

一类非凸-强凹极小极大问题的零阶优化算法^{*}

高瑞成, 谢 涛, 李觉友

(重庆师范大学 数学科学学院, 重庆 401331)

摘要:极小极大问题是博弈论和机器学习中的一类重要问题。目前已有大量基于目标函数的梯度和 Hessian 阵信息的优化算法来求解这类问题。但在有些应用中,目标函数的梯度或 Hessian 阵信息往往是计算昂贵或难以获取的。为此,针对一类非凸-强凹极小极大问题,在极小极大三次正则化牛顿算法的框架下,通过基于 Stein 恒等式的高斯平滑化方法来近似梯度与 Hessian 阵信息,进而提出一类零阶极小极大三次正则化牛顿算法。分析算法的收敛性,并得到算法达到一个二阶平稳点时的迭代复杂度为 $O(\epsilon^{-3/2})$,其中 ϵ 是算法终止所达到的精度。数值仿真实验结果表明:在相同的精度下,所提出的算法在 CPU 运行时间上优于极小极大三次正则化牛顿算法。

关键词:非凸-凹极小极大问题;三次正则化牛顿算法;零阶算法;复杂度分析

中图分类号:O221.2

文献标志码:A

文章编号:1672-6693(2024)02-0016-10

1 背景介绍

极小极大优化在博弈论和机器学习中扮演着重要的角色,如生成对抗网络(GANs)^[1]、多代理强化学习^[2]等。目前针对凸-凹极小极大问题的算法和理论已经有大量研究成果。但对非凸-凹极小极大优化问题的算法研究还不多见,这主要是由于问题的非凸性对算法设计带来诸多挑战^[3]。对于求解非凸优化问题,为了逃离鞍点,Nesterov 等人^[4]在传统牛顿算法基础上,通过引入一个三次正则项设计出了一类三次正则化牛顿算法,并得到该算法的全局收敛性。基于三次正则化牛顿法,Luo 等人^[5]设计了一类极小极大三次正则化牛顿算法(min-max cubic regularized Newton method,MCN 算法)来求解一类非凸-强凹极小极大问题,并得到算法达到二阶平稳点的迭代复杂度为 $O(\epsilon^{-3/2})$, ϵ 是算法终止所达到的精度。但 MCN 算法需要在每一步迭代时计算精确的梯度与 Hessian 阵信息,而在有些问题中的目标函数的梯度和 Hessian 阵信息往往计算昂贵或难以获取,例如机器学习中的黑盒对抗攻击问题^[2]。

近些年来,零阶优化算法在信号处理和机器学习中受到了广泛关注。零阶算法能处理函数梯度和 Hessian 阵信息难以获取或精确计算时代价过大的情况。基于目标函数的函数值信息,Xu 等人^[6]在 2021 年提出了一类零阶交替梯度投影算法来求解一类非凸-凹优化问题,并得到算法的迭代复杂度与对应的梯度算法同阶。但上述方法仅使用了函数的近似梯度信息,收敛速度较慢。基于 Stein 恒等式,借助高斯平滑化方法用函数值信息来近似函数的梯度和 Hessian 阵,文献[7]提出了一种零阶的三次正则化牛顿法来求解一类非凸优化问题,该算法的好处在于只需计算函数值,不需要精确计算梯度与 Hessian 阵。如何利用函数值信息来近似目标函数的 Hessian 阵,进而设计一类零阶的牛顿算法来求解极小极大优化问题是非常有意义的。

受文献[5,7]的启发,针对一类非凸-强凹优化问题中梯度和 Hessian 阵信息计算昂贵或难以获取的情况,基于 Stein 恒等式的高斯平滑化方案^[7],本文提出了一类零阶极小极大三次正则化牛顿算法(zeroth-order min-max cubic regularized Newton method,ZO-MCN 算法),分析了算法的收敛性,并给出了算法的迭代复杂度为 $O(\epsilon^{-3/2})$ 。数值仿真验证了算法的有效性。与文献[5]的算法相比,本文的算法不需要计算目标函数的精确梯度和 Hessian 阵,仅通过函数值信息利用 Stein 恒等式来近似目标函数的梯度和 Hessian 阵。

* 收稿日期:2022-10-17 修回日期:2023-12-29 网络出版时间:2023-06-25T11:13

资助项目:重庆市自然科学基金(No. cstc2020jcyj-msxmX0287)

第一作者简介:高瑞成,男,研究方向极小极大理论与算法,E-mail:gao18523940728@163.com;通信作者:李觉友,男,教授,博士,E-mail:lijueyou@cqnu.edu.cn

网络出版地址:https://link.cnki.net/urlid/50.1165.N.20230623.1116.008

2 问题与算法

2.1 问题描述

考虑如下一类极小极大优化问题:

$$\min_{x \in \mathbf{R}^{d_x}} \{Q(x) \triangleq \max_{y \in \mathbf{R}^{d_y}} f(x, y)\}, \quad (1)$$

其中: $f(x, y)$ 是一个二次连续可微函数, 且对任意固定的 y 关于 x 是非凸的, 对任意固定的 x 关于 y 是 μ -强凹的。非凸-凹极小极大问题是机器学习领域的主要研究热点之一^[3]。下面给出关于问题(1)的一些假设。

假设 1 1) 函数 $f(x, y)$ 的梯度 $\nabla f(x, y)$ 关于 (x, y) 是 l -Lipschitz 连续的, 即存在一个常数 $l > 0$, 使得对于任意的 $x, x' \in \mathbf{R}^{d_x}$ 和任意的 $y, y' \in \mathbf{R}^{d_y}$ 有 $\|\nabla f(x, y) - \nabla f(x', y')\|^2 \leq l^2 (\|x - x'\|^2 + \|y - y'\|^2)$ 。另外假设存在一个常数 $B > 0$, 使得关于 x 的梯度 $\nabla_x f(x, y)$ 满足 $\|\nabla_x f(x, y)\| \leq B$ 。

2) 对任意给定的 $y \in \mathbf{R}^{d_y}$, $f(x, y)$ 关于 x 是二次可微的。并且关于 x 的 Hessian 阵 $\nabla_{xx}^2 f(x, y)$ 是 γ -Lipschitz 连续的, 即对任意 $x, x' \in \mathbf{R}^{d_x}$, 存在一个常数 $\gamma > 0$, 有 $\|\nabla_{xx}^2 f(x, y) - \nabla_{xx}^2 f(x', y)\| \leq \gamma \|x - x'\|$ 。

3) 对任意给定的 $x \in \mathbf{R}^{d_x}$, $f(x, y)$ 关于 y 是 μ -强凹的, 即对任意 $y, y' \in \mathbf{R}^{d_y}$, 存在一个常数 $\mu > 0$, 有

$$f(x, y) \leq f(x, y') + \nabla_y f(x, y)^\top (y - y') - \frac{\mu}{2} \|y - y'\|^2.$$

