第43卷 第4期 《新疆师范大学学报》(自然科学版) Vol.43,No.4
2024年12月 Journal of Xinjiang Normal University Dec. 2024
(Natural Sciences Edition)

一种非精确非光滑信赖域算法

李祉赟,王湘美*,马德乐
(贵州大学 数学与统计学院,贵州 贵阳 550025)

摘 要: Aravkin 等人提出了求解非光滑优化问题 $\min\limits_{x\in R^d} f(x)+h(x)$ 的非光滑信赖域算法(采用 f 的精确梯度),其中 f 是连续可微函数,h 是邻近有界且下半连续的真函数。文章研究当该问题中 $f:=\frac{1}{n}\sum\limits_{i=1}^{n} f_i$($n$ 很大且每个分量函数 f_i 是连续可微)时,求解这类大规模可分离非光滑优化问题的有效算法。结合非精确算法和非光滑信赖域算法的思想,提出了用非精确梯度代替精确梯度的非精确非光滑信赖域算法。与非光滑信赖域算法(采用精确梯度)相比,该算法降低了每次迭代的计算量。在一定的假设条件下,证明了算法的迭代复杂度。

关键词: 大规模可分离非光滑优化;非精确信赖域算法;邻近梯度算法
中图分类号: O224 **文献标识码:** A **文章编号:** 1008-9659(2024)04-0044-09

考虑欧氏空间 R^d 上的非光滑优化问题

$$\min_{x\in R^d} f(x)+h(x) \tag{1}$$

其中 $f:R^d\to R$ 是连续可微,$h:R^d\to R\cup\{+\infty\}$ 是下半连续真函数。文献[1-2]称问题(1)为一种复合优化问题,这类问题在机器学习和统计学中有广泛应用,例如组稀疏优化 GLASSO[3]、ℓ_1-逻辑回归[4]、BPDN 问题[5]等。特别地,问题(1)中 $h:=0$ 和 $f:=0$ 分别对应光滑和非光滑优化问题。关于光滑优化问题的研究有丰富的研究成果,其中优化数值算法可参见文献[6-8]。对于非光滑优化问题($f:=0$),Dennis 等人[9]在 h 为 Lipschitz 连续函数的条件下提出了求解问题(1)的信赖域算法,并研究了算法的收敛性。后来,Qi 等人[10]基于 Dennis 等人的研究,在 h 为局部 Lipschitz 连续函数且有有界水平集条件下,证明了信赖域算法的收敛性。

对于问题(1)的一般情形($h\neq 0$ 且 $f\neq 0$),在假设 f 和 h 是凸函数的条件下,Kim 等人[11]提出了非光滑信赖域算法,并证明了算法的迭代复杂度为 $O(\varepsilon^{-1})(0<\varepsilon<1)$. Catis 等人[12]考虑了 f 非凸,$h:=g(c(x))$ 的情形,其中 g 是局部 Lipschitz 连续凸函数,c 是连续可微可能非凸的函数。他们提出了求解该非光滑优化问题的信赖域算法及一种二次正则化变体,并建立了算法的收敛性。Lee 等人[2]讨论了在 f 和 h 均为凸函数的情形,采用精确 Hessian 和非精确 Hessian 邻近牛顿法的全局及局部收敛性。Bolte 等人[13]讨论了求解一类形如 $Q(x,y)+g(x)+h(y)$ 的非凸非光滑最优化问题的邻近交替算法(其核心思想是采用了非凸函数的邻近映射,详见定义2),其中 g 和 h 是下半连续真函数,Q 连续可微,并且基于 Kurdyka-Lojasiewicz 假设条件[14-15](简称 K-L 条件)对算法的收敛性进行深入分析。

求解问题(1)的另一类常用算法是邻近梯度法。当目标函数是凸函数时,文献[16-22]分别提出并研究了邻近梯度算法、单调邻近梯度算法及加速邻近梯度算法。后来 Li 等人[23]提出了求解非凸情形下单调及非

[收稿日期]2024-03-01 [修回日期]2024-04-11
[基金项目]国家自然科学基金项目(12161017);贵州省省级科技计划项目(ZK[2022]110)。
[作者简介]李祉赟(1999-),女,硕士研究生,主要从事优化理论方面研究,E-mail:Lzylllady@hotmail.com.
*[通讯作者]王湘美(1972-),女,教授,主要从事优化理论方面研究,E-mail:xmwang2@gzu.edu.cn.

单调的加速邻近梯度法,并证明了满足K-L条件时单调邻近梯度算法次线性收敛。当问题是病态的,为了使目标函数值不增加,单调算法的步长可能变短甚至呈锯齿状,从而收敛缓慢。相较于单调算法,非单调算法在使用线搜索时可以采用更大的步长,计算量更少且进一步加快收敛速度。此外,Themelis等人[24]提出了一种非单调线搜索邻近拟牛顿法,并在满足K-L条件时分析了该算法的收敛性。

文献[25]结合信赖域算法和邻近梯度算法的思想提出了求解问题(1)的非光滑信赖域算法,在一定条件下证明了非光滑信赖域算法的迭代复杂度为(ε^{-2}). 以上算法都是求解问题(1)的有效算法,然而它们都没有考虑问题(1)中f的可分离性。受文献[25]启发,文章考虑以下大规模可分离非光滑优化问题

$$\min_{x\in R^d} f(x)+h(x)=\frac{1}{n}\sum_{i=1}^{n}f_i(x)+h(x) \tag{2}$$

其中$f=\frac{1}{n}\sum_{i=1}^{n}f_i$,每个分量函数$f_i:R^d\to R$是连续可微的,$h:R^d\to R\cup\{+\infty\}$是下半连续真函数。当问题的规模很大时(即$n\gg 1$),每次迭代要求出每个分量函数$f_i(i=1,2,\cdots,n)$的梯度,导致计算成本太高。因此,文章考虑用非精确梯度代替函数f的全梯度,提出求解问题(2)的非精确非光滑信赖域算法,希望在减少计算量的同时使得算法仍然具有良好的收敛性质。在一定条件下,证明了算法的迭代复杂度为$O(\varepsilon^{-1})$.

