# 循环客户端参与下通信高效的联邦 AUC 最大化

**Umesh Vangapally**[*] *Z2008841@students.niu.edu*
**Wenhan Wu**[†] *wwu25@uncc.edu*
**Chen Chen**[‡] *chen.chen@crcv.ucf.edu*
**Zhishuai Guo**[*] *zguo@niu.edu*
[*]计算机科学系，北伊利诺伊大学
[†]计算机科学系，北卡罗来纳大学夏洛特分校
[‡]中佛罗里达大学计算机视觉研究中心

**在 OpenReview 上评审：** *https://openreview.net/forum?id=18yPFLbVRy*

arXiv:2601.01649v1 [cs.LG] 4 Jan 2026

## 摘要

联邦 AUC 最大化是从联邦学习（FL）中的不平衡数据进行学习的一种有效方法。然而，现有方法通常假设客户端始终可用，这在实际中很少成立。在现实世界的联邦学习系统中，客户端往往以循环方式参与训练：按照固定且重复的计划加入训练。这种设置为不可分解的 AUC 目标函数带来了独特的优化挑战。本文针对这些挑战，开发并分析了适用于循环客户端参与场景的联邦 AUC 最大化通信高效算法。我们研究了两种关键设置：首先，我们研究采用平方代理损失的 AUC 最大化问题，并将其重新表述为非凸-强凹极小极大优化问题。利用 Polyak–Łojasiewicz（PL）条件，我们建立了当前最优的通信复杂度 $\widetilde{O}(1/\epsilon^{1/2})$ 和迭代复杂度 $\widetilde{O}(1/\epsilon)$。其次，我们考虑一般的成对 AUC 损失，并建立了通信复杂度 $O(1/\epsilon^3)$ 和迭代复杂度 $O(1/\epsilon^4)$。此外，在 PL 条件下，这些界可改进为通信复杂度 $\widetilde{O}(1/\epsilon^{1/2})$ 和迭代复杂度 $\widetilde{O}(1/\epsilon)$。在图像分类、医学影像和欺诈检测基准任务上进行的大量实验表明，我们提出的方法在效率和效果方面均具有优势。

## 1 引言

联邦学习（FL）无需集中原始数据即可实现协同模型训练，因此在医疗保健、金融和移动计算等注重隐私的领域具有重要价值（Pati 等，2022；McMahan 等，2017；Konečn`y 等，2016；Hard 等，2018；Kairouz 等，2021c）。然而，大多数联邦学习研究都聚焦于经验风险最小化（ERM）（Yu 等，2019b；Stich，2018；Mohri 等，2019），而这种方法并不适用于不平衡数据，因为它最小化的是总体代理损失，无法直接解决模型对多数类的偏差问题（Elkan，2001）。相比之下，研究表明，直接优化 ROC 曲线下面积（AUC）等指标，更能反映类别不平衡情况下的模型性能（Ying 等，2016；Cortes 和 Mohri，2003；Rakotomamonjy，2004；Yuan 等，2021b）。

基于先前的研究（Gao 等人，2013；Ying 等人，2016年；Zhao等人，2011b；Kotlowski等人，2011；Gao和Zhou，2015；Calders和Jaroszewicz，2007；Charoenphakdee等人，2019a），深度 AUC 最大化问题可以表述为：

$$\min_{\mathbf{w}\in\mathbb{R}^d} \mathbb{E}_{\mathbf{z}\in\mathcal{S}_+}\mathbb{E}_{\mathbf{z}'\in\mathcal{S}_-}\psi(h(\mathbf{w};\mathbf{z}),h(\mathbf{w};\mathbf{z}')). \tag{1}$$

其中，$\mathcal{S}_+$ 和 $\mathcal{S}_-$ 分别表示正样本和负样本集合。函数 $\psi(\cdot,\cdot)$ 是代理损失，而 $h(\mathbf{w};\mathbf{z})$ 表示深度神经网络对输入数据 $\mathbf{z}$ 的预测分数

其参数为 w。近期研究已将 AUC 最大化拓展到联邦设置（Guo 等人，2020；Yuan 等人，2021a；Guo 等人（2023a））。Guo 等人（2020）和 Yuan 等（2021a）研究了采用平方代理损失的 AUC 最大化，并将问题重新表述为一种不含显式成对耦合的极小极大优化。Guo 等（2023a）则通过允许客户端共享预测分数，研究了更一般的成对 AUC 公式。然而，这些方法的一项主要局限是，假设每个通信轮次中所有客户端都可用。它们要么要求所有客户端参与，要么每轮随机采样一个子集——而这些条件在实践中很少能够满足。现实世界中的联邦学习系统通常呈现循环客户端参与的特点，即客户端按照固定且重复的计划加入训练（Cho 等人，2023）。循环客户端参与（CyCP）是一种结构化方法，旨在推动联邦学习从理想化模型走向实际的现实世界部署，因为受电池限制和网络连接不稳定等因素影响，客户端可用性本身具有间歇性（Huba 等人，2022；Paulik 等人，2021）。通过确保所有客户端或客户端组在规定的元纪元内均参与，CyCP 相较于纯随机采样具有显著优势：它确保覆盖整个群体的数据，这对于缓解非独立同分布数据偏差十分重要（Cho 等人，2023；Zhu等人，2023）；此外，严格限制每个客户端与全局模型共同参与的频率，这种受控且可预测的参与频率也能增强隐私保护（Kairouz 等人，2021a）。虽然此类设置下经验风险最小化的收敛保证已经得到确立（Cho 等人，2023；Crawshaw 和 Liu，2024；Wang 和 Ji，2022），但AUC目标不可分解的特性带来了独特挑战，尚未得到探索。

为了使联邦 AUC 最大化在现实部署条件下具有实用性，我们针对两类关键的代理损失，在循环客户端参与的条件下研究这一问题。首先，对于 平方代理损失$\psi(a,\ b) = (1 - a + b)^2$, 问题（1）可以重新表述为一个极小极大优化问题（Ying 等人，2016年）。极小极大形式不需要显式构造正负样本对，因此天然适用于在线学习（数据按序到达）以及联邦设置（数据分散在众多设备上），因为它简化了实现并避免了跨客户端成对管理。以往研究（Guo 等人，2020年；Yuan 等人，2021a；Sharma 等人，2022年；Deng 和 Mahdavi，2021年）针对这类极小极大问题开发了通信高效算法，但都假设全客户端参与。

其次，我们针对问题（1）考虑一般的成对代理损失（例如 sigmoid）。上述极小极大重新表述并不涵盖所有 AUC 代理损失（Zhao等人，2011a；Kotlowski等人，2011；Gao 和 Zhou，2015；Calders 和 Jaroszewicz，2007；Charoenphakdee 等，2019a）。特别是，已有研究表明，与平方代理损失相比，对称成对损失对标签噪声具有更强的鲁棒性（Charoenphakdee 等，2019b；Zhu等人，2022年）。它还可推广至二部排序和度量学习等任务（Cohen等人，1997；Clémençon等人，2008；Kotlowski等人，2011；Dembczynski等人，2012；Radenović等人，2016；Wu等人，2017）。尽管 Guo 等人（2023a）为这一目标提出了通信高效的联邦算法，但其分析依赖于随机客户端采样。相比之下，循环参与会引入确定性延迟，因为每一步只有特定客户端处于活跃状态。这两类问题的具体形式分别在第4节和第5节中给出。我们在附录G中讨论了所提出的算法与多个具体 AUC 一致损失的兼容性。

我们的贡献如下。

**联邦非凸强凹极小极大问题**。在循环参与下，原始（v）和对偶（$\alpha$更新均因确定性客户端调度而变得复杂，这打破了常见的独立采样假设。因此，全局梯度估计会产生偏差，使得（Guo 等人，2020年；Yuan 等人，2021a）中现有的收敛分析失效。为此，我们采用分阶段算法，每个阶段都包含多个通信完整周期，并按预定义的顺序遍历所有客户端组。我们的分析构造了辅助序列，用于追踪周期边界处的模型状态。这些锚点使我们能够界定每个周期内有偏更新导致的漂移，并厘清客户端组之间复杂的误差传播。利用 $\alpha$中的强凹性，我们进一步控制对偶动态。在 Polyak–Łojasiewicz（PL）条件下，我们的方法实现了$\widetilde{O}(1/\epsilon^{1/2})$的通信复杂度，达到全员参与设置下已知的最佳速率（Guo 等人，2020年），这表明循环参与并不会天然降低效率。

**联邦成对优化**。对于一般的成对目标，关键挑战在于处理不同客户端之间的耦合关系，以构建损失函数。我们设计了一种主动–被动算法，将梯度分解为两部分：一部分在本地计算（主动），另一部分通过其他客户端共享的信息计算

（被动）。我们的分析将模型更新追溯至上一个循环的起点，在该起点，所有组件实际上都是在独立于当前更新的状态下进行评估的，从而减小了偏差。不同于（Guo 等人，2023a），我们的方法必须处理整整两个循环的延迟，这使误差动态更加复杂。尽管如此，我们的算法仍实现了相同的 $\tilde{O}(1/\epsilon^3)$ 通信复杂度（并且在 PL 条件下具有更优的收敛速率），这证实了在循环参与下，高效的成对优化仍然可行。

**实证评估** 我们在多个领域的基准数据集上验证了所提方法，包括 CIFAR-10、CIFAR-100（图像分类）、ChestMNIST（医学影像）以及一个大规模保险欺诈检测数据集。结果一致表明，在循环客户端参与场景下，我们提出的算法兼具有效性和效率。

## 2 相关工作

我们的工作处于针对非经验风险最小化目标的联邦优化与客户端参与方案研究的交叉领域。下面简要回顾这两个领域。

### 2.1 联邦非经验风险最小化优化

尽管联邦学习文献主要关注经验风险最小化（ERM），但受曲线下面积最大化（Ying 等人，2016年）以及分布鲁棒优化（Namkoong 和 Duchi，2016年）等应用的推动，人们对极小极大和复合优化等非经验风险最小化目标的兴趣日益浓厚。

联邦极小极大优化。已有多项研究探讨了联邦极小极大问题（Guo 等人，2020；Yuan 等（2021a）；Deng 和 Mahdavi，2021；Sharma 等人，2022）。与我们的工作最接近的是，Guo 等人（2020）和 Yuan 等（2021a）通过将曲线下面积最大化重新表述为非凸—强凹极小极大问题来解决该问题。他们利用 PL条件，并采用分阶段算法求解一系列子问题。Guo 等人（2020）实现了 $O(1/(\mu^2\epsilon))$的迭代复杂度和 $O(1/(\mu^{3/2}\epsilon^{1/2}))$的通信复杂度，其中 $\mu$ 为 PL 模量。随后，Yuan 等（2021a）引入方差缩减技术，进一步降低了通信复杂度。这些方法的一个关键局限在于假设客户端全部参与或独立随机参与，而我们的工作放宽了这一假设。

**联邦复合优化**。复合目标引入了嵌套结构，这在联邦学习中尤具挑战性。Gao 等人(2022年)分析了一种较简单的情形，其中每个客户端 $k$ 都拥有完全本地化的内层（$g_k$）和外层（$f_k$）函数，从而避免了客户端间的耦合。相比之下，Guo 等人(2023a)考虑了更一般的成对目标，其中不同客户端的数据天然存在耦合。他们的算法采用主动—被动梯度分解策略，并在随机客户端采样下实现了线性加速。其他启发式方法也已提出（Wu等人，2022年；Li & Huang，2022年），但通常缺乏严格的理论保证。

### 2.2 联邦学习中的客户端参与

联邦学习中已探索了各种客户端参与方案，以平衡效率、可扩展性和鲁棒性。最简单的策略是全参与，即所有客户端在每一轮中都执行本地更新（Stich，2018年；Yu 等人，2019a；b）。无偏客户端采样每轮随机选择一部分客户端（Karimireddy 等人，2020年；Jhunjhunwala 等人，2022年；Yang 等人，2021a），而有偏采样则依据预定义规则（Ruan 等人，2021年）或本地损失值等数据特征（Cho 等人，2020年；Goetz 等人，2019年）优先选择客户端。为适应异构客户端速度和可用性，也有人提出异步参与（Yu 等人，2019b；Wang 和 Ji，2022）。近年来，循环客户端参与——客户端按照固定且重复的顺序加入——因其实际应用价值和隐私优势而受到关注（Kairouz 等人，2021b）。这种结构化调度能够应对各种可用性挑战。在跨设备联邦学习中，客户端可能位于不同时区，或倾向于在特定时间为设备充电（Cho 等人，2023；Yang 等人，2018）。在跨机构联邦学习中，医院等客户端面临预先安排的限制，例如管理大型本地数据集、内部网络安全要求或计划中的 IT 维护时段。Cho 等人（2023）已为经验风险最小化中的循环参与建立收敛保证，Crawshaw 和 Liu（2024）随后通过方差缩减进一步强化了这些保证。然而，所有这些方法都只关注经验风险最小化。我们的

工作通过开发通信高效算法，并针对循环客户端参与这一现实场景下的联邦 AUC 最大化提供严格的收敛分析，弥合了这一空白。

## 3 预备知识与符号说明

我们首先给出全文中使用的标准定义和符号。若对于其定义域中的任意$\mathbf{x}$、$\mathbf{x}'$，均有 $|f(\mathbf{x}) - f(\mathbf{x}')| \leq C|\mathbf{x} - \mathbf{x}'|$，则称函数 $f$ 是 $C$-Lipschitz连续的。若可微函数 $f$ 的梯度是Lipschitz连续的，即 $|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}')| \leq L|\mathbf{x} - \mathbf{x}'|$，则称其为 $L$-光滑函数。

我们考虑一个包含 $N$ 个客户端的联邦学习环境，这些客户端进一步划分为 $K$ 个组。全局数据集$\mathcal{D}^1$ （正样本集）和$\mathcal{D}^2$ （负样本集）被划分为 $N$ 个互不重叠的子集，并分配到这些客户端上，分别记为$\mathcal{D}_1 = \mathcal{D}_1^1 \cup \mathcal{D}_1^2 \cup \ldots \cup \mathcal{D}_1^N$和$\mathcal{D}_2 = \mathcal{D}_2^1 \cup \mathcal{D}_2^2 \cup \ldots \cup \mathcal{D}_2^N$。其中，$\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}$ 表示在数据集$\mathcal{D}$上的经验期望。

我们将 $N$ 个客户端划分为 $K$ 个互不相交的组，每组包含 $N/K$ 个客户端，记为$\mathcal{G}^k$ ，其中 $k \in [K]$。训练期间，服务器按照固定的预定顺序依次遍历这些组（$\mathcal{G}^1$，...，$\mathcal{G}^K$），从而实现循环参与模式。在每个通信轮次中，从活跃组中均匀随机地无放回抽取 $M$ 个客户端。

该框架与现实世界中的联邦学习应用尤为相关。例如，在跨设备联邦学习中，可以按地区或时区对移动设备进行分组，而在每个组内进行随机采样则能提供灵活性，尤其是在组内客户端数量较多时。为简化起见，在 $K = N$ 或 $M = N/K$的特殊情况下，该框架也可以理解为完全确定性调度，即活跃组中的所有客户端都参与每一轮；在这种情况下，理论保证仍然成立。

评分函数 $h: \mathcal{X} \to \mathbb{R}$ 的曲线下面积定义如下：

$$AUC(h) = \Pr(h(\mathbf{x}) \geq h(\mathbf{x}')|y = 1, y' = -1), \tag{2}$$

其中，$\mathbf{z} = (\mathbf{x}\ y)$ 和 $\mathbf{z}' = (\mathbf{x}', y')$ 独立地从数据分布中抽取 $\mathbb{P}$.

