Operations Research and Fuzziology 运筹与模糊学, 2023, 13(6), 6441-6449
Published Online December 2023 in Hans. https://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2023.136634

Hans汉斯

非精确牛顿法在自适应三次正则化牛顿方法中的应用

张　林，何清龙，张海芳

贵州大学数学与统计学院，贵州　贵阳

收稿日期：2023年9月22日；录用日期：2023年12月4日；发布日期：2023年12月12日

摘　要

本文基于自适应三次正则化牛顿方法提出了非精确牛顿法自适应三次正则化牛顿方法，并且通过数值实验验证了该方法的单调性和收敛性。本文给出了3种算法。本文使用不同的非精确求解器来求解子优化问题，并且通过数值实验对比了在不同绝对截断误差和不同相对截断误差下非精确求解与精确求解的收敛情况。数值实验结果表明，在绝对截断误差过大时，会导致算法收敛速度变慢，随着绝对截断误差的减少算法的收敛速度逐渐加快。相对截断误差过大时也会出现收敛速度较慢的情况。此外，不同的非精确求解器在数值实验中在算法1上表现差异不大，但在算法2和算法3中却差异较为明显。

关键词

非精确牛顿法，Levenberg-Marquardt正则化方法，三次正则化牛顿方法，无约束优化

Application of Inexact Newton Method in Adaptive Third-Order Regularized Newton Method

Lin Zhang, Qinglong He, Haifang Zhang

School of Mathematics and Statistics, Guizhou University, Guiyang Guizhou

Received: Sep. 22nd, 2023; accepted: Dec. 4th, 2023; published: Dec. 12th, 2023

Abstract

In this paper, based on the adaptive cubic regularized Newton method, an inexact adaptive cubic

文章引用: 张林, 何清龙, 张海芳. 非精确牛顿法在自适应三次正则化牛顿方法中的应用[J]. 运筹与模糊学, 2023, 13(6): 6441-6449. DOI: 10.12677/orf.2023.136634

regularized Newton method is proposed, and the monotonicity and convergence of the method are verified by numerical experiments. Three algorithms are given in this paper. In this paper, different imprecise solvers are used to solve sub-optimization problems, and the convergence of imprecise solutions and exact solutions under different absolute and relative truncation errors is compared by numerical experiments. The numerical results show that when the absolute truncation error is too large, the convergence rate of the algorithm will slow down, and the convergence rate of the algorithm will gradually accelerate with the reduction of the absolute truncation error. When the relative truncation error is too large, the convergence speed will be slow. In addition, different imprecise solvers show little difference in algorithm 1 in numerical experiments, but the difference is obvious in algorithm 2 and algorithm 3.

Keywords

Inexact Newton Method, Levenberg-Marquardt Regularization Method, Third-Order Regularized Newton Method, Unconstrained Optimization

1. 引言

本文主要研究无约束优化问题：

$$\min_{x\in R^d} f(x)$$

其中， $f:R^d\to R$ 是二次可微函数，且其 Hessian 矩阵满足利普希茨条件(Lipschitz condition)。

无约束优化问题是优化理论中最为基本的问题之一，在诸多数学和工程问题中有着广泛应用。

梯度下降法和随机梯度下降法是求解无约束优化问题常用的方法[1] [2] [3]。在机器学习任务中，由于该类算法良好的并行性和较低的计算成本使其得到了广泛的引用。对于具有病态 Hessian 矩阵的问题，一阶方法的迭代收敛速度往往比较慢，极大地限制了梯度下降型方法在此类问题中的应用。相对于一阶梯度型方法，利用 Hessian 矩阵信息的 Newton 法在收敛速度方面具有较快的速度。然而，经典牛顿法仅适用于强凸问题，并且在初始值远离最优值点时可能出现不收敛情况。此外，每次迭代需要求解的 Hessian 矩阵可能存在病态问题，进而导致数值精度误差大和数值求解严重的不稳定性问题。为了解决这些问题，人们提出了多种方法来弥补 Newton 法的缺陷。