假设 2 函数 $Q(x) \triangleq \max_{y \in \mathbf{R}^{d_y}} f(x, y)$ 有下界, 即 $Q^* = \inf_{x \in \mathbf{R}^{d_x}} Q(x) > -\infty$ 。

2.2 梯度估计和 Hessian 阵估计

对于某些极小极大优化问题^[2], 梯度和 Hessian 阵信息往往是不容易得到的, 甚至不可用。为此, 本文借助 Nesterov 的高斯平滑化方法^[8], 用函数值信息近似估计函数的梯度和 Hessian 阵。考虑 f 的平滑化函数:

$$f_v(x, y) = \frac{1}{(2\pi)^{d_x/2}} \int f(x + v\mathbf{u}, y) e^{-\|\mathbf{u}\|^2/2} d\mathbf{u} = E[f(x + v\mathbf{u}, y)],$$

这里 $\mathbf{u} \sim N(0, I_{d_x})$ 为标准正态分布的随机向量, $v > 0$ 为平滑化参数, E 表示数学期望。 $f_v(x, y)$ 的梯度估计定义为:

$$G_v(x, y; \mathbf{u}) = \frac{f(x + v\mathbf{u}, y) - f(x, y)}{v} \mathbf{u}, \mathbf{u} \sim N(0, I_{d_x}). \quad (2)$$

$f_v(x, y)$ 的 Hessian 阵估计定义为:

$$H_v(x, y; \mathbf{u}) = \frac{1}{2v^2} (\mathbf{u}\mathbf{u}^\top - I_{d_x}) [f(x + v\mathbf{u}, y) + f(x - v\mathbf{u}, y) - 2f(x, y)], \mathbf{u} \sim N(0, I_{d_x}). \quad (3)$$

在假设 1 的条件下, 根据一阶和二阶 Stein 恒等式^[7]有:

$$E[G_v(x, y; \mathbf{u})] = \nabla_x f_v(x, y), E[H_v(x, y; \mathbf{u})] = \nabla_{xx}^2 f_v(x, y),$$

即 $G_v(x, y; \mathbf{u})$ 是 $\nabla_x f_v(x, y)$ 的一个无偏估计, $H_v(x, y; \mathbf{u})$ 是 $\nabla_{xx}^2 f_v(x, y)$ 的一个无偏估计。由此, 仅需利用计算较为简单的函数值信息即可估计函数的梯度和 Hessian 阵。下面的 2 个引理给出了 $Q(x)$ 的一些性质。

引理 1^[9] 若 f 满足假设 1 中的 1) 和 3), 那么有:

1) $Q(x)$ 的梯度 $\nabla Q(x) = \nabla_x f(x, y^*(x))$ 是 $(\zeta + 1)l$ -Lipschitz 连续的, 这里 $\zeta = l/\mu$;

2) $y^*(x) = \arg \max_{y \in \mathbf{R}^{d_y}} f(x, y)$ 是唯一, 且是 ζ -Lipschitz 连续的。

引理 2^[5] 若 f 满足假设 1, 那么有:

1) $Q(x)$ 在点 $y^*(x)$ 处有

$$\nabla^2 Q(x) = \nabla_{xx}^2 f(x, y) - \nabla_{xy}^2 f(x, y) [\nabla_{yy}^2 f(x, y)]^{-1} \nabla_{yx}^2 f(x, y);$$

2) $\nabla^2 Q(x)$ 是 ω -Lipschitz 连续的, 其中 $\omega = 4\sqrt{2}\zeta^3\gamma$, 即对任意的 $x, x' \in \mathbf{R}^{d_x}$, 有:

$$\|\nabla^2 Q(x) - \nabla^2 Q(x')\| \leq \omega \|x - x'\|.$$

在后面的证明中, 本文将用到平滑函数 f_v 的一些性质。

引理 3^[8] 1) 设 $\mathbf{u} \in N(0, I_{d_x})$, 则对任意的 $t \geq 2$ 有 $E[\|\mathbf{u}\|^t] = \frac{1}{(2\pi)^{d_x/2}} \int \|\mathbf{u}\|^t e^{-\|\mathbf{u}\|^2/2} d\mathbf{u} \leq (d_x + t)^{t/2}$

成立。

2) 如果假设 1 成立, 则函数 f_v 关于 x 的梯度 $\nabla_x f_v$ 是 l_{f_v} -Lipschitz 连续的, 即存在一个常数 $l_{f_v} > 0$, 使得对任意的 $x, x' \in \mathbf{R}^{d_x}$, 有 $\|\nabla_x f_v(x, y) - \nabla_x f_v(x', y)\| \leq l_{f_v} \|x - x'\|$ 成立, 其中 $l_{f_v} \leq l$ 。

下面 2 个引理表明 $H_v(x, y; \mathbf{u})$ 和 $\nabla_{xx}^2 f_v(x, y) - \nabla_{xx}^2 f(x, y)$ 的范数有上界。

引理 4^[7] 若假设 1 满足, $H_v(x, y; \mathbf{u})$ 由式(3)给出, 则有:

$$E[\|H_v(x, y; \mathbf{u})\|_F^4] \leq (d_x + 16)^8 \left(\frac{\gamma^4 (d_x + 16)^2 v^4}{9} + l^4 \right). \quad (4)$$

与文献[7]中引理 6 的证明相似, 可以得到 $\|\nabla_{xx}^2 f_v(x, y) - \nabla_{xx}^2 f(x, y)\|$ 的一个上界。

引理 5 若假设 1 满足, $H_v(x, y; \mathbf{u})$ 由式(3)给出, 则有:

$$\|\nabla_{xx}^2 f_v(x, y) - \nabla_{xx}^2 f(x, y)\| \leq \frac{\gamma v (d_x + 6)^{5/2}}{4}. \quad (5)$$

2.3 算法

定义 1^[5] 对任意给定的 $\epsilon > 0$ 和 $\omega > 0$, 如果点 $x^* \in \mathbf{R}^{d_x}$ 满足 $\|\nabla Q(x^*)\| \leq \epsilon$, 则称点 x^* 是 $Q(x)$ 的一个 ϵ 一阶平稳点。若同时还满足 $\nabla^2 Q(x^*) \geq -\sqrt{\omega\epsilon} I$, 则称 x^* 是 $Q(x)$ 的一个 $(\epsilon, \sqrt{\omega\epsilon})$ 二阶平稳点。

基于 MCN 算法^[5], 借助有关梯度和 Hessian 阵的近似估计, 本文设计一类 ZO-MCN 算法来寻找问题(1)中 $Q(x)$ 的一个二阶平稳点。在此之前, 先给加速梯度算法(accelerated gradient descent algorithm, AGD)。

算法 1 $y_t = \text{AGD}(x_{t-1}, y_{t-1}, M_t, \alpha, \beta)$ 的伪代码如下。

步骤 1, 初始化 $\hat{y}_0 = y_{t-1}$ 和 $\tilde{y}_0 = y_{t-1}$, 设置 $\alpha = \frac{1}{l}$ 和 $\beta = \frac{\sqrt{\xi} - 1}{\sqrt{\xi} + 1}$ 。

