

凸-凹极小极大优化问题的零阶梯度下降上升算法*

谢涛, 高瑞成, 童殷, 李觉友
(重庆师范大学 数学科学学院, 重庆 401331)

摘要:【目的】为了解决基于梯度下降上升算法在某些应用中,目标函数的梯度信息计算昂贵或难以获取的问题。【方法】基于此,针对一类凸-凹极小极大优化问题,在梯度下降上升算法(OGDA)的框架下,基于均匀分布的平滑化方法用差商来近似函数梯度信息,提出了一类零阶梯度下降上升算法(ZO-OGDA)。【结果】基于带误差的邻近点算法的收敛性分析理论,证明得到所提算法 ZO-OGDA 取得 ϵ -稳定点的迭代复杂度为 $\mathcal{O}(\epsilon^{-1})$ 。【结论】最后通过数值仿真,实验结果表明所提出的算法 ZO-OGDA 在数值上与算法 OGDA 表现相近。

关键词:凸-凹极小极大;零阶算法;梯度下降上升算法;复杂度分析

中图分类号:O221.2

文献标志码:A

文章编号:1672-6693(2023)01-0105-09

考虑如下的极小极大优化问题(或称鞍点问题):

$$\min_{u \in \mathcal{U}} \max_{v \in \mathcal{V}} f(u, v), \quad (1)$$

其中: $f: \mathcal{U} \times \mathcal{V} \rightarrow \mathbf{R}$ 是一个连续可微函数, $\mathcal{U} \subset \mathbf{R}^{d_u}$ 和 $\mathcal{V} \subset \mathbf{R}^{d_v}$ 是非空闭凸集, d_u 和 d_v 分别是变量 u 和 v 的维数。特别地,当问题(1)中的目标函数 $f(u, v)$ 是凸-凹函数时,即对任意的 $v \in \mathcal{V}$, 函数 $f(u, v)$ 关于变量 u 是凸函数,对任意的 $u \in \mathcal{U}$, 函数 $f(u, v)$ 关于变量 v 是凹函数,则称此类问题为凸-凹极小极大优化问题。对任意的 $u \in \mathcal{U}$, $v \in \mathcal{V}$, 如果满足: $f(u^*, v) \leq f(u^*, v^*) \leq f(u, v^*)$, 称点 $(u^*, v^*) \in \mathcal{U} \times \mathcal{V}$ 为问题(1)的鞍点。当 f 是凸-凹函数时,如果对任意的 $\epsilon > 0$, 点 $(u, v) \in \mathcal{U} \times \mathcal{V}$ 满足 $|f(u, v) - f(u^*, v^*)| \leq \epsilon$, 则点 (u, v) 为问题(1)的 ϵ -稳定点。问题(1)的应用比较广泛,已经出现在众多研究领域,包括零和博弈^[1]、机器学习中的生成对抗网络模型(GANs)^[2]等。另外,该问题与拉格朗日对偶问题以及变分不等式等问题密切相关,因此引起了诸如在最优化、计算机和人工智能等领域一大批学者的关注和研究^[3]。

对于求解凸-凹极小极大优化问题,梯度下降上升(GDA)算法由于简单有效而倍受青睐。但有研究表明,GDA 算法即使是在求解双线性凸-凹极小极大优化问题时也难以保证收敛性^[4]。因此,大量改进的 GDA 算法被学者们相继提出。2019年, Lin 等人^[5]提出了一个基于双时间尺度的 GDA 算法,并得到了算法收敛性结果。2020年, Mokhtari 等人^[6]在传统的 GDA 算法基础上,通过引入正则项进而提出了一类 GDA 算法(OGDA),通过将 OGDA 算法的迭代过程看作邻近点算法(PPA)框架下的一种带误差的特殊形式,借由 PPA 算法的收敛性理论证得了 OGDA 算法的迭代复杂度为 $\mathcal{O}(\epsilon^{-1})$ 。但以上这些算法需要计算精确的梯度信息,而在许多实际情况中梯度信息往往是难以获取的。

近些年来,零阶算法(仅使用函数值信息,而不需要梯度信息)在机器学习的众多领域得到广泛关注^[7-13],例如深层神经网络系统的黑盒攻击问题^[7]和 Bandit 优化问题^[8]。零阶算法能处理函数的梯度信息难以获取或精确计算梯度代价过大的情况。Nesterov^[9]较早给出了一种基于高斯分布的平滑化方案,设计出了一类基于高斯平滑化方法的零阶优化算法来求解无约束极小化问题。高斯分布的平滑化方案通常用来处理无约束优化问题,于是针对带有约束的极小极大优化问题, Xu 等人^[10]引入一种基于均匀分布的平滑化方案,提出了零阶交替随机梯度投影算法(AGP),并得到了该算法的复杂度结果。Liu 等人^[7]研究了在黑盒情形下的极小极大优化问题,提出了一类零阶算法(ZO-Min-Max),得到了获得局部鲁棒解时调用目标函数值所需要的计算次数估计。针对凸-

* 收稿日期:2022-09-02 修回日期:2022-12-03 网络出版时间:2023-02-24 11:31

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

第一作者简介:谢涛,男,研究方向为最优化理论与算法,E-mail:1647365355@qq.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://kns.cnki.net/kcms/detail/50.1165.N.20230223.1706.006.html

凹极小极大优化问题,Beznošikov 等人^[11]提出了一类零阶鞍点优化算法(ZO-SPA),并给出了该算法的 Oracle 复杂度结果为 $\mathcal{O}(\epsilon^{-2})$,还给出了梯度估计所需的函数值计算次数估计。基于镜像下降方法,Sadiev 等人^[12]提出了一类零阶镜像下降优化算法,得到了算法的迭代复杂度为 $\mathcal{O}(\epsilon^{-2})$ 。但这些零阶优化算法的迭代复杂度还有待进一步改进。

受文献[6]的工作的启发,针对带有约束的凸-凹极小极大优化问题,基于均匀分布的平滑化方案^[10],提出一类零阶梯度下降上升算法(ZO-OGDA)。基于带误差的邻近点算法的收敛性分析理论,得到本文所提出算法 ZO-OGDA 取得 ϵ -稳定点的迭代复杂度为 $\mathcal{O}(\epsilon^{-1})$ 。

1 凸-凹极小极大优化问题的零阶算法

1.1 梯度估计

为了设计零阶算法,首先引入一个函数的梯度估计^[10]。设 $f: \mathbf{R}^{d_u+d_v} \rightarrow \mathbf{R}$,对任意的 $\mathbf{u} \in \mathcal{U}, \mathbf{v} \in \mathcal{V}$,定义:

$$\begin{aligned}\hat{\nabla}_{\mathbf{u}} f(\mathbf{u}, \mathbf{v}) &:= \frac{d_{\mathbf{u}}(f(\mathbf{u} + \mu_1 \mathbf{p}, \mathbf{v}) - f(\mathbf{u}, \mathbf{v}))}{\mu_1} \mathbf{p}, \mathbf{p} \sim S_{\mathbf{u}}, \\ \hat{\nabla}_{\mathbf{v}} f(\mathbf{u}, \mathbf{v}) &:= \frac{d_{\mathbf{v}}(f(\mathbf{u}, \mathbf{v} + \mu_2 \mathbf{q}) - f(\mathbf{u}, \mathbf{v}))}{\mu_2} \mathbf{q}, \mathbf{q} \sim S_{\mathbf{v}},\end{aligned}\quad (2)$$

其中: $\mu_1, \mu_2 > 0$ 为平滑化参数, $\mathbf{p} \sim S_{\mathbf{u}}$ 表示随机向量 \mathbf{p} 在 $d_{\mathbf{u}}$ 维的单位欧氏球面 $S_{\mathbf{u}}$ 服从均匀分布, $\mathbf{q} \sim S_{\mathbf{v}}$ 类似。定义 $f(\mathbf{u}, \mathbf{v})$ 关于变量 \mathbf{u} 和 \mathbf{v} 的两个平滑化函数:

$$f_{\mu_1}(\mathbf{u}, \mathbf{v}) := \mathbb{E}[f(\mathbf{u} + \mu_1 \mathbf{p}, \mathbf{v})], f_{\mu_2}(\mathbf{u}, \mathbf{v}) := \mathbb{E}[f(\mathbf{u}, \mathbf{v} + \mu_2 \mathbf{q})],$$

其中: \mathbb{E} 表示数学期望,再根据文献[11]的引理 4.1 可得:

$$\mathbb{E}[\hat{\nabla}_{\mathbf{u}} f(\mathbf{u}, \mathbf{v})] = \nabla_{\mathbf{u}} f_{\mu_1}(\mathbf{u}, \mathbf{v}), \mathbb{E}[\hat{\nabla}_{\mathbf{v}} f(\mathbf{u}, \mathbf{v})] = \nabla_{\mathbf{v}} f_{\mu_2}(\mathbf{u}, \mathbf{v}).\quad (3)$$