为解决 Hessian 矩阵非正定问题，学者提出了 LM 方法。

Levenberg-Marquardt 正则化[4] [5]：如果 $f''(x)$ 不是正定的，可以用一个单位矩阵将其正则化。即，使用 $G=f''(x)+\gamma I\succ 0$ 和 $-G^{-1}f'(x)$ 来进行迭代：

$$x_{k+1}=x_k-\left[f''(x)+\gamma I\right]^{-1}f'(x)$$

然而 Levenberg-Marquardt 正则化方法存在正则化参数 γ 选择的问题。

为解决算法对初始值的依赖的问题，就推动了 Line-search 和 Trust-region 方法的研究。

线搜索(Line search) [6]：这是一种全局优化策略，它采用牛顿方法的更新方向，并在该方向上找到最佳的步长 $h_k>0$：

$$x_{k+1}=x_k-h_k\left[f''(x)\right]^{-1}f'(x)$$

但线搜索方法也存在计算量较大以及当 Hessian 矩阵负定时导致函数值无法下降的缺点。

信赖域方法(Trust-region approach) [7]：根据这种方法，在 x_k 点，必须形成它的邻域。其中函数的二阶近似是可靠的。这是一个信任区域 $\Delta(x_k)$。例如 $\Delta(x_k)=\{x:\|x-x_k\|\le\varepsilon\}(\varepsilon>0)$。然后选择下一个点 x_{k+1} 作为以下辅助问题的解：

$$\min_{x\in\Delta(x_k)}\left[\left\langle f'(x_k),x-x_k\right\rangle+\frac{1}{2}\left\langle f''(x_k)(x-x_k),x-x_k\right\rangle\right]$$

信赖域方法每次迭代需要求解子问题，子问题求解需要求出函数的一阶和二阶信息以及矩阵求逆等操作，计算复杂度较高。

三次正则化(CR)牛顿方法：采用一个初始化 $x_0\in R^d$，一个适当的参数 $M>0$，并生成一个序列 $\{x_k\}_k$ 通过以下更新规则来求解目标函数：

$$s_{k+1}=\arg\min_{s\in R^d}\nabla f(x_k)^{\mathrm{T}}s+\frac{1}{2}s^{\mathrm{T}}\nabla^2 f(x_k)s+\frac{M}{6}\|s\|^3,$$

$$x_{k+1}=x_k+s_{k+1}.$$

三次正则化(CR)牛顿方法相较其他方法在寻找二阶稳定点具有较好的优势[8]。然而，在大规模优化问题中，求解三次子优化问题的精确解是一个非线性问题，所以该方法在求解三次子优化问题的解时有巨大的计算负担。

为了降低三次正则化方法计算量大难题，Richtárik and Doikov [9]提出了二阶块坐标下降且增加了三次正则项的 RBCN 算法，在许多问题上其收敛速度和计算量有了明显的改进；Song and Liu [10]使用近似三次正则化牛顿法(PCNM)来求解目标函数由光滑凸函数和非光滑凸函数组成的优化问题，PCNM 方法通过构造一个近似目标函数，来避免求解 Hessian 矩阵的逆矩阵，并且通过增加一个正则化项，来保证算法的收敛性和稳定性，该方法收敛速率比线性速率收敛更快，但比二次速率收敛更慢的收敛速度。其次，众多的学者还研究了通过各种采样方案降低计算复杂度，如小批量抽样[11]、子抽样[12]、方差减少抽样[13]。

Konstantin Mishchenko 基于 Levenberg-Marquardt 正则化思想提出了正则化 Newton 方法，证明算法收敛性并给出了算法的全局收敛阶数和局部收敛阶数[14]。

经典的三次牛顿更新可以隐式地写成：

$$x_{k+1}=x_k-\left(\nabla^2 f(x_k)+H\|x_{k+1}-x_k\|I\right)^{-1}\nabla f(x_k),$$