步骤 2, 对于 $m = 0, 1, \dots, M_t$, 按照如下规则更新:

$$\begin{aligned} \hat{y}_{m+1} &= \tilde{y}_m + \alpha \nabla_y f(x_{t-1}, \tilde{y}_m), \\ \tilde{y}_{m+1} &= \hat{y}_{m+1} + \beta(\hat{y}_{m+1} - \hat{y}_m). \end{aligned}$$

步骤 3, 输出 $y_t = \tilde{y}_{M_t+1}$ 。

算法 2 ZO-MCN 算法的伪代码如下。

步骤 1, 初始化 $x_0 \in \mathbf{R}^{d_x}$ 和 $y_0 \in \mathbf{R}^{d_y}$ 。对给定的 $\epsilon > 0$, 选取合适的 $T \geq 1$, 平滑化参数 $v > 0$, 非负序列 $\{r_t\}_{t=1}^T$ 满足 $r_t \geq \omega$, 正整数序列 $\{a_t\}_t^T$ 和 $\{b_t\}_{t=1}^T$ 以及 $\{M_t\}_{t=1}^T$ 。

步骤 2, 对 $t = 1, \dots, T$,

i) 更新 $y_t: y_t = \text{AGD}(x_{t-1}, y_{t-1}, M_t, \alpha, \beta)$ 。

ii) 随机产生 $u_{t,j}^G \sim N(0, I_{d_x})$ 和 $u_{t,i}^H \sim N(0, I_{d_x})$, 计算 $\bar{G}_v^t$ 和 $\bar{H}_v^t$:

$$\bar{G}_v^t = \frac{1}{a_t} \sum_{j=1}^{a_t} \frac{f(x_{t-1} + v u_{t,j}^G, y_t) - f(x_{t-1}, y_t)}{v} u_{t,j}^G, \quad (6)$$

$$\bar{H}_v^t = \frac{1}{b_t} \sum_{i=1}^{b_t} \frac{f(x_{t-1} + v u_{t,i}^H, y_t) + f(x_{t-1} - v u_{t,i}^H, y_t) - 2f(x_{t-1}, y_t)}{2v^2} (u_{t,i}^H (u_{t,i}^H)^T - I_{d_x}). \quad (7)$$

更新 $x_t: x_t = \arg \min_{x \in \mathbf{R}^{d_x}} \hat{f}(x, x_{t-1}, \bar{G}_v^t, \bar{H}_v^t, r_t)$, 其中

$$\hat{f}(x, x_{t-1}, \bar{G}_v^t, \bar{H}_v^t, r_t) = \langle \bar{G}_v^t, x - x_{t-1} \rangle + \frac{1}{2} \langle \bar{H}_v^t (x - x_{t-1}), x - x_{t-1} \rangle + \frac{r_t}{6} \|x - x_{t-1}\|^3.$$

如果 $\|x_t - x_{t-1}\| < \frac{1}{2} \sqrt{\epsilon/r_t}$, 则算法终止。

步骤 3, 输出 (x_t, y_t) 。

注 1 1) ZO-MCN 算法是先固定变量 $x = x_{t-1}$, 利用 AGD 算法更新 y_t ; 然后固定 $y = y_t$ 计算近似的梯度和 Hessian 阵, 并利用三次正则化牛顿算法更新 x_t 。

2) 对比 MCN 算法^[5], 本文借助函数值信息来近似计算函数的梯度和 Hessian 阵, 好处是能处理梯度信息

和 Hessian 阵信息难以获取的情况。

2.4 收敛性分析

为了证明算法的收敛性,先根据文献[4]的结果给出算法 1 中的三次正则化子问题的最优性条件。

引理 6 令 $x_t = \arg \min_{x \in \mathbf{R}^{d_x}} \hat{f}(x, x_{t-1}, \bar{G}_v^t, \bar{H}_v^t, r_t)$, 则有:

$$\bar{G}_v^t + \bar{H}_v^t(x_t - x_{t-1}) + \frac{r_t}{2} \|x_t - x_{t-1}\| (x_t - x_{t-1}) = 0,$$

$$\bar{H}_v^t + \frac{r_t}{2} \|x_t - x_{t-1}\| I_{d_x} \geq 0,$$

$$(\bar{G}_v^t)^\top (x_t - x_{t-1}) + \frac{1}{2} (x_t - x_{t-1})^\top \bar{H}_v^t (x_t - x_{t-1}) + \frac{r_t}{6} \|x_t - x_{t-1}\|^3 \leq -\frac{r_t}{12} \|x_t - x_{t-1}\|^3.$$

基于上述最优性条件,下面分别给出梯度和 Hessian 阵估计量方差的一个上界。

引理 7^[7] 在假设 1 条件下, $\bar{G}_v^t$ 由算法 1 给出。有:

$$E[\|\bar{G}_v^t - \nabla_x f(x_{t-1}, y^*(x_{t-1}))\|^2] \leq \frac{4(d_x + 5)B^2}{a_t} + \frac{3v^2}{2} l^2 (d_x + 3)^3 := (\eta_t^g)^2. \quad (8)$$

引理 8 在假设 1 条件下,如果 b_t 满足 $b_t \geq 4(1 + 2\log(2d_x))$, $\bar{H}_v^t$ 由算法 1 给出,则有:

$$\begin{aligned} E[\|\bar{H}_v^t - \nabla_{xx}^2 f(x_{t-1}, y^*(x_{t-1}))\|^2] &\leq \\ \frac{64(1 + 2\log(2d_x))(d_x + 16)^4 l^2}{b_t} + 21\gamma^2 (d_x + 16)^5 v^2 &:= 2(\gamma + r_t) \eta_t^H, \end{aligned} \quad (9)$$

$$E[\|\bar{H}_v^t - \nabla_{xx}^2 f_v(x_{t-1}, y^*(x_{t-1}))\|^3] \leq \frac{90\sqrt{1 + 2\log(2d_x)}(d_x + 16)^6 l^2}{b_t^{3/2}} + 20\gamma^4 (d_x + 16)^8 v^2. \quad (10)$$

证明 首先证明式(9)成立。根据文献[10]中定理 1,有:

$$\begin{aligned} E[\|\bar{H}_v^t - \nabla_{xx}^2 f_v(x_{t-1}, y^*(x_{t-1}))\|^2] &\leq \\ \frac{8(1 + 2\log(2d_x))}{b_t^2} (E[\sum_{i=1}^{b_t} E[\Delta_{t,i}^2]] + 4(1 + 2\log(2d_x))E[\max_i \|\Delta_{t,i}\|^2]) \end{aligned} \quad (11)$$

其中: $\Delta_{t,i} = H_v(x_{t-1}, y^*(x_{t-1}); u_{t,i}^H) - \nabla_{xx}^2 f_v(x_{t-1}, y^*(x_{t-1}))$ 。由引理 4 可得:

$$E[\|H_v(x_{t-1}, y^*(x_{t-1}); u_{t,i}^H)\|^2] \leq (d_x + 16)^4 \left(\frac{\gamma^2 (d_x + 16)^2 v^2}{3} + l^2 \right), \quad (12)$$

