arXiv:2304.02441v1 [math.OC] 5 Apr 2023

复合非凸强凹极小极大问题的去中心化梯度下降最大化方法

YANGYANG XU*

摘要。 近年来，极小极大问题引起了广泛关注。已有一些研究尝试解决去中心化非凸强凹（NCSC）极小极大结构优化问题；然而，这些研究都聚焦于光滑问题，且至多仅在最大化变量上施加约束。本文首次尝试求解复合NCSC极小极大问题，该问题可在最小化变量和最大化变量上均包含凸非光滑项。我们基于一种新颖的去中心化极小极大问题重新表述来设计算法，该表述引入一个乘子以吸收对偶共识约束。移除对偶共识约束后，可以采用最激进的对偶更新方式（即局部最大化，而非梯度上升步骤），从而能够采用更大的原始步长，并获得更好的复杂度结果。此外，将对偶变量上的非光滑性与共识约束解耦，简化了去中心化算法的分析；因此，我们的重新表述为研究者设计求解NCSC极小极大问题的新型（并可能更高效的）去中心化方法开辟了新途径。我们证明了所提算法的全局收敛性，并给出了生成该重新表述的（近似）平稳点的迭代复杂度结果。此外，我们建立了该重新表述与原始表述的（近似）平稳性之间的关系。基于这一关系，我们证明，当对偶正则项为光滑项时，与现有算法相比，我们的算法在生成原始表述的近似驻点时可实现更低的复杂度（对条件数的依赖有所降低）。我们在分布鲁棒逻辑回归问题上进行了数值实验，以展示所提算法的性能。

关键词： 复合极小极大、非凸强凹问题、去中心化算法、迭代复杂度问题

y.

1. 引言。 本文研究极小极大结构问题

$$\min_{\mathbf{x}\in\mathbb{R}^{n_1}} \max_{\mathbf{y}\in\mathbb{R}^{n_2}} f(\mathbf{x},\mathbf{y}) + g(\mathbf{x}) - h(\mathbf{y}), \text{ with } f(\mathbf{x},\mathbf{y}) := \tfrac{1}{m}\textstyle\sum_{i=1}^m f_i(\mathbf{x},\mathbf{y}), \tag{1.1}$$

其中 f_i 是一个非凸强凹（NCSC）可微函数，且对每个 $i = 1,\dots,m$, g和 h都是闭凸函数。极小极大结构问题出现在许多应用中，例如功率控制与收发器设计 [26]，分布鲁棒优化 [42],生成对抗网络 [11],最优传输 [14]，以及强化学习 [5]。

近年来，设计用于求解极小极大结构非凸问题的算法受到越来越多的关注，例如 [12, 13, 15, 18, 20, 21, 23, 34, 36, 40, 44, 46–48]。不同于大多数现有研究，我们关注设计一种去中心化算法来求解（1.1）。假设在一个连通图或网络上有 m 个智能体 $\mathcal{G} = (\mathcal{N}, \mathcal{E})$，其中 $\mathcal{N}$表示节点（或智能体）集合，$\mathcal{E}$ 表示 $\mathcal{G}$的边集合。对每个 $i \in \mathcal{N}$，函数 f_i 由第 个 智能体 i私有持有。为了使这 m 个智能体协作求解（1.1），每个智能体都会保留一份原始变量和对偶变量的副本，并通过共识约束使所有本地副本保持一致。为实现共识，智能体会与其一跳邻居发送和接收本地信息，并进行邻居（加权）平均。

1.1. 极小极大问题的现有去中心化方法。大多数关于求解极小极大结构优化的去中心化方法的现有研究，主要针对（强）凸-（强）凹问题，或带耦合约束的凸多智能体优化的鞍点或变分不等式（VI）重新表述，例如 [2, 8, 9, 16, 17, 29–32, 35, 37]。近期有若干研究探讨了非凸-凹（甚至非凹）极小极大问题，与我们的工作关系更为密切。下面，我们回顾这些关于去中心化极小极大结构非凸优化的研究。

该工作 [41] 研究一类特殊的去中心化极小极大问题，其中一致性约束仅施加于原变量或对偶变量之一，而非同时施加于二者。文中给出了两种去中心化梯度下降上升（D-GDA）方法。然而，其收敛性结果仅针对求解不含对偶共识约束问题的方法建立。当涉及的函数光滑且强

*xuy21@rpi.edu，伦斯勒理工学院数学科学系，特洛伊，纽约州 12180

关于被限制在紧致定义域内的对偶变量为凹函数时， [41] 中的 D-GDA ε可在 $O(\varepsilon^{-2})$ 次外层迭代和通信轮次内得到一个 -驻点解。由于对偶变量不存在共识，且每次外层迭代都会将对偶部分最大化至一定精度，因此该方法可视为一种用于求解带有非精确梯度的非凸光滑最小化问题的去中心化方法。文献 [27] 给出了一种用于求解多智能体强化学习（MARL）的去中心化策略 GDA。该算法基于 MARL 的拉格朗日函数构建，该函数同样不对对偶变量（即拉格朗日乘子）施加共识约束。由于拉格朗日函数仅关于对偶变量为凹函数，文献 [27] 中建立的复杂度结果为 $O(\varepsilon^{-4})$，高于 [41] 中的结果，但对于非凸-凹问题而言，这是目前已知的最佳结果。

受合作式 MARL 的启发， [49] 提出了一种基于梯度跟踪的 D-GDA，用于求解有限和结构的非凸强凹极小极大问题，并在原始变量和对偶变量上都施加共识约束。其梯度计算和通信复杂度均为 $O(\varepsilon^{-2})$，可求得一个ε-驻点解。此外，为降低计算复杂度对分量函数数量的依赖，该方法采用基于动量的方差缩减（VR）技术，以利用有限和结构。 [49] 中的方法需要定期计算完整梯度，而 [10] 中的去中心化方法采用 SAGA 型 VR [7], ，无需定期计算完整梯度，而是维护所有分量梯度的估计值。其计算复杂度和通信复杂度与 [49]中的方法处于同一阶。 [10, 49] 都考虑不带硬约束或非光滑正则项的光滑问题。相比之下， [4] 考虑了稍具一般性的一类非凸强凹极小极大问题，其中对偶变量可以带有凸约束。与 [7, 49], 中的结果相比， [4] 的计算复杂度相近，但其通信复杂度对问题条件数和图拓扑的依赖更优。不过， [4] 中的方法每个计算步骤都需要进行多次通信，因此需要智能体之间进行更多协调。

去中心化非凸非凹（NCNC）极小极大问题已在 [24, 25]中得到研究。文献 [25] 提出了一种基于近端点的方法。该方法首先构建所考虑极小极大问题的最优性条件，其中涉及一个非单调算子 $\mathcal{B}+\mathcal{R}$ ；详情见 [25] 中的式（8）。随后，该方法旨在通过每个智能体处的 $\mathcal{B}+\mathcal{R}$ 的预解算子求解所得系统的根。在假设 $\mathcal{B}+\mathcal{R}$ 弱单调且满足Minty VI条件 [23], 的情况下，文献 [25] 建立了一个 $O(\varepsilon^{-2})$迭代复杂度结果，以生成一个 ε-驻点解。然而， $\mathcal{B}+\mathcal{R}$的预解算子通常计算开销很大，这限制了文献 [25]中算法的应用。文献 [24] 中的方法是对文献 [22], 中乐观随机梯度方法的去中心化扩展，后者是一种随机超梯度方法。与上述方法不同，该方法只需利用 O个样本对每个梯度进行随机近似。然而，其计算复杂度要高得多；为求得一个 ε-驻点解，复杂度达到 $O(\varepsilon^{-12})$。针对包含极小极大问题这一特例的随机VI问题，文献 [1]提出了一种去中心化随机超梯度方法，适用于强单调、单调和非单调VI。对于非单调情形，该方法假设满足Minty VI条件；然而，即使使用确定性梯度，平稳性违背也不会趋于零。Minty VI条件对于NCSC极小极大问题可能不成立（附录中给出了一个简单示例）；因此，现有结果不适用于我们的情形。

1.2. 贡献。 我们的贡献主要有三点。 首先，我们在去中心化设置下，对 (1.1) 给出了一个等价的重新表述，并且只需一个温和的假设：dom(h) 的相对内部非空。该重新表述包含原变量的一致性约束，但通过引入拉格朗日乘子消除了对偶变量的一致性约束。 其次，这一重新表述使我们能够在多个智能体之间无需协调，即可对对偶变量进行局部（近似）最大化。利用这一优势，我们提出了一种新颖的去中心化算法，通过对原变量执行梯度下降，并对对偶变量执行（近似）最大化。该算法每次迭代无需多次通信。此外，局部（近似）最大化可根据可检验的停止条件终止。因此，该算法易于

TABLE 1

Comparison of our decentralized method to a few existing ones for solving NCSC minimax problems in the form of (1.1)。 Because of different settings of problems and algorithms, we specialize the complexity of the compared methods on smoothproblems, i.e., $g \equiv 0$ and $h \equiv 0$, by using deterministic gradients. Here, $\iota_{\mathcal{X}}$ and $\iota_{\mathcal{Y}}$ denote the indicator functions on convex sets $\mathcal{X}$ and $\mathcal{Y}$; the results in [41, 49] do not have an explicit dependence on κ, so we use a, b and c to denote the unknown orders; the more conditions required on $\mathbf{W}$ are in addition to those in Assumption 3 below.

参考文献	问题与算法的关键假设				((1.1) 若 $g \equiv 0$, $h \equiv 0$		
	(g, h) 条件下	关于W的更多内容	共识	多重通信	通信次数	#x梯度	#y梯度
[41]	$g \equiv 0, h = \iota_{\mathcal{Y}}$ $g = \iota_{\mathcal{X}}, h \equiv 0$	对称	仅 $\mathbf{x}$ 仅 $\mathbf{y}$	✗	$O(\frac{\kappa^a}{\varepsilon^2})$	$O(\frac{\kappa^a}{\varepsilon^2})$	$\tilde{O}(\frac{\kappa^c}{\varepsilon^2})$
					尚未建立		
[49]	$g \equiv 0, h \equiv 0$	对称	$\mathbf{x}$ 和 $\mathbf{y}$	✗	$O(\frac{\kappa^b}{\varepsilon^2})$	$O(\frac{\kappa^b}{\varepsilon^2})$	
[10]	$g \equiv 0, h \equiv 0$	对称 非负	$\mathbf{x}$ 和 $\mathbf{y}$	✗	$O(\frac{\kappa^3}{\varepsilon^2})$	$O(\frac{\kappa^3}{\varepsilon^2})$	
[4]	$g \equiv 0, h = \iota_{\mathcal{Y}}$	对称 $\mathbf{0} \preccurlyeq \mathbf{W} \preccurlyeq \mathbf{I}$	$\mathbf{x}$ 且 $\mathbf{y}$	✓	$O(\frac{\kappa^2}{\varepsilon^2})$	$O(\frac{\kappa^2}{\varepsilon^2})$	
本文	凸函数	none	$\mathbf{x}$ 和 $\mathbf{y}$	✗	$O(\frac{\kappa^2}{\varepsilon^2})$	$O(\frac{\kappa^2}{\varepsilon^2})$	$\tilde{O}(\frac{\kappa^2}{\varepsilon^2}\sqrt{\kappa_y})$

实现。据我们所知，所提出的算法是首个能够求解原变量和对偶变量均含有正则化项的去中心化 NCSC 极小极大问题的算法，而现有方法 [4, 10, 49] 仅适用于光滑问题，且对偶变量至多只能带有凸约束。第三，我们建立了所提算法的全局收敛性和迭代复杂度结果。该算法能在期望意义下，于 $O(\frac{\kappa}{\varepsilon^2})$ 次迭代内得到一个 ε-平稳解（见下文定义 3.12）。这对应于 $O(\frac{\kappa}{\varepsilon^2})$ 次通信轮次和 x-梯度，以及 $\tilde{O}(\frac{\kappa}{\varepsilon^2}\sqrt{\kappa_y})$[1] y-梯度，其中$\kappa$ 和 κ_y 是两个条件数；见（3.6）。此外，我们证明，当 h 光滑时，通过建立两种表述的近似平稳性之间的关系，只需 $O(\frac{\kappa^2}{\varepsilon^2})$ 次迭代，即可在期望意义下得到原始表述的一个 ε-平稳点（见注 3.4）。与光滑 NCSC 极小极大问题的现有结果相比，忽略对数因子后，我们的结果在 x-梯度方面提升了 κ 倍，在 y-梯度方面提升了 $\sqrt{\kappa_y}$ 倍；更详细的比较见表 1。

1.3. 记号。 我们用粗体小写字母 $\mathbf{x}$、$\mathbf{y}$ ..... 表示向量，用粗体大写字母 $\mathbf{X}$、$\mathbf{Y}$ .....表示矩阵。$\mathbf{0}$ 和 $\mathbf{1}$ 分别表示全零向量和全一向量；$\mathbf{I}$ 保留为单位矩阵。 $[m]$ 表示集合 $\{1, 2, \ldots, m\}$。我们用下标 $_i$ 表示第 i 个智能体拥有的局部向量，例如 $\mathbf{x}_i$；上标 ^{t} 用于第 t 个迭代点。我们令

$$\text{(1.2a)}\quad \mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_m]^\top, \quad \mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_m]^\top, \quad \mathbf{z}_i := [\mathbf{x}_i; \mathbf{y}_i] \in \mathbb{R}^{n_1+n_2}, \forall\, i \in [m], \quad \mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_m]^\top,$$

$$\text{(1.2b)}\quad \nabla_{\mathbf{x}} F(\mathbf{Z}) = [\nabla_{\mathbf{x}} f_1(\mathbf{z}_1), \ldots, \nabla_{\mathbf{x}} f_m(\mathbf{z}_m)]^\top, \quad \nabla_{\mathbf{y}} F(\mathbf{Z}) = [\nabla_{\mathbf{y}} f_1(\mathbf{z}_1), \ldots, \nabla_{\mathbf{y}} f_m(\mathbf{z}_m)]^\top,$$

其中 $\nabla_{\mathbf{x}} f_i$ 取关于 $\mathbf{x}$ 的偏梯度， $\nabla_{\mathbf{y}} f_i$ 取关于 $\mathbf{y}$的偏梯度。另外，我们定义

$$\text{(1.3)}\qquad \mathbf{x}_{\text{avg}} = \tfrac{1}{m}\textstyle\sum_{i=1}^m \mathbf{x}_i,\ \mathbf{X}_{\text{avg}} = \mathbf{1}\mathbf{x}_{\text{avg}}^\top,\ \mathbf{X}_\perp = \mathbf{X} - \mathbf{X}_{\text{avg}}.$$

对于闭凸函数 r，$\mathbf{prox}_r(\mathbf{x}) := \arg\min_{\mathbf{y}} \{r(\mathbf{y}) + \frac{1}{2}\|\mathbf{y} - \mathbf{x}\|^2\}$ 是 r 的近端映射。

1.4. 文章结构。 本文其余部分安排如下。第2 节介绍我们的算法。第3 节给出收敛结果，第4节给出数值结果。最后，第5 节总结全文。

[1] Throughout this paper, $\tilde{O}$-notation may hide a factor of $\ln \frac{1}{\varepsilon}$ within the big-O.

2. 去中心化表述与所提出的算法。 引入局部变量$\{(\mathbf{x}_i,\mathbf{y}_i)\}$后，可以将（1.1）等价地表述为以下共识约束问题

$$\min_{\mathbf{x}_1,\ldots,\mathbf{x}_m}\max_{\mathbf{y}_1,\ldots,\mathbf{y}_m}\frac{1}{m}\sum_{i=1}^m\big(f_i(\mathbf{x}_i,\mathbf{y}_i)+g(\mathbf{x}_i)-h(\mathbf{y}_i)\big),\ \text{s.t.}\ (\mathbf{x}_i,\mathbf{y}_i)=(\mathbf{x}_j,\mathbf{y}_j),\forall\,(i,j)\in\mathcal{E}. \tag{2.1}$$

为实现共识，通常使用混合矩阵（或 gossip 矩阵）$\mathbf{W}\in\mathbb{R}^{m\times m}$ 对邻居进行（加权）平均。当 W 满足某些条件（见下文假设 3）时，y-共识可表示为 $\mathbf{WY}=\mathbf{Y}$。因此，引入乘子 $\mathbf{\Lambda}\in\mathbb{R}^{m\times n_2}$，根据强对偶性 [38]，（当 dom($h$) 的相对内部非空时成立），对于每个 X，有 $\max_{\mathbf{Y}}\{\frac{1}{m}\sum_{i=1}^m(f_i(\mathbf{x}_i,\mathbf{y}_i)-h(\mathbf{y}_i))$，满足 $\mathbf{WY}=\mathbf{Y}\}=\min_{\mathbf{\Lambda}}\max_{\mathbf{Y}}\Phi(\mathbf{X},\mathbf{\Lambda},\mathbf{Y})$，其中

$$\Phi(\mathbf{X},\mathbf{\Lambda},\mathbf{Y}):=\frac{1}{m}\sum_{i=1}^m\big(f_i(\mathbf{x}_i,\mathbf{y}_i)-h(\mathbf{y}_i)\big)-\frac{L}{2\sqrt{m}}\langle\mathbf{\Lambda},(\mathbf{W}-\mathbf{I})\mathbf{Y}\rangle. \tag{2.2}$$

因此，我们可以将 (2.1) 等价地改写为

$$\min_{\mathbf{X},\mathbf{\Lambda}}\max_{\mathbf{Y}}\Phi(\mathbf{X},\mathbf{\Lambda},\mathbf{Y})+\frac{1}{m}\sum_{i=1}^m g(\mathbf{x}_i),\ \text{s.t.}\ \mathbf{x}_i=\mathbf{x}_j,\forall\,(i,j)\in\mathcal{E}. \tag{2.3}$$

与 (2.1) 相比，(2.3) 中没有 y-共识约束。这使得我们能够采用更积极的 Y 更新方式，从而可以进一步增大 x 更新的步长；参见备注 3.1。更具体地说，固定 X 和 $(\mathbf{W}-\mathbf{I})^\top\mathbf{\Lambda}$ 后，所有智能体上的 y-子问题彼此独立，因此我们可以通过（近似）最大化每个局部函数来贪心地更新 Y，而无需与邻居通信。利用（近似）Y-最大化解，我们可以得到 $\nabla Q(\mathbf{X},\mathbf{\Lambda})$ 或其估计值，其中 $Q(\mathbf{X},\mathbf{\Lambda}):=\max_{\mathbf{Y}}\Phi(\mathbf{X},\mathbf{Y},\mathbf{\Lambda})$。

基于这些观察，我们提出一种去中心化梯度下降最大化（D-GDMax）方法来求解（1.1）。伪代码见算法 1，其中 $d_i^t(\cdot)$ 定义为

$$d_i^t(\mathbf{y}):=f_i(\mathbf{x}_i^t,\mathbf{y})-h(\mathbf{y})-\frac{L\sqrt{m}}{2}\langle\tilde{\boldsymbol{\lambda}}_i^t,\mathbf{y}\rangle,\forall\,i\in[m],\forall\,t\geq 0. \tag{2.4}$$

在 D-GDMax 的每次迭代中，各智能体并行地首先对各自的局部 $\mathbf{x}$ 变量进行一轮邻居通信，并执行局部近端梯度步，沿跟踪的 $\mathbf{x}$-梯度更新 $\mathbf{x}$ 变量；其次，通过对局部 $\mathbf{y}$ 变量进行一轮邻居通信，采用梯度下降更新 $\boldsymbol{\lambda}$ 变量；然后，通过（近似）最大化局部函数来更新$\mathbf{y}$ 变量；最后，利用（近似）$\mathbf{y}$ 最大化解计算局部 $\mathbf{x}$ 梯度，并执行梯度跟踪步骤。

算法 1： 用于（1.1）的去中心化梯度下降最大化（D-GDMax）方法

---

1 输入：$\mathbf{x}^0\in\text{dom}(g)$，$\eta_x>0$，$\eta_\lambda>0$，非负数列 $\{\delta_t\}_{t\geq 0}$。2 初始化：对于每个 $i\in[m]$，令 $\mathbf{x}_i^0=\mathbf{x}^0$，$\boldsymbol{\lambda}_i^0=\tilde{\boldsymbol{\lambda}}_i^0=\mathbf{0}$，求得 $\mathbf{y}_i^0$，使得 $\text{dist}(\mathbf{0},\partial d_i^0(\mathbf{y}_i^0))\leq\delta_0$，并令 $\mathbf{v}_i^0=\nabla_{\mathbf{x}}f_i(\mathbf{x}_i^0,\mathbf{y}_i^0)$ $\forall\,i\in[m]$。3 对于 $t=0,1,\ldots,$ 执行 4 对于 $i=1,\ldots,m$ 并行执行 5 令 $\tilde{\mathbf{x}}_i^t=\sum_{j=1}^m w_{ij}\mathbf{x}_j^t$，并更新 $\mathbf{x}_i^{t+1}=\text{prox}_{\eta_x g}(\tilde{\mathbf{x}}_i^t-\eta_x\mathbf{v}_i^t)$。6 更新 $\boldsymbol{\lambda}_i^{t+1}=\boldsymbol{\lambda}_i^t+\frac{L\eta_\lambda}{2\sqrt{m}}(\sum_{j=1}^m w_{ij}\mathbf{y}_j^t-\mathbf{y}_i^t)$，并令 $\tilde{\boldsymbol{\lambda}}_i^{t+1}=\sum_{j=1}^m w_{ji}\boldsymbol{\lambda}_j^{t+1}-\boldsymbol{\lambda}_i^{t+1}$。7 求得 $\mathbf{y}_i^{t+1}$，使得 $\text{dist}(\mathbf{0},\partial d_i^{t+1}(\mathbf{y}_i^{t+1}))\leq\delta_{t+1}$，其中 d_i^t 定义于（2.4）中。8 令 $\mathbf{v}_i^{t+1}=\sum_{j=1}^m w_{ij}\mathbf{v}_j^t+\nabla_{\mathbf{x}}f_i(\mathbf{x}_i^{t+1},\mathbf{y}_i^{t+1})-\nabla_{\mathbf{x}}f_i(\mathbf{x}_i^t,\mathbf{y}_i^t)$。

我们算法的一个集中式版本已在 [15] 中针对光滑极小极大问题提出。然而，[15] 未利用强凹性，因此在特殊的中心化设置下，其复杂度高于我们的方法。

此外，尽管我们在 (2.3) 中的重新表述没有关于 $\mathbf{Y}$ 的一致性约束，但原问题不会化为标准的去中心化最小化问题，因为项 $\langle \mathbf{\Lambda}, (\mathbf{W}-\mathbf{I})\mathbf{Y}\rangle$在通过最大化消去 $\mathbf{Y}$ 后会导致原变量之间发生耦合。因此，我们的算法与用于求解一致性最小化问题的去中心化方法有本质区别。

3. 收敛性分析。本节分析算法 1中的 D-GDMax 方法。以下（标量或向量）函数将在我们的分析中使用或用于陈述结果：

$$p(\mathbf{x}) := \max_{\mathbf{y}} \left\{ \tfrac{1}{m}\textstyle\sum_{i=1}^m f_i(\mathbf{x},\mathbf{y}) - h(\mathbf{y}) \right\}, \quad P(\mathbf{x},\mathbf{\Lambda}) := \max_{\mathbf{Y}} \Phi(\mathbf{1}\mathbf{x}^\top, \mathbf{Y}, \mathbf{\Lambda}), \tag{3.1}$$

$$\phi(\mathbf{x},\mathbf{\Lambda}) := P(\mathbf{x},\mathbf{\Lambda}) + g(\mathbf{x}), \quad S_\Phi(\mathbf{X},\mathbf{\Lambda}) := \arg\max_{\mathbf{Y}} \Phi(\mathbf{X},\mathbf{\Lambda},\mathbf{Y}). \tag{3.2}$$

这里， p 是问题（1.1）原问题中的光滑项； P 可视为问题（2.3）原问题中的光滑项，而 ϕ 是目标函数。

3.1. 假设。对于所考虑的问题（1.1），我们作出以下假设：

ASSUMPTION 1. 问题 *(1.1)* 存在有限解 $x^* := \arg\min_{\mathbf{x}}(p+g)(x)$。对于每个 $i \in [m]$, f_i 在开集 $\mathcal{Z} \supseteq dom(g) \times dom(h)$ 内是 L-光滑的，并且 $f_i(x,\cdot)$ 是 μ-强凹的，其中 $\mu > 0$，即：

$$\|\nabla f_i(\mathbf{z}) - \nabla f_i(\tilde{\mathbf{z}})\| \le L\|\mathbf{z}-\tilde{\mathbf{z}}\|, \forall\, \mathbf{z},\tilde{\mathbf{z}} \in \mathcal{Z}, \tag{3.3}$$

$$\left\langle \mathbf{y}-\tilde{\mathbf{y}}, \nabla_{\mathbf{y}} f_i(\mathbf{x},\mathbf{y}) - \nabla_{\mathbf{y}} f_i(\mathbf{x},\tilde{\mathbf{y}}) \right\rangle \le -\mu\|\mathbf{y}-\tilde{\mathbf{y}}\|^2, \forall\, \mathbf{x} \in \mathrm{dom}(g), \forall\, \mathbf{y},\tilde{\mathbf{y}} \in \mathrm{dom}(h). \tag{3.4}$$