1 预备知识

首先介绍文章使用的符号以及运用的基本概念和结论[6,13,25,26]。

用R表示实数集合,N表示自然数集合,并记$\bar{R}=R\cup\{+\infty\}$. 分别用$\|\cdot\|$和$\langle\cdot,\cdot\rangle$表示欧氏空间$R^d$上的$\ell_2$范数和内积。设$\Delta>0$,$B(0,\Delta)$(简记为$\Delta B$)是以0点为中心,$\Delta$为半径的一个球,定义如下

$$\Delta B=B(0,\Delta):=\{x\in R^d:\|x\|\leqslant\Delta\}$$

记$B=B(0,1)$. 设$A\subseteq R^d$为非空子集,$x\in R^d$,x到A的欧氏距离$\mathrm{dist}(x;A)$定义如下

$$\mathrm{dist}(x;A):=\inf_{a\in A}\|a-x\|$$

用$\chi(\cdot;A)$表示集合A的示性函数[14],即

$$\chi(x;A):=\begin{cases}0, & x\in A\\ +\infty, & x\notin A\end{cases}$$

特别地,球ΔB的示性函数记为$\chi(\cdot;\Delta)$. 函数$f:R^d\to\bar{R}$的定义域为

$$\mathrm{dom}f:=\{x\in R^d|f(x)<+\infty\}$$

若对所有的x有$f(x)>-\infty$且至少存在一个x使得$f(x)<+\infty$,则称f为真函数。设f为真函数,如果对任意x,$y\in\mathrm{dom}f$,$0\leqslant\lambda\leqslant 1$有

$$f(\lambda x+(1-\lambda)y)\leqslant\lambda f(x)+(1-\lambda)f(y)$$

则称f为凸函数。如果存在常数$c>0$,使得对任意$x,y\in\mathrm{dom}f$,$0\leqslant\lambda\leqslant 1$有

$$f(\lambda x+(1-\lambda)y)\leqslant\lambda f(x)+(1-\lambda)f(y)-\frac{1}{2}c\lambda(1-\lambda)\|x-y\|^2$$

则称f为常数,为c的强凸函数。设$\bar{x}\in R^d$,若在点$\bar{x}$处有$\liminf_{x\to\bar{x}}f(x)=f(\bar{x})$(也就是对每个$\varepsilon>0$均存在$\bar{x}$的开邻域$U(\bar{x})$,使得对于任意$x\in U$有$f(x)>f(\bar{x})-\varepsilon$),则称$f$是下半连续的。设$\alpha\in R$,函数$f$的$\alpha$-水平集定义如下

$$L_\alpha(h):=\{x\in R^d|f(x)\leqslant\alpha\}$$

若函数f的所有水平集均有界,则称f是水平有界的。若f是下半连续真函数且水平有界,则$\arg\min_{x\in R^d}f(x)$是非空紧集。若函数f在点x处可微,下文用$\nabla f(x)$表示f在点x的梯度。

定义1(次梯度和次微分) 设$\phi:R^d\to\bar{R}$,$\bar{x}\in R^d$,$\bar{x}\in\mathrm{dom}\phi$,对任意$v\in R^d$

(1)如果满足 $\liminf\limits_{x\to\bar{x},x\neq\bar{x}}\dfrac{\phi(x)-\phi(\bar{x})-v^T(x-\bar{x})}{\|x-\bar{x}\|}\geqslant 0$,则称$v$是$\phi$在点$\bar{x}$处的一个正则次梯度,$\hat{\partial}\phi(\bar{x})$是$\phi$在点$\bar{x}$处所有正则次梯度集合(即$v\in\hat{\partial}\phi(\bar{x})$),称为Fréchet次微分;

(2)如果存在序列$\{x_k\}$和$\{v_k\}$使得$x_k\to\bar{x}$,$\phi(x_k)\to\phi(\bar{x})$,$v_k\in\hat{\partial}\phi(x_k)$且$v_k\to v$,则称$v$是$\phi$在点$\bar{x}$处的一个一般次梯度,$\partial\phi(\bar{x})$是$\phi$在点$\bar{x}$处所有一般次梯度集合(即$v\in\partial\phi(\bar{x})$),称为极限次微分。

若ϕ为凸函数,则Fréchet次微分和极限次微分与凸分析中的次微分一致。若ϕ在点$\bar{x}$处可微,则$\hat{\partial}\phi(\bar{x})=\{\nabla\phi(\bar{x})\}$;若$\phi$在点$\bar{x}$处连续可微,则$\partial\phi(\bar{x})=\{\nabla\phi(\bar{x})\}$[26]. 此外,当$\hat{\partial}\phi(\bar{x})$为闭凸集且$\partial\phi(\bar{x})$为闭集时,有$\hat{\partial}\phi(\bar{x})\subset\partial\phi(\bar{x})$[26].