其中 $H>0$ 是一个常数，I 是单位矩阵。Konstantin Mishchenko 取 $\gamma=\sqrt{H\|\nabla f(x_k)\|}\approx H\|x_{k+1}-x_k\|$，即更新规则为：

$$x_{k+1}=x_k-\left(\nabla^2 f(x_k)+\sqrt{H\|\nabla f(x_k)\|}I\right)^{-1}\nabla f(x_k),$$

然而这一方法仍然具有线性方程组精确求解较难并且求解精确解需要较大计算量，其次也存在 Hessian 矩阵难求，占用内存大等难题。

本文基于正则化 Newton 方法，采用迭代法近似求解正则化 Newton 法中的高维线性方程组，提出了非精确三次正则化 Newton 方法，通过数值实验验证了方法收敛性和有效性。

2. 非精确三次正则化牛顿方法

2.1. 非精确牛顿法

非精确牛顿法的基本思路是利用迭代方法近似求解 Newton 方程组的解，当精度满足一定要求时就可以停止迭代，以此计算出一个满足一定精度要求的优化解。具体而言，在每次迭代中，非精确牛顿法不需要准确计算出 Hessian 矩阵 H_k，只需要计算出 Hessian 矩阵 H_k 与解向量 s，再结合梯度矢量 g_k 即可以计算出非精确解[15]：

$$s_k^j = s_k^{j-1} + f\left(H_k s_k^{j-1}, g_k\right)$$

然后，当截断误差满足一定条件时就停止迭代。

考虑到非精确牛顿法具有计算量较少以及存储成本低的优点，所以本文拟将非精确牛顿法运用到求解三次正则化牛顿方法的子优化问题中，以此进一步降低计算量。

下面简要介绍一下一些非精确求解线性方程组的方法：

CG (共轭梯度法)：算法的核心思想是通过迭代来寻找最优解。在每个迭代步骤中，它寻找一个共轭方向(conjugate direction)，该方向与之前的方向是共轭的，然后在该方向上进行线性搜索以确定步长，从而更新当前的优化变量 x。

GMRES (Generalized Minimal Residual)方法：该方法的思想是利用 Krylov 子空间的性质，该子空间由 A 和初始残差 b-Ax$_0$ 生成，其中 x$_0$ 是初始猜测解向量。GMRES 迭代寻找一个近似解，它在 Krylov 子空间中最小化残差范数。

2.2. 更新迭代式

本文提出的非精确牛顿法运用在 Konstantin Mishchenko 所提出的正则化方法中的更新迭代式为：

$$\left(\nabla^2 f\left(x_k\right)+\lambda_k I\right) s_k = -\nabla f\left(x_k\right)+\delta_k,$$

$$x_{k+1} = x_k + s_{k+1}.$$

其中 $\lambda_k = \sqrt{H\left\|\nabla f\left(x_k\right)\right\|}$，$\delta_k$ 为非精确牛顿法计算的截断误差。

2.3. 算法

在 Konstantin Mishchenko 学者研究的基础上，根据本文提出的非精确更新迭代式，本文给出了在已知参数 H 情况下的算法 1。

算法 1：H 已知的非精确三次正则化牛顿法

1：Input: $x_0 \in R^d$，$H>0$，$\eta>0$

2：for $k=0,1,\cdots$ do

3：$\lambda_k = \sqrt{H\left\|\nabla f\left(x_k\right)\right\|}$

4：$s_k = \text{inexact\_solver}\left(\left(\nabla^2 f\left(x_k\right)+\lambda_k I\right)s = -\nabla f\left(x_k\right)\right)$

$\delta_k = \left(\nabla^2 f\left(x_k\right)+\lambda_k I\right)s_k = -\nabla f\left(x_k\right)$

satisfy $\left\|\delta_k\right\| \le \eta\left(\text{absolute error}\right)$ or $\left\|\delta_k\right\| \le \eta\left\|s_k\right\|\left(\text{relative error}\right)$

