arXiv:2303.02532v1 [cs.LG] 5 Mar 2023

PRECISION：低通信复杂度和样本复杂度的去中心化约束极小极大学习

Zhuqing Liu[+]、Xin Zhang[*]、Songtao Lu[◦] 和 Jia Liu[++]电气与计算机工程系，俄亥俄州立大学
[*]统计系，爱荷华州立大学
[◦]IBM 人工智能研究院，托马斯·J·沃森研究中心

摘要

近年来，极小极大优化问题因其在机器学习（ML）中的广泛应用而日益受到关注。然而，现有的大多数极小极大求解技术要么基于单机，要么是由中央服务器协调的分布式算法。本文聚焦于带定义域约束的去中心化极小极大优化学习问题，其中多个智能体在不依赖任何服务器协调的情况下，共同求解非凸-强凹极小极大鞍点问题。带定义域约束的去中心化极小极大优化问题支撑着许多重要的机器学习应用，包括多智能体机器学习中的公平性保障，以及多智能体强化学习中的策略评估。我们提出了一种名为 PRECISION（近端梯度跟踪和随机递归方差缩减）的算法，其收敛速率为 $\mathcal{O}(1/T)$，其中 T 是最大迭代次数。为进一步降低样本复杂度，我们提出了采用自适应批量大小技术的 PRECISION$^+$ 。我们证明，PRECISION 和 PRECISION$^+$ 快速的 $\mathcal{O}(1/T)$ 收敛到 ϵ-驻点，意味着 $\mathcal{O}(\epsilon^{-2})$ 通信复杂度和 $\mathcal{O}(m\sqrt{n}\epsilon^{-2})$ 样本复杂度，其中 m 是智能体数量，n 是每个智能体的数据集大小。据我们所知，这是首项在带定义域约束的去中心化极小极大学习中，同时在样本复杂度和通信复杂度方面实现 $\mathcal{O}(\epsilon^{-2})$ 的工作。我们的实验也验证了理论结果。

1 引言

近年来，机器学习（ML）在许多领域取得了巨大成功，包括机器人学 Siau and Wang [2018],图像识别 Özyurt [2020],自然语言处理 Nozaki and Nakamoto [2018]，推荐系统 Deldjoo 等 [2020]，，仅举几例。传统上，机器学习模型的训练通常部署在大型数据中心内、易于访问大型训练数据集的高性能计算集群上。然而，随着更多样化的机器学习应用不断涌现，由于以下原因，机器学习的部署也逐渐向计算网络和通信网络的边缘迁移：首先，在许多机器学习应用中，数据由多种数据源生成和收集，而这些数据源在地理位置上分散（例如，智能移动设备、车辆、环境传感器、卫星图像）。其次，由于设备的通信能力有限且存在数据隐私方面的顾虑，将边缘网络收集的数据发送到云端进行集中处理，成本高昂，甚至不可行。近年来，这些现实限制推动了去中心化学习在边缘网络上的快速发展；这种学习方式能够利用拓扑结构任意且高度灵活的点对点边缘计算网络 Nedic 和 Ozdaglar [2009], Lian 等 [2017]。此外，由于去中心化学习具有抵御单点故障、保护数据隐私和易于实施等优势，近年来愈发受到关注，并已应用于各种科学和工程领域，例如分布式机器人控制 Ren 等 [2007], Zhou and Roumeliotis [2011] 、网络资源分配 Jiang 等 [2018], Rhee 等 [2012],，以及字典学习 Chen 等 [2014],、多智能体系统 Cao 等 [2012],Zhou and Roumeliotis [2011],、多任务学习 Wang 等[2018], Zhang 等 [2019],和信息检索 Ali and Van Stam [2004]。

从数学角度来看，在计算网络上开展去中心化学习，实质上就是由网络中的一组智能体以分布式方式并协同地求解一个优化问题。然而，在现有的去中心化学习文献中，大多数研究都聚焦于标准的损失最小化表述，即 $\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$，其中 $f(\cdot)$ 表示学习的损失目标函数，$\mathbf{x}$ 表示待学习的全局模型参数，d 为模型维度。尽管这种标准的损失最小化表述足够通用，可以

涵盖广泛的机器学习应用（例如，机器人网络 Smart 和 Kaelbling [2002],Kober 等 [2013],Polydoros 和 Nalpantidis [2017]；传感器网络 Cortes 等 [2004], Ogren 等 [2004],Rabbat 和 Nowak [2004]；电力网络 Callaway 和 Hiskens [2010], Dall'Anese 等 [2013], Ernst 等 [2004], Glavic 等 [2017]），但其数学结构不足以刻画机器学习应用不断涌现的新需求，这一点已日益显而易见。值得注意的是，如今许多复杂的机器学习问题都需要采用所谓的"极小极大"优化，其形式为 $\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x},\mathbf{y})$，其中$\mathbf{x}$ 和$\mathbf{y}$ 都是待学习的参数（维度可能不同），而 $\mathcal{X}$和 $\mathcal{Y}$分别是与$\mathbf{x}$ 和$\mathbf{y}$相对应的实子空间。尽管极小极大优化的历史也可追溯至 1945 年 Wald [1945],的研究，但去中心化极小极大优化目前仍处于起步阶段，相关研究成果也十分有限。

在本文中，我们不研究 Liu 等人 [2019, 2020],所讨论的无结构一般去中心化极小极大问题，而是聚焦于一类有趣的去中心化极小极大优化问题：多个智能体协同求解域约束非凸-强凹（NCX-SCV）极小极大问题。去中心化约束 NCX-SCV 极小极大问题十分重要，因为它自然地出现在许多新兴的多智能体机器学习应用中，例如对抗训练中的多智能体公平性约束（Xu 等人 [2021],）、多智能体强化学习（MARL）中的策略评估（Qiu 等人 [2020],），以及机器学习中的多智能体公平性保障（Baharlouei 等人 [2019], ； Sattigeri 等人 [2018] ）（更深入的讨论见第 2 节）。

然而，由于以下技术挑战，为去中心化约束 NCX-SCV 极小极大问题设计有效且高效的算法并非易事：首先，极小极大优化处理的是内层最大化问题与外层最小化问题的复合。这种紧密耦合的内外层数学结构，加上去中心化特性以及外层问题的非凸性，使算法设计和理论分析相当困难。此外，内层和外层问题中的约束结构，又为去中心化约束 NCX-SCV 极小极大问题的算法设计带来了额外挑战。其次，边缘计算网络中的去中心化面临两项根本相互冲突的 性能指标。一方面，鉴于深度学习模型的高维性和大规模数据集，算法设计中无法利用一阶随机梯度以外的信息来确定搜索方向。虽然可以通过增加小批量中的训练样本数量来降低随机梯度的方差，但这会增加随机梯度的计算成本。另一方面，如果每次迭代使用较少的训练样本以降低计算成本，随机梯度中较大的方差就不可避免地会导致需要更多通信轮次才能达到特定的训练准确率（即收敛速度更慢）。高通信复杂度在无线边缘网络中尤其棘手，因为通信连接可能速度较慢且极不可靠。第三，与无约束情形相比，受约束的去中心化极小极大优化面临着显著更大的挑战。这主要是因为定义域约束具有非光滑特性，而且这些约束与极小极大问题结构之间存在复杂的耦合关系。

本文的主要贡献是提出了一系列新的算法技术，以应对上述挑战，并在去中心化约束 NCX-SCV 极小极大问题中实现较低的样本复杂度和通信复杂度。我们的主要技术成果及其意义总结如下：

• 我们提出了一种名为 PRECISION（近端梯度跟踪与随机递归方差缩减）的去中心化约束极小极大优化算法，并证明, –为了达到一个 ϵ-驻点，PRECISION-的收敛速率为$\mathcal{O}(1/T)$（T 是最大迭代次数）。这一结果进一步表明，其样本-通信复杂度为 $[\mathcal{O}(m\sqrt{n}\epsilon^{-2})$ $\mathcal{O}(\epsilon^{-2})]$，其中 m 是智能体数量， n 是每个智能体本地数据集的大小。

•为放宽 PRECISION 对全梯度计算的要求，我们提出了一种基于自适应批量大小技术的增强算法，称为 PRECISION$^+$ 。PRECISION$^+$ 进一步降低了 PRECISION 的样本复杂度，同时保留了与 PRECISION 相同的 $[\mathcal{O}(m\sqrt{n}\epsilon^{-2})$、 $\mathcal{O}(\epsilon^{-2})]$ 样本-通信复杂度缩放规律。此外，PRECISION$^+$还可通过适当牺牲通信复杂度来降低样本复杂度（这种权衡仅体现在隐藏的大 O 常数中）。

• 我们注意到，PRECISION 和 PRECISION$^+$ 算法均结合了两个近端算子，分别用于内层和外层约束（针对 $\mathbf{x}$ 和 $\mathbf{y}$）；同时，它们还对内层和外层更新采用方差缩减技术，并对内层和外层变量采用基于梯度跟踪的更新。从这个意义上说，这两种基于 PRECISION 的算法都可视为一种三重混合方法，因此需要新的性能分析和证明技术。还值得指出的是，PRECISION 中提出的算法和证明技术，或许本身也会引起去中心化极小极大学习理论领域的兴趣。

本文其余部分的组织如下。第 2 节首先介绍去中心化

极小极大优化问题的预备知识，并讨论相关工作。第 3 节提出两种随机方差缩减算法，即 PRECISION 和 PRECISION$^+$。PRECISION 和 PRECISION$^+$ 的收敛速率、通信复杂度及样本复杂度也将在第 3 节给出。第 4 节提供数值结果，以验证我们的理论发现；第 5 节总结全文。

2 预备知识与相关工作

为便于后续的技术讨论，我们首先在第 2.1 节介绍去中心化极小极大优化及其共识形式的基础知识。随后，我们正式定义去中心化极小极大优化问题的共识形式中的样本复杂度和通信复杂度。接下来，在第 2.2 节，我们概述用于求解极小极大学习问题的现有优化算法及其在样本复杂度和通信复杂度方面的性能，从而以比较的视角审视我们的工作。

2.1 去中心化极小极大优化的预备知识

1) 网络共识形式： 考虑一个无向连通网络 $\mathcal{G} = (\mathcal{N}, \mathcal{L})$，其中 $\mathcal{N}$ 和 $\mathcal{L}$ 分别是节点（智能体）和边的集合，且有 $|\mathcal{N}| = m$。每个智能体都具备本地计算能力，并能够与其邻居智能体集合 $\mathcal{N}_i \triangleq \{i' \in \mathcal{N}, \text{中的智能体进行通信，该集合定义为：} (i, i') \in \mathcal{L}\}$。为便于表述，我们假设每个智能体 i 都有 n 个数据样本，因此总共有 mn 个数据样本[1]。在去中心化极小极大优化中，网络中的智能体以分布式、协作的方式求解以下去中心化极小极大优化问题：

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \Big[\frac{1}{m}\sum_{i=1}^{m} F_i(\mathbf{x},\mathbf{y}) + h(\mathbf{x})\Big], \tag{1}$$

其中，$\mathbf{x} \in \mathcal{X}$ 和 $\mathbf{y} \in \mathcal{Y}$ 分别是外层极小问题和内层极大问题中待训练的参数，集合 $\mathcal{X} \subseteq \mathbb{R}^{p_1}$ 和 $\mathcal{Y} \subseteq \mathbb{R}^{p_2}$ 是闭凸集，$F_i(\mathbf{x},\mathbf{y}) \triangleq \frac{1}{n}\sum_{j=1}^{n} f_{ij}(\mathbf{x}_i,\mathbf{y}_i|\boldsymbol{\xi}_{ij})$ 表示局部目标函数，而 $h(\mathbf{x}_i)$ 是一个适当凸函数（可能不可微），通常起正则化作用。这里，$F_i(\mathbf{x},\mathbf{y})$ 仅节点 i 可观测，并假设对于固定的 $\mathbf{x}$，它关于 $\mathbf{y}$ 是强凹的，而对于固定的 $\mathbf{x}$，它关于 $\mathbf{y}$ 是非凸的。为求解以去中心化方式求解问题 (1) 的一种常见方法是将其改写为以下等价形式：

$$\min_{\{\mathbf{x}_i\in\mathcal{X},\forall i\}} \max_{\{\mathbf{y}_i\in\mathcal{Y},\forall i\}} \left[\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}(\mathbf{x}_i,\mathbf{y}_i|\boldsymbol{\xi}_{ij}) + h(\mathbf{x}_i)\right],$$
$$\text{subject to} \quad \mathbf{x}_i = \mathbf{x}_{i'}, \mathbf{y}_i = \mathbf{y}_{i'}, \ \forall (i,i') \in \mathcal{L}, \tag{2}$$

其中，$\mathbf{x}_i$ 和 $\mathbf{y}_i$ 分别是智能体 i 处原始参数 $\mathbf{x}$ 和 $\mathbf{y}$ 的局部副本。(2) 中的等式约束确保所有智能体处的局部副本彼此相等，因此称为“共识形式”。显然，问题 (1) 和 (2) 具有相同的解。在本文余下部分，我们将重点求解问题 (2)，并将其称为去中心化非凸-强凹（NCX-SCV）共识极小极大优化问题。去中心化共识极小极大优化的目标是设计一种算法，以获得满足以下条件的集体 ϵ-驻点 $\{\mathbf{x}_i,\mathbf{y}_i,\forall i\}$：

$$\underbrace{\frac{1}{m}\sum_{i=1}^{m}\|\mathbf{x}_i - \overline{\mathbf{x}}\|^2}_{\substack{\text{Outer consensus}\\ \text{error}}} + \underbrace{\frac{1}{m}\sum_{i=1}^{m}\|\mathbf{y}_i - \overline{\mathbf{y}}\|^2}_{\substack{\text{Inner consensus}\\ \text{error}}}$$
$$+ \underbrace{\mathbb{E}\|\mathbf{y}^* - \bar{\mathbf{y}}\|^2}_{\substack{\text{Saddle point}\\ \text{error}}} + \underbrace{\|\frac{1}{m}\sum_{i=1}^{m}\nabla_{\mathbf{x}} F_i(\mathbf{x},\mathbf{y})\|^2}_{\text{Global gradient magnitude}} \le \epsilon^2,$$

其中，$\bar{\mathbf{x}} \triangleq \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i$、$\bar{\mathbf{y}} \triangleq \frac{1}{m}\sum_{i=1}^{m}\mathbf{y}_i$和$\mathbf{y}^*$ 分别表示在 $\mathbf{y}$上，F 的最大化点，其中$\mathbf{y}^*(\bar{\mathbf{x}}) \in \arg\max_{\mathbf{y}\in\mathcal{Y}} F(\bar{\mathbf{x}},\mathbf{y})$,

如第1节所述，去中心化优化中最重要的两项性能指标是样本复杂度和通信复杂度。在本文中，我们采用去中心化优化文献中广泛使用的样本复杂度和通信复杂度的两种定义（例如 Sun 等人 [2020]），以衡量我们的算法效率：

定义 1（样本复杂度）。样本复杂度定义为：在算法收敛到 ϵ-驻点之前，所有节点所需调用增量一阶预言机（*IFO*）的总次数，其中一次 *IFO* 调用在节点 i 处计算一对梯度（$\nabla_{\mathbf{x}} f_{ij}(\mathbf{x},\mathbf{y})$ $\nabla_{\mathbf{y}} f_{ij}(\mathbf{x},\mathbf{y})$）。

定义 2（通信复杂度）。将一个通信轮次定义为一个时间窗口，在此期间，每个节点向其相邻节点发送一个向量，同时接收来自所有相邻节点的一组向量。通信复杂度则定义为算法收敛至一个 ϵ-驻点所需的通信轮次总数。

[1]我们注意到，只需采用更复杂的记号，所有证明和结果在各本地数据集大小不相等的情况下仍然成立。

2）应用示例的动机：基于去中心化约束 NCX-SCV 极小极大优化的基础知识，我们提供两个示例，进一步说明其实际相关性：

• 多智能体公平机器学习：考虑一个机器学习任务，其数据集 $\{b_{ij}, [\tilde{\boldsymbol{\xi}}_{ij}^\top, \boldsymbol{\xi}_{ij}^{*\top}]^\top\}$ 分布在一个多智能体网络上，其中 b_{ij} 是第i个智能体的第 j个样本的观测标签，$\tilde{\boldsymbol{\xi}}_{ij} \in \mathbb{R}^{d_1}$ 表示相应的非敏感特征，$\boldsymbol{\xi}_{ij} \in \mathbb{R}^{d_2}$ 表示敏感特征。在公平机器学习问题中，通过添加正则化项来惩罚学习模型输出 b_{ij} 与敏感属性 $\boldsymbol{\xi}_{ij}^*$之间的统计相关性，从而施加公平性约束。在二元情形中，一个例子是将 Rényi 相关性 Baharlouei 等人[2019] 用作正则化项以实现公平性；在此设定下，多智能体公平机器学习问题可写成去中心化 NCX-SCV 极小极大问题 Baharlouei 等人 [2019]:
$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \frac{1}{c}\sum_{i=1}^{c} \mathbb{E}_i[\mathbb{L}(F_i(\mathbf{x},\mathbf{y};\boldsymbol{\xi}_i), b_i)] + \lambda_l \mathbf{f}(\mathbf{x}, \mathbf{y}) - \lambda_l \sum_{j=1}^{c} y_{ij}\tilde{S}\mathbf{f}_{ij}(\mathbf{x}, \boldsymbol{\xi}_i)]$，其中 $\tilde{S} = 2S - 1, S = \{0, 1\}$ 表示敏感属性，$\mathbb{L}$ 是损失函数，λ_l是用于平衡公平性与拟合优度的正标量，c 是类别标签，而 $\mathbf{f}_{ij}(\mathbf{x}, \boldsymbol{\xi}_i)$ 表示神经网络经过 softmax 层后的向量值输出。

• 数据投毒攻击：考虑一个去中心化学习问题，其中 m 个智能体试图学习一个共同模型。攻击者能够向部分智能体的训练样本中注入噪声。令 $\mathbf{y}_i$ 表示模型参数，令 $\mathbf{x}_i$ 表示注入的投毒数据参数。在此问题中，攻击者试图最大化损失函数，而其他智能体则旨在最小化损失函数。因此，数据投毒攻击问题具有如下 NCX-SCV 极小极大问题：$\max_{\mathbf{x}\in\mathcal{X}} \min_{\mathbf{y}\in\mathcal{Y}} \sum_{i=1}^{m} \frac{1}{|\boldsymbol{\xi}_i|}\sum_{\ell\in\boldsymbol{\xi}_i} \log\left(1 + \exp\left(\left(-v_\ell \mathbf{y}_i^T(w_\ell + \mathbf{x}_i)\right)\right)\right)$，其中 $v_\ell \in \mathbb{R}$ 和 $w_\ell \in \mathbb{R}^d$ 分别表示第 ℓ 个数据点的标签和特征向量。

2.2 相关工作

1）集中式 NCX-SCV 极小极大优化：在文献中，集中式设置下求解 NCX-SCV 优化问题的最先进算法包括 GDA（Lin 等人 [2020a], ）、min-max-PPA（Lin 等人 [2020b], ）和 SREDA（Luo 等人 [2020]）。具体而言，Lin 等人 [2020a] 提出了一种基于梯度的 GDA 方法，用于寻找一阶纳什均衡点。在每次迭代中，GDA 对$\mathbf{x}$变量执行梯度下降，对$\mathbf{y}$变量执行梯度上升。对于 NCX-SCV 极小极大优化问题，GDA 的收敛速率为 $\mathcal{O}(1/T)$，其中 T 为最大迭代次数。此外，该方法每次迭代都需要进行一次全梯度计算，这意味着要达到 ϵ 收敛误差，样本复杂度为 $\mathcal{O}(n\epsilon^{-2})$。Lin 等人 [2020b] 提出了 Minimax-PPA 方法来求解 NCX-NCV 问题，其样本复杂度为 $\tilde{\mathcal{O}}\left(n\varepsilon^{-2}\right)$。当 n较大、处于大数据场景时，这些方法的样本复杂度较高。为解决这一问题，人们还提出了几种方差缩减方法。例如，Luo 等人 [2020],提出了一种名为 SREDA 的方差缩减算法，Xu 等人 [2020] 进一步对其进行了改进，使其能够采用更大的步长。对于较大的 n，SREDA 的样本复杂度为 $\tilde{\mathcal{O}}\left(n + \sqrt{n}\epsilon^{-2}\right)$，因此其样本复杂度低于 GDA 和 min-max-PPA。然而，SREDA 只能处理对$\mathbf{x}$ 施加约束、但不对$\mathbf{y}$施加约束的极小极大问题。我们在表 1 中总结了上述比较。尽管上述算法在求解 NCX-SCV 极小极大问题方面取得了不同程度的成功，但它们都是针对集中式设置开发的，这与我们的工作有本质区别。

2）去中心化极小极大优化：如第1节所述，关于去中心化极小极大优化的现有研究结果相当有限。最早的尝试是 Mateos-Núnez 和 Cortés [2015],提出的 CSPSG 方法，该方法考虑了最理想的凸-凹（CX-CV）设置。由于采用了简单的 SGD 型更新，CSPSG 的样本复杂度和通信复杂度较高，达到 $\mathcal{O}(\epsilon^{-4})$。DPOSG Liu 等人 [2020] 在大规模 GAN 的背景下研究了无结构、非凸-非凹（NCX-NCV）、无约束的去中心化极小极大问题，并提出采用经典的 DSGD 方法 Nedic 和 Ozdaglar[2009] ，将名为 OGDA 的集中式对应算法 Mokhtari 等人 [2020]去中心化。由于 DSGD 固有的局限性，DPOSG 的样本复杂度较高，达到 $\mathcal{O}(\epsilon^{-12})$。相比之下，DPPSP Liu 等人[2019] 也研究了带约束的无结构 NCX-NCV 去中心化极小极大优化问题。由于采用了基础的近端 SGD 型更新，DPPSP 的样本复杂度和通信复杂度也较高，达到 $\mathcal{O}(\epsilon^{-4})$。

相比于 Liu 等人 [2019, 2020],较为简单的算法技术，我们的 PRECISION 算法是一种三重混合算法，融合了近端算子、方差缩减和梯度跟踪，因此大幅降低了样本复杂度和通信复杂度。需要指出的是，尽管我们是在结构更丰富的 NCX-SCV 场景下实现了显著更低的样本复杂度和通信复杂度，但我们认为，我们的方法也可应用于 NCX-NCV，以改善现有工作的样本复杂度和通信复杂度。这将留待未来工作探讨。

与我们最相关的工作是 Saknakis 等人的 GT-GDA[2020], , 该工作同样研究了受约束的去中心化 NCX-SCV 极小极大优化。GT-GDA 与我们工作的关键区别在于，它只对 $\mathbf{x}$ 或$\mathbf{y}$中的一个施加约束集，而不是同时对二者施加约束。相比之下，我们考虑

表 1：NCX-SCV 极小极大问题算法对比（m 为智能体数量， n 为每个智能体的数据集大小， ϵ为收敛误差。我们提出的算法以粗体标出。

算法 *	近端算子	样本复杂度。	通信复杂度。	去中心化-式
GDA Lin 等人 [2020a]	$\mathbf{y}$	$\tilde{\mathcal{O}}\left(n\varepsilon^{-2}\right)$	-	✗
Minmax-PPA Lin 等人 [2020b]	$\mathbf{x}$ 和$\mathbf{y}$	$\tilde{\mathcal{O}}\left(n\varepsilon^{-2}\right)$	-	✗
SREDA Luo 等人 [2020]	$\mathbf{x}$	$\tilde{\mathcal{O}}\left(n+\sqrt{n}\varepsilon^{-2}\right)$	-	✗
PRECISION PRECISION$^+$	$\mathbf{x}$ 和$\mathbf{y}$	$\mathcal{O}(m\sqrt{n}\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-2})$	✓

表 2：去中心化极小极大问题算法比较。

算法 *	近端算子	样本复杂度。	通信复杂度。	问题
DPOSG Liu 等人 [2020]	-	$\mathcal{O}(\epsilon^{-12})$	$\mathcal{O}(\log(1/\epsilon))$	NCX-NCV
CSPSG Mateos-Núnez 和 Cortés [2015]	$\mathbf{x}$和 $\mathbf{y}$	$\mathcal{O}(\epsilon^{-4})$	$\mathcal{O}(\epsilon^{-4})$	CX-CV
DPPSP Liu 等人 [2019]	$\mathbf{x}$和 $\mathbf{y}$	$\mathcal{O}(\epsilon^{-4})$	$\mathcal{O}(\epsilon^{-4})$	NCX-NCV
GT-GDA Tsaknakis 等人 [2020]	$\mathbf{x}$或$\mathbf{y}$	$\mathcal{O}(mn\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-2})$	NCX-SCV
PRECISION PRECISION$^+$	$\mathbf{x}$和 $\mathbf{y}$	$\mathcal{O}(m\sqrt{n}\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-2})$	NCX-SCV