假设 1 中的强凹性条件也被关于求解非凸极小极大问题的现有去中心化方法所采用（例如， [4, 49]）。与这些现有工作相比，我们的设定更为一般，因为它纳入了凸正则项 g 和 h。为了收紧复杂度对某些常数的依赖，我们令 L_y 为 $f_i(\mathbf{x},\cdot)$ 的光滑常数，即：

$$\|\nabla_{\mathbf{y}} f_i(\mathbf{x},\mathbf{y}) - \nabla_{\mathbf{y}} f_i(\mathbf{x},\tilde{\mathbf{y}})\| \le L_y\|\mathbf{y}-\tilde{\mathbf{y}}\|, \forall\, \mathbf{y},\tilde{\mathbf{y}} \in \mathrm{dom}(h),\ \forall\, \mathbf{x} \in \mathrm{dom}(g),\ \forall\, i \in [m]. \tag{3.5}$$

根据 L_y 假设1，可以保证 L的存在。我们定义

$$\kappa = \tfrac{L}{\mu}, \quad \kappa_y = \tfrac{L_y}{\mu}. \tag{3.6}$$

ASSUMPTION 2. $\mathrm{dom}(h)$ 的相对内部非空。

上述条件较为宽松，并且意味着（2.1）与（2.3）等价，且 $p(\mathbf{x}) = \min_{\mathbf{\Lambda}} P(\mathbf{x},\mathbf{\Lambda})\ \forall \mathbf{x}$。ASSUMPTION 3. 混合矩阵 $\mathbf{W} \in \mathbb{R}^{m\times m}$ 满足以下条件：（i） $w_{ij} = 0$ 如果 j 不是 i的邻居；（ii） $\rho := \|\mathbf{W} - \frac{1}{m}\mathbf{1}\mathbf{1}^\top\|_2 < 1$；（iii）$\mathrm{Null}(\mathbf{W}-\mathbf{I}) = \mathrm{Span}\{\mathbf{1}\}$ 且 $\mathbf{W}^\top\mathbf{1} = \mathbf{1}$；（iv） $\|\mathbf{W}-\mathbf{I}\|_2 \le 2$。

在上述假设中，条件（i）表示， j 只有在 i 当且仅当（i, j）是$\mathcal{G}$的一条边时，才能向其发送消息；条件（ii）和（iii）对于实现共识至关重要，并且当 $\mathcal{G}$ 连通时可以成立；条件（iv）仅为方便起见，只要 $\|\mathbf{W}-\mathbf{I}\|_2$ 有界，我们的分析仍然可以进行。

算法 1 中的更新是从每个智能体的角度进行描述的。可以很容易地得到如下更为紧凑的矩阵形式，为便于分析和提高可读性，本文将采用该形式：

$$\mathbf{x}_i^{t+1} = \mathbf{prox}_{\eta_x g}(\tilde{\mathbf{x}}_i^t - \eta_x \mathbf{v}_i^t), \forall\, i \in [m], \text{ with } \widetilde{\mathbf{X}}^t := \mathbf{W}\mathbf{X}^t \tag{3.7a}$$

$$\mathbf{\Lambda}^{t+1} = \mathbf{\Lambda}^t + \tfrac{L\eta_\lambda}{2\sqrt{m}}(\mathbf{W}-\mathbf{I})\mathbf{Y}^t, \quad \mathbf{V}^{t+1} = \mathbf{W}\mathbf{V}^t + \nabla_{\mathbf{x}} F(\mathbf{X}^{t+1},\mathbf{Y}^{t+1}) - \nabla_{\mathbf{x}} F(\mathbf{X}^t,\mathbf{Y}^t). \tag{3.7b}$$

分析路线图：我们将首先证明 P 和 S_Φ 的一些重要性质，然后根据 ϕ的变化，推导出一次更新迭代后的关键不等式。此外，我们利用 $\mathbf{W} - \frac{1}{m}\mathbf{1}\mathbf{1}^\top$的收缩性，界定 $\mathbf{X}$和 $\mathbf{V}$迭代点的共识误差。最后，我们将共识误差界与单次迭代的进展结合起来，证明生成的$\mathbf{X}$ 和 $\overline{\mathbf{\Lambda}}$ 迭代点满足平方可和结果，据此证明全局收敛并建立迭代复杂度结果。

3.2. 预备结果。下面的引理展示了 P 和 S_Φ的一些重要性质。

PROPOSITION 3.1 (的光滑性 P 以及 S_Φ). 设 P 和 S_Φ 定义于 (3.1)和 (3.2)中。则 P是 L_P-光滑的，且 $L_P = L\sqrt{4\kappa^2+1}$ 并且 S_Φ 是 *Lipschitz* 连续的。更准确地说，

$$\|\nabla P(\mathbf{x},\mathbf{\Lambda}) - \nabla P(\tilde{\mathbf{x}},\tilde{\mathbf{\Lambda}})\|_F^2 \le L_P^2\left(\|\mathbf{x}-\tilde{\mathbf{x}}\|^2 + \|\mathbf{\Lambda}-\tilde{\mathbf{\Lambda}}\|_F^2\right), \forall\, \mathbf{x},\tilde{\mathbf{x}} \in \mathrm{dom}(g);\ \forall\, \mathbf{\Lambda},\tilde{\mathbf{\Lambda}}, \tag{3.8a}$$

$$\|S_\Phi(\mathbf{X}\ \mathbf{\Lambda}) - S_\Phi(\widetilde{\mathbf{X}}\ \mathbf{\Lambda})\|_F^2 \le \kappa^2\|\mathbf{X}-\widetilde{\mathbf{X}}\|_F^2,\quad \forall \mathbf{X},\ \widetilde{\mathbf{X}} \in \text{定义域}(g)^m;\quad \forall \mathbf{\Lambda}, \tag{3.8b}$$

$$\|S_\Phi(\mathbf{X}\ \mathbf{\Lambda}) - S_\Phi(\tilde{\mathbf{X}}\ \tilde{\mathbf{\Lambda}})\|_F^2 \le 2\kappa^2\|\mathbf{X}-\tilde{\mathbf{X}}\|_F^2 + 2m\kappa^2\|\mathbf{\Lambda}-\tilde{\mathbf{\Lambda}}\|_F^2, \forall \mathbf{X}\ \tilde{\mathbf{X}} \in \text{定义域}(g)^m;\quad \forall \mathbf{\Lambda},\tilde{\mathbf{\Lambda}}\ . \tag{3.8c}$$

~ ~ ~ ~ ~证明。 对于任意满足 $\mathbf{X}$,$\mathbf{X}$ $\in \mathrm{dom}(g)^m$ 的 $(\mathbf{X}\ \mathbf{\Lambda})$ 和 $(\mathbf{X}\ \mathbf{\Lambda})$，令 $\mathbf{Y} = S_\Phi(\mathbf{X}\ \mathbf{\Lambda})$ 和 $\mathbf{Y} = S_\Phi(\mathbf{X}\ \mathbf{\Lambda})$。根据最优性条件，存在 $\boldsymbol{\zeta}_i \in \partial h(\mathbf{y}_i)\ \forall\, i \in [m]$ 使得

$$\tfrac{1}{m}\left[\nabla_{\mathbf{y}} f_1(\mathbf{x}_1,\mathbf{y}_1) - \boldsymbol{\zeta}_1, \cdots, \nabla_{\mathbf{y}} f_m(\mathbf{x}_m,\mathbf{y}_m) - \boldsymbol{\zeta}_m\right]^\top - \tfrac{L}{2\sqrt{m}}(\mathbf{W}-\mathbf{I})^\top\mathbf{\Lambda} = \mathbf{0}. \tag{3.9}$$

因此，$\frac{1}{m}\sum_{i=1}^m \langle \tilde{\mathrm{y}}_i - \mathrm{y}_i, \nabla_{\mathbf{y}} f_i(\mathrm{x}_i,\mathrm{y}_i) - \boldsymbol{\zeta}_i\rangle - \frac{L}{2\sqrt{m}}\langle \tilde{\mathrm{Y}} - \mathrm{Y}, (\mathrm{W}-\mathrm{I})^\top\mathbf{\Lambda}\rangle = 0$. 类似地，对每个 i，存在 $\tilde{\boldsymbol{\zeta}}_i \in \partial h(\tilde{\mathrm{y}}_i)$ 使得 $\frac{1}{m}\sum_{i=1}^m \langle \mathrm{y}_i - \tilde{\mathrm{y}}_i, \nabla_{\mathbf{y}} f_i(\tilde{\mathrm{x}}_i,\tilde{\mathrm{y}}_i) - \tilde{\boldsymbol{\zeta}}_i\rangle - \frac{L}{2\sqrt{m}}\langle \mathrm{Y} - \tilde{\mathrm{Y}}, (\mathrm{W}-\mathrm{I})^\top\tilde{\mathbf{\Lambda}}\rangle = 0$. 将两个等式相加，得到

$$\tfrac{1}{m}\sum_{i=1}^m \langle \tilde{\mathbf{y}}_i - \mathbf{y}_i, \nabla_{\mathbf{y}} f_i(\mathbf{x}_i,\mathbf{y}_i) - \nabla_{\mathbf{y}} f_i(\tilde{\mathbf{x}}_i,\tilde{\mathbf{y}}_i) + \tilde{\boldsymbol{\zeta}}_i - \boldsymbol{\zeta}_i\rangle + \tfrac{L}{2\sqrt{m}}\langle \widetilde{\mathbf{Y}} - \mathbf{Y}, (\mathbf{W}-\mathbf{I})^\top(\tilde{\mathbf{\Lambda}}-\mathbf{\Lambda})\rangle = 0. \tag{3.10}$$

每个 $f_i(\mathrm{x}_i,\cdot)$ 的 μ 强凹性意味着 $\frac{1}{m}\sum_{i=1}^m \langle \tilde{\mathrm{y}}_i - \mathrm{y}_i, \nabla_{\mathbf{y}} f_i(\mathrm{x}_i,\mathrm{y}_i) - \nabla_{\mathbf{y}} f_i(\mathrm{x}_i,\tilde{\mathrm{y}}_i)\rangle \ge \frac{\mu}{m}\sum_{i=1}^m \|\tilde{\mathrm{y}}_i - \mathrm{y}_i\|^2$. 此外，由 h 的凸性可得 $\frac{1}{m}\sum_{i=1}^m \langle \tilde{\mathrm{y}}_i - \mathrm{y}_i, \tilde{\boldsymbol{\zeta}}_i - \boldsymbol{\zeta}_i\rangle \ge 0$. 将这两个不等式代入 (3.10)，得到

$$\tfrac{1}{m}\sum_{i=1}^m \langle \tilde{\mathbf{y}}_i - \mathbf{y}_i, \nabla_{\mathbf{y}} f_i(\mathbf{x}_i,\mathbf{y}_i) - \nabla_{\mathbf{y}} f_i(\tilde{\mathbf{x}}_i,\tilde{\mathbf{y}}_i)\rangle + \tfrac{L}{2\sqrt{m}}\langle \widetilde{\mathbf{Y}} - \mathbf{Y}, (\mathbf{W}-\mathbf{I})^\top(\tilde{\mathbf{\Lambda}}-\mathbf{\Lambda})\rangle \ge \tfrac{\mu}{m}\sum_{i=1}^m \|\tilde{\mathbf{y}}_i - \mathbf{y}_i\|^2. \tag{3.11}$$

此外，由于 f_i 光滑，有 $\|\nabla_{\mathbf{y}} f_i(\mathrm{x}_i,\tilde{\mathrm{y}}_i) - \nabla_{\mathbf{y}} f_i(\tilde{\mathrm{x}}_i,\tilde{\mathrm{y}}_i)\| \le L\|\mathrm{x}_i - \tilde{\mathrm{x}}_i\|$，因此利用杨氏不等式可得 $\langle \tilde{\mathrm{y}}_i - \mathrm{y}_i, \nabla_{\mathbf{y}} f_i(\mathrm{x}_i,\tilde{\mathrm{y}}_i) - \nabla_{\mathbf{y}} f_i(\tilde{\mathrm{x}}_i,\tilde{\mathrm{y}}_i)\rangle \le \frac{a}{4}\|\tilde{\mathrm{y}}_i - \mathrm{y}_i\|^2 + \frac{L^2}{a}\|\mathrm{x}_i - \tilde{\mathrm{x}}_i\|^2$，对任意 $a > 0$ 均成立。因此，(3.11) 推出

$$\tfrac{\mu}{m}\sum_{i=1}^m \|\tilde{\mathbf{y}}_i - \mathbf{y}_i\|^2 \le \tfrac{1}{m}\sum_{i=1}^m \left(\tfrac{a}{4}\|\tilde{\mathbf{y}}_i - \mathbf{y}_i\|^2 + \tfrac{L^2}{a}\|\mathbf{x}_i - \tilde{\mathbf{x}}_i\|^2\right) + \tfrac{L}{2\sqrt{m}}\langle \widetilde{\mathbf{Y}} - \mathbf{Y}, (\mathbf{W}-\mathbf{I})^\top(\tilde{\mathbf{\Lambda}}-\mathbf{\Lambda})\rangle. \tag{3.12}$$

下面我们讨论两种不同的情况，以得到所需结果。

情况 I. 当 $\mathbf{\Lambda} = \tilde{\mathbf{\Lambda}}$时，我们取 $a = 2\mu$ 于（3.12）并重排各项，得到 $\|\tilde{\mathbf{Y}} - \mathbf{Y}\|_F^2 \le \frac{L^2}{\mu^2}\|\tilde{\mathbf{X}} - \mathbf{X}\|_F^2$，由 3.8b）根据 $\kappa = \frac{L}{\mu}$ 在（3.6）中的定义可知。

情况 II. 一般而言，我们首先应用杨氏不等式，得到

$$\tfrac{L}{2\sqrt{m}}\langle \widetilde{\mathbf{Y}} - \mathbf{Y}, (\mathbf{W}-\mathbf{I})^\top(\tilde{\mathbf{\Lambda}}-\mathbf{\Lambda})\rangle \le \tfrac{\mu}{4m}\|\widetilde{\mathbf{Y}} - \mathbf{Y}\|_F^2 + \tfrac{L^2}{4\mu}\|(\mathbf{W}-\mathbf{I})^\top(\tilde{\mathbf{\Lambda}}-\mathbf{\Lambda})\|_F^2.$$

然后，我们取 $a = \mu$ 于（3.12）并通过重排各项得到 $\frac{1}{2m}\|\tilde{\mathbf{Y}} - \mathbf{Y}\|_F^2 \le \frac{L^2}{m\mu^2}\|\mathbf{X} - \tilde{\mathbf{X}}\|_F^2 + \frac{L^2}{4\mu^2}\|(\mathbf{W}-\mathbf{I})^\top(\tilde{\mathbf{\Lambda}}-\mathbf{\Lambda})\|_F^2$，结合 $\|\mathbf{W}-\mathbf{I}\|_2 \le 2$ 这一事实，

$$\tfrac{1}{2m}\|\widetilde{\mathbf{Y}} - \mathbf{Y}\|_F^2 \le \tfrac{L^2}{m\mu^2}\|\mathbf{X} - \widetilde{\mathbf{X}}\|_F^2 + \tfrac{L^2}{\mu^2}\|\tilde{\mathbf{\Lambda}} - \mathbf{\Lambda}\|_F^2 = \tfrac{\kappa^2}{m}\|\mathbf{X} - \widetilde{\mathbf{X}}\|_F^2 + \kappa^2\|\tilde{\mathbf{\Lambda}} - \mathbf{\Lambda}\|_F^2. \tag{3.13}$$

因此，通过对（3.8c）两边应用 $2m$，我们得到（3.13）。

现在，令$\mathbf{X}=\mathbf{1}\mathbf{x}^\top$和$\widetilde{\mathbf{X}}=\mathbf{1}\tilde{\mathbf{x}}^\top$，我们注意到$\nabla P(\mathbf{x},\mathbf{\Lambda})=\left(\frac{1}{m}\sum_{i=1}^m\nabla_{\mathbf{x}}f_i(\mathbf{x},\mathbf{y}_i),\ -\frac{L}{2\sqrt{m}}(\mathbf{W}-\mathbf{I})\mathbf{Y}\right)$根据Danskin定理 [6]。因此，

$$\|\nabla P(\mathbf{x},\mathbf{\Lambda})-\nabla P(\tilde{\mathbf{x}},\tilde{\mathbf{\Lambda}})\|_F^2=\tfrac{1}{m^2}\big\|\sum_{i=1}^m\nabla_{\mathbf{x}}f_i(\mathbf{x},\mathbf{y}_i)-\sum_{i=1}^m\nabla_{\mathbf{x}}f_i(\tilde{\mathbf{x}},\tilde{\mathbf{y}}_i)\big\|^2+\tfrac{L^2}{4m}\big\|(\mathbf{W}-\mathbf{I})(\mathbf{Y}-\widetilde{\mathbf{Y}})\big\|_F^2$$
$$\le\tfrac{L^2}{m}\sum_{i=1}^m\left(\|\mathbf{x}-\tilde{\mathbf{x}}\|^2+\|\mathbf{y}_i-\tilde{\mathbf{y}}_i\|^2\right)+\tfrac{L^2}{m}\|\mathbf{Y}-\widetilde{\mathbf{Y}}\|_F^2\overset{(3.13)}{\le}L^2\|\mathbf{x}-\tilde{\mathbf{x}}\|^2+4L^2\kappa^2\left(\|\mathbf{x}-\tilde{\mathbf{x}}\|^2+\|\tilde{\mathbf{\Lambda}}-\mathbf{\Lambda}\|_F^2\right),$$

这意味着（3.8a）。证明完毕。 □

下面的命题给出了近似最大化器$\mathbf{Y}^t$与精确解之间的界。

PROPOSITION 3.2. 设 $\{\mathbf{Y}^t\}$ 由算法 1 生成。则对于任意 $t\ge0$，有 $\left\|\mathbf{Y}^t-S_\Phi(\mathbf{X}^t,\mathbf{\Lambda}^t)\right\|_F^2\le\frac{m\delta_t^2}{\mu^2}$。证明。设 $\mathbf{Y}^{t\star}=S_\Phi(\mathbf{X}^t,\mathbf{\Lambda}^t)$，即对于每个 $i\in[m]$，$\mathbf{y}_i^{t\star}=\arg\max_{\mathbf{y}_i}d_i^t(\mathbf{y}_i)$。由 d_i^t 的 μ-强凹性可知，$\mu\|\mathbf{y}_i^{t\star}-\mathbf{y}_i^t\|^2\le\langle\mathbf{y}_i^{t\star}-\mathbf{y}_i^t$，对于任意 $\boldsymbol{\xi}_i^t\in\partial d_i^t(\mathbf{y}_i^t)$ 均成立。因此，利用柯西-施瓦茨不等式，并根据条件 $\operatorname{dist}(\mathbf{0},\partial d_i^t(\mathbf{y}_i^t))\le\delta_t$，可得 $\mu\|\mathbf{y}_i^{t\star}-\mathbf{y}_i^t\|^2\le\delta_t\|\mathbf{y}_i^{t\star}-\mathbf{y}_i^t\|$，从而对于每个 $i\in[m]$，都有$\|\mathbf{y}_i^{t\star}-\mathbf{y}_i^t\|\le\frac{\delta_t}{\mu}$。这就得到了所需结果。□

3.3. 全局收敛和收敛速率结果。 我们首先利用以下两个引理建立一个关键的单次迭代进展不等式。

LEMMA 3.3 ([28, 引理 C.3]). 对于任意 $t\ge0$ 和每个 $i\in[m]$，成立

$$g(\mathbf{x}_i^{t+1})-g(\mathbf{x}_{\text{avg}}^t)+\left\langle\mathbf{x}_i^{t+1}-\mathbf{x}_{\text{avg}}^t,\mathbf{v}_i^t\right\rangle\le-\tfrac{1}{2\eta_x}\left(\|\mathbf{x}_i^{t+1}-\mathbf{x}_{\text{avg}}^t\|^2+\|\mathbf{x}_i^{t+1}-\tilde{\mathbf{x}}_i^t\|^2-\|\tilde{\mathbf{x}}_i^t-\mathbf{x}_{\text{avg}}^t\|^2\right).\tag{3.14}$$

LEMMA 3.4. 对于任意 $t\ge0$，在 $L_P=L\sqrt{4\kappa^2+1}$ 下成立

$$\begin{aligned}&\left\langle\nabla P(\mathbf{x}_{\text{avg}}^t,\mathbf{\Lambda}^t),(\mathbf{x}_{\text{avg}}^{t+1}-\mathbf{x}_{\text{avg}}^t,\mathbf{\Lambda}^{t+1}-\mathbf{\Lambda}^t)\right\rangle-\tfrac{1}{m}\sum_{i=1}^m\left\langle\mathbf{x}_i^{t+1}-\mathbf{x}_{\text{avg}}^t,\mathbf{v}_i^t\right\rangle\\&\le\tfrac{1}{2m}\left(L_P\|\mathbf{X}^{t+1}-\mathbf{X}_{\text{avg}}^t\|_F^2+\tfrac{1}{L_P}\|\mathbf{V}_\perp^t\|_F^2\right)+\tfrac{L}{2m}\left((2\kappa+1)\|\mathbf{X}_\perp^t\|_F^2+(\kappa+1)\|\mathbf{X}^{t+1}-\mathbf{X}_{\text{avg}}^t\|_F^2\right)+\tfrac{2\delta_t^2}{\mu}\\&\quad+\tfrac{L\kappa}{m}\|\mathbf{X}_\perp^t\|_F^2-\left(\tfrac{1}{\eta_\lambda}-\tfrac{L\kappa}{2}\right)\|\mathbf{\Lambda}^{t+1}-\mathbf{\Lambda}^t\|_F^2.\end{aligned}\tag{3.15}$$

证明。令 $\widehat{\mathbf{Y}}^t:=S_\Phi\left(\mathbf{1}(\mathbf{x}_{\text{avg}}^t)^\top,\mathbf{\Lambda}^t\right)$。根据Danskin定理 [6], 有

$$\nabla P(\mathbf{x}_{\text{avg}}^t,\mathbf{\Lambda}^t)=\left(\tfrac{1}{m}\sum_{i=1}^m\nabla_{\mathbf{x}}f_i(\mathbf{x}_{\text{avg}}^t,\hat{\mathbf{y}}_i^t),\ -\tfrac{L}{2\sqrt{m}}(\mathbf{W}-\mathbf{I})\widehat{\mathbf{Y}}^t\right).\tag{3.16}$$

因此，我们有

$$\begin{aligned}&\left\langle\nabla P(\mathbf{x}_{\text{avg}}^t,\mathbf{\Lambda}^t),(\mathbf{x}_{\text{avg}}^{t+1}-\mathbf{x}_{\text{avg}}^t,\mathbf{\Lambda}^{t+1}-\mathbf{\Lambda}^t)\right\rangle\\&=\left\langle\tfrac{1}{m}\sum_{i=1}^m\nabla_{\mathbf{x}}f_i(\mathbf{x}_{\text{avg}}^t,\hat{\mathbf{y}}_i^t),\mathbf{x}_{\text{avg}}^{t+1}-\mathbf{x}_{\text{avg}}^t\right\rangle-\tfrac{L}{2\sqrt{m}}\left\langle(\mathbf{W}-\mathbf{I})\widehat{\mathbf{Y}}^t,\mathbf{\Lambda}^{t+1}-\mathbf{\Lambda}^t\right\rangle.\end{aligned}\tag{3.17}$$

对于右侧（3.17）中的第一个内积项，我们将其拆分如下：

$$\begin{aligned}\left\langle\tfrac{1}{m}\sum_{i=1}^m\nabla_{\mathbf{x}}f_i(\mathbf{x}_{\text{avg}}^t,\hat{\mathbf{y}}_i^t),\mathbf{x}_{\text{avg}}^{t+1}-\mathbf{x}_{\text{avg}}^t\right\rangle&=\tfrac{1}{m}\sum_{i=1}^m\left\langle\nabla_{\mathbf{x}}f_i(\mathbf{x}_i^t,\mathbf{y}_i^t),\mathbf{x}_{\text{avg}}^{t+1}-\mathbf{x}_{\text{avg}}^t\right\rangle\\&\quad+\tfrac{1}{m}\sum_{i=1}^m\left\langle\nabla_{\mathbf{x}}f_i(\mathbf{x}_{\text{avg}}^t,\hat{\mathbf{y}}_i^t)-\nabla_{\mathbf{x}}f_i(\mathbf{x}_i^t,\hat{\mathbf{y}}_i^t),\mathbf{x}_{\text{avg}}^{t+1}-\mathbf{x}_{\text{avg}}^t\right\rangle\\&\quad+\tfrac{1}{m}\sum_{i=1}^m\left\langle\nabla_{\mathbf{x}}f_i(\mathbf{x}_i^t,\hat{\mathbf{y}}_i^t)-\nabla_{\mathbf{x}}f_i(\mathbf{x}_i^t,\mathbf{y}_i^t),\mathbf{x}_{\text{avg}}^{t+1}-\mathbf{x}_{\text{avg}}^t\right\rangle.\end{aligned}\tag{3.18}$$

由每个 f_i 的光滑性和杨氏不等式，我们得到

$$\begin{aligned}&\tfrac{1}{m}\sum_{i=1}^m\left\langle\nabla_{\mathbf{x}}f_i(\mathbf{x}_{\text{avg}}^t,\hat{\mathbf{y}}_i^t)-\nabla_{\mathbf{x}}f_i(\mathbf{x}_i^t,\hat{\mathbf{y}}_i^t),\mathbf{x}_{\text{avg}}^{t+1}-\mathbf{x}_{\text{avg}}^t\right\rangle\\&\le\tfrac{L}{2m}\sum_{i=1}^m\left(\|\mathbf{x}_{\text{avg}}^t-\mathbf{x}_i^t\|^2+\|\mathbf{x}_{\text{avg}}^{t+1}-\mathbf{x}_{\text{avg}}^t\|^2\right)=\tfrac{L}{2m}\left(\|\mathbf{X}_\perp^t\|_F^2+\|\mathbf{X}^{t+1}-\mathbf{X}_{\text{avg}}^t\|_F^2\right),\end{aligned}\tag{3.19}$$