5：$x_k = x_{k-1} + s_k$

6：End for

算法 1 在实际运用中，由于需要提前知道先验参数 H，这对于现实情况是很困难的。所以需要对算法进行适当调整，使得算法能够自适应的寻找参数 H。受到 Konstantin Mishchenko 研究的启示，本文对算法 1 进行了调整，算法 2 通过线搜索的方式来寻找合适的参数 H。

算法 2：H 线搜索的非精确三次正则化牛顿法

1：Input: $x_0 \in R^d$， $H_0 > 0$， $\eta > 0$

2：for $k = 0,1,\cdots$ do

3：Initialize $H_k = H_{k-1}/4$

4：repeat

5：$H_k = 2H_{k-1}$， $\lambda_k = \sqrt{H_k \|\nabla f(x_k)\|}$

6：$s_k = \text{inexact\_solver}\left(\left(\nabla^2 f(x_k) + \lambda_k I\right)s = -\nabla f(x_k)\right)$

$\delta_k = \left(\nabla^2 f(x_k) + \lambda_k I\right)s_k = -\nabla f(x_k)$

satisfy $\|\delta_k\| \le \eta(\text{absolute error})$ or $\|\delta_k\| \le \eta\|s_k\|(\text{relative error})$

7：until $f(x_k + \delta_k) \le f(x_k)$

8：$x_k = x_{k-1} + s_k$

9：End for

由于线搜索方式过于消耗计算量，所以基于 Konstantin Mishchenko 的研究，给出算法 3，可以自适应的估计参数 H。

算法 3：H 自适应的非精确三次正则化牛顿法

1：Input: $x_0 \ne x_1 \in R^d$， $\eta > 0$

2：Initialize $H_0 = \dfrac{\left\|\nabla f(x_1) - \nabla f(x_0) - \nabla^2 f(x_0)(x_1 - x_0)\right\|}{\|x_1 - x_0\|^2}$

3：for $k = 0,1,\cdots$ do

4：$M_k = \dfrac{\left\|\nabla f(x_k) - \nabla f(x_{k-1}) - \nabla^2 f(x_{k-1})(x_k - x_{k-1})\right\|}{\|x_k - x_{k-1}\|^2}$

5：$H_k = \max\left\{M_k, \dfrac{H_{k-1}}{2}\right\}$， $\lambda_k = \sqrt{H_k \|\nabla f(x_k)\|}$

6：$s_k = \text{inexact\_solver}\left(\left(\nabla^2 f(x_k) + \lambda_k I\right)s = -\nabla f(x_k)\right)$

$\delta_k = \left(\nabla^2 f(x_k) + \lambda_k I\right)s_k = -\nabla f(x_k)$

satisfy $\|\delta_k\| \le \eta(\text{absolute error})$ or $\|\delta_k\| \le \eta\|s_k\|(\text{relative error})$

7：$x_k = x_{k-1} + s_k$

8：End for

3. 数值实验

在本节中，本文通过数值实验探究本文提出方法的数值计算表现。本节将主要探究在不同绝对误差、不同相对误差以及不同非精确求解器在数值实验中的表现。本文数据集将使用 mashrooms 数据集(8124条 × 22 维)。

带二次正则化的逻辑回归：

$$\min_{x\in R^d}\frac{1}{n}\sum_{i=1}^{n}\left(-b_i\log\left(\frac{1}{1+\mathrm{e}^{-a_i^{\mathrm{T}}x}}\right)-(1-b_i)\log\left(1-\frac{1}{1+\mathrm{e}^{a_i^{\mathrm{T}}x}}\right)\right)+\frac{l}{2}\|x\|^2,$$

$A=\left(a_{ij}\right)\in R^{n\times d}$ 是特征矩阵，$b_i\in\{0,1\}$ 是第 i 个样本的标签。本文设置正则化系数 $l=10^{-10}$ 使得问题是病态的。