**符号说明。** 令 $\mathbf{w} \in \mathbf{R}^d$ 表示模型参数。在极小极大形式中，我们将$\mathbf{v} = (\mathbf{w}\ a, b)$定义为原变量，其中 $a, b \in \mathbf{R}$。对于分阶段算法，$\mathbf{v}_0^s$ 和 $\alpha_0^s$ 表示阶段 $s$开始时的初始化。

在每个阶段内，$\mathbf{v}_m^{e,k}{}_{,,t}$ 表示轮次 $e$、客户端组 $k$、客户端 $m$和迭代 $t$对应的变量。符号$\mathbf{v}_m^{e,k}{}_{,}$表示组 $k$ 中客户端 $m$在轮次 $e$的最终输出。组级聚合定义为

$$\mathbf{v}^{e,k} = \frac{1}{|\mathcal{G}^{e,k}|} \sum_{m \in \mathcal{G}^{e,k}} \mathbf{v}_m^{e,k},$$

它是轮次 $e$期间组 $k$ 中所有参与客户端输出的平均值。

对于$\mathbf{w}_m^{e,k}{}_{,,t}$ 和 $\alpha^{e,k}_{m,t}{}_{,}$，同样的符号约定也适用。本文使用的完整符号列表见附录A。

## 4 极小极大优化

为最大化曲线下面积，我们采用式（1）中的代理损失形式。遵循Ying 等人(2016年)，我们使用平方代理损失 $\psi(a, b) = (1 - a + b)^2,$，将曲线下面积最大化问题重写为如下极小极大优化问题：

$$\min_{\mathbf{w},a,b} \max_{\alpha} f(\mathbf{w}, a, b, \alpha) = \mathbb{E}_{\mathbf{z}}[F(\mathbf{w}, a, b, \alpha; \mathbf{z})], \tag{3}$$

其中

$$\begin{aligned} F(\mathbf{w}, a, b, \alpha; \mathbf{z}) =& (1-p)(h(\mathbf{w}; \mathbf{z}) - a)^2 \mathbb{I}_{[y=1]} + p(h(\mathbf{w}; \mathbf{z}) - b)^2 \mathbb{I}_{[y=-1]} \\ &+ 2(1+\alpha)(ph(\mathbf{w}; \mathbf{z})\mathbb{I}_{[y=-1]} - (1-p)h(\mathbf{w}; \mathbf{z})\mathbb{I}_{[y=1]}) - p(1-p)\alpha^2. \end{aligned}$$

令 $\mathbf{v} := (\mathbf{w}\ a,\ b)$ 表示原始变量集合。该重新表述实现了在客户端之间进行自然分解，从而得到以下联邦优化问题：

$$\min_{\substack{\mathbf{w}\in\mathbb{R}^d \\ (a,b)\in\mathbb{R}^2}} \max_{\alpha\in\mathbb{R}} f(\mathbf{w},a,b,\alpha) = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{|\mathcal{G}^k|}\sum_{m\in\mathcal{G}^k} f_{k,m}(\mathbf{w},a,b,\alpha), \tag{4}$$

其中，$f_{k,\ m}(\mathbf{w}\ a,\ b,\ \alpha) = \mathbb{E}_{\mathbf{z}\sim\mathbb{P}_{k,\ m}}[F(\mathbf{w}\ a,\ b,\ \alpha;\mathbf{z}^k)]$ 且 $\mathbb{P}_{k,\ m}$ 表示组 $k$中客户端 $m$的本地数据分布。

现在，我们介绍一种在循环客户端参与这一实际约束下实现联邦 AUC 最大化的算法。在第 $s$ 阶段，我们构造一个以前一阶段输出 $\mathbf{v}_0^s$ 为中心的强凸–强凹子问题。组 $k$ 中客户端 $m$ 的本地目标定义为：

$$\min_{\mathbf{v}}\max_{\alpha} f_{k,m}(\mathbf{v},\alpha) + \frac{\gamma}{2}|\mathbf{v}-\mathbf{v}_0^s|^2. \tag{5}$$

尽管我们的方法建立在 Guo 等人 （2020年）的阶段式框架以及 Yuan 等人 （2021a）的框架之上，但其关键创新在于阶段内优化：我们引入多个循环轮次，并给出一种能够处理确定性客户端排序的理论分析——这一方面带来了重大的技术挑战。单阶段的详细流程如算法 1所示，而总体的多阶段框架则见算法2。每个阶段都使用前一阶段的输出初始化原始变量和对偶变量。在一个阶段内，算法运行多个训练轮数，在此期间所有客户端组会被多次访问。每个客户端组采样一部分客户端参与，执行 $I$ 本地更新步数，然后将更新后的原始变量和对偶变量传递给下一个客户端组。原始变量通过随机梯度下降进行更新，而对偶变量通过随机梯度上升进行更新。完成一个阶段后，会调整步长和训练轮数，然后进入下一阶段。

我们定义以下用于分析的关键量：

$$\begin{aligned}
&\phi(\mathbf{v}) = \max_{\alpha} f(\mathbf{v},\alpha), \quad \phi_s(\mathbf{v}) = \phi(\mathbf{v}) + \frac{\gamma}{2}\|\mathbf{v}-\mathbf{v}_{s-1}\|^2, f^s(\mathbf{v},\alpha) = f(\mathbf{v},\alpha) + \frac{\gamma}{2}\|\mathbf{v}-\mathbf{v}_{s-1}\|^2, \\
&F_{k,m}^s(\mathbf{v},\alpha;\mathbf{z}_{k,m}) = F(\mathbf{v},\alpha;\mathbf{z}_{k,m}) + \frac{\gamma}{2}\|\mathbf{v}-\mathbf{v}_{s-1}\|^2, \mathbf{v}_\phi^* = \arg\min_{\mathbf{v}}\phi(\mathbf{v}), \quad \mathbf{v}_{\phi_s}^* = \arg\min_{\mathbf{v}}\phi_s(\mathbf{v}).
\end{aligned} \tag{6}$$

我们的收敛分析基于以下标准假设：

假设 *4.1. (i)* 初始化：存在 $v0, \Delta_0 > 0$ 使得 $\phi(v_0) - \phi(v_\phi^*) \le \Delta_0$. *(ii) PL*条件： $\phi(v)$ 满足 $\mu$-*PL* 条件，即 $\mu(\phi(v) - \phi(v_*)) \le \frac{1}{2}\|\nabla\phi(v)\|^2$; *(iii)* 光滑性：对于任意 $z$， $f(v\ \alpha; z)$ 关于 $v$ 是 $\ell$-光滑的，且 $\alpha$. $\phi(v)$ 是 $L$-光滑的，即 $\|\nabla\phi(v_1) - \nabla\phi(v_2)\| \le L\|v_1 - v_2\|$;*(iv)* 有界梯度：$\|\nabla_{\mathbf{v}} f_{k,m}^s(v\ \alpha; z)\|^2 \le G^2$; *(v)* 有界方差：

$$\begin{aligned}
&\mathbb{E}[\|\nabla_{\mathbf{v}} f_{k,m}(\mathbf{v},\alpha) - \nabla_{\mathbf{v}} F_{k,m}(\mathbf{v},\alpha;\mathbf{z})\|^2] \le \sigma^2, \\
&\mathbb{E}[|\nabla_{\alpha} f_{k,m}(\mathbf{v},\alpha) - \nabla_{\alpha} F_{k,m}(\mathbf{v},\alpha;\mathbf{z})|^2] \le \sigma^2.
\end{aligned} \tag{7}$$

**备注**。 这些假设与先前文献中的设定一致（Guo 等人，2020年；Yuan 等人，2021a）。极小极大 AUC 最大化的 PL条件已得到理论和实证验证（Guo 等人，2023b）。由于 $f$（$\mathbf{v}\ \alpha$；$\mathbf{z}$）关于 $\ell$-光滑于 $\mathbf{v}$，因此关于 $\ell$-弱凸于 $\mathbf{v}$。相应地，在后续分析中我们选择 $\gamma = 2\ell$，从而保证每个阶段的子问题（5）关于 $\ell$-强凸于 $\mathbf{v}$。

我们的 rst 首先要建立一个阶段内次优性差距的界（引理 4.2）。

**算法 1** 单阶段联邦极小极大（OSFM）

**在服务器上：**
初始化：$\mathbf{v}^{1,0}$ , $\alpha^{1,0}$
**对于** $e \in [E]$ 循环周期 **执行**
  **对于** $k \in [K]$ **执行**
    从第 $k$ 个客户端集合中均匀随机采样 $M$ 个客户端（不放回），得到客户端集合 $\mathcal{G}^{e,k}$
    将全局模型 $\mathbf{v}^{e,k-1}$ 发送给$\mathcal{G}^{e,k}$中的客户端
    每个客户端 $m \in \mathcal{G}^{e,k}$ 并行执行：
      $\mathbf{v}_m^{e,k}, \alpha_m^{e,k} \leftarrow$ 本地更新$(\mathbf{v}^{e,k-1}, \alpha^{e,k-1})$
    $\mathbf{v}^{e,k} = \frac{1}{|\mathcal{G}^{e,k}|} \sum_{m \in \mathcal{G}^{e,k}} \mathbf{v}_m^{e,k}$
    $\alpha^{e,k} = \frac{1}{|\mathcal{G}^{e,k}|} \sum_{m \in \mathcal{G}^{e,k}} \alpha_m^{e,k}$
  **结束循环**
  $\mathbf{v}^{e+1,0} = \mathbf{v}^{e,K}$, $\alpha^{e+1,0} = \alpha^{(e,K)}$
**结束循环**
输出：$\frac{1}{E}\sum_e \frac{1}{K}\sum_k \mathbf{v}^{e,k}$, $\frac{1}{E}\sum_e \frac{1}{K}\sum_k \alpha^{e,k}$

本地更新$(\mathbf{v}_0, \alpha_0)$：
**对于** $t \in [I]$，**执行**
  采样小批次 $\mathbf{z}_m^{e,k}$ 来自本地数据集
  更新 $\mathbf{v}_{m,t}^{e,k} = \mathbf{v}_{m,t-1}^{e,k} - \eta \nabla_{\mathbf{v}} f(\mathbf{v}_{m,t}^{e,k}, \alpha_{m,t}^{e,k}; \mathbf{z}_{m,t}^{e,k})$
  更新 $\alpha_{m,t}^{e,k} = \alpha_{m,t-1}^{e,k} + \eta \nabla_{\alpha} f(\mathbf{v}_{m,t}^{e,k}, \alpha_{m,t}^{e,k}; \mathbf{z}_{m,t}^{e,k})$
**结束循环**
返回 $\mathbf{v}_I$, $\alpha_I$

**算法 2** 具有循环客户端参与的联邦极小极大算法（CyCp-Minimax）

初始化：原变量 $\mathbf{x}^{1,0}$, 客户端组 $\delta(k)$, $k \in [K]$**对于** $s \in [S]$ **执行**
$\mathbf{v}_s, \alpha_s =$ OSFM$(\mathbf{v}_{s-1}, \alpha_{s-1}, \eta_s, E_s)$**结束循环**

**引理 4.2**。 假设假设条件 *4.1* 成立，并运行算法 *1* 一个阶段，输出记为 $(\bar{\mathbf{v}}, \bar{\alpha})$，则有

$$
\begin{aligned}
f^s(\bar{\mathbf{v}}, \alpha) - f^s(\mathbf{v}, \bar{\alpha}) &\le \frac{1}{E}\sum_{e=0}^{E-1}[f^s(\mathbf{v}^{e+1,0}, \alpha) - f^s(\mathbf{v}, \alpha^{e+1,0})] \\
&\le \frac{1}{E}\sum_{e=0}^{E-1}\Bigg[\underbrace{\langle \partial_{\mathbf{v}} f^s(\mathbf{v}^{e,0}, \alpha^{e,0}), \mathbf{v}^{e+1,0} - \mathbf{v}\rangle}_{A_1} + \underbrace{\langle \partial_{\alpha} f^s(\mathbf{v}^{e,0}, \alpha^{e,0}), \alpha - \alpha^{e+1,0}\rangle}_{A_2} \\
&\qquad + \underbrace{\frac{3\ell + 3\ell^2/\mu_2}{2}\|\mathbf{v}^{e+1,0} - \mathbf{v}^{e,0}\|^2 + 2\ell(\alpha^{e+1,0} - \alpha^{e,0})^2}_{A_3} - \frac{\ell}{3}\|\mathbf{v} - \mathbf{v}^{e,0}\|^2 - \frac{\mu_2}{3}(\alpha^{e,0} - \alpha)^2\Bigg].
\end{aligned}
$$

