0.797260	0.000002	8.44
ψ_6	0.373946	0.001343	8.24

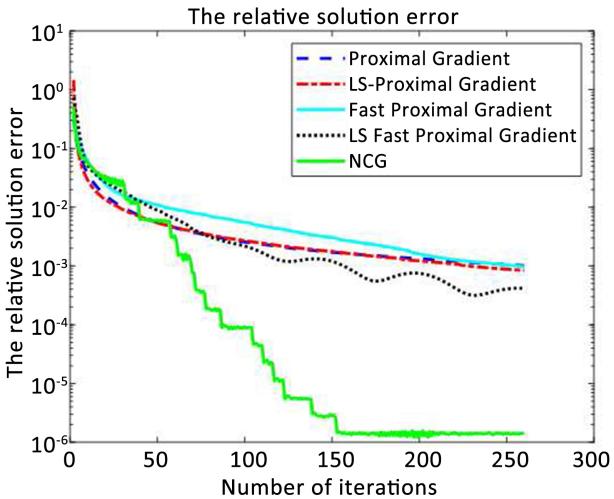

Figure 1. The plot of the relative errors about the approximate solutions of the models corresponding to the six smoothing functions

图 1. 六种光滑函数对应模型近似解相对误差变化图

数值实验中我们所用到的数据信息: $A = (a_1, a_2, \cdots, a_m)^T$ 为7366 × 300 阶的稀疏矩阵，$y =$

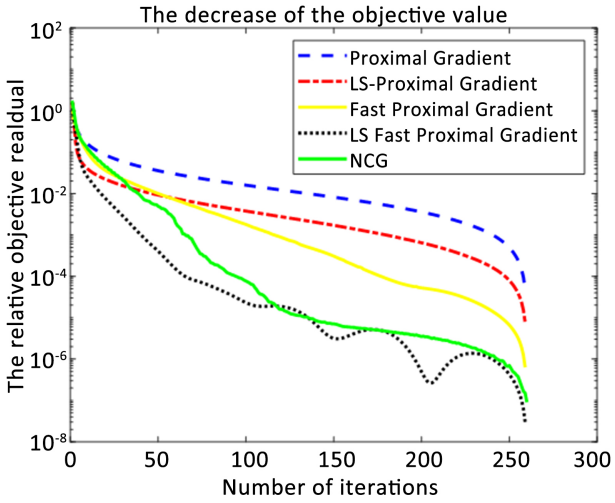

Figure 2. The plot of the relative errors about the objective function values of the models corresponding to the six smoothing functions

图 2. 六种光滑函数对应模型目标函数值相对误差变化图

Table 2. Comparison of numerical results of five algorithms with 260 iterations

表 2. 迭代260次的五种算法的数值结果比较

算法	目标函数值	相对误差	CPU时间(单位：s)
PGM	0.468644	0.001012	7.61
LS-PGM	0.455093	0.000839	10.96
FPGM	0.454174	0.000978	7.65
LS-FPGM	0.454130	0.000418	15.29
NCG	0.794706	0.000001	8.44

$(b_1, b_2, \cdots, b_m)^T$为$7366 \times 1$的向量, 其中$(a_i, b_i) \in R^n \times \{-1, 1\}, i = 1, 2, \ldots, m, m > 0$. 在MATLAB运行过程中, 我们将最大迭代次数设置为260 次, 算法参数取值如下：

$$x^0 = zeros(300, 1), \lambda = 0.008, \rho = 0.5, \delta = 0.002, \gamma = 0.2, \overline{\gamma} = 0.4, \mu_0 = 0.1, \mu = 2.4.$$

当满足下列相对误差时, 算法终止.

$$\frac{\|x_{k+1} - x_k\|}{\|x_k\|} < 1.0e^{-8}.$$

结合以上参数利用MATLAB编程实现对光滑函数近似后的稀疏逻辑回归模型进行求解, 我们得到利用六种光滑函数近似稀疏逻辑回归模型l_1- 范数的数值结果如表1, 相对残差、迭代序列变化如图1、图2 所示, 其中图1、图2中的“1、2、3、4、5、6”分别对应光滑函数$\psi_1, \ldots, \psi_6$.