其中我们使用了式（1.3）中的记号。类似地，有

$$\begin{aligned}(3.20)\quad &\tfrac{1}{m}\textstyle\sum_{i=1}^m \big\langle \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^t,\hat{\mathbf{y}}_i^t) - \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^t,\mathbf{y}_i^t), \mathbf{x}_{\rm avg}^{t+1} - \mathbf{x}_{\rm avg}^t\big\rangle \\ &\le \tfrac{L}{2m}\textstyle\sum_{i=1}^m \big(\tfrac{1}{\kappa}\|\hat{\mathbf{y}}_i^t - \mathbf{y}_i^t\|^2 + \kappa\|\mathbf{x}_{\rm avg}^{t+1} - \mathbf{x}_{\rm avg}^t\|^2\big) = \tfrac{L}{2m}\big(\tfrac{1}{\kappa}\|\widehat{\mathbf{Y}}^t - \mathbf{Y}^t\|_F^2 + \kappa\|\mathbf{X}_{\rm avg}^{t+1} - \mathbf{X}_{\rm avg}^t\|_F^2\big).\end{aligned}$$

By 根据 $\widehat{\ }\mathbf{Y}^t$ 的定义以及 $\mathbf{Y}^t$的选取，由命题 3.1 和 3.2 可得

$$(3.21)\qquad \|\widehat{\mathbf{Y}}^t - \mathbf{Y}^t\|_F^2 \le 2\|\widehat{\mathbf{Y}}^t - S_\Phi(\mathbf{X}^t,\mathbf{\Lambda}^t)\|_F^2 + 2\|S_\Phi(\mathbf{X}^t,\mathbf{\Lambda}^t) - \mathbf{Y}^t\|_F^2 \le 2\kappa^2\|\mathbf{X}_\perp^t\|_F^2 + \tfrac{2m\delta_t^2}{\mu^2},$$

因此，结合（3.20）以及 $\kappa = \frac{L}{\mu}$ 的定义，可得 –

$$(3.22)\quad \tfrac{1}{m}\textstyle\sum_{i=1}^m \big\langle \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^t,\hat{\mathbf{y}}_i^t) - \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^t,\mathbf{y}_i^t), \mathbf{x}_{\rm avg}^{t+1} - \mathbf{x}_{\rm avg}^t\big\rangle \le \tfrac{L}{2m}\big(2\kappa\|\mathbf{X}_\perp^t\|_F^2 + \kappa\|\mathbf{X}_{\rm avg}^{t+1} - \mathbf{X}_{\rm avg}^t\|_F^2\big) + \tfrac{\delta_t^2}{\mu}.$$

将（3.19）和（3.22）代入（3.18），得

$$\begin{aligned}(3.23)\quad &\big\langle \tfrac{1}{m}\textstyle\sum_{i=1}^m \nabla_{\mathbf{x}} f_i(\mathbf{x}_{\rm avg}^t,\hat{\mathbf{y}}_i^t), \mathbf{x}_{\rm avg}^{t+1} - \mathbf{x}_{\rm avg}^t\big\rangle \\ &\le \tfrac{1}{m}\textstyle\sum_{i=1}^m \big\langle \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^t,\mathbf{y}_i^t), \mathbf{x}_{\rm avg}^{t+1} - \mathbf{x}_{\rm avg}^t\big\rangle + \tfrac{L}{2m}\big((2\kappa+1)\|\mathbf{X}_\perp^t\|_F^2 + (\kappa+1)\|\mathbf{X}_{\rm avg}^{t+1} - \mathbf{X}_{\rm avg}^t\|_F^2\big) + \tfrac{\delta_t^2}{\mu}.\end{aligned}$$

此外，注意到 $\mathbf{v}_{\rm avg}^t = \frac{1}{m}\sum_{i=1}^m \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^t,\mathbf{y}_i^t)$。因此，

$$\begin{aligned}&\tfrac{1}{m}\textstyle\sum_{i=1}^m \big\langle \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^t,\mathbf{y}_i^t), \mathbf{x}_{\rm avg}^{t+1} - \mathbf{x}_{\rm avg}^t\big\rangle - \tfrac{1}{m}\textstyle\sum_{i=1}^m \big\langle \mathbf{x}_i^{t+1} - \mathbf{x}_{\rm avg}^t, \mathbf{v}_i^t\big\rangle \\ &= \tfrac{1}{m}\textstyle\sum_{i=1}^m \big\langle \mathbf{x}_i^{t+1} - \mathbf{x}_{\rm avg}^t, \mathbf{v}_{\rm avg}^t - \mathbf{v}_i^t\big\rangle \le \tfrac{1}{2m}\textstyle\sum_{i=1}^m \big(L_P\|\mathbf{x}_i^{t+1} - \mathbf{x}_{\rm avg}^t\|^2 + \tfrac{1}{L_P}\|\mathbf{v}_{\rm avg}^t - \mathbf{v}_i^t\|^2\big),\end{aligned}$$

其中， L_P 由命题 3.1给出。上述不等式结合（3.23）可得

$$\begin{aligned}(3.24)\quad &\big\langle \tfrac{1}{m}\textstyle\sum_{i=1}^m \nabla_{\mathbf{x}} f_i(\mathbf{x}_{\rm avg}^t,\hat{\mathbf{y}}_i^t), \mathbf{x}_{\rm avg}^{t+1} - \mathbf{x}_{\rm avg}^t\big\rangle - \tfrac{1}{m}\textstyle\sum_{i=1}^m \big\langle \mathbf{x}_i^{t+1} - \mathbf{x}_{\rm avg}^t, \mathbf{v}_i^t\big\rangle \\ &\le \tfrac{1}{2m}\big(L_P\|\mathbf{X}^{t+1} - \mathbf{X}_{\rm avg}^t\|_F^2 + \tfrac{1}{L_P}\|\mathbf{V}_\perp^t\|_F^2\big) + \tfrac{L}{2m}\big((2\kappa+1)\|\mathbf{X}_\perp^t\|_F^2 + (\kappa+1)\|\mathbf{X}_{\rm avg}^{t+1} - \mathbf{X}_{\rm avg}^t\|_F^2\big) + \tfrac{\delta_t^2}{\mu}.\end{aligned}$$

此外，根据杨氏不等式以及 $\mathbf{\Lambda}$ 在（3.7b）中的更新公式，有

$$\begin{aligned}&-\tfrac{L}{2\sqrt{m}}\big\langle(\mathbf{W}-\mathbf{I})\widehat{\mathbf{Y}}^t, \mathbf{\Lambda}^{t+1} - \mathbf{\Lambda}^t\big\rangle \\ &= -\tfrac{L}{2\sqrt{m}}\big\langle(\mathbf{W}-\mathbf{I})(\widehat{\mathbf{Y}}^t - \mathbf{Y}^t), \mathbf{\Lambda}^{t+1} - \mathbf{\Lambda}^t\big\rangle - \tfrac{L}{2\sqrt{m}}\big\langle(\mathbf{W}-\mathbf{I})\mathbf{Y}^t, \mathbf{\Lambda}^{t+1} - \mathbf{\Lambda}^t\big\rangle \\ (3.25)\quad &\le \tfrac{L}{2m\kappa}\|\widehat{\mathbf{Y}}^t - \mathbf{Y}^t\|_F^2 + \tfrac{L\kappa}{8}\|(\mathbf{W}-\mathbf{I})^\top(\mathbf{\Lambda}^{t+1} - \mathbf{\Lambda}^t)\|_F^2 - \tfrac{1}{\eta_\lambda}\|\mathbf{\Lambda}^{t+1} - \mathbf{\Lambda}^t\|_F^2.\end{aligned}$$

现在将 (3.24) 和 (3.25) 加到 (3.17) 上，并使用 (1.3) 中的记号，得到

$$\begin{aligned}&\big\langle \nabla P(\mathbf{x}_{\rm avg}^t,\mathbf{\Lambda}^t), (\mathbf{x}_{\rm avg}^{t+1} - \mathbf{x}_{\rm avg}^t, \mathbf{\Lambda}^{t+1} - \mathbf{\Lambda}^t)\big\rangle - \tfrac{1}{m}\textstyle\sum_{i=1}^m \big\langle \mathbf{x}_i^{t+1} - \mathbf{x}_{\rm avg}^t, \mathbf{v}_i^t\big\rangle \\ (3.26)\quad &\le \tfrac{1}{2m}\big(L_P\|\mathbf{X}^{t+1} - \mathbf{X}_{\rm avg}^t\|_F^2 + \tfrac{1}{L_P}\|\mathbf{V}_\perp^t\|_F^2\big) + \tfrac{L}{2m}\big((2\kappa+1)\|\mathbf{X}_\perp^t\|_F^2 + (\kappa+1)\|\mathbf{X}_{\rm avg}^{t+1} - \mathbf{X}_{\rm avg}^t\|_F^2\big) + \tfrac{\delta_t^2}{\mu} \\ &\quad + \tfrac{L}{2m\kappa}\|\widehat{\mathbf{Y}}^t - \mathbf{Y}^t\|_F^2 + \tfrac{L\kappa}{8}\|(\mathbf{W}-\mathbf{I})^\top(\mathbf{\Lambda}^{t+1} - \mathbf{\Lambda}^t)\|_F^2 - \tfrac{1}{\eta_\lambda}\|\mathbf{\Lambda}^{t+1} - \mathbf{\Lambda}^t\|_F^2,\end{aligned}$$

这与 $\|\mathbf{W} - \mathbf{I}\|_2 \le 2$ 以及 (3.21) 一起，给出了 (3.15) 中的期望结果。 □

现在我们准备展示单次迭代进展不等式。

THEOREM 3.5. 设 $\{(\mathbf{X}^t,\mathbf{\Lambda}^t,\mathbf{V}^t)\}$ 由算法 *1*生成。则我们有，在 $L_P = L\sqrt{4\kappa^2+1}$ 下 t

$$\begin{aligned}(3.27)\quad &\tfrac{1}{2m}\big(\tfrac{1}{\eta_x} - 2L_P - L(\kappa+1)\big)\|\mathbf{X}^{t+1} - \mathbf{X}_{\rm avg}^t\|^2 + \tfrac{1}{2m\eta_x}\|\mathbf{X}^{t+1} - \widetilde{\mathbf{X}}^t\|_F^2 + \big(\tfrac{1}{\eta_\lambda} - \tfrac{L_P+L\kappa}{2}\big)\|\mathbf{\Lambda}^{t+1} - \mathbf{\Lambda}^t\|_F^2 \\ &\le \phi(\mathbf{x}_{\rm avg}^t,\mathbf{\Lambda}^t) - \phi(\mathbf{x}_{\rm avg}^{t+1},\mathbf{\Lambda}^{t+1}) + \tfrac{1}{2mL_P}\|\mathbf{V}_\perp^t\|_F^2 + \big(\tfrac{\rho^2}{2m\eta_x} + \tfrac{L(4\kappa+1)}{2m}\big)\|\mathbf{X}_\perp^t\|_F^2 + \tfrac{2\delta_t^2}{\mu}.\end{aligned}$$

证明。 根据 L_P-光滑性，以及命题 P 中的 3.1，可得

$$\begin{aligned}\phi(\mathbf{x}_{\text{avg}}^{t+1},\mathbf{\Lambda}^{t+1})-\phi(\mathbf{x}_{\text{avg}}^{t},\mathbf{\Lambda}^{t}) \le \tfrac{L_P}{2}\left(\|\mathbf{x}_{\text{avg}}^{t+1}-\mathbf{x}_{\text{avg}}^{t}\|^2+\|\mathbf{\Lambda}^{t+1}-\mathbf{\Lambda}^{t}\|_F^2\right)\\ +\left\langle\nabla P(\mathbf{x}_{\text{avg}}^{t},\mathbf{\Lambda}^{t}),(\mathbf{x}_{\text{avg}}^{t+1}-\mathbf{x}_{\text{avg}}^{t},\mathbf{\Lambda}^{t+1}-\mathbf{\Lambda}^{t})\right\rangle+g(\mathbf{x}_{\text{avg}}^{t+1})-g(\mathbf{x}_{\text{avg}}^{t}).\end{aligned}\tag{3.28}$$

此外，由于 g，有 $g(\mathbf{x}_{\text{avg}}^{t+1})-g(\mathbf{x}_{\text{avg}}^{t}) \le \frac{1}{m}\sum_{i=1}^m\left(g(\mathbf{x}_i^{t+1})-g(\mathbf{x}_{\text{avg}}^{t})\right)$。因此，由（3.14）和（3.28）可得

$$\begin{aligned}\phi(\mathbf{x}_{\text{avg}}^{t+1},\mathbf{\Lambda}^{t+1})-\phi(\mathbf{x}_{\text{avg}}^{t},\mathbf{\Lambda}^{t}) \le \tfrac{L_P}{2}\left(\|\mathbf{x}_{\text{avg}}^{t+1}-\mathbf{x}_{\text{avg}}^{t}\|^2+\|\mathbf{\Lambda}^{t+1}-\mathbf{\Lambda}^{t}\|_F^2\right)\\ +\left\langle\nabla P(\mathbf{x}_{\text{avg}}^{t},\mathbf{\Lambda}^{t}),(\mathbf{x}_{\text{avg}}^{t+1}-\mathbf{x}_{\text{avg}}^{t},\mathbf{\Lambda}^{t+1}-\mathbf{\Lambda}^{t})\right\rangle-\tfrac{1}{m}\sum_{i=1}^m\left\langle\mathbf{x}_i^{t+1}-\mathbf{x}_{\text{avg}}^{t},\mathbf{v}_i^{t}\right\rangle\\ -\tfrac{1}{2m\eta_x}\sum_{i=1}^m\left(\|\mathbf{x}_i^{t+1}-\mathbf{x}_{\text{avg}}^{t}\|^2+\|\mathbf{x}_i^{t+1}-\tilde{\mathbf{x}}_i^{t}\|^2-\|\tilde{\mathbf{x}}_i^{t}-\mathbf{x}_{\text{avg}}^{t}\|^2\right).\end{aligned}\tag{3.29}$$

将（3.15）加到（3.29）并合并同类项，得到

$$\begin{aligned}\phi(\mathbf{x}_{\text{avg}}^{t+1},\mathbf{\Lambda}^{t+1})-\phi(\mathbf{x}_{\text{avg}}^{t},\mathbf{\Lambda}^{t}) \le \tfrac{L_P+L(\kappa+1)}{2m}\|\mathbf{X}_{\text{avg}}^{t+1}-\mathbf{X}_{\text{avg}}^{t}\|^2-\left(\tfrac{1}{\eta_\lambda}-\tfrac{L\kappa}{2}-\tfrac{L_P}{2}\right)\|\mathbf{\Lambda}^{t+1}-\mathbf{\Lambda}^{t}\|_F^2\\ +\tfrac{1}{2m}\left(L_P\|\mathbf{X}^{t+1}-\mathbf{X}_{\text{avg}}^{t}\|_F^2+\tfrac{1}{L_P}\|\mathbf{V}_\perp^{t}\|_F^2\right)+\tfrac{L(4\kappa+1)}{2m}\|\mathbf{X}_\perp^{t}\|_F^2+\tfrac{2\delta_t^2}{\mu}\\ -\tfrac{1}{2m\eta_x}\left(\|\mathbf{X}^{t+1}-\mathbf{X}_{\text{avg}}^{t}\|_F^2+\|\mathbf{X}^{t+1}-\widetilde{\mathbf{X}}^{t}\|_F^2-\|\widetilde{\mathbf{X}}^{t}-\mathbf{X}_{\text{avg}}^{t}\|_F^2\right).\end{aligned}\tag{3.30}$$

由于 $\mathbf{W}\mathbf{1}=\mathbf{1}$，因此有 $\mathbf{W}-\frac{1}{m}\mathbf{1}\mathbf{1}^\top=(\mathbf{W}-\frac{1}{m}\mathbf{1}\mathbf{1}^\top)(\mathbf{I}-\frac{1}{m}\mathbf{1}\mathbf{1}^\top)$。于是，$\|\widetilde{\mathbf{X}}^t-\mathbf{X}_{\text{avg}}^t\|_F=\|(\mathbf{W}-\frac{1}{m}\mathbf{1}\mathbf{1}^\top)\mathbf{X}_\perp^t\|_F \le \rho\|\mathbf{X}_\perp^t\|_F$。此外，利用事实 $\|\mathbf{X}_{\text{avg}}^{t+1}-\mathbf{X}_{\text{avg}}^{t}\|_F^2 \le \|\mathbf{X}^{t+1}-\mathbf{X}_{\text{avg}}^{t}\|_F^2$，并整理（3.30）中的各项，即可完成证明。□

下面，我们界定 $\{\mathbf{X}_\perp^t\}$ 和 $\{\mathbf{V}_\perp^t\}$ 然后与 (3.27) 结合，以展示迭代点上的平方可和结果。以下结果来自引理 C.7 以及引理 C.13 证明的末尾，在 [28]:

$$\|\mathbf{X}_\perp^{t+1}\|_F^2 \le \rho\|\mathbf{X}_\perp^{t}\|_F^2+\tfrac{\eta_x^2}{1-\rho}\|\mathbf{V}_\perp^{t}\|_F^2,\quad \|\mathbf{X}^{t+1}-\mathbf{X}^{t}\|_F^2 \le 2\|\mathbf{X}^{t+1}-\widetilde{\mathbf{X}}^{t}\|_F^2+8\|\mathbf{X}_\perp^{t}\|_F^2.\tag{3.31}$$

有了上述两个不等式，我们准备界定一致性误差。

LEMMA 3.6. 对于任意正整数 T，成立

$$\begin{aligned}(1-\rho)\textstyle\sum_{t=0}^{T-1}\|\mathbf{V}_\perp^{t}\|_F^2 \le \left\|\mathbf{V}_\perp^{0}\right\|_F^2+\tfrac{6mL^2\kappa^2}{1-\rho}\sum_{t=1}^{T-1}\|\mathbf{\Lambda}^{t}-\mathbf{\Lambda}^{t-1}\|_F^2\\ +\tfrac{2L^2(1+6\kappa^2)}{1-\rho}\sum_{t=1}^{T-1}\left(\|\mathbf{X}^{t}-\widetilde{\mathbf{X}}^{t-1}\|_F^2+4\|\mathbf{X}_\perp^{t-1}\|_F^2\right)+\tfrac{3m\kappa^2}{1-\rho}\sum_{t=1}^{T-1}(\delta_t^2+\delta_{t-1}^2),\end{aligned}\tag{3.32}$$

and

$$\begin{aligned}\left(1-\rho-\tfrac{8\eta_x^2L^2(1+6\kappa^2)}{(1-\rho)^3}\right)\textstyle\sum_{t=0}^{T}\|\mathbf{X}_\perp^{t}\|_F^2 \le \|\mathbf{X}_\perp^{0}\|_F^2+\tfrac{3m\eta_x^2\kappa^2}{(1-\rho)^3}\sum_{t=1}^{T-1}(\delta_t^2+\delta_{t-1}^2)\\ +\tfrac{\eta_x^2}{(1-\rho)^2}\left(\left\|\mathbf{V}_\perp^{0}\right\|_F^2+\tfrac{6mL^2\kappa^2}{1-\rho}\sum_{t=1}^{T-1}\|\mathbf{\Lambda}^{t}-\mathbf{\Lambda}^{t-1}\|_F^2+\tfrac{2L^2(1+6\kappa^2)}{1-\rho}\sum_{t=1}^{T-1}\|\mathbf{X}^{t}-\widetilde{\mathbf{X}}^{t-1}\|_F^2\right).\end{aligned}\tag{3.33}$$

证明。 根据 $\mathbf{1}^\top\mathbf{W}=\mathbf{1}^\top$ 和 $\mathbf{W}\mathbf{1}=\mathbf{1}$，有 $(\mathbf{I}-\frac{1}{m}\mathbf{1}\mathbf{1}^\top)\mathbf{W}=\mathbf{W}-\frac{1}{m}\mathbf{1}\mathbf{1}^\top=(\mathbf{W}-\frac{1}{m}\mathbf{1}\mathbf{1}^\top)(\mathbf{I}-\frac{1}{m}\mathbf{1}\mathbf{1}^\top)$。因此，$(\mathbf{I}-\frac{1}{m}\mathbf{1}\mathbf{1}^\top)\mathbf{W}\mathbf{V}^{t-1}=(\mathbf{W}-\frac{1}{m}\mathbf{1}\mathbf{1}^\top)\mathbf{V}_\perp^{t-1}$，由此根据（3.7b），可得对任意 $a>0$,

$$\begin{aligned}\|\mathbf{V}_\perp^{t}\|_F^2=\left\|(\mathbf{W}-\tfrac{1}{m}\mathbf{1}\mathbf{1}^\top)\mathbf{V}_\perp^{t-1}+(\mathbf{I}-\tfrac{1}{m}\mathbf{1}\mathbf{1}^\top)\left(\nabla_{\mathbf{x}}F(\mathbf{X}^{t},\mathbf{Y}^{t})-\nabla_{\mathbf{x}}F(\mathbf{X}^{t-1},\mathbf{Y}^{t-1})\right)\right\|_F^2\\ \le(1+a)\left\|(\mathbf{W}-\tfrac{1}{m}\mathbf{1}\mathbf{1}^\top)\mathbf{V}_\perp^{t-1}\right\|_F^2+(1+\tfrac{1}{a})\left\|\nabla_{\mathbf{x}}F(\mathbf{X}^{t},\mathbf{Y}^{t})-\nabla_{\mathbf{x}}F(\mathbf{X}^{t-1},\mathbf{Y}^{t-1})\right\|_F^2,\end{aligned}\tag{3.34}$$

其中，该不等式是利用杨氏不等式以及以下事实 $\|\mathbf{I}-\frac{1}{m}\mathbf{1}\mathbf{1}^\top\|_2\le 1$。将 $a=\frac{1}{\rho}-1$ 代入（3.34），根据 $\|\mathbf{W}-\frac{1}{m}\mathbf{1}\mathbf{1}^\top\|_2\le\rho$，并利用每个 -光滑性 f_i，可得

$$\begin{aligned}\|\mathbf{V}_\perp^t\|_F^2 &\le \rho\big\|\mathbf{V}_\perp^{t-1}\big\|_F^2+\tfrac{1}{1-\rho}\big\|\nabla_{\mathbf{x}}F(\mathbf{X}^t,\ \mathbf{Y}^t)\ -\nabla_{\mathbf{x}}F(\mathbf{X}^{t-1},\ \mathbf{Y}^{t-1})\big\|_F^2\\ &\le \rho\big\|\mathbf{V}_\perp^{t-1}\big\|_F^2+\tfrac{L^2}{1-\rho}\big(\|\mathbf{X}^t-\mathbf{X}^{t-1}\|_F^2+\|\mathbf{Y}^t-\mathbf{Y}^{t-1}\|_F^2\big).\end{aligned}\tag{3.35}$$

此外，由（3.8c）、命题 3.2和杨氏不等式可得

$$\begin{aligned}&\|\mathbf{Y}^t-\mathbf{Y}^{t-1}\|_F^2\\ =&\|\mathbf{Y}^t-S_\Phi(\mathbf{X}^t,\mathbf{\Lambda}^t)+S_\Phi(\mathbf{X}^t,\mathbf{\Lambda}^t)-S_\Phi(\mathbf{X}^{t-1},\mathbf{\Lambda}^{t-1})+S_\Phi(\mathbf{X}^{t-1},\mathbf{\Lambda}^{t-1})-\mathbf{Y}^{t-1}\|_F^2\\ \le&3\|\mathbf{Y}^t-S_\Phi(\mathbf{X}^t,\mathbf{\Lambda}^t)\|_F^2+3\|S_\Phi(\mathbf{X}^t,\mathbf{\Lambda}^t)-S_\Phi(\mathbf{X}^{t-1},\mathbf{\Lambda}^{t-1})\|_F^2+3\|S_\Phi(\mathbf{X}^{t-1},\mathbf{\Lambda}^{t-1})-\mathbf{Y}^{t-1}\|_F^2\\ \le&\tfrac{3m\delta_t^2}{\mu^2}+\tfrac{3m\delta_{t-1}^2}{\mu^2}+6\kappa^2\|\mathbf{X}^t-\mathbf{X}^{t-1}\|_F^2+6m\kappa^2\|\mathbf{\Lambda}^t-\mathbf{\Lambda}^{t-1}\|_F^2.\end{aligned}\tag{3.36}$$

因此，（3.35）推出

$$\begin{aligned}\|\mathbf{V}_\perp^t\|_F^2 &\le \rho\big\|\mathbf{V}_\perp^{t-1}\big\|_F^2+\tfrac{L^2}{1-\rho}\Big((1+6\kappa^2)\|\mathbf{X}^t-\mathbf{X}^{t-1}\|_F^2+6m\kappa^2\|\mathbf{\Lambda}^t-\mathbf{\Lambda}^{t-1}\|_F^2+\tfrac{3m\delta_t^2}{\mu^2}+\tfrac{3m\delta_{t-1}^2}{\mu^2}\Big)\\ &\overset{(3.31)}{\le} \rho\big\|\mathbf{V}_\perp^{t-1}\big\|_F^2+\tfrac{6mL^2\kappa^2}{1-\rho}\|\mathbf{\Lambda}^t-\mathbf{\Lambda}^{t-1}\|_F^2+\tfrac{2L^2(1+6\kappa^2)}{1-\rho}\big(\|\mathbf{X}^t-\widetilde{\mathbf{X}}^{t-1}\|_F^2+4\|\mathbf{X}_\perp^{t-1}\|_F^2\big)+\tfrac{3m\kappa^2}{1-\rho}(\delta_t^2+\delta_{t-1}^2).\end{aligned}\tag{3.37}$$