由此得到的界包含如下项 $A_1$ 和 $A_2$，它们将不同客户端组的随机性耦合在一起。为了解耦这些依赖关系，我们引入了新颖的虚拟序列 $\hat{\mathbf{v}}^{e,0}$, $\tilde{\mathbf{v}}^{e,0}$, $\hat{\alpha}^{e,0}$以及 $\tilde{\alpha}^{e,0}$（引理4.3和 B.4）。这些序列纯属概念性构造，无需任何实际计算。与（Guo 等人，2020年；Yuan 等人，2021a）不同，这些虚拟序列还依赖于虚拟梯度估计，例如，使用当前数据和旧模型计算梯度。这些序列经过精心设计，旨在将“信号”（真实梯度）与“噪声”（随机梯度误差）分离开来，并处理循环调度引入的偏差。

$K$ $I$**引理 4.3.** 定义 $\hat{\mathbf{v}}^{e+1,0} = \mathbf{v}^{e,0} - \eta \sum_{k=1}^{K} \frac{1}{M} \sum_{m \in \mathcal{G}^{e,k}} \sum_{t=1}^{I} \nabla_{\mathbf{v}} f_k^s(\mathbf{v}^{e,0}, \alpha^{e,0})$，$\alpha^{e,0}$) 以及

$$\tilde{\mathbf{v}}^{e+1,0} = \tilde{\mathbf{v}}^{e,0} - \frac{\eta}{M} \sum_{k=1}^{K} \sum_{m \in \mathcal{G}^{e,k}} \sum_{t=1}^{I} \left( \nabla_{\mathbf{v}} F_k^s(\mathbf{v}^{e,0}, \alpha^{e,0}; z_{m,t}^{e,k}) - \nabla_{\mathbf{v}} f_k^s(\mathbf{v}^{e,0}, \alpha^{e,0}) \right), \ for\ t > 0;\ \tilde{\mathbf{v}}_0 = \mathbf{v}_0. \tag{8}$$

于是我们有

$$\mathbb{E}_{e,0}[A_1] \le \frac{6\tilde{\eta} K \sigma^2}{MKI} + \frac{6\ell}{KMI} \sum_k \sum_{m \in \mathcal{G}^{e,k}} \sum_t \mathbb{E}(\|\mathbf{v}^{e,0} - \mathbf{v}_{m,t}^{e,k}\|^2 + \|\alpha^{e,0} - \alpha_{m,t}^{e,k}\|^2) + \frac{\ell}{3}\|\mathbf{v}^{e+1,0} - \mathbf{v}\|^2$$
$$+ \frac{1}{2\eta KI}\mathbb{E}(\|\mathbf{v}^{e,0} - \mathbf{v}\|^2 - \|\mathbf{v}^{e,0} - \mathbf{v}^{e+1,0}\|^2 - \|\mathbf{v}^{e+1,0} - \mathbf{v}\|^2) + \frac{1}{2\eta KI}(\|\mathbf{v} - \tilde{\mathbf{v}}^{e,0}\|^2 - \|\mathbf{v} - \tilde{\mathbf{v}}^{e+1,0}\|^2),$$

其中$\mathbf{E}_{e,0}$ 表示对第 $e$ 轮次之前实现的所有随机性取期望。

借助这两个虚拟序列 $\hat{\mathbf{v}}^{e,0}$ 和 $\tilde{\mathbf{v}}^{e,0}$，我们成功解耦了不同客户端之间的相互依赖，并将界转化为标准方差界加模型漂移，即右端（RHS）的第二项。

将各项 合并起来，我们得到算法一个阶段的收敛分析如下：

**引理 4.4**。 假设 *4.1* 成立。运行算法 *1* 的一个阶段可确保

$$\mathbb{E}[f^s(\bar{\mathbf{v}}, \alpha) - f^s(\mathbf{v}, \bar{\alpha})] \le \frac{1}{4\eta EKI}\|\mathbf{v}_0 - \mathbf{v}\|^2 + \frac{1}{4\eta EKI}\|\alpha_0 - \alpha\|^2 + \left(\frac{3\ell^2}{2\mu_2} + \frac{3\ell}{2}\right) 36\eta^2 I^2 K^2 G^2 + \frac{3\eta\sigma^2}{M},$$

**备注**。 上述引理表明，一个阶段的输出（算法 1）的界取决于其输入的质量，即 $\mathbf{v}_0$ 和 $\alpha_0$，它们是前一阶段的输出。通过采用分阶段算法（算法2）并适当设置参数，可以确保对偶间隙在各阶段之间呈指数下降。将单阶段结果扩展到完整的双循环（算法 2）过程，可得到以下定理。

定理 *4.5*。定义 $\hat{L} = L + 2\ell, c = \mu/\hat{L}$ 设 $\gamma = 2\ell$, $\eta_s = \eta_0 \exp(-(s-1)c)$, $T_s = \frac{212}{\eta_0 \epsilon} \max(\ell, \mu_2) \exp((s-1)c)$。要返回满足 $\mathbb{E}[\phi(v_S) - \phi(v_\phi^*)] \le \epsilon$ 的 $v_S$，只需选择 $S \ge O\left(\frac{\hat{L}+\mu}{\mu}\right.$ $\left.\max\left\{\log\left(\frac{2\Delta_0}{\epsilon}\right), \log S + \log\left[\frac{2\eta_0}{\epsilon}\left(\frac{12\sigma^2}{5K}\right)\right]\right\}\right)$。迭代复杂度为 $\tilde{O}\left(\frac{\hat{L}}{\mu^2 M\epsilon}\right)$，通过设置 $I_s = \Theta\left(\frac{1}{K\sqrt{M\eta_s}}\right)$，通信复杂度为 $\tilde{O}\left(\frac{K}{\mu^{3/2}\epsilon^{1/2}}\right)$，其中 $\tilde{O}$ 表示忽略对数因子。

**备注**。**线性加速与效率：** 迭代复杂度相对于同时参与的客户端数量 $M$呈线性加速，即 $O(1/(M\epsilon))$。这表明，即使在循环参与的情况下，所提出的算法也能高效利用并行性。随着组数增加，需要缩短通信间隔（按 $K$ 缩放），这是循环参与协议自然带来的结果。较大的 $K$ 会使连续客户端更新之间的延迟变长，从而加剧陈旧性和漂移。根据 $K$ 调整本地更新次数 $I_s$，能够有效平衡通信效率与由此产生的误差。

**与先前工作的比较：** 我们的通信复杂度对 $\mu$ 和 $\epsilon$ 的依赖与 Guo 等人(2020年)；Yuan 等人(2021a)在随机客户端采样这一更理想化假设下得到的结果相匹配。因此，我们的分析表明，循环参与不会降低渐近收敛速率。据我们所知，这是联邦极小极大优化领域中首个此类保证。

## 5 成对目标

本节中，我们研究一类更广泛的联邦成对优化问题，将曲线下面积最大化推广到任意成对损失函数。我们提出了一种新算法，并针对

这一设定提供收敛分析，适用于循环客户端参与，无论是否假设满足 PL 条件。我们考虑以下联邦成对目标函数：

$$\min_{\mathbf{w}\in\mathbb{R}^d} F(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\mathbf{z}\in\mathcal{D}_1^i}\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}_{\mathbf{z}'\in\mathcal{D}_2^j}\psi(h(\mathbf{w};\mathbf{z}),h(\mathbf{w};\mathbf{z}')). \tag{9}$$

优化方程 9 的一项重大挑战在于，其梯度天然地将所有客户端的数据耦合在一起。设$\nabla_1\psi(\cdot、\cdot)$和 $\nabla_2\psi(\cdot、\cdot)$分别表示 $\psi$ 相对于其第一和第二个参数的偏导数。我们将全局梯度分解为：

$$\begin{aligned}\nabla F(\mathbf{w}) =& \underbrace{\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\mathbf{z}\in\mathcal{D}_1^i}\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}_{\mathbf{z}'\in\mathcal{D}_2^j}\nabla_1\psi(h(\mathbf{w};\mathbf{z}),h(\mathbf{w};\mathbf{z}'))\nabla h(\mathbf{w};\mathbf{z})}_{\Delta_{i1}}\\ &+\underbrace{\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\mathbf{z}'\in\mathcal{D}_2^i}\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}_{\mathbf{z}\in\mathcal{D}_1^j}\nabla_2\psi(h(\mathbf{w};\mathbf{z}),h(\mathbf{w};\mathbf{z}'))\nabla h(\mathbf{w};\mathbf{z}')}_{\Delta_{i2}}.\end{aligned}$$

将局部梯度分量定义为 $\nabla F_i(\mathbf{w}) := \Delta_{i1} + \Delta_{i2}$，全局梯度可简洁地表示为 $\nabla F(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{w})$。

计算 $\Delta_{i1}$ 和 $\Delta_{i2}$ 的难点在于其依赖全局信息——具体而言，是对所有客户端数据集的期望 $\mathcal{D}_2^j$ 和$\mathcal{D}_1^j$：

$$\Delta_{i1} = \mathbb{E}_{\mathbf{z}\in\mathcal{D}_1^i}\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}_{\mathbf{z}'\in\mathcal{D}_2^j}\nabla_1\psi(\underbrace{h(\mathbf{w};\mathbf{z})}_{\text{local}},\underbrace{h(\mathbf{w};\mathbf{z}')}_{\text{global}})\underbrace{\nabla h(\mathbf{w};\mathbf{z})}_{\text{local}}, \tag{10}$$

其中，局部分量可由每个客户端使用自身数据进行计算，而全局分量则依赖所有客户端的数据。

为应对这一挑战，同时保持循环参与约束，我们采用主动–被动策略。为估计 $\Delta_{i,1}$，每个客户端 $i$ 使用自身数据计算主动部分，而被动部分则由其他客户端通过先前共享的全局信息提供。具体而言，对于第 $e$ 轮次中的第 $k$ 组，我们通过无放回采样 $h_2^{e-1}{}_{,\xi}\in\mathcal{H}_2^{e-1}$来估计全局项，并计算 $\Delta_{i1}$ 的估计量：

$$G_{i,t,1}^{e,k} = \nabla_1\psi(\underbrace{h(\mathbf{w}_{m,t}^{e,k};\mathbf{z}_{m,t,1}^{e,k})}_{\text{active}},\underbrace{h_{2,\xi}^{e-1}}_{\text{passive}})\underbrace{\nabla h(\mathbf{w}_{m,t}^{e,k};\mathbf{z}_{m,t,1}^{e,k})}_{\text{active}}, \tag{11}$$

并类似地 $\Delta_{i2}$ 通过

$$G_{m,t,2}^{e,k} = \nabla_1\psi(\underbrace{h_{1,\xi}^{e-1}}_{\text{passive}},\underbrace{h(\mathbf{w}_{m,t}^{e,k};\mathbf{z}_{m,t,2}^{e,k})}_{\text{active}},)\underbrace{\nabla h(\mathbf{w}_{m,t}^{e,k};\mathbf{z}_{m,t,2}^{e,k})}_{\text{active}}, \tag{12}$$

其中，$G^{e,k}_{i,t,1}$ 中的局部分量称为主动部分，而全局分量称为被动部件。被动部件 $h_2^{e-1}{}_{,\xi}$ 和 $h_1^{e-1}{}_{,\xi}$ 通过对上一轮次（$e-1$）的得分进行采样构造。而 $\xi = (\tilde{k},\tilde{m},\tilde{t},\mathbf{z}^{e-1}_{\tilde{m}\,\tilde{t}\,2},,,)$ 表示一个随机变量，用于捕获从采样组$\tilde{k}\in\{1,\ldots,K\}$、采样客户端 $\tilde{m}\in\{1,\ldots,G^{\tilde{k}}\}$、迭代索引 $\tilde{t}\in\{1,\ldots,I\}$、数据样本 $\mathbf{z}^{e-1\,k}_{\tilde{m}\,\tilde{t}},,,_1\in\mathbf{D}_1^j$ 和$\mathbf{z}^{e-1\,k}_{\tilde{m}\,\tilde{t}},,,_2\in\mathbf{D}_2^j$ 中所包含的随机性，以估计式（10）中的全局分量。这种延迟估计带来两项挑战：使用旧估计值产生的延迟误差，以及不同轮次和客户端的随机性之间的相互依赖。

我们的算法设计和分析与 Guo 等人（2023a）在两个关键方面存在差异：（1）被动组件来自上一 轮次，而非上一 轮次，这会引入更大的延迟，但对于

循环分析而言是必要的；（2）确定性循环顺序相比随机采样在客户端之间造成了更复杂的相互依赖关系。不同于我们的极小极大分析可以追溯到当前循环的起点，这里我们必须参考上一循环的初始状态（$\mathbf{w}^{e-1,\ 0}$）以建立独立性，这增加了分析的复杂性。算法 3展示了一个单循环算法，而4中展示了一个利用PL条件的双循环算法。单阶段联邦成对算法（算法3：OSFP）通过迭代多个轮次执行单阶段联邦优化。在每个轮次中，它依次处理每个客户端组。对于每个组，服务器采样一部分客户端，并广播当前全局模型以及参考预测集。随后，每个选定的客户端执行本地更新 过程：维护过去预测的缓冲区，采样新的数据点，计算成对损失及其梯度，并使用随机梯度步更新本地模型。本地计算完成后，客户端将更新后的模型和预测集返回服务器，由服务器聚合这些结果以更新全局模型。所有轮次完成后，OSFP 输出全局模型平均值。此外，如果满足PL条件，外层循环（算法4）会在多个阶段中反复调用OSFP，同时调整学习率和轮次计划，以逐步改进模型。