由上式可知,可以通过差商的形式(2)来近似估计平滑化以后的函数 $f_{\mu_1}(\mathbf{u}, \mathbf{v})$ 和 $f_{\mu_2}(\mathbf{u}, \mathbf{v})$ 的梯度^[10]。

1.2 ZO-OGDA

OGDA 算法的主要思想是极小极大化对应的目标函数的一阶近似与一个正则项之和。但该算法需要计算精确梯度。由于在有些应用中梯度计算是比较昂贵的,或者是难以获取的。为此,借助上一小节有关梯度估计的方法,设计出 ZO-OGDA 算法来求解问题(1)。下面给出具体的算法框架:

算法 ZO-OGDA

步骤 0: 设置 $\mathbf{u}_{-1} = \mathbf{u}_0 \in \mathcal{U}$ 和 $\mathbf{v}_{-1} = \mathbf{v}_0 \in \mathcal{V}$ 。设置步长 $\eta > 0$, 采样次数序列 $\{m_{1,k}\}, \{m_{2,k}\}$ 和正的单调递减的平滑化参数序列 $\{\mu_{1,k}\}, \{\mu_{2,k}\}$ 。令 $k=0$ 。

步骤 1: 计算

$$\begin{aligned}\hat{\nabla}_{\mathbf{u}} f(\mathbf{u}_k, \mathbf{v}_k) &:= \frac{1}{m_{1,k}} \sum_{i=1}^{m_{1,k}} \frac{d_{\mathbf{u}}(f(\mathbf{u}_k + \mu_{1,k} \mathbf{p}_i, \mathbf{v}_k) - f(\mathbf{u}_k, \mathbf{v}_k))}{\mu_{1,k}} \mathbf{p}_i, \mathbf{p}_i \sim S_{\mathbf{u}}, \\ \hat{\nabla}_{\mathbf{v}} f(\mathbf{u}_k, \mathbf{v}_k) &:= \frac{1}{m_{2,k}} \sum_{i=1}^{m_{2,k}} \frac{d_{\mathbf{v}}(f(\mathbf{u}_k, \mathbf{v}_k + \mu_{2,k} \mathbf{q}_i) - f(\mathbf{u}_k, \mathbf{v}_k))}{\mu_{2,k}} \mathbf{q}_i, \mathbf{q}_i \sim S_{\mathbf{v}}.\end{aligned}\quad (4)$$

步骤 2: 更新迭代点 $\mathbf{w}_{k+1} := [\mathbf{u}_{k+1}; \mathbf{v}_{k+1}]$:

$$\begin{aligned}\mathbf{u}_{k+1} &:= \arg \min_{\mathbf{u} \in \mathcal{U}} \langle 2\hat{\nabla}_{\mathbf{u}} f(\mathbf{u}_k, \mathbf{v}_k) - \hat{\nabla}_{\mathbf{u}} f(\mathbf{u}_{k-1}, \mathbf{v}_{k-1}), \mathbf{u} - \mathbf{u}_k \rangle + \frac{1}{2\eta} \|\mathbf{u} - \mathbf{u}_k\|^2, \\ \mathbf{v}_{k+1} &:= \arg \max_{\mathbf{v} \in \mathcal{V}} \langle 2\hat{\nabla}_{\mathbf{v}} f(\mathbf{u}_k, \mathbf{v}_k) - \hat{\nabla}_{\mathbf{v}} f(\mathbf{u}_{k-1}, \mathbf{v}_{k-1}), \mathbf{v} - \mathbf{v}_k \rangle - \frac{1}{2\eta} \|\mathbf{v} - \mathbf{v}_k\|^2.\end{aligned}\quad (5)$$

步骤 3: 当算法的迭代点是问题(1)的 ϵ -稳定点时,停止迭代。否则,取 $k := k+1$,并回到步骤 1。

注 1 1) 在 ZO-OGDA 算法中,(4)式表示用多次采样的函数值信息来近似梯度信息;(5)式的第一个公式右边的第一项是分别关于变量 \mathbf{u} 和 \mathbf{v} 对函数 f 的一阶近似,第二项为正则项; $\hat{\nabla}_{\mathbf{u}} f(\mathbf{u}_k, \mathbf{v}_k) - \hat{\nabla}_{\mathbf{u}} f(\mathbf{u}_{k-1}, \mathbf{v}_{k-1})$ 视为关于变量 \mathbf{u} 的“动量项”。2) 对比 OGDA 算法,本文提出的算法是将 OGDA 算法对应的梯度信息 $\nabla_{\mathbf{u}} f(\mathbf{u}_k, \mathbf{v}_k)$ 和 $\nabla_{\mathbf{v}} f(\mathbf{u}_k, \mathbf{v}_k)$ 换成了(4)式的 $\hat{\nabla}_{\mathbf{u}} f(\mathbf{u}_k, \mathbf{v}_k)$ 和 $\hat{\nabla}_{\mathbf{v}} f(\mathbf{u}_k, \mathbf{v}_k)$,这样做的优点是能处理梯度信息难以获取的问题。

引入 Euclidean 投影算子 $P_{\mathcal{U}}$ 和 $P_{\mathcal{V}}$, (5)式可以等价的写成如下形式:

$$\begin{aligned} \mathbf{u}_{k+1} &= P_{\mathcal{U}}(\mathbf{u}_k - 2\eta \hat{\nabla}_{\mathbf{u}} f(\mathbf{u}_k, \mathbf{v}_k) + \eta \hat{\nabla}_{\mathbf{u}} f(\mathbf{u}_{k-1}, \mathbf{v}_{k-1})), \\ \mathbf{v}_{k+1} &= P_{\mathcal{V}}(\mathbf{v}_k + 2\eta \hat{\nabla}_{\mathbf{v}} f(\mathbf{u}_k, \mathbf{v}_k) - \eta \hat{\nabla}_{\mathbf{v}} f(\mathbf{u}_{k-1}, \mathbf{v}_{k-1})). \end{aligned} \quad (6)$$

为了证明方便,令 $\mathbf{w} := [\mathbf{u}; \mathbf{v}] \in \mathbf{R}^{d_u+d_v}$ 和 $\nabla F(\mathbf{w}_k) := [\nabla_{\mathbf{u}} f(\mathbf{u}_k, \mathbf{v}_k); -\nabla_{\mathbf{v}} f(\mathbf{u}_k, \mathbf{v}_k)]$ 。类似地,有:

$$\hat{\nabla} F(\mathbf{w}_k) := [\hat{\nabla}_{\mathbf{u}} f(\mathbf{u}_k, \mathbf{v}_k); -\hat{\nabla}_{\mathbf{v}} f(\mathbf{u}_k, \mathbf{v}_k)], \nabla F_{\mu}(\mathbf{w}_k) := [\nabla_{\mathbf{u}} f_{\mu_{1,k}}(\mathbf{u}_k, \mathbf{v}_k); -\nabla_{\mathbf{v}} f_{\mu_{2,k}}(\mathbf{u}_k, \mathbf{v}_k)],$$

由此,(6)式可简写为:

$$\mathbf{w}_{k+1} = P_{\mathcal{U} \times \mathcal{V}}\{\mathbf{w}_k - 2\eta \hat{\nabla} F(\mathbf{w}_k) + \eta \hat{\nabla} F(\mathbf{w}_{k-1})\}.$$

为了证明 ZO-OGDA 算法的收敛性,将上式表示成 PPA 算法的特殊形式,如下:

$$\mathbf{w}_{k+1} = P_{\mathcal{U} \times \mathcal{V}}\{\mathbf{w}_k - \eta \hat{\nabla} F(\mathbf{w}_{k+1}) + \xi_k\}, \quad (7)$$

其中:

$$\xi_k = \eta [(\hat{\nabla} F(\mathbf{w}_{k+1}) - \hat{\nabla} F(\mathbf{w}_k)) - (\hat{\nabla} F(\mathbf{w}_k) - \hat{\nabla} F(\mathbf{w}_{k-1}))]. \quad (8)$$

1.3 算法收敛性分析

假设 1 函数 $f(\mathbf{u}, \mathbf{v})$ 关于 \mathbf{u} 和 \mathbf{v} 连续可微。对任意 $\mathbf{v} \in \mathcal{V}$, 函数 $f(\mathbf{u}, \mathbf{v})$ 关于 $\mathbf{u} \in \mathcal{U}$ 是凸函数; 对任意 $\mathbf{u} \in \mathcal{U}$, 函数 $f(\mathbf{u}, \mathbf{v})$ 对于 $\mathbf{v} \in \mathcal{V}$ 是凹函数。