从而有:

$$E[\|\Delta_{t,i}\|^2] \leq 4E[\|H_v(x_{t-1}, y^*(x_{t-1}); u_{t,i}^H)\|^2] \leq \frac{4(d_x + 16)^4 (\gamma^2 (d_x + 16) v^2 + 3l^2)}{3}, \quad (13)$$

其中:第 1 个不等式由 $(m - n)^2 \leq 2m^2 + 2n^2$ 以及 $E[H_v(x, y; u)] = \nabla_{xx}^2 f_v(x, y)$ 可得,第 2 个不等式由式(12)得到。注意到 $b_t \geq 4(1 + 2\log(2d_x))$, 并由 Jensen 不等式

$$\left\| \sum_{i=1}^{b_t} E[\Delta_{t,i}^2] \right\| \leq \sum_{i=1}^{b_t} \|E[\Delta_{t,i}^2]\| \leq \sum_{i=1}^{b_t} E[\|\Delta_{t,i}\|^2],$$

以及式(11)和(13)可得

$$E[\|\bar{H}_v^t - \nabla_{xx}^2 f_v(x_{t-1}, y^*(x_{t-1}))\|^2] \leq \frac{64(1 + 2\log(2d_x))(d_x + 16)^4}{3b_t} (\gamma^2 (d_x + 16) v^2 + 3l^2), \quad (14)$$

结合式(14)和(5)即证得式(9)。

接下来证式(10)成立。通过 Hölder 不等式可得:

$$\begin{aligned} E[\|\bar{H}_v^t - \nabla_{xx}^2 f_v(x_{t-1}, y^*(x_{t-1}))\|^3] &\leq \\ E[\|\bar{H}_v^t - \nabla_{xx}^2 f_v(x_{t-1}, y^*(x_{t-1}))\| \cdot \|\bar{H}_v^t - \nabla_{xx}^2 f_v(x_{t-1}, y^*(x_{t-1}))\|_F^2] &\leq \\ \sqrt{E[\|\bar{H}_v^t - \nabla_{xx}^2 f_v(x_{t-1}, y^*(x_{t-1}))\|^2] \cdot E[\|\bar{H}_v^t - \nabla_{xx}^2 f_v(x_{t-1}, y^*(x_{t-1}))\|_F^4]}. \end{aligned} \quad (15)$$

由 Rosenthal 不等式^[11]以及式(4)可得:

$$E[\|\bar{H}_v^t - \nabla_{xx}^2 f_v(x_{t-1}, y^*(x_{t-1}))\|_F^4] \leq \frac{18E[\|\Delta_{t,i}\|_F^4]}{b_i^2} \leq \frac{18(d_x + 16)^8(\gamma^4(d_x + 16)^2 v^2 + 9l^4)}{b_i^2},$$

由上述不等式,并结合式(14)和(15)可得式(10).

证毕

为了证明主要结果,下面将分别给出 $\|x_t - x_{t-1}\|$ 的一个下界和上界估计.

引理 9 如果假设 1 成立,序列 $\{(x_t, y_t)\}$ 由算法 1 产生.有:

$$\sqrt{E[\|x_t - x_{t-1}\|^2]} \geq \max \left\{ \sqrt{\frac{E[\|\nabla_x f(x_t, y^*(x_{t-1}))\|] - \eta_t^g - \eta_t^H}{\gamma + r_t}}, \frac{-2}{2\gamma + r_t} [E[\lambda_{\min}(\nabla_{xx}^2 f(x_t, y^*(x_{t-1})))] + \sqrt{2(\gamma + r_t)\eta_t^H}] \right\}.$$

证明 由文献[4]中引理 1 可知:

$$\|\nabla_x f(x_t, y^*(x_{t-1})) - \nabla_x f(x_{t-1}, y^*(x_{t-1})) - \nabla_{xx}^2 f(x_{t-1}, y^*(x_{t-1}))(x_t - x_{t-1})\| \leq \frac{\gamma}{2} \|x_t - x_{t-1}\|^2. \quad (16)$$

由三角不等式和引理 6 有:

$$\begin{aligned} & \|\nabla_x f(x_t, y^*(x_{t-1}))\| \leq \\ & \|\nabla_x f(x_t, y^*(x_{t-1})) - \nabla_x f(x_{t-1}, y^*(x_{t-1})) - \nabla_{xx}^2 f(x_{t-1}, y^*(x_{t-1}))(x_t - x_{t-1})\| + \\ & \|\nabla_x f(x_{t-1}, y^*(x_{t-1})) - \bar{G}_v^t\| + \|\nabla_{xx}^2 f(x_{t-1}, y^*(x_{t-1})) - \bar{H}_v^t\| \|x_t - x_{t-1}\| + \frac{r_t}{2} \|x_t - x_{t-1}\|^2 \leq \\ & \frac{\gamma + r_t}{2} \|x_t - x_{t-1}\|^2 + \|\nabla_x f(x_{t-1}, y^*(x_{t-1})) - \bar{G}_v^t\| + \|\nabla_{xx}^2 f(x_{t-1}, y^*(x_{t-1})) - \bar{H}_v^t\| \|x_t - x_{t-1}\| \leq \\ & (\gamma + r_t) \|x_t - x_{t-1}\|^2 + \|\nabla_x f(x_{t-1}, y^*(x_{t-1})) - \bar{G}_v^t\| + \frac{\|\nabla_{xx}^2 f(x_{t-1}, y^*(x_{t-1})) - \bar{H}_v^t\|^2}{2(\gamma + r_t)}. \end{aligned}$$

其中第 2 个不等式由式(16)可得,第 3 个不等式由 Hölder 不等式可得.对上述不等式两边同时取期望,并注意到式(8)和(9),可得:

$$\frac{(E[\|\nabla_x f(x_t, y^*(x_{t-1}))\|] - \eta_t^g - \eta_t^H)}{\gamma + r_t} \leq E[\|x_t - x_{t-1}\|^2]. \quad (17)$$

由假设 1 可得:

$$\begin{aligned} \nabla_{xx}^2 f(x_t, y^*(x_{t-1})) & \geq \nabla_{xx}^2 f(x_{t-1}, y^*(x_{t-1})) - \gamma \|x_t - x_{t-1}\| I_{d_x} \geq \\ \nabla_{xx}^2 f(x_{t-1}, y^*(x_{t-1})) - \bar{H}_v^t & - \frac{(2\gamma + r_t) \|x_t - x_{t-1}\|}{2} I_{d_x}, \end{aligned}$$

其中第 2 个不等式由引理 6 得到.将上式整理可得:

$$\frac{(2\gamma + r_t) \|x_t - x_{t-1}\|}{2} \geq \lambda_{\min}(\nabla_{xx}^2 f(x_{t-1}, y^*(x_{t-1})) - \bar{H}_v^t) - \lambda_{\min}(\nabla_{xx}^2 f(x_t, y^*(x_{t-1}))).$$

上述不等式两边同时取期望,并结合 Jensen 不等式和式(9),有:

$$\sqrt{E[\|x_t - x_{t-1}\|^2]} \geq E[\|x_t - x_{t-1}\|] \geq \frac{-2}{2\gamma + r_t} [\sqrt{2(r_t + \gamma)\eta_t^H} + E[\lambda_{\min}(\nabla_{xx}^2 f(x_t, y^*(x_{t-1})))]].$$

最后结合上述不等式与式(17)可证得引理 9 成立.