对（3.37）从 $t=1$ 到 $T-1$求和，可得（3.32）。现在对（3.31）中的第一个不等式从 $t=0$ 到 $T-1$ 求和，并利用（3.32），得到

$$\begin{aligned}&\|\mathbf{X}_\perp^T\|_F^2+(1-\rho)\textstyle\sum_{t=0}^{T-1}\|\mathbf{X}_\perp^t\|_F^2\le\|\mathbf{X}_\perp^0\|_F^2+\tfrac{\eta_x^2}{1-\rho}\sum_{t=0}^{T-1}\|\mathbf{V}_\perp^t\|_F^2\\ \le&\|\mathbf{X}_\perp^0\|_F^2+\tfrac{\eta_x^2}{(1-\rho)^2}\Big(\big\|\mathbf{V}_\perp^0\big\|_F^2+\tfrac{6mL^2\kappa^2}{1-\rho}\textstyle\sum_{t=1}^{T-1}\|\mathbf{\Lambda}^t-\mathbf{\Lambda}^{t-1}\|_F^2\\ &+\tfrac{2L^2(1+6\kappa^2)}{1-\rho}\textstyle\sum_{t=1}^{T-1}\big(\|\mathbf{X}^t-\widetilde{\mathbf{X}}^{t-1}\|_F^2+4\|\mathbf{X}_\perp^{t-1}\|_F^2\big)+\tfrac{3m\kappa^2}{1-\rho}\textstyle\sum_{t=1}^{T-1}(\delta_t^2+\delta_{t-1}^2)\Big),\end{aligned}\tag{3.38}$$

这显然表明（3.33）。证明完毕。 □

By T定理 3.5 和引理 3.6，我们可以很容易地证明以下平方可和结果。

THEOREM 3.7. 设 $\{(\mathbf{X}^t,\mathbf{\Lambda}^t,\ \mathbf{V}^t)\}$ 由算法 *1*生成。那么对任意正整数 T，均有

$$\begin{aligned}&\alpha_x\textstyle\sum_{t=0}^{T-1}\|\mathbf{X}^{t+1}-\mathbf{X}_{\text{avg}}^t\|^2+\tilde{\alpha}_x\sum_{t=0}^{T-1}\|\mathbf{X}^{t+1}-\widetilde{\mathbf{X}}^t\|_F^2+\alpha_\lambda\sum_{t=0}^{T-1}\|\mathbf{\Lambda}^{t+1}-\mathbf{\Lambda}^t\|_F^2\\ \le&\phi(\mathbf{x}_{\text{avg}}^0,\mathbf{\Lambda}^0)-\phi(\mathbf{x}_{\text{avg}}^T,\mathbf{\Lambda}^T)+c\|\mathbf{X}_\perp^0\|_F^2+\Big(\tfrac{1}{2mL_P(1-\rho)}+\tfrac{c\eta_x^2}{(1-\rho)^2}\Big)\|\mathbf{V}_\perp^0\|_F^2\\ &+\Big(\tfrac{3\kappa^2}{2L_P(1-\rho)^2}+\tfrac{3cm\eta_x^2\kappa^2}{(1-\rho)^3}\Big)\textstyle\sum_{t=1}^{T-1}(\delta_t^2+\delta_{t-1}^2)+\sum_{t=0}^{T-1}\frac{2\delta_t^2}{\mu},\end{aligned}\tag{3.39}$$

其中，常数 α_x、$\tilde{\alpha}_x$、α_λ 和 c 定义为

$$\alpha_x=\tfrac{1}{2m}\big(\tfrac{1}{\eta_x}-2L_P-L(\kappa+1)\big),\quad \tilde{\alpha}_x=\tfrac{1}{2m\eta_x}-\tfrac{L^2(1+6\kappa^2)}{mL_P(1-\rho)^2}-\tfrac{2cL^2(1+6\kappa^2)\eta_x^2}{(1-\rho)^3}\tag{3.40a}$$

$$\alpha_\lambda=\big(\tfrac{1}{\eta_\lambda}-\tfrac{L\kappa}{2}-\tfrac{L_P}{2}\big)-\tfrac{3L^2\kappa^2}{L_P(1-\rho)^2}-\tfrac{6cmL^2\kappa^2\eta_x^2}{(1-\rho)^3},\quad c=\frac{\frac{\rho^2}{2m\eta_x}+\frac{L(4\kappa+1)}{2m}+\frac{4L^2(1+6\kappa^2)}{mL_P(1-\rho)^2}}{1-\rho-\frac{8\eta_x^2L^2(1+6\kappa^2)}{(1-\rho)^3}}.\tag{3.40b}$$

证明。 将（3.27）对 $t=0$ 至 $T-1$ 求和，并使用（3.32）。可得

$$\begin{aligned}
&\tfrac{1}{2m}\big(\tfrac{1}{\eta_x}-2L_P-L(\kappa+1)\big)\textstyle\sum_{t=0}^{T-1}\|\mathbf{X}^{t+1}-\mathbf{X}^t_{\rm avg}\|^2+\big(\tfrac{1}{\eta_\lambda}-\tfrac{L\kappa}{2}-\tfrac{L_P}{2}\big)\sum_{t=0}^{T-1}\|\mathbf{\Lambda}^{t+1}-\mathbf{\Lambda}^t\|_F^2\\
\le\ &\phi(\mathbf{x}^0_{\rm avg},\mathbf{\Lambda}^0)-\phi(\mathbf{x}^T_{\rm avg},\mathbf{\Lambda}^T)+\big(\tfrac{\rho^2}{2m\eta_x}+\tfrac{L(4\kappa+1)}{2m}\big)\textstyle\sum_{t=0}^{T-1}\|\mathbf{X}^t_\perp\|_F^2-\frac{1}{2m\eta_x}\sum_{t=0}^{T-1}\|\mathbf{X}^{t+1}-\widetilde{\mathbf{X}}^t\|_F^2\\
&+\textstyle\sum_{t=0}^{T-1}\frac{2\delta_t^2}{\mu}+\frac{1}{2mL_P(1-\rho)}\Big(\big\|\mathbf{V}^0_\perp\big\|_F^2+\frac{6mL^2\kappa^2}{1-\rho}\sum_{t=1}^{T-1}\|\mathbf{\Lambda}^t-\mathbf{\Lambda}^{t-1}\|_F^2\\
&\qquad+\tfrac{2L^2(1+6\kappa^2)}{1-\rho}\textstyle\sum_{t=1}^{T-1}\big(\|\mathbf{X}^t-\widetilde{\mathbf{X}}^{t-1}\|_F^2+4\|\mathbf{X}^{t-1}_\perp\|_F^2\big)+\frac{3m\kappa^2}{1-\rho}\sum_{t=1}^{T-1}(\delta_t^2+\delta_{t-1}^2)\Big)\\
\le\ &\phi(\mathbf{x}^0_{\rm avg},\mathbf{\Lambda}^0)-\phi(\mathbf{x}^T_{\rm avg},\mathbf{\Lambda}^T)+\tfrac{1}{2mL_P(1-\rho)}\big\|\mathbf{V}^0_\perp\big\|_F^2-\Big(\tfrac{1}{2m\eta_x}-\tfrac{L^2(1+6\kappa^2)}{mL_P(1-\rho)^2}\Big)\textstyle\sum_{t=0}^{T-1}\|\mathbf{X}^{t+1}-\widetilde{\mathbf{X}}^t\|_F^2\\
&+\tfrac{3L^2\kappa^2}{L_P(1-\rho)^2}\textstyle\sum_{t=1}^{T-1}\|\mathbf{\Lambda}^t-\mathbf{\Lambda}^{t-1}\|_F^2+\Big(\tfrac{\rho^2}{2m\eta_x}+\tfrac{L(4\kappa+1)}{2m}+\tfrac{4L^2(1+6\kappa^2)}{mL_P(1-\rho)^2}\Big)\textstyle\sum_{t=0}^{T-1}\|\mathbf{X}^t_\perp\|_F^2\\
&+\tfrac{3\kappa^2}{2L_P(1-\rho)^2}\textstyle\sum_{t=1}^{T-1}(\delta_t^2+\delta_{t-1}^2)+\sum_{t=0}^{T-1}\frac{2\delta_t^2}{\mu}.
\end{aligned}\tag{3.41}$$

现在利用（3.33）对 $\sum_{t=0}^{T-1}\|\mathbf{X}^t_\perp\|_F^2$ 进行界定，并将其代入（3.41）。我们得到

$$\begin{aligned}
&\tfrac{1}{2m}\big(\tfrac{1}{\eta_x}-2L_P-L(\kappa+1)\big)\textstyle\sum_{t=0}^{T-1}\|\mathbf{X}^{t+1}-\mathbf{X}^t_{\rm avg}\|^2+\big(\tfrac{1}{\eta_\lambda}-\tfrac{L\kappa}{2}-\tfrac{L_P}{2}\big)\sum_{t=0}^{T-1}\|\mathbf{\Lambda}^{t+1}-\mathbf{\Lambda}^t\|_F^2\\
\le\ &\phi(\mathbf{x}^0_{\rm avg},\mathbf{\Lambda}^0)-\phi(\mathbf{x}^T_{\rm avg},\mathbf{\Lambda}^T)+\tfrac{1}{2mL_P(1-\rho)}\big\|\mathbf{V}^0_\perp\big\|_F^2-\Big(\tfrac{1}{2m\eta_x}-\tfrac{L^2(1+6\kappa^2)}{mL_P(1-\rho)^2}\Big)\textstyle\sum_{t=0}^{T-1}\|\mathbf{X}^{t+1}-\widetilde{\mathbf{X}}^t\|_F^2\\
&+\tfrac{3L^2\kappa^2}{L_P(1-\rho)^2}\textstyle\sum_{t=1}^{T-1}\|\mathbf{\Lambda}^t-\mathbf{\Lambda}^{t-1}\|_F^2+\sum_{t=0}^{T-1}\frac{2\delta_t^2}{\mu}+\Big(\frac{3\kappa^2}{2L_P(1-\rho)^2}+\frac{3cm\eta_x^2\kappa^2}{(1-\rho)^3}\Big)\sum_{t=1}^{T-1}(\delta_t^2+\delta_{t-1}^2)+c\|\mathbf{X}^0_\perp\|_F^2\\
&+\tfrac{c\eta_x^2}{(1-\rho)^2}\Big(\big\|\mathbf{V}^0_\perp\big\|_F^2+\tfrac{6mL^2\kappa^2}{1-\rho}\textstyle\sum_{t=1}^{T-1}\|\mathbf{\Lambda}^t-\mathbf{\Lambda}^{t-1}\|_F^2+\frac{2L^2(1+6\kappa^2)}{1-\rho}\sum_{t=1}^{T-1}\|\mathbf{X}^t-\widetilde{\mathbf{X}}^{t-1}\|_F^2\Big),
\end{aligned}$$

其中 c 在d 义于（3.40b）。合并上面不等式中的同类项，即得所需结果 t. □

在下面的引理中，我们指定步长，使（3.39）左侧的系数为正。其证明只涉及基本（但繁琐的）代数运算，因此为便于阅读，我们将其放在附录中。

LEMMA 3.8. 设步长参数 η_x 和 η_λ 取为

$$\eta_x=\tfrac{(1-\rho)^2}{5L\sqrt{1+6\kappa^2}},\quad \eta_\lambda=\tfrac{(1-\rho)^2}{L(9\kappa+2)}.\tag{3.42}$$

设 c、α_x、$\tilde{\alpha}_x$ 和 α_λ 定义于 (3.40)。则有

$$\alpha_x\ge\tfrac{L(4\kappa+1)}{2m(1-\rho)^2},\ \tilde{\alpha}_x\ge\tfrac{3L(1+4\kappa)}{25m(1-\rho)^2},\ \alpha_\lambda\ge\tfrac{L(4\kappa+1)}{(1-\rho)^2},\tag{3.43a}$$

$$\tfrac{3\kappa^2}{2L_P(1-\rho)^2}+\tfrac{3cm\eta_x^2\kappa^2}{(1-\rho)^3}\le\tfrac{3\kappa}{2L(1-\rho)^2},\ \tfrac{1}{2mL_P(1-\rho)}+\tfrac{c\eta_x^2}{(1-\rho)^2}\le\tfrac{1}{2mL\kappa(1-\rho)}.\tag{3.43b}$$

REMARK 3.1. 式 (3.42) 中的步长比 *[20]*中的 *GDA* 方法和 *[4,10,49],* 中的去中心化梯度类方法的步长大 κ 或 κ^2倍。这将使我们的方法具有更好的数值性能和更低的复杂度，尤其是在 κ 较大时；数值结果见第 *4*节。

下面， 我们给出误差容限 $\{\delta_t\}$ 平方可和时的全局收敛结果 .

THEOREM 3.9 (全局收敛). 在假设 *1–3* 下，设 $\{(X^t,\mathbf{\Lambda}^t)\}_{t\ge0}$ 由算法 *1* 在 η_x 和 η_λ 取为式 *(3.42)* 中的值时生成。假设 $\sum_{t=0}^{\infty}\delta_t^2<+\infty$。则该序列 $\{(X^t,\mathbf{\Lambda}^t)\}_{t\ge0}$ 的任意极限点 *(* $\bar{X}$ $\bar{\mathbf{\Lambda}}$*)* 满足：*(i)* $\bar{X}_\perp=\mathbf{0}$；*(ii)* *(*$\bar{x}_{\rm avg}$, $\bar{\mathbf{\Lambda}}$*)* 是 ϕ的一个驻点解。

证明。 当 η_x 和 η_λ 取为（3.42）中的值时，由（3.43a）可得 α_x、$\tilde{\alpha}_x$ 和 α_λ 均为正。因此，在（ $T\to\infty$ ）中取 3.39，并利用 ϕ 的下有界性，可得

$$\textstyle\sum_{t=0}^{\infty}\|\mathbf{X}^{t+1}-\mathbf{X}^t_{\rm avg}\|_F^2+\sum_{t=0}^{\infty}\|\mathbf{X}^{t+1}-\widetilde{\mathbf{X}}^t\|_F^2+\sum_{t=0}^{\infty}\|\mathbf{\Lambda}^{t+1}-\mathbf{\Lambda}^t\|_F^2<\infty,\tag{3.44}$$

结合（3.32）和（3.33）可推出 $\sum_{t=0}^{\infty}\|\mathbf{V}_{\perp}^{t}\|_F^2<\infty$ 以及 $\sum_{t=0}^{\infty}\|\mathbf{X}_{\perp}^{t}\|_F^2<\infty$。因此，$\mathbf{X}_{\perp}^{t}\to\mathbf{0}$ 当 $t\to\infty$ 时。故极限点必满足 $\mathbf{X}_{\perp}=\mathbf{0}$。

此外，我们有 $-\frac{L}{2\sqrt{m}}(\mathbf{W}-\mathbf{I})\mathbf{Y}^{t}=\mathbf{\Lambda}^{t+1}-\mathbf{\Lambda}^{t}\to\mathbf{0}$ 根据 $t\to\infty$ 式（3.44）。由式（3.21）可得

$$\|\widehat{\mathbf{Y}}^{t}-\mathbf{Y}^{t}\|_F\le\sqrt{2}\kappa\|\mathbf{X}_{\perp}^{t}\|_F+\frac{\sqrt{2m}\delta_t}{\mu}\to 0. \tag{3.45}$$

因此，

$$\nabla_{\mathbf{\Lambda}}P(\mathbf{x}_{\mathrm{avg}}^{t},\mathbf{\Lambda}^{t})=-\tfrac{L}{2\sqrt{m}}(\mathbf{W}-\mathbf{I})\widehat{\mathbf{Y}}^{t}=-\tfrac{L}{2\sqrt{m}}(\mathbf{W}-\mathbf{I})\mathbf{Y}^{t}+\tfrac{L}{2\sqrt{m}}(\mathbf{W}-\mathbf{I})(\mathbf{Y}^{t}-\widehat{\mathbf{Y}}^{t})\to\mathbf{0},\ \text{as } t\to\infty. \tag{3.46}$$

此外，由式（3.44）可知 $\mathbf{X}^{t+1}-\mathbf{X}_{\mathrm{avg}}^{t}\to\mathbf{0}$ 且 $\mathbf{X}^{t+1}-\tilde{\mathbf{X}}^{t}\to\mathbf{0}$，由于 $t\to\infty$。因此，由 $\mathbf{X}_{\perp}^{t}\to\mathbf{0}$ 可得 $\mathbf{x}_i^{t+1}-\mathbf{x}_{\mathrm{avg}}^{t}\to\mathbf{0}$ 且 $\tilde{\mathbf{x}}_i^{t}-\mathbf{x}_{\mathrm{avg}}^{t}\to\mathbf{0}$ 对任意 $i\in[m]$ 成立。因此，由式（3.7a）可得

$$\mathbf{x}_{\mathrm{avg}}^{t}-\mathbf{prox}_{\eta_x g}\big(\mathbf{x}_{\mathrm{avg}}^{t}-\eta_x\mathbf{v}_i^{t}\big)\to\mathbf{0},\ \text{as } t\to\infty,\forall\, i\in[m]. \tag{3.47}$$

此外，$\mathbf{V}_{\perp}^{t}\to\mathbf{0}$ 因此 $\mathbf{v}_i^{t}-\mathbf{v}_{\mathrm{avg}}^{t}\to\mathbf{0}$ 由于 $t\to\infty$。现在注意到 $\mathbf{v}_{\mathrm{avg}}^{t}=\frac{1}{m}\sum_{i=1}^{m}\nabla_{\mathbf{x}}f_i(\mathbf{x}_i^{t},\ \mathbf{y}_i^{t})$ 以及

$$\begin{aligned}&\|\tfrac{1}{m}\textstyle\sum_{i=1}^{m}\nabla_{\mathbf{x}}f_i(\mathbf{x}_i^{t},\mathbf{y}_i^{t})-\tfrac{1}{m}\sum_{i=1}^{m}\nabla_{\mathbf{x}}f_i(\mathbf{x}_{\mathrm{avg}}^{t},\hat{\mathbf{y}}_i^{t})\|^2\le\tfrac{L^2}{m}\textstyle\sum_{i=1}^{m}\big(\|\mathbf{x}_i^{t}-\mathbf{x}_{\mathrm{avg}}^{t}\|^2+\|\mathbf{y}_i^{t}-\hat{\mathbf{y}}_i^{t}\|^2\big)\\&=\tfrac{L^2}{m}\big(\|\mathbf{X}_{\perp}^{t}\|_F^2+\|\mathbf{Y}^{t}-\widehat{\mathbf{Y}}^{t}\|_F^2\big)\overset{(3.45)}{\longrightarrow}0,\ \text{as } t\to\infty.\end{aligned} \tag{3.48}$$

因此，$\mathbf{v}_{\mathrm{avg}}^{t}-\nabla_{\mathbf{x}}P(\mathbf{x}_{\mathrm{avg}}^{t},\mathbf{\Lambda}^{t})=\mathbf{v}_{\mathrm{avg}}^{t}-\frac{1}{m}\sum_{i=1}^{m}\nabla_{\mathbf{x}}f_i(\mathbf{x}_{\mathrm{avg}}^{t},\hat{\mathbf{y}}_i^{t})\to\mathbf{0}$，从而 $\mathbf{v}_i^{t}-\nabla_{\mathbf{x}}P(\mathbf{x}_{\mathrm{avg}}^{t},\mathbf{\Lambda}^{t})\to\mathbf{0}$ 对任意 i 均成立。因此，式（3.47）表明 $\mathbf{x}_{\mathrm{avg}}^{t}$ 近端算子 $_{\eta_x g}\big(\mathbf{x}_{\mathrm{avg}}^{t}-\eta_x\nabla_{\mathbf{x}}P(\mathbf{x}_{\mathrm{avg}}^{t},\mathbf{\Lambda}^{t})\big)\to\mathbf{0}$，结合式（3.46）可得（$\bar{\mathbf{x}}_{\mathrm{avg}},\mathbf{\Lambda}$）的驻点性。证明完毕。□

接下来，我们对每个序列的有限平方和进行界定。

LEMMA 3.10. 设 $\{(\mathbf{X}^{t},\mathbf{\Lambda}^{t},\mathbf{V}^{t})\}$ 由算法 *1* 生成，其中 η_x 和 η_λ 取为 (3.42)中的值。则

$$\textstyle\sum_{t=0}^{T-1}\|\mathbf{X}^{t+1}-\mathbf{X}_{\mathrm{avg}}^{t}\|_F^2\le\frac{2mC_{0,T}(1-\rho)^2}{L(4\kappa+1)} \tag{3.49a}$$

$$\textstyle\sum_{t=0}^{T-1}\|\mathbf{X}^{t+1}-\widetilde{\mathbf{X}}^{t}\|_F^2\le\frac{25mC_{0,T}(1-\rho)^2}{3L(1+4\kappa)}, \tag{3.49b}$$

$$\textstyle\sum_{t=0}^{T-1}\|\mathbf{\Lambda}^{t+1}-\mathbf{\Lambda}^{t}\|_F^2\le\frac{C_{0,T}(1-\rho)^2}{L(4\kappa+1)}, \tag{3.49c}$$

$$\textstyle\sum_{t=0}^{T-1}\|\mathbf{X}_{\perp}^{t}\|_F^2\le\frac{3(1-\rho)}{50L^2(1+6\kappa^2)}\big\|\mathbf{V}_{\perp}^{0}\big\|_F^2+\frac{53mC_{0,T}(1-\rho)^2}{50L(4\kappa+1)}+\frac{3m}{50L^2}\Delta_T, \tag{3.49d}$$

$$\textstyle\sum_{t=0}^{T-1}\|\mathbf{V}_{\perp}^{t}\|_F^2\le\frac{3}{2(1-\rho)}\big\|\mathbf{V}_{\perp}^{0}\big\|_F^2+\frac{m(1+9\kappa^2)}{(1-\rho)^2}\Delta_T+40mC_{0,T}L\kappa, \tag{3.49e}$$

其中 $\Delta_T=\sum_{t=0}^{T-1}\delta_t^2$ 且 $C_{0,T}=\phi(\mathbf{x}_{\mathrm{avg}}^{0},\mathbf{\Lambda}^{0})-(p+g)(\mathbf{x}^{*})+\frac{1}{2mL\kappa(1-\rho)}\|\mathbf{V}_{\perp}^{0}\|_F^2+\frac{5\kappa}{L(1-\rho)^2}\Delta_T$。证明。注意到 $\min_{\mathbf{\Lambda}}P(\mathbf{x},\mathbf{\Lambda})=p(\mathbf{x})$。因此，$\phi(\mathbf{x},\mathbf{\Lambda})\ge(p+g)(\mathbf{x}^{*})$，从而由 (3.39)、(3.43b) 中的两个结果、设定 $\mathbf{x}_i^{0}=\mathbf{x}^{0},\forall\, i\in[m]$ 和 $\frac{1}{\mu}=\frac{\kappa}{L}$，可得

$$\alpha_x\textstyle\sum_{t=0}^{T-1}\|\mathbf{X}^{t+1}-\mathbf{X}_{\mathrm{avg}}^{t}\|^2+\tilde{\alpha}_x\sum_{t=0}^{T-1}\|\mathbf{X}^{t+1}-\widetilde{\mathbf{X}}^{t}\|_F^2+\alpha_\lambda\sum_{t=0}^{T-1}\|\mathbf{\Lambda}^{t+1}-\mathbf{\Lambda}^{t}\|_F^2\le C_{0,T}. \tag{3.50}$$

由 α_x、$\tilde{\alpha}_x$ 和 α_λ 在（3.43a）中的界，根据上述不等式，我们立即得到（3.49a）–（3.49c）。利用（A.1）中的三个不等式，由（3.33）可得

$$\textstyle\sum_{t=0}^{T}\|\mathbf{X}_{\perp}^{t}\|_F^2\le\frac{3\eta_x^2}{2(1-\rho)^3}\big\|\mathbf{V}_{\perp}^{0}\big\|_F^2+\frac{9m\eta_x^2\kappa^2}{2(1-\rho)^4}\sum_{t=1}^{T-1}(\delta_t^2+\delta_{t-1}^2)+\frac{3m}{50}\sum_{t=1}^{T-1}\|\mathbf{\Lambda}^{t}-\mathbf{\Lambda}^{t-1}\|_F^2+\frac{3}{25}\sum_{t=1}^{T-1}\|\mathbf{X}^{t}-\widetilde{\mathbf{X}}^{t-1}\|_F^2.$$

将（3.49b）和（3.49c）代入上述不等式，得到

$$\sum_{t=0}^{T}\|\mathbf{X}_\perp^t\|_F^2 \le \tfrac{3\eta_x^2}{2(1-\rho)^3}\big\|\mathbf{V}_\perp^0\big\|_F^2 + \tfrac{9m\eta_x^2\kappa^2}{2(1-\rho)^4}\sum_{t=1}^{T-1}(\delta_t^2+\delta_{t-1}^2) + \tfrac{3mC_{0,T}(1-\rho)^2}{50L(4\kappa+1)} + \tfrac{mC_{0,T}(1-\rho)^2}{L(1+4\kappa)},$$