两者均受约束的更复杂情形。GT-GDA 还要求对$\mathbf{y}$ 进行多次内层更新，然后对$\mathbf{x}$执行一次更新，这类似于 Boyd 等人 [2011] 提出的交替方向乘子法（ADMM）更新方案。此外，与 GT-GDA 中的 $\mathcal{O}(mn\epsilon^{-2})$ 相比，我们的算法实现了更低的样本复杂度 $\mathcal{O}(m\sqrt{n}\epsilon^{-2})$。最后，我们在表 2 中总结上述比较。另一项密切相关的工作是 Zhang 等人 [2021], 的研究，其中作者针对多智能体强化学习策略评估问题，基于均方投影贝尔曼误差（MSPBE）开发了一种去中心化优化方法；该问题可以表述为有限和极小极大问题。然而，我们的工作与 Zhang 等人 [2021]存在以下差异：（i）不同于 Zhang 等人 [2021], ，我们的方法能够处理非光滑目标函数。然而，将 Zhang 等人 [2021] 中的算法直接扩展为近端形式，在求解去中心化问题时可能会发散 Hong 等人 [2022]。为应对这一挑战，我们提出了一种专用的近端算子 $\tilde{\mathbf{x}}_i(\mathbf{x}_{i,t})$，详见备注 1 中的讨论；（ii）我们的方法适用于一般的去中心化极小极大优化问题，而 Zhang 等人 [2021]的方法仅限于强化学习策略评估。

3 求解方法

本节首先分别在第 3.1 节和第 3.2 节介绍我们的 PRECISION 和 PRECISION$^+$ 算法。随后，我们在第 3.3 节给出 PRECISION 和 PRECISION$^+$ 算法的主要理论结果及关键见解。由于篇幅有限且为提高可读性，我们将部分理论结果的证明细节放在附录中。

3.1 PRECISION算法

为求解问题 (2) 中去中心化极小极大问题的共识形式，我们采用文献 Nedic 和 Ozdaglar [2009] 中的网络共识混合方法。为此，我们令 $\mathbf{M} \in \mathbb{R}^{m\times m}$ 表示共识权重矩阵，并令 $[\mathbf{M}]_{ii'}$ 表示 $\mathbf{M}$ 中第 i 行、第 i' 列的元素。$\mathbf{M}$ 满足以下性质 Nedic 和 Ozdaglar [2009],Wai 等人 [2018]:

(a) 双随机： $\sum_{i=1}^{m}[\mathbf{M}]_{ii'} = \sum_{i'=1}^{m}[\mathbf{M}]_{ii'} = 1$; (b) 对称： $[\mathbf{M}]_{ii'} = [\mathbf{M}]_{i'i}, \quad \forall i, i' \in \mathcal{N}$; (c) 网络定义的稀疏性： $[\mathbf{M}]_{ii'} > 0$ 若 $(i, i') \in \mathcal{L}$; 否则 $[\mathbf{M}]_{ii'} = 0, \quad \forall i, i' \in \mathcal{N}$。

请注意，上述性质意味着$\mathbf{M}$的特征值均为实数，且可按 $-1<\lambda_m(\mathbf{M}) \leq \cdots \leq \lambda_2(\mathbf{M}) < \lambda_1(\mathbf{M}) = 1$的顺序排列。为方便记号，我们将$\mathbf{M}$的绝对值第二大的特征值定义为 $\lambda \triangleq \max\{|\lambda_2(\mathbf{M})|, \ldots, |\lambda_m(\mathbf{M})|\}$，它将在算法收敛速率的步长选择与分析中发挥重要作用。采用上述记号后，我们现在可以介绍所提出的算法。

如第1节所述，我们的 PRECISION 算法可以视为近端、梯度跟踪和方差缩减技术的三重混合。接下来，我们将看到，这些技术可以组织成三个关键的算法步骤：

• 步骤 *1*（局部近端操作）：在每次迭代 t中，每个智能体 i 首先执行以下近端操作，以分别处理外层变量和内层变量的约束集 $\mathcal{X}$ 和 $\mathcal{Y}$：

$$\begin{aligned}\tilde{\mathbf{x}}_i(\mathbf{x}_{i,t}) =& \arg\min_{\mathbf{x}_i\in\mathcal{X}} \langle \mathbf{p}_{i,t}, \mathbf{x}_i - \mathbf{x}_{i,t}\rangle \\ &+ \frac{\tau}{2}\|\mathbf{x}_i - \mathbf{x}_{i,t}\|^2 + h(\mathbf{x}_i), \end{aligned} \tag{3}$$

$$\tilde{\mathbf{y}}_i(\mathbf{y}_{i,t}) = \arg\min_{\mathbf{y}_i\in\mathcal{Y}} \big\|\mathbf{y}_i - \big(\mathbf{y}_{i,t} + \alpha \mathbf{d}_{i,t}\big)\big\|^2, \tag{4}$$

其中，$\mathbf{p}_{i,t}$ 和 $\mathbf{d}_{i,t}$ 是用于梯度跟踪的两个辅助向量，稍后将予以定义；$\tau > 0$ 是一个常数近端控制参数，$\alpha > 0$ 是用于控制 $\mathbf{y}$更新幅度的常数参数。

• 步骤 *2*（共识更新）：接下来，每个智能体 i 更新外层和内层模型参数 $\mathbf{x}_i$,$\mathbf{y}_i$:

$$\mathbf{x}_{i,t+1} = \underbrace{\sum_{i'\in\mathcal{N}_i} [\mathbf{M}]_{ii'}\mathbf{x}_{i',t}}_{\text{(a)}} + \underbrace{\nu\,(\tilde{\mathbf{x}}_i(\mathbf{x}_{i,t}) - \mathbf{x}_{i,t})}_{\text{(b)}}, \tag{5}$$

$$\mathbf{y}_{i,t+1} = \underbrace{\sum_{i'\in\mathcal{N}_i} [\mathbf{M}]_{ii'}\mathbf{y}_{i',t}}_{\text{(a)}} + \underbrace{\eta(\tilde{\mathbf{y}}_i(\mathbf{y}_{i,t}) - \mathbf{y}_{i,t})}_{\text{(b)}}, \tag{6}$$

其中，ν 和 η 分别是更新 $\mathbf{x}$变量和 $\mathbf{y}$变量的步长。注意，在式（5）和（6）中，分量（a）是智能体 i 处的局部加权平均，也称为“共识步骤”；分量（b）则基于近端点 $\tilde{\mathbf{x}}$ 和 $\tilde{\mathbf{y}}$，按照 Frank-Wolfe 的思路执行局部更新，这不同于传统的去中心化随机梯度更新 Nedi´c 等人 [2018]。

• 步骤 *3*（局部梯度估计）：在下一步中，每个智能体 i 使用以下方法估计其局部梯度梯度估计器：

$$\mathbf{v}_{i,t} = \begin{cases} \nabla_{\mathbf{x}} F_i(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}), & \text{if } \mathrm{mod}(t,q) = 0, \\ \mathbf{v}_{i,t-1} + \frac{1}{|\mathcal{S}_{i,t}|}\sum_{j\in\mathcal{S}_{i,t}}\big(\nabla_{\mathbf{x}} f_{ij}(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}) \\ \qquad -\nabla_{\mathbf{x}} f_{ij}(\mathbf{x}_{i,t-1}, \mathbf{y}_{i,t-1})\big), & \text{o.w.} \end{cases} \tag{7a}$$

$$\mathbf{u}_{i,t} = \begin{cases} \nabla_{\mathbf{y}} F_i(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}), & \text{if } \mathrm{mod}(t,q) = 0, \\ \mathbf{u}_{i,t-1} + \frac{1}{|\mathcal{S}_{i,t}|}\sum_{j\in\mathcal{S}_{i,t}}\big(\nabla_{\mathbf{y}} f_{ij}(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}) \\ \qquad -\nabla_{\mathbf{y}} f_{ij}(\mathbf{x}_{i,t-1}, \mathbf{y}_{i,t-1})\big), & \text{o.w.} \end{cases} \tag{7b}$$

这里，$\mathcal{S}_{i,\,t}$ 是第 t 次迭代中的样本小批量，而 q 是预设的内循环迭代次数。

• 步骤 *4*（梯度跟踪）：每个智能体 i 通过对其邻居的跟踪梯度取平均来更新$\mathbf{p}_i$ 和$\mathbf{d}_i$：

$$\begin{cases} \mathbf{p}_{i,t} = \sum_{i'\in\mathcal{N}_i} [\mathbf{M}]_{ii'}\mathbf{p}_{i',t-1} + \mathbf{v}_{i,t} - \mathbf{v}_{i,t-1}, \\ \mathbf{d}_{i,t} = \sum_{i'\in\mathcal{N}_i} [\mathbf{M}]_{ii'}\mathbf{d}_{i',t-1} + \mathbf{u}_{i,t} - \mathbf{u}_{i,t-1}. \end{cases} \tag{8}$$

我们可以直观地理解 PRECISION 算法如下：在 PRECISION 中，每个智能体都执行下降步骤和上升步骤，因为问题 (2) 对 x 求最小值、对 y 求最大值。请注意，式 (7) 中的 $\mathbf{v}_{i,t}$ 和 $\mathbf{u}_{i,t}$ 仅包含局部目标函数 F_i (x,y) 的梯度信息。仅沿方向 $\mathbf{v}_{i,t}$ 和 $\mathbf{u}_{i,t}$ 更新，无法保证全局目标函数 F(x, y) 收敛。因此，我们引入两个辅助变量 $\mathbf{p}_{i,t}$ 和 $\mathbf{d}_{i,t}$，用于全局梯度跟踪。每个智能体 i 按照式 (8) 所示执行局部加权聚合来更新这两个变量，因此 $\mathbf{p}_{i,t}$ 和 $\mathbf{d}_{i,t}$ 会跟踪全局梯度的方向。

将 PRECISION 与我们最相关的工作 Tsaknakis 等人 [2020]提出的 GT-GDA 方法进行比较颇有启发。在 GT-GDA 中，智能体 i 在第t次迭代计算局部全梯度如下：

$$\mathbf{v}_{i,t} = \nabla_{\mathbf{x}} F_i(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}), \quad \mathbf{u}_{i,t} = \nabla_{\mathbf{y}} F_i(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}). \tag{9}$$

与 GT-GDA Tsaknakis 等人不同，PRECISION 在智能体处估计式 (7) 中的局部梯度。式 (7) 中，算法仅每隔 q 步计算一次全梯度 ∇F_i(x$_{i,t}$,y$_{i,t}$)。对于满足 mod(t, q) $\neq$ 0 的其他迭代，PRECISION 使用由小批量 $|\mathcal{S}_i^1{}_{,t}|\sum_{j\in\mathcal{S}_{i,t}}\nabla_{\mathbf{y}} f_{ij}$(x$_{i,t}$,y$_{i,t}$) 估计的局部随机梯度，以及递归校正项 $\mathbf{u}_{i,t-1} - {}_{|\mathcal{S}_i^1{}_{,t}|}\sum_{j\in\mathcal{S}_{i,t}}\nabla_{\mathbf{y}} f_{ij}$(x$_{i,t-1}$, y$_{i,t-1}$)。得益于周期性计算的全梯度和递归校正项，PRECISION 能够达到 $\mathcal{O}(1/T)$ 的收敛速率。此外，由于对 $\mathcal{S}_{i,t}$ 进行随机子采样，PRECISION 的样本复杂度低于 GT-GDA Tsaknakis 等人的方法。PRECISION 的完整描述见算法 1。

算法 1：智能体 i上的 PRECISION/PRECISION$^+$ 。

若 PRECISION：$|\mathcal{R}_{i,\ t}|=n$; 若 PRECISION$^+$:

$$|\mathcal{R}_{i,t}| = \min\{c_\gamma \sigma^2 (\gamma_t)^{-1}, c_\epsilon \sigma^2 \epsilon^{-1}, n\}.$$

1：设置原始-对偶参数对 $(\mathbf{x}_{i,\ 0},\ \mathbf{y}_{i,\ 0}) = (\mathbf{x}^0, \mathbf{y}^0)$。2：有放回地抽取 $\mathcal{R}_{i,\ 0}$ 个样本，并计算局部随机梯度估计量，如下所示

$$\mathbf{p}_{i,0} = \mathbf{v}_{i,0} = \frac{1}{|\mathcal{R}_{i,0}|} \sum_{j \in \mathcal{R}_{i,0}} \nabla_{\mathbf{x}} f_{ij}(\mathbf{x}_{i,0}, \mathbf{y}_{i,0});$$

$$\mathbf{d}_{i,0} = \mathbf{u}_{i,0} = \frac{1}{|\mathcal{R}_{i,0}|} \sum_{j \in \mathcal{R}_{i,0}} \nabla_{\mathbf{y}} f_{ij}(\mathbf{x}_{i,0}, \mathbf{y}_{i,0});$$

3: **对于** $t = 1, \cdots, T$，**执行**

4：更新局部参数 $(\mathbf{x}_{i,\ t+1},\ \mathbf{y}_{i,\ t+1})$，如公式（3）-（6）所示；5：计算局部估计量 $(\mathbf{v}_{i,\ t+1},\ \mathbf{u}_{i,\ t+1})$，如公式（7）所示；6：跟踪全局梯度 $(\mathbf{p}_{i,\ t+1},\ \mathbf{d}_{i,\ t+1})$，如公式（8）所示；7：**结束循环**

3.2 PRECISION$^+$ 算法

请注意，在 PRECISION 中，每隔 q 步就需要计算一次全梯度，这在某些情况下仍可能导致较高的计算成本。此外，在 PRECISION 的初始化阶段（主循环开始之前），智能体需要计算全梯度，这可能非常耗时。为应对这些挑战，我们通过采用自适应批量大小技术对 PRECISION 进行了改进，并将改进后的版本称为 PRECISION $^+$。具体而言，我们在迭代 t 中，将式（7a）和（7b）中的梯度估计量修改为 $\mathrm{mod}(t,\ q) = 0$，如下所示：

$$\mathbf{v}_{i,t} = \frac{1}{|\mathcal{R}_{i,t}|} \sum_{j \in \mathcal{R}_{i,t}} \nabla_{\mathbf{x}} f_{ij}(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}), \tag{10}$$

$$\mathbf{u}_{i,t} = \frac{1}{|\mathcal{R}_{i,t}|} \sum_{j \in \mathcal{R}_{i,t}} \nabla_{\mathbf{y}} f_{ij}(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}), \tag{11}$$

其中 $\mathcal{R}_{i,\ t}$ 为子采样集（无放回抽样），其大小选为

$$|\mathcal{R}_{i,t}| = \min\{c_\gamma \sigma^2 (\gamma_t)^{-1}, c_\epsilon \sigma^2 \epsilon^{-1}, n\}. \tag{12}$$

这里，c_γ 和 c_ϵ 是稍后定义的问题相关常数，σ^2 是智能体之间数据异质性的方差界（也将在稍后定义），γ_t而 $\triangleq$ $\frac{1}{q}\sum_{i=\ n_t-1\ q}^{t} \|\tilde{}_t - 1 \otimes \bar{}_t\|^2 \quad \otimes \quad ()\mathbf{xx},$ $-$

其中 表示克罗内克积算子。

$|\mathcal{R}_{i,t}|$ 的选择基于这样一个事实：周期性全梯度评估只在收敛过程的后期发挥重要作用：在收敛过程的后期，我们需要更精确的更新方向。稍后我们将看到，在一些温和的假设和参数设置下，PRECISION$^+$ 具有与 PRECISION 相同的收敛速率。PRECISION$^+$ 算法的完整描述也见算法 1。

3.3 PRECISION 和PRECISION$^+$ 算法的理论结果

在介绍我们算法的理论结果之前，我们先给出以下假设：

假设 *1*（全局目标函数）。函数 $F(x, y) = \frac{1}{m}\sum_{i=1}^{m}[F_i(x_i, y_i)]$ 和 $J(x) = max_{\mathbf{y}\in\mathcal{Y}} F(x, y)$ 满足：*(a)*（下界有界）：存在一个有限下界 $Q^* = Q(x^*) = inf_x(J(x) + h(x)) > -\infty$;

(b)（关于 y 强凹）：对于固定的 x，局部目标函数$F_i(x\ \cdot)$ 是 μ-强凹的，即存在一个正数常量 μ，使得 $\|\nabla_{\mathbf{y}} F_i(x, y) - \nabla_{\mathbf{y}} F_i(x, y')\| \geq \mu\|y - y'\|, \forall\ x, y, y' \in \mathbb{R}^{p_2}$, $i \in [m]$。*(c)*（最大值处梯度有界）：每个 $(x\ \nabla_{\mathbf{x}} F(x, y^*(x)))$ 处的偏梯度均有界，即$\|\nabla_{\mathbf{x}} F(x, y^*(x))\| < \infty, \forall\ x \in \mathbb{R}^{p_1}$。

假设 1(a) 和 1(b) 在文献中是标准假设。假设 1(c) 保证 $\nabla J(\mathbf{x}) = \nabla_{\mathbf{x}} F(\mathbf{x}、\ \mathbf{y}^*(\mathbf{x}))$。

假设 *2*（局部目标函数的利普希茨光滑性）。函数 $f_{ij}(x\ \cdot)$ 是 L_f-利普希茨光滑的，即存在一个常数 $L_f > 0$，使得 $\nabla f_{ij}(x, y) = [\nabla_{\mathbf{x}} f_{ij}(x, y]^\top, \nabla_{\mathbf{y}} f_{ij}(x, y)^\top)^\top$ 满足 $\|\nabla f_{ij}(x, y) - \nabla f_{ij}\ '\ '\|^2 \leq L_f^2\| - '\|^2 + L_f^2\| - '\|^2\ \forall\ ,\ ' \in (x, y)$ $\mathcal{X}\ ,\ ' \in \mathcal{Y}\ i \in [m]\ j \in [n]$ $x\ x\ y\ y\ ,\ x\ x\ ,\ y\ y\ ,\ ,$。

此外，我们仅对算法 PRECISION$^+$ 作出以下假设：

假设3（方差有界）。存在常数$\sigma^2 > 0$，使得 $\mathbb{E}\|\nabla f_{ij}$（$\mathbf{x},\mathbf{y}$）$- \nabla F_i$（$\mathbf{x}, \mathbf{y}$）$\|^2 \leq \sigma^2, \forall\, \mathbf{x},\mathbf{y} \in \mathbb{R}^p, i \in [m], j \in [n]$。

为应对刻画 NCX-SCV 去中心化约束极小极大问题收敛速率所面临的挑战，我们提出以下新 度量，它是本文建立所有收敛结果的关键：

$$\mathfrak{M}_t \triangleq \mathbb{E}[\|\tilde{\mathbf{x}}_t - 1 \otimes \bar{\mathbf{x}}_t\|^2 + \|\mathbf{x}_t - 1 \otimes \bar{\mathbf{x}}_t\|^2 + \|\mathbf{y}_t - 1 \otimes \bar{\mathbf{y}}_t\|^2 + \|\mathbf{y}_t^* - \bar{\mathbf{y}}_t\|^2], \tag{13}$$

其中，$\mathbf{y}_t^*$表示$\mathbf{y}^*(\bar{\mathbf{x}}_t) = \arg\max_{y\in\mathbb{R}^p} F(\bar{\mathbf{x}}_t, \mathbf{y})$。(13)中的前两项受到 SONATA Scutari 和 Sun [2019], 中度量的启发，该度量用于衡量非凸去中心化极小化问题（而非极小极大问题）的收敛进展。(13)中的第三项衡量$\mathbf{y}$上局部副本的一致性误差。第四项量化 $\bar{\mathbf{y}}_t$ 收敛到$\mathbf{y}_t^*$点的程度，该点对应于 $F(\bar{\mathbf{x}}_t, \cdot)$。因此，随着 $\mathfrak{M}_t \to 0$，算法会在原始去中心化约束极小极大优化问题的一个一阶驻点（FOSP）上达成一致。

采用式 (13) 中的度量，算法 PRECISION / PRECISION$^+$ 的收敛速率可表征如下：

定理 1 (PRECISION的收敛)。在假设 *1 (a)-(d)* 和假设 *2* 下，假设 $\beta \le$最小 $\left\{\frac{\tau}{12}, \frac{1}{3}\right\}$，

$\alpha \le \frac{1}{4L_f}$,$q = |\mathcal{S}_{i,t}| = \lceil\sqrt{n}\rceil$ 成立，并令 $c_1 = \frac{1-\lambda^2}{1+\lambda^2}$，若步长满足：$\eta \le$最小 $\left\{\frac{1}{8}, \frac{c_1 m\mu}{375\alpha L_f^2}, \frac{15L_f^2}{\beta\mu\alpha^2 c_1}, \frac{3c_1^2 m}{10(1+c_1)\mu\alpha}\right\}$, $\nu \le$最小 $\left\{\frac{c_1 m\beta}{40L_f^2}, \frac{2c_1 m\beta}{5\tau}, \frac{2c_1\beta\mu^2 m}{375L_f^4}, \frac{5\tau}{3mc_1}, 6m\frac{\tau}{(1+1/c_1)}, \frac{3\mu\eta\alpha\tau}{17L_f^2}, 3\frac{\tau}{(L_f+L_f^2/\mu)}\right\}$，则算法 PRECISION 的以下收敛结果成立：

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\mathfrak{M}_t] \le \frac{\mathbb{E}[\mathfrak{p}_0 - Q^*]}{\min\{C_1, C_2, C_3, \nu L_f^2/2\}(T+1)},$$

其中 $Q^* = Q(\mathbf{x}^*)$ 和 $\mathfrak{p}_t$ 是定义如下的势函数：

$$\mathfrak{p}_t \triangleq Q(\bar{\mathbf{x}}_t) + \frac{4\nu L_f^2}{\beta\mu\eta^2}\|\bar{\mathbf{y}}_t - \mathbf{y}_t^*\|^2 + \frac{1}{m}\sum_{i=1}^{m}[\|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 + \|\mathbf{y}_{i,t} - \bar{\mathbf{y}}_t\|^2], \quad (14)$$

其中 C_1、C_2、$C_3 \ge 0$ 为常数。由于篇幅限制，这些常数的详细定义列于附录中。此外，在（*14*）中，$Q(\mathbf{x}_t) \triangleq \max_{\mathbf{y}} F(\mathbf{x}_t, \mathbf{y}) + h(\mathbf{x}_t)$，且 $\mathbf{y}_t^* = \arg\max_{\mathbf{y}} F(\mathbf{x}_t, \mathbf{y})$。

定理 2（PRECISION$^+$的收敛）。在假设 *1*（*a*）–（*d*）、假设 *2–3*，以及与定理 *1* 相同的参数设置下，若附加参数c_γ和 c_ϵ 满足以下条件：

$$c_\gamma \ge \left(\frac{75\eta\alpha}{8\mu}\frac{1}{m} + \frac{\nu}{\beta}\frac{1}{m}\right)\frac{\nu\tau}{12}, \quad c_\epsilon > 0, \quad (15)$$

以及定理 *1* 中所述的势函数，PRECISION$^+$ 满足以下收敛结果：

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\mathfrak{M}_t] \le \frac{\mathbb{E}[\mathfrak{p}_0 - Q^*]}{(T+1)\min\{C_1, C_2', C_3, \nu L_f^2/2\}} + \left(\frac{75\eta\alpha}{16\mu}\frac{2}{m} + \frac{\nu}{2\beta}\frac{2}{m}\right)\frac{\epsilon}{c_\epsilon}, \quad (16)$$

其中常数 $C_2' \ge$ 以及 C_2' 的定义见附录。

注释 1。与现有的去中心化极小极大优化工作 Tsaknakis 等人 [2020], 、Zhang 等[2021], 相比，值得注意的是，建立定理 1 和定理 2 中收敛结果的主要困难源于外层子问题中的近端算子。正如附录中的引理 3 所述，该算子使得无法使用传统的下降引理进行收敛分析。此外，不同于单智能体约束双层优化，Hong 等人 [2022]提出的算法的直接近端扩展 $\tilde{\mathbf{x}}_{i,t} = \arg\min_{\mathbf{x}\in\mathcal{X}}\|\mathbf{x} - (\mathbf{x}_{i,t} - \tau\mathbf{p}_{i,t})\|^2$) 在本文的去中心化约束极小极大问题中将发散。为应对这一挑战，我们在（3）中采用了特殊的近端更新规则。（3）中的近端算子 $\tilde{\mathbf{x}}_{i,t}$、一致性更新（5）以及相应的局部更新（5），是解决去中心化学习中所遇到的非光滑 目标函数挑战的关键。