假设 2 梯度 $\nabla_{\mathbf{u}} f(\mathbf{u}, \mathbf{v})$ 分别关于变量 \mathbf{u} 和 \mathbf{v} 是 L_u 利普希茨连续的, 且梯度 $\nabla_{\mathbf{v}} f(\mathbf{u}, \mathbf{v})$ 分别关于变量 \mathbf{u} 和 \mathbf{v} 是 L_v 利普希茨连续的, 即:

$$\begin{aligned} \|\nabla_{\mathbf{u}} f(\mathbf{u}, \mathbf{v}_1) - \nabla_{\mathbf{u}} f(\mathbf{u}, \mathbf{v}_2)\| &\leq L_u \|\mathbf{v}_1 - \mathbf{v}_2\|, \forall \mathbf{u} \in \mathcal{U}; \|\nabla_{\mathbf{u}} f(\mathbf{u}_1, \mathbf{v}) - \nabla_{\mathbf{u}} f(\mathbf{u}_2, \mathbf{v})\| \leq L_u \|\mathbf{u}_1 - \mathbf{u}_2\|, \forall \mathbf{v} \in \mathcal{V}, \\ \|\nabla_{\mathbf{v}} f(\mathbf{u}_1, \mathbf{v}) - \nabla_{\mathbf{v}} f(\mathbf{u}_2, \mathbf{v})\| &\leq L_v \|\mathbf{u}_1 - \mathbf{u}_2\|, \forall \mathbf{v} \in \mathcal{V}; \|\nabla_{\mathbf{v}} f(\mathbf{u}, \mathbf{v}_1) - \nabla_{\mathbf{v}} f(\mathbf{u}, \mathbf{v}_2)\| \leq L_v \|\mathbf{v}_1 - \mathbf{v}_2\|, \forall \mathbf{u} \in \mathcal{U}, \end{aligned}$$

这里让 $L = 4\max\{L_u, L_v\}$ 。

假设 3 对任意 $\mathbf{u} \in \mathcal{U}, \mathbf{v} \in \mathcal{V}$, 存在常数 $G > 0$, 使得 $\|\nabla_{\mathbf{u}} f(\mathbf{u}, \mathbf{v})\|^2 \leq G^2$ 和 $\|\nabla_{\mathbf{v}} f(\mathbf{u}, \mathbf{v})\|^2 \leq G^2$ 成立。

假设 4 令最优解集 $\mathcal{Z}^* := \{\mathbf{w}^* = [\mathbf{u}^*; \mathbf{v}^*] \in \mathbf{R}^{d_u+d_v} : (\mathbf{u}^*, \mathbf{v}^*) \text{ 是问题(1)的鞍点}\}$, 假设 \mathcal{Z}^* 非空。

为了得到算法的收敛性结果, 引入下面几个引理。

引理 1^[10] 若假设 1 和假设 2 成立, 则:

1) 对任意的 $\mu_1 > 0$ 和 $\mu_2 > 0$, 有:

$$|f_{\mu_1}(\mathbf{u}, \mathbf{v}) - f(\mathbf{u}, \mathbf{v})| \leq \frac{L_u \mu_1^2}{2}, |f_{\mu_2}(\mathbf{u}, \mathbf{v}) - f(\mathbf{u}, \mathbf{v})| \leq \frac{L_v \mu_2^2}{2};$$

2) $f_{\mu_1}(\mathbf{u}, \mathbf{v})$ 关于 \mathbf{u} 是凸函数, $f_{\mu_2}(\mathbf{u}, \mathbf{v})$ 关于 \mathbf{v} 是凹函数; 且 $\nabla f_{\mu_1}(\mathbf{u}, \mathbf{v})$ 和 $\nabla f_{\mu_2}(\mathbf{u}, \mathbf{v})$ 分别具有 L_{μ_1} 和 L_{μ_2} 利普希茨连续性, 满足 $L_{\mu_1} \leq L$ 和 $L_{\mu_2} \leq L$;

3) 对于 ZO-OGDA 算法产生的迭代序列 $\{(\mathbf{u}_k, \mathbf{v}_k)\}$ 满足:

$$\|\nabla_{\mathbf{u}} f_{\mu_{1,k}}(\mathbf{u}_k, \mathbf{v}) - \nabla_{\mathbf{u}} f(\mathbf{u}_k, \mathbf{v})\|^2 \leq \frac{L_u^2 \mu_{1,k}^2 d_u^2}{4}, \|\nabla_{\mathbf{v}} f_{\mu_{2,k}}(\mathbf{u}, \mathbf{v}_k) - \nabla_{\mathbf{v}} f(\mathbf{u}, \mathbf{v}_k)\|^2 \leq \frac{L_v^2 \mu_{2,k}^2 d_v^2}{4};$$

4) 对于 ZO-OGDA 算法产生的迭代序列 $\{(\mathbf{u}_k, \mathbf{v}_k)\}$ 满足:

$$\begin{aligned} \mathbb{E} \left[\left\| \frac{d_{\mathbf{u}} [f(\mathbf{u}_k + \mu_{1,k} \mathbf{p}_i, \mathbf{v}) - f(\mathbf{u}_k, \mathbf{v})]}{\mu_{1,k}} \mathbf{p}_i \right\|^2 \right] &\leq 2d_{\mathbf{u}} \|\nabla_{\mathbf{u}} f(\mathbf{u}_k, \mathbf{v})\|^2 + \frac{\mu_{1,k}^2 d_{\mathbf{u}}^2 L_u^2}{2}, i=1, \dots, m_{1,k}, \\ \mathbb{E} \left[\left\| \frac{d_{\mathbf{v}} [f(\mathbf{u}, \mathbf{v}_k + \mu_{2,k} \mathbf{q}_i) - f(\mathbf{u}, \mathbf{v}_k)]}{\mu_{2,k}} \mathbf{q}_i \right\|^2 \right] &\leq 2d_{\mathbf{v}} \|\nabla_{\mathbf{v}} f(\mathbf{u}, \mathbf{v}_k)\|^2 + \frac{\mu_{2,k}^2 d_{\mathbf{v}}^2 L_v^2}{2}, i=1, \dots, m_{2,k}. \end{aligned}$$

引理 2^[10] 令 $\{(\mathbf{u}_k, \mathbf{v}_k)\}$ 是由 ZO-OGDA 算法产生的迭代序列。若假设 3 成立, 则对于 $k \geq 1$, 有:

$$\mathbb{E}[\|\hat{\nabla}_{\mathbf{u}} f(\mathbf{u}_k, \mathbf{v}) - \nabla_{\mathbf{u}} f_{\mu_{1,k}}(\mathbf{u}_k, \mathbf{v})\|^2] \leq \psi_{\mathbf{u},k}^2, \mathbb{E}[\|\hat{\nabla}_{\mathbf{v}} f(\mathbf{u}, \mathbf{v}_k) - \nabla_{\mathbf{v}} f_{\mu_{2,k}}(\mathbf{u}, \mathbf{v}_k)\|^2] \leq \psi_{\mathbf{v},k}^2,$$

其中: $\psi_{\mathbf{u},k}^2 = \frac{1}{m_{1,k}} \left(2d_{\mathbf{u}} G^2 + \frac{\mu_{1,k}^2 d_{\mathbf{u}}^2 L_u^2}{2} \right)$ 和 $\psi_{\mathbf{v},k}^2 = \frac{1}{m_{2,k}} \left(2d_{\mathbf{v}} G^2 + \frac{\mu_{2,k}^2 d_{\mathbf{v}}^2 L_v^2}{2} \right)$ 。

引理 3 令 $\{\mathbf{w}_k\}$ 是由 ZO-OGDA 算法产生的迭代序列, 其中 $\mathbf{w}_k = [\mathbf{u}_k; \mathbf{v}_k]$, 则对任意向量 $\mathbf{w} \in \mathcal{U} \times \mathcal{V}$ 和 $k \geq 1$ 有:

$$\hat{\nabla} F(\mathbf{w}_{k+1})^\top (\mathbf{w}_{k+1} - \mathbf{w}) \leq \frac{1}{2\eta} \|\mathbf{w}_k - \mathbf{w}\|^2 - \frac{1}{2\eta} \|\mathbf{w}_{k+1} - \mathbf{w}\|^2 - \frac{1}{2\eta} \|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2 + \frac{1}{\eta} \xi_k^\top (\mathbf{w}_{k+1} - \mathbf{w}).$$

证明 由(7)式的最优性条件可知,对任意 $\mathbf{w} \in \mathcal{U} \times \mathcal{V}$ 有:

$$\langle \mathbf{w}_{k+1} - (\mathbf{w}_k - \eta \hat{\nabla} F(\mathbf{w}_{k+1}) + \boldsymbol{\xi}_k), \mathbf{w} - \mathbf{w}_{k+1} \rangle \geq 0,$$

整理上式并由(8)式可得:

$$\eta \hat{\nabla} F(\mathbf{w}_{k+1})^\top (\mathbf{w}_{k+1} - \mathbf{w}) \leq \langle \mathbf{w}_k - \mathbf{w}_{k+1}, \mathbf{w}_{k+1} - \mathbf{w} \rangle + \langle \boldsymbol{\xi}_k, \mathbf{w}_{k+1} - \mathbf{w} \rangle,$$