这意味着（3.49d），这是由 η_x的选取可得。因此，由（3.49b）–（3.49d），由（3.32）可得

$$\begin{aligned}\sum_{t=0}^{T-1}\|\mathbf{V}_\perp^t\|_F^2 &\le \tfrac{1}{1-\rho}\Big(\big\|\mathbf{V}_\perp^0\big\|_F^2 + \tfrac{6mL^2\kappa^2}{1-\rho}\sum_{t=1}^{T-1}\|\mathbf{\Lambda}^t-\mathbf{\Lambda}^{t-1}\|_F^2\\ &\qquad + \tfrac{2L^2(1+6\kappa^2)}{1-\rho}\sum_{t=1}^{T-1}\big(\|\mathbf{X}^t-\widetilde{\mathbf{X}}^{t-1}\|_F^2+4\|\mathbf{X}_\perp^{t-1}\|_F^2\big)+\tfrac{6m\kappa^2}{1-\rho}\Delta_T\Big)\\ &\le \tfrac{1}{1-\rho}\big\|\mathbf{V}_\perp^0\big\|_F^2 + \tfrac{6mL^2\kappa^2}{(1-\rho)^2}\tfrac{C_{0,T}(1-\rho)^2}{L(4\kappa+1)} + \tfrac{2L^2(1+6\kappa^2)}{(1-\rho)^2}\tfrac{25mC_{0,T}(1-\rho)^2}{3L(1+4\kappa)} + \tfrac{6m\kappa^2}{(1-\rho)^2}\Delta_T\\ &\quad + \tfrac{8L^2(1+6\kappa^2)}{(1-\rho)^2}\Big(\tfrac{3(1-\rho)}{50L^2(1+6\kappa^2)}\big\|\mathbf{V}_\perp^0\big\|_F^2 + \tfrac{53mC_{0,T}(1-\rho)^2}{50L(4\kappa+1)} + \tfrac{3m}{50L^2}\Delta_T\Big)\\ &\le \tfrac{3}{2(1-\rho)}\big\|\mathbf{V}_\perp^0\big\|_F^2 + \tfrac{m(1+9\kappa^2)}{(1-\rho)^2}\Delta_T + \tfrac{6mC_{0,T}L\kappa^2}{4\kappa+1} + \tfrac{50mC_{0,T}L(1+6\kappa^2)}{3(1+4\kappa)} + \tfrac{9mC_{0,T}L(1+6\kappa^2)}{1+4\kappa},\end{aligned}$$

这意味着（3.49e）。至此，证明完成。 □

下面，我们准备基于驻点性违反量给出收敛速率结果

THEOREM 3.11. 在假设 1–3下，令 $\{(\mathbf{X}^t,\mathbf{\Lambda}^t)\}_{t=0}^T$ 由算法 1 生成，且 η_x 和 η_λ 取为（(3.42)）中的值，其中 T 为正整数。从 τ 中 $\{0,\dots,T-1\}$ 均匀随机抽取。则

$$\mathbb{E}_\tau\|\nabla_{\mathbf{\Lambda}}P(\mathbf{x}_{\rm avg}^\tau,\mathbf{\Lambda}^\tau)\|_F \le \tfrac{7\sqrt{L\kappa C_{0,T}}}{(1-\rho)\sqrt{T}} + \tfrac{\sqrt{2(1-\rho)}}{10\sqrt{mT}}\big\|\mathbf{V}_\perp^0\big\|_F + \tfrac{2\kappa}{\sqrt{T}}\sqrt{\Delta_T}, \tag{3.51a}$$

$$\tfrac{1}{\eta_x}\mathbb{E}_\tau\big\|\mathbf{x}_{\rm avg}^\tau - \mathbf{prox}_{\eta_x g}\big(\mathbf{x}_{\rm avg}^\tau - \eta_x\nabla_{\mathbf{x}}P(\mathbf{x}_{\rm avg}^\tau,\mathbf{\Lambda}^\tau)\big)\big\| \le \tfrac{2}{\sqrt{T}}\sqrt{C_{1,T}+\tfrac{156C_{0,T}L\kappa}{(1-\rho)^2}}, \tag{3.51b}$$

其中 $C_{0,T}$ 定义见引理 3.10，并且

$$C_{1,T} = \tfrac{1}{m}\left(\tfrac{3}{2(1-\rho)} + \tfrac{3(1-\rho)(1+2\kappa^2)}{50(1+6\kappa^2)} + \tfrac{3}{2(1-\rho)^2}\right)\big\|\mathbf{V}_\perp^0\big\|_F^2 + \left(\tfrac{3(1+2\kappa^2)}{50} + \tfrac{3(1+6\kappa^2)}{2(1-\rho)^4} + \tfrac{1+9\kappa^2}{(1-\rho)^2} + 2\kappa^2\right)\Delta_T. \tag{3.52}$$

证明。 根据 $\widehat{\mathbf{Y}}^t$ 在引理 3.4 的证明中和（3.16）的定义，利用三角不等式可得

$$\begin{aligned}\|\nabla_{\mathbf{\Lambda}}P(\mathbf{x}_{\rm avg}^\tau,\mathbf{\Lambda}^\tau)\|_F &\le \tfrac{L}{2\sqrt{m}}\|(\mathbf{W}-\mathbf{I})\mathbf{Y}^\tau\|_F + \tfrac{L}{2\sqrt{m}}\|(\mathbf{W}-\mathbf{I})(\mathbf{Y}^\tau-\widehat{\mathbf{Y}}^\tau)\|_F\\ &= \tfrac{1}{\eta_\lambda}\|\mathbf{\Lambda}^{\tau+1}-\mathbf{\Lambda}^\tau\|_F + \tfrac{L}{2\sqrt{m}}\|(\mathbf{W}-\mathbf{I})(\mathbf{Y}^\tau-\widehat{\mathbf{Y}}^\tau)\|_F.\end{aligned} \tag{3.53}$$

此外，由（3.21）和 $\|\mathbf{W}-\mathbf{I}\|_2\le 2$ 可得 $\|(\mathbf{W}-\mathbf{I})(\mathbf{Y}^\tau-\widehat{\mathbf{Y}}^\tau)\|_F \le 2\sqrt{2\kappa}\|\mathbf{X}_\perp^\tau\|_F + \frac{2\sqrt{2m\bar{\delta}_\tau}}{\mu}$。根据 τ的选取以及（3.49c），有 $\mathbb{E}_\tau\|\mathbf{\Lambda}^{\tau+1}-\mathbf{\Lambda}^\tau\|_F \le \sqrt{\mathbb{E}_\tau\big[\|\mathbf{\Lambda}^{\tau+1}-\mathbf{\Lambda}^\tau\|_F^2\big]} \le \sqrt{\frac{C_{0,T}}{TL}\frac{1-\rho}{4\kappa+1}}$，类似地，由（3.49d）可得 $\mathbb{E}_\tau\|\mathbf{X}_\perp^\tau\|_F \le \frac{1}{\sqrt{T}}\sqrt{\frac{3}{50L^2}(1-\rho)(1+6\kappa^2)\big\|\mathbf{V}_\perp^0\big\|_F^2 + \frac{53mC_{0,T}}{50L}\frac{1-\rho}{4\kappa+1} + \frac{3m}{50L^2}\Delta_T}$。此外，$\mathbb{E}_\tau[\delta_\tau] \le \sqrt{\mathbb{E}_\tau[\delta_\tau^2]} = \frac{1}{\sqrt{T}}\sqrt{\Delta_T}$。因此，由（3.42）给出的 η_λ，不等式（3.53）表明

$$\begin{aligned}&\mathbb{E}_\tau\|\nabla_{\mathbf{\Lambda}}P(\mathbf{x}_{\rm avg}^\tau,\mathbf{\Lambda}^\tau)\|_F\\ &\le \tfrac{L(9\kappa+2)}{(1-\rho)^2}\sqrt{\tfrac{C_{0,T}(1-\rho)^2}{TL(4\kappa+1)}} + \tfrac{\sqrt{2}L\kappa}{\sqrt{mT}}\sqrt{\tfrac{3(1-\rho)}{50L^2(1+6\kappa^2)}\big\|\mathbf{V}_\perp^0\big\|_F^2 + \tfrac{53mC_{0,T}(1-\rho)^2}{50L(4\kappa+1)} + \tfrac{3m}{50L^2}\Delta_T} + \tfrac{\sqrt{2}\kappa}{\sqrt{T}}\sqrt{\Delta_T}\\ &\le \tfrac{9\sqrt{LC_{0,T}(4\kappa+1)}}{4(1-\rho)\sqrt{T}} + \tfrac{\sqrt{2(1-\rho)}}{10\sqrt{mT}}\big\|\mathbf{V}_\perp^0\big\|_F + \tfrac{(1-\rho)\sqrt{L\kappa}}{\sqrt{T}}\sqrt{C_{0,T}} + \tfrac{\kappa\sqrt{3\Delta_T}}{5\sqrt{T}} + \tfrac{\sqrt{2}\kappa}{\sqrt{T}}\sqrt{\Delta_T},\end{aligned} \tag{3.54}$$

其中最后一个不等式由 $\frac{9\kappa+2}{\sqrt{4\kappa+1}} \le \frac{9}{4}\sqrt{4\kappa+1}$、$\sqrt{a+b} \le \sqrt{a}+\sqrt{b}$、$\forall a、\ b \in \mathbb{R}_+$、$\frac{\kappa}{\sqrt{1+6\kappa^2}} \le \frac{1}{\sqrt{6}}$ 以及 $\frac{\sqrt{53}\kappa}{5\sqrt{4\kappa+1}} \le \sqrt{\kappa}$得到。现在，注意到和 $\sqrt[3]{5}+\sqrt{2} \le 2$，由（3.51a）可得（3.54）。此外，由（3.16）可得

$$\left\|\mathbf{x}_{\text{avg}}^{\tau} - \mathbf{prox}_{\eta_x g}\left(\mathbf{x}_{\text{avg}}^{\tau} - \eta_x \nabla_{\mathbf{x}} P(\mathbf{x}_{\text{avg}}^{\tau}, \mathbf{\Lambda}^{\tau})\right)\right\| = \left\|\mathbf{x}_{\text{avg}}^{\tau} - \mathbf{prox}_{\eta_x g}\left(\mathbf{x}_{\text{avg}}^{\tau} - \tfrac{\eta_x}{m}\sum_{i=1}^{m} \nabla_{\mathbf{x}} f_i(\mathbf{x}_{\text{avg}}^{\tau}, \hat{\mathbf{y}}_i^{\tau})\right)\right\|.$$

因此，根据三角不等式以及**近端算子**$\eta_x g$的非扩张性，我们有

$$\begin{aligned}
&\left\|\mathbf{x}_{\text{avg}}^{\tau} - \mathbf{prox}_{\eta_x g}\left(\mathbf{x}_{\text{avg}}^{\tau} - \eta_x \nabla_{\mathbf{x}} P(\mathbf{x}_{\text{avg}}^{\tau}, \mathbf{\Lambda}^{\tau})\right)\right\| \\
\le\ &\left\|\mathbf{x}_{\text{avg}}^{\tau} - \mathbf{prox}_{\eta_x g}\left(\tilde{\mathbf{x}}_i^{\tau} - \eta_x \mathbf{v}_i^{\tau}\right)\right\| + \left\|\mathbf{x}_{\text{avg}}^{\tau} - \tfrac{\eta_x}{m}\textstyle\sum_{i=1}^{m} \nabla_{\mathbf{x}} f_i(\mathbf{x}_{\text{avg}}^{\tau}, \hat{\mathbf{y}}_i^{\tau}) - \left(\tilde{\mathbf{x}}_i^{\tau} - \eta_x \mathbf{v}_i^{\tau}\right)\right\| \\
\le\ &\left\|\mathbf{x}_{\text{avg}}^{\tau} - \mathbf{prox}_{\eta_x g}\left(\tilde{\mathbf{x}}_i^{\tau} - \eta_x \mathbf{v}_i^{\tau}\right)\right\| + \eta_x \left\|\tfrac{1}{m}\textstyle\sum_{i=1}^{m} \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^{\tau}, \mathbf{y}_i^{\tau}) - \tfrac{1}{m}\sum_{i=1}^{m} \nabla_{\mathbf{x}} f_i(\mathbf{x}_{\text{avg}}^{\tau}, \hat{\mathbf{y}}_i^{\tau})\right\| \\
&+ \left\|\mathbf{x}_{\text{avg}}^{\tau} - \tilde{\mathbf{x}}_i^{\tau}\right\| + \eta_x \left\|\tfrac{1}{m}\textstyle\sum_{i=1}^{m} \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^{\tau}, \mathbf{y}_i^{\tau}) - \mathbf{v}_i^{\tau}\right\|.
\end{aligned}$$

将上述不等式两边平方，并应用杨氏不等式，得到

$$\begin{aligned}
&\left\|\mathbf{x}_{\text{avg}}^{\tau} - \mathbf{prox}_{\eta_x g}\left(\mathbf{x}_{\text{avg}}^{\tau} - \eta_x \nabla_{\mathbf{x}} P(\mathbf{x}_{\text{avg}}^{\tau}, \mathbf{\Lambda}^{\tau})\right)\right\|^2 \\
\le\ &4\left\|\mathbf{x}_{\text{avg}}^{\tau} - \mathbf{prox}_{\eta_x g}\left(\tilde{\mathbf{x}}_i^{\tau} - \eta_x \mathbf{v}_i^{\tau}\right)\right\|^2 + 4\eta_x^2 \left\|\tfrac{1}{m}\textstyle\sum_{i=1}^{m} \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^{\tau}, \mathbf{y}_i^{\tau}) - \tfrac{1}{m}\sum_{i=1}^{m} \nabla_{\mathbf{x}} f_i(\mathbf{x}_{\text{avg}}^{\tau}, \hat{\mathbf{y}}_i^{\tau})\right\|^2 \\
&+ 4\left\|\mathbf{x}_{\text{avg}}^{\tau} - \tilde{\mathbf{x}}_i^{\tau}\right\|^2 + 4\eta_x^2 \left\|\tfrac{1}{m}\textstyle\sum_{i=1}^{m} \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^{\tau}, \mathbf{y}_i^{\tau}) - \mathbf{v}_i^{\tau}\right\|^2.
\end{aligned} \tag{3.55}$$

将（3.48）连同 $t=\tau$代入（3.55），并对 $i=1,\ldots,m$ 求和，得

$$\begin{aligned}
&m\left\|\mathbf{x}_{\text{avg}}^{\tau} - \mathbf{prox}_{\eta_x g}\left(\mathbf{x}_{\text{avg}}^{\tau} - \eta_x \nabla_{\mathbf{x}} P(\mathbf{x}_{\text{avg}}^{\tau}, \mathbf{\Lambda}^{\tau})\right)\right\|^2 \\
\le\ &4\left\|\mathbf{X}_{\text{avg}}^{\tau} - \mathbf{X}^{\tau+1}\right\|_F^2 + 4\eta_x^2 L^2\left(\|\mathbf{X}_{\perp}^{\tau}\|_F^2 + \|\mathbf{Y}^{\tau} - \widehat{\mathbf{Y}}^{\tau}\|_F^2\right) + 4\left\|\mathbf{X}_{\text{avg}}^{\tau} - \widetilde{\mathbf{X}}^{\tau}\right\|_F^2 + 4\eta_x^2 \|\mathbf{V}_{\perp}^{\tau}\|_F^2,
\end{aligned} \tag{3.56}$$

其中用到了（3.7a）以及如下事实：$\mathbf{v}_{\text{avg}}^{t} = \frac{1}{m}\sum_{i=1}^{m} \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t)$ 对任意 t均成立。利用（3.21），并注意到 $\left\|\mathbf{X}_{\text{avg}}^{\tau} - \widetilde{\mathbf{X}}^{\tau}\right\|_F^2 \le \|\mathbf{X}_{\perp}^{\tau}\|_F^2$。由（3.56）可得

$$\begin{aligned}
&m\left\|\mathbf{x}_{\text{avg}}^{\tau} - \mathbf{prox}_{\eta_x g}\left(\mathbf{x}_{\text{avg}}^{\tau} - \eta_x \nabla_{\mathbf{x}} P(\mathbf{x}_{\text{avg}}^{\tau}, \mathbf{\Lambda}^{\tau})\right)\right\|^2 \\
\le\ &4\left\|\mathbf{X}_{\text{avg}}^{\tau} - \mathbf{X}^{\tau+1}\right\|_F^2 + 4\eta_x^2 L^2(1+2\kappa^2)\left\|\mathbf{X}_{\perp}^{\tau}\right\|_F^2 + 8m\eta_x^2\kappa^2\delta_{\tau}^2 + 4\left\|\mathbf{X}_{\perp}^{\tau}\right\|_F^2 + 4\eta_x^2 \|\mathbf{V}_{\perp}^{\tau}\|_F^2.
\end{aligned} \tag{3.57}$$

现在对 τ 对（3.57）两边取期望，并利用（3.49a）、（3.49d）和（3.49e），可得

$$\begin{aligned}
&m\mathbb{E}_{\tau}\left[\left\|\mathbf{x}_{\text{avg}}^{\tau} - \mathbf{prox}_{\eta_x g}\left(\mathbf{x}_{\text{avg}}^{\tau} - \eta_x \nabla_{\mathbf{x}} P(\mathbf{x}_{\text{avg}}^{\tau}, \mathbf{\Lambda}^{\tau})\right)\right\|^2\right] \\
\le\ &\frac{4}{T}\left[\tfrac{2mC_{0,T}(1-\rho)^2}{L(4\kappa+1)} + \left(\eta_x^2 L^2(1+2\kappa^2)+1\right)\left(\tfrac{3(1-\rho)}{50L^2(1+6\kappa^2)}\left\|\mathbf{V}_{\perp}^{0}\right\|_F^2 + \tfrac{53mC_{0,T}(1-\rho)^2}{50L(4\kappa+1)} + \tfrac{3m}{50L^2}\Delta_T\right)\right. \\
&\left. + \eta_x^2\left(\tfrac{3}{2(1-\rho)}\left\|\mathbf{V}_{\perp}^{0}\right\|_F^2 + \tfrac{m(1+9\kappa^2)}{(1-\rho)^2}\Delta_T + 40mC_{0,T}L\kappa\right) + 2m\eta_x^2\kappa^2\Delta_T\right].
\end{aligned}$$

因此，将上面不等式的两边除以 $m\eta_x^2$，可得

$$\begin{aligned}
&\tfrac{1}{\eta_x^2}\mathbb{E}_\tau\left[\left\|\mathbf{x}_{\rm avg}^\tau-\mathbf{prox}_{\eta_x g}\big(\mathbf{x}_{\rm avg}^\tau-\eta_x\nabla_{\mathbf{x}}P(\mathbf{x}_{\rm avg}^\tau,\mathbf{\Lambda}^\tau)\big)\right\|^2\right]\\
\le&\tfrac{4}{\eta_x^2 mT}\Bigg[\tfrac{2mC_{0,T}(1-\rho)^2}{L(4\kappa+1)}+\big(\eta_x^2L^2(1+2\kappa^2)+1\big)\Big(\tfrac{3(1-\rho)}{50L^2(1+6\kappa^2)}\big\|\mathbf{V}_\perp^0\big\|_F^2+\tfrac{53mC_{0,T}(1-\rho)^2}{50L(4\kappa+1)}+\tfrac{3m}{50L^2}\Delta_T\Big)\\
&\qquad+\eta_x^2\Big(\tfrac{3}{2(1-\rho)}\big\|\mathbf{V}_\perp^0\big\|_F^2+\tfrac{m(1+9\kappa^2)}{(1-\rho)^2}\Delta_T+40mC_{0,T}L\kappa\Big)+2m\eta_x^2\kappa^2\Delta_T\Bigg]\\
=&\tfrac{4}{T}\Big[C_{1,T}+40C_{0,T}L\kappa+\tfrac{53C_{0,T}L(1-\rho)^2(1+2\kappa^2)}{50(4\kappa+1)}+\tfrac{50LC_{0,T}(1+6\kappa^2)}{(4\kappa+1)(1-\rho)^2}+\tfrac{53LC_{0,T}(1+6\kappa^2)}{2(1-\rho)^2(4\kappa+1)}\Big]\\
\le&\tfrac{4}{T}\Big[C_{1,T}+\tfrac{156C_{0,T}L\kappa}{(1-\rho)^2}\Big],
\end{aligned}\tag{3.58}$$

其中，最后一个不等式由 $(1-\rho)^2\le 1$ 和 $\frac{53(1+2\kappa^2)}{50(4\kappa+1)}+\frac{50(1+6\kappa^2)}{4\kappa+1}+\frac{53(1+6\kappa^2)}{2(4\kappa+1)}\le 116\kappa$得到。现在利用詹森不等式，由（3.51b）可得（3.58），证明完毕。□

3.4. 复杂度结果。本小节将建立算法1生成问题（2.3）近似驻点的复杂度结果。

DEFINITION 3.12. 给定 $\varepsilon>0$，点（$\mathbf{X}$ $\mathbf{\Lambda}$）是问题 ε-平稳点（2.3），如果存在某个 $\eta>0$，

$$\tfrac{1}{\eta}\left\|\mathbf{x}_{\rm avg}-\mathbf{prox}_{\eta g}\big(\mathbf{x}_{\rm avg}-\eta\nabla_{\mathbf{x}}P(\mathbf{x}_{\rm avg},\mathbf{\Lambda})\big)\right\|\le\varepsilon,\quad \tfrac{L}{\sqrt{m}}\|\mathbf{X}_\perp\|_F\le\varepsilon,\quad \|\nabla_{\mathbf{\Lambda}}P(\mathbf{x}_{\rm avg},\mathbf{\Lambda})\|_F\le\varepsilon.\tag{3.59}$$

THEOREM 3.13. 设 $\varepsilon>0$ 已知。假设 $\delta_t\le\frac{(1-\rho^2)}{8\kappa}\frac{()}{1+t}\ \forall t\ge 0$ 且 $\|\mathbf{V}_\perp^0\|_F^2\le 2mL\kappa(1-\rho)$。令

$$T=\left\lceil\frac{1}{\varepsilon^2}\max\left\{\frac{64\big(10L\kappa(\phi_0+1)+1\big)}{(1-\rho)^2},\ \frac{4096L\kappa}{1-\rho},\ 800\right\}\right\rceil,\tag{3.60}$$

其中 $\phi_0=\phi(\mathbf{X}_{\rm avg}^0,\mathbf{\Lambda}^0)-(p+g)(x^*)$。令 $\{(X^t,\mathbf{\Lambda}^t)\}_{t=0}^T$ 由算法 *1* 生成，其中 η_x 和 η_λ 取为式 *(3.42)* 中的值。令 τ 从 $\{0,1,\dots,T-1\}$ 中均匀随机选取。则在假设 *1–3* 下，$(X^\tau,\mathbf{\Lambda}^\tau)$ 是问题 *(2.3)* 的一个 ε-平稳点（就期望而言）。

证明。根据引理 3.10 中 Δ_T 的定义，可得 $\Delta_T\le\sum_{t=0}^\infty\frac{1-\rho^4}{64\kappa^2}\frac{()}{1+t}\le\frac{(1-\rho^4)}{}$。因此，根据关于 $\|\mathbf{V}_\perp^0\|_F^2$ 的假设、引理 3.10 中 $C_{0,T}$和$C_{1,T}$ 的定义、式 (3.52) 以及 ϕ_0的定义，有 $C_{0,T}^{32\kappa^2}\le\phi_0+1+\frac{1}{6L\kappa}$ 和 $C_{1,T}\le\frac{8L\kappa}{1-\rho}+1$ 成立。由此，根据式 (3.51) 可得

$$\mathbb{E}_\tau\|\nabla_{\mathbf{\Lambda}}P(\mathbf{x}_{\rm avg}^\tau,\mathbf{\Lambda}^\tau)\|_F\le\tfrac{7\sqrt{L\kappa(\phi_0+1)+\frac{1}{6}}}{(1-\rho)\sqrt{T}}+\tfrac{(1-\rho)\sqrt{L\kappa}}{5\sqrt{T}}+\tfrac{(1-\rho)^2}{\sqrt{8T}},\tag{3.61a}$$

$$\tfrac{1}{\eta_x}\mathbb{E}_\tau\left\|\mathbf{x}_{\rm avg}^\tau-\mathbf{prox}_{\eta_x g}\big(\mathbf{x}_{\rm avg}^\tau-\eta_x\nabla_{\mathbf{x}}P(\mathbf{x}_{\rm avg}^\tau,\mathbf{\Lambda}^\tau)\big)\right\|\le\tfrac{2}{\sqrt{T}}\sqrt{\tfrac{8L\kappa}{1-\rho}+1+\tfrac{156L\kappa(\phi_0+1)+26}{(1-\rho)^2}}.\tag{3.61b}$$

此外，(3.49d)表明 $\sum_{t=0}^{T-1}\|\mathbf{X}_\perp^t\|_F^2\le\frac{3m\kappa}{25L}\frac{(1-\rho^2)()}{1+6\kappa^2}+\frac{53m(1-\rho)^2(\phi_0+1+\frac{1}{6L\kappa})50L(4\kappa+1)}{}+\frac{3m}{50L^2}\frac{(1-\rho^4)}{32\kappa^2}$，因此

$$\mathbb{E}_\tau\|\mathbf{X}_\perp^\tau\|_F\le\sqrt{\tfrac{m}{LT}}\left(\tfrac{\sqrt{3\kappa}(1-\rho)}{5\sqrt{1+6\kappa^2}}+\tfrac{\sqrt{53(\phi_0+1+\frac{1}{6L\kappa})}(1-\rho)}{5\sqrt{2+8\kappa}}+\tfrac{\sqrt{3}(1-\rho)^2}{40\kappa\sqrt{L}}\right).\tag{3.61c}$$