下文中通常不使用精确的次梯度定义,而是使用以下性质。

命题1[26] 设$\phi:R^d\to R$是真函数,且有局部最优解$\bar{x}$,则$0\in\hat{\partial}\phi(\bar{x})\subseteq\partial\phi(\bar{x})$. 若$\phi$为凸函数,则局部最优解也为全局最优解。若$\phi=f+h$,其中$f$在$\bar{x}$的一个邻域$U(\bar{x})$上连续可微且$h(\bar{x})$为有限值,则$\partial\phi(\bar{x})=\nabla f(\bar{x})+\partial h(\bar{x})$.

定义2(Moreau包络和邻近算子) 设$x\in R^d$,$h:R^d\to\bar{R}$是下半连续真函数,参数$\nu>0$,函数h的Moreau包络$e_\nu h$和邻近算子$p_\nu h$定义如下

$$e_\nu h(x):=\inf_\omega\frac{1}{2\nu}\|\omega-x\|^2+h(\omega)=\nu^{-1}\inf_\omega\left(\frac{1}{2}\|\omega-x\|^2+\nu h(\omega)\right)$$

$$p_\nu h(x):=\arg\min_\omega\frac{1}{2\nu}\|\omega-x\|^2+h(\omega)=\arg\min_\omega\frac{1}{2}\|\omega-x\|^2+\nu h(\omega)$$

在一定的假设下,若$p_\nu h(x)$是强凸函数,则它是单集。通常情况下,集合$p_\nu h$可能非空或者含有多个元素。若h是下半连续真函数,以下定义给出了当h的Moreau包络$e_\nu h$为有限值时参数ν的取值范围。

定义3(邻近有界) 设$h:R^d\to\bar{R}$是下半连续真函数。若存在参数$\nu>0$,至少有一个$x\in R^d$使得$e_\nu h(x)>-\infty$,则称函数h是邻近有界的,并且将所有满足以上条件的参数$\nu>0$的上确界ν_h称为函数h的邻近有界阈值。

进一步地,若h是水平有界的,则$\inf h$是有限值。因此,对于所有的$x\in R^d$和$\nu>0$有$e_\nu h(x)>-\infty$,即函数h是邻近有界的。

命题2[26] 设$h:R^d\to\bar{R}$是下半连续真函数,h是邻近有界的且具有邻近有界阈值$\nu_h>0$. 对任意$x\in R^d$和任意$\nu\in(0,\nu_h)$满足以下性质:

(1)$p_\nu h(x)$是非空紧集;

(2)$e_\nu h(x)$在(ν,x)上连续,即随着ν单调减小趋于0时$e_\nu h(x)$单调增加趋于$h(x)$.

下节中的非光滑非精确信赖域算法需要用到邻近梯度算法[16,17,26],下面回顾该算法及其性质。

考虑一般非光滑优化问题

$$\min_s\varphi(s)+\psi(s)\tag{3}$$

其中$\varphi:R^d\to R$是连续可微函数,$\psi:R^d\to\bar{R}$是下半连续邻近有界真函数。邻近梯度法的具体迭代形式如下:给定初始迭代点$s_0\in\mathrm{dom}\psi$,对任意$j\in N$有

$$s_{j+1}\in p_\nu\psi\left(s_j-\nu\nabla\varphi\left(s_j\right)\right):=\arg\min_\omega\frac{1}{2\nu}\left\|\omega-s_j+\nu\nabla\varphi\left(s_j\right)\right\|^2+\psi(\omega)\tag{4}$$

其中$\nu>0$为步长,$\nabla\varphi\left(s_j\right)$是函数$\varphi$在点$s_j$处的梯度。该算法产生的迭代点列式(4)的一阶最优性条件如下

$$0\in s_{j+1}-s_j+\nu\nabla\varphi\left(s_j\right)+\nu\partial\psi\left(s_{j+1}\right)\tag{5}$$

其中$\partial\psi\left(s_{j+1}\right)$是函数$\psi$在点$s_{j+1}$处的Fréchet次微分。

以下命题表明由邻近梯度法产生的点列$\{s_j\}$使得$\{(\varphi+\psi)(s_j)\}$是单调下降序列。

命题3[13] 设$\varphi:R^d \to R$连续可微,$\nabla\varphi$为L-Lipschitz连续函数,$\psi:R^d \to \bar{R}$是邻近有界下半连续真函数。又设邻近梯度法的步长$0 < \nu < 1/L$,给定初始点$s_0 \in \mathrm{dom}\psi$,则迭代点列式(4)满足

$$(\varphi+\psi)(s_{j+1}) \leqslant (\varphi+\psi)(s_j) - \frac{1}{2}(\nu^{-1}-L)\|s_{j+1}-s_j\|^2, \quad j \in N \tag{6}$$

当$\psi = 0$时,式(3)为光滑优化问题,可得$s_1 = -\nu\nabla\varphi(s_0)$,则式(6)变为

$$\varphi(s_1) \leqslant \varphi(s_0) - \frac{1}{2}(\nu^{-1}-L)\|s_1-s_0\|^2$$

2 非精确梯度非光滑信赖域算法及其收敛性分析

文献[25]提出了非光滑优化问题的信赖域算法(未考虑问题的可分离性)。受文献[25]启发,文章考虑大规模可分离非光滑优化问题(2)

$$\min_{x\in R^d} f(x)+h(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x) + h(x)$$