通过表1和图1、图2的数值结果我们可以得出以下几点结论:

1) 由表1我们得到利用算法求解六个光滑函数对应的光滑稀疏逻辑回归模型, 在相同的迭代次

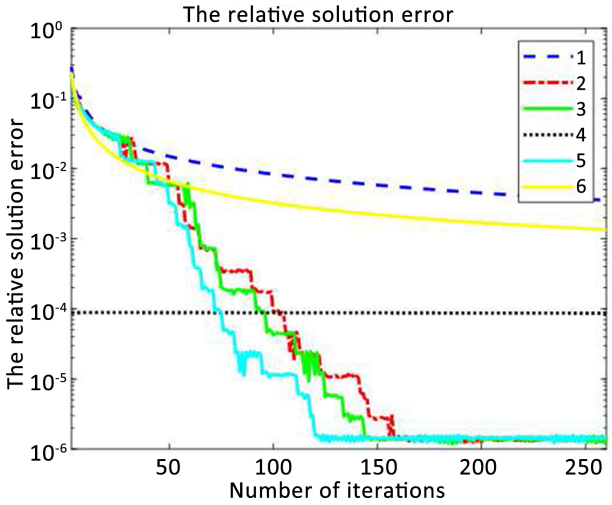

Figure 3. The plot of relative error changes about the approximate solution for the five methods

图 3. 五种方法近似解相对误差变化图

数条件下, 光滑函数ψ_1近似得到的目标函数值是趋于无穷大的,除光滑函数ψ_4 之外, 其余四个光滑函数求解得到的目标函数值相差不大, 并且光滑函数ψ_2，ψ_3，ψ_5目标函数值十分接近, 相差的误差也很小；光滑函数ψ_6的误差虽然相对于除光滑函数ψ_1 外的四个光滑函数的大, 但是它运行所需要消耗的CPU时间是五个光滑函数中最少的. 除此之外虽然光滑函数ψ_2 和ψ_5算法求解结果的误差相等, 但目标函数值稍有不同, 所用的CPU 时间相差很小.

2) 六种光滑函数对应的光滑稀疏逻辑回归模型迭代次数、相对残差和相对误差的收敛行为分别如图1、图2 所示. 从图1、图2可以看出光滑函数ψ_2, ψ_3, ψ_5 各自的相对残差与误差相差很小；结合表1 得到光滑函数ψ_2应用算法求解所消耗的CPU 时间是除光滑函数ψ_4，ψ_6之外最少的, 但是在相同迭代次数下光滑函数ψ_3 目标函数值的误差相较于ψ_2更为理想, 而且从图1、图2直观地可以得出光滑函数所求得目标函数的相对残差与迭代序列相对误差与其余几个光滑函数相比效果最佳，因此我们综合比较得出光滑函数ψ_3 对稀疏逻辑回归模型中的l_1- 范数的每个分量近似效果最好. 综上所述结合图1、图2, 我们可以得到六个光滑函数的相对误差数值表现如下：

$$\psi_3 \succ \psi_2 \approx \psi_5 \succ \psi_4 \succ \psi_6 \succ \psi_1.$$

根据上一部分利用共轭梯度法求解六种光滑函数模型的比较结果得出光滑函数ψ_3 对l_1- 范数每个分量的近似效果最佳, 因此我们将共轭梯度算法求解光滑函数ψ_3 的数值结果与简单邻近梯度法(Proximal Gradient Method)、线性简单邻近梯度法(LS-Proximal Gradient Method)、加速邻近梯度法(Fast Proximal Gradient Method)、线性加速邻近梯度法(LS-Fast Proximal Gradient Method) 的数值结果进行分析比较. 利用上一部分所用到的数据和参数值, 下面我们将最大迭代次数设置为260 次. 五种算法的数值结果比较如表2, 算法相对残差、迭代序列变化如图3、图4所示.

通过表2和图3、图4的数值结果我们可以得出以下几点结论:

1) 由表2我们得出在相同的迭代次数条件下, 简单邻近梯度法(PGM)、线性简单邻近梯度

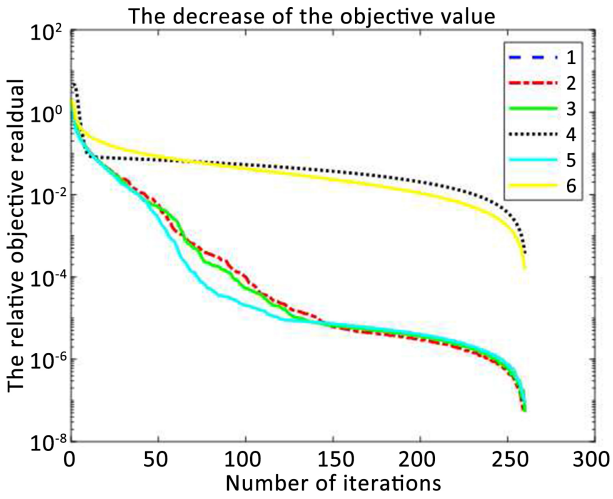

Figure 4. The plot of relative error changes about the objective function values for the five methods