现在，根据 T 的选择，可以直接由（3.61a）–（3.61c）验证

$$\tfrac{1}{\eta_x}\mathbb{E}_\tau\left\|\mathbf{x}_{\rm avg}^\tau-\mathbf{prox}_{\eta_x g}\big(\mathbf{x}_{\rm avg}^\tau-\eta_x\nabla_{\mathbf{x}}P(\mathbf{x}_{\rm avg}^\tau,\mathbf{\Lambda}^\tau)\big)\right\|\le\varepsilon,\ \mathbb{E}_\tau\|\nabla_{\mathbf{\Lambda}}P(\mathbf{x}_{\rm avg}^\tau,\mathbf{\Lambda}^\tau)\|_F\le\varepsilon,\ \tfrac{L}{\sqrt{m}}\mathbb{E}_\tau\|\mathbf{X}_\perp^\tau\|_F\le\varepsilon,$$

也就是说，（$\mathbf{X}^\tau$，$\mathbf{\Lambda}^\tau$）是问题（ε）在期望意义下的平稳点。 □

REMARK 3.2. 我们对定理 3.13作几点说明。首先，复杂度对 ε的依赖是最优的。此外，对于求解复合非凸问题（例如 *SONATA* [39]）且每次迭代不执行多次通信的去中心化方法，关于 ρ 的依赖是目前已知最好的。其次，假设条件 $\|\mathbf{V}_\perp^0\|_F^2 \le 2mL\kappa(1-\rho)$ 在 κ 足够大时成立；否则，我们可以在初始步骤执行多次通信。第三，该定理给出了以期望意义产生 ε-平稳点所需的通信轮次和 $\mathbf{x}$-梯度计算次数，该平稳点对应于 (2.3) 。如果每个局部 $\mathbf{y}$-子问题都有闭式解，例如我们将在下一节测试的分布鲁棒逻辑回归问题，那么 $\mathbf{y}$- 梯度计算总次数也为 T。一般而言，我们需要使用迭代求解器来求得每个局部近似解。下面，我们估算一般情形下 $\mathbf{y}$-梯度的总数。

对于强凸问题具有加速收敛性的加速近端梯度法（APG），例如 [19,33], 中的方法，可用于求解每个 $t \ge 0$ 和每个 $i \in [m]$对应的 $\mathbf{y}_i^t$。为保证算法 1 第 7 行所要求的驻点条件，我们采用 [19],中 APG 的修改版本，该版本见 [43,算法 2]。为简单起见，我们使用非自适应版本（即假设已知光滑常数，不进行线搜索），近似求解 $\max_{\mathbf{y}} d_i^t(\mathbf{y})$，或等价地求解 $\min(-d_i^t)(\mathbf{y})$，其中 d_i^t 在 (2.4) 中定义。令 $\mathbf{y}_i^{t\star}$为 d_i^t的唯一极大点。注意，由于 d_i^t具有 μ-强凹性，$d_i^t(\mathbf{y}_i^{t\star}) - d_i^t(\mathbf{y}_i) \ge \frac{\mu}{2}\|\mathbf{y}_i^{t\star} - \mathbf{y}_i\|^2, \forall\, \mathbf{y}_i \in \mathrm{dom}(h)$。因此，若采用 [43]中的算法 2 的非自适应版本求解 $\min(-d_i^t)(\mathbf{y})$，并从 $\mathbf{y}_i^{t-1}$ 开始，则根据 [43, 定理 2.2], ，经过 s 次迭代后，可得到一点 $\mathbf{y}^{t-1,s}_i$，使得

$$\mathrm{dist}\big(\mathbf{0}, \partial d_i^t(\mathbf{y}_i^{t-1,s})\big) \le 4\sqrt{L_y}\sqrt{d_i^t(\mathbf{y}_i^{t\star}) - d_i^t(\mathbf{y}_i^{t-1})}\left(1 - \frac{1}{\sqrt{\kappa_y}}\right)^{\frac{s}{2}}, \tag{3.62}$$

其中，L_y 和 κ_y 分别由（3.5）和（3.6）给出。因此，若要得到满足 $\mathrm{dist}\big(\mathbf{0},\ \partial d_i^t(\mathbf{y}_i^t)\big) \le \delta_t$ 的 $\mathbf{y}_i^t$，只需运行 s_t 次 APG 迭代，前提是 $s_t \ge \ln \frac{16L_y\left(d_i^t(\mathbf{y}_i^{t\star}) - d_i^t(\mathbf{y}_i^{t-1})\right)}{\delta_t^2} \Big/ \ln \frac{1}{1-1/\sqrt{\kappa_y}}$。因此，由于 $\ln \frac{1}{1-x} \ge x$, $\forall x \in (0,1)$，我们可以设定

$$s_t = \left\lceil \sqrt{\kappa_y} \ln \frac{16L_y\left(d_i^t(\mathbf{y}_i^{t\star}) - d_i^t(\mathbf{y}_i^{t-1})\right)}{\delta_t^2} \right\rceil. \tag{3.63}$$

We s如何得到 $d_i^t(\mathbf{y}_i^{t\star}) - d_i^t(\mathbf{y}_i^{t-1})$ 的上界如下。其证明见附录 x.

LEMMA 3.14. 令 $\mathbf{Y}^{t\star} = \arg\max_{\mathbf{Y}} \Phi(\mathbf{X}^t, \mathbf{\Lambda}^t, \mathbf{Y})$，即 $\mathbf{y}_i^{t\star} = \arg\max_{\mathbf{y}} d_i^t(\mathbf{y})\ \forall\, t \ge 0,\ i \in [m]$。那么

$$\begin{aligned} d_i^t(\mathbf{y}_i^{t\star}) - d_i^t(\mathbf{y}_i^{t-1}) \le{}& \left(L(11+\sqrt{m}) + 22m\kappa^2(2L + L\sqrt{m}+2)\right)\frac{C_{0,t-1}(1-\rho)^2}{L(4\kappa+1)} \\ &+ \frac{(1-\rho)\left(\frac{L}{2}+\kappa^2(2L+L\sqrt{m}+2)\right)}{2L^2(1+6\kappa^2)}\left\|\mathbf{V}_\perp^0\right\|_F^2 + \frac{m\left(\frac{L}{2}+\kappa^2(2L+L\sqrt{m}+2)\right)}{2L^2}\Delta_{t-1} + \left(L + \frac{L\sqrt{m}}{2} + 1\right)\frac{\delta_{t-1}^2}{\mu^2} + \frac{\delta_{t-1}^2}{2}, \end{aligned} \tag{3.64}$$

对所有 $t \ge 1$，其中 $C_{0,\,t}$ 且 Δ_t 定义于引理 3.10中。

结合（3.63）和（3.64），我们即可对 $\mathbf{y}$-梯度的总数进行界定。

THEOREM 3.15 (总计 $\mathbf{y}$-梯度). 在定理 3.13 的相同假设下，令$\delta_t = \frac{(1-\rho^2)()}{8\kappa\ 1+t}\forall\, t \ge 0$。则为得到 $(\mathbf{X}^\tau, \mathbf{\Lambda}^\tau)$，梯度计算次数总计满足

$$T_y \le \left\lceil \sqrt{\kappa_y} \ln \frac{1024\kappa L_y\left(d_i^{0\star} - d_i^0(\mathbf{y}_i^{-1})\right)}{(1-\rho)^4} + T\sqrt{\kappa_y} \ln \frac{1024\kappa L_y D(T+1)^2}{(1-\rho)^4} + T + 1 \right\rceil, \forall\, i \in [m],$$

其中 $\mathbf{y}_i^{-1}$ 是用于求解 $d_i^{0\star} := \max_{\mathbf{y}} d_i^0(\mathbf{y})$ 的初始点，并且

$$
\begin{aligned}
D := & \left(L(11+\sqrt{m}) + 22m\kappa^2(2L + L\sqrt{m}+2)\right)\tfrac{(\phi_0+1+\frac{1}{6L\kappa})(1-\rho)^2}{L(4\kappa+1)} \\
& + \tfrac{m\kappa(1-\rho)^2\left(\frac{L}{2}+\kappa^2(2L+L\sqrt{m}+2)\right)}{L(1+6\kappa^2)} + \tfrac{m\left(\frac{L}{2}+\kappa^2(2L+L\sqrt{m}+2)\right)(1-\rho)^4}{64L^2\kappa^2} + \tfrac{(L+\frac{L\sqrt{m}}{2}+1)(1-\rho)^4}{64\mu^2\kappa^2} + \tfrac{(1-\rho)^4}{128\kappa^2}.
\end{aligned} \tag{3.65}
$$

证明。根据定理 3.13 的证明，可得 $\Delta_t \le (1-\rho^4)$，$\forall t \ge 0$ 且 $C_{0,t} \le \phi_0 + 1 + \frac{1}{6L\kappa}$。此外，显然有 $\delta_t^2 \le (1-\rho^4)$。因此，将这些上界代入式 (3.64)，可轻易得到 $d_i^t(\mathbf{y}_i^{t\star}) - d_i^t(\mathbf{y}_i^{t-1}) \le D, \forall t \ge 1, i \in [m]$。因此，由式 (3.63) 可知，对每个 i，y-梯度评估的总次数有如下上界：

$$
\begin{aligned}
T_y = \textstyle\sum_{t=0}^{T} s_t &\le \sqrt{\kappa_y}\ln\tfrac{16L_y\left(d_i^{0\star}-d_i^0(\mathbf{y}_i^{-1})\right)}{\delta_0^2} + \textstyle\sum_{t=1}^{T}\sqrt{\kappa_y}\ln\tfrac{16L_yD}{\delta_t^2} + T + 1 \\
&\le \sqrt{\kappa_y}\ln\tfrac{16L_y\left(d_i^{0\star}-d_i^0(\mathbf{y}_i^{-1})\right)}{\delta_0^2} + T\sqrt{\kappa_y}\ln\tfrac{16L_yD}{\delta_T^2} + T + 1 \\
&= \sqrt{\kappa_y}\ln\tfrac{1024\kappa L_y\left(d_i^{0\star}-d_i^0(\mathbf{y}_i^{-1})\right)}{(1-\rho)^4} + T\sqrt{\kappa_y}\ln\tfrac{1024\kappa L_yD(T+1)^2}{(1-\rho)^4} + T + 1,
\end{aligned}
$$

至此，证明完成。 □

REMARK 3.3. 根据定理 3.13 和 3.15，算法 1 可在期望意义下生成一个(2.3)的 ε-平稳点，所需通信轮次为 $O\left(\frac{L\kappa}{\varepsilon^2(1-\rho)^2}\right)$，x-梯度计算次数为x-梯度计算次数，且 y-梯度评估次数为 $O\left(\frac{L\kappa\sqrt{\kappa_y}}{\varepsilon^2}(1-\rho)^2\ln\frac{\kappa}{\varepsilon}\right)$ y-梯度评估次数。

3.5. （2.1）与（2.3）的驻点性关系。 (2.1) 和 (2.3)。本小节将证明，ϕ 的（近似）平稳点也是 $p+g$ 的（近似）平稳点，其中 $p(\cdot)$ 定义于（3.1），而 $\phi(\cdot)$ 定义于（3.2）。据此，我们能够建立（2.3）与原始表述（2.1）之间驻点性的关系。因此，上一小节的结果可转化为求解（2.1）近似驻点的复杂度结果。下面的引理给出了精确情形下的关系。

LEMMA 3.16. 若 $(\bar{\mathbf{x}}\ \bar{\mathbf{\Lambda}})$ 是 ϕ 所定义的(3.2)的平稳点，即对于某个 $\eta > 0$,

$$
\bar{\mathbf{x}} = \mathbf{prox}_{\eta g}\left(\bar{\mathbf{x}} - \eta\nabla_{\mathbf{x}}P(\bar{\mathbf{x}},\bar{\mathbf{\Lambda}})\right), \quad \nabla_{\mathbf{\Lambda}}P(\bar{\mathbf{x}},\bar{\mathbf{\Lambda}}) = \mathbf{0}, \tag{3.66}
$$

则 $\bar{\mathbf{x}}$ 也是 $p+g$ 的平稳点，即 $\bar{\mathbf{x}} = \mathrm{prox}_{\eta g}(\bar{\mathbf{x}} - \eta\nabla p(\bar{\mathbf{x}}))$，其中 $p(\cdot)$ 定义于（3.1）。证明。令 $\bar{\mathbf{Y}} = S_\Phi(\mathbf{1}\bar{\mathbf{x}}^\top, \bar{\mathbf{\Lambda}})$。于是 $\nabla P(\bar{\mathbf{x}}\ \bar{\mathbf{\Lambda}}) = \left(\frac{1}{m}\sum_{i=1}^m \nabla_{\mathbf{x}} f_i(\bar{\mathbf{x}}\ \bar{\mathbf{y}}_i) - \frac{L}{2\sqrt{m}}(\mathbf{W}-\mathbf{I})\ \bar{\mathbf{Y}}\right)$。因此，由 $\nabla_{\mathbf{\Lambda}}P(\bar{\mathbf{x}}\ \bar{\mathbf{\Lambda}}) = \mathbf{0}$ 可得 $\bar{\mathbf{Y}} = \frac{1}{m}\mathbf{1}\mathbf{1}^\top\bar{\mathbf{Y}} = \mathbf{1}\bar{\mathbf{y}}_{\mathrm{avg}}^\top$；根据 $\bar{\mathbf{Y}}$ 的定义，有 $\Phi(\mathbf{1}\bar{\mathbf{x}}^\top, \bar{\mathbf{\Lambda}}, \mathbf{1}\bar{\mathbf{y}}_{\mathrm{avg}}^\top) \ge \Phi(\mathbf{1}\bar{\mathbf{x}}^\top, \bar{\mathbf{\Lambda}}, \mathbf{1}\mathbf{y}^\top)\ \forall \mathbf{y} \in \mathrm{dom}(h)$。注意到 $(\mathbf{W}-\mathbf{I})\mathbf{1}\mathbf{y}^\top = \mathbf{0}$。我们有 $\bar{\mathbf{y}}_{\mathrm{avg}} = \arg\max_{\mathbf{y}} \frac{1}{m}\sum_{i=1}^m f_i(\bar{\mathbf{x}}, \mathbf{y}) - h(\mathbf{y})$。因此 $\nabla p(\bar{\mathbf{x}}) = \frac{1}{m}\sum_{i=1}^m \nabla_{\mathbf{x}} f_i(\bar{\mathbf{x}}\ \bar{\mathbf{y}}_{\mathrm{avg}}) = \nabla_{\mathbf{x}}P(\bar{\mathbf{x}}\ \bar{\mathbf{\Lambda}})$。于是，（3.66）中的第一个条件化简为 $\bar{\mathbf{x}} = \mathrm{prox}_{\eta g}(\bar{\mathbf{x}} - \eta\nabla p(\bar{\mathbf{x}}))$，证明完毕。□

[{"id": 0, "output": "以下引理将用于展示近似平稳性情形下的关系。"}]

LEMMA 3.17. 在假设 3 下，$(\mathbf{W}-\mathbf{I})^\top(\mathbf{W}-\mathbf{I})$ 在 $\mathrm{Span}\{\mathbf{1}\}^\perp$ 上的特征值均为正，我们将其中最小的特征值记为 $\lambda^+\min > 0$。证明。设 $\mathbf{x} \in \mathrm{Span}\{\mathbf{1}\}^\perp$ 且 $\mathbf{x} \ne \mathbf{0}$。假设 $\mathbf{x}^\top(\mathbf{W}-\mathbf{I})^\top(\mathbf{W}-\mathbf{I})\mathbf{x} = 0$，即 $\|(\mathbf{W}-\mathbf{I})\mathbf{x}\|^2 = 0$，因此 $(\mathbf{W}-\mathbf{I})\mathbf{x} = \mathbf{0}$。于是，$\mathbf{x} \in \mathrm{Null}(\mathbf{W}-\mathbf{I})$。根据假设 3(iii)，可得 $\mathbf{x} \in \mathrm{Span}\{\mathbf{1}\}$。又因 $\mathbf{x} \in \mathrm{Span}\{\mathbf{1}\}^\perp$，故只能有 $\mathbf{x} = \mathbf{0}$，这与 $\mathbf{x} \ne \mathbf{0}$矛盾。因此，$\mathbf{x}^\top(\mathbf{W}-\mathbf{I})^\top(\mathbf{W}-\mathbf{I})\mathbf{x} > 0$，即 $(\mathbf{W}-\mathbf{I})^\top(\mathbf{W}-\mathbf{I})$ 在 $\mathrm{Span}\{\mathbf{1}\}^\perp$ 上正定。证明完毕。□

下一个引理表明，在 h满足不同条件时， ϕ 的近似驻点也是 $p+g$ 的近似驻定解。

LEMMA 3.18. 假设（$\bar{\mathbf{x}}$ $\bar{\mathbf{\Lambda}}$）是 ε 的平稳点，其中 ϕ，即对某个 $\eta>0$，

$$\frac{1}{\eta}\left\|\bar{\mathbf{x}}-\mathbf{prox}_{\eta g}\left(\bar{\mathbf{x}}-\eta\nabla_{\mathbf{x}}P(\bar{\mathbf{x}},\bar{\mathbf{\Lambda}})\right)\right\|\leq\varepsilon,\quad\|\nabla_{\mathbf{\Lambda}}P(\bar{\mathbf{x}},\bar{\mathbf{\Lambda}})\|_F\leq\varepsilon. \tag{3.67}$$

我们有如下结论：

(i) 如果 h 是 L_h-光滑的，那么 $\bar{x}$ 是 $p+g$ 的 ε_1-平稳点，其中 $\varepsilon_1=\varepsilon+\varepsilon\sqrt{\frac{8}{(\mu+L+L_h)\,\mu\lambda^+}}$ $\mu\lambda^+$ *min* 和 λ^+ *min* 由引理 *3.17* 给出，即 $\frac{1}{\eta}\left\|\bar{x}-\right.$ –近端算子$_{\eta g}\left(\bar{x}-\eta\nabla p(\bar{x})\right)\|\leq\varepsilon_1$。 *(ii)* 如果对于任意 $\boldsymbol{\xi}\in\partial h(y)$ 以及任意 $y\in dom(h)$，$\|\boldsymbol{\xi}\|\leq M_h$，那么 $\bar{x}$ 是 $p+g$ 的 ε_2-平稳点，其中 $\varepsilon_2=\varepsilon+\varepsilon\sqrt{8}$ $(\mu+L)\,\mu\lambda^+$ *min* $+\sqrt{\frac{16L\varepsilon}{\mu\lambda^+\ min}+\frac{4LM_h^2\varepsilon}{\mu}}$。

证明。对于给定的 $(\bar{\mathbf{x}}\ \bar{\mathbf{\Lambda}})$，我们定义

$$\psi(\mathbf{Y}):=\Phi(\mathbf{1}\bar{\mathbf{x}}^\top,\bar{\mathbf{\Lambda}},\mathbf{Y}),\quad\bar{\mathbf{Y}}=\arg\max_{\mathbf{Y}}\psi(\mathbf{Y}),\quad\hat{\mathbf{y}}=\arg\max_{\mathbf{y}}f(\bar{\mathbf{x}},\mathbf{y})-h(\mathbf{y}).$$

有 $\nabla_{\mathbf{\Lambda}}P(\bar{\mathbf{x}}\ \bar{\mathbf{\Lambda}})=-\frac{L}{2\sqrt{m}}(\mathbf{W}-\mathbf{I})\ \bar{\mathbf{Y}}$，因此

$$\varepsilon^2\geq\|\nabla_{\mathbf{\Lambda}}P(\bar{\mathbf{x}},\bar{\mathbf{\Lambda}})\|_F^2=\frac{L^2}{4m}\|(\mathbf{W}-\mathbf{I})\bar{\mathbf{Y}}\|_F^2=\frac{L^2}{4m}\|(\mathbf{W}-\mathbf{I})\bar{\mathbf{Y}}_\perp\|_F^2\geq\frac{\lambda_{\min}^+L^2}{4m}\|\bar{\mathbf{Y}}_\perp\|_F^2, \tag{3.68}$$

其中，第二个等号成立是因为 $\mathbf{W}\ -\mathbf{I}\ =(\mathbf{W}\ -\mathbf{I})(\mathbf{I}-\frac{1}{m}\mathbf{1}\mathbf{1}^\top)$，而第二个不等式由引理 3.17 得出。此外，我们有

$$\begin{aligned}&\|\nabla p(\bar{\mathbf{x}})-\nabla_{\mathbf{x}}P(\bar{\mathbf{x}},\bar{\mathbf{\Lambda}})\|^2=\left\|\tfrac{1}{m}\textstyle\sum_{i=1}^m\nabla_{\mathbf{x}}f_i(\bar{\mathbf{x}},\hat{\mathbf{y}})-\tfrac{1}{m}\sum_{i=1}^m\nabla_{\mathbf{x}}f_i(\bar{\mathbf{x}},\bar{\mathbf{y}}_i)\right\|^2\leq\tfrac{L^2}{m}\textstyle\sum_{i=1}^m\|\hat{\mathbf{y}}-\bar{\mathbf{y}}_i\|^2\\&\leq\tfrac{2L^2}{m}\textstyle\sum_{i=1}^m\left(\|\bar{\mathbf{y}}_{\rm avg}-\bar{\mathbf{y}}_i\|^2+\|\hat{\mathbf{y}}-\bar{\mathbf{y}}_{\rm avg}\|^2\right)=\tfrac{2L^2}{m}\|\bar{\mathbf{Y}}_\perp\|_F^2+2L^2\|\hat{\mathbf{y}}-\bar{\mathbf{y}}_{\rm avg}\|^2,\end{aligned}$$

因此　由 **近端算子**ηg 的非扩张性、条件（3.67）和（3.68），可得

$$\frac{1}{\eta}\left\|\bar{\mathbf{x}}-\mathbf{prox}_{\eta g}\left(\bar{\mathbf{x}}-\eta\nabla p(\bar{\mathbf{x}})\right)\right\|\leq\varepsilon+\left\|\nabla p(\bar{\mathbf{x}})-\nabla_{\mathbf{x}}P(\bar{\mathbf{x}},\bar{\mathbf{\Lambda}})\right\|\leq\varepsilon+\sqrt{\tfrac{8\varepsilon^2}{\lambda_{\min}^+}+2L^2\|\hat{\mathbf{y}}-\bar{\mathbf{y}}_{\rm avg}\|^2}. \tag{3.69}$$

因此，要证明 $\bar{\mathbf{x}}$ 对于 $p+g$ 的近似驻定性，只需对 $\|\hat{\mathbf{y}}\ -\bar{\mathbf{y}}_{\rm avg}\|^2$ 进行界定。由每个 f_i 的 L- 光滑性，可得

$$\frac{1}{m}\textstyle\sum_{i=1}^m f_i(\bar{\mathbf{x}},\bar{\mathbf{y}}_{\rm avg})\geq\frac{1}{m}\sum_{i=1}^m\left(f_i(\bar{\mathbf{x}},\bar{\mathbf{y}}_i)+\left\langle\nabla_{\mathbf{y}}f_i(\bar{\mathbf{x}},\bar{\mathbf{y}}_i),\bar{\mathbf{y}}_{\rm avg}-\bar{\mathbf{y}}_i\right\rangle-\frac{L}{2}\|\bar{\mathbf{y}}_{\rm avg}-\bar{\mathbf{y}}_i\|^2\|\right). \tag{3.70}$$

此外，根据 $\mathbf{Y}$ 处的最优性条件，对于每个 i，存在一个次梯度 $\boldsymbol{\xi}_i\in\partial h(\bar{\mathbf{y}}_i)$，使得

$$\mathbf{0}=\tfrac{1}{m}\left[\nabla_{\mathbf{y}}f_1(\bar{\mathbf{x}},\bar{\mathbf{y}}_1)-\bar{\boldsymbol{\xi}}_1,\ldots,\nabla_{\mathbf{y}}f_m(\bar{\mathbf{x}},\bar{\mathbf{y}}_m)-\bar{\boldsymbol{\xi}}_m\right]^\top-\tfrac{L}{2\sqrt{m}}(\mathbf{W}-\mathbf{I})^\top\bar{\mathbf{\Lambda}}.$$

上述方程与 (3.70) 给出

$$\begin{aligned}&\tfrac{1}{m}\textstyle\sum_{i=1}^m f_i(\bar{\mathbf{x}},\bar{\mathbf{y}}_{\rm avg})\\&\geq\tfrac{1}{m}\textstyle\sum_{i=1}^m\left(f_i(\bar{\mathbf{x}},\bar{\mathbf{y}}_i)+\left\langle\bar{\boldsymbol{\xi}}_i,\bar{\mathbf{y}}_{\rm avg}-\bar{\mathbf{y}}_i\right\rangle-\tfrac{L}{2}\|\bar{\mathbf{y}}_{\rm avg}-\bar{\mathbf{y}}_i\|^2\|\right)+\tfrac{L}{2\sqrt{m}}\left\langle(\mathbf{W}-\mathbf{I})^\top\bar{\mathbf{\Lambda}},\mathbf{1}\bar{\mathbf{y}}_{\rm avg}^\top-\bar{\mathbf{Y}}\right\rangle\\&=\tfrac{1}{m}\textstyle\sum_{i=1}^m\left(f_i(\bar{\mathbf{x}},\bar{\mathbf{y}}_i)+\left\langle\bar{\boldsymbol{\xi}}_i,\bar{\mathbf{y}}_{\rm avg}-\bar{\mathbf{y}}_i\right\rangle-\tfrac{L}{2}\|\bar{\mathbf{y}}_{\rm avg}-\bar{\mathbf{y}}_i\|^2\|\right)-\tfrac{L}{2\sqrt{m}}\left\langle(\mathbf{W}-\mathbf{I})^\top\bar{\mathbf{\Lambda}},\bar{\mathbf{Y}}\right\rangle\\&=\psi(\bar{\mathbf{Y}})-\tfrac{L}{2m}\|\bar{\mathbf{Y}}_\perp\|_F^2+\tfrac{1}{m}\textstyle\sum_{i=1}^m\left(h(\bar{\mathbf{y}}_i)+\left\langle\bar{\boldsymbol{\xi}}_i,\bar{\mathbf{y}}_{\rm avg}-\bar{\mathbf{y}}_i\right\rangle\right),\end{aligned} \tag{3.71}$$