证毕

引理 10 在引理 9 的条件下,序列 $\{(x_t, y_t)\}$ 由算法 1 产生.对任意给定的 $T \geq 1$,有:

$$E[\|x_t - x_{t-1}\|^3] \leq \frac{36}{\sum_{t=1}^T r_t} \left[Q(x_0) - Q^* + \sum_{t=1}^T \frac{4(\eta_t^g)^{3/2}}{\sqrt{3r_t}} + \sum_{t=1}^T \frac{324\sqrt{2}}{r_t^2} ((\gamma + r_t)\eta_t^H)^{3/2} \right].$$

证明 由引理 2 以及 $r_t \geq \omega$ 有:

$$Q(x_t) - Q(x_{t-1}) - \langle \nabla Q(x_{t-1}), x_t - x_{t-1} \rangle - \frac{1}{2} \langle x_t - x_{t-1}, \nabla^2 Q(x_{t-1})(x_t - x_{t-1}) \rangle \leq \frac{r_t}{6} \|x_t - x_{t-1}\|^3.$$

由算法 1 中 $\hat{f}(x, x_{t-1}, \bar{G}_v^t, \bar{H}_v^t, r_t)$ 定义,结合上式可得:

$$\begin{aligned} Q(x_t) & \leq Q(x_{t-1}) + \hat{f}(x_t, x_{t-1}, \bar{G}_v^t, \bar{H}_v^t, r_t) + \|\nabla Q(x_{t-1}) - \bar{G}_v^t\| \|x_t - x_{t-1}\| + \\ & \frac{1}{2} \|\nabla^2 Q(x_{t-1}) - \bar{H}_v^t\| \|x_t - x_{t-1}\|^2. \end{aligned} \quad (18)$$

由引理 6 有:

$$\hat{f}(x_t, x_{t-1}, \bar{G}_v^t, \bar{H}_v^t, r_t) \leq -\frac{r_t}{12} \|x_t - x_{t-1}\|^3,$$

将上述不等式与式(18)结合,可得:

$$\begin{aligned} \frac{r_t}{12} \|x_t - x_{t-1}\|^3 &\leq Q(x_{t-1}) - Q(x_t) + \|\nabla Q(x_{t-1}) - \bar{G}_v^t\| \|x_t - x_{t-1}\| + \frac{1}{2} \|\nabla^2 Q(x_{t-1}) - \bar{H}_v^t\| \|x_t - x_{t-1}\|^2 \leq \\ &Q(x_{t-1}) - Q(x_t) + \frac{4}{\sqrt{3}r_t} \|\nabla Q(x_{t-1}) - \bar{G}_v^t\|^{3/2} + \left(\frac{9\sqrt{2}}{r_t}\right)^2 \|\nabla^2 Q(x_{t-1}) - \bar{H}_v^t\|^3 + \frac{r_t}{18} \|x_t - x_{t-1}\|^3, \end{aligned} \quad (19)$$

其中第 2 个不等式由 Young 不等式得到。式(19)两边同时取期望,并利用式(8)和(9),整理可得:

$$\frac{r_t}{36} E[\|x_t - x_{t-1}\|^3] \leq Q(x_{t-1}) - Q(x_t) + \frac{4(\eta_t^g)^{3/2}}{\sqrt{3}r_t} + \frac{324\sqrt{2}}{r_t^2} ((\gamma + r_t)\eta_t^H)^{3/2}.$$

对上述不等式两端从 1 到 T 求和,然后两边同除以 $\sum_{t=1}^T r_t$, 可证得引理 10 成立。

证毕

基于以上引理,下面给出算法 1 的主要结果。

定理 1 如果假设 1 成立,序列 $\{(x_t, y_t)\}$ 由算法 1 产生。对任意给定的 $\epsilon > 0$, 在算法 1 中令:

$$\begin{aligned} a_t &= \frac{26(d_x + 5)B^2}{\epsilon^2}, b_t = \frac{2(1 + 2\log(2d_x))l^2}{\omega\epsilon} (4(d_x + 16))^4, r_t = \omega, \\ v &\leq \frac{1}{2} \min \left\{ \frac{\epsilon}{l(d_x + 3)^{3/2}}, \sqrt{\frac{\omega\epsilon}{36\gamma^2(d_x + 16)^5}} \right\}. \end{aligned} \quad (20)$$

则至多经过 $T = \left\lceil \frac{192\sqrt{\omega}(Q(x_0) - Q^*)}{\epsilon^{3/2}} \right\rceil$ 次迭代, x_T 是 $Q(x)$ 的一个 $(\epsilon, \sqrt{\omega\epsilon})$ 二阶平稳点, 即 x_T 满足:

$$\max \left\{ \sqrt{E[\|\nabla Q(x_T)\|]}, -\sqrt{\frac{2}{\omega}} E[\lambda_{\min}(\nabla^2 Q(x_T))] \right\} \leq 12\sqrt{\epsilon}. \quad (21)$$

证明 首先证明经过至多 T 次迭代后序列 $\{x_t\}$ 满足算法 1 的终止条件 $\|x_t - x_{t-1}\| < \frac{1}{2}\sqrt{\frac{\epsilon}{\omega}}$ 。利用引理 1

和引理 2, 根据定理 1 中 a_t 和 b_t 以及 v 的选取, 可以得到式(8)和(9)中的 $\eta_t^g = \frac{2\epsilon}{5}$ 和 $\eta_t^H = \frac{\epsilon}{38}$ 。这样, 式(8)和(9)可分别表示为:

$$E[\|\nabla Q(x_{t-1}) - \bar{G}_v^t\|] \leq C_g \epsilon, E[\|\nabla^2 Q(x_{t-1}) - \bar{H}_v^t\|] \leq C_H \sqrt{\omega\epsilon}, \quad (22)$$

这里 $C_g = \frac{1}{192}$, $C_H = \frac{1}{48}$ 。下面假设算法 1 在 T 步以前都有 $\|x_t - x_{t-1}\| \geq \frac{1}{2}\sqrt{\frac{\epsilon}{\omega}}$ 。考虑:

$$\begin{aligned} Q(x_t) - Q(x_{t-1}) &\leq \\ \nabla Q(x_{t-1})^T (x_t - x_{t-1}) &+ \frac{1}{2} (x_t - x_{t-1})^T \nabla^2 Q(x_{t-1}) (x_t - x_{t-1}) + \frac{\omega}{6} \|x_t - x_{t-1}\|^3 \leq \\ -\frac{\omega}{12} \|x_t - x_{t-1}\|^3 &+ C_g \epsilon \|x_t - x_{t-1}\| + \frac{C_H \sqrt{\omega\epsilon}}{2} \|x_t - x_{t-1}\|^2 \leq \\ -\frac{\omega}{12} \|x_t - x_{t-1}\|^3 &+ 4C_g \omega \|x_t - x_{t-1}\|^3 + C_H \omega \|x_t - x_{t-1}\|^3, \end{aligned}$$