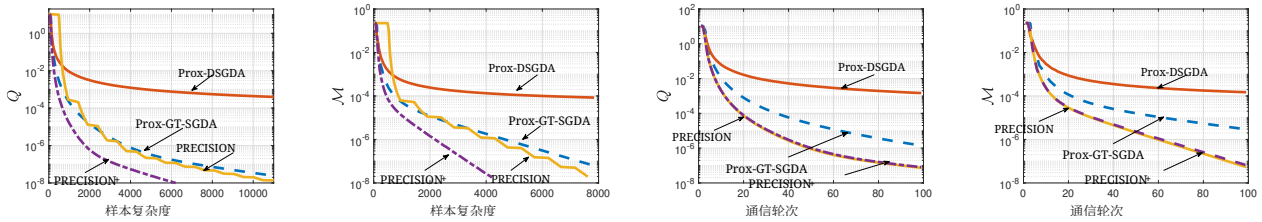
备注 2。在定理 1 和 2 中，步长和收敛速率取决于网络拓扑。对于稀疏网络，λ 接近（但不完全等于）1（请注意，$\lambda = \max\{|\lambda_2|, |\lambda_m|\} < 1$），当 λ 接近 1 时，步长需要更小，这会导致收敛速度变慢。此外，PRECISION$^+$的收敛性能会受到常数 $(\frac{75\eta\alpha}{16\mu}\frac{2}{m} + \frac{\nu}{2\beta}\frac{2}{m})\frac{\epsilon}{c_\epsilon}$的影响，该常数取决于第 t 次迭代时$\|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\| = 0$下的不精确梯度估计。直观而言，较大的 c_ϵ使我们能够采用更大的批量大小，如式 (12) 所示，从而加快收敛速度。从理论上看，较大的 c_ϵ 会使式 (16) 中的常数 $(\frac{75\eta\alpha}{16\mu}\frac{2}{m} + \frac{\nu}{2\beta}\frac{2}{m})\frac{\epsilon}{c_\epsilon}$变小，进而得到更准确的估计。

根据定理 1 和 2，我们立即得到 PRECISION 和 PRECISION$^+$ 算法的样本复杂度与通信复杂度结果：

推论 3. 在定理 *1* 和 *2* 的条件下，且满足 $q = \sqrt{n}$, 为获得一个 ϵ-驻点解，PRECISION 和 PRECISION$^+$ 算法满足以下结果：

- 通信复杂度：通信轮数总量的上界为 $\mathcal{O}(\epsilon^{-2})$
- 样本复杂度：网络中评估的样本总数的上界为 $\mathcal{O}(m\sqrt{n}\epsilon^{-2})$)。

注释 3. PRECISION/PRECISION$^+$ 算法的通信复杂度与 GT-GDA Tsak- nakis 等人 [2020], 相同，但样本复杂度低了 $\sqrt{n}$ 倍



(a) 损失函数值与样本复杂度的关系。(b) 收敛指标与样本复杂度的关系。(c) 损失函数值与通信复杂度的关系。(d) 收敛指标与通信复杂度的关系。

图1：去中心化 NCX-SCV 极小极大优化问题中各算法的比较。

低于 GT-GDA Tsaknakis 等人 [2020]。这在“大数据”场景中特别有利，因为此时 n较大（即本地数据集的规模较大）。尽管 PRECISION$^+$ 的理论复杂度界与 PRECISION 相同，但 PRECISION$^+$ 无需进行全梯度计算，这意味着 PRECISION$^+$ 在实践中使用的样本显著少于 PRECISION。下一节中的数值结果也将通过实证对此予以确认。

4 实验结果

本节中，我们使用 LIBSVM 仓库中的“a9a”数据集上的去中心化 NCX-SCV 回归问题进行数值实验，以展示所提出的 PRECISION 和 PRECISION$^+$ 算法的性能。该数据集由 Chang 和 Lin [2011]公开提供。补充材料中，我们还提供了在数据集“a9a”Chang 和 Lin [2011] 和“MNIST”LeCun等人 [1998]上进行 AUC 最大化问题的其他实验。由于目前缺少适用于同时具有外层和内层约束集的去中心化 NCX-SCV 的现有算法（详见第 2 节 2.2），我们在实验中将两种随机算法作为基线与我们的方法进行比较。这些基线可视为 PRECISION /PRECISION$^+$ 的“简化版”，即移除了梯度跟踪或方差缩减技术。受篇幅限制，详细的实验设置见附录。

1) 逻辑回归模型和数据集：我们采用以下具有数据集 $\{(\mathbf{a}_{ij}, b_{ij})\}_{j=1}^{n}$的去中心化 NCX-SCV 极小极大回归问题，其中 $\mathbf{a}_{ij} \in \mathbb{R}^d$ 是智能体 i的第 j个样本的特征，$b_{ij} \in \{1, -1\}$ 是相应的标签：

$$\min_{\mathbf{x}_i \in \mathcal{X}} \max_{\mathbf{y}_i \in \mathcal{Y}} \frac{1}{m} \sum_{i=1}^{m} F_i(\mathbf{x}_i, \mathbf{y}_i), \tag{17}$$

其中 $F_i(\mathbf{x}_i, \mathbf{y}_i)$ 定义如下：

$$F_i(\mathbf{x}_i, \mathbf{y}_i) \triangleq \frac{1}{n} \sum_{j=1}^{n} \left(y_{ij} l_{ij}(\mathbf{x}_i) - V(\mathbf{y}_i) + g(\mathbf{x}_i) \right). \tag{18}$$

在（18）中，损失函数为 $l_{ij}(\mathbf{x}_i) \triangleq \log\left(1 + \exp\left(-b_{ij}\, \mathbf{a}_{ij}^{\top} \mathbf{x}_i\right)\right)$，而 $g(\mathbf{x}_i)$ 是定义如下的非凸正则化项：$g(\mathbf{x}_i) \triangleq \lambda_2 \sum_{k=1}^{d} \frac{\alpha x_{ik}^2}{1+\alpha x_{ik}^2}$，其中 $V(\mathbf{y}_i) = \frac{1}{2}\lambda_1 \| m\mathbf{y}_i - \mathbf{1} \|_2^2$，并且我们设定约束为 $\mathcal{X} = [0, 10]^d$，$\mathcal{Y} = [0, 10]^n$。我们选择常数$\lambda_1 = 1/n^2$、$\lambda_2 = 10^{-3}$ 和 $\alpha = 10$。我们使用来自 LIBSVM 仓库的“a9a”数据集测试算法的收敛性能，该数据集由 Chang 和 Lin [2011]公开提供。

2）算法比较：由于文献中关于去中心化约束极小极大优化的研究结果非常有限，在实验中，我们采用以下算法作为基线进行性能比较：

•Prox-DSGDA（近端去中心化随机梯度下降上升）：该算法受 DSGD Nedic 和 Ozdaglar [2009],Jiang etal. [2017] 的启发。每个智能体将其本地参数更新为 $\boldsymbol{\theta}_{i,t+1} = \sum_{j \in \mathcal{N}_i} [\mathbf{M}]_{ij} \boldsymbol{\theta}_{j,t} - \gamma \frac{1}{|\mathcal{S}_i|} \sum_{j \in \mathcal{S}_i} \nabla_{\boldsymbol{\theta}} f_{ij}(\boldsymbol{\theta}_{i,t}, \boldsymbol{\omega}_{i,t})$ 和 $\boldsymbol{\omega}_{i,t+1} = \sum_{j \in \mathcal{N}_i} [\mathbf{M}]_{ij} \boldsymbol{\omega}_{j,t} - \eta \frac{1}{|\mathcal{S}_i|} \sum_{j \in \mathcal{S}_{i,t}} \nabla_{\boldsymbol{\omega}} f_{ij}(\boldsymbol{\theta}_{i,t}, \boldsymbol{\omega}_{i,t})$。

•Prox-GT-SGDA（基于近端梯度跟踪的随机梯度下降上升）：该算法的灵感来源于 Xin 等人提出的 GT-SGD 算法 [2020], Lu 等人 [2019]。GT-SGDA 与 GT-GDA 具有相同的结构，但它使用随机梯度更新 $\mathbf{v}_{i,t}$ 和 $\mathbf{u}_{i,t}$，具体如下：$\mathbf{v}_{i,t} = \frac{1}{|\mathcal{S}_i|} \sum_{j \in \mathcal{S}_{i,t}} \nabla_{\boldsymbol{\theta}} f_{ij}(\boldsymbol{\theta}_{i,t}, \boldsymbol{\omega}_{i,t})$ 和 $\mathbf{u}_{i,t} = \frac{1}{|\mathcal{S}_i|} \sum_{j \in \mathcal{S}_{i,t}} \nabla_{\boldsymbol{\omega}} f_{ij}(\boldsymbol{\theta}_{i,t}, \boldsymbol{\omega}_{i,t})$。

3）结果：从图1(a)和1(b)可以看出，就一阶预言机评估总次数而言，我们提出的PRECISION$^+$ 算法比其他算法（PRECISION、Prox-GT-SGDA和Prox-DSGDA）收敛得快得多。我们还可以观- 察到，PRECISION和PRECISION$^+$ 均具有更低的

样本复杂度，低于另外两种算- 法。如图1(c)和1(d)所示，PRECISION和PRECISION$^+$ 的通信成本远低于Prox-DSGDA和Prox-GT-SGDA。因此，我们的实- 验结果验证了我们的理论分析：PRECISION / PRECISION$^+$ 在去中心化约束极小极大优化问题中同时具有较低的样本复杂度和通- 信复杂度。

5 结论

本文研究了去中心化约束非凸-强凹（NCX-SCV）极小极大优化问题，并提出了两种算法，分别称为 PRECISION 和 PRECISION$^+$。我们证明，对于去中心化约束 NCX-SCV 极小极大问题，要达到一个 ϵ-驻点，PRECISION 和 PRECISION$^+$ 的通信复杂度为$\mathcal{O}(\epsilon^{-2})$，样本复杂度为 $\mathcal{O}(m\sqrt{n}\epsilon^{-2})$，其中 m是智能体数量，n 是每个智能体的数据集大小。数值实验也验证了所提算法的理论性能。我们指出，去中心化约束极小极大学习仍是一个研究不足的领域，而我们的工作为未来研究开辟了若干有趣的方向。例如，在我们的算法中，智能体需要向邻居发送外层和内层模型参数对，这两者的维度都可能很高。未来，我们可以尝试采用通信高效机制（例如压缩技术）来进一步降低通信成本，尤其是针对大规模深度学习模型。

参考文献

K. Ali 和 W. Van Stam。Tivo：利用分布式协同过滤架构进行节目推荐。载于第十届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集，第394–401页，2004年。S. Baharlouei、M. Nouiehed、A. Beirami 和 M. Razaviyayn。Rényi公平推断。*arXiv*预印本 *arXiv:1906.12005*，2019年。S. Boyd、N. Parikh 和 E. Chu。通过交替方向乘子法实现分布式优化和统计学习。Now Publishers Inc，2011年。D. S. Callaway 和 I. A. Hiskens。实现电力负荷可控性。*IEEE* 会刊，99(1)：184–199，2010年。Y. Cao、W. Yu、W. Ren 和 G. Chen。近年来分布式多智能体协调研究进展概述-

。 *IEEE*工业信息学汇刊，9(1)：427–438，2012年。

C.-C. Chang 和 C.-J. Lin。Libsvm：一个用于支持向量机的库。*ACM*智能系统与技术汇刊 *(TIST)*，2(3):1–27，2011年。

J. Chen、Z. J. Towfic 和 A. H. Sayed。分布式模型上的字典学习。*IEEE* 信号处理汇刊，63(4):1001–1016，2014年。

J. Cortes、S. Martinez、T. Karatas 和 F. Bullo。移动传感网络的覆盖控制。*IEEE* 机器人与自动化汇刊，20(2):243–255，2004年。

E. Dall'Anese、H. Zhu 和 G. B. Giannakis。智能微电网的分布式最优潮流。*IEEE* 智能电网汇刊，4(3):1464–1475，2013年。Y. Deldjoo、T. Di Noia 和 F. A. Merra。推荐系统中的对抗性机器学习（AML-Recsys）。载于第*13*届国际网络搜索与数据挖掘会议论文集，第869–872页，2020。D. Ernst、M. Glavic 和 L. Wehenkel。电力系统稳定性控制：强化学习框架。*IEEE* 电力系统汇刊，19(1):427–435，2004年。M. Glavic、R. Fonteneau 和 D. Ernst。电力系统决策与控制中的强化学习：过去的思考与展望。*IFAC-PapersOnLine*，50(1):6918–6927，2017年。M. Hong、S. Zeng、J. Zhang 和 H. Sun。关于去中心化非凸优化的发散性。*SIAM* 优化杂志，32(4):2879–2908，2022年。Z. Jiang、A. Balu、C. Hegde 和 S. Sarkar。固定拓扑网络中的协同深度学习。载于神经信息处理系统进展，第5904–5914页，2017年。Z. Jiang、K. Mukherjee 和 S. Sarkar。分布式优化中的共识—分歧权衡。载于*2018*年美国控制会议年会（*ACC*），第571–576页。IEEE，2018年。J. Kober、J. A. Bagnell 和 J. Peters。机器人学中的强化学习：综述。国际机器人研究杂志，32(11):1238–1274，2013年。Y. LeCun、C. Cortes 和 C. Burges。MNIST手写数字数据库。可访问：*http://yann. l ecun. com/exdb/mnist*，1998年。X. Lian、C. Zhang、H. Zhang、C.-J. Hsieh、W. Zhang 和 J. Liu。去中心化算法能否胜过中心化算法？去中心化并行

随机梯度下降的案例研究。载于神经信息处理系统进展，第5330–5340页，2017年。

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。载于国际机器学习会议，第6083–6093 页。PMLR，2020a。

T. Lin、C. Jin 和 M. I. Jordan。极小极大优化的近最优算法。载于学习理论会议论文集，第 2738–2779 页。PMLR，2020b。

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。载于神经信息处理系统进展论文集，第33 卷，2020。

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* 预印本 *arXiv:1910.14380*，2019年。S. Lu、X. Zhang、H. Sun 和 M. Hong。GNSD：一种基于梯度跟踪的非凸随机算法，用于去中心化优化。载于*2019*年 *IEEE* 数据科学研讨会（*DSW*），第315–321页。IEEE，2019年。L. Luo、H. Ye、Z. Huang 和 T. Zhang。用于随机非凸-强凹极小极大问题的随机递归梯度下降上升。*arXiv* 预印本 *arXiv:2001.03724*，2020。D. Mateos-Núnez 和 J. Cortés。用于鞍点问题的分布式次梯度方法。载于*2015*年第*54*届*IEEE*决策与控制会议（*CDC*），第5462–5467页。IEEE，2015年。A. Mokhtari、A. Ozdaglar 和 S. Pattathil。鞍点问题中额外梯度方法与乐观梯度方法的统一分析：近端点方法。载于国际人工智能与统计会议，第1497–1507页。PMLR，2020年。A. Nedic 和 A. Ozdaglar。用于多智能体优化的分布式次梯度方法。*IEEE* 自动控制汇刊，54(1):48，2009年。A. Nedi´c、A. Olshevsky 和 M. G. Rabbat。去中心化优化中的网络拓扑与通信-计算权衡。*IEEE* 会刊，106(5):953–976，2018年。Y. Nozaki 和 T. Nakamoto。结合机器学习与自然语言处理，对化学物质的气味特征进行预测建模。*PLOS ONE*，13 (6):e0198475，2018年。

P. Ogren、E. Fiorelli 和 N. E. Leonard。分布式环境中的移动传感器网络协同控制：自适应梯度爬升。*IEEE* 自动控制汇刊，49(8):1292–1302，2004年。

F. Özyurt。采用融合深度学习架构进行遥感图像识别的高效深度特征选择。《超级计算杂志》，76(11):8413–8431，2020。

A. S. Polydoros 和 L. Nalpantidis。《基于模型的强化学习综述：在机器人学中的应用》。《智能与机器人系统杂志》，86(2)：153–173，2017。

S. Qiu、Z. Yang、X. Wei、J. Ye 和 Z. Wang。用于平滑非线性时序差分学习的单时间尺度随机非凸-凹优化。*arXiv* 预印本 *arXiv:2008.10103*，2020。M. Rabbat 和 R. Nowak。传感器网络中的分布式优化。载于传感器网络信息处理国际研讨会论文集，第20–27页，2004。W. Ren、R. W. Beard 和 E. M. Atkins。多车辆协同控制中的信息共识。《*IEEE* 控制系统杂志》，27(2):71–82，2007。S. H. Rhee、H.-S. Kim 和 S.-W. Sohn。去中心化资源分配对网络中心战的影响。载于*2012*年国际信息网络会议，第478–481页。IEEE，2012。P. Sattigeri、S. C. Hoffman、V. Chenthamarakshan 和 K. R. Varshney。公平性生成对抗网络。*arXiv* 预印本 *arXiv:1805.09910*，2018。G. Scutari 和 Y. Sun。在时变有向图上进行分布式非凸约束优化。《数学规划》，176(1):497–544，2019。K. Siau 和 W. Wang。在人工智能、机器学习和机器人学领域建立信任。《*Cutter*商业技术杂志》，31(2):47–53，2018。W. D. Smart 和 L. P. Kaelbling。面向移动机器人的有效强化学习。载于*IEEE*机器人与自动化国际会议论文集，第4卷，第3404–3410页，2002。H. Sun、S. Lu 和 M. Hong。改进去中心化非凸优化的样本和通信复杂度：联合梯度估计与跟踪。载于机器学习国际会议论文集，第9217–9228页。PMLR，2020。I. Tsaknakis、M. Hong 和 S. Liu。去中心化极小极大优化：问题表述、算法与应用

在网络投毒攻击中。载于*IEEE*国际声学、语音与信号处理会议（*ICASSP*）论文集，第5755–5759页。IEEE，2020。

H.-T. Wai、Z. Yang、Z. Wang 和 M. Hong。通过双重平均原始-对偶优化实现多智能体强化学习。*arXiv* 预印本 *arXiv:1806.00877*，2018。

A. Wald。最小化最大风险的统计决策函数。数学年刊，第 265–280 页，1945。

W. Wang、J. Wang、M. Kolar 和 N. Srebro。具有图正则化的分布式随机多任务学习。*arXiv* 预印本 *arXiv:1802.03830*，2018。

R. Xin、U. A. Khan 和 S. Kar。去中心化在线随机非凸优化的改进收敛分析。*arXiv* 预印本 *arXiv:2008.04195*，2020。

H. Xu、X. Liu、Y. Li、A. Jain 和 J. Tang。选择鲁棒性还是公平性：迈向对抗训练中的公平性。发表于国际机器学习会议，第 11492–11501 页。PMLR，2021。

T. Xu、Z. Wang、Y. Liang 和 H. V. Poor。用于极小极大优化的增强型一阶和零阶方差缩减算法。*arXiv* 预印本 *arXiv:2006.09361*，2020。

X. Zhang、J. Liu 和 Z. Zhu。网络上的分布式线性模型聚类：基于树的融合套索ADMM方法。*arXiv* 预印本 *arXiv:1905.11549*，2019。

X. Zhang、Z. Liu、J. Liu、Z. Zhu 和 S. Lu。降低协作式多智能体强化学习去中心化策略评估中的通信和样本复杂度。神经信息处理系统进展，34：18825–18838，2021。

K. Zhou 和 S. I. Roumeliotis。结合相对观测的多机器人主动目标跟踪。*IEEE*机器人学汇刊，27(4)：678–695，2011年。

主要结果的证明概要

由于篇幅有限，我们概述定理 1 和定理 2 的证明关键步骤。完整的证明见附录。在深入进行理论分析之前，我们先给出以下符号：

- $\bar{\mathbf{x}}_t = \frac{1}{m}\sum_{i=1}^m \mathbf{x}_{i,t}$ and $\mathbf{x}_t = [\mathbf{x}_{1,t}^\top, \cdots, \mathbf{x}_{m,t}^\top]^\top$ for any vector $\mathbf{x}$;
- $\nabla_{\mathbf{x}} F_t = [\nabla_{\mathbf{x}} F(\mathbf{x}_{1,t}, \mathbf{y}_{1,t})^\top, \cdots, \nabla_{\mathbf{x}} F(\mathbf{x}_{m,t}, \mathbf{y}_{m,t})^\top]^\top$;
- $\nabla_{\mathbf{y}} F_t = [\nabla_{\mathbf{y}} F(\mathbf{x}_{1,t}, \mathbf{y}_{1,t})^\top, \cdots, \nabla_{\mathbf{y}} F(\mathbf{x}_{m,t}, \mathbf{y}_{m,t})^\top]^\top$;
- $\mathcal{E}(\mathbf{x}_t) = \frac{1}{m}\sum_{i=1}^m \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2$ for any vector $\mathbf{x}$.

此外，下面的结果对后续分析也很有用。

引理 1. 在假设 *1* 下，函数 $J(\mathbf{x}) = F(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))$ 关于 $\mathbf{x}$ 是利普希茨光滑的，即存在一个正数常量 L_J，使得

$$\|\nabla J(\mathbf{x}) - \nabla J(\mathbf{x}')\| \le L_J \|\mathbf{x} - \mathbf{x}'\|, \ \forall \mathbf{x}, \mathbf{x}' \in \mathbb{R}^d, \quad (19)$$

其中，对于算法 *1*，利普希茨常数为 $L_J = L_f + L_f^2/\mu$。该引理可由 *Lin* 等人的引理 *4.3* 直接推出。*[2020a]*。

引理 2.在假设 *1* 下，$\mathbf{y}^*(\mathbf{x}) = \arg\max_{\mathbf{y}} F(\mathbf{x}, \mathbf{y})$ 是利普希茨连续的，即存在一个正数常量 L_y，使得

$$\|\mathbf{y}^*(\mathbf{x}) - \mathbf{y}^*(\mathbf{x}')\| \le L_y \|\mathbf{x} - \mathbf{x}'\|, \ \forall \mathbf{x}, \mathbf{x}' \in \mathbb{R}^d, \quad (20)$$

其中，利普希茨常数为 $L_y = L_f/\mu$。

A.1 证明主要定理的重要引理

我们首先证明 PRECISION 算法在函数 $Q(\cdot)$ 上具有如下下降性质，该性质由以下引理给出：

引理 3 （关于 $Q(\mathbf{x})$ 的下降不等式。 在假设 *1* 下，以下下降不等式成立：

$$\begin{aligned} Q(\bar{\mathbf{x}}_{t+1}) - Q(\bar{\mathbf{x}}_t) &\le \frac{\nu L_F^2}{2\beta}\|\mathbf{y}_t^* - \bar{\mathbf{y}}_t\|^2 \\ &+ \frac{\nu}{2\beta}\|\nabla_{\mathbf{x}} F(\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_t) - \bar{\mathbf{p}}_t\|^2 + \frac{\nu\tau}{2\beta m}\|\mathbf{x}_t - 1\otimes\bar{\mathbf{x}}_t\|^2 \\ &- \left(\frac{\nu\tau}{m} - \frac{\nu^2 L_J}{2m} - \frac{\nu\beta}{m} - \frac{\nu\tau\beta}{2m}\right)\|\tilde{\mathbf{x}}_t - 1\otimes\bar{\mathbf{x}}_t\|^2. \quad (21) \end{aligned}$$

其中 $Q(\mathbf{x}_t) = \max_{\mathbf{y}} F(\mathbf{x}_t, \mathbf{y}) + h(\mathbf{x}_t)$ 且 $\mathbf{y}_t^* = \arg\max_{\mathbf{y}} F(\bar{\mathbf{x}}_t, \mathbf{y})$。

引理 *3* 的证明概要。令 $J(\mathbf{x}_t) = \max_{\mathbf{y}} F(\mathbf{x}_t, \mathbf{y})$。根据算法更新、$J$ 的利普希茨连续梯度以及 $h(\mathbf{x})$的最优性条件，可得：

$$\begin{aligned} &J(\bar{\mathbf{x}}_{t+1}) - J(\bar{\mathbf{x}}_t) \\ &\le \frac{\nu}{m}\sum_i \langle \nabla J(\bar{\mathbf{x}}_t) - \mathbf{p}_{i,t}, \tilde{\mathbf{x}}_{i,t} - \bar{\mathbf{x}}_t\rangle \\ &+ \frac{\nu\tau}{m}\sum_i \langle \mathbf{x}_{i,t} - \bar{\mathbf{x}}_t, \tilde{\mathbf{x}}_{i,t} - \bar{\mathbf{x}}_t\rangle + \frac{\nu^2 L_J}{2m}\|\tilde{\mathbf{x}}_t - 1\otimes\bar{\mathbf{x}}_t\|^2 \\ &- \frac{\nu\tau}{m}\|\tilde{\mathbf{x}}_t - 1\otimes\bar{\mathbf{x}}_t\|^2 - h(\bar{\mathbf{x}}_{t+1}) + h(\bar{\mathbf{x}}_t). \quad (22) \end{aligned}$$