然后再利用等式 $\frac{1}{\eta} \langle \mathbf{w}_k - \mathbf{w}_{k+1}, \mathbf{w}_{k+1} - \mathbf{w} \rangle = \frac{1}{2\eta} \|\mathbf{w}_k - \mathbf{w}\|^2 - \frac{1}{2\eta} \|\mathbf{w}_{k+1} - \mathbf{w}\|^2 - \frac{1}{2\eta} \|\mathbf{w}_k - \mathbf{w}_{k+1}\|^2$, 即可证得结论。

证毕

引理 4 若假设 1、假设 2 和假设 3 成立。令 $\{\mathbf{w}_k\}$ 是由 ZO-OGDA 算法产生的迭代序列 $\mathbf{w}_k = [\mathbf{u}_k; \mathbf{v}_k]$, 则有:

$$\begin{aligned} \mathbb{E}[f(\hat{\mathbf{u}}_K, \mathbf{v}) - f(\mathbf{u}, \hat{\mathbf{v}}_K)] &\leq \mathbb{E}\left[\frac{1}{K} \sum_{k=1}^K \hat{\nabla} F(\mathbf{w}_k)^\top (\mathbf{w}_k - \mathbf{w})\right] + \\ &\frac{11\eta}{2K} \sum_{k=1}^K (\psi_{\mathbf{u},k-2}^2 + \psi_{\mathbf{v},k-2}^2) + \frac{1}{K} \sum_{k=1}^K (L_u \mu_{1,k}^2 + L_v \mu_{2,k}^2), \end{aligned}$$

其中: $\hat{\mathbf{u}}_K = \frac{1}{K} \sum_{k=1}^K \mathbf{u}_k$, $\hat{\mathbf{v}}_K = \frac{1}{K} \sum_{k=1}^K \mathbf{v}_k$, $\psi_{\mathbf{u},k}^2$ 和 $\psi_{\mathbf{v},k}^2$ 由引理 2 给出。

证明 令 $\tilde{\mathbf{w}}_k = P_{\mathcal{U} \times \mathcal{V}}\{\mathbf{w}_{k-1} - 2\eta \nabla F_\mu(\mathbf{w}_{k-1}) + \eta \nabla F_\mu(\mathbf{w}_{k-2})\}$ 。由(3)式可知 $\mathbb{E}[\hat{\nabla} F(\mathbf{w}_k)] = \nabla F_\mu(\mathbf{w}_k)$, 从而有:

$$\begin{aligned} &\mathbb{E}\left[\frac{1}{K} \sum_{k=1}^K \hat{\nabla} F(\mathbf{w}_k)^\top (\mathbf{w}_k - \mathbf{w})\right] = \\ &\mathbb{E}\left[\frac{1}{K} \sum_{k=1}^K (\hat{\nabla} F(\mathbf{w}_k) - \nabla F_\mu(\mathbf{w}_k))^\top (\mathbf{w}_k - \tilde{\mathbf{w}}_k)\right] + \mathbb{E}\left[\frac{1}{K} \sum_{k=1}^K \nabla F_\mu(\mathbf{w}_k)^\top (\mathbf{w}_k - \mathbf{w})\right]. \end{aligned} \quad (9)$$

首先估计(9)式右边的第一项,有:

$$\begin{aligned} &\mathbb{E}\left[\frac{1}{K} \sum_{k=1}^K (\hat{\nabla} F(\mathbf{w}_k) - \nabla F_\mu(\mathbf{w}_k))^\top (\mathbf{w}_k - \tilde{\mathbf{w}}_k)\right] \geq \\ &-\mathbb{E}\left[\frac{\eta}{2K} \sum_{k=1}^K \|\hat{\nabla} F(\mathbf{w}_k) - \nabla F_\mu(\mathbf{w}_k)\|^2\right] - \mathbb{E}\left[\frac{1}{2\eta K} \sum_{k=1}^K \|\mathbf{w}_k - \tilde{\mathbf{w}}_k\|^2\right] \geq \\ &-\frac{\eta}{2K} \sum_{k=1}^K (\psi_{\mathbf{u},k}^2 + \psi_{\mathbf{v},k}^2) - \frac{1}{2\eta K} \mathbb{E}\left[\sum_{k=1}^K \|(2\eta(\nabla F_\mu(\mathbf{w}_{k-1}) - \hat{\nabla} F(\mathbf{w}_{k-1})) + \eta(\hat{\nabla} F(\mathbf{w}_{k-2}) - \nabla F_\mu(\mathbf{w}_{k-2})))\|^2\right] \geq \\ &-\frac{\eta}{2K} \sum_{k=1}^K (\psi_{\mathbf{u},k}^2 + \psi_{\mathbf{v},k}^2) - \mathbb{E}\left[\frac{4\eta}{K} \sum_{k=1}^K \|\nabla F_\mu(\mathbf{w}_{k-1}) - \hat{\nabla} F(\mathbf{w}_{k-1})\|^2\right] - \mathbb{E}\left[\frac{\eta}{K} \sum_{k=1}^K \|\hat{\nabla} F(\mathbf{w}_{k-2}) - \nabla F_\mu(\mathbf{w}_{k-2})\|^2\right] \geq \\ &-\frac{\eta}{2K} \sum_{k=1}^K (\psi_{\mathbf{u},k}^2 + \psi_{\mathbf{v},k}^2) - \frac{4\eta}{K} \sum_{k=1}^K (\psi_{\mathbf{u},k-1}^2 + \psi_{\mathbf{v},k-1}^2) - \frac{\eta}{K} \sum_{k=1}^K (\psi_{\mathbf{u},k-2}^2 + \psi_{\mathbf{v},k-2}^2) \geq -\frac{11\eta}{2K} \sum_{k=1}^K (\psi_{\mathbf{u},k-2}^2 + \psi_{\mathbf{v},k-2}^2). \end{aligned} \quad (10)$$

其中第一个不等式用到了 Young 不等式,第二个不等式用到了引理 2 和 Euclidean 投影算子 $P_{\mathcal{U}}$ 和 $P_{\mathcal{V}}$ 的非扩张性,第三个不等式用到了 $(a+b)^2 \leq 2a^2 + 2b^2$,再由步骤 0 中的参数设置可知 $\psi_{\mathbf{u},k}^2$ 和 $\psi_{\mathbf{v},k}^2$ 都是非负递减序列,由此得到最后一个不等式。

下面估计(9)式右边的第二项,有:

$$\begin{aligned} &\mathbb{E}\left[\frac{1}{K} \sum_{k=1}^K \nabla F_\mu(\mathbf{w}_k)^\top (\mathbf{w}_k - \mathbf{w})\right] \geq \mathbb{E}\left[\frac{1}{K} \sum_{k=1}^K (f_{\mu_1,k}(\mathbf{u}_k, \mathbf{v}_k) - f_{\mu_1,k}(\mathbf{u}, \mathbf{v}_k) + f_{\mu_2,k}(\mathbf{u}_k, \mathbf{v}) - f_{\mu_2,k}(\mathbf{u}_k, \mathbf{v}_k))\right] \geq \\ &\mathbb{E}\left[\frac{1}{K} \sum_{k=1}^K \left(f(\mathbf{u}_k, \mathbf{v}) - f(\mathbf{u}, \mathbf{v}_k) - \frac{L_v \mu_{2,k}^2 + L_u \mu_{1,k}^2}{2}\right)\right] + \mathbb{E}\left[\frac{1}{K} \sum_{k=1}^K \left(f(\mathbf{u}_k, \mathbf{v}_k) - f(\mathbf{u}_k, \mathbf{v}) - \frac{L_v \mu_{2,k}^2 + L_u \mu_{1,k}^2}{2}\right)\right] \geq \\ &\mathbb{E}[f(\hat{\mathbf{u}}_K, \mathbf{v}) - f(\mathbf{u}, \hat{\mathbf{v}}_K)] - \frac{1}{K} \sum_{k=1}^K (L_v \mu_{2,k}^2 + L_u \mu_{1,k}^2). \end{aligned} \quad (11)$$

其中第一个和第三个不等式分别用到了 f 关于 (\mathbf{u}, \mathbf{v}) 的凸-凹性,第二个不等式用到了引理 1 的结论 1)。最后将(10)和(11)式代入到(9)式并整理可得结论。