其中: 第 1 个不等式由引理 2 可得, 第 2 个不等式由式(22)并结合引理 6 可得, 第 3 个不等式由上面的假设

$\|x_t - x_{t-1}\| \geq \frac{1}{2}\sqrt{\frac{\epsilon}{\omega}}$ 得到。将上式整理可得:

$$\frac{1}{8} \left(\frac{1}{12} - 4C_g - C_H \right) \sqrt{\frac{\epsilon^3}{\omega}} \leq \left(\frac{1}{12} - 4C_g - C_H \right) \omega \|x_t - x_{t-1}\|^3 \leq Q(x_{t-1}) - Q(x_t). \quad (23)$$

注意到 $\frac{1}{8} \left(\frac{1}{12} - 4C_g - C_H \right) \sqrt{\frac{\epsilon^3}{\omega}} = \frac{1}{192\sqrt{\omega}} \sqrt{\frac{\epsilon^3}{\omega}} > 0$ 。将上述不等式(23)按 $t=1, \dots, T$ 求和, 有:

$$\frac{T}{192\sqrt{\frac{\epsilon^3}{\omega}}} \leq Q(x_0) - Q(x_T) \leq Q(x_0) - Q^*.$$

由上式求得迭代次数 $T \leq \frac{192\sqrt{\omega}(Q(x_0) - Q^*)}{\epsilon^{3/2}}$ 。这表明序列 $\{x_t\}$ 在至多 $T = \left\lceil \frac{192\sqrt{\omega}(Q(x_0) - Q^*)}{\epsilon^{3/2}} \right\rceil$ 次迭代后

都能满足算法 1 的终止条件 $\|x_t - x_{t-1}\| < \frac{1}{2}\sqrt{\frac{\epsilon}{\omega}}$ 。

下证经过 T 次迭代后 x_T 是 $Q(x)$ 的一个 $(\epsilon, \sqrt{\omega\epsilon})$ 二阶平稳点, 即证式(21)成立。将式(20)中的参数取值以及 $\eta_i^G = \frac{2\epsilon}{5}$ 和 $\eta_i^H = \frac{\epsilon}{38}$ 代入引理 10, 并注意到 $\gamma \leq \omega$, 有:

$$E[\|x_t - x_{t-1}\|^3] \leq \frac{1}{\omega^{3/2}} \left[\frac{36\sqrt{\omega}(Q(x_0) - Q^*)}{T} + 279\epsilon^{3/2} \right] \leq \frac{280}{\omega^{3/2}} \epsilon^{3/2}, \quad (24)$$

其中: 第 2 个不等式由选取 $T = \left\lceil \frac{192\sqrt{\omega}(Q(x_0) - Q^*)}{\epsilon^{3/2}} \right\rceil$ 得到。再将式(20)中的参数取值代入引理 9, 有

$$\max \left\{ \sqrt{\frac{E[\|\nabla Q(x_T)\|]}{2\omega}}, \frac{-E[\lambda_{\min}(\nabla^2 Q(x_T))]}{\omega} \right\} \leq \sqrt{E[\|x_T - x_{T-1}\|^2]} + \sqrt{\frac{\epsilon}{\omega}} \leq [E[\|x_T - x_{T-1}\|^3]]^{1/3} + \sqrt{\frac{\epsilon}{\omega}} \leq (1 + (280)^{1/3}) \sqrt{\frac{\epsilon}{\omega}},$$

其中: 第 2 个不等式由 Jensen 不等式可得, 第 3 个不等式由式(24)可得。最后, 上式可证得式(21)。再由定义 1 可知, 此时的 x_T 是 $Q(x)$ 的一个 $(\epsilon, \sqrt{\omega\epsilon})$ 二阶平稳点。证毕

注 2 1) 由定理 1 的结果知, ZO-MCN 算法的迭代复杂度为 $T = O(\epsilon^{-3/2})$ 。另外, 根据 a_t 和 b_t 的选取可得, 计算近似梯度和近似 Hessian 阵所需函数值的总采样次数分别为 $O(d_x \epsilon^{-7/2})$ 和 $O(d_x^4 \epsilon^{-5/2})$ 。

2) 通过对比, 本文提出的 ZO-MCN 算法与 MCN 算法^[5]的迭代复杂度是同阶的, 但本文提出的算法不需要计算精确的梯度和 Hessian 阵信息。

3 数值实验

下面通过求解 2 个不同维度的问题来测试 ZO-MCN 算法的数值表现。实验环境为笔记本电脑 AMD Ryzen 54600U 处理器, 16 GB RAM, 所有算法均用 Matlab 编码实现。

例 1 考虑一个非凸-强凹的合成极小极大问题^[5]:

$$\min_{x \in \mathbf{R}^3} \max_{y \in \mathbf{R}^2} f(x, y) := \varphi(x_3) - \frac{y_1^2}{20} + x_1 y_1 - \frac{5y_2^2}{2} + x_2 y_2,$$

这里 $x = [x_1, x_2, x_3]^T$, $y = [y_1, y_2]^T$, 且 $\varphi(\cdot)$ 如下所示:

$$\varphi(z) = \begin{cases} \sqrt{\theta}(z + (C+1)\sqrt{\theta})^2 - \frac{1}{3}(z + (C+1)\sqrt{\theta})^3 - \frac{1}{3}(3C+1)\theta^{3/2}, & z \leq -C\sqrt{\theta}; \\ \theta z + \frac{\theta^{3/2}}{3}, & -C\sqrt{\theta} < z \leq -\sqrt{\theta}; \\ -\sqrt{\theta}z^2 - \frac{z^3}{3}, & -\sqrt{\theta} < z \leq 0; \\ -\sqrt{\theta}z^2 + \frac{z^3}{3}, & 0 < z \leq \sqrt{\theta}; \\ -\theta z + \frac{\theta^{3/2}}{3}, & \sqrt{\theta} < z \leq C\sqrt{\theta}; \\ \sqrt{\theta}(z - (C+1)\sqrt{\theta})^2 + \frac{1}{3}(z - (C+1)\sqrt{\theta})^3 - \frac{1}{3}(3C+1)\theta^{3/2}, & C\sqrt{\theta} \leq z. \end{cases}$$