根据三角不等式和 $Q(\mathbf{x})$的定义，我们有：根据三角不等式和 $Q(\mathbf{x})$的定义，可得引理 3 中所述的结果。 □

接下来，考虑引理 3 中的误差界 $\|\bar{}\mathbf{y}_t - \mathbf{y}_t^*\|^2$，我们有以下引理：

引理 4（关于 $\mathbf{y}^*(\mathbf{x})$）。在假设 *1* 下，对于 PRECISION/PRECISION$^+$，以下不等式成立：

$$\begin{aligned}\left\|\bar{\mathbf{y}}_{t+1}-\mathbf{y}_{t+1}^*\right\|^2 \leq & \left(1-\frac{\mu\eta\alpha}{4}\right)\left\|\bar{\mathbf{y}}_t-\mathbf{y}_t^*\right\|^2 \\ & -\frac{3\eta}{4}\left\|\tilde{\mathbf{y}}_t-1\otimes\bar{\mathbf{y}}_t\right\|^2+\frac{75\eta\alpha}{16\mu}\left\|\bar{\mathbf{d}}_t-\nabla_{\mathbf{y}}F\left(\bar{\mathbf{x}}_t,\bar{\mathbf{y}}_t\right)\right\|^2 \\ & +\frac{17L_y^2\nu^2}{2\mu\eta\alpha m}\left\|\tilde{\mathbf{x}}_t-1\otimes\bar{\mathbf{x}}_t\right\|^2. \end{aligned} \tag{23}$$

引理 *4* 的证明概要。类似于 Qiu 等人 [2020],引理 B.2、B.3，并且由于对$\mathbf{y}$ 进行约束优化的最优性条件以及 μ-强凹性，我们有

$$\begin{aligned}\left\|\bar{\mathbf{y}}_{t+1}-\mathbf{y}_t^*\right\|^2 \leq & \frac{4\eta^2}{\mu}\left\|\nabla_{\mathbf{y}}F\left(\bar{\mathbf{x}}_t,\bar{\mathbf{y}}_t\right)-\bar{\mathbf{d}}_t\right\|^2 \\ & +\left(1-\frac{\eta^2\mu}{2}\right)\left\|\bar{\mathbf{y}}_t-\mathbf{y}_t^*\right\|^2-\frac{3\eta}{4}\left\|\tilde{\mathbf{y}}_t-1\otimes\bar{\mathbf{y}}_t\right\|^2. \end{aligned} \tag{24}$$

此外，我们有

$$\begin{aligned}\left\|\bar{\mathbf{y}}_{t+1}-\mathbf{y}_{t+1}^*\right\|^2 \leq & \left(1+\frac{\mu\eta^2}{4}\right)\left\|\bar{\mathbf{y}}_{t+1}-\mathbf{y}_t^*\right\|^2 \\ & -\left(1+\frac{4}{\mu\eta^2}\right)L_y^2\nu^2\left\|\frac{1}{m}\sum_i\tilde{\mathbf{x}}_{i,t}-\bar{\mathbf{x}}_t\right\|^2. \end{aligned} \tag{25}$$

根据三角不等式和 $Q(\mathbf{x})$ 的定义，我们有：

$$\begin{aligned}Q\left(\bar{\mathbf{x}}_{t+1}\right) \leq & Q\left(\bar{\mathbf{x}}_t\right)+\frac{\nu}{m}\sum_i\frac{1}{2\beta}\left\|\nabla J\left(\bar{\mathbf{x}}_t\right)-\nabla_{\mathbf{x}}F(\bar{\mathbf{x}}_t,\bar{\mathbf{y}}_t)\right\|^2 \\ & +\frac{\nu}{m}\sum_i\frac{\beta}{2}\left\|\tilde{\mathbf{x}}_{i,t}-\bar{\mathbf{x}}_t\right\|^2+\nu\frac{1}{2\beta}\left\|\nabla_{\mathbf{x}}F(\bar{\mathbf{x}}_t,\bar{\mathbf{y}}_t)-\bar{\mathbf{p}}_t\right\|^2 \\ & +\frac{\nu}{m}\sum_i\frac{\beta}{2}\left\|\tilde{\mathbf{x}}_{i,t}-\bar{\mathbf{x}}_t\right\|^2+\frac{\nu\tau}{m}\frac{1}{2\beta}\sum_i\left\|\bar{\mathbf{x}}_t-\mathbf{x}_{i,t}\right\|^2 \\ & +\frac{\nu\tau}{m}\sum_i\frac{\beta}{2}\left\|\tilde{\mathbf{x}}_{i,t}-\bar{\mathbf{x}}_t\right\|^2-\left(\frac{\nu\tau}{m}-\frac{\nu^2L_J}{2m}\right)\left\|\tilde{\mathbf{x}}_t-1\otimes\bar{\mathbf{x}}_t\right\|^2. \end{aligned}$$

对上述不等式进行一些整理后，我们得到引理 4 中所述的结果。

□

通过对前述引理从 0 到 $T+1$ 的综合结果进行望远镜求和，并经过一些整理后，我们得到以下结果：

引理 5。在假设 *1* 和条件$\eta \leq 1/2L_f$下，以下不等式对PRECISION/PRECISION$^+$ 成立：

$$\begin{aligned}& Q(\bar{\mathbf{x}}_{T+1})-Q(\bar{\mathbf{x}}_0)+\frac{4\nu L_F^2}{\beta\mu\eta^2}\left[\|\bar{\mathbf{y}}_{T+1}-\mathbf{y}_{T+1}^*\|^2-\|\mathbf{y}_0^*-\bar{\mathbf{y}}_0\|^2\right] \\ \leq & \frac{4\nu L_F^2}{\beta\mu\eta\alpha}\Big\{-\frac{3\eta}{4}\left\|\tilde{\mathbf{y}}_t-1\otimes\bar{\mathbf{y}}_t\right\|^2+\frac{17L_y^2\nu^2}{2\mu m\eta\alpha}\left\|\tilde{\mathbf{x}}_t-1\otimes\bar{\mathbf{x}}_t\right\|^2\Big\} \\ & +\frac{75\eta\alpha}{16\mu}\frac{2}{m}\|\nabla_{\mathbf{y}}F(\mathbf{x}_t,\mathbf{y}_t)-\bar{\mathbf{d}}_t\|^2 \\ & +\frac{\nu}{2\beta}\frac{2}{m}\|\nabla_{\mathbf{x}}F(\mathbf{x}_t,\mathbf{y}_t)-\bar{\mathbf{p}}_t\|^2 \\ & +\frac{\nu\tau}{2\beta m}\left\|\mathbf{x}_t-1\otimes\bar{\mathbf{x}}_t\right\|^2-\Big(\frac{\nu\tau}{m}-\frac{\nu^2L_J}{2m}-\frac{\nu\beta}{m}-\frac{\nu\tau\beta}{2m}\Big) \\ & \cdot\left\|\tilde{\mathbf{x}}_t-1\otimes\bar{\mathbf{x}}_t\right\|^2+\Big[\frac{\nu}{\beta}\frac{L_F^2}{m}+\frac{4\nu L_F^2}{\beta\mu\eta\alpha}\frac{75\eta\alpha}{16\mu}\frac{2L_F^2}{m}\Big] \\ & \cdot\sum_{i=1}^m[\|\bar{\mathbf{x}}_t-\mathbf{x}_{i,t}\|^2+\|\bar{\mathbf{y}}_t-\mathbf{y}_{i,t}\|^2] \\ & -\frac{\nu L_F^2}{2\beta}\left\|\bar{\mathbf{y}}_t-\mathbf{y}_t^*\right\|^2. \end{aligned} \tag{26}$$

接下来，我们对式 (26) 中的 $\|\mathbf{x}_t-1\otimes\bar{}\mathbf{x}_t\|^2$ 和 $\|\mathbf{y}_t-1\otimes\bar{}\mathbf{y}_t\|^2$ 的迭代收缩进行界定。

引理 6 (迭代收缩)。迭代满足以下收缩性质：

$$\begin{aligned}\|\mathbf{x}_t-1\otimes\bar{\mathbf{x}}_t\|^2 \leq & (1+c_1)\lambda^2\|\mathbf{x}_{t-1}-1\otimes\bar{\mathbf{x}}_{t-1}\|^2 \\ & +(1+\frac{1}{c_1})\nu^2\|\tilde{\mathbf{x}}_{t-1}-\mathbf{x}_{t-1}\|^2, \\ \|\mathbf{y}_t-1\otimes\bar{\mathbf{y}}_t\|^2 \leq & (1+c_2)\lambda^2\|\mathbf{y}_{t-1}-1\otimes\bar{\mathbf{y}}_{t-1}\|^2 \\ & +(1+\frac{1}{c_2})\eta^2\|\tilde{\mathbf{y}}_{t-1}-\mathbf{y}_{t-1}\|^2, \end{aligned} \tag{27}$$

其中，c_1 和 c_2 是任意正数常量。此外，我们还有

$$\begin{aligned}\|\mathbf{x}_t-\mathbf{x}_{t-1}\|^2 \leq & 8\mathcal{E}(\mathbf{x}_{t-1})+2\nu^2\|\tilde{\mathbf{x}}_{t-1}-\mathbf{x}_{t-1}\|^2, \\ \|\mathbf{y}_t-\mathbf{y}_{t-1}\|^2 \leq & 8\mathcal{E}(\mathbf{y}_{t-1})+2\eta^2\|\tilde{\mathbf{y}}_{t-1}-\mathbf{y}_{t-1}\|^2. \end{aligned} \tag{28}$$

接下来，我们对式（26）中的梯度跟踪误差 $\sum_{t=0}^T\|\bar{}\mathbf{d}_t-\nabla_{\mathbf{x}}F_t\|^2$ 和 $\sum_{t=0}^T\|\bar{}\mathbf{p}_t-\nabla_{\mathbf{y}}F_t\|^2$ 进行界定。

引理 7（梯度估计器的误差）。在假设 *2* 下，梯度跟踪器的误差界如下：

$$\begin{aligned}& \sum_{t=0}^T\|\bar{\mathbf{d}}_t-\nabla_{\mathbf{x}}F_t\|^2 \\ \leq & \sum_{t=1}^T\mathbb{E}\|\bar{\mathbf{d}}_{(n_t-1)q}-\nabla_{\mathbf{x}}F(\mathbf{x}_{(n_t-1)q},\mathbf{y}_{(n_t-1)q})\|^2 \\ & +L_f^2\big(\|\mathbf{x}_t-\mathbf{x}_{t-1}\|^2+\|\mathbf{y}_t-\mathbf{y}_{t-1}\|^2\big), \\ & \sum_{t=0}^T\|\bar{\mathbf{p}}_t-\nabla_{\mathbf{y}}F_t\|^2 \\ \leq & \sum_{t=1}^T\mathbb{E}\|\bar{\mathbf{p}}_{(n_t-1)q}-\nabla_{\mathbf{y}}F(\mathbf{x}_{(n_t-1)q},\mathbf{y}_{(n_t-1)q})\|^2 \end{aligned} \tag{29}$$

$$+ L_f^2\big(\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 + \|\mathbf{y}_t - \mathbf{y}_{t-1}\|^2\big), \tag{30}$$

其中 n_t 是满足 $(n_t-1)q \le t$ 的最大正整数。

引理 7 的证明概要。定义

$$A_{i,t} = \bar{\mathbf{d}}_{i,t} - \nabla_{\mathbf{x}} F_{i,t};\ B_{i,t} = \frac{1}{|\mathcal{S}_{i,t}|}\sum_{j\in\mathcal{S}_{i,t}} \nabla_{\mathbf{x}} f_{i,t}(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}) - \nabla_{\mathbf{x}} f_{i,t}(\mathbf{x}_{i,t-1}, \mathbf{y}_{i,t-1}) + \nabla_{\mathbf{x}} F_{i,t-1} - \nabla_{\mathbf{x}} F_{i,t}. \tag{31}$$

注意到 $\mathbb{E}_t[B_{i,t}] = 0$，其中期望是对第 t次迭代时数据采样的随机性取的。因此，

$$\mathbb{E}_t\|A_{i,t}\|^2 = \|A_{i,t-1}\|^2 + \mathbb{E}_t\|B_{i,t}\|^2. \tag{32}$$

此外，结合 $|\mathcal{S}_{i,\ t}| = q$，我们有

$$\mathbb{E}_t\|B_{i,t}\|^2 \le \frac{L_f^2}{q}\big(\|\mathbf{x}_{i,t} - \mathbf{x}_{i,t-1}\|^2 + \|\mathbf{y}_{i,t} - \mathbf{y}_{i,t-1}\|^2\big). \tag{33}$$

对 $(n_t-1)q+1$ 到 t 中的 t 取完全期望，并对 (33) 进行望远镜求和，其中 $t \le n_t q$ $\frac{L_f^2}{q}$ 1，我们得到 $\mathbb{E}\|A_t\|^2 \le \mathbb{E}\|A_{(n_t-1)q}\|^2 + \sum_{r=(n_t-1)q+1}^{t} \mathbb{E}\big(\|\mathbf{x}_r - \mathbf{x}_{r-1}\|^2 + \|\mathbf{y}_r - \mathbf{y}_{r-1}\|^2\big)$. 因此，$\sum_{k=0}^{t}\mathbb{E}\|A_k\|^2 \le \sum_{r=0}^{t}\|A_{(n_r-1)q}\|^2 + \sum_{r=1}^{t} L_f^2\big(\|\mathbf{x}_r - \mathbf{x}_{r-1}\|^2 + \|\mathbf{y}_r - \mathbf{y}_{r-1}\|^2\big)$. 当 $A_{i,\ t} = \bar{}\mathbf{p}_{i,\ t} - \nabla_{\mathbf{y}} F_{i,\ t}$ 时，也有类似结果。引理 7 的证明至此完成。 □

A.2 定理 1 的证明思路

证明。 根据定义的势函数 $\mathfrak{p}$ 和引理 3-7 的结果，我们有

$$\begin{aligned}
\mathbb{E}\mathfrak{p}_{T+1} - \mathfrak{p}_0 \le{}& \nu L_f{}^2 2\sum_{t=0}^{T}\|\bar{\mathbf{y}}_t - \mathbf{y}_t^*\|^2 \\
&- C_1\sum_{t=0}^{T}\sum_{i=1}^{m}\|\bar{\mathbf{x}}_t - \mathbf{x}_{i,t}\|^2 - C_2\sum_{t=0}^{T}\|\tilde{\mathbf{x}}_t - 1\otimes\bar{\mathbf{x}}_t\|^2 \\
&- C_3\sum_{t=0}^{T}\sum_{i=1}^{m}[\|\bar{\mathbf{y}}_t - \mathbf{y}_{i,t}\|^2] - C_4\sum_{t=0}^{T}\|\tilde{\mathbf{y}}_t - 1\otimes\bar{\mathbf{y}}_t\|^2,
\end{aligned} \tag{34}$$

其中

$$\begin{aligned}
C_1 ={}& \big[1 - 8L_f^2\big(\frac{75\eta\alpha}{16\mu}\frac{2}{m} + \frac{\nu}{2\beta}\frac{2}{m}\big) - \frac{\nu\tau}{2\beta m} \\
&- (1+c_1)\lambda^2 - \frac{\nu}{\beta}\frac{L_f^2}{m} - \frac{4\nu L_f^2}{\beta\mu\eta\alpha}\frac{75\eta\alpha}{16\mu}\frac{2L_f^2}{m}\big], \\
C_2 ={}& \big(-2\nu^2 L_f^2\big(\frac{75\eta\alpha}{16\mu}\frac{2}{m} + \frac{\nu}{2\beta}\frac{2}{m}\big) - (1+\frac{1}{c_1})\nu^2 \\
&- \frac{17L_{\mathbf{y}}^2\nu^2}{2\mu m\eta\alpha} + \frac{\nu\tau}{m} - \frac{\nu^2 L_J}{2m} - \frac{\nu\beta}{m} - \frac{\nu\tau\beta}{2m}\big),
\end{aligned}$$

$$\begin{aligned}
C_3 ={}& \big[1 - 8L_f^2\big(\frac{75\eta\alpha}{16\mu}\frac{2}{m} + \frac{\nu}{2\beta}\frac{2}{m}\big) - (1+c_2)\lambda^2 \\
&- \frac{\nu}{\beta}\frac{L_f^2}{m} - \frac{4\nu L_f^2}{\beta\mu\eta\alpha}\frac{75\eta\alpha}{16\mu}\frac{2L_f^2}{m}\big], \\
C_4 ={}& \frac{4\nu L_f^2}{\beta\mu\eta\alpha}\frac{3\eta}{4} \\
&- (1+\frac{1}{c_2})\eta^2 - 2\eta\alpha L_f^2\big(\frac{75\eta\alpha}{16\mu}\frac{2}{m} + \frac{\nu}{2\beta}\frac{2}{m}\big).
\end{aligned}$$

Suppose that $\beta \le \min\left\{\frac{\tau}{12}, \frac{1}{3}\right\}$，$\alpha \le \frac{1}{4L_f}$ 成立，并令 $c_1 = \frac{1-\lambda^2}{1+\lambda^2}$。若步长满足定理 1，以确保 $C_1,\ C_2,\ C_3,\ C_4 \ge 0$。我们可以得出结论：

$$\frac{1}{T+1}\sum_{t=0}^{T}\mathfrak{M}_t \le \frac{\mathbb{E}[\mathfrak{p}_0 - Q^*]}{\min\{C_1, C_2, \nu L_f^2/2\}(T+1)}. \tag{35}$$

定理 1 的证明至此完成。 □

A.3 定理 2 的证明思路

证明。 对于 PRECISION$^+$，我们有

$$\begin{aligned}
&\mathbb{E}\|\bar{\mathbf{d}}_{(n_t-1)q} - \nabla_{\mathbf{x}} F_{(n_t-1)q}\|^2 \\
&= \mathbb{E}\|\bar{\mathbf{p}}_{(n_t-1)q} - \nabla_{\mathbf{y}} F_{(n_t-1)q}\|^2 = \frac{I_{(\mathcal{N}_s<M)}}{\mathcal{N}_s}\sigma^2.
\end{aligned} \tag{36}$$

召回率 $\mathcal{N}_s = \min\{c_\gamma\sigma^2(\gamma^{(k)})^{-1}, c_\epsilon\sigma^2\epsilon^{-1}, M\}$，我们有

$$\frac{I_{(\mathcal{N}_s<M)}}{\mathcal{N}_s} \le \max\{\frac{\gamma^{(k)}}{c_\gamma\sigma^2}, \frac{\epsilon}{c_\epsilon\sigma^2}\} \le \frac{\gamma^{(k)}}{c_\gamma\sigma^2} + \frac{\epsilon}{c_\epsilon\sigma^2}. \tag{37}$$

由于 $\gamma_{t+1} = \frac{1}{q}\sum_{i=(n_t-1)\ q}^{t}\|\tilde{}\mathbf{x}_t - 1\otimes\bar{}\mathbf{x}_t\|^2$。将 (37) 代入引理 5，并采用额外的参数设置，可得以下结果

$$c_\gamma \ge (\frac{75\eta\alpha}{8\mu}\frac{1}{m} + \frac{\nu}{\beta}\frac{1}{m})\frac{\nu\tau}{12}.$$

对于 PRECISION$^+$，根据定义的势函数$\mathfrak{p}$ 以及引理 3-7 的结果，在 $\mathfrak{p}_{T+1} \ge Q^*$ 的条件下，我们得出结论：

$$\begin{aligned}
\frac{1}{(T+1)}\sum_{t=0}^{T}\mathfrak{M}_t \le{}& (\frac{75\eta\alpha}{16\mu}\frac{2}{m} + \frac{\nu}{2\beta}\frac{2}{m})\frac{\epsilon}{c_\epsilon} \\
&+ \frac{\mathbb{E}[\mathfrak{p}_0 - \mathfrak{p}_{T+1}]}{(T+1)\min\{C_1, C_2', \nu L_f^2/2\}},
\end{aligned} \tag{38}$$

其中 $C_1,\ C_2' \ge 0$。至此，定理 2 的证明完成。 □

B 更多实验与补充结果

下面我们详细介绍实验的设置：

1) AUC最大化模型和数据集：

给定数据集 $\{\mathbf{a}^{ij},\ b_{ij}\}_{j=1}^{n}$ 我们应用以下 AUC 最大化问题，其中$\mathbf{a}^{ij}$ 表示一个特征向量，$b_{ij} \in \{-1,+1\}$ 表示相应的标签。对于由$\mathbf{x_i} \in \mathcal{X}$参数化的分类模型，其函数为 $h_{\mathbf{x}}$，AUC 定义如下：

$$\max_{\mathbf{x}_i \in \mathcal{X}} \frac{1}{e^+e^-} \sum_{b_{ij}=+1, b_{ij}=-1} \mathbb{I}_{\{h_{\mathbf{x}_i}(\mathbf{a}_{ij}) \geq h_{\mathbf{x}_i}(\mathbf{a}_{jj})\}}, \tag{39}$$

其中，e^+(e^-) 表示正（负）样本的数量，$\mathbb{I}$ 表示指示函数。上述优化问题具有以下等价的极小极大形式：

$$\begin{aligned}
&\min_{\mathbf{x}_i, c_1, c_2} \max_{y_i} \frac{1}{m} \sum_{i=1}^{m} F_i(\mathbf{x}_i, c_1, c_2, \lambda) \\
:=&\frac{1}{m} \sum_{i=1}^{m} \Big\{ (1-\tau) \left(h_{\mathbf{x}_i}(\mathbf{a}_{ij}) - c_1\right)^2 \mathbb{I}_{\{b_{ij}=1\}} - \tau(1-\tau) y_i^2 + \tau \left(h_{\mathbf{x}_i}(\mathbf{a}_{ij}) - c_2\right)^2 \mathbb{I}_{\{b_{ij}=-1\}} \\
&+2(1+y_i)\tau h_{\mathbf{x_i}}(\mathbf{a}_{ij}) \mathbb{I}_{\{b_{ij}=-1\}} - 2(1+y_i)(1-\tau) h_{\mathbf{x}_i}(\mathbf{a}_{ij}) \mathbb{I}_{\{b_{ij}=1\}} \Big\},
\end{aligned}$$

其中，$\tau := e^+/(e^+ + e^-)$表示正样本数据的比例。

我们使用来自 LIBSVM 仓库的“a9a”数据集测试算法的收敛性能，该数据集由 Chang 和 Lin [2011] 公开提供；此外还使用了 LeCun等人的‘MNIST”数据集 [1998]。

2) 去中心化网络：我们使用一个包含五个节点的多智能体系统，其通信图 $\mathcal{G}$ 由 Erd¨os-R`eny 图生成，边连接概率为 $p_c = 0.6.$。网络共识矩阵选择为$\mathbf{W} = \mathbf{I} - \frac{2}{3\lambda_{\max}(\mathbf{L})}\mathbf{L}$，其中$\mathbf{L}$ 是 $\mathcal{G}$的拉普拉斯矩阵，$\lambda_{\max}(\mathbf{L})$ 表示$\mathbf{L}$的最大特征值。生成的拓扑如图 2 所示。

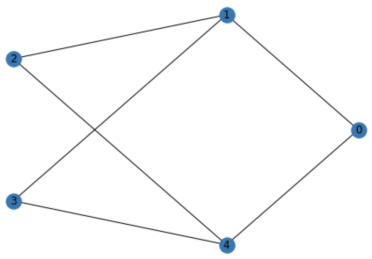


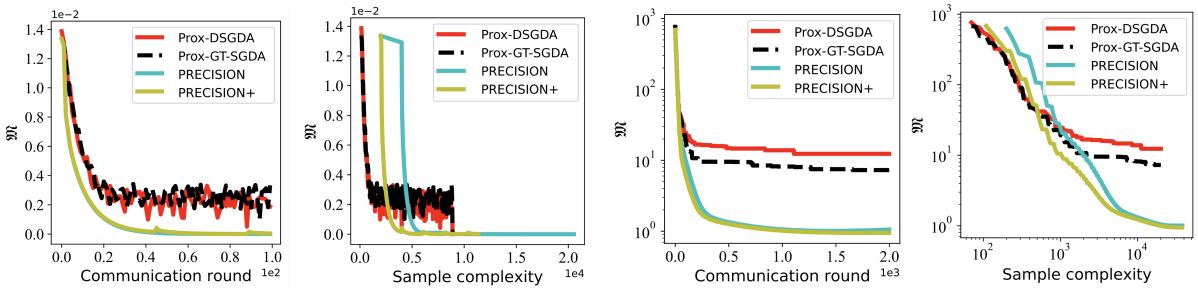
图 2：网络拓扑

B.1 算法比较

在本小节中，我们针对算法比较提供了一个额外实验。我们在 a9a 数据集和 MNIST 数据集上运行所有算法，以求解 AUC最大化问题。在该实验中，我们为所有算法从正态分布中初始化参数，并将学习率固定为 $\gamma = 10^{-1}$, $\eta = 10^{-1}$。从图 3 中可以观察到，在“a9a”数据集和“MNIST”数据集上求解 AUC最大化问题时，我们提出的算法 PRECISION/PRECISION$^+$ 具有较低的样本复杂度和通信复杂度。