我们对问题（9）中的成对优化分析依赖以下标准假设。

假设 *5.1*。*(i)* $\psi(\cdot)$ 可微、$L_\psi$-光滑且满足 $C_\psi$-利普希茨条件。*(ii)* 对任意 $z \in \ S_1 \ \cup S_2$，$h(\cdot; z)$ 关于 $w$ 可微、$L_h$-光滑且满足 $C_h$-利普希茨条件。*(iii)* $\mathbb{E}_{\mathbf{z}\in\mathcal{S}_1^i}\mathbb{E}_{j\in[1:N]}\mathbb{E}_{\mathbf{z}'\in\mathcal{S}_2^j}\|\nabla_1\psi(h(w; z)\ h(w; z'))\nabla h(w; z) + \nabla_2\psi(h(w; z)\ h(w; z'))\nabla h(w; z') - \nabla F_i(w)\|^2 \le \sigma^2$。*(iv)* $\exists D$ 使得 $\|\nabla F_i(w) - \nabla F(w)\|^2 \le D^2, \forall i$。

**定理 5.2**。 假设 *5.1* 成立。运行算法 *3* 可确保

$$\frac{1}{E}\sum_{e=1}^{E}\mathbb{E}\|\nabla F(\mathbf{w}^{e-1})\|^2 \le O\left(\frac{F(\mathbf{w}^{e,0}) - F(\mathbf{w}_*)}{\eta IKE} + \eta^2 I^2 K^2 D^2 + 24\eta\frac{\sigma^2}{M}\right). \tag{13}$$

**备注**。 由于 $\tilde{\eta} = \eta I$。通过设置 $\eta = \epsilon^2 M, I = \frac{1}{MK\epsilon^2}, E = \frac{1}{\epsilon^3}$，迭代复杂度为 $EKI = O(\frac{1}{M\epsilon^4})$，这体现了$M$带来的线性加速，通信成本为 $EK = O(\frac{K}{\epsilon^3})$。这些结果与全客户端参与设置下（Guo 等人，2023a）的结果一致。这表明，我们的算法在不牺牲渐近效率的情况下，成功应对了循环参与带来的额外挑战。值得注意的是，预测分数（$\mathcal{H}_1^e$，$\mathcal{H}_2^e$）不会产生额外的计算开销，因为它们只是重复利用本地更新期间生成的分数。由于（$\mathcal{H}_1^e$，$\mathcal{H}_2^e$）仅存储上一通信轮次的预测分数，所需内存为 $O(IMK)$，其中 $I$ 是通信间隔， $K$是客户端组的数量， $M$ 是每组同时参与的客户端数量。与现代神经网络的参数数量相比，这项存储开销可以忽略不计；在实践中，还可以通过调优 $M$ 或 $I$来调整该开销。每个客户端的本地缓冲区 $\mathcal{B}_{i,\ 1}$ 和$\mathcal{B}_{i,\ 2}$的大小为 $O(I)$，其中存储了足够的历史预测，用于在每次本地更新迭代中构造损失函数。打乱这些缓冲区至关重要，这样随着时间推移，每个客户端都有机会与其他每个客户端交互。

在额外假设 $F(\mathbf{w})$ 满足 $\mu$-PL条件（成对优化中常用的假设，Yangetal.，2021c）下，我们可以获得显著提升的收敛速率，如下所示。

**定理 5.3**。 假设 *5.1* 成立，且 $F(\cdot)$ 满足 $\mu$-*PL*条件。通过运行算法 $\epsilon$，要达到一个 -平稳点，迭代复杂度为 $\widetilde{O}\left(\frac{1}{\mu^2 M\epsilon}\right)$，通信复杂度为$\widetilde{O}\left(\frac{K}{\mu^{3/2}\epsilon^{1/2}}\right)$，其中令 $I_s = \Theta(\frac{1}{\sqrt{K\eta_s}})$，其中 $\widetilde{O}$ 隐藏了对数因子。

**备注**。**线性加速：** 定理 5.2 和5.3 均证实，随着同时参与的客户端数量增加，可实现线性加速 $M$，这表明即使在循环参与的情况下，也能有效利用并行资源。

**PL条件的影响：**PL条件将迭代复杂度从 $O(1/\epsilon^4)$ 大幅降低至 $\widetilde{O}(1/\epsilon)$，并将通信复杂度从 $O(1/\epsilon^3)$降低至 $\widetilde{O}(1/\mu^{3/2}\epsilon^{1/2})$。

**算法 3** 单阶段联邦成对算法 (OSFP)

在服务器端
通过一个初始轮次初始化 $\mathcal{H}_1^1$,$\mathcal{H}_2^1$
**对于** $e \in [E]$ **执行**
  $\boldsymbol{H}_1^{e+1}, \boldsymbol{H}_2^{e+1} = \emptyset$
  **对于** $k \in [K]$ **执行**
    从第 $k$ 个客户端集合中均匀随机地无放回抽取 $M$ 个客户端，得到客户端集合 $\mathcal{G}^{e,k}$
    将全局模型 $\mathbf{w}^{e,k-1}$ 发送给客户端集合 $\mathcal{S}^{e,k}$中的客户端
    客户端 $\mathcal{S}^{e,k}$ 并行执行:
      无放回抽样 $\mathcal{R}_i^{e,k}$ ，$\mathcal{R}_i^{e,k}$，$\mathcal{R}_i^{e,k}$，$_1$，$\mathcal{R}_i^{e,k}$，$_2$ 来自 $\mathcal{H}_1^e$，$\mathcal{H}_2^e$，分别。
      发送 $\mathcal{R}_i^{e,k}$，$\mathcal{R}_i^{e,k}$，$\mathcal{R}_i^{e,k}$，$_1$，$\mathcal{R}_i^{e,k}$，$_2$ 发送至客户端 $i$，适用于所有 $i \in [N]$
      $\mathbf{w}_m^{e,k}$, $\boldsymbol{H}_{m,1}^{e,k}$, $\boldsymbol{H}_{m,2}^{e,k} \leftarrow$ 本地更新($\mathbf{w}^{e,k-1}$)
    添加 $\cup_{m\in\mathcal{G}^{e,k}} \boldsymbol{H}_{m,1}^{e,k}$ 到 $\mathcal{H}_1^{e+1}$和 $\cup_{m\in\mathcal{G}^{e,k}} \boldsymbol{H}_{m,2}^{e,k}$ 至 $\mathcal{H}_2^{e+1}$
    计算$\mathbf{w}^{e,k} = \frac{1}{|\mathcal{G}^{e,k}|}\sum_{m\in\mathcal{G}^{e,k}} \mathbf{w}_m^{e,k}$。
  **结束循环**
**结束循环**
返回 $\frac{1}{E}\sum_e \frac{1}{K}\sum_k \mathbf{w}_{m,K}^{e,k}$

17：客户端本地更新 18：客户端 $i$：需要参数 $\eta, K$19：初始化模型 $\mathrm{w}_{i,K}^{0}$ 并初始化缓冲区 $\mathrm{B}_{i,1}, \mathrm{B}_{i,2} = \emptyset$ 20：从服务器接收 $\bar{\mathrm{w}}^{e,k-1}$，并设置 $\mathrm{w}_{i,0}^{e,k} = \bar{\mathrm{w}}^{e,k-1}$21：从服务器接收 $\mathrm{R}_{i,1}^{e,k}, , , \mathrm{R}_{i,2}^{e,k}, ,$ 22：随机打乱 $\mathrm{R}_{i,1}^{e,k}, , , \mathrm{R}_{i,2}^{e,k}, ,$ 并将结果放入缓冲区 $\mathrm{B}_{i,1}, \mathrm{B}_{i,2}$ 23：从 $\mathrm{D}_1^i$中采样 $K$ 个点，使用模型 $\mathrm{w}_{i,K}^{0}$ 计算其预测，记为 $\mathrm{H}_{i,1}^{0}$ 24：从 $\mathrm{D}_2^i$中采样 $K$ 个点，使用模型 $\mathrm{w}_{i,K}^{0}$ 计算其预测，记为 $\mathrm{H}_{i,2}^{0}$ 25：对于 $k = 0, .., K-1$，执行 26：从 $\mathrm{D}_1^m$中采样 $\mathrm{z}_{m,t}^{e,k}, , ,1$，从 $\mathrm{D}_2^m$ ◇ 中采样 $\mathrm{z}_{m,t}^{e,k}, , ,2$，或采样两个数据小批次 27：分别从 $\mathrm{B}_{m,1}$ 和 $\mathrm{B}_{m,2}$ 中取出下一个 $h_\xi^{e-1}$ 和 $h_\zeta^{e-1}$ 28：计算 $h(\mathrm{w}_{m,t}^{e,k}, , ; \mathrm{z}_{m,t}^{e,k}, , ,1)$ 和 $h(\mathrm{w}_{m,t}^{e,k}, , ; \mathrm{z}_{m,t}^{e,k}, , ,2)$ 29：将 $h(\mathrm{w}_m^{e,k}, , ,k; \mathrm{z}_{m,t}^{e,k}, , ,1)$ 加入 $\mathrm{H}_m^{e,k}, , ,1$，并将 $h(\mathrm{w}_{m,t}^{e,k}, , ; \mathrm{z}_{m,t}^{e,k}, , ,2)$ 加入 $\mathrm{H}_m^{e,k}, , ,2$ 30：根据 (11) 和 (12) 计算 $G^{e,k}_{m,t}, , ,1$ 和 $G^{e,k}_{m,t}, , ,2$ 31：$\mathrm{w}_m^{e,k}, , ,_{t+1} = \mathrm{w}_m^{e,k}, , ,_t - \eta(G^{e,k}_{m,t}, , ,_1 + G^{e,k}_{m,t}, , ,2)$ 32：结束循环 33：将 $\mathrm{w}_m^{e,k}, , ,_t$ 发送至服务器 34：将 $\mathrm{H}_{m,1}^{e,k}, , , \mathrm{H}_m^{e,k}, , ,2$ 发送至服务器

**算法 4** PL条件下的联邦成对 AUC（CyCP-Pairwise）

初始化：$\mathbf{w}_0$
**对于** $s \in [S]$ **执行**
  $\mathbf{w}_s, \ \alpha_s = \mathrm{OSFP}(\mathbf{w}_{s-1}, \ \eta_s, \ E_s)$
**结束循环**
返回 $\mathbf{w}_S, \alpha_S$

**总体影响：**我们考虑的问题具有足够的普适性，可扩展至 AUC 最大化之外的应用，例如二部排序、度量学习以及其他客户端耦合任务。例如，考虑度量学习中常用的对比损失（Hadsell 等人，2006）：

$$\psi(h(\mathbf{w};\mathbf{z}_1),h(\mathbf{w};\mathbf{z}_2)) = \frac{1}{2}\,\delta(\mathbf{z}_1,\mathbf{z}_2)\,\|h(\mathbf{w};\mathbf{z}_1)-h(\mathbf{w};\mathbf{z}_2)\|^2+\frac{1}{2}\,(1-\delta(\mathbf{z}_1,\mathbf{z}_2))\,\big(\max(0,m-\|h(\mathbf{w};\mathbf{z}_1)-h(\mathbf{w};\mathbf{z}_2)\|)\big)^2, \tag{14}$$

其中 $\delta(\mathbf{z}_1,\mathbf{z}_2) = 1$ 如果$\mathbf{z}_1$和$\mathbf{z}_2$相似，0 否则，且 $h(\mathbf{w};\ \mathbf{z})$ 表示模型$\mathbf{w}$生成的数据点$\mathbf{w}$的嵌入。若数据对分布在多个客户端上，该损失自然符合问题9 中的形式化定义。

隐私考量。本文提出的两种算法框架要求客户端共享模型参数；对于成对目标，还需额外共享预测分数，以便进行协同优化，这与以往研究一致（Guo 等，2023a；McMahan 等，2017）。然而，此类信息交换可能带来隐私风险，因为模型更新及相关信号可能泄露有关单个数据点的信息（Zhu 等，2019）。可通过多种技术降低这些风险，包括：1）添加噪声以实现差分隐私（Abadi 等，2016；McMahan 等，2018；Truex 等，2020；Wei 等，2020）；2）量化（Kang 等，2024；Youn 等，2023；Xu 等，2025）；3）丢弃法（Jain 等，2015）；以及 4）在聚合过程中使用同态加密（Jin 等，2023；Fang 和 Qian，2021）。

## 6 实验

我们在四个不同的数据集上评估所提出的方法：CIFAR-10、CIFAR-100、ChestMNIST 和保险欺诈数据集，以展示所提出的算法在具有挑战性的非独立同分布和类别不平衡条件下，跨领域（计算机视觉、医学影像和表格数据）的鲁棒性。

对于 CIFAR-10 和 CIFAR-100（Krizhevsky 等人，2009年），我们将多分类任务重新构建为二元分类问题，指定一个类别为正标签，其余类别为负标签。为引入类别不平衡，我们分别移除 CIFAR-10 中 95% 和 CIFAR-100 中 50% 的正样本。ChestMNIST（Yang 等人，2021b）被设定为一项二元任务，用于预测是否存在肿块。我们使用浓度参数为 $dir$ 的狄利克雷分布，将训练集分配给 100 个客户端，以模拟异质性。CIFAR-10/100 使用 ResNet18（He 等人，2016年），ChestMNIST 则使用 DenseNet121 Huang 等人（2017年）。

保险欺诈数据集由美国医疗保险和医疗补助服务中心（CMS）提供的 Medicare B 部分理赔数据（2017—2022年）构建（美国医疗保险和医疗补助服务中心（CMS），2016年）。根据排除个人与实体名单（LEIE）对提供者进行标记，识别其中的欺诈者（美国卫生与公众服务部监察长办公室，2016年）。每个美国州均视为一个独立客户端，因此共有 50 个客户端，且数据天然具有异质性。数据按时间顺序划分：2017—2020年用于训练，2021年用于验证，2022年用于测试。保险数据采用线性模型。

我们将所提方法与采用循环参与机制的基线方法进行比较：CyCp-FedAVG（Cho等人，2023）、放大 FedAVG（A-FedAVG）（Wang & Ji，2022年）和放大 SCAFFOLD（A-SCAFFOLD）（Crawshaw &Liu，2024）。我们还纳入了随机采样极小极大（RS-Minimax）和随机采样 PSM（RS-PSM）。这两种方法分别基于Guo 等人（2020年）和Guo 等人（2023a）的方法改编而来，并在每轮中随机采样客户端。步长在 [1e-1、1e-2、1e-3] 中调优。我们使用代理损失 $\psi(h1,\ h2) = 1+\frac{1}{\exp((h1-h2)/\lambda)}$，其中缩放因子 $\lambda$ 在 [1, 1e-1、1e-2、1e-3] 中调优。极小极大 AUC 的正则化因子 $\rho$ 在 [1e-1、1e-2、1e-3] 中调优。所有算法均训练 100 轮；分阶段算法的初始阶段长度 [1,10,50], 、衰减因子 [0.1、0.2、0.5], 和阶段缩放 [2,5,10]均经过调优。所有表格中报告的结果均为某算法运行三次所得结果的均值和标准差