其中第一个等式由$\mathbf{W1}=\mathbf{1}$得出，第二个等式使用了ψ的定义。将(3.71)两边减去$h(\bar{\mathbf{y}}_{\text{avg}})$并利用$f$的定义可得

$$f(\bar{\mathbf{x}},\bar{\mathbf{y}}_{\text{avg}})-h(\bar{\mathbf{y}}_{\text{avg}})\geq\psi(\bar{\mathbf{Y}})-\tfrac{L}{2m}\|\bar{\mathbf{Y}}_{\perp}\|_F^2+\tfrac{1}{m}\textstyle\sum_{i=1}^m\left(h(\bar{\mathbf{y}}_i)-h(\bar{\mathbf{y}}_{\text{avg}})+\left\langle\bar{\boldsymbol{\xi}}_i,\bar{\mathbf{y}}_{\text{avg}}-\bar{\mathbf{y}}_i\right\rangle\right). \tag{3.72}$$

此外，根据$\bar{\mathbf{Y}}$的定义，有$\psi(\bar{\mathbf{Y}})\geq\psi(\mathbf{1}\hat{\mathbf{y}}^\top)=f(\bar{\mathbf{x}},\hat{\mathbf{y}})-h(\hat{\mathbf{y}})$，并且由$f(\bar{\mathbf{x}}\,\cdot)-h(\cdot)$的$\mu$-强凹性，可得$f(\bar{\mathbf{x}}\,\bar{\mathbf{y}}_{\text{avg}})-h(\bar{\mathbf{y}}_{\text{avg}})\leq f(\bar{\mathbf{x}},\hat{\mathbf{y}})-h(\hat{\mathbf{y}})-\frac{\mu}{2}\|\hat{\mathbf{y}}-\bar{\mathbf{y}}_{\text{avg}}\|^2$。将这两个不等式与(3.72)结合，得到

$$\tfrac{\mu}{2}\|\hat{\mathbf{y}}-\bar{\mathbf{y}}_{\text{avg}}\|^2\leq\tfrac{L}{2m}\|\bar{\mathbf{Y}}_{\perp}\|_F^2+\tfrac{1}{m}\textstyle\sum_{i=1}^m\left(h(\bar{\mathbf{y}}_{\text{avg}})-h(\bar{\mathbf{y}}_i)-\left\langle\bar{\boldsymbol{\xi}}_i,\bar{\mathbf{y}}_{\text{avg}}-\bar{\mathbf{y}}_i\right\rangle\right). \tag{3.73}$$

现在，如果h是L_h-光滑的，那么次梯度$\bar{\boldsymbol{\xi}}_i$就化为梯度$\nabla h(\bar{\mathbf{y}}_i)$，并且对于每个$i$，$\left\langle\nabla h(\bar{\mathbf{y}}_i),\bar{\mathbf{y}}_{\text{avg}}-\bar{\mathbf{y}}_i\right\rangle\leq\frac{L_h}{2}\|\bar{\mathbf{y}}_{\text{avg}}-\bar{\mathbf{y}}_i\|^2\leq\frac{L+L_h}{m}\|\bar{\mathbf{Y}}_{\perp}\|_F^2$)，并且对于每个$i$，$h(\bar{\mathbf{y}}_i)-h(\bar{\mathbf{y}}_i)-\left\langle\bar{\boldsymbol{\xi}}_i,\bar{\mathbf{y}}_{\text{avg}}-\bar{\mathbf{y}}_i\right\rangle\leq\frac{\mu}{2}\|\bar{\mathbf{y}}_{\text{avg}}-\bar{\mathbf{y}}_i\|^2\leq\frac{L+L_h}{m}\|\bar{\mathbf{Y}}_{\perp}\|_F^2$)。因此，$(\cdot)$意味着$\hat{\mathbf{y}}\,\mathbf{y}\,\bar{\mathbf{Y}}$，结合(3.68)可得$\|\hat{\mathbf{y}}-\bar{\mathbf{y}}_{\text{avg}}\|^2\leq\frac{4(L+L_h)\varepsilon^2}{\mu L^2\lambda_+}$)$\min$。于是，由(3.69)得到

$$\tfrac{1}{\eta}\left\|\bar{\mathbf{x}}-\mathbf{prox}_{\eta g}\left(\bar{\mathbf{x}}-\eta\nabla p(\bar{\mathbf{x}})\right)\right\|\leq\varepsilon+\sqrt{\tfrac{8\varepsilon^2}{\lambda_{\min}^+}+\tfrac{8(L+L_h)\varepsilon^2}{\mu\lambda_{\min}^+}}=\varepsilon+\varepsilon\sqrt{\tfrac{8(\mu+L+L_h)}{\mu\lambda_{\min}^+}},$$

这证明了引理的第一个结论。

当$\|\bar{\boldsymbol{\xi}}_i\|\leq M_h$时，对每个$i$，有$h(\bar{\mathbf{y}}_{\text{avg}})-h(\bar{\mathbf{y}}_i)-\left\langle\bar{\boldsymbol{\xi}}_i,\bar{\mathbf{y}}_{\text{avg}}-\bar{\mathbf{y}}_i\right\rangle\leq 2M_h\|\bar{\mathbf{y}}_{\text{avg}}-\bar{\mathbf{y}}_i\|\leq\frac{M_h^2\varepsilon}{L}+\frac{L}{\varepsilon}\|\bar{\mathbf{y}}_{\text{avg}}-\bar{\mathbf{y}}_i\|^2$。因此，(3.73)意味着$\frac{\mu}{2}\|\hat{\mathbf{y}}-\bar{\mathbf{y}}_{\text{avg}}\|^2\leq\frac{L}{2m}\|\bar{\mathbf{Y}}_{\perp}\|_F^2+\frac{L}{m\varepsilon}\|\bar{\mathbf{Y}}_{\perp}\|_F^2+\frac{M_h^2\varepsilon}{L}$，结合(3.68)可得$\|\hat{\mathbf{y}}-\bar{\mathbf{y}}_{\text{avg}}\|^2\leq(1+\frac{2}{\varepsilon})\frac{4\varepsilon^2}{\mu\lambda^+}\min_L+\frac{2M_h^2\varepsilon}{\mu L}$。因此，由(3.69)可得

$$\tfrac{1}{\eta}\left\|\bar{\mathbf{x}}-\mathbf{prox}_{\eta g}\left(\bar{\mathbf{x}}-\eta\nabla p(\bar{\mathbf{x}})\right)\right\|\leq\varepsilon+\sqrt{\tfrac{8\varepsilon^2}{\lambda_{\min}^+}+(1+\tfrac{2}{\varepsilon})\tfrac{8L\varepsilon^2}{\mu\lambda_{\min}^+}+\tfrac{4LM_h^2\varepsilon}{\mu}}\leq\varepsilon+\varepsilon\sqrt{\tfrac{8(\mu+L)}{\mu\lambda_{\min}^+}}+\sqrt{\tfrac{16L\varepsilon}{\mu\lambda_{\min}^+}+\tfrac{4LM_h^2\varepsilon}{\mu}},$$

从而完成证明。 □

REMARK 3.4. 根据引理 *3.18*，为了生成原始表述 *(2.1)* 的一个 ε-解 X，即 $\frac{L}{\sqrt{m}}\|X_{\perp}\|_F\leq\varepsilon$ 和 $\frac{1}{\eta}\left\|x^{\text{avg}}-\text{近端算子}_{\eta g}\left(x_{\text{avg}}-\eta\nabla p(x^{\text{avg}})\right)\right\|\leq\varepsilon$ 对于某个 $\eta>0$，当 h 光滑时，复杂度在注 *3.3* 中所述复杂度的基础上还会增加一个 κ 因子；当 h非光滑但具有有界次梯度时，复杂度还会增加一个 $\frac{\kappa^2}{\varepsilon^2}$ 因子。与集中式—*GDMax* 方法相比，后一种情况的复杂度更高。我们推测，较差的结果是由我们的分析可能不够紧致所致。

4. 数值实验。本节中，我们测试所提出的算法，以求解以下分布鲁棒逻辑回归（DRLR）：

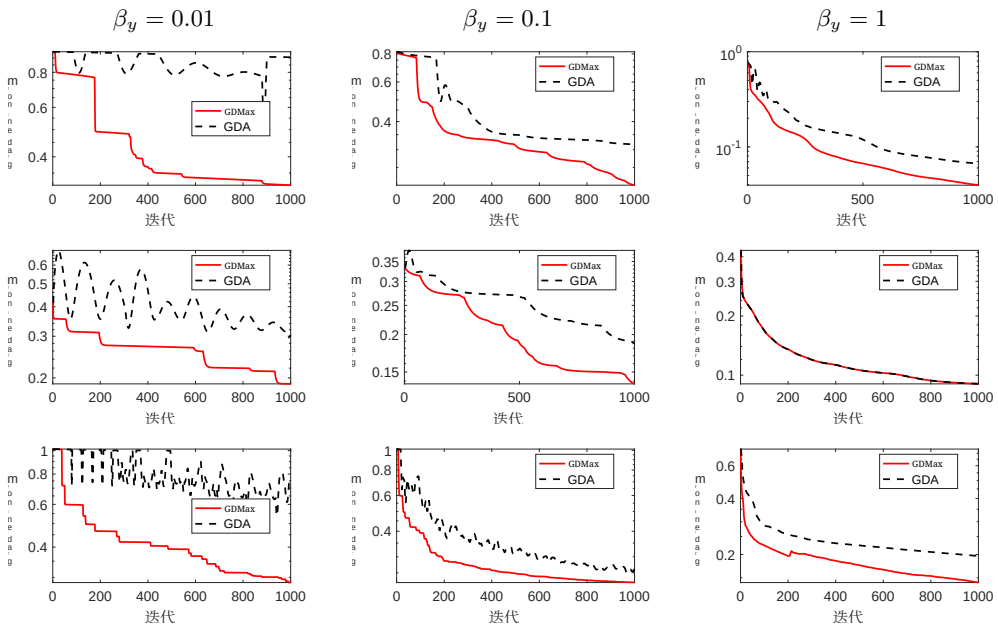
$$\min_{\mathbf{x}\in\mathbb{R}^n}\max_{\mathbf{y}\in\mathcal{Y}}\textstyle\sum_{j=1}^N y_j\ell(\mathbf{x};\mathbf{a}_j,b_j)+V_x(\mathbf{x})-V_y(\mathbf{y}), \tag{4.1}$$

其中，$\{(\mathbf{a}_j,b_j)\}_{j=1}^N$是数据集，其中每个$\mathbf{a}_j\in\mathbb{R}^n$都是一个特征向量，$b_j\in\{-1,+1\}$是相应的标签；$\mathcal{Y}=\{\mathbf{y}\in\mathbb{R}_+^N:\mathbf{1}^\top\mathbf{y}=1\}$是$N$维单纯形；$\ell(\mathbf{x};\mathbf{a}\;b)=\log\left(1+\exp(-b\cdot\mathbf{a}^\top\mathbf{x})\right)$是逻辑损失；$V_x(\mathbf{x})=\beta_x\sum_{i=1}^n\frac{\alpha x_i^2}{1+\alpha x_i^2}$，而$V_y(\mathbf{y})=\frac{\beta_y}{2}\|\mathbf{y}-\frac{1}{N}\|^2$是一个强凸项，用于控制概率向量$\mathbf{y}$与均匀分布向量$\frac{1}{N}$之间的距离。当强制令$\mathbf{y}=\frac{1}{N}$时，(4.1)化简为标准（非鲁棒）LR。假设数据集被划分为m个子集$\{\mathcal{D}_i\}_{i=1}^m$。令$\mathcal{J}_i\subset[N]$为每个$i\in[m]$对应于子数据集$\mathcal{D}_i$的索引集。定义

$$f_i(\mathbf{x},\mathbf{y})=m\textstyle\sum_{j\in\mathcal{J}_i}y_j\ell(\mathbf{x};\mathbf{a}_j,b_j)+V_x(\mathbf{x})-V_y(\mathbf{y}),\forall\,i\in[m],\quad g(\mathbf{x})=0,\quad h(\mathbf{y})=\iota_{\mathcal{Y}}.$$

因此，(4.1)可以表述为(1.1)的形式，并满足假设1和2中的条件。

4.1. 中心化设置下的比较。若在算法 1 中采用 $m = 1$，即可轻松得到我们方法的中心化（或非分布式）版本，并将其命名为 GDMax。为了展示对偶最大化的优势，我们仅将本方法与 [20] 中的 GDA 方法进行比较。我们使用了三个 LIBSVM 数据集：a9a、gisette 和 rcv1。对于每个数据集，我们遵循 [45]，在 (4.1) 中固定 $\alpha = 10$、$\beta_x = 10^{-3}$，并将 β_y 从 $\{0.01, 0.1, 1\}$ 开始变化，其中较小的 β_y 表示更难的问题实例。由于 y 子问题可以通过将某个点投影到单纯形上精确求解，因此无需迭代子程序；对于我们的方法，只需调整 x-步长。GDA 则需要调整 x 和 y 的步长。对于每个数据集，我们对我们的方法在 $\{0.5, 0.1, 0.05, 0.01, 0.005, 0.001\}$ 范围内网格搜索 η_x，并对 GDA 在 $\{0.5, 0.1, 0.05, 0.01, 0.005, 0.001\}^2$ 范围内网格搜索 (η_x, η_y)。报告其中的最佳结果。图1展示了结果，表中列出了最佳步长。我们发现，较小的 β_y（对应于较大的 κ）会减慢 GDMax 和 GDA 的收敛速度，但 GDMax 的表现显著优于 GDA，尤其是在 β_y 较小时。这表明了执行对偶最大化的优势。



	$\beta_y = 0.01$		$\beta_y = 0.1$		$\beta_y = 1$	
	GDMax	GDA	GDMax	GDA	GDMax	GDA
a9a	$\eta_x = 0.005$	$\eta_x = 0.001, \eta_y = 0.5$	$\eta_x = 0.01$	$\eta_x = 0.005, \eta_y = 0.5$	$\eta_x = 0.1$	$\eta_x = 0.05, \eta_y = 0.05$
gisette	$\eta_x = 0.005$	$\eta_x = 0.001, \eta_y = 0.5$	$\eta_x = 0.01$	$\eta_x = 0.005, \eta_y = 0.5$	$\eta_x = 0.05$	$\eta_x = 0.01, \eta_y = 0.5$
rcv1	$\eta_x = 0.01$	$\eta_x = 0.005, \eta_y = 0.5$	$\eta_x = 0.01$	$\eta_x = 0.01, \eta_y = 0.1$	$\eta_x = 0.5$	$\eta_x = 0.1, \eta_y = 0.5$

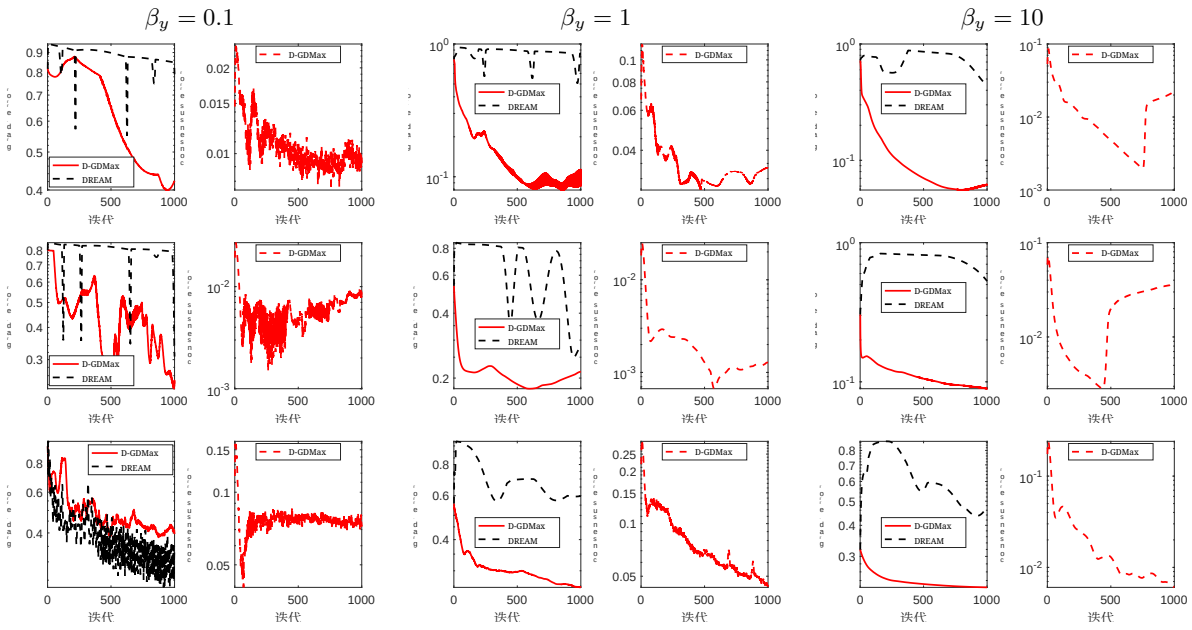
Primal gradient norm at each iterate by the centralized version of our method (called GDMax) and GDA in [20]

FIG. 1。 *for solving the DRLR problem in* 4 1 *on a9a gisette*

（。） ,, *and rcv1 (from top to bottom) with* β_y *varying from* $\{0.01, 0.1, 1\}$ *and,* $\alpha = 10$ $\beta_x = 10^{-3}$ *fixed. The corresponding best stepsizes are listed in the table.*

4.2. 与一种去中心化方法的比较。我们进一步将所提出的 D-GDMax 方法与 DREAM 方法进行比较，[4], 后者似乎是目前唯一一种在带有y约束的 NCSC 极小极大问题中，通过每次更新执行多轮通信来保证收敛的现有去中心化方法。同样，我们使用 a9a、gisette 和 rcv1，并将每个数据集均匀随机划分为 $m = 20$ 个子集。我们生成一个含 m 个节点的 Erd˝os-R´enyi 随机连通图。混合矩阵 $\mathbf{W}$ 设为 $\mathbf{I} - 0.8\mathbf{L}/\lambda_{\max}(\mathbf{L})$，其中 $\mathbf{L}$ 为图拉普拉斯矩阵。生成的 $\mathbf{W}$ 具有

$\rho = \|\mathbf{W} - \frac{1}{m}\mathbf{1}\mathbf{1}^\top\| \approx 0.9271$。D-GDMax 和 DREAM 使用相同的 $\mathbf{W}$。同样，我们固定 $\alpha = 10$，$\beta_x = 10^{-3}$在（4.1）中的值，但改变 β_y从 $\{0.1, 1, 10\}$。对于 DREAM，我们简单地在每次迭代中执行 10 轮通信。两种方法都有两个步长，我们通过网格搜索从 $\{0.5, 0.1, 0.05, 0.01, 0.005, 0.001\}^2$中选取。最佳结果绘于图2 中，相应的步长列于表中。由于多轮通信，DREAM 的共识误差在每个迭代点都处于 10^{-4} 的量级，因此未绘出。我们看到，除 rcv1 在`rcv1` 且 $\beta_y = 0.1$的情况外，DREAM 基于原始表述（2.1）衡量的梯度误差显著大于 D-GDMax，尽管前者在每次迭代中执行多轮通信。D-GDMax 的优势仍应归因于我们的重新表述所实现的对偶最大化。



	$\beta_y = 0.01$		$\beta_y = 0.1$		$\beta_y = 1$	
	D-GDMax	DREAM [4]	D-GDMax	DREAM [4]	D-GDMax	DREAM [4]
a9a	$\eta_x = 10^{-3}, \eta_\lambda = 0.05$	$\eta_x = 0.005, \eta_y = 10^{-3}$	$\eta_x = 0.005, \eta_\lambda = 10^{-3}$	$\eta_x = 0.005, \eta_y = 10^{-3}$	$\eta_x = 0.005, \eta_\lambda = 10^{-3}$	$\eta_x = 0.01, \eta_y = 10^{-3}$
gisette	$\eta_x = 10^{-3}, \eta_\lambda = 10^{-3}$	$\eta_x = 0.01, \eta_y = 10^{-3}$	$\eta_x = 10^{-3}, \eta_\lambda = 10^{-3}$	$\eta_x = 0.01, \eta_y = 0.005$	$\eta_x = 0.005, \eta_\lambda = 10^{-3}$	$\eta_x = 0.01, \eta_y = 0.5$
rcv1	$\eta_x = 0.005, \eta_\lambda = 10^{-3}$	$\eta_x = 0.1, \eta_y = 0.5$	$\eta_x = 0.01, \eta_\lambda = 10^{-3}$	$\eta_x = 0.05, \eta_y = 0.5$	$\eta_x = 0.01, \eta_\lambda = 10^{-3}$	$\eta_x = 0.1, \eta_y = 0.005$

FIG. 2. *Primal gradient norm and consensus error at each iterate by our method (called D-GDMax) and DREAM in [4] for solving a decentralized formulation of the DRLR problem in (4.1) on a9a, gisette, and rcv1 (from top to bottom) with β_y varying from $\{0.1, 1, 10\}$ and $\alpha = 10, \beta_x = 10^{-3}$ fixed. The best stepsizes are listed in the table. A random graph with 20 nodes is used. Consensus error for DREAM stands at the order of 10^{-4} by 10 rounds of communication for each iteration.*

5. 结论。我们提出了一种求解非凸强凹复合极小极大问题的去中心化算法，通过构建一种重新表述来解耦对偶变量上的非光滑性与共识。据我们所知，这是处理去中心化非光滑非凸强凹问题的首次尝试。我们证明了全局收敛（至驻点性）以及迭代复杂度结果。此外，在求解分布鲁棒逻辑回归问题时，所提出的算法相比最先进的去中心化方法给出了更优的数值性能。

附录A. 若干引理的证明。在本节中，我们给出分析中所使用的若干引理的证明。

A.1. 引理的证明 3.8。 选择 η_x后，很容易得到

$$\frac{2L^2(1+6\kappa^2)\eta_x^2}{(1-\rho)^3} \le \frac{2(1-\rho)}{25}, \quad \frac{6mL^2\kappa^2\eta_x^2}{(1-\rho)^3} \le \frac{m(1-\rho)}{25}, \quad 1-\rho-\frac{8\eta_x^2L^2(1+6\kappa^2)}{(1-\rho)^3} \ge \frac{2(1-\rho)}{3}. \tag{A.1}$$

根据 α_x 的定义并回顾 $L_P = L\sqrt{1+4\kappa^2}$，可得

$$\alpha_x \ge \frac{1}{2m}\left(\frac{5L\sqrt{1+6\kappa^2}}{(1-\rho)^2} - 2L\sqrt{1+4\kappa^2} - L(\kappa+1)\right) \ge \frac{L(4\kappa+1)}{2m(1-\rho)^2},$$

其中，第二个不等式由$(1-\rho)^2 \le 1$ 及简单计算得到。此外，由 $\tilde{\alpha}_x$ 的公式可得

$$\begin{aligned}
\tilde{\alpha}_x &\ge \frac{1}{2m\eta_x} - \frac{L^2(1+6\kappa^2)}{mL_P(1-\rho)^2} - \frac{2c(1-\rho)}{25} \ge \frac{1}{2m\eta_x} - \frac{L^2(1+6\kappa^2)}{mL_P(1-\rho)^2} - \frac{3}{25}\left(\frac{\rho^2}{2m\eta_x} + \frac{L(4\kappa+1)}{2m} + \frac{4L^2(1+6\kappa^2)}{mL_P(1-\rho)^2}\right) \\
&\ge \frac{11}{25m\eta_x} - \frac{37L^2(1+6\kappa^2)}{25mL_P(1-\rho)^2} - \frac{3L(4\kappa+1)}{50m} \ge \frac{11L\sqrt{1+6\kappa^2}}{5m(1-\rho)^2} - \frac{37L(1+6\kappa^2)}{25m\sqrt{1+4\kappa^2}(1-\rho)^2} - \frac{3L(4\kappa+1)}{50m} \\
&\ge \frac{L}{m(1-\rho)^2}\left(\left(\frac{11}{5} - \frac{37\sqrt{3}}{25\sqrt{2}}\right)\sqrt{1+6\kappa^2} - \frac{3(1+4\kappa)}{50}\right) \ge \frac{L(1+4\kappa)}{m(1-\rho)^2}\left(\frac{9}{50} - \frac{3}{50}\right) = \frac{3L(1+4\kappa)}{25m(1-\rho)^2}
\end{aligned}$$

其中，第一个不等式用到了（A.1）中的第一个结果；第二个不等式由 c的定义及（A.1）中的第三个结果得到；第三个不等式由 $\rho^2 < 1$成立；第四个不等式用到了 η_x 和$L_P = L\sqrt{1+4\kappa^2}$的取值；第五个不等式依据事实 $\frac{\sqrt{1+6\kappa^2}}{\sqrt{1+4\kappa^2}} \le \frac{\sqrt{3}}{\sqrt{2}}$ 和$(1-\rho)^2 \le 1$；第六个不等式由$\left(\frac{11}{5} - \frac{37\sqrt{3}}{25\sqrt{2}}\right)\sqrt{1+6\kappa^2} \ge \frac{9(1+4\kappa)}{50}$得到。