B.2 学习率设定

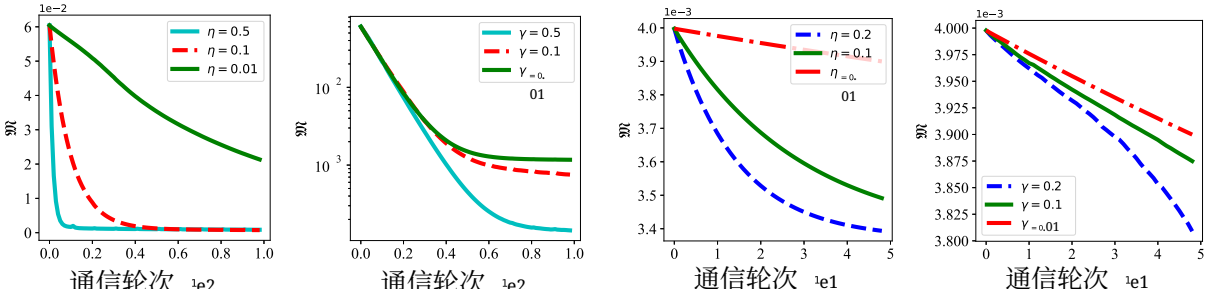
我们使用一个具有生成拓扑的 5 节点多智能体系统，如图 2 所示。在此实验中，我们选择数据量 $n = 2000$ 和小批量大小 $q = \lceil \sqrt{n} \rceil$。图 4 展示了 PRECISION 在不同设置下的收敛指标 $\mathfrak{M}$ 性能，



(a) 在“a9a”数据集上的算法比较。 (b) 在“MNIST”数据集上的算法比较。

图 3：AUC最大化问题的算法比较。

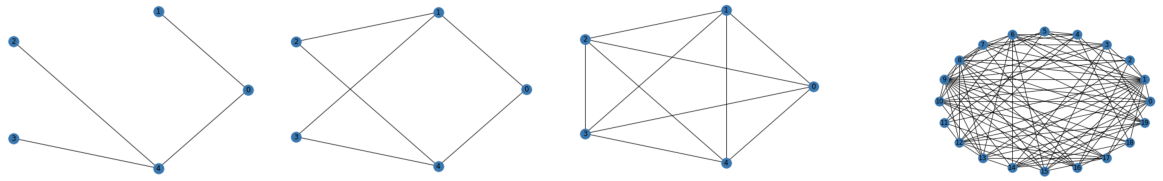
学习率分别为 γ 和 η。比较 η 时，我们固定一个相对较小的学习率 $\gamma = 10^{-1}$；比较 γ 时，则设置为 $\eta = 10^{-1}$。在此实验中，我们观察到，学习率较小的方法在图中的斜率也较小，因此收敛速度较慢。



(a) 回归的步长比较。 (b) AUC 最大化的步长比较。

图 4：不同步长下算法（PRECISION）的性能。

B.3 拓扑设置



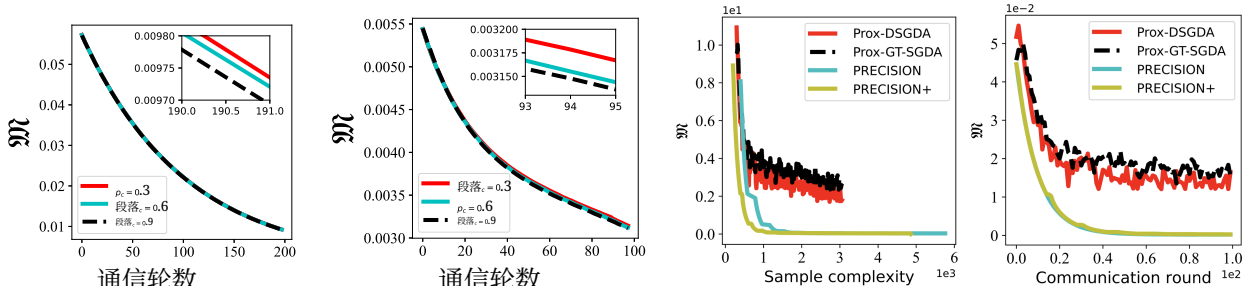
（a）拓扑稀疏度 $p_c =$ 0.3.（b）拓扑稀疏度 $p_c = 0.6$。（c）拓扑稀疏度 $p_c = 0.9$。（d）拓扑稀疏度 $p_c = 0.5$，包含 20 个节点 es.

图 5：拓扑。

我们采用一个 5 节点多智能体系统，并在三种不同的拓扑结构上进行实验。图 5 展示了生成的不同稀疏度的拓扑结构。每个智能体的数据量为 $n = 100$，我们设置恒定学习率 $\gamma = 0.1$、$\eta = 0.1$，以及小批量大小 $q = \lceil \sqrt{n} \rceil$。我们观察到，收敛指标 $\mathfrak{M}$ 对网络拓扑不敏感。图 6(a) 和图 6(b) 的子图显示，随着 p_c 减小，$\mathfrak{M}$ 略有增加。

B.4 节点设置

我们在不同的多智能体系统上进行了以下实验。图 5(d) 展示了一个 20 节点系统生成的拓扑结构。恒定学习率为 $\gamma = 0.1$、$\eta = 0.1$，小批量大小为 $q = \lceil \sqrt{n} \rceil$。我们根据式 (13) 中的收敛指标，将所提出的算法 PRECISION/PRECISION$^+$ 与两种基线算法 Prox-GT-SGDA 和 Prox-DSGDA 进行比较。我们观察到的结果与第 4 节所示的结果相似。因此，我们可以得出结论，所提出的算法 PRECISION/PRECISION$^+$ 通常具有较低的样本复杂度和通信复杂度。



(a) 拓扑稀疏性比较：回归

(b) 拓扑稀疏性比较：AUC 最大化

(c) 20 个节点的算法比较。

图 6：算法比较。

C 引理证明

在深入探讨理论分析之前，我们先定义以下符号：

- $\bar{\mathbf{x}}_t = \frac{1}{m}\sum_{i=1}^m \mathbf{x}_{i,t}$ and $\mathbf{x}_t = [\mathbf{x}_{1,t}^\top, \cdots, \mathbf{x}_{m,t}^\top]^\top$ for any vector $\mathbf{x}$;
- $\nabla_{\mathbf{x}} F_t = [\nabla_{\mathbf{x}} F(\mathbf{x}_{1,t}, \mathbf{y}_{1,t})^\top, \cdots, \nabla_{\mathbf{x}} F(\mathbf{x}_{m,t}, \mathbf{y}_{m,t})^\top]^\top$;
- $\nabla_{\mathbf{y}} F_t = [\nabla_{\mathbf{y}} F(\mathbf{x}_{1,t}, \mathbf{y}_{1,t})^\top, \cdots, \nabla_{\mathbf{y}} F(\mathbf{x}_{m,t}, \mathbf{y}_{m,t})^\top]^\top$;
- $\mathcal{E}(\mathbf{x}_t) = \frac{1}{m}\sum_{i=1}^m \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2$ for any vector $\mathbf{x}$.

C.1 引理 3 的证明：

我们的第一步是证明 PRECISION 算法在函数 $Q(\cdot)$ 上具有下降性质，如引理 3 所示。

证明。令 $J(\mathbf{x}_t) = \text{最大化}_{\mathbf{y}}\, F(\mathbf{x}_t,\ \mathbf{y})$。根据算法更新，我们有：

$$
\begin{aligned}
J(\bar{\mathbf{x}}_{t+1}) - J(\bar{\mathbf{x}}_t) &\overset{(a)}{\leq} \langle \nabla J(\bar{\mathbf{x}}_t), \bar{\mathbf{x}}_{t+1} - \bar{\mathbf{x}}_t \rangle + \frac{L_J}{2}\|\bar{\mathbf{x}}_{t+1} - \bar{\mathbf{x}}_t\|^2 \\
&\leq \left\langle \nabla J\left(\bar{\mathbf{x}}_t\right), \nu\left(\frac{1}{m}\sum_{i\in M}\tilde{\mathbf{x}}_{i,t} - \bar{\mathbf{x}}_t\right)\right\rangle + \frac{\nu^2 L_J}{2}\left\|\frac{1}{m}\sum_i \tilde{\mathbf{x}}_{i,t} - \bar{\mathbf{x}}_t\right\|^2 \\
&\leq \nu\frac{1}{m}\sum_i \left\langle \nabla J\left(\bar{\mathbf{x}}_t\right), \tilde{\mathbf{x}}_{i,t} - \bar{\mathbf{x}}_t\right\rangle + \frac{\nu^2 L_J}{2}\frac{1}{m}\left\|\tilde{\mathbf{x}}_t - 1\bar{\mathbf{x}}_t\right\|^2 \\
&\leq \nu\frac{1}{m}\sum_i \left\langle \nabla J\left(\bar{\mathbf{x}}_t\right) - \mathbf{p}_{i,t} - \tau\left(\bar{\mathbf{x}}_t - \mathbf{x}_{i,t}\right), \tilde{\mathbf{x}}_{i,t} - \bar{\mathbf{x}}_t\right\rangle + \frac{\nu^2 L_J}{2}\frac{1}{m}\left\|\tilde{\mathbf{x}}_t - 1\bar{\mathbf{x}}_t\right\|^2 \\
&\quad - \frac{\nu\tau}{m}\left\|\tilde{\mathbf{x}}_t - 1\bar{\mathbf{x}}_t\right\|^2 - h\left(\bar{\mathbf{x}}_{t+1}\right) + h\left(\bar{\mathbf{x}}_t\right)
\end{aligned}
$$

$$\leq \frac{\nu}{m}\sum_i \langle \nabla J(\bar{\mathbf{x}}_t) - \mathbf{p}_{i,t}, \tilde{\mathbf{x}}_{i,t} - \bar{\mathbf{x}}_t\rangle + \frac{\nu\tau}{m}\sum_i \langle \mathbf{x}_{i,t} - \bar{\mathbf{x}}_t, \tilde{\mathbf{x}}_{i,t} - \bar{\mathbf{x}}_t\rangle$$
$$+ \frac{\nu^2 L_J}{2m}\|\tilde{\mathbf{x}}_t - 1\bar{\mathbf{x}}_t\|^2 - \frac{\nu\tau}{m}\|\tilde{\mathbf{x}}_t - 1\bar{\mathbf{x}}_t\|^2 - h(\bar{\mathbf{x}}_{t+1}) + h(\bar{\mathbf{x}}_t), \tag{40}$$

其中 (a) 是由于 J 的利普希茨连续梯度。

□

$$\begin{aligned}
Q(\bar{\mathbf{x}}_{t+1}) &\leq Q(\bar{\mathbf{x}}_t) + \frac{\nu}{m}\sum_i \langle \nabla J(\bar{\mathbf{x}}_t) - \mathbf{p}_{i,t}, \tilde{\mathbf{x}}_{i,t} - \bar{\mathbf{x}}_t\rangle + \frac{\nu\tau}{m}\sum_i \langle \mathbf{x}_{i,t} - \bar{\mathbf{x}}_t, \tilde{\mathbf{x}}_{i,t} - \bar{\mathbf{x}}_t\rangle \\
&+ \frac{\nu^2 L_J}{2m}\|\tilde{\mathbf{x}}_t - 1\bar{\mathbf{x}}_t\|^2 - \frac{\nu\tau}{m}\|\tilde{\mathbf{x}}_t - 1\bar{\mathbf{x}}_t\|^2 \\
&\leq Q(\bar{\mathbf{x}}_t) + \frac{\nu}{m}\sum_i \langle \nabla J(\bar{\mathbf{x}}_t) - \nabla_{\mathbf{x}} F(\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_t), \tilde{\mathbf{x}}_{i,t} - \bar{\mathbf{x}}_t\rangle \\
&+ \frac{\nu}{m}\sum_i \langle \nabla_{\mathbf{x}} F(\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_t) - \bar{\mathbf{p}}_t, \tilde{\mathbf{x}}_{i,t} - \bar{\mathbf{x}}_t\rangle \\
&+ \frac{\nu\tau}{m}\sum_i \langle \mathbf{x}_{i,t} - \bar{\mathbf{x}}_t, \tilde{\mathbf{x}}_{i,t} - \bar{\mathbf{x}}_t\rangle + \frac{\nu^2 L_J}{2m}\|\tilde{\mathbf{x}}_t - 1\bar{\mathbf{x}}_t\|^2 - \frac{\nu\tau}{m}\|\tilde{\mathbf{x}}_t - 1\bar{\mathbf{x}}_t\|^2 \\
&\leq Q(\bar{\mathbf{x}}_t) + \frac{\nu}{m}\sum_i \frac{1}{2\beta}\|\nabla J(\bar{\mathbf{x}}_t) - \nabla_{\mathbf{x}} F(\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_t)\|^2 + \frac{\nu}{m}\sum_i \frac{\beta}{2}\|\tilde{\mathbf{x}}_{i,t} - \bar{\mathbf{x}}_t\|^2 \\
&+ \nu\frac{1}{2\beta}\|\nabla_{\mathbf{x}} F(\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_t) - \bar{\mathbf{p}}_t\|^2 + \frac{\nu}{m}\sum_i \frac{\beta}{2}\|\tilde{\mathbf{x}}_{i,t} - \bar{\mathbf{x}}_t\|^2 \\
&+ \frac{\nu\tau}{m}\frac{1}{2\beta}\sum_i \|\bar{\mathbf{x}}_t - \mathbf{x}_{i,t}\|^2 + \frac{\nu\tau}{m}\sum_i \frac{\beta}{2}\|\tilde{\mathbf{x}}_{i,t} - \bar{\mathbf{x}}_t\|^2 - \left(\frac{\nu\tau}{m} - \frac{\nu^2 L_J}{2m}\right)\|\tilde{\mathbf{x}}_t - 1\bar{\mathbf{x}}_t\|^2 \\
&\leq Q(\bar{\mathbf{x}}_t) + \frac{\nu L_F^2}{2\beta m}\sum_i \|\mathbf{y}_t^* - \bar{\mathbf{y}}_t\|^2 + \frac{\nu}{2\beta}\|\nabla_{\mathbf{x}} F(\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_t) - \bar{\mathbf{p}}_t\|^2 \\
&+ \frac{\nu\tau}{2\beta m}\|\mathbf{x}_t - 1\bar{\mathbf{x}}_t\|^2 + \left(\frac{\beta\nu}{m} + \frac{\nu\tau\beta}{2m}\right)\|\tilde{\mathbf{x}}_t - 1\bar{\mathbf{x}}_t\|^2 - \left(\frac{\nu\tau}{m} - \frac{\nu^2 L_J}{2m}\right)\|\tilde{\mathbf{x}}_t - 1\bar{\mathbf{x}}_t\|^2 \\
&\leq Q(\bar{\mathbf{x}}_t) + \frac{\nu L_F^2}{2\beta}\|\mathbf{y}_t^* - \bar{\mathbf{y}}_t\|^2 + \frac{\nu}{2\beta}\sum_i \|\nabla_{\mathbf{x}} F(\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_t) - \bar{\mathbf{p}}_t\|^2 \\
&+ \frac{\nu\tau}{2\beta m}\|\mathbf{x}_t - 1\bar{\mathbf{x}}_t\|^2 - \left(\frac{\nu\tau}{m} - \frac{\nu^2 L_J}{2m} - \frac{\nu\beta}{m} - \frac{\nu\tau\beta}{2m}\right)\|\tilde{\mathbf{x}}_t - 1\bar{\mathbf{x}}_t\|^2,
\end{aligned} \tag{41}$$

C.2 引理 4 的证明

注意，引理 3 右侧存在一个误差项 $\|\mathbf{y}_t^* - \bar{\mathbf{y}}_t\|^2$。这里，引理 4 给出了该误差项的收缩性质。

证明。 召回率：$\mathbf{y}_t^* = \mathbf{y}^*(\bar{\mathbf{x}}_t) = \arg\max_{\mathbf{y}} F(\bar{\mathbf{x}}_t, \mathbf{y})$。我们有：

$$\begin{aligned}
\|\bar{\mathbf{y}}_{t+1} - \mathbf{y}_t^*\|^2 &= \left\|\bar{\mathbf{y}}_t + \eta\left(\frac{1}{m}\sum_{i\in M}\widetilde{\mathbf{y}}_i - \bar{\mathbf{y}}_t\right) - \mathbf{y}_t^*\right\|^2 \\
&= \|\bar{\mathbf{y}}_t - \mathbf{y}_t^*\|^2 + 2\eta\left\langle \bar{\mathbf{y}}_t - \mathbf{y}_t^*, \frac{1}{m}\sum_{i\in M}\widetilde{\mathbf{y}}_i - \bar{\mathbf{y}}_t\right\rangle + \eta^2\left\|\frac{1}{m}\sum_{i\in M}\widetilde{\mathbf{y}}_i - \bar{\mathbf{y}}_t\right\|^2 \\
&\leq \|\bar{\mathbf{y}}_t - \mathbf{y}_t^*\|^2 + 2\eta\left\langle \bar{\mathbf{y}}_t - \mathbf{y}_t^*, \frac{1}{m}\sum_{i\in M}\widetilde{\mathbf{y}}_i - \bar{\mathbf{y}}_t\right\rangle + \eta^2\|\tilde{\mathbf{y}}_t - 1\bar{\mathbf{y}}_t\|^2.
\end{aligned} \tag{42}$$

根据投影运算，我们有

$$\tilde{\mathbf{y}}_i(\mathbf{y}_{i,t}) = arg\,min_{\mathbf{y}_i \in \mathcal{Y}} \left\| \mathbf{y}_i - \left(\mathbf{y}_{i,t} + \alpha \mathbf{d}_{i,t} \right) \right\|^2. \tag{43}$$

由于约束凸优化的最优性条件，我们有

$$\left\langle \widetilde{\mathbf{y}}_i - \left(\mathbf{y}_{i,t} + \alpha \mathbf{d}_{i,t} \right), \mathbf{y} - \widetilde{\mathbf{y}}_i \right\rangle \geq 0, \quad \forall \mathbf{y} \in \mathcal{Y}, i \in M. \tag{44}$$

因此，我们有

$$\left\langle -\mathbf{d}_{i,t} + \alpha^{-1} \left(\widetilde{\mathbf{y}}_i - \mathbf{y}_{i,t} \right), \mathbf{y} - \widetilde{\mathbf{y}}_i \right\rangle \geq 0, \forall \mathbf{y} \in \mathcal{Y}, i \in M. \tag{45}$$

此外，我们有

$$F\left(\bar{\mathbf{x}}_t, \mathbf{y} \right) - F\left(\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_t \right) - \left\langle \nabla_{\mathbf{y}} F\left(\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_t \right), \mathbf{y} - \bar{\mathbf{y}}_t \right\rangle \leq -\frac{\mu}{2} \left\| \mathbf{y} - \bar{\mathbf{y}}_t \right\|^2 \tag{46}$$

重新整理上述不等式中的各项，我们有

$$\begin{aligned} F\left(\bar{\mathbf{x}}_t, \mathbf{y} \right) + \frac{\mu}{2} \left\| \mathbf{y} - \bar{\mathbf{y}}_t \right\|^2 \leq & F\left(\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_t \right) + \left\langle \nabla_{\mathbf{y}} F\left(\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_t \right), \mathbf{y} - \bar{\mathbf{y}}_t \right\rangle \\ \leq & F\left(\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_t \right) + \frac{1}{\alpha} \left\langle \tilde{\mathbf{y}}_t - 1\bar{\mathbf{y}}_t, \mathbf{y} - \frac{1}{m} \sum_{i \in M} \widetilde{\mathbf{y}}_i \right\rangle + \left\langle \nabla_{\mathbf{y}} F\left(\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_t \right) - \bar{\mathbf{d}}_t, \mathbf{y} - \frac{1}{m} \sum_{i \in M} \widetilde{\mathbf{y}}_i \right\rangle \\ & + \left\langle \nabla_{\mathbf{y}} F\left(\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_t \right), \tilde{\mathbf{y}}_t - 1\bar{\mathbf{y}}_t \right\rangle - \frac{1}{2\alpha} \left\| \tilde{\mathbf{y}}_t - 1\bar{\mathbf{y}}_t \right\|^2 + \frac{1}{2\alpha} \left\| \tilde{\mathbf{y}}_t - 1\bar{\mathbf{y}}_t \right\|^2. \end{aligned} \tag{47}$$

由于 F(**x**、**y**) 具有梯度利普希茨性质，并且根据本引理中的条件

$$\alpha \leq \frac{1}{2L_F} \leq \frac{1}{L_F},$$

我们有

$$\begin{aligned} -\frac{1}{2\alpha} \left\| \tilde{\mathbf{y}}_t - 1\bar{\mathbf{y}}_t \right\|^2 \leq & -\frac{L_F}{2} \left\| \tilde{\mathbf{y}}_t - 1\bar{\mathbf{y}}_t \right\|^2 \\ & \leq F\left(\mathbf{x}_t, \tilde{\mathbf{y}}_t \right) - F\left(\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_t \right) - \left\langle \nabla_{\mathbf{y}} F\left(\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_t \right), \tilde{\mathbf{y}}_t - 1\bar{\mathbf{y}}_t \right\rangle. \end{aligned} \tag{48}$$

$$\begin{aligned} F\left(\bar{\mathbf{x}}_t, \mathbf{y} \right) + \frac{\mu}{2} \left\| \mathbf{y} - \bar{\mathbf{y}}_t \right\|^2 \leq & F\left(\mathbf{x}_t, \tilde{\mathbf{y}}_t \right) + \frac{1}{\alpha} \left\langle \tilde{\mathbf{y}}_t - 1\bar{\mathbf{y}}_t, \mathbf{y} - \tilde{\mathbf{y}}_t \right\rangle \\ & + \left\langle \nabla_{\mathbf{y}} F\left(\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_t \right) - \bar{\mathbf{d}}_t, \mathbf{y} - \tilde{\mathbf{y}}_t \right\rangle + \frac{1}{2\alpha} \left\| \tilde{\mathbf{y}}_t - 1\bar{\mathbf{y}}_t \right\|^2. \end{aligned} \tag{49}$$

请注意，在最后一个不等式中，我们有

$$\begin{aligned} & \frac{1}{\alpha} \left\langle \tilde{\mathbf{y}}_t - 1\bar{\mathbf{y}}_t, \mathbf{y} - \tilde{\mathbf{y}}_t \right\rangle + \frac{1}{2\alpha} \left\| \tilde{\mathbf{y}}_t - 1\bar{\mathbf{y}}_t \right\|^2 \\ & = \frac{1}{\alpha} \left\langle \tilde{\mathbf{y}}_t - 1\bar{\mathbf{y}}_t, \bar{\mathbf{y}}_t - \tilde{\mathbf{y}}_t \right\rangle + \frac{1}{\alpha} \left\langle \tilde{\mathbf{y}}_t - 1\bar{\mathbf{y}}_t, \mathbf{y} - \bar{\mathbf{y}}_t \right\rangle + \frac{1}{2\alpha} \left\| \tilde{\mathbf{y}}_t - 1\bar{\mathbf{y}}_t \right\|^2 \\ & = \frac{1}{\alpha} \left\langle \tilde{\mathbf{y}}_t - 1\bar{\mathbf{y}}_t, \mathbf{y} - \bar{\mathbf{y}}_t \right\rangle - \frac{1}{2\alpha} \left\| \tilde{\mathbf{y}}_t - 1\bar{\mathbf{y}}_t \right\|^2, \end{aligned} \tag{50}$$

从而导致

$$\begin{aligned} F\left(\bar{\mathbf{x}}_t, \mathbf{y} \right) + \frac{\mu}{2} \left\| \mathbf{y} - \bar{\mathbf{y}}_t \right\|^2 \leq & F\left(\bar{\mathbf{x}}_t, \tilde{\mathbf{y}}_t \right) + \frac{1}{\alpha} \left\langle \tilde{\mathbf{y}}_t - 1\bar{\mathbf{y}}_t, \mathbf{y} - \bar{\mathbf{y}}_t \right\rangle \\ & + \left\langle \nabla_{\mathbf{y}} F\left(\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_t \right) - \bar{\mathbf{d}}_t, \mathbf{y} - \tilde{\mathbf{y}}_t \right\rangle - \frac{1}{2\alpha} \left\| \tilde{\mathbf{y}}_t - 1\bar{\mathbf{y}}_t \right\|^2. \end{aligned} \tag{51}$$

我们令$\mathbf{y} = \mathbf{y}_t^*$，并得到

$$F\left(\bar{\mathbf{x}}_t, \mathbf{y}_t^*\right) + \frac{\mu}{2}\left\|\mathbf{y}_t^* - \bar{\mathbf{y}}_t\right\|^2 \leq F\left(\bar{\mathbf{x}}_t, \tilde{\mathbf{y}}_t\right) + \frac{1}{\alpha}\left\langle \tilde{\mathbf{y}}_t - 1\bar{\mathbf{y}}_t, \mathbf{y}_t^* - \bar{\mathbf{y}}_t\right\rangle + \left\langle \nabla_{\mathbf{y}} F\left(\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_t\right) - \bar{\mathbf{d}}_t, \mathbf{y}_t^* - \tilde{\mathbf{y}}_t\right\rangle - \frac{1}{2\alpha}\left\|\tilde{\mathbf{y}}_t - 1\bar{\mathbf{y}}_t\right\|^2, \tag{52}$$