其中$f = \frac{1}{n}\sum_{i=1}^{n} f_i$,每个分量函数$f_i:R^d \to R$连续可微,$n \gg 1$($n$很大),$h:R^d \to \bar{R}$是邻近有界下半连续真函数。为求解该问题,提出了非精确梯度非光滑信赖域算法(即算法1),该算法在每次迭代的子问题中,用函数φ和ψ分别近似函数f和h. 具体地,设$x_k \in R^d (k \in N)$,求解第k次迭代子问题

$$\min_{s\in R^d, \|s\|\leqslant\Delta_k} m_k(s,x_k) := \varphi(s,x_k) + \psi(s,x_k) + \chi(s,\Delta_k) \approx f(x_k+s) + h(x_k+s) \tag{7}$$

其中$\Delta_k > 0$为信赖域半径。本研究总做以下假设:

假设1(函数假设) 对任意$x \in R^d$,

(1)$\varphi(\cdot,x)$是连续可微函数,满足$\varphi(0,x) = f(x)$,$\nabla_s\varphi(0,x) \approx \nabla f(x)$且$\nabla_s\varphi(0,x)$是Lipschitz连续的,并且其Lipschitz常数$L > 0$;

(2)$\psi(\cdot,x)$是下半连续真函数,$\psi(0,x) = h(x)$且$\partial\psi(0,x) = \partial h(x)$.

注1 文献[25]中算法总假设$\nabla_s\varphi(0,x) = \nabla f(x)$. 考虑到函数$f$的可分离性,为减少算法的计算量,用$f$的近似梯度代替全梯度,即$\nabla_s\varphi(0,x) \approx \nabla f(x)$(与文献[25]中算法的不同之处)。

下面先分析子问题的一些性质。设$x \in R^d$,$\Delta > 0$. 一般地,和文献[25]类似,记

$$p(\Delta,x) := \min_s \varphi(s,x) + \psi(s,x) + \chi(s,\Delta) \tag{8-a}$$

$$P(\Delta,x) := \arg\min_s \varphi(s,x) + \psi(s,x) + \chi(s,\Delta) \tag{8-b}$$

命题4[25] 令$p(0,x) := \varphi(\cdot,x) + \psi(\cdot,x)$,$P(0,x) := \{0\}$,则$p(\cdot,x)$和$P(\cdot,x)$的定义域为$\{\Delta \mid \Delta \geqslant 0\}$,且有

(1)对每个$\Delta \geqslant 0$,$p(\cdot,x)$是下半连续真函数,$P(\cdot,x)$是非空紧集;

(2)若$\varphi(\cdot,x) + \psi(\cdot,x)$是严格凸函数,则$P(\Delta,x)$是单点集。

特别地,先考虑$\varphi(\cdot,x) = \varphi^{\nu}(\cdot,x)$为二次函数的情形,对任意$s \in R^d$定义$\varphi^{\nu}(\cdot,x)$如下

$$\varphi^{\nu}(s,x) = f(x) + g(x)^T s + \frac{1}{2\nu}\|s\|^2 = \frac{1}{2\nu}\|s+\nu g(x)\|^2 + f(x) - \frac{1}{2}\nu\|g(x)\|^2 \tag{9}$$

其中$\nu > 0$是正则化参数,$g(x)$是$\nabla f(x)$的近似,即$g(x) \approx \nabla f(x)$. 记

$$\begin{aligned} p(\Delta,x,\nu) &:= e_{\nu}(\psi(\cdot,x)+\chi(\cdot,\Delta))(-\nu g(x)) + f(x) - \frac{1}{2}\nu\|g(x)\|^2 \\ &= \min_s f(x) + g(x)^T s + \frac{1}{2\nu}\|s\|^2 + \psi(s,x) + \chi(s,\Delta) \end{aligned} \tag{10-a}$$

$$P(\Delta,x,\nu) := p_{\nu}(\psi(\cdot,x)+\chi(\cdot,\Delta))(-\nu g(x)) = \arg\min_s f(x) + g(x)^T s + \frac{1}{2\nu}\|s\|^2 + \psi(s,x) + \chi(s,\Delta) \tag{10-b}$$

其中$p(\Delta,x,\nu)$和$e_{\nu}(\psi(\cdot,x)+\chi(\cdot,\Delta))(-\nu g(x))$只相差一个常数。特别地,当式(10-a)及(10-b)中的$g(x) = \nabla f(x)$时,把$p(\Delta,x,\nu)$和$P(\Delta,x,\nu)$分别记为$\bar{p}(\Delta,x,\nu)$和$\bar{P}(\Delta,x,\nu)$. 进一步,记

$$\xi(\Delta,x,\nu):=f(x)+h(x)-p(\Delta,x,\nu) \tag{11-a}$$

由$p(\Delta,x,\nu)$的定义,$p(\Delta,x,\nu)\leqslant\varphi(0,x)+\psi(0,x)+\chi(0,\Delta)=f(x)+h(x)$,故$\xi(\Delta,x,\nu)\geqslant 0$. 特别地,当$g(x)=\nabla f(x)$时,记

$$\bar{\xi}(\Delta,x,\nu):=f(x)+h(x)-\bar{p}(\Delta,x,\nu) \tag{11-b}$$

下面给出求解式(2)的非精确梯度非光滑信赖域算法,并结合以上理论对其收敛性进行分析。

算法1 欧氏空间上的非精确梯度非光滑信赖域算法。

步骤1 输入初始点$x_0\in\mathrm{dom}h$,初始半径$\Delta_0>0$,其他参数$\Delta_{\min}>0$,$0<\eta_1\leqslant\eta_2<1$,$0<\gamma_1\leqslant\gamma_2<1<\gamma_3\leqslant\gamma_4$,$\alpha>0$,$\beta\geqslant 1$,$\varepsilon>0$,置$k=0$;