我们首先设置 dir= 0.5 用于 CIFAR-10/100，并评估两种设置：(1) 原始标签，使用自然的不平衡标签；(2) 翻转标签，其中 20% 的正标签被随机翻转为负标签。测试 AUC 结果如表 1 和 2所示。我们提出的方法（极小极大和 PSM）在各个数据集和设置下始终优于所有基线方法。例如，在 CIFAR-100 上，PSM 比最佳基线方法的 AUC 提高了 5–6%；在保险数据集上，PSM 比 FedAVG 高出 4% 以上。在标签翻转情况下，极小极大和 PSM 也显著优于基线方法。我们提出的方法 CyCP-Minimax 和

CyCP-PSM 在大多数情况下都优于这些随机采样基线，因为随机采样无法确保总体覆盖，而总体覆盖对于减轻非独立同分布偏差至关重要。我们还强调，随机采样隐含地假设所有被采样的客户端都能够参与，而这一假设在现实世界的联邦环境中往往难以成立。相比之下，循环客户端参与通过利用可预测的参与模式克服了这些局限。

**对异质性和标签噪声的敏感性** 我们进一步考察异质性（Dirichlet 集中度 $dir$）和标签噪声（翻转比例 $flip$）对 CIFAR-10、CIFAR-100 和 ChestMNIST 的影响（表格3、4 和5）。整体趋势一致：1）随着异质性增加（Dirichlet 参数 $\alpha$减小），基线方法的曲线下面积显著下降，而极小极大和 PSM 保持稳定。2）在标签翻转率较高时（$flip$=0.2），我们的方法仍能保持有竞争力的表现，有时甚至超过干净标签基线（例如 PS M 在 CIFAR-10 上的表现）。3）这些观察结果证实，尽管存在数据不一致和标签污染，所提出的算法仍能更好地协调客户端目标，从而验证了其针对鲁棒联邦曲线下面积优化的理论设计。

**通信效率消融研究** 图1 展示了针对通信间隔 $I$的消融研究，该间隔控制同步前的本地更新次数。即使 $I$ 增加，极小极大和 PSM 仍能保持稳定的测试 AUC，表明所提出的方法能够容忍频繁通信，而不会损失收敛质量。这一特性使其非常适用于带宽受限的联邦环境，因为在这些环境中，通信成本占据训练时间的主要部分。

总体而言，实验结果表明：1）性能：在多种模态上，极小极大和 PSM 相较于最先进的基线方法均取得了稳定且显著的曲线下面积提升。2）鲁棒性：在严重的类别不平衡和标签噪声条件下，它们仍保持良好的适应能力。3）可扩展性：这些方法能够有效扩展到大规模异质客户端群体，并且能够容忍稀疏通信。综上，这些结果证实了我们提出的算法在涉及非独立同分布、不平衡和含噪数据分布的现实联邦学习场景中的实际优势。

表 1：原始标签上的实验结果。（Cifar-10/100 的 $dir$=0.5）

| | Cifar-10 | Cifar-100 | ChestMNIST | 保险 |
|---|---|---|---|---|
| CyCp-FedAVG | $0.7836 \pm 0.0020$ | $0.8894 \pm 0.0012$ | $0.6012 \pm 0.0016$ | $0.7339 \pm 0.0035$ |
| A-FedAVG | $0.7950 \pm 0.0014$ | $0.9137 \pm 0.0028$ | $0.5994 \pm 0.0025$ | $0.7384 \pm 0.0030$ |
| A-SCAFFOLD | $0.7993 \pm 0.0018$ | $0.9105 \pm 0.0017$ | $0.6015 \pm 0.0022$ | $0.7412 \pm 0.0021$ |
| RS-Minimax | $0.8153 \pm 0.0025$ | $0.9392 \pm 0.0029$ | $0.5697 \pm 0.0031$ | $0.7419 \pm 0.0013$ |
| RS-Pairwise | $0.8357 \pm 0.0012$ | $\mathbf{0.9688 \pm 0.0010}$ | $0.6118 \pm 0.0020$ | $\mathbf{0.7619 \pm 0.0011}$ |
| CyCp-Minimax | $\mathbf{0.8446 \pm 0.0015}$ | $0.9318 \pm 0.0014$ | $\mathbf{0.6163 \pm 0.0026}$ | $0.7616 \pm 0.0017$ |
| CyCp-Pairwise | $\mathbf{0.8480 \pm 0.0008}$ | $\mathbf{0.9694 \pm 0.0013}$ | $\mathbf{0.6256 \pm 0.0015}$ | $\mathbf{0.7778 \pm 0.0002}$ |

T表 2：翻转标签下的实验结果（Cifar-10/100 为 $dir$=0.5， $flip$=20%）

| | Cifar-10 | Cifar-100 | ChestMNIST | 保险 |
|---|---|---|---|---|
| CyCp-FedAVG | $0.7739 \pm 0.0031$ | $0.8781 \pm 0.0019$ | $0.5927 \pm 0.0008$ | $0.7292 \pm 0.0029$ |
| A-FedAVG | $0.7891 \pm 0.0020$ | $0.8951 \pm 0.0023$ | $0.5901 \pm 0.0013$ | $0.7298 \pm 0.0035$ |
| A-SCAFFOLD | $0.7950 \pm 0.0025$ | $0.9066 \pm 0.0024$ | $0.5987 \pm 0.0011$ | $0.7307 \pm 0.0036$ |
| RS-Minimax | $0.8170 \pm 0.0021$ | $0.9185 \pm 0.0019$ | $0.5587 \pm 0.0026$ | $\mathbf{0.7517 \pm 0.0038}$ |
| RS-Pairwise | $0.8201 \pm 0.0017$ | $\mathbf{0.9480 \pm 0.0020}$ | $0.6004 \pm 0.0009$ | $0.7173 \pm 0.0043$ |
| 极小极大 | $\mathbf{0.8316 \pm 0.0006}$ | $0.9275 \pm 0.0015$ | $\mathbf{0.6150 \pm 0.0007}$ | $0.7505 \pm 0.0041$ |
| CyCp-Pairwise | $\mathbf{0.8424 \pm 0.0038}$ | $\mathbf{0.9455 \pm 0.0020}$ | $\mathbf{0.6077 \pm 0.0012}$ | $\mathbf{0.7722 \pm 0.0018}$ |

我们在图1中展示了关于通信间隔的消融实验。我们可以看到，算法能够容忍大量的本地更新步数 $I$ 而不会导致性能下降，这验证了所提出算法的通信效率。

有关更多消融实验，请参阅附录F。

表3：不同狄利克雷参数 $dir$ 和翻转比例 $flip$ 下在 Cifar-10 上的结果

| | $dir$=0.1, $flip$=0 | $dir$=0.1, $flip$=0.2 | $dir$=10, $flip$=0 | $dir$=10, $flip$=0.2 |
|---|---|---|---|---|
| CyCp-FedAVG | $0.7581 \pm 0.0023$ | $0.6800 \pm 0.0030$ | $0.8227 \pm 0.0025$ | $0.8199 \pm 0.0019$ |
| A-FedAVG | $0.7732 \pm 0.0029$ | $0.7015 \pm 0.0028$ | $0.8354 \pm 0.0033$ | $0.8254 \pm 0.0015$ |
| A-SCAFFOLD | $0.7815 \pm 0.0016$ | $0.7119 \pm 0.0024$ | $0.8336 \pm 0.0019$ | $0.8315 \pm 0.0011$ |
| RS-Minimax | $0.7976 \pm 0.0021$ | $\mathbf{0.8083 \pm 0.0025}$ | $0.8154 \pm 0.0031$ | $0.8086 \pm 0.0016$ |
| RS-Pairwise | $0.8059 \pm 0.0014$ | $0.7712 \pm 0.0027$ | $0.8304 \pm 0.0032$ | $0.8217 \pm 0.0023$ |
| CyCp-Minimax | $\mathbf{0.8184 \pm 0.0018}$ | $\mathbf{0.8095 \pm 0.0017}$ | $\mathbf{0.8702 \pm 0.0023}$ | $\mathbf{0.8511 \pm 0.0008}$ |
| CyCp-Pairwise | $\mathbf{0.8283 \pm 0.0012}$ | $0.7828 \pm 0.0020$ | $\mathbf{0.8626 \pm 0.0027}$ | $\mathbf{0.8540 \pm 0.0014}$ |

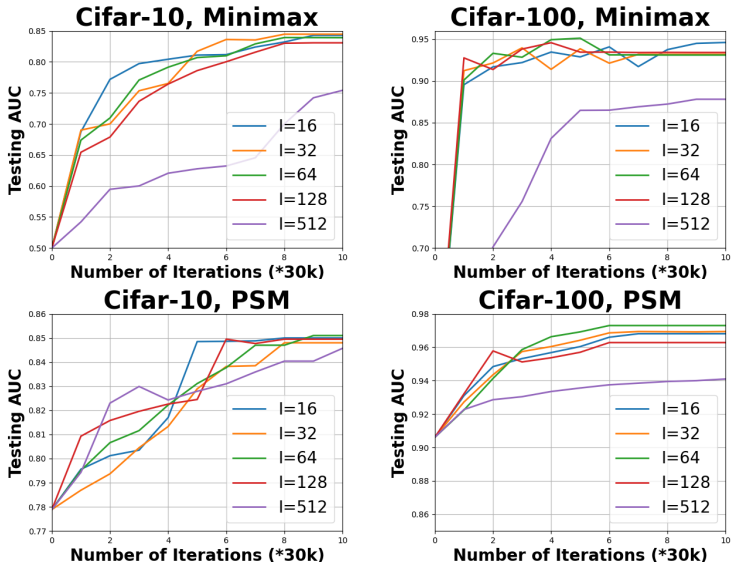



图1：消融研究：通信间隔 $I$的影响。

表4：狄利克雷参数 $dir$ 和翻转比例 $flip$ 不同情况下 Cifar-1 上的结果 00

| | $dir$=0.1,$flip$=0 | $dir$=0.1,$flip$=0.2 | $dir$=10,$flip$=0 | $dir$=10,$flip$=0.2 |
|---|---|---|---|---|
| CyCp-FedAVG | $0.9203 \pm 0.0009$ | $0.8774 \pm 0.0018$ | $0.9587 \pm 0.0025$ | $0.9363 \pm 0.0024$ |
| A-FedAVG | $0.9230 \pm 0.0014$ | $0.8825 \pm 0.0026$ | $0.9605 \pm 0.0030$ | $0.9402 \pm 0.0023$ |
| A-SCAFFOLD | $0.9199 \pm 0.0021$ | $0.8906 \pm 0.0015$ | $\mathbf{0.9627 \pm 0.0027}$ | $0.9421 \pm 0.0016$ |
| RS-Minimax | $0.9216 \pm 0.0011$ | $0.9123 \pm 0.0022$ | $0.9520 \pm 0.0018$ | $0.9266 \pm 0.0030$ |
| RS-Pairwise | $0.9243 \pm 0.0015$ | $\mathbf{0.9180 \pm 0.0024}$ | $0.9601 \pm 0.0017$ | $0.9475 \pm 0.0026$ |
| CyCp-Minimax | $\mathbf{0.9299 \pm 0.0013}$ | $\mathbf{0.9148 \pm 0.0017}$ | $0.9610 \pm 0.0023$ | $\mathbf{0.9564 \pm 0.0019}$ |
| CyCp-Pairwise | $\mathbf{0.9308 \pm 0.0016}$ | $0.9123 \pm 0.0028$ | $\mathbf{0.9669 \pm 0.0015}$ | $\mathbf{0.9525 \pm 0.0021}$ |

表 5：ChestMNIST 上不同狄利克雷参数 $dir$ 和翻转比例 $flip$ 下的结果

| | $dir$=0.1,$flip$=0 | $dir$=0.1,$flip$=0.2 | $dir$=10,$flip$=0 | $dir$=10,$flip$=0.2 |
|---|---|---|---|---|
| CyCp-FedAVG | $\mathbf{0.5584 \pm 0.0006}$ | $0.5329 \pm 0.0003$ | $0.6561 \pm 0.0013$ | $0.6197 \pm 0.0017$ |
| A-FedAVG | $0.5580 \pm 0.0004$ | $0.5334 \pm 0.0006$ | $\mathbf{0.6572 \pm 0.0015}$ | $0.6201 \pm 0.0013$ |
| A-SCAFFOLD | $0.5573 \pm 0.0004$ | $0.5341 \pm 0.0002$ | $0.6568 \pm 0.0016$ | $0.6233 \pm 0.0019$ |
| RS-Minimax | $0.5245 \pm 0.0011$ | $0.5155 \pm 0.0004$ | $0.6541 \pm 0.0008$ | $0.6232 \pm 0.0017$ |
| RS-Pairwise | $0.5502 \pm 0.0007$ | $0.5338 \pm 0.0009$ | $0.6545 \pm 0.0012$ | $0.6293 \pm 0.0014$ |
| CyCp-Minimax | $0.5553 \pm 0.0010$ | $\mathbf{0.5362 \pm 0.0005}$ | $0.6431 \pm 0.0012$ | $\mathbf{0.6318 \pm 0.0018}$ |
| CyCp-Pairwise | $\mathbf{0.5598 \pm 0.0009}$ | $\mathbf{0.5498 \pm 0.0004}$ | $\mathbf{0.6595 \pm 0.0010}$ | $\mathbf{0.6356 \pm 0.0023}$ |