从而得到

$$\begin{aligned}&\frac{\mu}{2}\|\mathbf{y}_t^* - \bar{\mathbf{y}}_t\|^2 + \frac{1}{2\alpha}\left\|\tilde{\mathbf{y}}_t - 1\bar{\mathbf{y}}_t\right\|^2 \\ &\quad\leq \frac{1}{\alpha}\left\langle \tilde{\mathbf{y}}_t - 1\bar{\mathbf{y}}_t, \mathbf{y}_t^* - \bar{\mathbf{y}}_t\right\rangle + \left\langle \nabla_{\mathbf{y}} F\left(\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_t\right) - \bar{\mathbf{d}}_t, \mathbf{y}_t^* - \tilde{\mathbf{y}}_t\right\rangle.\end{aligned} \tag{53}$$

$F(\bar{\mathbf{x}}_t,\ \mathbf{y}_t^*) \geq F(\bar{\mathbf{x}}_t,\ \tilde{\mathbf{y}}_t)$ 源于强凹性，且 $\mathbf{y}_t^* = \operatorname{argmax}_{\mathbf{y}\in\mathcal{Y}} F(\bar{\mathbf{x}}_t,\ \mathbf{y})$。此外，对于上述不等式的最后一项，我们进一步将其界定如下

$$\begin{aligned}&\left\langle \nabla_{\mathbf{y}} F\left(\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_t\right) - \bar{\mathbf{d}}_t, \mathbf{y}_t^* - \tilde{\mathbf{y}}_t\right\rangle \\ &\leq \frac{2}{\mu}\left\|\nabla_{\mathbf{y}} F\left(\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_t\right) - \bar{\mathbf{d}}_t\right\|^2 + \frac{\mu}{4}\left\|\mathbf{y}_t^* - \bar{\mathbf{y}}_t\right\|^2 + \frac{\mu}{4}\left\|\tilde{\mathbf{y}}_t - 1\bar{\mathbf{y}}_t\right\|^2.\end{aligned} \tag{54}$$

于是，我们有

$$\begin{aligned}&2\eta\left\langle \tilde{\mathbf{y}}_t - 1\bar{\mathbf{y}}_t, \bar{\mathbf{y}}_t - \mathbf{y}_t^*\right\rangle \\ &\leq -\frac{\eta\alpha\mu}{2}\left\|\bar{\mathbf{y}}_t - \mathbf{y}_t^*\right\|^2 - \frac{2\alpha - \eta\alpha\mu}{2}\left\|\tilde{\mathbf{y}}_t - 1\bar{\mathbf{y}}_t\right\|^2 + \frac{4\eta\alpha}{\mu}\left\|\nabla_{\mathbf{y}} F\left(\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_t\right) - \bar{\mathbf{d}}_t\right\|^2,\end{aligned} \tag{55}$$

这给出了式（42）右侧第二项的上界。于是，我们有

$$\left\|\bar{\mathbf{y}}_{t+1} - \mathbf{y}_t^*\right\|^2 \leq \frac{2 - \eta\alpha\mu}{2}\left\|\bar{\mathbf{y}}_t - \mathbf{y}_t^*\right\|^2 - \frac{2\eta - \eta\alpha\mu - 2\eta^2}{2}\left\|\tilde{\mathbf{y}}_t - 1\bar{\mathbf{y}}_t\right\|^2 + \frac{4\eta\alpha}{\mu}\left\|\nabla_{\mathbf{y}} F\left(\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_t\right) - \bar{\mathbf{d}}_t\right\|^2.. \tag{56}$$

因此，根据引理的条　　件 $\eta \leq 1/8$ and $\alpha \leq (4L_F)^{-1} \leq (4\mu)^{-1}$ 以及事实 $L_F \geq \mu > 0$，我们有　　e

$$-\frac{2\eta - \eta\alpha\mu - 2\eta^2}{2} \leq -\frac{3\eta}{4}, \tag{57}$$

最终得到

$$\left\|\bar{\mathbf{y}}_{t+1} - \mathbf{y}_t^*\right\|^2 \leq \left(1 - \frac{\eta\alpha\mu}{2}\right)\left\|\bar{\mathbf{y}}_t - \mathbf{y}_t^*\right\|^2 - \frac{3\eta}{4}\left\|\tilde{\mathbf{y}}_t - 1\bar{\mathbf{y}}_t\right\|^2 + \frac{4\eta\alpha}{\mu}\left\|\nabla_{\mathbf{y}} F\left(\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_t\right) - \bar{\mathbf{d}}_t\right\|^2. \tag{58}$$

为简洁起见，将$\mathbf{y}^*(\mathbf{x}_t)$和$\mathbf{y}^*(\mathbf{x}_{t+1})$分别记作$\mathbf{y}_t^*$和$\mathbf{y}_{t+1}^*$，我们先将项$\left\|\bar{\mathbf{y}}_{t+1} - \mathbf{y}_{t+1}^*\right\|^2$ 分解如下，以此开始证明

$$\begin{aligned}\left\|\bar{\mathbf{y}}_{t+1} - \mathbf{y}_{t+1}^*\right\|^2 &= \left\|\bar{\mathbf{y}}_{t+1} - \mathbf{y}_t^* + \mathbf{y}_t^* - \mathbf{y}_{t+1}^*\right\|^2 \\ &\leq \left(1 + \frac{\mu\eta\alpha}{4}\right)\left\|\bar{\mathbf{y}}_{t+1} - \mathbf{y}_t^*\right\|^2 + \left(1 + \frac{4}{\mu\eta\alpha}\right)\left\|\mathbf{y}_t^* - \mathbf{y}_{t+1}^*\right\|^2 \\ &\leq \left(1 + \frac{\mu\eta\alpha}{4}\right)\left\|\bar{\mathbf{y}}_{t+1} - \mathbf{y}_t^*\right\|^2 + \left(1 + \frac{4}{\mu\eta\alpha}\right)L_{\mathbf{y}}^2\left\|\bar{\mathbf{x}}_{t+1} - \bar{\mathbf{x}}_t\right\|^2.\end{aligned} \tag{59}$$

接下来，将更新规则 $\bar{\mathbf{x}}_{t+1} = \bar{\mathbf{x}}_t + \nu\left(\frac{1}{m}\sum_i \tilde{\mathbf{x}}_{i,t} - \bar{\mathbf{x}}_t\right)$ 代入上述不等式，可得

$$\left\|\bar{\mathbf{y}}_{t+1} - \mathbf{y}_{t+1}^*\right\|^2 \leq \left(1 + \frac{\mu\eta\alpha}{4}\right)\left\|\bar{\mathbf{y}}_{t+1} - \mathbf{y}_t^*\right\|^2 + \left(1 + \frac{4}{\mu\eta\alpha}\right)L_{\mathbf{y}}^2\nu^2\left\|\frac{1}{m}\sum_i \tilde{\mathbf{x}}_{i,t} - \bar{\mathbf{x}}_t\right\|^2. \tag{60}$$

此外，我们有

$$\begin{aligned}&\|\bar{\mathbf{y}}_{t+1}-\mathbf{y}_t^*\|^2\\&\leq\left(1-\frac{\eta\alpha\mu}{2}\right)\|\bar{\mathbf{y}}_t-\mathbf{y}_t^*\|^2-\frac{3\eta}{4}\|\tilde{\mathbf{y}}_t-1\bar{\mathbf{y}}_t\|^2+\frac{4\eta\alpha}{\mu}\left\|\nabla_{\mathbf{y}}F\left(\bar{\mathbf{x}}_t,\bar{\mathbf{y}}_t\right)-\bar{\mathbf{d}}_t\right\|^2.\end{aligned}\tag{61}$$

根据条件 $0<\alpha\leq(4L_F)^{-1}$， $0<\eta\leq 1/8$ 以及 $L_F\geq\mu>0$，我们有

$$\alpha\leq\frac{1}{4L_F}\leq\frac{1}{4\mu},\quad\text{and}\quad\eta\alpha\leq\frac{1}{32\mu},\tag{62}$$

从而得到

$$\left(1+\frac{\mu\eta\alpha}{4}\right)\left(1-\frac{\mu\eta\alpha}{2}\right)=1-\frac{\mu\eta\alpha}{2}+\frac{\mu\eta\alpha}{4}-\frac{\mu^2\eta^2\alpha^2}{4}\leq 1-\frac{\mu\eta\alpha}{4}\tag{63}$$

$$-\left(1+\frac{\mu\eta\alpha}{4}\right)\frac{3\eta}{4}\leq-\frac{3\eta}{4},\quad\frac{4\eta\alpha}{\mu}\left(1+\frac{\mu\eta\alpha}{4}\right)=\frac{4\eta\alpha}{\mu}+\eta^2\alpha^2<\frac{75\eta\alpha}{16\mu}\tag{64}$$

$$\text{and }\left(1+\frac{4}{\mu\eta\alpha}\right)L_{\mathbf{y}}^2\nu^2\leq\frac{129}{32}\frac{L_{\mathbf{y}}^2\nu}{\mu\eta\alpha}<\frac{17L_{\mathbf{y}}^2\nu^2}{2\mu\eta\alpha}\tag{65}$$

最终我们得到

$$\begin{aligned}\left\|\bar{\mathbf{y}}_{t+1}-\mathbf{y}_{t+1}^*\right\|^2\leq&\left(1-\frac{\mu\eta\alpha}{4}\right)\|\bar{\mathbf{y}}_t-\mathbf{y}_t^*\|^2-\frac{3\eta}{4}\|\tilde{\mathbf{y}}_t-1\bar{\mathbf{y}}_t\|^2\\&+\frac{75\eta\alpha}{16\mu}\left\|\bar{\mathbf{d}}_t-\nabla_{\mathbf{y}}F\left(\bar{\mathbf{x}}_t,\bar{\mathbf{y}}_t\right)\right\|^2+\frac{17L_{\mathbf{y}}^2\nu^2}{2\mu\eta\alpha m}\|\tilde{\mathbf{x}}_t-1\bar{\mathbf{x}}_t\|^2.\end{aligned}\tag{66}$$

这就完成了证明。 □

C.3 引理 5 的证明

Next，通过结合引理3和引理4的结果，我们得到引理 5 所示的下降结果。

证明。由引理3和引理4可得

$$\begin{aligned}&Q(\bar{\mathbf{x}}_{t+1})-Q(\bar{\mathbf{x}}_t)+\frac{4\nu L_F^2}{\beta\mu\eta\alpha}\left[\|\bar{\mathbf{y}}_{t+1}-\mathbf{y}_{t+1}^*\|^2-\|\mathbf{y}_t^*-\bar{\mathbf{y}}_t\|^2\right]\\\leq&\frac{4\nu L_F^2}{\beta\mu\eta\alpha}\Big[\left(-\frac{\mu\eta\alpha}{4}\right)\|\bar{\mathbf{y}}_t-\mathbf{y}_t^*\|^2-\frac{3\eta}{4}\|\tilde{\mathbf{y}}_t-1\bar{\mathbf{y}}_t\|^2+\frac{75\eta\alpha}{16\mu}\left\|\bar{\mathbf{d}}_t-\nabla_{\mathbf{y}}F\left(\bar{\mathbf{x}}_t,\bar{\mathbf{y}}_t\right)\right\|^2\\&+\frac{17L_{\mathbf{y}}^2\nu^2}{2\mu m\eta\alpha}\left\|\tilde{\mathbf{x}}_t-1\bar{\mathbf{x}}_t\right\|^2\Big]+\frac{\nu L_F^2}{2\beta}\|\bar{\mathbf{y}}_t-\mathbf{y}_t^*\|^2+\frac{\nu}{2\beta}\|\nabla_{\mathbf{x}}F(\bar{\mathbf{x}}_t,\bar{\mathbf{y}}_t)-\bar{\mathbf{p}}_t\|^2\\&+\frac{\nu\tau}{2\beta m}\|\mathbf{x}_t-1\bar{\mathbf{x}}_t\|^2-\left(\frac{\nu\tau}{m}-\frac{\nu^2L_J}{2m}-\frac{\nu\beta}{m}-\frac{\nu\tau\beta}{2m}\right)\|\tilde{\mathbf{x}}_t-1\bar{\mathbf{x}}_t\|^2\\=&\frac{4\nu L_F^2}{\beta\mu\eta\alpha}\Big[-\frac{3\eta}{4}\|\tilde{\mathbf{y}}_t-1\bar{\mathbf{y}}_t\|^2+\frac{75\eta\alpha}{16\mu}\left\|\bar{\mathbf{d}}_t-\nabla_{\mathbf{y}}F\left(\bar{\mathbf{x}}_t,\bar{\mathbf{y}}_t\right)\right\|^2\\&+\frac{17L_{\mathbf{y}}^2\nu^2}{2\mu m\eta\alpha}\left\|\tilde{\mathbf{x}}_t-1\bar{\mathbf{x}}_t\right\|^2\Big]-\frac{\nu L_F^2}{2\beta}\|\bar{\mathbf{y}}_t-\mathbf{y}_t^*\|^2+\frac{\nu}{2\beta}\|\nabla_{\mathbf{x}}F(\bar{\mathbf{x}}_t,\bar{\mathbf{y}}_t)-\bar{\mathbf{p}}_t\|^2\\&+\frac{\nu\tau}{2\beta m}\|\mathbf{x}_t-1\bar{\mathbf{x}}_t\|^2-\left(\frac{\nu\tau}{m}-\frac{\nu^2L_J}{2m}-\frac{\nu\beta}{m}-\frac{\nu\tau\beta}{2m}\right)\|\tilde{\mathbf{x}}_t-1\bar{\mathbf{x}}_t\|^2.\end{aligned}\tag{67}$$

请注意

$$\|\nabla_{\mathbf{x}}F(\bar{\mathbf{x}}_t,\bar{\mathbf{y}}_t)-\bar{\mathbf{p}}_t\|^2$$

$$
\begin{aligned}
=&\|\nabla_{\mathbf{x}} F(\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_t) - \frac{1}{m}\sum_{i=1}^{m} \nabla_{\mathbf{x}} F_i(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}) + \frac{1}{m}\sum_{i=1}^{m} \nabla_{\mathbf{x}} F_i(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}) - \bar{\mathbf{p}}_t\|^2 \\
\leq& 2\|\nabla_{\mathbf{x}} F(\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_t) - \frac{1}{m}\sum_{i=1}^{m} \nabla_{\mathbf{x}} F_i(\mathbf{x}_{i,t}, \mathbf{y}_{i,t})\|^2 + 2\|\frac{1}{m}\sum_{i=1}^{m} \nabla_{\mathbf{x}} F_i(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}) - \bar{\mathbf{p}}_t\|^2 \\
\leq& \frac{2}{m}\sum_{i=1}^{m} \|\nabla_{\mathbf{x}} F(\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_t) - \nabla_{\mathbf{x}} F_i(\mathbf{x}_{i,t}, \mathbf{y}_{i,t})\|^2 + 2\|\frac{1}{m}\sum_{i=1}^{m} \nabla_{\mathbf{x}} F_i(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}) - \bar{\mathbf{p}}_t\|^2 \\
\leq& \frac{2L_F^2}{m}\sum_{i=1}^{m} [\|\bar{\mathbf{x}}_t - \mathbf{x}_{i,t}\|^2 + \|\bar{\mathbf{y}}_t - \mathbf{y}_{i,t}\|^2] + 2\|\frac{1}{m}\sum_{i=1}^{m} \nabla_{\mathbf{x}} F_i(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}) - \bar{\mathbf{p}}_t\|^2. \qquad (68)
\end{aligned}
$$

同理，我们有：

$$
\begin{aligned}
\|\nabla_{\mathbf{y}} F(\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_t) - \bar{\mathbf{d}}_t\|^2 \leq \frac{2L_F^2}{m}\sum_{i=1}^{m} &[\|\bar{\mathbf{x}}_t - \mathbf{x}_{i,t}\|^2 + \|\bar{\mathbf{y}}_t - \mathbf{y}_{i,t}\|^2] \\
&+ 2\|\frac{1}{m}\sum_{i=1}^{m} \nabla_{\mathbf{y}} F_i(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}) - \bar{\mathbf{d}}_t\|^2. \qquad (69)
\end{aligned}
$$

因此，我们有

$$
\begin{aligned}
& Q(\bar{\mathbf{x}}_{t+1}) - Q(\bar{\mathbf{x}}_t) + \frac{4\nu L_F^2}{\beta\mu\eta\alpha}\big[\|\bar{\mathbf{y}}_{t+1} - \mathbf{y}_{t+1}^*\|^2 - \|\mathbf{y}_t^* - \bar{\mathbf{y}}_t\|^2\big] \\
\leq& \frac{4\nu L_F^2}{\beta\mu\eta\alpha}\Big\{ -\frac{3\eta}{4}\|\tilde{\mathbf{y}}_t - 1\bar{\mathbf{y}}_t\|^2 + \frac{75\eta\alpha}{16\mu}\big[\frac{2L_F^2}{m}\sum_{i=1}^{m}(\|\bar{\mathbf{x}}_t - \mathbf{x}_{i,t}\|^2 + \|\bar{\mathbf{y}}_t - \mathbf{y}_{i,t}\|^2) + 2\|\frac{1}{m}\sum_{i=1}^{m}\nabla_{\mathbf{y}} F_i(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}) - \bar{\mathbf{d}}_t\|^2\big] \\
&+ \frac{17L_{\mathbf{y}}^2\nu^2}{2\mu m\eta\alpha}\|\tilde{\mathbf{x}}_t - 1\bar{\mathbf{x}}_t\|^2\Big\} + \frac{\nu}{2\beta}\Big\{\frac{2L_F^2}{m}\sum_{i=1}^{m}[\|\bar{\mathbf{x}}_t - \mathbf{x}_{i,t}\|^2 + \|\bar{\mathbf{y}}_t - \mathbf{y}_{i,t}\|^2] + 2\|\frac{1}{m}\sum_{i=1}^{m}\nabla_{\mathbf{x}} F_i(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}) - \bar{\mathbf{p}}_t\|^2\Big\} \\
&+ \frac{\nu\tau}{2\beta m}\|\mathbf{x}_t - 1\bar{\mathbf{x}}_t\|^2 - \left(\frac{\nu\tau}{m} - \frac{\nu^2 L_J}{2m} - \frac{\nu\beta}{m} - \frac{\nu\tau\beta}{2m}\right)\|\tilde{\mathbf{x}}_t - 1\bar{\mathbf{x}}_t\|^2 - \frac{\nu L_F^2}{2\beta}\|\bar{\mathbf{y}}_t - \mathbf{y}_t^*\|^2 \\
\overset{(a)}{\leq}& \frac{4\nu L_F^2}{\beta\mu\eta\alpha}\Big\{ -\frac{3\eta}{4}\|\tilde{\mathbf{y}}_t - 1\bar{\mathbf{y}}_t\|^2 + \frac{75\eta\alpha}{16\mu}\big[\frac{2L_F^2}{m}\sum_{i=1}^{m}(\|\bar{\mathbf{x}}_t - \mathbf{x}_{i,t}\|^2 + \|\bar{\mathbf{y}}_t - \mathbf{y}_{i,t}\|^2) + \frac{2}{m}\|\nabla_{\mathbf{y}} F(\mathbf{x}_t, \mathbf{y}_t) - \bar{\mathbf{d}}_t\|^2\big] \\
&+ \frac{17L_{\mathbf{y}}^2\nu^2}{2\mu m\eta\alpha}\|\tilde{\mathbf{x}}_t - 1\bar{\mathbf{x}}_t\|^2\Big\} + \frac{\nu}{2\beta}\Big\{\frac{2L_F^2}{m}\sum_{i=1}^{m}[\|\bar{\mathbf{x}}_t - \mathbf{x}_{i,t}\|^2 + \|\bar{\mathbf{y}}_t - \mathbf{y}_{i,t}\|^2] + \frac{2}{m}\|\nabla_{\mathbf{x}} F(\mathbf{x}_t, \mathbf{y}_t) - \bar{\mathbf{p}}_t\|^2\Big\} \\
&+ \frac{\nu\tau}{2\beta m}\|\mathbf{x}_t - 1\bar{\mathbf{x}}_t\|^2 - \left(\frac{\nu\tau}{m} - \frac{\nu^2 L_J}{2m} - \frac{\nu\beta}{m} - \frac{\nu\tau\beta}{2m}\right)\|\tilde{\mathbf{x}}_t - 1\bar{\mathbf{x}}_t\|^2 - \frac{\nu L_F^2}{2\beta}\|\bar{\mathbf{y}}_t - \mathbf{y}_t^*\|^2 \\
=& \frac{4\nu L_F^2}{\beta\mu\eta\alpha}\Big\{ -\frac{3\eta}{4}\|\tilde{\mathbf{y}}_t - 1\bar{\mathbf{y}}_t\|^2 + \frac{75\eta\alpha}{16\mu}\frac{2}{m}\|\nabla_{\mathbf{y}} F(\mathbf{x}_t, \mathbf{y}_t) - \bar{\mathbf{d}}_t\|^2 \\
&+ \frac{17L_{\mathbf{y}}^2\nu^2}{2\mu m\eta\alpha}\|\tilde{\mathbf{x}}_t - 1\bar{\mathbf{x}}_t\|^2\Big\} + \frac{\nu}{2\beta}\frac{2}{m}\|\nabla_{\mathbf{x}} F(\mathbf{x}_t, \mathbf{y}_t) - \bar{\mathbf{p}}_t\|^2 \\
&+ \frac{\nu\tau}{2\beta m}\|\mathbf{x}_t - 1\bar{\mathbf{x}}_t\|^2 - \left(\frac{\nu\tau}{m} - \frac{\nu^2 L_J}{2m} - \frac{\nu\beta}{m} - \frac{\nu\tau\beta}{2m}\right)\|\tilde{\mathbf{x}}_t - 1\bar{\mathbf{x}}_t\|^2 \\
&+ \big[\frac{\nu}{\beta}\frac{L_F^2}{m} + \frac{4\nu L_F^2}{\beta\mu\eta\alpha}\frac{75\eta\alpha}{16\mu}\frac{2L_F^2}{m}\big]\sum_{i=1}^{m}[\|\bar{\mathbf{x}}_t - \mathbf{x}_{i,t}\|^2 + \|\bar{\mathbf{y}}_t - \mathbf{y}_{i,t}\|^2] - \frac{\nu L_F^2}{2\beta}\|\bar{\mathbf{y}}_t - \mathbf{y}_t^*\|^2, \qquad (70)
\end{aligned}
$$

其中 (a) 是由于 $\|\frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_{i,\ t} - \bar{x}_t\|^2 \leq \frac{1}{m}\sum_{i=1}^{m}\|\mathbf{x}_{i,\ t} - \bar{x}_t\|^2$。

对上述不等式进行望远镜求和，即得所述结果。 □

C.4 引理6的证明

接下来，我们证明以下引理中的迭代收缩性质，这对分析去中心化梯度跟踪算法很有帮助。

证明。 首先，对于迭代值$\mathbf{x}_t$，有如下收缩性质：

$$\|\widetilde{\mathbf{M}}\mathbf{x}_t - 1\bar{\mathbf{x}}_t\|^2 = \|\widetilde{\mathbf{M}}(\mathbf{x}_t - 1\bar{\mathbf{x}}_t)\|^2 \leq \lambda^2\|\mathbf{x}_t - 1\bar{\mathbf{x}}_t\|^2. \tag{71}$$

$\widetilde{\ }$这是因为$\mathbf{x}_t - 1\mathbf{x}_t$ 与$\mathbf{1}$, 正交，而$\mathbf{1}$是与$\mathbf{M}$的最大特征值对应的特征向量, 并且$\lambda = \max\{|\lambda_2|, |\lambda_m|\}$. 因此,

$$\begin{aligned}\|\mathbf{x}_t - 1\bar{\mathbf{x}}_t\|^2 &= \|\widetilde{\mathbf{M}}\mathbf{x}_{t-1} + \nu(\tilde{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}) - 1[\bar{\mathbf{x}}_{t-1} + \nu(\frac{1}{m}\sum_i \tilde{\mathbf{x}}_i - \mathbf{x}_{t-1})]\|^2 \\ &\leq (1+c_1)\lambda^2\|\mathbf{x}_{t-1} - 1\bar{\mathbf{x}}_{t-1}\|^2 + (1+\frac{1}{c_1})\nu^2\|\tilde{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}\|^2. \end{aligned}\tag{72}$$

对于$\mathbf{y}_t$，有

$$\|\mathbf{y}_t - 1\bar{\mathbf{y}}_t\|^2 \leq (1+c_2)\lambda^2\|\mathbf{y}_{t-1} - 1\bar{\mathbf{y}}_{t-1}\|^2 + (1+\frac{1}{c_2})\eta^2\|\tilde{\mathbf{y}}_{t-1} - \bar{\mathbf{y}}_{t-1}\|^2. \tag{73}$$