步骤2 取$\nu_k\in(0,\dfrac{1}{L+\alpha^{-1}\Delta_k^{-1}}]$,选择子问题$m_k(s,x_k):=\varphi(s,x_k)+\psi(s,x_k)$满足假设1;

步骤3 若$\xi(\Delta_k,x_k,\nu_k)<\varepsilon$,停止并输出$x_k$;

步骤4 近似求解式(7)得s_k满足$\|s_k\|\leqslant\min\left(\Delta_k,\beta\|s_{k,1}\|\right)$,其中$s_{k,1}$是式(7)中$\varphi(\cdot,x)=\varphi^{\nu_k}(\cdot,x)$的解;

步骤5 计算$\rho_k:=\dfrac{f(x_k)+h(x_k)-\left(f(x_k+s_k)+h(x_k+s_k)\right)}{m_k(0,x_k)-m_k(s,x_k)}$;

步骤6 令$x_{k+1}=\begin{cases}x_k+s_k & \text{如果}\quad \rho_k\geqslant\eta_1\\ x_k & \text{否则}\end{cases}$

步骤7 令$\Delta_{k+1}\in\begin{cases}[\gamma_3\Delta_k,\gamma_4\Delta_k], & \text{如果}\quad \rho_k\geqslant\eta_2\text{(非常成功迭代)}\\ [\gamma_2\Delta_k,\Delta_k], & \text{如果}\quad \eta_1\leqslant\rho_k<\eta_2\text{(一般成功迭代)}\\ [\gamma_1\Delta_k,\gamma_2\Delta_k], & \text{如果}\quad \rho_k<\eta_1\text{(失败迭代)}\end{cases}$

步骤8 令$k=k+1$,转步骤3.

引理1 对任意$k\in N$,有$(\varphi+\psi)(s_{k,1},x_k)\leqslant(f+h)(x_k)-\dfrac{1}{2}\left(\nu^{-1}-L(x_k)\right)\|s_{k,1}\|^2$,其中$L(x_k)>0$为$\nabla_s\varphi(0,x_k)$的Lipschitz常数。

证明 设$k\in N$,由定义有$s_{k,1}\in\arg\min\limits_s f(x_k)+g(x_k)^Ts+\dfrac{1}{2\nu}\|s\|^2+\psi(s,x_k)+\chi(s,\Delta_k)$.

设$s_0=0$,则由式(6)得

$$(\varphi+\psi)(s_{k,1},x_k)\leqslant(\varphi+\psi)(s_0,x_k)-\frac{1}{2}\left(\nu^{-1}-L(x_k)\right)\|s_{k,1}-s_0\|^2=(f+h)(x_k)-\frac{1}{2}\left(\nu^{-1}-L(x_k)\right)\|s_{k,1}\|^2$$

其中等号由假设1得到。

为保证算法1的收敛性,还需要以下假设。

假设2 存在常数$k_m>0$,$k_{mdc}\in(0,1)$,使得对任意$k\in N$有

$$\left|f(x_k+s_k)+h(x_k+s_k)-m_k(s_k,x_k)\right|\leqslant k_m\|s_k\|^2 \tag{12-a}$$

$$m_k(0,x_k)-m_k(s_k,x_k)\geqslant k_{mdc}\xi(\Delta_k,x_k,\nu_k) \tag{12-b}$$

注2 (1) 当函数f和φ二阶可导,且二阶导数均有界,即存在$M>0$使得

$$\|\nabla^2 f(x)\|\leqslant M,\quad \|\nabla^2\varphi(0,x)\|\leqslant M,\quad \forall x\in R^d \tag{13}$$

同时取$\psi(\cdot,x_k)=h(x_k+\cdot)$,且存在常数$C>0$使

$$\|\nabla f(x_k)-g(x_k)\|\leqslant C\|s_k\| \tag{14}$$

则不等式(12-a)成立。事实上,由$\psi(s,x_k)=h(x_k+s)$和m_k的定义有

$$\left|f(x_k+s_k)+h(x_k+s_k)-m_k(s_k,x_k)\right|=\left|f(x_k+s_k)-\varphi(s_k,x_k)\right|$$

又由微分中值定理,存在$\bar{x}_k,\tilde{x}_k$使得

$$f(x_k+s_k)=f(x_k)+\nabla f(x_k)^T s_k+\frac{1}{2}s_k{}^T\nabla^2 f(\bar{x}_k)s_k$$

$$\varphi(s_k,x_k)=\varphi(0,x_k)+\nabla\varphi(0,x_k)^T s_k+\frac{1}{2}s_k{}^T\nabla^2\varphi(0,\tilde{x}_k)s_k$$

由上面两个等式得

$$\begin{aligned}\left|f(x_k+s_k)+h(x_k+s_k)-m_k(s_k,x_k)\right|&=\left|f(x_k+s_k)-\varphi(s_k,x_k)\right|\\&\leqslant\left\|\nabla f(x_k)-g(x_k)\right\|\|s_k\|+M\|s_k\|^2\leqslant(C+M)\|s_k\|^2\end{aligned}$$

其中第一个不等式由式(13)及假设1中(1)得到,第二个不等式由式(14)得到。故不等式(12-a)成立,其中$k_m=C+M$.

(2)当精确求解子问题式(7)时,不等式(12-b)自然成立,其中$k_{mdc}=1$.