容易证明点 $(x_0, y_0) = ([0, 0, 0]^T, [0, 0]^T)$ 是该问题的一个一阶平稳点。

首先考虑采样次数对 ZO-MCN 算法收敛性结果和 CPU 运行时间的影响。图 1 与图 2 给出了 4 种不同采

样次数设置情形下,ZO-MCN 算法达到预设精度 $\epsilon=1e-4$ 时所需迭代次数和 CPU 运行时间的对比结果。由图 1 可以看到,ZO-MCN 算法所需的迭代次数随着采样次数 a 和 b 的增加而减少,其中采样次数 $a=5$ 和 $b=2$ 时收敛效果较差,采样次数 $a=35$ 和 $b=20$ 时收敛效果较好。从图 2 可以看到,算法所需的 CPU 运行时间随着采样次数 a 和 b 的增加而增加,其中采样次数 $a=35$ 和 $b=20$ 时 CPU 运行时间较多,采样次数 $a=5$ 和 $b=2$ 时 CPU 运行时间较短。综合图 1 与图 2 来看,随着采样次数的增加,ZO-MCN 算法达到预设精度时所需的迭代次数减少,但 CPU 运行时间会增加,这表明 ZO-MCN 算法需要在采样次数和 CPU 运行时间上进行权衡。

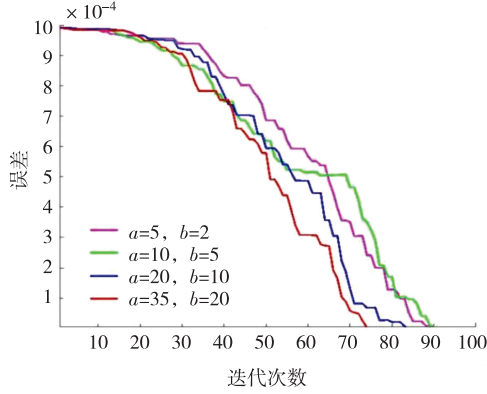


图 1 采样次数对 ZO-MCN 算法收敛性影响

Fig. 1 Influence of sampling times on convergence for algorithm ZO-MCN

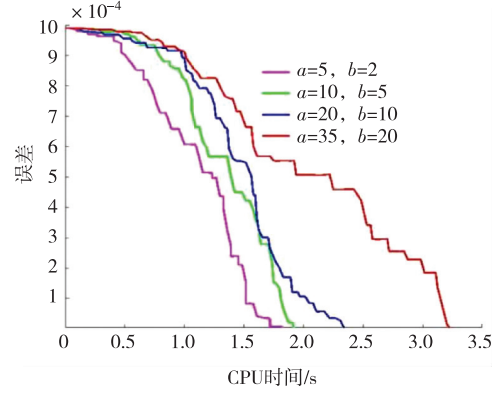


图 2 采样次数对 ZO-MCN 算法运行时间影响

Fig. 2 Influence of sampling times on CPU time for algorithm ZO-MCN

接下来对比 ZO-MCN 算法、MCN 算法^[5]和 GDA 算法^[9]在收敛速度和 CPU 运行时间上的数值表现。在实验中,取 $\theta=0.01$ 和 $C=5$;选取初始点 $(x_1, y_1) = ([10^{-3}, 10^{-3}, 10^{-3}]^T, [0, 0]^T)$,以及精度 $\epsilon=1e-4$ 。在 ZO-MCN 算法中,取参数 $\omega=10$,梯度采样次数 $a=5$,Hessian 阵采样次数 $b=2$,这里采样次数的设置可参考灵敏度分析权衡得到。将 3 种算法计算的误差 $Q(x) - Q^*$ 与迭代次数、运行时间分别进行对比,见图 3 和图 4。由图 3 可以看到随着迭代次数的增加,3 种算法都能渐进收敛到最优值。MCN 算法收敛效果稍好于 ZO-MCN 算法,这是因为 MCN 算法使用了精确的梯度和 Hessian 阵信息,但本文提出的 ZO-MCN 算法仅仅使用了目标函数值信息。在相同的精度下,MCN 算法和 ZO-MCN 算法均好于 GDA 算法。从图 4 可以看出,达到相同的精度下 ZO-MCN 算法的 CPU 时间相比 GDA 算法缩短了 87% 左右,相较 MCN 算法所用 CPU 时间缩短了 60% 左右。

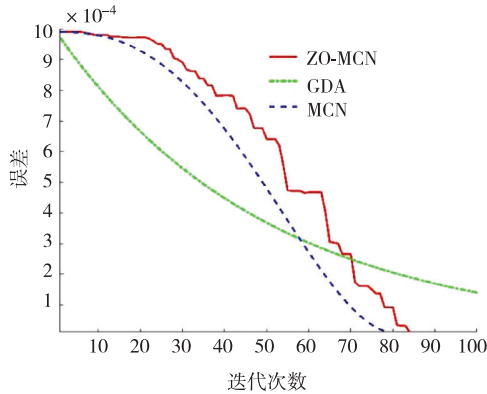


图 3 ZO-MCN 算法与 MCN、GDA 算法的迭代次数对比

Fig. 3 Comparison of iterations among algorithms ZO-MCN, MCN and GDA

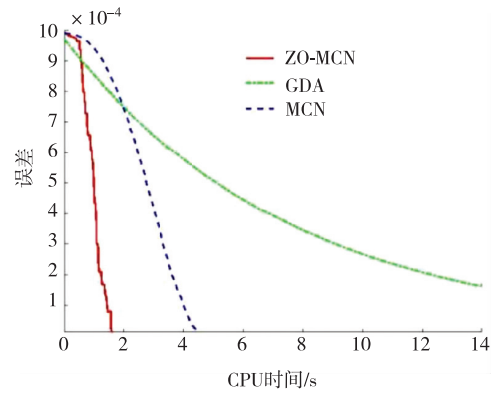


图 4 ZO-MCN 算法与 MCN、GDA 算法的运行时间对比

Fig. 4 Comparison of CPU time among algorithms ZO-MCN, MCN and GDA

例 2 考虑如下一个非凸-强凹的极小极大问题^[5]:

$$\min_{x \in \mathbf{R}^d} \max_{y \in \mathbf{R}^d} f(x, y) :=$$

$$\left(\sum_{i=1}^{d_x} x_i - \sum_{j=1}^{d_y} y_j\right)^3 - \left(\sum_{j=1}^{d_y} y_j\right)^2 + \left(\sum_{j=1}^{d_y} y_j\right)^3 + \sum_{i=1}^{d_x} x_i \cdot \sum_{j=1}^{d_y} y_j + \left(\sum_{i=1}^{d_x} x_i\right)^2 - 3 \sum_{i=1}^{d_x} x_i \cdot \left(\sum_{j=1}^{d_y} y_j\right)^2 - \sum_{j=1}^{d_y} y_j^2.$$