证毕

引理 5 若假设 1、假设 2、假设 3 和假设 4 成立。设置步长 η 满足 $0 < \eta \leq 1/(L^2 + 3)$, 迭代序列 $\{\mathbf{w}_k\}$ 由 ZO-OGDA 算法产生, 则有 $\{\mathbf{w}_k\}_{k=1}^K \subseteq \mathcal{D}_K$, 其中:

$$\mathcal{D}_K := \left\{ \mathbf{w} \mid \mathbb{E}[\|\mathbf{w} - \mathbf{w}^*\|^2] \leq \frac{L^2 + 3}{L^2} \mathbb{E}[\|\mathbf{w}_0 - \mathbf{w}^*\|^2] + \frac{2L^2 + 17}{L^2(L^2 + 3)} \sum_{k=0}^K (\psi_{u,k-1}^2 + \psi_{v,k-1}^2) + \frac{1}{2} \sum_{k=0}^K [(1 + d_u^2) \mu_{1,k-1}^2 + (1 + d_v^2) \mu_{2,k-1}^2] \right\},$$

这里 $\mathbf{w}^* = [\mathbf{u}^*; \mathbf{v}^*] \in \mathcal{Z}^*$ 。

证明 由引理 3 及(8)式可得:

$$\begin{aligned} \mathbb{E}[\hat{\nabla}F(\mathbf{w}_{k+1})^\top(\mathbf{w}_{k+1} - \mathbf{w})] &\leq \frac{1}{2\eta} \mathbb{E}[\|\mathbf{w}_k - \mathbf{w}\|^2] - \frac{1}{2\eta} \mathbb{E}[\|\mathbf{w}_{k+1} - \mathbf{w}\|^2] - \frac{1}{2\eta} \mathbb{E}[\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2] + \\ &\mathbb{E}[(\hat{\nabla}F(\mathbf{w}_{k+1}) - \hat{\nabla}F(\mathbf{w}_k))^\top(\mathbf{w}_{k+1} - \mathbf{w})] - \mathbb{E}[(\hat{\nabla}F(\mathbf{w}_k) - \hat{\nabla}F(\mathbf{w}_{k-1}))^\top(\mathbf{w}_{k+1} - \mathbf{w})]. \end{aligned}$$

将上式右边加减项 $\mathbb{E}[(\hat{\nabla}F(\mathbf{w}_k) - \hat{\nabla}F(\mathbf{w}_{k-1}))^\top(\mathbf{w}_k - \mathbf{w})]$ 有:

$$\begin{aligned} \mathbb{E}[\hat{\nabla}F(\mathbf{w}_{k+1})^\top(\mathbf{w}_{k+1} - \mathbf{w})] &\leq \frac{1}{2\eta} \mathbb{E}[\|\mathbf{w}_k - \mathbf{w}\|^2] - \frac{1}{2\eta} \mathbb{E}[\|\mathbf{w}_{k+1} - \mathbf{w}\|^2] - \frac{1}{2\eta} \mathbb{E}[\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2] + \\ &\mathbb{E}[(\hat{\nabla}F(\mathbf{w}_{k+1}) - \hat{\nabla}F(\mathbf{w}_k))^\top(\mathbf{w}_{k+1} - \mathbf{w})] - \mathbb{E}[(\hat{\nabla}F(\mathbf{w}_k) - \hat{\nabla}F(\mathbf{w}_{k-1}))^\top(\mathbf{w}_k - \mathbf{w})] + \\ &\mathbb{E}[(\hat{\nabla}F(\mathbf{w}_k) - \hat{\nabla}F(\mathbf{w}_{k-1}))^\top(\mathbf{w}_k - \mathbf{w}_{k+1})]. \end{aligned} \quad (12)$$

首先估计(12)式中的项 $\mathbb{E}[(\hat{\nabla}F(\mathbf{w}_k) - \hat{\nabla}F(\mathbf{w}_{k-1}))^\top(\mathbf{w}_k - \mathbf{w}_{k+1})]$, 有:

$$\begin{aligned} \mathbb{E}[(\hat{\nabla}F(\mathbf{w}_k) - \hat{\nabla}F(\mathbf{w}_{k-1}))^\top(\mathbf{w}_k - \mathbf{w}_{k+1})] &= \\ \mathbb{E}[(\hat{\nabla}F(\mathbf{w}_k) - \nabla F_{\mu}(\mathbf{w}_k) + \nabla F_{\mu}(\mathbf{w}_k) - \nabla F_{\mu}(\mathbf{w}_{k-1}) + \nabla F_{\mu}(\mathbf{w}_{k-1}) - \hat{\nabla}F(\mathbf{w}_{k-1}))^\top(\mathbf{w}_k - \mathbf{w}_{k+1})] &\leq \\ \mathbb{E}\left[\frac{1}{2} \|\hat{\nabla}F(\mathbf{w}_k) - \nabla F_{\mu}(\mathbf{w}_k)\|^2 + \frac{1}{2} \|\nabla F_{\mu}(\mathbf{w}_{k-1}) - \hat{\nabla}F(\mathbf{w}_{k-1})\|^2 + \frac{1}{2} \|\nabla F_{\mu}(\mathbf{w}_k) - \nabla F_{\mu}(\mathbf{w}_{k-1})\|^2 + \frac{3}{2} \|\mathbf{w}_k - \mathbf{w}_{k+1}\|^2\right] &\leq \\ (\psi_{u,k-1}^2 + \psi_{v,k-1}^2) + \frac{3}{2} \mathbb{E}[\|\mathbf{w}_k - \mathbf{w}_{k+1}\|^2] + \frac{1}{2} \mathbb{E}[\|\nabla F_{\mu}(\mathbf{w}_k) - \nabla F_{\mu}(\mathbf{w}_{k-1})\|^2]. \end{aligned} \quad (13)$$

其中第一个不等式用到了 $\mathbf{a}^\top \mathbf{b} \leq \frac{\|\mathbf{a}\|^2}{2} + \frac{\|\mathbf{b}\|^2}{2}$, 第二个不等式用到了引理 2 以及序列 $\{\psi_{u,k}^2\}$ 和 $\{\psi_{v,k}^2\}$ 的非负递减性。注意到:

$$\begin{aligned} \mathbb{E}[\|\nabla_u f_{\mu_{1,k}}(\mathbf{u}_k, \mathbf{v}_k) - \nabla_u f_{\mu_{1,k-1}}(\mathbf{u}_{k-1}, \mathbf{v}_{k-1})\|^2] &\leq \\ 4\mathbb{E}[\|\nabla_u f_{\mu_{1,k}}(\mathbf{u}_k, \mathbf{v}_k) - \nabla_u f_{\mu_{1,k}}(\mathbf{u}_k, \mathbf{v}_{k-1})\|^2 + \|\nabla_u f_{\mu_{1,k}}(\mathbf{u}_k, \mathbf{v}_{k-1}) - \nabla_u f(\mathbf{u}_k, \mathbf{v}_{k-1})\|^2 + \\ \|\nabla_u f(\mathbf{u}_k, \mathbf{v}_{k-1}) - \nabla_u f(\mathbf{u}_{k-1}, \mathbf{v}_{k-1})\|^2 + \|\nabla_u f(\mathbf{u}_{k-1}, \mathbf{v}_{k-1}) - \nabla_u f_{\mu_{1,k-1}}(\mathbf{u}_{k-1}, \mathbf{v}_{k-1})\|^2] &\leq \\ 4L_u^2 \mathbb{E}[\|\mathbf{v}_k - \mathbf{v}_{k-1}\|^2] + L_u^2 \mu_{1,k}^2 d_u^2 + 4L_u^2 \mathbb{E}[\|\mathbf{u}_k - \mathbf{u}_{k-1}\|^2] + L_u^2 \mu_{1,k-1}^2 d_u^2 &\leq \frac{L^2}{2} \mathbb{E}[\|\mathbf{w}_k - \mathbf{w}_{k-1}\|^2] + \frac{L^2 \mu_{1,k-1}^2 d_u^2}{2}, \end{aligned}$$

其中第二个不等式用到了引理 1 的 2) 和 3) 以及 $\{\mu_{1,k}\}$ 的单调递减性。同理可得:

$$\mathbb{E}[\|\nabla_v f_{\mu_{2,k}}(\mathbf{u}_k, \mathbf{v}_k) - \nabla_v f_{\mu_{2,k-1}}(\mathbf{u}_{k-1}, \mathbf{v}_{k-1})\|^2] \leq \frac{L^2}{2} \|\mathbf{w}_k - \mathbf{w}_{k-1}\|^2 + \frac{L^2 \mu_{2,k-1}^2 d_v^2}{2}.$$

由上面两个不等式可得(13)式右边最后一项的估计, 如下:

$$\mathbb{E}[\|\nabla F_{\mu}(\mathbf{w}_k) - \nabla F_{\mu}(\mathbf{w}_{k-1})\|^2] \leq L^2 \|\mathbf{w}_k - \mathbf{w}_{k-1}\|^2 + \frac{L^2 \mu_{1,k-1}^2 d_u^2 + L^2 \mu_{2,k-1}^2 d_v^2}{2}. \quad (14)$$