图 4. 五种方法目标函数值相对误差变化图

法(LS-PGM)、加速邻近梯度法(FPGM)、线性加速邻近梯度法(LS- FPGM)四种算法计算得出的目标函数值比光滑化共轭梯度法(NCG)的相对误差值较大, 其误差结果也是比光滑化共轭梯度算法大；在消耗CPU时间方面, 光滑化共轭梯度算法所消耗的CPU时间比LS-PGM、LS-FPGM 这两个算法所要消耗的时间短, 比PGM、FPGM 需要的CPU 时间稍长.

2) 五种算法迭代次数、相对残差和相对误差的收敛行为分别如图3、图4所示. 从图3、图4可以看出, 利用这五种算法对稀疏逻辑回归问题求解，在相同迭代次数的条件下所得到的目标函数的相对残差结果相差不是很大, 光滑化共轭梯度算法在求解稀疏逻辑回归问题所得到的目标函数值的解要比其他四个算法更准确, 误差更小，结合图4可以直观地看到在迭代60次之后, 光滑化共轭梯度算法的相对误差与其余四个算法的迭代序列相对误差差距明显, 光滑化共轭梯度算法求解的相对误差优于其他四个算法, 这表明光滑化共轭梯度算法求解稀疏逻辑回归问题是有效的.

基金项目

国家级大学生创新训练项目: 202110595047.

参考文献

[1] Beck, A. and Teboulle, M. (2009) A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM Journal on Imaging Sciences*, **2**, 183-202. https://doi.org/10.1137/080716542

[2] Scheinberg, K. and Tang, X. (2016) Practical Inexact Proximal Quasi-Newton Method with Global Complexity Analysis. *Mathematical Programming*, **160**, 495-529.

https://doi.org/10.1007/s10107-016-0997-3

[3] Koh, K., Kim, S., Boyd, S., *et al.* (2007) An interior-Point Method for Large-Scale -Regularized Logistic Regression. *Journal of Machine Learning Research*, **8**, 1519-1555.

[4] Friedman, J., Hastie, T. and Tibshirani, R. (2010) Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, **33**, 1-22. https://doi.org/10.18637/jss.v033.i01

[5] Yuan, G.X., Ho, C.H. and Lin, C.J. (2012) An Improved GLMNET for ℓ_1-Regularized Logistic Regression. *The Journal of Machine Learning Research*, **13**, 1999-2030. https://doi.org/10.1145/2020408.2020421

[6] Yuan, G.X., Chang, K.W., Hsieh, C.J., *et al.* (2010) A Comparison of Optimization Methods and Software for Large-Scale ℓ_1-Regularized Linear Classification. *Journal of Machine Learning Research*, **11**, 3183-3234. https://www.jmlr.org/papers/v11/yuan10c.html

[7] Bian, Y., Li, X., Cao, M., *et al.* (2013) Bundle CDN: A Highly Parallelized Approach for Large-Scale ℓ_1-Regularized Logistic Regression. In: Blockeel, H., Kersting, K., Nijssen, S. and Železný, F., Eds., *Machine Learning and Knowledge Discovery in Databases*, Springer, Berlin, Heidelberg, 81-95. https://doi.org/10.1007/978-3-642-40994-3_6

[8] Peng, H., Wang, Z., Chang, E.Y., *et al.* (2012) Sublinear Algorithms for Penalized Logistic Regression in Massive Datasets. In: Flach, P.A., De Bie, T. and Cristianini, N., Eds., *Machine Learning and Knowledge Discovery in Databases*, Springer, Berlin, Heidelberg, 553-568. https://doi.org/10.1007/978-3-642-33460-3_41

[9] Peng, H., Liang, D. and Choi, C. (2013) Evaluating Parallel Logistic Regression Models. *2013 IEEE International Conference on Big Data*, Silicon Valley, CA, 6-9 October 2013, 119-126. https://doi.org/10.1109/BigData.2013.6691743

[10] Yu, H.F., Huang, F.L. and Lin, C.J. (2011) Dual Coordinate Descent Methods for Logistic Regression and Maximum Entropy Models. *Machine Learning*, **85**, 41-75. https://doi.org/10.1007/s10994-010-5221-8