从图 1 中可以看出，在算法 1 中不同绝对截断误差下的非精确牛顿法的数值实验结果相差不大。但在算法 2 中可以明显地发现非精确牛顿法可以在前期获得更快的收敛速度，在后期时由于绝对截断误差过大，导致其所能达到的最小值精度不够。在算法 3 中可以发现不同程度的绝对截断误差对于前期优化影响不大，但后期的收敛速度却比不上精确解。

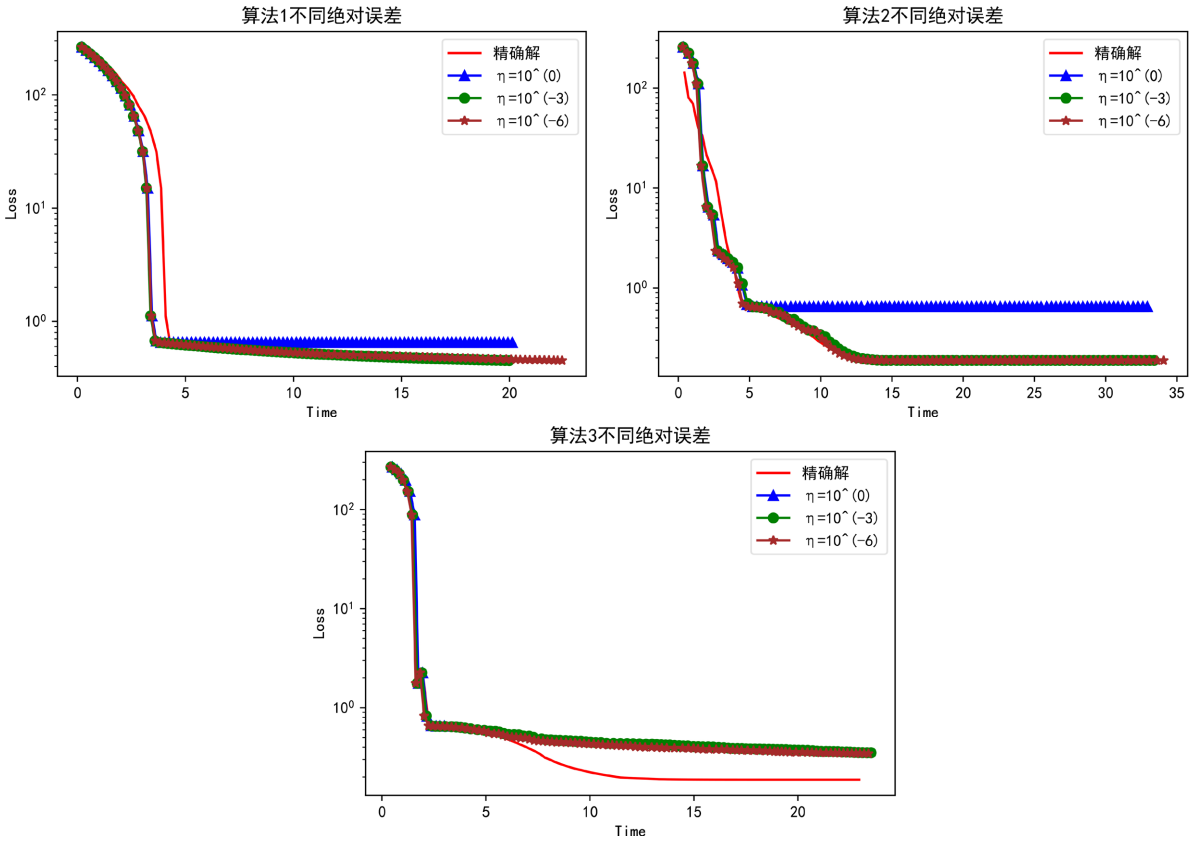


Figure 1. Numerical results for L2 regularized logistic regression problems on the “mushrooms” dataset using exact solution and three algorithms with absolute truncation errors ($\eta = 10^0, 10^{-3}, 10^{-6}$)