将(13)、(14)式代入(12)式, 并利用引理的条件 $0 < \eta < 1/(L^2 + 3)$, 整理可得:

$$\begin{aligned} \mathbb{E}[\hat{\nabla}F(\mathbf{w}_{k+1})^\top(\mathbf{w}_{k+1} - \mathbf{w})] &\leq \\ \frac{1}{2\eta} \mathbb{E}[\|\mathbf{w}_k - \mathbf{w}\|^2] - \frac{1}{2\eta} \mathbb{E}[\|\mathbf{w}_{k+1} - \mathbf{w}\|^2] - \frac{L^2}{2} \mathbb{E}[\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2] + \frac{L^2}{2} \mathbb{E}[\|\mathbf{w}_k - \mathbf{w}_{k-1}\|^2] + \\ \mathbb{E}[(\hat{\nabla}F(\mathbf{w}_{k+1}) - \hat{\nabla}F(\mathbf{w}_k))^\top(\mathbf{w}_{k+1} - \mathbf{w})] - \mathbb{E}[(\hat{\nabla}F(\mathbf{w}_k) - \hat{\nabla}F(\mathbf{w}_{k-1}))^\top(\mathbf{w}_k - \mathbf{w})] + \\ (\psi_{u,k-1}^2 + \psi_{v,k-1}^2) + \frac{L^2 \mu_{1,k-1}^2 d_u^2 + L^2 \mu_{2,k-1}^2 d_v^2}{4}. \end{aligned}$$

对上式将 k 从 0 到 $K-1$ 相加, 并注意到 $\mathbf{w}_0 = \mathbf{w}_{-1}$, 有:

$$\begin{aligned} & \sum_{k=0}^{K-1} \mathbb{E}[\hat{\nabla} F(\mathbf{w}_{k+1})^\top (\mathbf{w}_{k+1} - \mathbf{w})] \leq \\ & \frac{1}{2\eta} \mathbb{E}[\|\mathbf{w}_0 - \mathbf{w}\|^2] - \frac{1}{2\eta} \mathbb{E}[\|\mathbf{w}_K - \mathbf{w}\|^2] - \frac{L^2}{2} \mathbb{E}[\|\mathbf{w}_K - \mathbf{w}_{K-1}\|^2] + \\ & \sum_{k=0}^{K-1} (\psi_{u,k-1}^2 + \psi_{v,k-1}^2) + \sum_{k=0}^{K-1} \frac{L^2 \mu_{1,k-1}^2 d_u^2 + L^2 \mu_{2,k-1}^2 d_v^2}{4} + \mathbb{E}[(\hat{\nabla} F(\mathbf{w}_K) - \hat{\nabla} F(\mathbf{w}_{K-1}))^\top (\mathbf{w}_K - \mathbf{w})], \end{aligned} \quad (15)$$

类似于(13)式, 估计(15)式右边的最后一项, 如下:

$$\begin{aligned} & \mathbb{E}[(\hat{\nabla} F(\mathbf{w}_K) - \hat{\nabla} F(\mathbf{w}_{K-1}))^\top (\mathbf{w}_K - \mathbf{w})] \leq \\ & (\psi_{u,K-1}^2 + \psi_{v,K-1}^2) + \frac{L^2}{2} \mathbb{E}[\|\mathbf{w}_K - \mathbf{w}_{K-1}\|^2] + \frac{3}{2} \mathbb{E}[\|\mathbf{w}_K - \mathbf{w}\|^2] + \frac{L^2 \mu_{1,K-1}^2 d_u^2 + L^2 \mu_{2,K-1}^2 d_v^2}{4}. \end{aligned}$$

将上式代入(15)式并整理可得:

$$\begin{aligned} & \sum_{k=0}^{K-1} \mathbb{E}[\hat{\nabla} F(\mathbf{w}_{k+1})^\top (\mathbf{w}_{k+1} - \mathbf{w})] \leq \\ & \frac{1}{2\eta} \mathbb{E}[\|\mathbf{w}_0 - \mathbf{w}\|^2] - \left(\frac{1}{2\eta} - \frac{3}{2}\right) \mathbb{E}[\|\mathbf{w}_K - \mathbf{w}\|^2] + \sum_{k=0}^{K-1} (\psi_{u,k-1}^2 + \psi_{v,k-1}^2) + \sum_{k=0}^{K-1} \frac{L^2 \mu_{1,k-1}^2 d_u^2 + L^2 \mu_{2,k-1}^2 d_v^2}{4}, \end{aligned} \quad (16)$$

再由引理 4 的证明过程可得:

$$\begin{aligned} & \sum_{k=1}^K \mathbb{E}[f(\mathbf{u}_k, \mathbf{v}) - f(\mathbf{u}, \mathbf{v}_k)] \leq \sum_{k=0}^{K-1} \mathbb{E}[\hat{\nabla} F(\mathbf{w}_{k+1})^\top (\mathbf{w}_{k+1} - \mathbf{w})] + \\ & \frac{11\eta}{2} \sum_{k=1}^K (\psi_{u,k-2}^2 + \psi_{v,k-2}^2) + \sum_{k=1}^K (L_u \mu_{1,k}^2 + L_v \mu_{2,k}^2), \end{aligned} \quad (17)$$

结合(16)和(17)式, 整理可得:

$$\begin{aligned} & \sum_{k=1}^K \mathbb{E}[f(\mathbf{u}_k, \mathbf{v}) - f(\mathbf{u}, \mathbf{v}_k)] \leq \frac{1}{2\eta} \mathbb{E}[\|\mathbf{w}_0 - \mathbf{w}\|^2] - \left(\frac{1}{2\eta} - \frac{3}{2}\right) \mathbb{E}[\|\mathbf{w}_K - \mathbf{w}\|^2] + \\ & \frac{2 + 11\eta}{2} \sum_{k=0}^{K-1} (\psi_{u,k-1}^2 + \psi_{v,k-1}^2) + \frac{L^2}{4} \sum_{k=0}^{K-1} [(1 + d_u^2) \mu_{1,k-1}^2 + (1 + d_v^2) \mu_{2,k-1}^2]. \end{aligned} \quad (18)$$

在(18)式中令 $\mathbf{u} = \mathbf{u}^*$, $\mathbf{v} = \mathbf{v}^*$, 并注意到 $f(\mathbf{u}_k, \mathbf{v}^*) - f^* + f^* - f(\mathbf{u}^*, \mathbf{v}_k) \geq 0, \forall k \geq 0$, 可得:

$$\begin{aligned} & \frac{1 - 3\eta}{2\eta} \mathbb{E}[\|\mathbf{w}_K - \mathbf{w}^*\|^2] \leq \frac{1}{2\eta} \mathbb{E}[\|\mathbf{w}_0 - \mathbf{w}^*\|^2] + \frac{2 + 11\eta}{2} \sum_{k=0}^{K-1} (\psi_{u,k-1}^2 + \psi_{v,k-1}^2) + \\ & \frac{L^2}{4} \sum_{k=0}^{K-1} [(1 + d_u^2) \mu_{1,k-1}^2 + (1 + d_v^2) \mu_{2,k-1}^2]. \end{aligned}$$

利用条件 $0 < \eta \leq 1/(L^2 + 3)$, 将上式整理可得结论. 证毕

基于以上假设和引理, 下面将给出 ZO-OGDA 算法具体的迭代复杂度结果.

定理 1 若假设 1~4 成立. 在 ZO-OGDA 算法, 令迭代步长 η 满足 $0 < \eta \leq 1/(L^2 + 3)$, 设置采样次数序列 $\{m_{1,k}\}, \{m_{2,k}\}$ 和平滑化参数序列满足:

$$\begin{aligned} & m_{1,k} := \mathcal{O}(d_u(k+3)\epsilon^{-1}), m_{2,k} := \mathcal{O}(d_v(k+3)\epsilon^{-1}), \\ & \mu_{1,k} := \mathcal{O}((d_u(k+3))^{1/2})^{-1} \epsilon^{1/2}, \mu_{2,k} := \mathcal{O}((d_v(k+3))^{1/2})^{-1} \epsilon^{1/2}, \end{aligned} \quad (19)$$

设迭代序列 $\{(\mathbf{u}_k, \mathbf{v}_k)\}$ 是由 ZO-OGDA 算法产生, 则当迭代次数 K 至多为 $\mathcal{O}(\epsilon^{-1})$ 时, $(\hat{\mathbf{u}}_K, \hat{\mathbf{v}}_K)$ 是问题(1)的 ϵ -稳

定点, 其中 $\hat{\mathbf{u}}_K = \frac{1}{K} \sum_{k=1}^K \mathbf{u}_k, \hat{\mathbf{v}}_K = \frac{1}{K} \sum_{k=1}^K \mathbf{v}_k$.