以下引理表明,在一定假设条件下$\|s_k\|$有正的下界。这样,当$\left\|\nabla f(x_k)-g(x_k)\right\|\leqslant\frac{C\varepsilon}{M_k}$时,式(14)成立。

引理2 设$\varphi(s_k,x_k)$如式(9)中所定义,$\psi(s_k,x_k)=h(x_k+s_k)$,$M_k>0$为$\left\|g(x_k)\right\|$和$\left\|\partial h(x_k)\right\|$在信赖域$\Delta_k$B上的界,则当$\xi(\Delta_k,x_k,\nu_k)\geqslant\varepsilon$时,有$\|s_k\|\geqslant\frac{\varepsilon}{M_k}$.

证明 由ξ和p的定义及$\xi(\Delta_k,x_k,\nu_k)\geqslant\varepsilon$,得

$$\xi(\Delta_k,x_k,\nu_k)=f(x_k)+h(x_k)-p(\Delta_k,x_k,\nu_k)=-g(x_k)^T s_k-\frac{1}{2\nu_k}\|s_k\|^2+h(x_k)-h(x_k+s_k)\geqslant\varepsilon$$

当$h(x_k)-h(x_k+s_k)\geqslant\varepsilon$时,由微分中值定理,存在$\bar{x}_k\in(x_k,x_k+s_k)$,使$h(x_k)-h(x_k+s_k)=\zeta_k^T s_k$,其中$\zeta_k\in\partial h(\bar{x}_k)$,则有

$$\varepsilon\leqslant h(x_k)-h(x_k+s_k)\leqslant\|\zeta_k\|\|s_k\|\leqslant M_k\|s_k\|\Rightarrow\|s_k\|\geqslant\frac{\varepsilon}{M_k}$$

当$-g(x_k)^T s_k-\frac{1}{2\nu_k}\|s_k\|^2\geqslant\varepsilon$时,可得$\frac{1}{2\nu_k}\|s_k\|^2-\left\|g(x_k)\right\|\|s_k\|+\varepsilon\leqslant 0$,把这个不等式看作一个关于$\|s_k\|$的一元二次方程并求解,则有

$$\frac{\left\|g(x_k)\right\|-\sqrt{\left\|g(x_k)\right\|^2-2\varepsilon(\nu_k)^{-1}}}{(\nu_k)^{-1}}\leqslant\|s_k\|\leqslant\frac{\left\|g(x_k)\right\|+\sqrt{\left\|g(x_k)\right\|^2-2\varepsilon(\nu_k)^{-1}}}{(\nu_k)^{-1}}$$

故有

$$\|s_k\|\geqslant\nu_k\left[\left\|g(x_k)\right\|-\sqrt{\left\|g(x_k)\right\|^2-2\varepsilon(\nu_k)^{-1}}\right]\geqslant\frac{2\varepsilon}{\left\|g(x_k)\right\|+\sqrt{\left\|g(x_k)\right\|^2-2\varepsilon(\nu_k)^{-1}}}\geqslant\frac{\varepsilon}{M_k}$$

下面分析算法的收敛性,为此记常数

$$\Delta_{succ}:=\frac{k_{mdc}(1-\eta_2)}{2k_m\alpha\beta^2}>0\text{及}\Delta_{\min}=\min\{\Delta_0,\gamma_1\Delta_{succ}\}>0$$

其中$k_m>0$,$k_{mdc}\in(0,1)$为假设2中的参数,$\alpha>0$,$\beta\geqslant 1$,$\eta_2\in(0,1)$为算法1的步骤1中给定的参数。

引理3 设算法1第k次迭代的信赖域半径为Δ_k,则以下结论成立:

(1)如果$\Delta_k\leqslant\Delta_{succ}$,则第$k$次迭代非常成功,从而要么$\xi(\Delta_{k+1},x_k,\nu_k)<\varepsilon$,算法在$k+1$次次迭代停止,要么有$\Delta_{k+1}>\Delta_k$;

(2)$\Delta_k\geqslant\Delta_{\min}$,从而$\xi(\Delta_k,x_k,\nu_k)\geqslant\xi(\Delta_{\min},x_k,\nu_k)$.

证明 引理3中讨论(1)的证明同文献[25]。结论(2)由结论(1)算法的定义以及ξ的定义得到。

注意文献[25]表明,对任意$\Delta>0,\nu>0$,有$\bar{\xi}(\Delta,x,\nu)=0\Leftrightarrow 0\in\bar{P}(\Delta,x,\nu)\Rightarrow x$是式(2)的一阶稳定点,其中$\bar{\xi}(\Delta,x,\nu)$如式(11-b)所定义。采用文献[25]中的方式,称$\bar{\xi}(\Delta_{\min},x,\nu)$为最优度量。下面的引理表明,可以用$\xi(\Delta_{\min},x,\nu)$来估计最优度量$\bar{\xi}(\Delta_{\min},x,\nu)$,这也是算法1中步骤3停机准则的选择依据。

引理4 设$\varepsilon>0,\Delta\geqslant\Delta_{\min},\left\|g(x)-\nabla f(x)\right\|\leqslant\dfrac{\varepsilon}{\Delta}$,则

$$\xi(\Delta,x,\nu)<\varepsilon\Rightarrow\bar{\xi}(\Delta_{\min},x,\nu)<2\varepsilon$$