图 1. 对“mashrooms”数据集使用精确解和有绝对截断误差($\eta = 10^0, 10^{-3}, 10^{-6}$)的 3 个算法的 L2 正则化逻辑回归问题的数值结果

与图 1 类似，从图 2 中可以看出，在算法 1 中不同相对截断误差下的非精确牛顿法的数值实验结果相差不大。同样可以在算法 2 中发现有相对截断误差可以在前期获得更快的收敛速度，由于是相对截断

误差，所以在后期时也并未出现最小值精度无法达到的情况。同样不同程度的相对截断误差在算法 3 中前期优化影响不大，但后期可以发现相对误差越小其收敛速度越快。

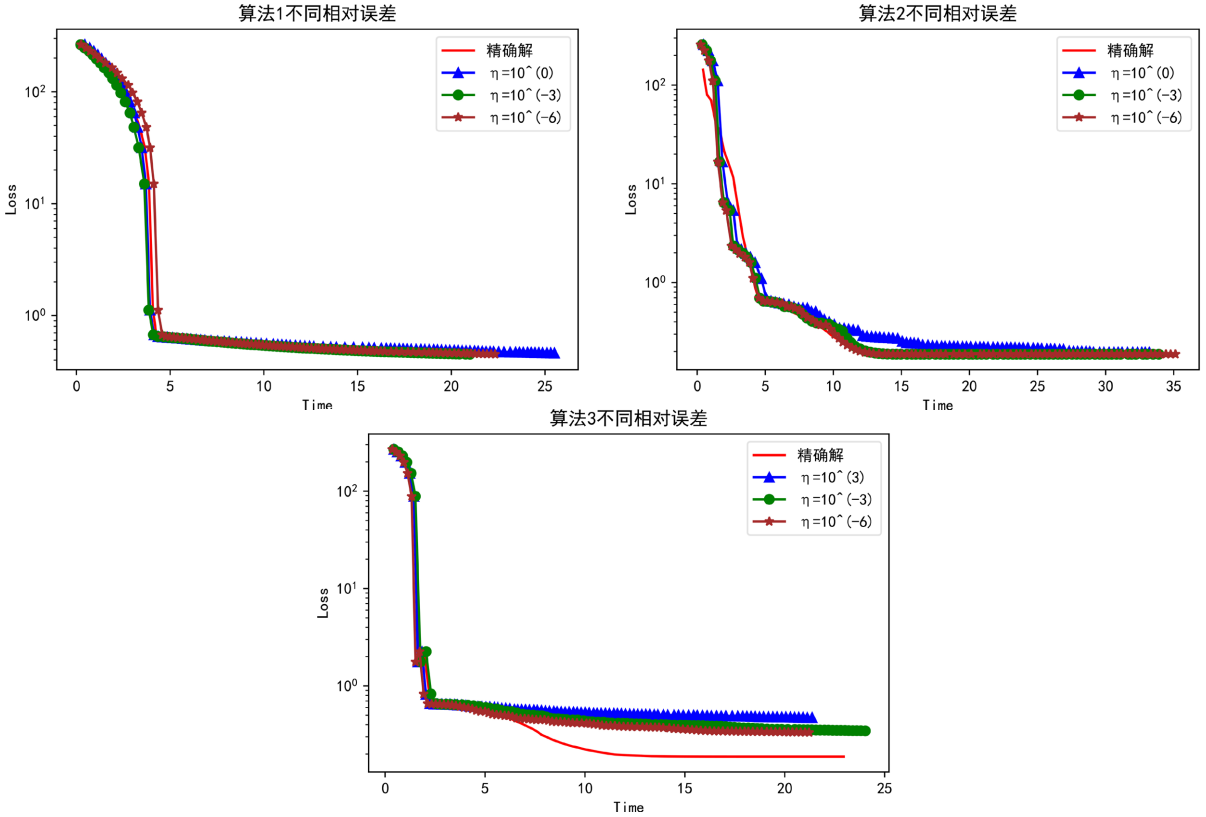
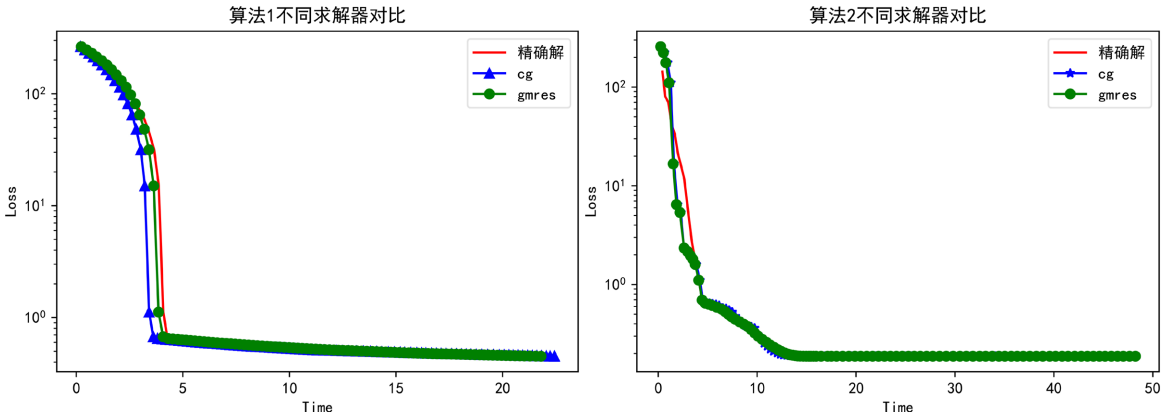