证明 令 $f^* = f(\mathbf{u}^*, \mathbf{v}^*)$, $D = \|\mathbf{w}_0 - \mathbf{w}^*\|^2$. 由引理 5, 对任意 $\mathbf{w} = [\mathbf{u}; \mathbf{v}] \in \mathcal{D}_K$ 有:

$$\begin{aligned} & \mathbb{E}[\|\mathbf{w}_0 - \mathbf{w}\|^2] \leq 2\mathbb{E}[\|\mathbf{w}_0 - \mathbf{w}^*\|^2] + 2\mathbb{E}[\|\mathbf{w} - \mathbf{w}^*\|^2] \leq \\ & \frac{4L^2 + 6}{L^2} D + \frac{4L^2 + 34}{L^2(L^2 + 3)} \sum_{k=0}^{K-1} (\psi_{u,k-1}^2 + \psi_{v,k-1}^2) + \sum_{k=0}^{K-1} [(1 + d_u^2) \mu_{1,k-1}^2 + (1 + d_v^2) \mu_{2,k-1}^2]. \end{aligned}$$

将上式代入到(18)式中, 两边同乘 $1/K$, 并利用 f 关于变量 (\mathbf{u}, \mathbf{v}) 的凸-凹性, 对任意 $\mathbf{w} = [\mathbf{u}; \mathbf{v}] \in \mathcal{D}_K$ 有:

$$\begin{aligned} & \mathbb{E} \left[\max_{\mathbf{v}: (\hat{\mathbf{u}}_K, \mathbf{v}) \in \mathcal{D}_K} f(\hat{\mathbf{u}}_K, \mathbf{v}) - f^* \right] + \mathbb{E} \left[f^* - \min_{\mathbf{u}: (\mathbf{u}, \hat{\mathbf{v}}_K) \in \mathcal{D}_K} f(\mathbf{u}, \hat{\mathbf{v}}_K) \right] \leq \\ & \frac{2L^2 + 3}{K\eta L^2} D + \frac{1}{K} \left(\frac{2L^2 + 17}{\eta L^2 (L^2 + 3)} + \frac{2L^2 + 17}{2(L^2 + 3)} \right) \sum_{k=0}^K (\psi_{\mathbf{u}, k-1}^2 + \psi_{\mathbf{v}, k-1}^2) + \\ & \frac{1}{2K} \left(\frac{1}{\eta} + \frac{L^2}{2} \right) \sum_{k=0}^K [(1 + d_{\mathbf{u}}^2) \mu_{1, k-1}^2 + (1 + d_{\mathbf{v}}^2) \mu_{2, k-1}^2]. \end{aligned} \quad (20)$$

由于:

$$\begin{aligned} & \mathbb{E} \left[\max_{\mathbf{v}: (\hat{\mathbf{u}}_K, \mathbf{v}) \in \mathcal{D}_K} f(\hat{\mathbf{u}}_K, \mathbf{v}) \right] \geq \mathbb{E} [f(\hat{\mathbf{u}}_K, \mathbf{v}^*)] \geq \mathbb{E} [f(\mathbf{u}^*, \mathbf{v}^*)], \\ & \mathbb{E} \left[\min_{\mathbf{u}: (\mathbf{u}, \hat{\mathbf{v}}_K) \in \mathcal{D}_K} f(\mathbf{u}, \hat{\mathbf{v}}_K) \right] \leq \mathbb{E} [f(\mathbf{u}^*, \hat{\mathbf{v}}_K)] \leq \mathbb{E} [f(\mathbf{u}^*, \mathbf{v}^*)], \end{aligned}$$

所以 $\mathbb{E} \left[\max_{\mathbf{v}: (\hat{\mathbf{u}}_K, \mathbf{v}) \in \mathcal{D}_K} f(\hat{\mathbf{u}}_K, \mathbf{v}) - f^* \right]$ 和 $\mathbb{E} \left[f^* - \min_{\mathbf{u}: (\mathbf{u}, \hat{\mathbf{v}}_K) \in \mathcal{D}_K} f(\mathbf{u}, \hat{\mathbf{v}}_K) \right]$ 均为非负, 从而有:

$$\mathbb{E} [f(\hat{\mathbf{u}}_K, \hat{\mathbf{v}}_K) - f^*] \leq \mathbb{E} \left[\max_{\mathbf{v}: (\hat{\mathbf{u}}_K, \mathbf{v}) \in \mathcal{D}_K} f(\hat{\mathbf{u}}_K, \mathbf{v}) - f^* \right], \quad \mathbb{E} [f^* - f(\hat{\mathbf{u}}_K, \hat{\mathbf{v}}_K)] \leq \mathbb{E} \left[f^* - \min_{\mathbf{u}: (\mathbf{u}, \hat{\mathbf{v}}_K) \in \mathcal{D}_K} f(\mathbf{u}, \hat{\mathbf{v}}_K) \right], \quad (21)$$

结合(20)和(21)式可得:

$$\begin{aligned} \mathbb{E} [|f(\hat{\mathbf{u}}_K, \hat{\mathbf{v}}_K) - f^*|] & \leq \frac{2L^2 + 3}{K\eta L^2} D + \frac{1}{K(L^2 + 3)} \left(\frac{2L^2 + 17}{\eta L^2} + \frac{2L^2 + 17}{2} \right) \sum_{k=0}^K (\psi_{\mathbf{u}, k-1}^2 + \psi_{\mathbf{v}, k-1}^2) + \\ & \frac{1}{2K} \left(\frac{1}{\eta} + \frac{L^2}{2} \right) \sum_{k=0}^K [(1 + d_{\mathbf{u}}^2) \mu_{1, k-1}^2 + (1 + d_{\mathbf{v}}^2) \mu_{2, k-1}^2], \end{aligned}$$

最后由 $\frac{1}{K} \sum_{k=2}^{K+2} \frac{1}{k} \leq \frac{1}{K} \int_1^{K+2} \frac{1}{k} dk = \frac{\ln(K+2)}{K} \leq 1 (K \geq 2)$ 及(19)式, 结合上式可证得定理 1 结论成立。证毕

注 2 1) 由定理 1 可知, ZO-OGDA 的迭代复杂度为 $K = \mathcal{O}(\epsilon^{-1})$, 表明该算法在遍历平均意义下达到 ϵ -稳定点所需要的迭代次数 K 至多为 $\mathcal{O}(\epsilon^{-1})$ 。且函数值的总调用次数为:

$$K_{\text{ZO}} = \sum_{k=1}^K (m_{1,k} + m_{2,k}) \sim \mathcal{O}((d_{\mathbf{u}} + d_{\mathbf{v}})(\epsilon^{-2} + \epsilon^{-3})).$$

2) 通过对比, 本文提出的 ZO-OGDA 算法的迭代复杂度与 OGDA 算法^[6]是同阶的, 但本文提出的 ZO-OGDA 算法不需要计算梯度信息。

2 数值实验

考虑矩阵博弈问题^[12], 即两个玩家 $\mathbf{x} \in \mathbf{R}^{d_x}$ 和 $\mathbf{y} \in \mathbf{R}^{d_y}$, 玩家 \mathbf{y} 的目标是通过做出 1 到 d_y 个动作来获得尽可能大的增益; 而玩家 \mathbf{x} 的目标则是通过做出的 1 到 d_x 个动作, 尽可能的最小化玩家 \mathbf{y} 的增益。该博弈问题可以表示为如下的一个双线性极小极大问题^[6]:

$$\min_{\mathbf{x} \in \mathcal{X}} \max_{\mathbf{y} \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y}) := \langle \mathbf{x}, \mathbf{B}\mathbf{y} \rangle, \quad (22)$$

其中: $\mathcal{X} = \{\mathbf{x} \mid \|\mathbf{x}\| \leq c_x, \mathbf{x} \in \mathbf{R}^{d_x}\}$, $\mathcal{Y} = \{\mathbf{y} \mid \|\mathbf{y}\| \leq c_y, \mathbf{y} \in \mathbf{R}^{d_y}\}$, 稀疏矩阵 $\mathbf{B} \in \mathbf{R}^{d_x \times d_y}$ 。

下面将通过数值实验测试 ZO-OGDA 算法在求解极小极大问题(22)时的数值表现。具体实验环境为, CPU: AMD Ryzen 7 5800H 处理器, 内存: 16 GB。所有算法均在 Matlab 环境下实现。另外, 也将本文提出的算法与 ZO-Min-Max^[7]算法和 ZO-AGP^[10]算法做数值对比。在仿真中, 设置 $d_x = 10\,000$, $d_y = 100$, $c_x = c_y = 1$ 。让稀疏矩阵 $\mathbf{B} \in \mathbf{R}^{d_x \times d_y}$ 的生成与文献[6]类似, 即每一项元素非零的概率为 $p = 0.01$, 当该元素确认非零时, 取值独立且均匀的在区间 $[-1, 1]$ 内生成。为了对比, 设置 3 种算法 ZO-OGDA 算法、ZO-Min-Max 算法和 ZO-AGP 算法的函数值采样次数相同 $m_{1,k} = m_{2,k} = m = 3$, 平滑化参数 $\mu_{1,k} = \mu_{2,k} = \frac{0.1}{(k+3)^{1/2}}$ 。