对于 α_λ，我们有

$$\begin{aligned}
\alpha_\lambda &\ge \frac{1}{\eta_\lambda} - \frac{L\kappa}{2} - \frac{L_P}{2} - \frac{3L^2\kappa^2}{L_P(1-\rho)^2} - \frac{cm(1-\rho)}{25} \\
&\ge \frac{1}{\eta_\lambda} - \frac{L\kappa}{2} - \frac{L_P}{2} - \frac{3L^2\kappa^2}{L_P(1-\rho)^2} - \frac{3m}{50}\left(\frac{\rho^2}{2m\eta_x} + \frac{L(4\kappa+1)}{2m} + \frac{4L^2(1+6\kappa^2)}{mL_P(1-\rho)^2}\right) \\
&= \frac{1}{\eta_\lambda} - \frac{3L\rho^2\sqrt{1+6\kappa^2}}{20(1-\rho)^2} - \frac{L}{2}\left(\frac{62\kappa}{50} + \frac{3}{50}\right) - \frac{L\sqrt{1+4\kappa^2}}{2} - \frac{3L}{\sqrt{1+4\kappa^2}(1-\rho)^2}\left(\frac{37\kappa^2}{25} + \frac{2}{25}\right) \\
&\ge \frac{1}{\eta_\lambda} - \frac{L}{(1-\rho)^2}\left(\frac{3(\sqrt{6}\kappa+1)}{20} + \frac{62\kappa}{100} + \frac{3}{100} + \kappa + \frac{1}{2} + 3\kappa\right) \ge \frac{1}{\eta_\lambda} - \frac{L(5\kappa+1)}{(1-\rho)^2} = \frac{L(4\kappa+1)}{(1-\rho)^2},
\end{aligned}$$

其中，第一个不等式使用了（A.1）中的第二个结果；第二个不等式根据 c 的定义以及（A.1）中的第三个结果得出；第三个不等式由 $\sqrt{1+6\kappa^2} \le \sqrt{6}\kappa+1$、$\sqrt{1+4\kappa^2} \le 2\kappa+1$、$\frac{37\kappa^2+2}{25\sqrt{1+4\kappa^2}} \le \kappa$以及 $(1-\rho)^2 \le 1$得出。此外，我们有

$$\begin{aligned}
\frac{3\kappa^2}{2L_P(1-\rho)^2} + \frac{3cm\eta_x^2\kappa^2}{(1-\rho)^3} &\le \frac{3\kappa^2}{2L_P(1-\rho)^2} + \frac{9m\eta_x^2\kappa^2}{2(1-\rho)^4}\left(\frac{\rho^2}{2m\eta_x} + \frac{L(4\kappa+1)}{2m} + \frac{4L^2(1+6\kappa^2)}{mL_P(1-\rho)^2}\right) \\
&= \frac{\kappa^2}{L\sqrt{1+4\kappa^2}(1-\rho)^2}\left(\frac{3}{2} + \frac{18}{25}\right) + \frac{9\kappa^2\rho^2}{20L\sqrt{1+6\kappa^2}(1-\rho)^2} + \frac{9\kappa^2(1+4\kappa)}{100L(1+6\kappa^2)} \\
&\le \frac{\kappa^2}{L(1-\rho)^2}\left(\left(\frac{3}{2} + \frac{18}{25}\right)\frac{1}{2\kappa} + \frac{9}{20\sqrt{6}\kappa} + \frac{9}{140\kappa}\right) \le \frac{3\kappa}{2L(1-\rho)^2},
\end{aligned}$$

其中，第一个不等式由 c 在（3.40b）中的公式及（A.1）中的第三个结果得出；第二个不等式用到了 $\sqrt{1+4\kappa^2} \ge 2\kappa$、$\sqrt{1+6\kappa^2} \ge \sqrt{6}\kappa$、$\frac{1+4\kappa}{1+6\kappa^2} \le \frac{5}{7\kappa}$和 $1-\rho \le 1$。

最后，有

$$\begin{aligned}\tfrac{1}{2mL_P(1-\rho)}+\tfrac{c\eta_x^2}{(1-\rho)^2}&\le\tfrac{1}{2mL_P(1-\rho)}+\tfrac{3\eta_x^2}{2(1-\rho)^3}\left(\tfrac{\rho^2}{2m\eta_x}+\tfrac{L(4\kappa+1)}{2m}+\tfrac{4L^2(1+6\kappa^2)}{mL_P(1-\rho)^2}\right)\\&=\tfrac{1}{2mL_P(1-\rho)}+\tfrac{3(1-\rho)}{50L^2(1+6\kappa^2)}\left(\tfrac{5L\rho^2\sqrt{1+6\kappa^2}}{2m(1-\rho)^2}+\tfrac{L(4\kappa+1)}{2m}+\tfrac{4L^2(1+6\kappa^2)}{mL_P(1-\rho)^2}\right)\\&=\tfrac{37}{50mL_P(1-\rho)}+\tfrac{3\rho^2}{20mL\sqrt{1+6\kappa^2}(1-\rho)}+\tfrac{3(1-\rho)(1+4\kappa)}{100mL(1+6\kappa^2)}\\&=\tfrac{37}{50mL\sqrt{1+4\kappa^2}(1-\rho)}+\tfrac{3\rho^2}{20mL\sqrt{1+6\kappa^2}(1-\rho)}+\tfrac{3(1-\rho)(1+4\kappa)}{100mL(1+6\kappa^2)}\\&\le\tfrac{1}{mL(1-\rho)}\left(\tfrac{37}{100\kappa}+\tfrac{3}{20\sqrt{6}\kappa}+\tfrac{3}{140\kappa}\right)\le\tfrac{1}{2mL\kappa(1-\rho)},\end{aligned}$$

其中，第一个不等式用到了 c 的定义以及（A.1）中的第三个结果；第二个不等式由 $\sqrt{1+4\kappa^2}\ge 2\kappa$、$\sqrt{1+6\kappa^2}\ge\sqrt{6}\kappa$、$\frac{1+4\kappa}{1+6\kappa^2}\le\frac{5}{7\kappa}$ 和 $1-\rho\le 1$ 得出。证明完毕。

A.2. 引理3.14的证明。 由 d_i^t 的凹性可知，$d_i^t(\mathbf{y}_i^{t\star})-d_i^t(\mathbf{y}_i^{t-1})\le\langle\boldsymbol{\zeta}_i^{t-1},\mathbf{y}_i^{t\star}-\mathbf{y}_i^{t-1}\rangle$ 对任意 $\boldsymbol{\zeta}_i^{t-1}\in\partial d_i^t(\mathbf{y}_i^{t-1})$ 均成立，即：

$$d_i^t(\mathbf{y}_i^{t\star})-d_i^t(\mathbf{y}_i^{t-1})\le\left\langle\nabla_{\mathbf{y}}f_i(\mathbf{x}_i^t,\mathbf{y}_i^{t-1})-\boldsymbol{\xi}_i^{t-1}-\tfrac{L\sqrt{m}}{2}\tilde{\boldsymbol{\lambda}}_i^t,\mathbf{y}_i^{t\star}-\mathbf{y}_i^{t-1}\right\rangle,\forall\,\boldsymbol{\xi}_i^{t-1}\in\partial h(\mathbf{y}_i^{t-1}).\tag{A.2}$$

由条件 $\mathrm{dist}(\mathbf{0},\ \partial d_i^{t-1}(\mathbf{y}_i^{t-1}))\le\delta_{t-1}$ 可知，存在一个次梯度 $\boldsymbol{\xi}_i^{t-1}\in\partial h(\mathbf{y}_i^{t-1})$，使得 $\|\nabla_{\mathbf{y}}f_i(\mathbf{x}_i^{t-1},\ \mathbf{y}_i^{t-1})-\boldsymbol{\xi}_i^{t-1}-\frac{L\sqrt{m}}{2}\tilde{\boldsymbol{\lambda}}_i^{t-1}\|\le\delta_{t-1}$；结合（A.2）可得

$$\begin{aligned}&d_i^t(\mathbf{y}_i^{t\star})-d_i^t(\mathbf{y}_i^{t-1})\\&\le\left\langle\nabla_{\mathbf{y}}f_i(\mathbf{x}_i^t,\mathbf{y}_i^{t-1})-\nabla_{\mathbf{y}}f_i(\mathbf{x}_i^{t-1},\mathbf{y}_i^{t-1})-\tfrac{L\sqrt{m}}{2}(\tilde{\boldsymbol{\lambda}}_i^t-\tilde{\boldsymbol{\lambda}}_i^{t-1}),\mathbf{y}_i^{t\star}-\mathbf{y}_i^{t-1}\right\rangle+\delta_{t-1}\|\mathbf{y}_i^{t\star}-\mathbf{y}_i^{t-1}\|.\end{aligned}$$

现在利用 L-光滑性以及 f_i 的杨氏不等式，由上述不等式可得

$$d_i^t(\mathbf{y}_i^{t\star})-d_i^t(\mathbf{y}_i^{t-1})\le\tfrac{L}{2}\|\mathbf{x}_i^t-\mathbf{x}_i^{t-1}\|^2+\tfrac{L\sqrt{m}}{4}\|\tilde{\boldsymbol{\lambda}}_i^t-\tilde{\boldsymbol{\lambda}}_i^{t-1}\|^2+\left(\tfrac{L}{2}+\tfrac{L\sqrt{m}}{4}+\tfrac{1}{2}\right)\|\mathbf{y}_i^{t\star}-\mathbf{y}_i^{t-1}\|^2+\tfrac{\delta_{t-1}^2}{2}.\tag{A.3}$$

此外，$\|\mathbf{y}_i^{t\star}-\mathbf{y}_i^{t-1}\|^2\le 2\|\mathbf{y}_i^{t\star}-\mathbf{y}_i^{t-1\star}\|^2+2\|\mathbf{y}_i^{t-1\star}-\mathbf{y}_i^{t-1}\|^2\le 2\|\mathbf{y}_i^{t\star}-\mathbf{y}_i^{t-1\star}\|^2+\frac{2\delta_{t-1}^2}{\mu^2}$，其中第二个不等式由命题 3.2 的证明可得。因此，（A.3）表明

$$\begin{aligned}&d_i^t(\mathbf{y}_i^{t\star})-d_i^t(\mathbf{y}_i^{t-1})\\&\le\tfrac{L}{2}\|\mathbf{x}_i^t-\mathbf{x}_i^{t-1}\|^2+\tfrac{L\sqrt{m}}{4}\|\tilde{\boldsymbol{\lambda}}_i^t-\tilde{\boldsymbol{\lambda}}_i^{t-1}\|^2+\left(L+\tfrac{L\sqrt{m}}{2}+1\right)\left(\|\mathbf{y}_i^{t\star}-\mathbf{y}_i^{t-1*}\|^2+\tfrac{\delta_{t-1}^2}{\mu^2}\right)+\tfrac{\delta_{t-1}^2}{2}\\&\le\tfrac{L}{2}\|\mathbf{X}^t-\mathbf{X}^{t-1}\|_F^2+\tfrac{L\sqrt{m}}{4}\|\tilde{\boldsymbol{\Lambda}}^t-\tilde{\boldsymbol{\Lambda}}^{t-1}\|_F^2+\left(L+\tfrac{L\sqrt{m}}{2}+1\right)\left(\|\mathbf{Y}^{t\star}-\mathbf{Y}^{t-1*}\|_F^2+\tfrac{\delta_{t-1}^2}{\mu^2}\right)+\tfrac{\delta_{t-1}^2}{2}.\end{aligned}\tag{A.4}$$

此外，由（3.8c），可得 $\|\mathbf{Y}^{t\star}-\mathbf{Y}^{t-1*}\|_F^2\le 2\kappa^2\|\mathbf{X}^t-\mathbf{X}^{t-1}\|_F^2+2m\kappa^2\|\boldsymbol{\Lambda}^t-\boldsymbol{\Lambda}^{t-1}\|_F^2$。因此，由（A.4）和 $\|\tilde{\boldsymbol{\Lambda}}^t-\tilde{\boldsymbol{\Lambda}}^{t-1}\|_F^2\le 4\|\boldsymbol{\Lambda}^t-\boldsymbol{\Lambda}^{t-1}\|_F^2$，可得

$$\begin{aligned}d_i^t(\mathbf{y}_i^{t\star})-d_i^t(\mathbf{y}_i^{t-1})&\le\left(\tfrac{L}{2}+\kappa^2(2L+L\sqrt{m}+2)\right)\|\mathbf{X}^t-\mathbf{X}^{t-1}\|_F^2\\&\quad+\left(L\sqrt{m}+m\kappa^2(2L+L\sqrt{m}+2)\right)\|\boldsymbol{\Lambda}^t-\boldsymbol{\Lambda}^{t-1}\|_F^2+\left(L+\tfrac{L\sqrt{m}}{2}+1\right)\tfrac{\delta_{t-1}^2}{\mu^2}+\tfrac{\delta_{t-1}^2}{2}.\end{aligned}\tag{A.5}$$

最后，我们 在上面的不等式中使用 (3.31) 和 (3.49b)–(3.49d) 中的第二个不等式，可得 ain

$$\begin{aligned}&d_i^t(\mathbf{y}_i^{t\star})-d_i^t(\mathbf{y}_i^{t-1})\\&\le\left(\tfrac{L}{2}+\kappa^2(2L+L\sqrt{m}+2)\right)\left(\tfrac{50mC_{0,t-1}(1-\rho)^2}{3L(1+4\kappa)}+\tfrac{12(1-\rho)}{25L^2(1+6\kappa^2)}\left\|\mathbf{V}_\perp^0\right\|_F^2+\tfrac{212mC_{0,t-1}(1-\rho)^2}{25L(4\kappa+1)}+\tfrac{12m}{25L^2}\Delta_{t-1}\right)\\&\quad+\left(L\sqrt{m}+m\kappa^2(2L+L\sqrt{m}+2)\right)\tfrac{C_{0,t-1}(1-\rho)^2}{L(4\kappa+1)}+\left(L+\tfrac{L\sqrt{m}}{2}+1\right)\tfrac{\delta_{t-1}^2}{\mu^2}+\tfrac{\delta_{t-1}^2}{2}.\end{aligned}$$

合并上面不等式中的同类项，并注意到 $\frac{50}{3} + \frac{212}{25} \leq 21$，$\frac{12}{25} \leq \frac{1}{2}$，即可得到所需结果。

附录B. 非凸强凹问题中Minty变分不等式条件的失效。我们给出一个简单例子，说明为处理非凸非凹问题而在 [24] 中假设的Minty变分不等式条件，可能不适用于NCSC极小极大问题。设对任意 $x \in \mathbb{R}$、$y \in \mathbb{R}$，$f(x,y) = -x^2y + \frac{1}{2}y^2$。那么，$\min_{-1\leq x\leq 1} \max_y f(x,y)$是一个定义良好的NCSC极小极大问题，并满足我们所作的假设1。为说明这一点，任取 x，有 $y^* := \arg\max_y f(x,y) = x^2$。因此，原问题为$\min_{-1\leq x\leq 1} p(x) := -\frac{1}{2}x^4$，它有两个有限个极小解 $x^* = \pm 1$。然而，Minty条件，即存在$(\bar{x},\bar{y})$满足 $-1 \leq \bar{x} \leq 1$，使得$\langle (f'_x(x,y) - f'_y(x,y)), (x,y) - (\bar{x},\bar{y})\rangle = -2xy(x-\bar{x}) + (x^2-y)(y-y^*) \geq 0$，对$\forall -1 \leq x \leq 1$、$\forall y \in \mathbb{R}$成立，并不成立，因为对任意$(\bar{x},\bar{y})$及固定的 $x \in [-1,1]$，容易得到 $-2xy(x-\bar{x}) + (x^2-y)(y-y^*) \to -\infty$，其中 $|y| \to \infty$。

参考文献

[1] A. Beznosikov、P. Dvurechensky、A. Koloskova、V. Samokhin、S. U. Stich 和 A. Gasnikov。用于变分不等式的去中心化局部随机外梯度法。 *arXiv preprint arXiv:2106.08315*，2021年。[2] A. Beznosikov、G. Scutari、A. Rogozin 和 A. Gasnikov。数据相似性条件下的分布式鞍点问题。*Advances in Neural Information Processing Systems*，34:8172–8184，2021年。[3] C.-C. Chang 和 C.-J. Lin。LIBSVM：支持向量机库。 *ACM transactions on intelligent systems and technology (TIST)* 2 3 1 27, 2011 ，( ): – 。

[4] L. Chen、H. Ye 和 L. Luo。用于去中心化非凸强凹极小极大优化的简单高效随机算法。 *arXiv preprint arXiv:2212.02387*，2022年。[5] B. Dai、A. Shaw、L. Li、L. Xiao、N. He、Z. Liu、J. Chen 和 L. Song。Sbeed：采用非线性函数逼近的收敛型强化学习。载于 *International Conference on Machine Learning*，第1125–1134页。PMLR，2018年。[6] J. M. Danskin。极大极小理论及其应用。 *SIAM Journal on Applied Mathematics*，14(4):641–664，1966年。[7] A. Defazio、F. Bach 和 S. Lacoste-Julien。SAGA：一种支持非强凸复合目标函数的快速增量梯度法。 *Advances in neural information processing systems*，27，2014年。[8] A. Falsone、K. Margellos、S. Garatti 和 M. Prandini。带耦合约束的多智能体分布式优化的对偶分解。 *Automatica*，84:149–158，2017年。[9] A. Falsone、I. Notarnicola、G. Notarstefano 和 M. Prandini。用于分布式约束耦合优化的跟踪交替方向乘子法。 *Automatica*，117:108962，2020年。[10] H. Gao。用于有限和极小极大问题的去中心化随机梯度下降上升法。 *arXiv preprint arXiv:2212.02724* 2022 ，。[11]

I. Goodfellow、J. Pouget-Abadie、M. Mirza、B. Xu、D. Warde-Farley、S. Ozair、A. Courville 和 Y. Bengio。生成对抗网络。 *Communications of the ACM*，63(11):139–144，2020年。[12] B. Grimmer、H. Lu、P. Worah 和 V. Mirrokni。非凸非凹极小极大优化中近端点法的图景。 *Mathematical Programming*，第1–35页，2022年。[13] F. Huang、S. Gao、J. Pei 和 H. Huang。从小规模到极小极大优化的加速零阶和一阶动量方法。 *J. Mach. Learn. Res.*，23:36–1，2022年。[14] M. Huang、S. Ma 和 L. Lai。关于公平最优传输中投影交替最大化的收敛性。 *arXiv preprint arXiv:2109.15030*，2021年。[15] C. Jin、P. Netrapalli 和 M. Jordan。非凸非凹极小极大优化中的局部最优性是什么？载于*International conference on machine learning*，第4880–4889页。PMLR，2020年。[16] J. Koshal、A. Nedi

´ *SIAM* c 和 U. *Journal on Optimization* 21 3 1046 1081, 2011
V. Shanbhag。多用户优化：分布式算法与误差分析。，( ): – 。[17] D. Kovalev、A. Beznosikov、A. Sadiev、M. I. Persiianov、P. Richt´arik 和 A. Gasnikov。去中心化随机变分不等式的最优算法。载于 *Advances in Neural Information Processing Systems*，2022年。[18] J. Li、L. Zhu 和 A. M.-C. So。非光滑复合非凸凹极小极大优化。 *arXiv preprint arXiv:2209.10825* 2022 ，。[19]

Q. Lin 和 L. Xiao。稀疏优化的自适应加速近端梯度法及其同伦延拓。 *Computational Optimization and Applications*，60(3):633–674，2015年。[20] T. Lin、C. Jin 和 M. Jordan。关于非凸凹极小极大问题的梯度下降上升法。载于 *International Conference on Machine Learning* 6083 6093 2020 ，第–页。

PMLR，。[21] T. Lin、C. Jin 和 M. I. Jordan。极小极大优化的近最优算法。载于 *Conference on Learning Theory*，第2738–2779页。PMLR，2020年。[22] M. Liu、Y. Mroueh、J. Ross、W. Zhang、X. Cui、P. Das 和 T. Yang。深入理解生成对抗网络中的自适应梯度算法。 *arXiv preprint arXiv:1912.11940*，2019年。

[23] M. Liu、H. Rafique、Q. Lin 和 T. Yang。弱凸弱凹极小极大问题的一阶收敛理论。 *The Journal of Machine Learning Research*，22(1):7651–7684，2021年。[24] M. Liu、W. Zhang、Y. Mroueh、X. Cui、J. Ross、T. Yang 和 P. Das。用于训练生成对抗网络的去中心化并行算法。 *Advances in Neural Information Processing Systems*，33:11056–11070，2020年。[25] W. Liu、A. Mokhtari、A. Ozdaglar、S. Pattathil、Z. Shen 和 N. Zheng。用于鞍点问题的去中心化近端点型方法。 *arXiv preprint arXiv:1910.14380*，2019年。[26] S. Lu、I. Tsaknakis、M. Hong 和 Y. Chen。用于单侧非凸极小极大问题的混合分块逐次逼近：算法与应用。 *IEEE Transactions on Signal Processing*，68:3676–3691，2020年。[27] S. Lu、K. Zhang、T. Chen、T. Ba¸sar 和 L. Horesh。用于安全多智能体强化学习的去中心化策略梯度下降上升。载于 *Proceedings of the AAAI Conference on Artificial Intelligence* 35， 8767 8775, ，卷页码–。[28] G. 2021
Mancino-Ball、S. Miao、Y. Xu 和 J. Chen。用于非凸复合去中心化优化的近端随机递归动量方法。 *arXiv preprint arXiv:2211.11954*，2022年。[29] D. Mateos-N ´ ´ *2015 54th IEEE Con-* unez
ference on Decision and Control (CDC) 5462 5467 2015
和 J. Cortes。用于鞍点问题的分布式次梯度方法。载于，页码–。IEEE，。[30] D. Mateos-N
´ ´ *IEEE Trans-* unez 和 J. Cortes。采用
actions on Automatic Control 62 6 2720 2735, 2016
拉普拉斯平均的分布式鞍点次梯度算法。，( ): – 。[31] D. Metelev、A. Rogozin、A. Gasnikov 和 D. Kovalev。强凸与强凹参数不同的去中心化鞍点问题。 *arXiv preprint arXiv:2206.00090*，2022年。[32] S. Mukherjee 和 M. Chakraborty。用于大规模极小极大问题的去中心化算法。载于 *2020 59th IEEE Conference on Decision and Control (CDC)* 2967 2972 2020 ，页码–。IEEE，。[33]

Y. Nesterov。用于最小化复合函数的梯度方法。 *Mathematical Programming*，140(1):125–161，2013年。[34] D. M. Ostrovskii、A. Lowy 和 M. Razaviyayn。在非凸凹光滑极小极大问题中高效搜索一阶纳什均衡。 *SIAM Journal on Optimization*，31(4):2508–2538，2021年。[35] D. Paccagnan、B. Gentile、F. Parise、M. Kamgarpour 和 J. Lygeros。带仿射耦合约束的二次聚合博弈中广义纳什均衡的分布式计算。载于 *2018 IEEE 57th conference on decision and control (CDC)* 6123 6128 2016 ，页码–。IEEE，。[36]

H. Rafique、M. Liu、Q. Lin 和 T. Yang。弱凸凹极小极大优化：可证明算法及其在机器学习中的应用。 *Optimization Methods and Software*，37(3):1087–1121，2022年。[37] A. Richards 和 J. How。用于鲁棒约束模型预测控制的去中心化算法。载于 *Proceedings of the 2004 American control conference* 5， 4261 4266 2004 ，卷页码–。IEEE，。

[38] R. T. Rockafellar。 *Convex analysis*，第11卷。普林斯顿大学出版社，1997年。[39] G. Scutari 和 Y. Sun。时变有向图上的分布式非凸约束优化。*Programming* 176 497 544, 2019 *Mathematical* ，：– 。[40] K. K.

Thekumparampil、P. Jain、P. Netrapalli 和 S. Oh。光滑极小极大优化的高效算法。 *Advances in Neural Information Processing Systems* 32 2019 ，，。[41] I.

Tsaknakis、M. Hong 和 S. Liu。去中心化极小极大优化：网络投毒攻击中的建模、算法与应用。载于 *ICASSP 2020-2020 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)* 5755 5759 2020 ，，页码–。IEEE，。
[42] W. Wiesemann、D. Kuhn 和 M. Sim。分布鲁棒凸优化。 *Operations Research*，62(6):1358–1376，2014年。[43] Y. Xu。具有 o (1) 个函数约束的问题的一阶方法，其收敛速率几乎可与无约束问题相同。 *SIAM Journal on Optimization*，32(3):1759–1790，2022年。[44] Z. Xu、H. Zhang、Y. Xu 和 G. Lan。用于非凸凹和凸非凹极小极大问题的统一单循环交替梯度投影算法。 *Mathematical Programming*，第1–72页，2023年。[45] Y. Yan、Y. Xu、Q. Lin、L. Zhang 和 T. Yang。对于不具有双线性结构的问题，收敛速度快于 $O(1/\sqrt{T})$的随机原始-对偶算法。 *arXiv preprint arXiv:1904.10112*，2019年。[46] J. Zhang、P. Xiao、R. Sun 和 Z. Luo。用于非凸凹极小极大问题的单循环平滑梯度下降-上升算法。 *Advances in neural information processing systems*，33:7377–7389，2020年。[47] S. Zhang、J. Yang、C. Guzm´an、N. Kiyavash 和 N. He。非凸强凹极小极大优化的复杂度。载于 *Uncertainty in Artificial Intelligence*，第482–492页。PMLR，2021年。[48] X. Zhang、N. S. Aybat 和 M. Gurbuzbalaban。Sapd+：用于非凸凹极小极大问题的加速随机方法。 *arXiv preprint arXiv:2205.15084*，2022年。[49] X. Zhang、Z. Liu、J. Liu、Z. Zhu 和 S. Lu。降低合作式多智能体强化学习中去中心化策略评估的通信与样本复杂度。 — *Advances in Neural Information Processing Systems*，34 18825: –18838，2021年。