Figure 2. Numerical results for L2 regularized logistic regression problems on the "mushrooms" dataset using exact solution and three algorithms with relative truncation errors ($\eta = 10^0, 10^{-3}, 10^{-6}$)
图 2. 对"mashrooms"数据集使用精确解和有相对截断误差($\eta = 10^0, 10^{-3}, 10^{-6}$)的 3 个算法的 L2 正则化逻辑回归问题的数值结果

见图 3，其中非精确求解器为 cg (共轭梯度法)、gmres (Generalized Minimal Residual)。与图 1 类似，在算法 1 中同一绝对截断误差下使用不同求解器的非精确牛顿法的数值实验结果相差不大。同样可以在算法 2 中发现 cg 方法略优于 gmres 方法。在算法 3 中不同求解器在前期优化速度基本相同，但后期在收敛速度上 cg 求解器略逊于 gmres 方法。



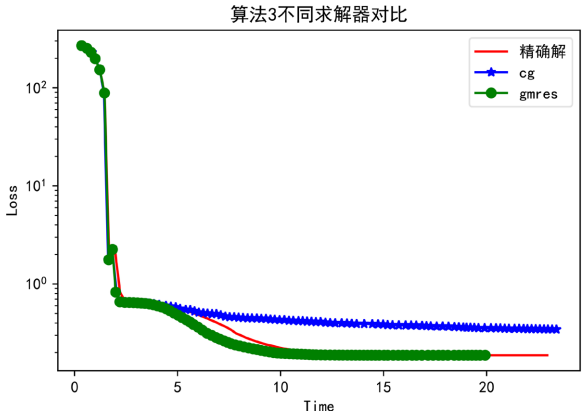


Figure 3. Numerical results for L2 regularized logistic regression problems on the "mushrooms" dataset using exact solution and two algorithms with absolute truncation errors ($\eta = 10^{-6}$)
图 3. 对"mashrooms"数据集使用精确解和有绝对截断误差($\eta = 10^{-6}$)的 2 个算法的 L2 正则化逻辑回归问题的数值结果