## 7 结论

本文研究非经验风险最小化目标的联邦优化，重点关注循环客户端参与这一实际约束下的曲线下面积最大化问题。我们提出了适用于这一场景的通信高效算法。对于极小极大重构的 AUC，我们引入了一种带辅助序列的分阶段方法，以处理循环依赖。对于一般的成对损失，我们开发了一种主动-被动策略，在客户端之间共享预测分数。我们为迭代复杂度和通信复杂度提供了理论保证，并在涉及 50–100 个客户端的多种任务上验证了所提算法的有效性。

## 8 局限性与未来工作

我们的工作存在若干局限性。首先，$\widetilde{O}(1/\epsilon^{1/2})$ 的快速通信速率依赖于 PL 条件，这可能限制相关算法的泛化能力。其次，无论是在 PL 还是非 PL 情形下，收敛速率都尚未达到已知下界，因此仍有改进空间。第三，如何开发能够正式保证差分隐私的方法，仍是一个悬而未决的问题。最后，将我们的方法从曲线下面积最大化拓展到更广泛的非经验风险最小化目标，是未来研究的一个重要方向。

## 参考文献

Martin Abadi、Andy Chu、Ian Goodfellow、H BrendanMcMahan、Ilya Mironov、Kunal Talwar 和 Li Zhang。差分隐私深度学习。载于*2016*年 *ACM SIGSAC* 计算机与通信安全会议论文集，第 308–318 页，2016年。

Toon Calders 和 Szymon Jaroszewicz。面向分类的高效 AUC 优化。载于数据库中的知识发现：*PKDD 2007*，第*11*届欧洲数据库知识发现原理与实践会议，波兰华沙，*2007*年*9*月*17*日至*21*日，会议论文集，第《计算机科学讲义》4702卷，第42–53页。Springer，2007年。

美国医疗保险和医疗补助服务中心（CMS）。医疗保险提供者利用与支付数据：医师及其他供应商。2016年。访问日期：2024年3月25日。

Nontawat Charoenphakdee、Jongyeong Lee 和 Masashi Sugiyama。利用受污染标签进行学习的对称损失。载于第*36*届国际机器学习会议，*ICML 2019*，*2019*年*6*月*9*日至*15*日，美国加利福尼亚州长滩，第97卷《机器学习研究论文集》，第961–970页。PMLR，2019a。

Nontawat Charoenphakdee、Jongyeong Lee 和 Masashi Sugiyama。关于使用受污染标签进行学习的对称损失。载于国际机器学习会议，第 961–970 页。PMLR，2019b。

Yae Jee Cho、Jianyu Wang 和 Gauri Joshi。联邦学习中的客户端选择：收敛分析与择优选择策略。*arXiv* 预印本 *arXiv:2010.01243*，2020年。

Yae Jee Cho、Pranay Sharma、Gauri Joshi、Zheng Xu、Satyen Kale 和 Tong Zhang。关于循环客户端参与下联邦平均的收敛性。载于国际机器学习会议，第 5677–5721 页。PMLR，2023。

Stéphan Clémençon、GáborLugosi 和 NicolasVayatis。U统计量的排序与经验最小化。统计年鉴，36(2)：844–874，2008年。

William W Cohen、Robert E Schapire 和 Yoram Singer。学习排序。神经信息处理系统进展，10，1997。

Corinna Cortes 和 Mehryar Mohri。曲线下面积优化与错误率最小化。神经信息处理系统进展，16，2003。

Michael Crawshaw 和 Mingrui Liu。周期性客户端参与和异质数据下的联邦学习：一种新的通信高效算法及其分析。神经信息处理系统进展，37：8240–8299，2024。

Krzysztof Dembczynski、Wojciech Kotlowski 和 Eyke Hüllermeier。通过单变量损失实现一致的多标签排序。*arXiv* 预印本 *arXiv:1206.6401*，2012 年。Yuyang Deng 和 Mehrdad Mahdavi。局部随机梯度下降上升：收敛分析与通信效率。载于国际人工智能与统计学会议，第 1387–1395 页。PMLR，2021 年。Charles Elkan。代价敏感学习的基础。载于国际人工智能联合会议，第 17 卷，第 973–978 页。Lawrence Erlbaum Associates Ltd，2001 年。Haokun Fang 和 Quan Qian。基于同态加密和联邦学习的隐私保护机器学习。未来互联网，13(4):94，2021 年。Hongchang Gao、Junyi Li 和 Heng Huang。关于带动量的局部随机复合梯度下降的收敛性。载于国际机器学习会议，第 7017–7035 页。PMLR，2022 年。

Wei Gao 和 Zhi-Hua Zhou。论 AUC成对优化的一致性。载于 Qiang Yang 和 MichaelJ. Wooldridge（编），第二十四届国际人工智能联合会议论文集，*IJCAI 2015*，阿根廷布宜诺斯艾利斯，*2015*年*7*月*25*日至*31*日，第 939–945 页。AAAI出版社，2015年。

Wei Gao、Rong Jin、Shenghuo Zhu 和 Zhi-Hua Zhou。单遍 AUC 优化。载于 国际机器学习会议（第 *3* 卷），第 906–914 页，2013年。

Jack Goetz、Kshitiz Malik、Duc Bui、Seungwhan Moon、Honglei Liu 和 Anuj Kumar。主动联邦学习。*arXiv* 预印本 *arXiv:1909.12641*，2019年。

Zhishuai Guo、Mingrui Liu、Zhuoning Yuan、Li Shen、Wei Liu 和 Tianbao Yang。基于深度神经网络的通信高效分布式随机AUC最大化。载于国际机器学习会议，第 3864–3874 页。PMLR，2020年。

Zhishuai Guo、Rong Jin、Jiebo Luo 和 Tianbao Yang。Fedxl：用于深度X风险优化的可证明联邦学习。载于《国际机器学习会议》，第 11934–11966 页。PMLR，2023a。

Zhishuai Guo、Yan Yan、Zhuoning Yuan 和 Tianbao Yang。满足 PL条件的非凸强凹极小极大问题的快速目标函数与对偶间隙收敛。机器学习研究期刊，24(148):1–63，2023b。Raia Hadsell、Sumit Chopra 和 Yann LeCun。通过学习不变映射进行降维。载于*2006*年*6*月*17*日至*22*日，美国纽约，*2006*年*IEEE* 计算机学会计算机视觉与模式识别会议（*CVPR 2006*），第1735–1742页。电气电子工程师学会计算机学会，2006年。doi: 10。1109/CVPR。2006。100。网址 https://doi.org/10.1109/CVPR.2006.100。Andrew Hard、Kanishka Rao、Rajiv Mathews、Swaroop Ramaswamy、Françoise Beaufays、Sean Augenstein、Hubert Eichner、Chloé Kiddon 和 Daniel Ramage。用于移动键盘预测的联邦学习。*arXiv*预印本 *arXiv:1811.03604*，2018年。Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。用于图像识别的深度残差学习。载于*IEEE*计算机视觉与模式识别会议论文集，第770–778页，2016年。Gao Huang、Zhuang Liu、Laurens Van Der Maaten 和 Kilian Q Weinberger。密集连接卷积网络。载于*IEEE*计算机视觉与模式识别会议论文集，第4700–4708页，2017年。

DzmitryHuba、John Nguyen、KshitizMalik、Ruiyu Zhu、Mike Rabbat、Ashkan Yousefpour、Carole-Jean Wu、Hongyuan Zhan、Pavel Ustinov、Harish Srinivas、Kaikai Wang、Anthony Shoumikhin、Jesik Min 和 Mani Malek。PAPAYA：实用、私密且可扩展的联邦学习。载于 Diana Marculescu、Yuejie Chi 和 Carole-Jean Wu（编），第五届机器学习与系统会议论文集，*MLSys 2022*，圣克拉拉，美国，加利福尼亚州，*2022*年*8*月*29*日至*9*月*1*日。mlsys.org，2022年。网址 https://proceedings.mlsys.org/paper_files/paper/2022/hash/a8bc4cb14a20f20d1f96188bd61eec87-Abstract.html。

Prateek Jain、Vivek Kulkarni、Abhradeep Thakurta 和 Oliver Williams。该不该丢弃：丢弃法的鲁棒性、一致性和差分隐私特性。*CoRR*，abs/1503.02031，2015年。URLhttp://arxiv.org/abs/1503.02031。

Divyansh Jhunjhunwala、Pranay Sharma、Aushim Nagarkatti 和 Gauri Joshi。Fedvarp：解决联邦学习中客户端部分参与导致的方差问题。载于人工智能中的不确定性，第 906–916 页。PMLR，2022年。Weizhao Jin、Yuhang Yao、Shanshan Han、Jiajun Gu、Carlee Joe-Wong、Srivatsan Ravi、Salman Avestimehr 和 Chaoyang He。Fedml-he：一种基于同态加密的高效隐私保护联邦学习系统。*arXiv* 预印本 *arXiv:2303.10837*，2023。Peter Kairouz、Brendan McMahan、Shuang Song、Om Thakkar、Abhradeep Thakurta 和 Zheng Xu。无需采样或打乱的实用且隐私保护的（深度）学习。载于 Marina Meila 和 Tong Zhang （编），第 *38* 届国际机器学习会议论文集（*ICML 2021*），*2021* 年 *7* 月 *18*日至*24*日，线上会议，第139卷，《机器学习研究论文集》，第 5213–5225 页。PMLR，2021a。网址：http://proceedings.mlr.press/v139/kairouz21b.html。Peter Kairouz、Brendan McMahan、Shuang Song、Om Thakkar、Abhradeep Thakurta 和 Zheng Xu。无需采样或打乱的实用且隐私保护的（深度）学习。载于国际机器学习会议，第 5213–5225 页。PMLR，2021b。Peter Kairouz、H Brendan McMahan、Brendan Avent、Aurélien Bellet、Mehdi Bennis、Arjun Nitin Bhagoji、Kallista Bonawitz、Zachary Charles、Graham Cormode、Rachel Cummings 等。联邦学习的进展与未决问题。基础与趋势® 机器学习，14(1–2)：1–210，2021c。Tianqu Kang、Lumin Liu、Hengtao He、Jun Zhang、Shenghui Song 和 Khaled B. Letaief。联邦学习中量化的影响：Rényi差分隐私视角。载于*IEEE*国际地中海通信与网络会议（*MeditCom 2024*），西班牙马德里，*2024*年*7*月*8*日至*11*日，第 233–238 页。IEEE，2024。doi：10.1109/MEDITCOM61057.2024.10621075。网址 https://doi.org/10.1109/MeditCom61057.2024.10621075。Sai Praneeth Karimireddy、Satyen Kale、Mehryar Mohri、Sashank Reddi、Sebastian Stich 和 Ananda Theertha Suresh。Scaffold：用于联邦学习的随机控制平均。载于国际机器学习会议，第 5132–5143 页。PMLR，2020年。Jakub Konečn`y、H Brendan McMahan、Daniel Ramage 和 Peter Richtárik。联邦优化：用于端侧智能的分布式机器学习。*arXiv* 预印本 *arXiv:1610.02527*，2016年。Wojciech Kotlowski、Krzysztof Dembczynski 和 Eyke Hüllermeier。通过最小化单变量损失进行二部排序。载于第*28*届国际机器学习会议论文集（*ICML 2011*），美国华盛顿州贝尔维尤，*2011*年*6*月*28*日至*7*月*2*日，第 1113–1120 页。Omnipress，2011年。Alex Krizhevsky、Geoffrey Hinton 等。利用微小图像学习多层特征。（2009），2009年。Junyi Li 和 Heng Huang。Fedgrec：通过潜在嵌入的延迟更新实现的联邦图推荐系统。*arXiv* 预印本 *arXiv:2210.13686*，2022年。Brendan McMahan、Eider Moore、Daniel Ramage、Seth Hampson 和 Blaise Aguera y Arcas。基于去中心化数据的深度网络通信高效学习。载于人工智能与统计学，第 1273–1282 页。PMLR，2017年。

H Brendan McMahan、Daniel Ramage、Kunal Talwar 和 Li Zhang。差分隐私循环语言模型的学习。载于国际学习表征会议，2018年。Mehryar Mohri、Gary Sivek 和 Ananda Theertha Suresh。无关先验联邦学习。载于国际机器学习会议，第 4615–4625 页。PMLR，2019年。Hongseok Namkoong 和 John C Duchi。采用 f-散度进行分布鲁棒优化的随机梯度方法。载于神经信息处理系统进展（*NeurIPS*），第 2208–2216 页，2016年。Yurii E. Nesterov。凸优化导论——基础课程，第87卷，应用优化。Springer，2004年。

美国卫生与公众服务部监察长办公室。LEIE：监察长办公室 LEIE可下载数据库。https://oig.hhs.gov/exclusions/authorities.asp，2016年。访问日期：2024-03- 25。

Sarthak Pati、Ujjwal Baid、Brandon Edwards、Micah Sheller、Shih-Han Wang、G Anthony Reina、Patrick Foley、Alexey Gruzdev、Deepthi Karkada、Christos Davatzikos 等。联邦学习助力大数据实现罕见癌症边界检测。《自然·通讯》，13(1):7346，2022年。

Matthias Paulik、Matt Seigel、Henry Mason、Dominic Telaar、Joris Kluivers、Rogier C. van Dalen、Chi Wai Lau、Luke Carlson、Filip Granqvist、Chris Vandevelde、Sudeep Agarwal、Julien Freudiger、Andrew Byde、Abhishek Bhowmick、Gaurav Kapoor、Si Beaumont、Áine Cahill、Dominic Hughes、Omid Javidbakht、Fei Dong、Rehan Rishi 和 Stanley Hung。面向设备端个性化的联邦评估与调优：系统设计与应用。*CoRR*，abs/2102.08503，2021年。URL https://arxiv.org/abs/2102.08503。

Filip Radenović、Giorgos Tolias 和 Ondřej Chum。CNN图像检索从词袋中学习：使用困难样本进行无监督微调。载于欧洲计算机视觉会议，第 3–20 页。Springer，2016年。

Alain Rakotomamonjy。使用支持向量机优化 ROC曲线下面积。载于*ROCAI*，第 71–80 页，2004年。

Yichen Ruan、Xiaoxi Zhang、Shu-Che Liang 和 Carlee Joe-Wong。迈向联邦学习中的灵活设备参与。载于国际人工智能与统计学会议，第 3403–3411 页。PMLR，2021年。