根据更新规则，有

$$\begin{aligned}&\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 = \|\widetilde{\mathbf{M}}\mathbf{x}_{t-1} + \nu(\tilde{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}) - \mathbf{x}_{t-1}\|^2 \\ =&\|(\widetilde{\mathbf{M}} - \mathbf{I})\mathbf{x}_{t-1} + \nu(\tilde{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1})\|^2 \leq 2\|(\widetilde{\mathbf{M}} - \mathbf{I})\mathbf{x}_{t-1}\|^2 + 2\nu^2\|\tilde{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}\|^2 \\ =&2\|(\widetilde{\mathbf{M}} - \mathbf{I})(\mathbf{x}_{t-1} - 1\bar{\mathbf{x}}_{t-1})\|^2 + 2\nu^2\|\tilde{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}\|^2 \\ \leq&8\|(\mathbf{x}_{t-1} - 1\bar{\mathbf{x}}_{t-1})\|^2 + 2\nu^2\|\tilde{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}\|^2 \\ \leq&8\mathcal{E}(\mathbf{x}_{t-1}) + 2\nu^2\|\tilde{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}\|^2 \end{aligned}\tag{74}$$

并且还有

$$\|\mathbf{y}_t - \mathbf{y}_{t-1}\|^2 \leq 8\mathcal{E}(\mathbf{y}_{t-1}) + 2\eta^2\|\tilde{\mathbf{y}}_{t-1} - \mathbf{y}_{t-1}\|^2 \tag{75}$$

□

Lem引理 8（估计器的微分界）。在假设 *1* 下，以下不等式成立：

$$\sum_{t=1}^T \mathbb{E}\|\mathbf{v}_t - \mathbf{v}_{t-1}\|^2 \leq \sum_{t=1}^T 3L_F^2\mathbb{E}\|\mathbf{x}_{t-1} - \mathbf{x}_t\|^2 + 3L_F^2\mathbb{E}\|\mathbf{y}_{t-1} - \mathbf{y}_t\|^2, \tag{76}$$

$$\sum_{t=1}^T \mathbb{E}\|\mathbf{u}_t - \mathbf{u}_{t-1}\|^2 \leq \sum_{t=1}^T 3L_F^2\mathbb{E}\|\mathbf{x}_{t-1} - \mathbf{x}_t\|^2 + 3L_F^2\mathbb{E}\|\mathbf{y}_{t-1} - \mathbf{y}_t\|^2. \tag{77}$$

证明。对于 $\|\mathbf{v}_t - \mathbf{v}_{t-1}\|^2$，我们有

$$\begin{aligned}&\mathbb{E}\|\mathbf{v}_t - \mathbf{v}_{t-1}\|^2 = \mathbb{E}\|\mathbf{v}_t - \nabla_{\mathbf{x}}\mathbf{F}_t + \nabla_{\mathbf{x}}\mathbf{F}_t - \nabla_{\mathbf{x}}\mathbf{F}_{t-1} + \nabla_{\mathbf{x}}\mathbf{F}_{t-1} - \mathbf{v}_{t-1}\|^2 \\ \leq&3\mathbb{E}\|\mathbf{v}_t - \nabla_{\mathbf{x}}F_t\|^2 + 3\mathbb{E}\|\nabla_{\mathbf{x}}F_t - \nabla_{\mathbf{x}}F_{t-1}\|^2 + 3\mathbb{E}\|\nabla_{\mathbf{x}}F_{t-1} - \mathbf{v}_{t-1}\|^2 \\ \leq&3L_F\mathbb{E}\|\mathbf{x}_{t-1} - \mathbf{x}_t\|^2 + 3L_F^2\mathbb{E}\|\mathbf{y}_{t-1} - \mathbf{y}_t\|^2. \end{aligned}\tag{78}$$

因此，我们有：$\sum_{t=1}^T \mathbb{E}\|\mathbf{v}_t - \mathbf{v}_{t-1}\|^2 \leq \sum_{t=1}^T 3L_F^2\mathbb{E}\|\mathbf{x}_{t-1} - \mathbf{x}_t\|^2 + 3L_F^2\mathbb{E}\|\mathbf{y}_{t-1} - \mathbf{y}_t\|^2$，类似地，$\sum_{t=1}^T \mathbb{E}\|\mathbf{u}_t - \mathbf{u}_{t-1}\|^2 \leq \sum_{t=1}^T 3L_F^2\mathbb{E}\|\mathbf{x}_{t-1} - \mathbf{x}_t\|^2 + 3L_F^2\mathbb{E}\|\mathbf{y}_{t-1} - \mathbf{y}_t\|^2$。 □

C.5 引理7的证明

接下来，我们对梯度估计量的误差进行如下界定：

证明。根据算法更新，我们有：

$$\begin{aligned}
\|\underbrace{\bar{\mathbf{d}}_{i,t}-\nabla_{\boldsymbol{x}}F_{i,t}}_{A_{i,t}}\|^2 &= \|\bar{\mathbf{d}}_{i,t-1}+\frac{1}{|\mathcal{S}_{i,t}|}\sum_{j\in\mathcal{S}_{i,t}}\nabla_{\boldsymbol{x}}f_{i,j}(\boldsymbol{x}_{i,t},\mathbf{y}_{i,t})-\nabla_{\boldsymbol{x}}f_{i,j}(\boldsymbol{x}_{i,t-1},\mathbf{y}_{i,t-1})-\nabla_{\boldsymbol{x}}F_{i,t}\|^2 \\
=\|\underbrace{\bar{\mathbf{d}}_{i,t-1}-\nabla_{\boldsymbol{x}}F_{i,t-1}}_{A_{i,t-1}}&+\underbrace{\frac{1}{|\mathcal{S}_{i,t}|}\sum_{j\in\mathcal{S}_{i,t}}\nabla_{\boldsymbol{x}}f_{i,t}(\boldsymbol{x}_{i,t},\mathbf{y}_{i,t})-\nabla_{\boldsymbol{x}}f_{i,t}(\boldsymbol{x}_{i,t-1},\mathbf{y}_{i,t-1})+\nabla_{\boldsymbol{x}}F_{i,t-1}-\nabla_{\boldsymbol{x}}F_{i,t}}_{B_{i,t}}\|^2 \\
&=\|A_{i,t-1}\|^2+\|B_{i,t}\|^2+2\langle A_{i,t-1},B_{i,t}\rangle. \qquad (79)
\end{aligned}$$

No注意， $\mathbb{E}_t[B_{i,t}]=0$，其中期望是对第 t次迭代中的随机性取的。因此，

$$\mathbb{E}_t\|A_{i,t}\|^2=\|A_{i,t-1}\|^2+\mathbb{E}_t\|B_{i,t}\|^2. \qquad (80)$$

此外，由于$|\mathcal{S}_{i,t}|=q$，我们有

$$\begin{aligned}
\mathbb{E}_t\|B_{i,t}\|^2 &= \mathbb{E}_t\|\frac{1}{|\mathcal{S}_{i,t}|}\sum_{j\in\mathcal{S}_{i,t}}\nabla_{\boldsymbol{x}}f_{i,j}(\boldsymbol{x}_{i,t},\mathbf{y}_{i,t})-\nabla_{\boldsymbol{x}}f_{i,j}(\boldsymbol{x}_{i,t-1},\mathbf{y}_{i,t-1})-\nabla_{\boldsymbol{x}}F_{i,t}+\nabla_{\boldsymbol{x}}F_{i,t-1}\|^2 \\
&\le\frac{1}{|\mathcal{S}_{i,t}|^2}\sum_{j\in\mathcal{S}_{i,t}}\mathbb{E}_t\|\nabla_{\boldsymbol{x}}f_{i,j}(\boldsymbol{x}_{i,t},\mathbf{y}_{i,t})-\nabla_{\boldsymbol{x}}f_{i,j}(\boldsymbol{x}_{i,t-1},\mathbf{y}_{i,t-1})-\nabla_{\boldsymbol{x}}F_{i,t}+\nabla_{\boldsymbol{x}}F_{i,t-1}\|^2 \\
&\le\frac{L_f^2}{q}\big(\|\boldsymbol{x}_{i,t}-\boldsymbol{x}_{i,t-1}\|^2+\|\mathbf{y}_{i,t}-\mathbf{y}_{i,t-1}\|^2\big). \qquad (81)
\end{aligned}$$

对 t 从 $(n_t-1)q+1$ 到 t 取全期望并对 (81) 进行望远镜求和，其中 $t\le n_tq-1$，我们有

$$\begin{aligned}
\mathbb{E}\|A_t\|^2 &\le \mathbb{E}\|A_{t-1}\|^2+\frac{L_f^2}{q}\mathbb{E}\big(\|\boldsymbol{x}_t-\boldsymbol{x}_{t-1}\|^2+\|\mathbf{y}_t-\mathbf{y}_{t-1}\|^2\big) \\
&\le \mathbb{E}\|A_{(n_t-1)q}\|^2+\sum_{r=(n_t-1)q+1}^{t}\frac{L_f^2}{q}\mathbb{E}\big(\|\boldsymbol{x}_r-\boldsymbol{x}_{r-1}\|^2+\|\mathbf{y}_r-\mathbf{y}_{r-1}\|^2\big). \qquad (82)
\end{aligned}$$

因此，我们有：

$$\begin{aligned}
&\sum_{k=0}^{t}\mathbb{E}\|A_k\|^2=\sum_{k=0}^{q-1}\mathbb{E}\|A_k\|^2+\cdots+\sum_{k=(n_t-1)q}^{t}\mathbb{E}\|A_k\|^2 \\
\le&q\|A_0\|^2+\sum_{k=1}^{q-1}\sum_{r=1}^{k}\frac{L_f^2}{q}\big(\|\boldsymbol{x}_r-\boldsymbol{x}_{r-1}\|^2+\|\mathbf{y}_r-\mathbf{y}_{r-1}\|^2\big) \\
&+\cdots \\
&+\big(t-(n_t-1)q\big)\|A_{(n_t-1)q}\|^2+\sum_{k=(n_t-1)q+1}^{t}\sum_{r=(n_t-1)q+1}^{k}\frac{L_f^2}{q}\big(\|\boldsymbol{x}_r-\boldsymbol{x}_{r-1}\|^2+\|\mathbf{y}_r-\mathbf{y}_{r-1}\|^2\big) \\
\le&q\|A_0\|^2+\sum_{r=1}^{q-1}\sum_{k=r}^{q-1}\frac{L_f^2}{q}\big(\|\boldsymbol{x}_r-\boldsymbol{x}_{r-1}\|^2+\|\mathbf{y}_r-\mathbf{y}_{r-1}\|^2\big) \\
&+\cdots \\
&+\big(t-(n_t-1)q\big)\|A_{(n_t-1)q}\|^2+\sum_{r=(n_t-1)q+1}^{t}\sum_{k=r}^{t}\frac{L_f^2}{q}\big(\|\boldsymbol{x}_r-\boldsymbol{x}_{r-1}\|^2+\|\mathbf{y}_r-\mathbf{y}_{r-1}\|^2\big)
\end{aligned}$$

$$
\begin{aligned}
&\leq q\|A_0\|^2+\sum_{r=1}^{q-1}L_f^2\big(\|\boldsymbol{x}_r-\boldsymbol{x}_{r-1}\|^2+\|\mathbf{y}_r-\mathbf{y}_{r-1}\|^2\big)\\
&\quad+\cdots\\
&\quad+\big(t-(n_t-1)q\big)\|A_{(n_t-1)q}\|^2+\sum_{r=(n_t-1)q+1}^{t}L_f^2\big(\|\boldsymbol{x}_r-\boldsymbol{x}_{r-1}\|^2+\|\mathbf{y}_r-\mathbf{y}_{r-1}\|^2\big)\\
&=\sum_{r=0}^{t}\|A_{(n_r-1)q}\|^2+\sum_{r=1}^{t}L_f^2\big(\|\boldsymbol{x}_r-\boldsymbol{x}_{r-1}\|^2+\|\mathbf{y}_r-\mathbf{y}_{r-1}\|^2\big).
\end{aligned}\tag{83}
$$

因此，我们有:

$$
\sum_{t=0}^{T}\|\bar{\mathbf{d}}_t-\nabla_{\boldsymbol{x}}F_t\|^2\leq\sum_{t=0}^{T}\mathbb{E}\|\bar{\mathbf{d}}_{(n_t-1)q}-\nabla_{\boldsymbol{x}}F_{(n_t-1)q})\|^2+\sum_{t=1}^{T}L_f^2\big(\|\boldsymbol{x}_t-\boldsymbol{x}_{t-1}\|^2+\|\mathbf{y}_t-\mathbf{y}_{t-1}\|^2\big)\tag{84}
$$

同理，我们有:

$$
\sum_{t=0}^{T}\|\bar{\mathbf{p}}_t-\nabla_{\mathbf{y}}F_t\|^2\leq\sum_{t=0}^{T}\mathbb{E}\|\bar{\mathbf{p}}_{(n_t-1)q}-\nabla_{\mathbf{y}}F_{(n_t-1)q})\|^2+\sum_{t=1}^{T}L_f^2\big(\|\boldsymbol{x}_t-\boldsymbol{x}_{t-1}\|^2+\|\mathbf{y}_t-\mathbf{y}_{t-1}\|^2\big).\tag{85}
$$

证明至此完成。 □

D 定理 1 和定理 2 的证明

由引理 1–8 和定义的势函数，我们有:

$$
\begin{aligned}
&Q(\bar{\mathbf{x}}_{T+1})-Q(\bar{\mathbf{x}}_0)+\frac{4\nu L_F^2}{\beta}\big[\|\bar{\mathbf{y}}_{T+1}-\mathbf{y}^*_{T+1}\|^2-\|\mathbf{y}^*_0-\bar{\mathbf{y}}_0\|^2\big]\\
&\leq\underbrace{\frac{75\eta\alpha}{16\mu}\frac{2}{m}\sum_{t=0}^{T}\|\nabla_{\mathbf{y}}F(\mathbf{x}_t,\mathbf{y}_t)-\bar{\mathbf{d}}_t\|^2+\frac{\nu}{2\beta}\frac{2}{m}\sum_{t=0}^{T}\|\nabla_{\mathbf{x}}F(\mathbf{x}_t,\mathbf{y}_t)-\bar{\mathbf{p}}_t\|^2}_{R_1}-\frac{\nu {L_F}^2}{2}\sum_{t=0}^{T}\|\bar{\mathbf{y}}_t-\mathbf{y}^*_t\|^2\\
&\quad+\frac{\nu\tau}{2\beta m}\sum_{t=0}^{T}\|\mathbf{x}_t-1\bar{\mathbf{x}}_t\|^2+\big[\frac{\nu}{\beta}\frac{L_F^2}{m}+\frac{4\nu L_F^2}{\beta\mu\eta\alpha}\frac{75\eta\alpha}{16\mu}\frac{2L_F^2}{m}\big]\sum_{t=0}^{T}\sum_{i=1}^{m}[\|\bar{\mathbf{x}}_t-\mathbf{x}_{i,t}\|^2+\|\bar{\mathbf{y}}_t-\mathbf{y}_{i,t}\|^2]\\
&\quad-\left(-\frac{17L_{\mathbf{y}}^2\nu^2}{2\mu m\eta\alpha}+\frac{\nu\tau}{m}-\frac{\nu^2L_J}{2m}-\frac{\nu\beta}{m}-\frac{\nu\tau\beta}{2m}\right)\sum_{t=0}^{T}\|\tilde{\mathbf{x}}_t-1\bar{\mathbf{x}}_t\|^2-\frac{4\nu L_F^2}{\beta\mu\eta\alpha}\frac{3\eta}{4}\|\tilde{\mathbf{y}}_t-1\bar{\mathbf{y}}_t\|^2,
\end{aligned}\tag{86}
$$

由定义的势函数 $\mathfrak{p}$，我们有

$$
\begin{aligned}
\mathbb{E}\mathfrak{p}_{T+1}-\mathfrak{p}_0\leq&\underbrace{\frac{75\eta\alpha}{16\mu}\frac{2}{m}\sum_{t=0}^{T}\|\nabla_{\mathbf{y}}F(\boldsymbol{x}_t,\mathbf{y}_t)-\bar{\mathbf{d}}_t\|^2+\frac{\nu}{2\beta}\frac{2}{m}\sum_{t=0}^{T}\|\nabla_{\boldsymbol{x}}F(\boldsymbol{x}_t,\mathbf{y}_t)-\bar{\mathbf{p}}_t\|^2}_{R_1}-\frac{\nu {L_f}^2}{2}\sum_{t=0}^{T}\|\bar{\mathbf{y}}_t-\mathbf{y}^*_t\|^2\\
&+\frac{\nu\tau}{2\beta m}\sum_{t=0}^{T}\|\boldsymbol{x}_t-1\bar{\boldsymbol{x}}_t\|^2+\big[\frac{\nu}{\beta}\frac{L_f^2}{m}+\frac{4\nu L_f^2}{\beta\mu\eta\alpha}\frac{75\eta\alpha}{16\mu}\frac{2L_f^2}{m}\big]\sum_{t=0}^{T}\sum_{i=1}^{m}[\|\bar{\boldsymbol{x}}_t-\boldsymbol{x}_{i,t}\|^2+\|\bar{\mathbf{y}}_t-\mathbf{y}_{i,t}\|^2]\\
&-\left(-\frac{17L_{\mathbf{y}}^2\nu^2}{2\mu m\eta\alpha}+\frac{\nu\tau}{m}-\frac{\nu^2L_J}{2m}-\frac{\nu\beta}{m}-\frac{\nu\tau\beta}{2m}\right)\sum_{t=0}^{T}\|\tilde{\boldsymbol{x}}_t-1\bar{\boldsymbol{x}}_t\|^2-\frac{4\nu L_f^2}{\beta\mu\eta\alpha}\frac{3\eta}{4}\sum_{t=0}^{T}\|\widetilde{\mathbf{y}}_{t+1}-\bar{\mathbf{y}}_t\|^2,
\end{aligned}\tag{87}
$$

对于术语 R_1，我们有

$$
\frac{75\eta\alpha}{16\mu}\frac{2}{m}\sum_{t=0}^{T}\mathbb{E}\|\nabla_{\mathbf{y}}F_t-\bar{\mathbf{d}}_t\|^2+\frac{\nu}{2\beta}\frac{2}{m}\sum_{t=0}^{T}\mathbb{E}\|\nabla_{\boldsymbol{x}}F_t-\bar{\mathbf{p}}_t\|^2
$$

$$
\begin{aligned}
&\leq \frac{75\eta\alpha}{16\mu}\frac{2}{m}\mathbb{E}\Big(\sum_{t=0}^{T}\|\bar{\mathbf{d}}_{(n_t-1)q}-\nabla_{\boldsymbol{x}}F_{(n_t-1)q}\|^2+\sum_{t=1}^{T}L_f^2\big(\|\boldsymbol{x}_t-\boldsymbol{x}_{t-1}\|^2+\|\mathbf{y}_t-\mathbf{y}_{t-1}\|^2\big)\Big)\\
&\quad+\frac{\nu}{2\beta}\frac{2}{m}\mathbb{E}\Big(\sum_{t=0}^{T}\|\bar{\mathbf{p}}_{(n_t-1)q}-\nabla_{\mathbf{y}}F_{(n_t-1)q}\|^2+\sum_{t=1}^{T}L_f^2\big(\|\boldsymbol{x}_t-\boldsymbol{x}_{t-1}\|^2+\|\mathbf{y}_t-\mathbf{y}_{t-1}\|^2\big)\Big)\\
&=L_f^2\big(\frac{75\eta\alpha}{16\mu}\frac{2}{m}+\frac{\nu}{2\beta}\frac{2}{m}\big)\sum_{t=1}^{T}\mathbb{E}\big(\|\boldsymbol{x}_t-\boldsymbol{x}_{t-1}\|^2+\|\mathbf{y}_t-\mathbf{y}_{t-1}\|^2\big)\\
&\quad+\frac{75\eta\alpha}{16\mu}\frac{2}{m}\sum_{t=0}^{T}\mathbb{E}\|\bar{\mathbf{d}}_{(n_t-1)q}-\nabla_{\boldsymbol{x}}F_{(n_t-1)q}\|^2+\frac{\nu}{2\beta}\frac{2}{m}\sum_{t=0}^{T}\mathbb{E}\|\bar{\mathbf{p}}_{(n_t-1)q}-\nabla_{\mathbf{y}}F_{(n_t-1)q}\|^2. \qquad (88)
\end{aligned}
$$

代入上述结果，我们得到

$$
\begin{aligned}
\mathbb{E}\mathfrak{p}_{T+1}-\mathfrak{p}_0 \leq&-\frac{\nu {L_f}^2}{2}\sum_{t=0}^{T}\|\bar{\mathbf{y}}_t-\mathbf{y}_t^*\|^2+\frac{\nu\tau}{2\beta m}\sum_{t=0}^{T}\|\boldsymbol{x}_t-1\bar{\boldsymbol{x}}_t\|^2\\
&-\big[1-(1+c_1)\lambda^2-\frac{\nu}{\beta}\frac{L_f^2}{m}-\frac{4\nu L_f^2}{\beta\mu\eta\alpha}\frac{75\eta\alpha}{16\mu}\frac{2L_f^2}{m}\big]\sum_{t=0}^{T}\sum_{i=1}^{m}[\|\bar{\boldsymbol{x}}_t-\boldsymbol{x}_{i,t}\|^2]\\
&-\big[1-(1+c_2)\lambda^2-\frac{\nu}{\beta}\frac{L_f^2}{m}-\frac{4\nu L_f^2}{\beta\mu\eta\alpha}\frac{75\eta\alpha}{16\mu}\frac{2L_f^2}{m}\big]\sum_{t=0}^{T}\sum_{i=1}^{m}[\|\bar{\mathbf{y}}_t-\mathbf{y}_{i,t}\|^2]\\
&-\left(-(1+\frac{1}{c_1})\nu^2-\frac{17L_{\mathbf{y}}^2\nu^2}{2\mu m\eta\alpha}+\frac{\nu\tau}{m}-\frac{\nu^2L_J}{2m}-\frac{\nu\beta}{m}-\frac{\nu\tau\beta}{2m}\right)\sum_{t=0}^{T}\|\tilde{\boldsymbol{x}}_t-1\bar{\boldsymbol{x}}_t\|^2\\
&-[\frac{4\nu L_f^2}{\beta\mu\eta\alpha}\frac{3\eta}{4}-(1+\frac{1}{c_2})\eta^2]\sum_{t=0}^{T}\|\widetilde{\mathbf{y}}_{t+1}-\bar{\mathbf{y}}_t\|^2\\
&+L_f^2\big(\frac{75\eta\alpha}{16\mu}\frac{2}{m}+\frac{\nu}{2\beta}\frac{2}{m}\big)\sum_{t=1}^{T}\mathbb{E}\big(\|\boldsymbol{x}_t-\boldsymbol{x}_{t-1}\|^2+\|\mathbf{y}_t-\mathbf{y}_{t-1}\|^2\big)\\
&+\frac{75\eta\alpha}{16\mu}\frac{2}{m}\sum_{t=0}^{T}\mathbb{E}\|\bar{\mathbf{d}}_{(n_t-1)q}-\nabla_{\boldsymbol{x}}F_{(n_t-1)q}\|^2+\frac{\nu}{2\beta}\frac{2}{m}\sum_{t=0}^{T}\mathbb{E}\|\bar{\mathbf{p}}_{(n_t-1)q}-\nabla_{\mathbf{y}}F_{(n_t-1)q}\|^2\\
=&-\frac{\nu {L_f}^2}{2}\sum_{t=0}^{T}\|\bar{\mathbf{y}}_t-\mathbf{y}_t^*\|^2\\
&-\underbrace{\big[1-8L_f^2\big(\frac{75\eta\alpha}{16\mu}\frac{2}{m}+\frac{\nu}{2\beta}\frac{2}{m}\big)-\frac{\nu\tau}{2\beta m}-(1+c_1)\lambda^2-\frac{\nu}{\beta}\frac{L_f^2}{m}-\frac{4\nu L_f^2}{\beta\mu\eta^2}\frac{75\eta\alpha}{16\mu}\frac{2L_f^2}{m}\big]}_{C_1}\sum_{t=0}^{T}\sum_{i=1}^{m}[\|\bar{\boldsymbol{x}}_t-\boldsymbol{x}_{i,t}\|^2]\\
&-\underbrace{\left(-2\nu^2L_f^2\big(\frac{75\eta\alpha}{16\mu}\frac{2}{m}+\frac{\nu}{2\beta}\frac{2}{m}\big)-(1+\frac{1}{c_1})\nu^2-\frac{17L_{\mathbf{y}}^2\nu^2}{2\mu m\eta\alpha}+\frac{\nu\tau}{m}-\frac{\nu^2L_J}{2m}-\frac{\nu\beta}{m}-\frac{\nu\tau\beta}{2m}\right)}_{C_2}\sum_{t=0}^{T}\|\tilde{\boldsymbol{x}}_t-1\bar{\boldsymbol{x}}_t\|^2\\
&-\underbrace{\big[1-8L_f^2\big(\frac{75\eta\alpha}{16\mu}\frac{2}{m}+\frac{\nu}{2\beta}\frac{2}{m}\big)-(1+c_2)\lambda^2-\frac{\nu}{\beta}\frac{L_f^2}{m}-\frac{4\nu L_f^2}{\beta\mu\eta\alpha}\frac{75\eta\alpha}{16\mu}\frac{2L_f^2}{m}\big]}_{C_3}\sum_{t=0}^{T}\sum_{i=1}^{m}[\|\bar{\mathbf{y}}_t-\mathbf{y}_{i,t}\|^2]\\
&-\underbrace{[\frac{4\nu L_f^2}{\beta\mu\eta\alpha}\frac{3\eta}{4}-(1+\frac{1}{c_2})\eta^2-2\eta\alpha L_f^2\big(\frac{75\eta\alpha}{16\mu}\frac{2}{m}+\frac{\nu}{2\beta}\frac{2}{m}\big)]}_{C_4}\sum_{t=0}^{T}\|\widetilde{\mathbf{y}}_t-\bar{\mathbf{y}}_t\|^2 \qquad (89)
\end{aligned}
$$