4. 结论与展望

本文基于 Konstantin Mishchenko 学者提出自适应牛顿方法给出了非精确自适应三次正则化牛顿方法。本文针对参数 H 已知和未知的情况提出了 3 种算法，并通过数值实验验证了这些算法的单调性和收敛性。本文对比了不同相对截断误差、不同绝对截断误差、不同求解器几种情况下算法的实际表现。通过这些对比实验得出，在绝对截断误差过大时，算法 2 会导致函数达到的最小值精度有限的情况，但随着绝对误差的减少函数的精度可以得到提升。相对截断误差过大时也会出现一些迭代步时效果不好的情况。此外，不同的非精确求解器在数值实验中在算法 1 上表现差异不大，但在算法 2 和算法 3 中有一定的差异。本文后续将从理论上给出其保持单调性的截断条件、证明收敛阶数并给出在高维数据下的数值实验表现。

参考文献

[1] Bottou, L., Curtis, F.E., and Nocedal, J. (2018) Optimization Methods for Large-Scale Machine Learning. *SIAM Review*, **60**, 223-311. https://doi.org/10.1137/16M1080173

[2] Gower, R.M., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E., and Richtárik, P. (2019) SGD: General Analysis and Improved Rates. *Proceedings of the* 36*th International Conference on Machine Learning*, Volume 97, Long Beach, 9-15 June 2019, 5200-5209.

[3] Mishchenko, K., Khaled, A. and Richtárik, P. (2020) Random Reshuffling: Simple Analysis with Vast Improvements. *Advances in Neural Information Processing Systems*, **33**, 17309-17320.

[4] Levenberg, K. (1944) A Method for the Solution of Certain Problems in Least Squares. *Quarterly of Applied Mathematics*, **2**,164-168. https://doi.org/10.1090/qam/10666

[5] Marquardt, D. (1963) An Algorithm for Least-Squares Estimation of Nonlinear Parameters. *SIAM Journal on Applied Mathematics*, **11**, 431-441. https://doi.org/10.1137/0111030

[6] Grippo, L., Lampariello, F. and Luclidi, S. (1986) A Nonmonotone Line Search Technique for Newton's Method. *SIAM Journal on Numerical Analysis*, **23**, 707-716. https://doi.org/10.1137/0723046

[7] Conn, A.B., Gould, N.I.M. and Toint, P.L. (2000) Trust Region Methods. SIAM, Philadelphia. https://doi.org/10.1137/1.9780898719857

[8] Carmon, Y. and Duchi, J.C. (2016) Gradient Descent Efficiently Finds the Cubic-Regularized Non-Convex Newton Step. arXiv: 1803.09357.

[9] Richtárik, P. and Doikov, N. (2018) Randomized Block Cubic Newton Method. arXiv Preprint: 1802.04084.

[10] Song, C. and Liu, J. (2019) Inexact Proximal Cubic Regularized Newton Methods for Convex Optimization. arXiv Preprint: 1902.02388.

[11] Xu, P., Roosta-Khorasani, F. and Mahoney, M.W. (2017) Newton-Type Methods for Non-Convex Optimization under Inexact Hessian Information. arXiv: 1708.07164.

[12] Kohler, J.M. and Lucchi, A. (2017) Sub-Sampled Cubic Regularization for Non-Convex Optimization. *Proceedings of the* 34*th International Conference on Machine Learning* (*ICML*), Volume 70, Naha, 16-18 April 2019, 1895-1904.

[13] Wang, Z., Zhou, Y., Liang, Y. and Lan, G. (2019) Sample Complexity of Stochastic Variance-Reduced Cubic Regularization for Nonconvex Optimization. *Proceedings of the* 22*nd International Conference on Artificial Intelligence and Statistics* (*AISTATS*), Volume 33, Sydney, 6-11 August 2017, 1440-1462.

[14] Mishchenko, K. (2023) Regularized Newton Method with Global O(1/k2) Convergence. arXiv: 2112.02089.

[15] Wright, S.J. (1999) Numerical Optimization. Springer Science & Business Media, Berlin.