证明 由$\left\|g(x)-\nabla f(x)\right\|\leqslant\dfrac{\varepsilon}{\Delta}$有$\left|\left(g(x)-\nabla f(x)\right)^{T}s\right|\leqslant\left\|g(x)-\nabla f(x)\right\|\left\|s\right\|\leqslant\varepsilon,\forall s\in\Delta\mathrm{B}$,于是

$$f(x)+g(x)^{T}s-\varepsilon+\frac{1}{2\nu}\left\|s\right\|^{2}+\psi(s,x)+\chi(s,\Delta)\leqslant f(x)+\nabla f(x)^{T}s+\frac{1}{2\nu}\left\|s\right\|^{2}+\psi(s,x)+\chi(s,\Delta)$$

$$\leqslant f(x)+g(x)^{T}s+\varepsilon+\frac{1}{2\nu}\left\|s\right\|^{2}+\psi(s,x)+\chi(s,\Delta)$$

对上式取“min”,再结合式(10-a)得$\left|p(\Delta,x,\nu)-\bar{p}(\Delta,x,\nu)\right|\leqslant\varepsilon$,从而有

$$\left|\xi(\Delta,x,\nu)-\bar{\xi}(\Delta,x,\nu)\right|=\left|p(\Delta,x,\nu)-\bar{p}(\Delta,x,\nu)\right|\leqslant\varepsilon$$

所以$\xi(\Delta,x,\nu)<\varepsilon\Rightarrow\bar{\xi}(\Delta,x,\nu)<2\varepsilon$,又由引理3中的结论(2),引理4的结论成立。

下面分析算法的迭代复杂度,为此设算法1迭代$k(\varepsilon)$次后满足终止条件

$$\xi(\Delta_k,x_k,\nu_k)<\varepsilon \tag{15}$$

并令

$$S:=\left\{k\in\mathbb{N}\middle|\rho_k\geqslant\eta_1\right\}$$

$$S(\varepsilon):=\left\{k\in S\middle|k<k(\varepsilon)\right\}$$

$$U(\varepsilon):=\left\{k\in\mathbb{N}\middle|k\notin S,\quad k<k(\varepsilon)\right\}$$

其中$S,S(\varepsilon),U(\varepsilon)$分别代表算法1产生的所有成功迭代的集合,算法1并未达到式(15)之前产生的所有成功迭代的集合以及算法1并未达到式(15)之前产生的所有失败迭代的集合。假设函数$f+h$有下界,记为$\left(f+h\right)_{low}$.

引理5 设算法1中参数$\nu_k\geqslant\nu_{\min}>0$,则有

$$\left|S(\varepsilon)\right|\leqslant\frac{\left(f+h\right)(x_0)-\left(f+h\right)_{low}}{\eta_1 k_{mdc}\varepsilon}=\mathrm{O}\left(\varepsilon^{-1}\right) \tag{16}$$

$$\left|U(\varepsilon)\right|\leqslant\log\gamma_2\left(\frac{\Delta_{\min}}{\Delta_0}\right)+\left|S(\varepsilon)\right|\left|\log\gamma_2(\gamma_4)\right|=\mathrm{O}\left(\varepsilon^{-1}\right) \tag{17}$$

证明 设$k\in S(\varepsilon)$,则$\rho_k\geqslant\eta_1$. 结合ρ_k的定义有

$$f(x_k)+h(x_k)-f(x_{k+1})-h(x_{k+1})=f(x_k)+h(x_k)-f(x_k+s_k)-h(x_k+s_k)$$

$$\geqslant\eta_1(m_k(0,x_k)-m_k(s_k,x_k))\geqslant\eta_1 k_{mdc}\xi(\Delta_k,x_k,\nu_k)\geqslant\eta_1 k_{mdc}\varepsilon$$

对上述不等式中所有的$k\in S(\varepsilon)$求和得

$$\left(f+h\right)(x_0)-\left(f+h\right)_{low}\geqslant\sum_{k\in S(\varepsilon)}\left(f+h\right)(x_k)-\left(f+h\right)(x_{k+1})\geqslant\left|S(\varepsilon)\right|\eta_1 k_{mdc}\varepsilon$$

即式(16)成立。

由算法1的定义,每次失败迭代信赖域半径至少减少至当前半径的γ_2倍,即$\Delta_{k+1}\leqslant\gamma_2\Delta_k$;每次成功迭代信赖域半径总小于当前半径的$\gamma_4$倍,即$\Delta_{k+1}\leqslant\gamma_4\Delta_k$. 因为算法在第$k(\varepsilon)$次迭代第一次满足条件式(15),结合引理3中的结论(1),对前$k(\varepsilon)-1$次迭代有

$$\Delta_{\min}\leqslant\Delta_{k(\varepsilon)-1}\leqslant\Delta_0\gamma_2^{|U(\varepsilon)|}\gamma_4^{|S(\varepsilon)|}$$

对上式两边取对数(注意$0<\gamma_2<1$)得式(17)。

最后,在以下定理中给出算法1的迭代复杂度。

定理1 设引理5中条件均满足,则算法1有限步终止且迭代次数N有以下估计

$$N=\left|S(\varepsilon)\right|+\left|U(\varepsilon)\right|=\mathrm{O}\left(\varepsilon^{-1}\right)$$

参考文献:

[1] CATIS C,GOULD N,TOINT P L. On the Evaluation Complexity of Composite Function Minimization with Applications to Nonconvex Nonlinear Programming[J]. SIAM Journal on Optimization,2011,21(04):1721-1739.

[2] LEE J D,SUN Y,SAUNDERS M A. Proximal Newton-type Methods for Minimizing Composite Functions[J]. SIAM Journal on Optimization,2014,24(03):1420-1443.

[3] YUAN M,LIN Y. Model Selection and Estimation in Regression with Grouped Variables[J]. Journal of the Royal Statistical Society Series B:Statistical Methodology,2006,68(01):49-67.