对于PRECISION，外循环计算全梯度。因此，我们有 $\mathbb{E}\|\bar{}\mathbf{d}_{(n_t-1)q}-\nabla_{\boldsymbol{x}}F_{(n_t-1)q}\|^2=\mathbb{E}\|\bar{}\mathbf{p}_{(n_t-1)q}-$

$\nabla_{\mathbf{y}} F_{(n_t-1)q}\|^2 = 0$。

取 $c_1 = c_2 = \frac{1-\lambda^2}{1+\lambda^2}$，我们有

$$\begin{aligned} C_1 &= \big[1 - 8L_f^2\big(\frac{75\eta\alpha}{16\mu}\frac{2}{m} + \frac{\nu}{2\beta}\frac{2}{m}\big) - \frac{\nu\tau}{2\beta m} - (1+c_1)\lambda^2 - \frac{\nu}{\beta}\frac{L_f^2}{m} - \frac{4\nu L_f^2}{\beta\mu\eta\alpha}\frac{75\eta\alpha}{16\mu}\frac{2L_f^2}{m}\big] \\ &\geq \big[1 - 8L_f^2\big(\frac{75\eta\alpha}{16\mu}\frac{2}{m} + \frac{\nu}{2\beta}\frac{2}{m}\big) - \frac{\nu\tau}{2\beta m} - (1-c_1) - \frac{\nu}{\beta}\frac{L_f^2}{m} - \frac{4\nu L_f^2}{\beta\mu\eta\alpha}\frac{75\eta\alpha}{16\mu}\frac{2L_f^2}{m}\big] \\ &\geq c_1 - \frac{c_1}{5} - \frac{c_1}{5} - \frac{c_1}{5} - \frac{c_1}{5} - \frac{c_1}{5} = 0 \end{aligned} \tag{90}$$

$$\begin{aligned} C_2 &= \left(-2\nu^2 L_f^2\big(\frac{75\eta\alpha}{16\mu}\frac{2}{m} + \frac{\nu}{2\beta}\frac{2}{m}\big) - (1+\frac{1}{c_1})\nu^2 - \frac{17L_{\mathbf{y}}^2\nu^2}{2\mu m\eta\alpha} + \frac{\nu\tau}{m} - \frac{\nu^2 L_J}{2m} - \frac{\nu\beta}{m} - \frac{\nu\tau\beta}{2m}\right) \\ &\geq -\frac{\nu\tau}{6m} - \frac{\nu\tau}{6m} - \frac{\nu\tau}{6m} + \frac{\nu\tau}{m} - \frac{\nu\tau}{6m} - \frac{\nu\tau}{12m} - \frac{\nu\tau}{6m} > 0 \end{aligned} \tag{91}$$

$$\begin{aligned} C_3 =& \big[1 - 8L_f^2\big(\frac{75\eta\alpha}{16\mu}\frac{2}{m} + \frac{\nu}{2\beta}\frac{2}{m}\big) - (1+c_2)\lambda^2 - \frac{\nu}{\beta}\frac{L_f^2}{m} - \frac{4\nu L_f^2}{\beta\mu\eta\alpha}\frac{75\eta\alpha}{16\mu}\frac{2L_f^2}{m}\big] \\ &\geq \big[1 - 8L_f^2\big(\frac{75\eta\alpha}{16\mu}\frac{2}{m} + \frac{\nu}{2\beta}\frac{2}{m}\big) - (1-c_2) - \frac{\nu}{\beta}\frac{L_f^2}{m} - \frac{4\nu L_f^2}{\beta\mu\eta\alpha}\frac{75\eta\alpha}{16\mu}\frac{2L_f^2}{m}\big] \\ &\geq c_2 - \frac{c_2}{5} - \frac{c_2}{5} - \frac{c_2}{5} - \frac{c_2}{5} - \frac{c_2}{5} = 0 \end{aligned} \tag{92}$$

$$C_4 = [\frac{4\nu L_f^2}{\beta\mu\eta\alpha}\frac{3\eta}{4} - (1+\frac{1}{c_2})\eta^2 - 2\eta\alpha L_f^2\big(\frac{75\eta\alpha}{16\mu}\frac{2}{m} + \frac{\nu}{2\beta}\frac{2}{m}\big)] \geq \frac{4\nu L_f^2}{\beta\mu\eta\alpha}\frac{3\eta}{4} - \frac{4\nu L_f^2}{\beta\mu\eta\alpha}\frac{3\eta}{8} - \frac{4\nu L_f^2}{\beta\mu\eta\alpha}\frac{3\eta}{8} = 0 \tag{93}$$

对于参数

$$\begin{aligned} \eta &\leq \min\{\frac{c_1 m\mu}{375\alpha L_f^2}, \frac{15L_f^2}{\beta\mu\alpha^2 c_1}, \frac{3c_1^2 m}{10(1+c_1)\mu\alpha}\} \\ \nu &\leq \min\{\frac{c_1 m\beta}{40L_f^2}, \frac{2c_1 m\beta}{5\tau}, \frac{2c_1\beta\mu^2 m}{375L_f^4}, \frac{5\tau}{3mc_1}, \frac{\tau}{6m(1+1/c_1)}, \frac{3\mu\eta\alpha\tau}{17L_f^2}, \frac{\tau}{3(L_f + \frac{L_f^2}{\mu})}\} \\ \beta &\leq \min\{\frac{\tau}{12}, \frac{1}{3}\}, \end{aligned} \tag{94}$$

我们得到 PRECISION 的上述结果:

$$\sum_{t=0}^{T}\Big(\mathbf{E}[\|\tilde{\boldsymbol{x}}_t - 1\bar{\boldsymbol{x}}_t\|^2 + \|\boldsymbol{x}_t - 1\bar{\boldsymbol{x}}_t\|^2] + \mathbb{E}\|\mathbf{y}_t^* - \bar{\mathbf{y}}_t\|^2\Big) \leq \frac{\mathbb{E}[\mathfrak{p}_0 - \mathfrak{p}_{T+1}]}{\min\{C_1, C_2, \nu L_f^2/2\}}. \tag{95}$$

对于 PRECISION$^+$，我们有

$$\mathbb{E}\|\bar{\mathbf{d}}_{(n_t-1)q} - \nabla_{\boldsymbol{x}} F_{(n_t-1)q}\|^2 = \mathbb{E}\|\bar{\mathbf{p}}_{(n_t-1)q} - \nabla_{\mathbf{y}} F_{(n_t-1)q}\|^2 = \frac{I_{(\mathcal{N}_s<M)}}{\mathcal{N}_s}\sigma^2 \tag{96}$$

回顾一下，$\mathcal{N}_s = \min\{c_\gamma\sigma^2(\gamma^{(k)})^{-1}, c_\epsilon\sigma^2\epsilon^{-1}, M\}$。那么我们有

$$\begin{aligned} \frac{I_{(\mathcal{N}_s<M)}}{\mathcal{N}_s} &\leq \frac{1}{\min\{c_\epsilon\sigma^2(\epsilon)^{-1}, c_\gamma\sigma^2(\gamma^{(k)})^{-1}\}} \\ &= \max\{\frac{\gamma^{(k)}}{c_\gamma\sigma^2}, \frac{\epsilon}{c_\epsilon\sigma^2}\} \leq \frac{\gamma^{(k)}}{c_\gamma\sigma^2} + \frac{\epsilon}{c_\epsilon\sigma^2}. \end{aligned} \tag{97}$$

因此，我们有

$$
\begin{aligned}
&Q(\bar{\boldsymbol{x}}_{T+1})-Q(\bar{\boldsymbol{x}}_0)+\frac{4\nu L_F^2}{\beta}\big[\|\bar{\mathbf{y}}_{T+1}-\mathbf{y}_{T+1}^*\|^2-\|\mathbf{y}_0^*-\bar{\mathbf{y}}_0\|^2\big]\\
\le&\underbrace{\frac{75\eta\alpha}{16\mu}\frac{2}{m}\sum_{t=0}^{T}\|\nabla_{\mathbf{y}}F(\boldsymbol{x}_t,\mathbf{y}_t)-\bar{\mathbf{d}}_t\|^2+\frac{\nu}{2\beta}\frac{2}{m}\sum_{t=0}^{T}\|\nabla_{\boldsymbol{x}}F(\boldsymbol{x}_t,\mathbf{y}_t)-\bar{\mathbf{p}}_t\|^2}_{R_1}-\frac{\nu {L_F}^2}{2}\sum_{t=0}^{T}\|\bar{\mathbf{y}}_t-\mathbf{y}_t^*\|^2\\
&+\frac{\nu\tau}{2\beta m}\sum_{t=0}^{T}\|\boldsymbol{x}_t-1\bar{\boldsymbol{x}}_t\|^2+\big[\frac{\nu}{\beta}\frac{L_F^2}{m}+\frac{4\nu L_F^2}{\beta\mu\eta\alpha}\frac{75\eta\alpha}{16\mu}\frac{2L_F^2}{m}\big]\sum_{t=0}^{T}\sum_{i=1}^{m}[\|\bar{\boldsymbol{x}}_t-\boldsymbol{x}_{i,t}\|^2+\|\bar{\mathbf{y}}_t-\mathbf{y}_{i,t}\|^2]\\
&-\left(-\frac{17L_{\mathbf{y}}^2\nu^2}{2\mu m\eta\alpha}+\frac{\nu\tau}{m}-\frac{\nu^2L_J}{2m}-\frac{\nu\beta}{m}-\frac{\nu\tau\beta}{2m}\right)\sum_{t=0}^{T}\|\tilde{\boldsymbol{x}}_t-1\bar{\boldsymbol{x}}_t\|^2-\frac{4\nu L_F^2}{\beta\mu\eta\alpha}\frac{3\eta}{4}\|\tilde{\mathbf{y}}_t-1\bar{\mathbf{y}}_t\|^2\\
\le&(\frac{75\eta\alpha}{16\mu}\frac{2}{m}+\frac{\nu}{2\beta}\frac{2}{m})\sum_{t=0}^{T}(\frac{\gamma^{(t)}}{c_\gamma}+\frac{\epsilon}{c_\epsilon})-\frac{\nu {L_F}^2}{2}\sum_{t=0}^{T}\|\bar{\mathbf{y}}_t-\mathbf{y}_t^*\|^2\\
&+\frac{\nu\tau}{2\beta m}\sum_{t=0}^{T}\|\boldsymbol{x}_t-1\bar{\boldsymbol{x}}_t\|^2+\big[\frac{\nu}{\beta}\frac{L_F^2}{m}+\frac{4\nu L_F^2}{\beta\mu\eta\alpha}\frac{75\eta\alpha}{16\mu}\frac{2L_F^2}{m}\big]\sum_{t=0}^{T}\sum_{i=1}^{m}[\|\bar{\boldsymbol{x}}_t-\boldsymbol{x}_{i,t}\|^2+\|\bar{\mathbf{y}}_t-\mathbf{y}_{i,t}\|^2]\\
&-\left(-\frac{17L_{\mathbf{y}}^2\nu^2}{2\mu m\eta\alpha}+\frac{\nu\tau}{m}-\frac{\nu^2L_J}{2m}-\frac{\nu\beta}{m}-\frac{\nu\tau\beta}{2m}\right)\sum_{t=0}^{T}\|\tilde{\boldsymbol{x}}_t-1\bar{\boldsymbol{x}}_t\|^2-\frac{4\nu L_F^2}{\beta\mu\eta\alpha}\frac{3\eta}{4}\|\tilde{\mathbf{y}}_t-1\bar{\mathbf{y}}_t\|^2
\end{aligned}
\tag{98}
$$

由于 $\gamma_{t+1}=\frac{1}{q}\sum_{i=(n_k-1)q}^{k}\|\tilde{\boldsymbol{x}}_t-1\bar{\boldsymbol{x}}_t\|^2$。

$$
\begin{aligned}
&\mathbb{E}\mathfrak{p}_{T+1}-\mathfrak{p}_0\le-\frac{\nu {L_f}^2}{2}\sum_{t=0}^{T}\|\bar{\mathbf{y}}_t-\mathbf{y}_t^*\|^2\\
&\underbrace{-\big[1-8L_f^2\big(\frac{75\eta\alpha}{16\mu}\frac{2}{m}+\frac{\nu}{2\beta}\frac{2}{m}\big)-\frac{\nu\tau}{2\beta m}-(1+c_1)\lambda^2-\frac{\nu}{\beta}\frac{L_f^2}{m}-\frac{4\nu L_f^2}{\beta\mu\eta^2}\frac{75\eta\alpha}{16\mu}\frac{2L_f^2}{m}\big]}_{C_1}\sum_{t=0}^{T}\sum_{i=1}^{m}[\|\bar{\boldsymbol{x}}_t-\boldsymbol{x}_{i,t}\|^2]\\
&\underbrace{-\left(c_\gamma(\frac{75\eta\alpha}{16\mu}\frac{2}{m}+\frac{\nu}{2\beta}\frac{2}{m})-2\nu^2L_f^2\big(\frac{75\eta\alpha}{16\mu}\frac{2}{m}+\frac{\nu}{2\beta}\frac{2}{m}\big)-(1+\frac{1}{c_1})\nu^2-\frac{17L_{\mathbf{y}}^2\nu^2}{2\mu m\eta\alpha}+\frac{\nu\tau}{m}-\frac{\nu^2L_J}{2m}-\frac{\nu\beta}{m}-\frac{\nu\tau\beta}{2m}\right)}_{C_2}\sum_{t=0}^{T}\|\tilde{\boldsymbol{x}}_t-1\bar{\boldsymbol{x}}_t\|^2\\
&\underbrace{-\big[1-8L_f^2\big(\frac{75\eta\alpha}{16\mu}\frac{2}{m}+\frac{\nu}{2\beta}\frac{2}{m}\big)-(1+c_2)\lambda^2-\frac{\nu}{\beta}\frac{L_f^2}{m}-\frac{4\nu L_f^2}{\beta\mu\eta\alpha}\frac{75\eta\alpha}{16\mu}\frac{2L_f^2}{m}\big]}_{C_3}\sum_{t=0}^{T}\sum_{i=1}^{m}[\|\bar{\mathbf{y}}_t-\mathbf{y}_{i,t}\|^2]\\
&\underbrace{-[\frac{4\nu L_f^2}{\beta\mu\eta\alpha}\frac{3\eta}{4}-(1+\frac{1}{c_2})\eta^2-2\eta\alpha L_f^2\big(\frac{75\eta\alpha}{16\mu}\frac{2}{m}+\frac{\nu}{2\beta}\frac{2}{m}\big)]}_{C_4}\sum_{t=0}^{T}\|\tilde{\mathbf{y}}_t-\bar{\mathbf{y}}_t\|^2
\end{aligned}
\tag{99}
$$

选择 $c_1=c_2=\frac{2\lambda_m\ (M)\ \lambda_+}{}$ 我们有

$$
\begin{aligned}
C_1&=\big[1-8L_f^2\big(\frac{75\eta\alpha}{16\mu}\frac{2}{m}+\frac{\nu}{2\beta}\frac{2}{m}\big)-\frac{\nu\tau}{2\beta m}-(1+c_1)\lambda^2-\frac{\nu}{\beta}\frac{L_f^2}{m}-\frac{4\nu L_f^2}{\beta\mu\eta\alpha}\frac{75\eta\alpha}{16\mu}\frac{2L_f^2}{m}\big]\\
&\ge\big[1-8L_f^2\big(\frac{75\eta\alpha}{16\mu}\frac{2}{m}+\frac{\nu}{2\beta}\frac{2}{m}\big)-\frac{\nu\tau}{2\beta m}-(1-c_1)-\frac{\nu}{\beta}\frac{L_f^2}{m}-\frac{4\nu L_f^2}{\beta\mu\eta\alpha}\frac{75\eta\alpha}{16\mu}\frac{2L_f^2}{m}\big]
\end{aligned}
$$

$$\geq c_1 - \frac{c_1}{5} - \frac{c_1}{5} - \frac{c_1}{5} - \frac{c_1}{5} - \frac{c_1}{5} = 0 \tag{100}$$

$$\begin{aligned} C_2' &= \left(c_\gamma (\frac{75\eta\alpha}{16\mu}\frac{2}{m} + \frac{\nu}{2\beta}\frac{2}{m}) - 2\nu^2 L_f^2 (\frac{75\eta\alpha}{16\mu}\frac{2}{m} + \frac{\nu}{2\beta}\frac{2}{m}) - (1+\frac{1}{c_1})\nu^2 - \frac{17L_{\mathbf{y}}^2\nu^2}{2\mu m\eta\alpha} + \frac{\nu\tau}{m} - \frac{\nu^2 L_J}{2m} - \frac{\nu\beta}{m} - \frac{\nu\tau\beta}{2m} \right) \\ &\geq -\frac{\nu\tau}{12m} - \frac{\nu\tau}{6m} - \frac{\nu\tau}{6m} - \frac{\nu\tau}{6m} + \frac{\nu\tau}{m} - \frac{\nu\tau}{6m} - \frac{\nu\tau}{12m} - \frac{\nu\tau}{6m} = 0 \end{aligned} \tag{101}$$

$$\begin{aligned} C_3 =& \big[1 - 8L_f^2 (\frac{75\eta\alpha}{16\mu}\frac{2}{m} + \frac{\nu}{2\beta}\frac{2}{m}) - (1+c_2)\lambda^2 - \frac{\nu}{\beta}\frac{L_f^2}{m} - \frac{4\nu L_f^2}{\beta\mu\eta\alpha}\frac{75\eta\alpha}{16\mu}\frac{2L_f^2}{m}\big] \\ &\geq \big[1 - 8L_f^2 (\frac{75\eta\alpha}{16\mu}\frac{2}{m} + \frac{\nu}{2\beta}\frac{2}{m}) - (1-c_2) - \frac{\nu}{\beta}\frac{L_f^2}{m} - \frac{4\nu L_f^2}{\beta\mu\eta\alpha}\frac{75\eta\alpha}{16\mu}\frac{2L_f^2}{m}\big] \\ &\geq c_2 - \frac{c_2}{5} - \frac{c_2}{5} - \frac{c_2}{5} - \frac{c_2}{5} - \frac{c_2}{5} = 0 \end{aligned} \tag{102}$$

$$C_4 = [\frac{4\nu L_f^2}{\beta\mu\eta\alpha}\frac{3\eta}{4} - (1+\frac{1}{c_2})\eta^2 - 2\eta\alpha L_f^2 (\frac{75\eta\alpha}{16\mu}\frac{2}{m} + \frac{\nu}{2\beta}\frac{2}{m})] \geq \frac{4\nu L_f^2}{\beta\mu\eta\alpha}\frac{3\eta}{4} - \frac{4\nu L_f^2}{\beta\mu\eta\alpha}\frac{3\eta}{8} - \frac{4\nu L_f^2}{\beta\mu\eta\alpha}\frac{3\eta}{8} = 0 \tag{103}$$

使用参数

$$\begin{aligned} &c_\gamma \geq (\frac{75\eta\alpha}{8\mu}\frac{1}{m} + \frac{\nu}{\beta}\frac{1}{m})\frac{\nu\tau}{12} \\ &\eta \leq \min\{\frac{c_1 m\mu}{375\alpha L_f^2}, \frac{15L_f^2}{\beta\mu\alpha^2 c_1}, \frac{3c_1^2 m}{10(1+c_1)\mu\alpha}\} \\ &\nu \leq \min\{\frac{c_1 m\beta}{40L_f^2}, \frac{2c_1 m\beta}{5\tau}, \frac{2c_1\beta\mu^2 m}{375L_f^4}, \frac{5\tau}{3mc_1}, \frac{\tau}{6m(1+1/c_1)}, \frac{3\mu\eta\alpha\tau}{17L_f^2}, \frac{\tau}{3(L_f + \frac{L_f^2}{\mu})}\} \\ &\beta \leq \min\{\frac{\tau}{12}, \frac{1}{3}\}, \end{aligned} \tag{104}$$

因此，对于 PRECISION$^+$，我们有以下收敛结果:

$$\begin{aligned} &\frac{1}{(T+1)}\sum_{t=0}^{T}\left(\mathbf{E}[\|\tilde{\boldsymbol{x}}_t - 1\bar{\boldsymbol{x}}_t\|^2 + \|\boldsymbol{x}_t - 1\bar{\boldsymbol{x}}_t\|^2] + \mathbb{E}\|\mathbf{y}_t^* - \bar{\mathbf{y}}_t\|^2\right) \\ \leq& \frac{\mathbb{E}[\mathfrak{p}_0 - \mathfrak{p}_{T+1}]}{(T+1)\min\{C_1, C_2', \nu L_f^2/2\}} + (\frac{75\eta\alpha}{16\mu}\frac{2}{m} + \frac{\nu}{2\beta}\frac{2}{m})\frac{\epsilon}{c_\epsilon}. \end{aligned} \tag{105}$$

借助 $\mathfrak{p}_{T+1} \geq Q^*$，我们得出结论。

E 辅助引理

引理 9. 在假设 *1* 下， $\boldsymbol{y}^*(\boldsymbol{x}) = \arg\max_{\boldsymbol{y}} F(\boldsymbol{x}, \boldsymbol{y})$ 是利普希茨连续的，即存在一个正数常量 $L_{\boldsymbol{y}}$，使得

$$\|\boldsymbol{y}^*(\boldsymbol{x}) - \boldsymbol{y}^*(\boldsymbol{x}')\| \leq L_{\boldsymbol{y}}\|\boldsymbol{x} - \boldsymbol{x}'\|, \ \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d, \tag{106}$$

其中，利普希茨常数为 $L_{\boldsymbol{y}} = L_f/\mu$。

证明。 参见 Lin 等人的引理 4.3。 [2020a]。 □

引理 10. 在假设 *1* 下，函数 $J(\boldsymbol{x}) = F(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x}))$ 满足 $\nabla J(\boldsymbol{x}) = \nabla_{\boldsymbol{x}} F(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x}))$。

证明. 由于 $J(\boldsymbol{x}) = F(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x}))$，根据链式法则，我们有

$$dJ(\boldsymbol{x}) = \frac{\partial F(\boldsymbol{x}, \boldsymbol{y})}{\partial \boldsymbol{x}}\Big|_{\boldsymbol{y}=\boldsymbol{y}^*(\boldsymbol{x})} \cdot d\boldsymbol{x} + \frac{\partial F(\boldsymbol{x}, \boldsymbol{y})}{\partial \boldsymbol{y}}\Big|_{\boldsymbol{y}=\boldsymbol{y}^*(\boldsymbol{x})} \cdot \frac{\partial \omega^*(\boldsymbol{x})}{\partial \boldsymbol{x}} \cdot d\boldsymbol{x}, \tag{107}$$

其中，$\partial F(\boldsymbol{x}, \boldsymbol{y})/\partial \boldsymbol{x}$ 和 $\partial F(\boldsymbol{x}, \boldsymbol{y})/\partial \boldsymbol{y}$ 分别是 F 关于第一变量 $\boldsymbol{x}$ 和第二变量 $\boldsymbol{y}$ 的偏导数。注意，$\boldsymbol{y}^*(\boldsymbol{x})$ 是使 $F(\boldsymbol{x}, \boldsymbol{y})$ 达到最大值的唯一最优点。因此，对于所有 $\boldsymbol{x}$，有 $\frac{\partial F(}{\partial \boldsymbol{y}}\Big|_{\boldsymbol{y}=\boldsymbol{y}^*, \ )}(\boldsymbol{x}) = 0$。此外，根据引理 9，$\partial \omega^*(\boldsymbol{x})/\partial \boldsymbol{x}$ 有界。因此可得

$$dJ(\boldsymbol{x}) = \frac{\partial F(\boldsymbol{x}, \boldsymbol{y})}{\partial \boldsymbol{x}}\Big|_{\boldsymbol{y}=\boldsymbol{y}^*(\boldsymbol{x})} \cdot d\boldsymbol{x}, \tag{108}$$

即 $\nabla J(\boldsymbol{x}) = \nabla_{\boldsymbol{x}} F(\boldsymbol{x}, \ \boldsymbol{y}^*(\boldsymbol{x}))$.

□