Pranay Sharma、Rohan Panda、Gauri Joshi 和 Pramod Varshney。联邦极小极大优化：改进的收敛分析与算法。载于国际机器学习会议，第 19683–19730 页。PMLR，2022年。

Sebastian U Stich。本地随机梯度下降收敛迅速，通信开销小。发表于国际学习表征会议，2018年。

Stacey Truex、Ling Liu、Ka-Ho Chow、Mehmet Emre Gursoy 和 Wenqi Wei。LDP-Fed：采用本地差分隐私的联邦学习。发表于第三届 *ACM* 边缘系统、分析与网络国际研讨会论文集，第 61–66 页，2020年。

Shiqiang Wang 和 Mingyue Ji。对任意客户端参与情况下的联邦学习进行统一分析。神经信息处理系统进展，35：19124–19137，2022年。Kang Wei、Jun Li、Ming Ding、Chuan Ma、Howard H Yang、Farhad Farokhi、Shi Jin、Tony QS Quek 和 H Vincent Poor。差分隐私联邦学习：算法与性能分析。*IEEE*《信息取证与安全汇刊》，15：3454–3469，2020年。

Chao-Yuan Wu、R Manmatha、Alexander J Smola 和 Philipp Krahenbuhl。深度嵌入学习中的采样至关重要。载于*IEEE*国际计算机视觉大会论文集，第 2840–2848 页，2017年。

Yawen Wu、Zhepeng Wang、Dewen Zeng、Meng Li、Yiyu Shi 和 Jingtong Hu。结合特征融合与邻域匹配的联邦对比表示学习，2022年。网址 https://openreview.net/forum?id=6LNPEcJAGWe。

Zhibo Xu、Jianhao Zhu、Jingwen Xu、Changze Lv、Zisu Huang、Xiaohua Wang、Muling Wu、Qi Qian、Xiaoqing Zheng 和 Xuanjing Huang。通过随机掩码和量化增强联邦学习中的模型隐私。*CoRR*，abs/2508.18911，2025年。doi：10.48550/ARXIV.2508.18911。URL https://doi.org/10.48550/arXiv.2508.18911。Yan Yan、Yi Xu、Qihang Lin、Wei Liu 和 Tianbao Yang。用于极小极大优化的最优轮次随机梯度下降上升方法。载于神经信息处理系统进展 *33 (NeurIPS)*，2020年。Haibo Yang、Minghong Fang 和 Jia Liu。通过非独立同分布联邦学习中的部分工作节点参与实现线性加速。*arXiv* 预印本 *arXiv:2101.11203*，2021a。Jiancheng Yang、Rui Shi、Donglai Wei、Zequan Liu、Lin Zhao、Bilian Ke、Hanspeter Pfister 和 Bingbing Ni。MedMNIST v2：用于二维和三维生物医学图像分类的大规模轻量级基准。*arXiv* 预印本 *arXiv:2110.14795*，2021b。Timothy Yang、Galen Andrew、Hubert Eichner、Haicheng Sun、Wei Li、Nicholas Kong、Daniel Ramage 和 Françoise Beaufays。应用联邦学习：改进 Google 键盘的查询建议。*CoRR*，abs/1812.02903，2018年。URL http://arxiv.org/abs/1812.02903。Zhenhuan Yang、Yunwen Lei、Puyu Wang、Tianbao Yang 和 Yiming Ying。用于成对学习的简单随机梯度下降和在线梯度下降算法。神经信息处理系统进展，34：20160–20171，2021c。Yiming Ying、Longyin Wen 和 Siwei Lyu。随机在线 AUC 最大化。载于神经信息处理系统进展，第 451–459 页，2016年。Yeojoon Youn、Zihao Hu、Juba Ziani 和 Jacob D. Abernethy。随机量化是联邦学习差分隐私所需的一切。*CoRR*，abs/2306.11913，2023。doi：10.48550/ARXIV.2306.11913。URL https://doi.org/10.48550/arXiv.2306.11913。Hao Yu、Rong Jin 和 Sen Yang。关于分布式非凸优化中通信高效动量随机梯度下降的线性加速分析。载于国际机器学习会议，第 7184–7193 页。PMLR，2019a。Hao Yu、Sen Yang 和 Shenghuo Zhu。采用更快收敛和更少通信的并行重启随机梯度下降：揭示模型平均为何适用于深度学习。载于人工智能促进协会年会论文集，第 5693–5700 页，2019b。Zhuoning Yuan、Zhishuai Guo、Yi Xu、Yiming Ying 和 Tianbao Yang。面向异构数据的联邦深度 AUC 最大化，具有常数通信复杂度。载于国际机器学习会议，第 12219–12229 页。PMLR，2021a。Zhuoning Yuan、Yan Yan、Milan Sonka 和 Tianbao Yang。大规模鲁棒深度 AUC 最大化：一种新的代理损失及医学图像分类实证研究。载于*IEEE/CVF* 国际计算机视觉会议论文集，2021b。Peilin Zhao、Steven C. H. Hoi、Rong Jin 和 Tianbao Yang。在线 AUC 最大化。载于第 *28* 届国际机器学习会议，*ICML 2011*，美国华盛顿州贝尔维尤，*2011*年*6*月*28*日至*7*月*2*日，第 233–240 页，2011a。Peilin Zhao、Steven CH Hoi、Rong Jin 和 Tianbo Yang。在线 AUC 最大化。2011b。Dixian Zhu、Xiaodong Wu 和 Tianbao Yang。深度 AUROC 优化基准测试：损失函数与算法选择。*CoRR*，abs/2203.14177，2022年。doi：10.48550/ARXIV.2203.14177。URL https://doi.org/10.48550/arXiv.2203.14177。

Feng Zhu、Jingjing Zhang 和 Xin Wang。采用基于工作节点年龄的选择机制的通信高效本地随机梯度下降。超级计算杂志，79(12)：13794–13816，2023年。doi：10.1007/S11227-023-05190-7。URL https://doi.org/10.1007/s11227-023-05190-7。

Ligeng Zhu、Zhijian Liu 和 Song Han。《梯度泄露》。载于 Hanna M. Wallach、Hugo Larochelle、Alina Beygelzimer、Florence d'Alché-Buc、Emily B. Fox 和 Roman Garnett（编），神经信息处理系统进展第*32*卷：*2019*年神经信息处理系统年度大会，*NeurIPS 2019*，*2019*年*12*月*8*日至*14*日，加拿大不列颠哥伦比亚省温哥华，第14747–14756页，2019年。URL https://proceedings.neurips.cc/paper/2019/hash/60a6c4002cc7b29142def8871531281a-Abstract.html。

# A 符号说明

表 6：符号说明

| | |
|---|---|
| $\mathbf{w} \in \mathbb{R}^d$<br>$\mathbf{w}_{m,t}^{e,k} \in \mathbb{R}^d$ | 神经网络中待训练的模型参数和变量<br>第 $e$ 轮次、第 $t$ 次迭代时，组 $k$ 中机器 $m$ 的模型参数；上下文明确时省略阶段索引 |
| $a, b, \alpha \in \mathbb{R}^1$<br>$\mathbf{v} = (\mathbf{w}\ a,\ b) \in \mathbb{R}^{d+2}$<br>$\mathbf{w}^{e,k},\ \mathbf{v}^{e,k},\ \alpha^{e,k}$ | 极小极大目标函数中的引入变量<br>极小极大目标函数中的原变量。<br>组/轮次 $k$（轮次 $e$）的输出。所有参与客户端的平均值。 |
| $\mathbf{z}$<br>$\mathbf{z}_i$<br>$\mathbf{z}_{m,t}^{e,k}$<br>$\mathbf{z}^{e,k}_{m,t,\ 1}, \mathbf{z}^{e,k}_{m,t,\ 2}$ | 一个数据点<br>来自机器 $i$ 的一个数据点<br>在轮次 $e$、迭代 $t$ 时于组 $k$ 的机器 $m$ 上采样的一个数据点<br>在第 $e$ 轮次、第 $t$ 次迭代中，从第 $k$ 组的机器 $m$ 上采样的两个相互独立的数据点（分别为正标签和负标签） |
| $h(\mathbf{w};\mathbf{z}) \in \mathbb{R}^1$<br>$[X]$<br>$G^{e,k}_{m,t,\ 1},\ G^{e,k}_{m,t,\ 2}$ | 网络 $\mathbf{w}$ 对数据 $\mathbf{z}$ 的预测分数<br>整数集合 $\{1, 2, ..., X\}$，其中 $X$ 是一个整数。<br>梯度各分量的局部随机估计量 |
| $H_1^e, H_2^e$<br>$\text{离}_m^{e,k}, \text{离}_m^{e,k}$<br>$\mathcal{R}_{i\ 1}^{e,k}, \mathcal{R}_{i\ 2}^{e,k}$ | 正样本和负样本数据的历史预测分数全局缓冲区。<br>在第 $e$ 轮次，收集组/轮次 $k$ 中机器 $m$ 上的历史预测分数<br>从服务器传输到客户端的缓冲区 |
| $h_\epsilon^{e-1}, h_\zeta^{e-1}$ | 从第 $e-1$ 轮次收集的分数中采样得到的预测分数 |

# B CyCP 极小极大分析

## B.1 辅助引理

对于极小极大问题的分阶段算法，我们将点（$\mathbf{v}$ $\alpha$）处的第 $s$阶段对偶间隙定义为

$$Gap_s(\mathbf{v}, \alpha) = \max_{\alpha'} f^s(\mathbf{v}, \alpha') - \min_{\mathbf{v}'} f^s(\mathbf{v}', \alpha). \tag{15}$$

在给出证明之前，我们首先介绍 Yan 等人 （2020年）中的辅助引理。

**引理 B.1** （Yan 等人 （2020年）的引理 1）。 设函数 $f$（$\mathbf{v}$ $\alpha$） 关于 $\lambda_1$-强凸，关于 $\mathbf{v}$ 且$\lambda_2$-强凹，关于 $\alpha$。考虑以下问题

$$\min_{\mathbf{v}\in X} \max_{\alpha\in Y} f(\mathbf{v}, \alpha),$$

其中 $X$ 和 $Y$ 是凸紧集。记 $\hat{}\mathbf{v}f$（$\alpha$） $=\arg\min_{\mathbf{v}'\in X} f$（$\mathbf{v}'$, $\alpha$） 以及 $\hat{}\alpha_f$（$\mathbf{v}$） $=\arg\max_{\alpha'\in Y} f$（$\mathbf{v}$ $\alpha'$）。设我们有两个解 （$\mathbf{v}_0$， $\alpha_0$） 和 （$\mathbf{v}_1$， $\alpha_1$）。则变量距离与对偶间隙之间满足以下关系

$$\begin{aligned}\frac{\lambda_1}{4}\|\hat{\mathbf{v}}_f(\alpha_1) - \mathbf{v}_0\|^2 + \frac{\lambda_2}{4}\|\hat{\alpha}_f(\mathbf{v}_1) - \alpha_0\|^2 \le &\max_{\alpha'\in Y} f(\mathbf{v}_0, \alpha') - \min_{\mathbf{v}'\in X} f(\mathbf{v}', \alpha_0) \\ &+ \max_{\alpha'\in Y} f(\mathbf{v}_1, \alpha') - \min_{\mathbf{v}'\in X} f(\mathbf{v}', \alpha_1).\end{aligned} \tag{16}$$

□

**引理 B.2 (**Yan 等人 (2020年）**的引理 5）**。我们得到 *Gap* 的以下下界$s(\mathbf{v}_s, \alpha_s)$

$$Gap_s(\mathbf{v}_s, \alpha_s) \ge \frac{3}{50} Gap_{s+1}(\mathbf{v}_0^{s+1}, \alpha_0^{s+1}) + \frac{4}{5}(\phi(\mathbf{v}_0^{s+1}) - \phi(\mathbf{v}_0^s)),$$

其中 $\mathbf{v}_0^{s+1} = \mathbf{v}_s$ 且 $\alpha_0^{s+1} = \alpha_s$，即，第 $s+1$ 阶段的初始化是第 $s$ 阶段的输出 .

## B.2 引理

利用 根据 $f(\cdot, \cdot)$ 的强凸性/凹性和光滑性，我们有如下引理 .