[4] KOH K,KIM S J,BOYD S. An Interior-point Method for Large-scale L1-regularized Logistic Regression[J]. Journal of Machine Learning Research,2007,(08):1519-1555.

[5] DONOHO D L. Compressed Sensing[J]. IEEE Transactions on Information Theory,2006,52(04):1289-1306.

[6] CONN A R,GOULD N,TOINT P L. Trust Region Methods[M]. Philadelphia:Society for Industrial and Applied Mathematics,2000.

[7] GRAPIGLIA G N,YUAN J,YUAN Y. Nonlinear Stepsize Control Algorithms:Complexity Bounds for First and Second-order Optimality[J]. Journal of Optimization Theory and Applications,2016,171:980-997.

[8] TOINT P L. Nonlinear stepsize Control,Trust Regions and Regularizations for Unconstrained Optimization[J]. Optimization Methods and Software,2013,28(01):82-95.

[9] DENNIS J E,LI S B,TAPIA R A. A Unified Approach to Global Convergence of Trust Region Methods for Nonsmooth Optimization[J]. Mathematical Programming,1995,68(01-03):319-346.

[10] QI L,SUN J. A Trust Region Algorithm for Minimization of Locally Lipschitzian Functions[J]. Mathematical Programming,1994,66(01):25-43.

[11] KIM D,SRA S,DHILLON I. A Scalable Trust Region Algorithm with Application to Mixed-norm Regression[C]. Proceedings of the 27th International Conference on Machine Learning (ICML 2010):Stroudsburg,2010:519-526.

[12] CATIS C,GOULD N,TOINT P L. On the Evaluation Complexity of Composite Function Minimization with Applications to Nonconvex Nonlinear Programming[J]. SIAM Journal on Optimization,2011,21(04):1721-1739.

[13] BOLTE J,SABACH S,TEBOULLE M. Proximal Alternating Linearized Minimization for Nonconvex and Nonsmooth Problems[J]. Mathematical Programming,2014,146(01-02):459-494.

[14] KURDYKA K. On Gradients of Functions Definable In O-minimal Structures[C]. Annales de l'institut Fourier. Grenoble,1998,48(03):769-783.

[15] LOJASIEWICZ S. Une Proprié Té Topologique Des Sous-ensembles Analytiques RéEls[J]. Les Équations Aux Dérivées Partielles,1963,117:87-89.

[16] BAUSCHKE H H,COMBETTES P L. Convex Analysis and Monotone Operator Theory in Hilbert Spaces[M]. Switzerland:Springer,2011.

[17] LIONS P L,MERCIER B. Splitting Algorithms for the Sum of Two Nonlinear Operators[J]. SIAM Journal on Numerical Analysis,1979,16(06):964-979.

[18] NESTEROV Y. Smooth Minimization of Non-smooth Functions[J]. Mathematical Programming,2005,103:127-152.

[19] NESTEROV Y. Gradient Methods Minimizing Functions[J]. Mathematical Programming,2013,140(01):125-161.

[20] BECK A,TEBOULLE M. Fast Gradient-based Algorithms for Constrained Total Variation Image Denoising and Deblurring Problems[J]. IEEE Transactions on Image Processing,2009,18(11):2419-2434.

[21] BECK A,TEBOULLE M. A Fast Iterative Shrinkage-thresholding Algorithm for Linear Inverse Problems[J]. SIAM Journal on Imaging Sciences,2009,2(01):183-202.

[22] TSENG P. On Accelerated Proximal Gradient Methods for Convex-concave Optimization[J]. SIAM Journal on Optimization,2008,2(03):581-589.

[23] LI H,LIN Z. Accelerated Proximal Gradient Methods for Nonconvex Programming[J]. Advances in Neural Information Processing Systems,2015,(28):379-387.

[24] THEMELIS A,STELLA L,PATRINOS P. Forward-backward Envelope for the Sum of Two Nonconvex Functions:Further Properties and Nonmonotone Linesearch Algorithms[J]. SIAM Journal on Optimization,2018,28(03):2274-2303.

[25] ARAVKIN A Y,BARALDI R,ORBAN D. A Proximal Quasi-newton Trust-region Method for Nonsmooth Regularized Optimization[J].

SIAM Journal on Optimization,2022,32(02):900-929.
[26] ROCKAFELLAR R T,WETS R J B. Variational Analysis[M]. Berlin:Springer,1998.

An Inexact Trust Region Algorithm for Nonsmooth Optimization

LI Zhi-yun,WANG Xiang-mei*,MA De-le
(*College of Mathematics and Statistics,Guizhou University,Guiyang,Guizhou,550025,China*)

Abstract: Aravkin et al proposed the trust region algorithm (employing exact gradients of f) for solving the nonsmooth optimization problem $\min\limits_{x\in R^d} f(x)+h(x)$, where f is a continuously differentiable function and h is a lower semicontinuous and prox-bounded proper function. In the case when $f:=\frac{1}{n}\sum\limits_{i=1}^{n}f_i$ (n is quite big, and each component f_i is continuously differentiable), the efficient algorithm for solving such kind of large-scale separable nonsmooth optimization problem is studied. Combining the concepts of the inexact algorithm and the above trust-region algorithm, it is proposed that the inexact trust-region algorithm replaces the exact gradients with the inexact gradients for solving this nonsmooth problem. Comparing with the trust-region algorithm (employing the exact gradients of f), the new algorithm can reduce the computational cost at each iteration. Under certain assumptions, the iteration complexity of this algorithm is established.

Keywords: Large-scale separable nonsmooth optimization; Inexact trust-region algorithm; Proximal gradient method