下面通过数值实验考察 ZO-MCN 算法与 MCN 算法^[5]在较高维数上的数值表现。在实验中,设置 $d_x = 1\ 000$, $d_y = 100$ 和精度 $\varepsilon = 1e-3$; 初始点设为 $x = (1/d_x, 1/d_x, \dots, 1/d_x)^\top$ 和 $y = (1/d_y, 1/d_y, \dots, 1/d_y)^\top$ 。在 ZO-MCN 算法中,取参数 $\omega = 10$, $a = 8$ 和 $b = 3$ 。图 5 和图 6 给出了 2 种算法的迭代次数与 CPU 运行时间的对比结果。由图 5 可以看到,在达到相同精度下,ZO-MCN 算法所需迭代次数比算法 MCN 稍微多一些,这是由于 MCN 算法使用了精确的梯度和 Hessian 阵信息,而 ZO-MCN 算法仅使用了函数值信息。但是,MCN 算法所用的 CPU 时间大致为 ZO-MCN 算法所用时间的 3 倍,见图 6。

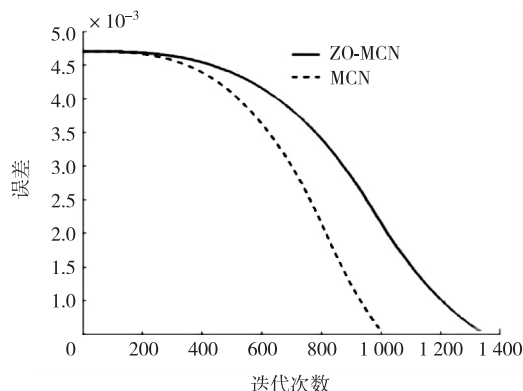


图 5 ZO-MCN 算法与 MCN 算法的收敛性对比

Fig. 5 Comparison of convergence for algorithms ZO-MCN and MCN

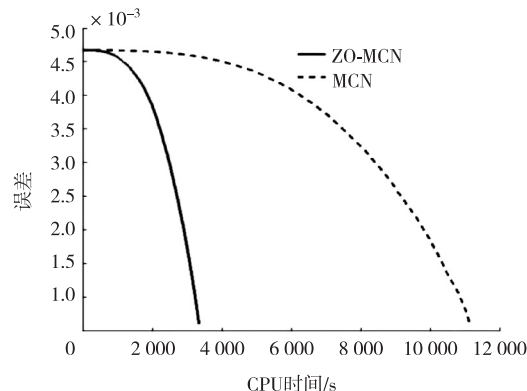


图 6 ZO-MCN 算法与 MCN 算法的运行时间对比

Fig. 6 Comparison of CPU time for algorithms ZO-MCN and MCN

4 结论

考虑到许多问题中目标函数的梯度和 Hessian 阵信息往往是难以计算或难以获取的情形,本文针对一类非凸-凹极小极大优化问题,提出了一类零阶极小极大三次正则化牛顿算法,分析了算法的收敛性,并给出了算法的迭代复杂度。数值仿真表明了算法的有效性。与文献[5]的算法相比,本文的算法不需要精确计算目标函数的梯度和 Hessian 阵信息,仅仅需要计算目标函数的函数值信息。

参考文献:

- [1] 张柯,白富生,吴至友,等. 基于对抗生成网络的人脸照片去网纹技术[J]. 重庆师范大学学报(自然科学版),2019,36(6):110-116.
ZHANG K, BAI F S, WU Z Y, et al. Method of elimination mesh of face photograph based on generative adversarial networks [J]. Journal of Chongqing Normal University (Natural Science), 2019, 36(6): 110-116.
- [2] LIU S J, CHEN P Y, KAILKHURA B, et al. A primer on zeroth-order optimization in signal processing and machine learning: principals, recent advances, and applications[J]. IEEE Signal Processing Magazine, 2020, 37(5): 43-54.
- [3] 徐姿,张慧灵. 非凸极小极大优化问题的优化算法与复杂度分析[J]. 运筹学学报, 2021, 25(3): 74-86.
XU Z, ZHANG H L. Optimization algorithms and their complexity analysis for non-convex minimax problems[J]. Operations Research Transactions, 2021, 25(3): 74-86.
- [4] NESTEROV Y, POLYAK B T. Cubic regularization of Newton method and its global performance[J]. Mathematical Programming, 2006, 108(1): 177-205.
- [5] LUO L, LI Y J, CHEN C. Finding second-order stationary point for nonconvex-strongly-concave minimax problem[EB/OL]. (2021-10-10)[2022-10-17]. https://arxiv.org/abs/2110.04814v3.
- [6] XU Z, SHEN J J, WANG Z Q, et al. Zeroth-order alternating randomized gradient projection algorithms for general nonconvex-concave minimax problems[EB/OL]. (2021-08-01)[2022-10-17]. https://arxiv.org/abs/2108.00473v1.
- [7] BALASUBRAMANIAN K, GHADIMI S. Zeroth-order nonconvex stochastic optimization: handling constraints, high dimensionality,

- and saddle points[J]. Foundations of Computational Mathematics, 2022, 22(1): 35-76.
- [8] NESTEROV Y, SPOKOINY V. Random gradient-free minimization of convex functions[J]. Foundations of Computational Mathematics, 2017, 17(2): 527-566.
- [9] LIN T Y, JIN C, JORDAN M I. On gradient descent ascent for nonconvex-concave minimax problems[EB/OL]. (2019-06-02) [2022-10-17]. <https://arxiv.org/abs/1906.00331v4>.
- [10] TROPP J A. The expected norm of a sum of independent random matrices: an elementary approach[C]//HOUDRÉ C, MASON D, REYNAUD-BOURET P, et al. High Dimensional Probability VII: Progress in Probability, vol 71. Cham: Birkhäuser; 2016: 173-202.
- [11] PINELIS I. Optimum bounds for the distributions of martingales in Banach spaces[J]. The Annals of Probability, 1994, 22(4): 1679-1706.

Special Topics on Optimization Theory and Algorithm

Zeroth-Order Optimization Algorithm for a Class of Non-Convex Strongly Concave Mini-Max Problems

GAO Ruicheng, XIE Tao, LI Jueyou

(School of Mathematical Sciences, Chongqing Normal University, Chongqing 401331, China)

Abstract: A class of mini-max optimization problems is frequently encountered in machine learning. A large number of optimization algorithms have been designed for mini-max problems based on the gradient and Hessian matrix information of the objective function. However, in some applications, the gradient and Hessian matrix information are possibly computationally expensive or difficult to obtain. Therefore, under the framework of mini-max cubic Newton algorithm (MCN), a zeroth-order mini-max cubic regularized Newton algorithm (ZO-MCN) is proposed for solving the non-convex strongly concave mini-max problem by using Gaussian smoothing method to approximate the gradient and Hessian matrix based on Stein identity. The convergence and iteration complexity of the proposed algorithm are analyzed. The iteration complexity of the algorithm is the order of $O(\epsilon^{-3/2})$ for achieving a second-order stable point, where ϵ is the accuracy. The results of numerical simulations show that: under the same accuracy, the proposed algorithm is better than the algorithm MCN in CPU running time.

Keywords: non-convex concave mini-max problem; cubic regularized Newton algorithm; zeroth-order algorithm; complexity analysis

(责任编辑 黄 颖)