图 1 中纵轴表示目标函数在迭代点处函数值与最优值的绝对误差, 横轴表示算法的迭代次数, 其中蓝色短实线代表 ZO-OGDA 算法、红色实线代表原一阶 OGDA 算法、黄色实点交叉线代表 ZO-Min-Max 算法、紫色点线为 ZO-AGP 算法。由图 1 可以看出, 随着迭代次数的增加, 4 种算法都能渐进收敛到最优值, 且 OGDA 算法的收敛效果最好。但 OGDA 算法使用了函数的梯度信息。相比于 ZO-Min-Max 算法和 ZO-AGP 算法, ZO-OGDA 算法达到同样的误差所需要的算法迭代次数较少。对比 ZO-OGDA 算法和 OGDA 算法, 虽然 ZO-

OGDA 算法的收敛效果略差,但 ZO-OGDA 算法仅需要函数值信息,而不需要使用梯度信息。

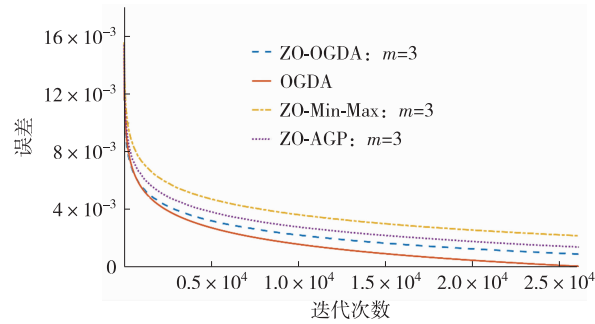


图 1 不同算法的误差对比 ($m=3$)

Fig. 1 Error comparisons using different algorithms with sampling times $m=3$

下面考虑采样次数对 ZO-OGDA 算法收敛性结果的影响,结果见表 1。为了简便,设置 $m_{1,k} = m_{2,k} = m$,稀疏度 $s=0.01$,分两组情形考虑:情形 1: $d_x=1\ 000, d_y=100$;情形 2: $d_x=10\ 000, d_y=100$ 。

表 1 采样次数 m 对 ZO-OGDA 算法收敛性的影响

Tab. 1 Influence of sampling times m on the convergence of ZO-OGDA algorithm

ZO-OGDA		$m=3$	$m=7$	$m=10$	$m=20$
d_x/d_y	ϵ	$k/\text{CPU 时间}$	$k/\text{CPU 时间}$	$k/\text{CPU 时间}$	$k/\text{CPU 时间}$
1 000/100	0.01	17 178/1.8	13 119/2.9	11 470/3.3	7 948/4.4
	0.001	27 490/2.8	22 454/5.0	15 271/6.7	13 850/7.8
10 000/1 000	0.01	35 321/99.3	32 857/234.0	26 914/256.3	22 964/417.3
	0.001	47 919/180.4	44 150/322.7	40 163/467.3	32 967/621.8

表 1 给出了在不同采样次数 m 的设置下,ZO-OGDA 算法要达到不同终止精度(ϵ)所需要的迭代次数(k),以及算法达到终止精度时的 CPU 运行时间的对比结果。其中,第一列代表变量维数 d_x, d_y ,第二列为算法终止时达到的精度 ϵ ,后面几列分别表示采样次数 m 取 3,7,10,20 时 ZO-OGDA 算法的数值实验结果。从表 1 中可以看到,当达到相同的算法精度时,算法所需的迭代次数随着问题的维数增加而增加,对应的运行时间也明显增加。对于两种不同问题维数情形,在达到相同算法精度时,算法所需要的迭代次数随着采样次数 m 的增加而减少,但对应算法的运行时间会增加。其中对比采样次数 $m=3$ 和 $m=20$ 时可以看到,虽 $m=20$ 时算法所需迭代次数最少,但相应的算法运行时间是最多的;采样次数 $m=3$ 时算法所需迭代次数虽最多,但相应的算法运行时间是最少的。这表明 ZO-OGDA 算法需要在采样次数和运算时间上进行权衡。

3 结束语

本文提出了 ZO-OGDA 算法求解凸-凹极小极大问题,并给出了算法的复杂度分析结果。最后,通过数值实验验证了算法有效性。本文提出的零阶优化算法适用于目标函数的梯度难以获取或计算昂贵的情形。

参考文献:

- [1] 徐姿,张慧灵.非凸极小极大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 algorithm and complexity analysis of nonconvex minimax problems[J]. Journal of operations research,2021,25(3):74-86.
- [2] 羊英,钟力炜,罗守成,等.基于博弈的手术综合评价方法[J].重庆师范大学学报(自然科学版),2013,30(1):7-11.
YANG Y,ZHONG L W,LUO S C,et al. Comprehensive evaluation of surgical operations based on game theory[J]. Journal of Chongqing Normal University (Natural Science),2013,30(1):7-11.
- [3] 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):110-118.
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): 110-118.
- [4] LETCHER A, BALDUZZI D, RACANIÈRE S, et al. Differentiable game mechanics[J]. Journal of Machine Learning Research, 2019, 20(84): 1-40.
- [5] LIN T, JIN C, JORDAN M. On gradient descent ascent for nonconvex-concave minimax problems[EB/OL]. (2021-11-16)[2022-09-02]. <https://arxiv.org/abs/1906.00331>.
- [6] MOKHTARI A, OZDAGLAR A, PATTATHIL S. Convergence Rate of $\mathcal{O}(1/k)$ for Optimistic Gradient and Extra-gradient Methods in Smooth Convex-Concave Saddle Point Problems[J]. SIAM Journal on Optimization, 2020, 30(4): 3230-3251.
- [7] LIU S, LU S, CHEN X, et al. Min-Max optimization without gradients; convergence and applications to adversarial ML[EB/OL]. (2020-06-17)[2022-09-02]. <https://arxiv.org/abs/1909.13806>.
- [8] BUBECK S, ELDAN R, LEE Y T. Kernel-based methods for bandit convex optimization[EB/OL]. (2016-07-11)[2022-09-02]. <https://arxiv.org/abs/1607.03084v1>.
- [9] NESTEROV Y. Random gradient-free minimization of convex functions[J]. Foundations of Computational Mathematics, 2017, 17(2): 527-566.
- [10] XU Z, SHEN J J, WANG Z Q, et al. Zeroth-order algorithms randomized gradient projection algorithms for general nonconvex-concave minimax problems[EB/OL]. (2021-08-05)[2022-09-02]. <https://arxiv.org/abs/2108.00473v2>.
- [11] BEZNOSIKOV A, SADIEV A, GASNIKOV A. Gradient-free methods with inexact oracle for convex-concave stochastic saddle-point problem[EB/OL]. (2020-09-09)[2022-09-02]. <https://arxiv.org/abs/2005.05913v3>.
- [12] SADIEV A, BEZNOSIKOV A, DVURECHENSKY P, et al. Zeroth-order algorithms for smooth saddle-point problems[EB/OL]. (2021-02-27)[2022-09-02]. <https://arxiv.org/abs/2009.09908>.
- [13] GAO X, JIANG B, ZHANG S. On the information-adaptive variants of the ADMM: an iteration complexity perspective[J]. Journal of Scientific Computing, 2018, 76: 327-363.

Operations Research and Cybernetics

Zerth-Order Gradient Descent Ascent Algorithm for General Convex Concave Min-Max Problems

XIE Tao, GAO Ruicheng, TONG Yin, LI Jueyou

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

Abstract: [Purposes] The generate adversarial network models in machine learning are proposed by a min-max optimization problem, which has attracted extensive attention of scholars. At present, most of optimization algorithms are designed based on the standard gradient descent ascent algorithm. However, in some applications, the gradient information of the objective function is often computationally expensive or difficult to obtain. [Methods] Therefore, for a class of convex-concave min-max optimization problems, a zero-order optimistic gradient descent ascent algorithm (ZO-OGDA) is proposed by using the information of function values to approximate the gradient information based on a smoothing method. The proposed ZO-OGDA algorithm extends the OGDA algorithm to the gradient-free case. [Findings] Then, based on the convergence analysis theory of the proximal point algorithm with errors, the iteration complexity of the proposed ZO-OGDA algorithm to obtain ϵ -stationary points with order of $O(\epsilon^{-1})$ is obtained. [Conclusions] Finally, numerical experiments on the matrix game model is performed. The numerical results show that the performance of the proposed ZO-OGDA algorithm is similar to that of the OGDA algorithm.

Keywords: convex-concave min-max problem; zeroth-order algorithm; gradient descent ascent algorithm; complexity analysis

(责任编辑 许 甲)