[11] Balamurugan, P. (2013) Large-Scale Elastic Net Regularized Linear Classification SVMs and Logistic Regression. *2013 IEEE 13th International Conference on Data Mining*, Dallas, TX, 7-10 December 2013, 949-954. https://doi.org/10.1109/ICDM.2013.126

[12] Figueiredo, M., Nowak, R. and Wright, S.J. (2008) Gradient Projection for Sparse Reconstruction: Application to Compressed Sensing and Other Inverse Problems. *Journal of Selected Topics in Signal Processing*, **1**, 586-597. https://doi.org/10.1109/JSTSP.2007.910281

[13] Nowak, R. and Figueiredo, M. (2001) Fast Wavelet-Based Image Deconvolution Using the EM Algorithm. *Conference Record of Thirty-Fifth Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, 4-7 November 2001, 371-375.

[14] Hale, E.T., Yin, W. and Zhang, Y. (2009) Fixed-Point Continuation for ℓ_1-Minimization: Methodology and Convergence. *SIAM Journal on Optimization*, **19**, 1107-1130. https://doi.org/10.1137/070698920

[15] Bioucas-Dias, J.M. and Figueiredo, M. (2007) A New TwIST: Two-Step Iterative Shrinkage/Thresholding Algorithms for Image Restoration. *Transactions on Image Processing*, **16**, 2992-3004. https://doi.org/10.1109/TIP.2007.909319

[16] Becker, S., Bobin, J. and Candes, E. (2011) NESTA: A Fast and Accurate First-Order Method for Sparse Recovery. *SIAM Journal Imaging Sciences*, **4**, 1-39. https://doi.org/10.1137/090756855

[17] Tsuruoka, Y., Tsujii, J. and Ananiadou, S. (2009) Stochastic Gradient Descent Training for *L*1-Regularized Log-Linear Models with Cumulative Penalty. *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, **1**, 477-485. https://doi.org/10.3115/1687878.1687946

[18] 陈宝林. 最优化理论与算法[M]. 北京: 清华大学出版社, 2005: 294-305.

[19] Wu, C.Y., Zhan, J.M., Lu, Y., *et al.* (2019) Signal Reconstruction by Conjugate Gradient Algorithm Based on Smoothing ℓ_1-Norm. *Calcolo*, **56**, Article No. 42. https://doi.org/10.1007/s10092-019-0340-5

[20] Polak, E. and Ribiere, G. (1969) Note sur la convergence de méthodes de directions conjuguées. *Revue Française d'Informatique et de Recherche Opérationnelle*, **16**, 35-43. https://doi.org/10.1051/m2an/196903R100351

[21] Polyak, B.T. (1969) The Conjugate Gradient Methods in Extreme Problems. *USSR Computational Mathematics Physics*, **9**, 94-112. https://doi.org/10.1016/0041-5553(69)90035-4

[22] Hu, Q.J., Zhang, H.R. and Chen, Y. (2022) Global Convergence of a Descent PRP Type Conjugate Gradient Method for Nonconvex Optimization. *Applied Numerical Mathematics*, **173**, 38-50. https://doi.org/10.1016/j.apnum.2021.11.001

[23] Touati-Ahmed, D. and Storey, C. (1990) Efficient Hybrid Conjugate Gradient Techniques. *Journal of Optimization Theory and Applications*, **64**, 379-397. https://doi.org/10.1007/BF00939455

[24] Yuan, G.L., Li, T.T. and Hu, W.J. (2020) A Conjugate Gradient Algorithm for Large-Scale Nonlinear Equations and Image Restoration Problems. *Applied Numerical Mathematics*, **147**, 129-141. https://doi.org/10.1016/j.apnum.2019.08.022

[25] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016: 57-60.

[26] Wang, R., Xiu, N. and Zhang, C. (2019) Greedy Projected Gradient-Newton Method for Sparse Logistic Regression. *Transactions on Neural Networks and Learning Systems*, **31**, 527-538. https://doi.org/10.1109/TNNLS.2019.2905261

[27] Saheya, B., Yu, C.H. and Chen, J.S. (2018) Numerical Comparisons Based on Four Smoothing Functions for Absolute Value Equation. *Journal of Applied Mathematics and Computing*, **56**, 131-149. https://doi.org/10.1007/s12190-016-1065-0

[28] Voronin, S., Ozkaya, G. and Yoshida, D. (2014) Convolution Based Smooth Approximations to the Absolute Value Function with Application to Non-Smooth Regularization. arXiv: 1408.6795