**引理 B.3.** 假设 *4.1* 成立，并以 $(\mathbf{v}^{0,0}, \alpha^{0,0})$ 为输入运行算法 *1* 。对于任意$(\mathbf{v}\ \alpha)$，输出 $(\bar{\mathbf{v}}\ \bar{\alpha})$ 满足

$$
\begin{aligned}
&f^s(\bar{\mathbf{v}}, \alpha) - f^s(\mathbf{v}, \bar{\alpha}) \le \frac{1}{E}\sum_{e=0}^{E-1}[f^s(\mathbf{v}^{e+1,0}, \alpha) - f^s(\mathbf{v}, \alpha^{e+1,0})] \\
&\le \frac{1}{E}\sum_{e=0}^{E-1}\Bigg[\underbrace{\langle \partial_{\mathbf{v}} f^s(\mathbf{v}^{e,0}, \alpha^{e,0}), \mathbf{v}^{e+1,0} - \mathbf{v}\rangle}_{A_1} + \underbrace{\langle \partial_{\alpha} f^s(\mathbf{v}^{e,0}, \alpha^{e,0}), \alpha - \alpha^{e+1,0}\rangle}_{A_2} \\
&\qquad + \underbrace{\frac{3\ell + 3\ell^2/\mu_2}{2}\|\mathbf{v}^{e+1,0} - \mathbf{v}^{e,0}\|^2 + 2\ell(\alpha^{e+1,0} - \alpha^{e,0})^2}_{A_3} - \frac{\ell}{3}\|\mathbf{v} - \mathbf{v}^{e,0}\|^2 - \frac{\mu_2}{3}(\alpha^{e,0} - \alpha)^2\Bigg].
\end{aligned}
$$

证明。 对于任意 $\mathbf{v}$ 和 $\alpha$，利用詹森不等式，以及 $f^s(\mathbf{v}\ \alpha)$ 关于 $\mathbf{v}$ 是凸函数、关于$\alpha$是凹函数这一事实，

$$
f^s(\bar{\mathbf{v}}, \alpha) - f^s(\mathbf{v}, \bar{\alpha}) \le \frac{1}{E}\sum_{e=1}^{E}\left(f^s(\mathbf{v}^{e,0}, \alpha) - f^s(\mathbf{v}, \alpha^{e,0})\right). \tag{17}
$$

由 $f^s(\mathbf{v}\ \alpha)$ 关于 $\mathbf{v}$的 $\ell$-强凸性可得

$$
f^s(\mathbf{v}^{e,0}, \alpha^{e,0}) + \langle \partial_{\mathbf{v}} f^s(\mathbf{v}^{e,0}, \alpha^{e,0}), \mathbf{v} - \mathbf{v}^{e,0}\rangle + \frac{\ell}{2}\|\mathbf{v}^{e,0} - \mathbf{v}\|^2 \le f(\mathbf{v}, \alpha^{e,0}). \tag{18}
$$

由 $f^s(\mathbf{v}\ \alpha)$ 关于 $\mathbf{v}$的 $3\ell$-光滑性可得

$$
\begin{aligned}
&f^s(\mathbf{v}^{e+1,0}, \alpha) \le f^s(\mathbf{v}^{e,0}, \alpha) + \langle \partial_{\mathbf{v}} f^s(\mathbf{v}^{e,0}, \alpha), \mathbf{v}^{e+1,0} - \mathbf{v}^{e,0}\rangle + \frac{3\ell}{2}\|\mathbf{v}^{e+1,0} - \mathbf{v}^{e,0}\|^2 \\
&= f^s(\mathbf{v}^{e,0}, \alpha) + \langle \partial_{\mathbf{v}} f^s(\mathbf{v}^{e,0}, \alpha^{e,0}), \mathbf{v}^{e+1,0} - \mathbf{v}^{e,0}\rangle + \frac{3\ell}{2}\|\mathbf{v}^{e+1,0} - \mathbf{v}^{e,0}\|^2 \\
&\quad + \langle \partial_{\mathbf{v}} f^s(\mathbf{v}^{e,0}, \alpha) - \partial_{\mathbf{v}} f^s(\mathbf{v}^{e,0}, \alpha^{e,0}), \mathbf{v}^{e+1,0} - \mathbf{v}^{e,0}\rangle \\
&\overset{(a)}{\le} f^s(\mathbf{v}^{e,0}, \alpha) + \langle \partial_{\mathbf{v}} f^s(\mathbf{v}^{e,0}, \alpha^{e,0}), \mathbf{v}^{e+1,0} - \mathbf{v}^{e,0}\rangle + \frac{3\ell}{2}\|\mathbf{v}^{e+1,0} - \mathbf{v}^{e,0}\|^2 \\
&\quad + \ell|\alpha^{e,0} - \alpha|\|\mathbf{v}^{e+1,0} - \mathbf{v}^{e,0}\| \\
&\overset{(b)}{\le} f^s(\mathbf{v}^{e,0}, \alpha) + \langle \partial_{\mathbf{v}} f^s(\mathbf{v}^{e,0}, \alpha^{e,0}), \mathbf{v}^{e+1,0} - \mathbf{v}^{e,0}\rangle + \frac{3\ell}{2}\|\mathbf{v}^{e+1,0} - \mathbf{v}^{e,0}\|^2 \\
&\quad + \frac{\mu_2}{6}(\alpha^{e,0} - \alpha)^2 + \frac{3\ell^2}{2\mu_2}\|\bar{\mathbf{v}}^{e+1,0} - \mathbf{v}^{e,0}\|^2,
\end{aligned} \tag{19}
$$

其中，$(a)$成立是因为我们知道 $\partial_{\mathbf{v}} f(\mathbf{v}\ \alpha)$ 关于 $\alpha$是 $\ell$-利普希茨的，这是因为 $f(\mathbf{v}\ \alpha)$ 是 $\ell$-光滑的；$(b)$由杨不等式得出；而 $\mu_2 = 2p(1-p)$ 是 $f^s$ 关于 $\alpha$的强凹系数。
将（18）和（19）相加并整理各项，得

$$
\begin{aligned}
&f^s(\mathbf{v}^{e,0}, \alpha^{e,0}) + f^s(\mathbf{v}^{e+1,0}, \alpha) \\
&\le f(\mathbf{v}, \alpha^{e,0}) + f(\mathbf{v}^{e,0}, \alpha) + \langle \partial_{\mathbf{v}} f(\mathbf{v}^{e,0}, \alpha^{e,0}), \mathbf{v}^{e+1,0} - \mathbf{v}\rangle + \frac{3\ell + 3\ell^2/\mu_2}{2}\|\mathbf{v}^{e+1,0} - \mathbf{v}^{e,0}\|^2 \\
&\quad - \frac{\ell}{2}\|\mathbf{v}^{e,0} - \mathbf{v}\|^2 + \frac{\mu_2}{6}(\alpha^{e,0} - \alpha)^2.
\end{aligned} \tag{20}
$$

我们知道，$-f(\mathbf{v}\ \alpha)$ 在 $\alpha$ 中是 $\mu_2$-强凸的。因此，我们有

$$
-f^s(\mathbf{v}^{e,0}, \alpha^{e,0}) - \partial_{\alpha} f^s(\mathbf{v}^{e,0}, \alpha^{e,0})^{\top}(\alpha - \alpha^{e,0}) + \frac{\mu_2}{2}(\alpha - \alpha^{e,0})^2 \le -f^s(\mathbf{v}^{e,0}, \alpha). \tag{21}
$$

由于 $f(\mathbf{v}\ \alpha)$ 在 $\alpha$ 中是 $\ell$-光滑的，我们得到

$$
\begin{aligned}
&-f^s(\mathbf{v},\alpha^{e+1,0}) \le -f^s(\mathbf{v},\alpha^{e,0}) - \langle \partial_\alpha f^s(\mathbf{v},\alpha^{e,0}), \alpha^{e+1,0}-\alpha^{e,0}\rangle + \frac{\ell}{2}(\alpha^{e+1,0}-\alpha^{e,0})^2 \\
&= -f^s(\mathbf{v},\alpha^{e,0}) - \langle \partial_\alpha f^s(\mathbf{v}^{e,0},\alpha^{e,0}), \alpha^{e+1,0}-\alpha^{e,0}\rangle + \frac{\ell}{2}(\alpha^{e+1,0}-\alpha^{e,0})^2 \\
&\quad - \langle \partial_\alpha f^s(\mathbf{v},\alpha^{e,0}) - \partial_\alpha f^s(\mathbf{v}^{e,0},\alpha^{e,0}), \alpha^{e+1,0}-\alpha^{e,0}\rangle \\
&\overset{(a)}{\le} -f^s(\mathbf{v},\alpha^{e,0}) - \langle \partial_\alpha f^s(\mathbf{v}^{e,0},\alpha^{e,0}), \alpha^{e+1,0}-\alpha^{e,0}\rangle + \frac{\ell}{2}(\alpha^{e+1,0}-\alpha^{e,0})^2 + \ell\|\mathbf{v}-\mathbf{v}^{e,0}\||\alpha^{e+1,0}-\alpha^{e,0}| \\
&\le -f^s(\mathbf{v},\alpha^{e,0}) - \langle \partial_\alpha f^s(\mathbf{v}^{e,0},\alpha^{e,0}), \alpha^{e+1,0}-\alpha^{e,0}\rangle + \frac{\ell}{2}(\alpha^{e+1,0}-\alpha^{e,0})^2 + \frac{\ell}{6}\|\mathbf{v}^{e,0}-\mathbf{v}\|^2 + \frac{3\ell}{2}(\alpha^{e+1,0}-\alpha^{e,0})^2,
\end{aligned}
\tag{22}
$$

其中，(a) 成立是因为 $\partial_\alpha f^s(\mathbf{v}\ \alpha)$ 关于$\mathbf{v}$是 $\ell$-利普希茨的。

将（21）、（22）相加并整理各项，可得

$$
\begin{aligned}
-f^s(\mathbf{v}^{e,0},\alpha^{e,0}) - f^s(\mathbf{v},\alpha^{e+1,0}) &\le -f^s(\mathbf{v}^{e,0},\alpha) - f^s(\mathbf{v},\alpha^{e,0}) - \langle \partial_\alpha f^s(\mathbf{v}^{e,0},\alpha^{e,0}), \alpha^{e+1,0}-\alpha\rangle \\
&\quad + 2\ell(\alpha^{e+1,0}-\alpha^{e,0})^2 + \frac{\ell}{6}\|\mathbf{v}^{e,0}-\mathbf{v}\|^2 - \frac{\mu_2}{2}(\alpha-\alpha^{e,0})^2.
\end{aligned}
\tag{23}
$$

将（20）和（23）相加，得到

$$
\begin{aligned}
f^s(\mathbf{v}^{e+1,0},\alpha) - f^s(\mathbf{v},\alpha^{e+1,0}) &\le \langle \partial_{\mathbf{v}} f(\mathbf{v}^{e,0},\alpha^{e,0}), \mathbf{v}^{e+1,0}-\mathbf{v}\rangle - \langle \partial_\alpha f(\mathbf{v}^{e,0},\alpha^{e,0}), \alpha^{e+1,0}-\alpha\rangle \\
&\quad + \frac{3\ell+3\ell^2/\mu_2}{2}\|\mathbf{v}^{e+1,0}-\mathbf{v}^{e,0}\|^2 + 2\ell(\alpha^{e+1,0}-\alpha^{e-1,0})^2 - \frac{\ell}{3}\|\mathbf{v}^{e,0}-\mathbf{v}\|^2 - \frac{\mu_2}{3}(\alpha^{e,0}-\alpha)^2.
\end{aligned}
\tag{24}
$$

对 $e$、$k$、$t$取平均，得到

$$
\begin{aligned}
f^s(\bar{\mathbf{v}},\alpha) - f^s(\mathbf{v},\bar{\alpha}) &\le \frac{1}{E}\sum_{e=0}^{E-1}[f^s(\mathbf{v}^{e+1,0},\alpha) - f^s(\mathbf{v},\alpha^{e+1,0})] \\
&\le \frac{1}{E}\sum_{e=0}^{E-1}\Bigg[\underbrace{\langle \partial_{\mathbf{v}} f^s(\mathbf{v}^{e,0},\alpha^{e,0}), \mathbf{v}^{e+1,0}-\mathbf{v}\rangle}_{A_1} + \underbrace{\langle \partial_\alpha f^s(\mathbf{v}^{e,0},\alpha^{e,0}), \alpha-\alpha^{e+1,0}\rangle}_{A_2} \\
&\quad + \underbrace{\frac{3\ell+3\ell^2/\mu_2}{2}\|\mathbf{v}^{e+1,0}-\mathbf{v}^{e,0}\|^2 + 2\ell(\alpha^{e+1,0}-\alpha^{e,0})^2}_{A_3} - \frac{\ell}{3}\|\mathbf{v}-\mathbf{v}^{e,0}\|^2 - \frac{\mu_2}{3}(\alpha^{e,0}-\alpha)^2\Bigg].
\end{aligned}
$$

□

接下来，我们将根据引理 $A_1$ 对该项进行界定，并根据引理 4.3， $A_2$ 对该项进行界定 B.4。

**引理的证明4.3** 。

证明。

$$
\begin{aligned}
\langle \nabla_{\mathbf{v}} f^s(\mathbf{v}^{e,0},\alpha^{e,0}), \mathbf{v}^{e+1,0}-\mathbf{v}\rangle &= \left\langle \frac{1}{K}\sum_{k=1}^{K}\frac{1}{M}\sum_{m\in\mathcal{G}^{e,k}} \nabla_{\mathbf{v}} f^s_{k,m}(\mathbf{v}^{e,0},\alpha^{e,0}), \mathbf{v}^{e+1,0}-\mathbf{v}\right\rangle \\
&= \left\langle \frac{1}{K}\sum_{k=1}^{K}\frac{1}{M}\sum_{m\in\mathcal{G}^{e,k}} [\frac{1}{I}\sum_t \nabla_{\mathbf{v}} f^s_{k,m}(\mathbf{v}^{e,0},\alpha^{e,0})] - \frac{1}{K}\sum_{k=1}^{K}\frac{1}{M}\sum_{m\in\mathcal{G}^{e,k}} [\frac{1}{I}\sum_t \nabla_{\mathbf{v}} F^s_{k,m}(\mathbf{v}^{e,k}_{m,t},\alpha^{e,k}_{m,t};\mathbf{z}^{e,k}_{m,t})], \mathbf{v}^{e+1,0}-\mathbf{v}\right\rangle \quad ③ \\
&\quad + \left\langle \frac{1}{K}\sum_{k=1}^{K}\frac{1}{M}\sum_{m\in\mathcal{G}^{e,k}} [\frac{1}{I}\sum_t \nabla_{\mathbf{v}} F^s_{k,m}(\mathbf{v}^{e,k}_{m,t},\alpha^{e,k}_{m,t};\mathbf{z}^{e,k}_{m,t})], \mathbf{v}^{e+1,0}-\mathbf{v}\right\rangle \qquad ④.
\end{aligned}
\tag{25}
$$

$K$利用 $\hat{\mathbf{v}}^{e+1,0} = \mathbf{v}^{e,0} - \eta I \sum_{k=1}^{K} \frac{1}{M} \sum_{m \in \mathcal{G}^{e,k}} \nabla_{\mathbf{v}} f^s(\mathbf{v}^{e,0}, \alpha^{e,0})$，则有

$$\mathbf{v}^{e+1,0} - \hat{\mathbf{v}}^{e+1,0} = \eta \Bigg( \sum_{k=1}^{K} \frac{1}{M} \sum_{m \in \mathcal{G}^{e,k}} \sum_{t=1}^{I} \nabla_{\mathbf{v}} f^s_{k,m}(\mathbf{v}^{e,0}, \alpha^{e,0}) - \sum_{k=1}^{K} \frac{1}{M} \sum_{m \in \mathcal{G}^{e,k}} \sum_{t=1}^{I} \nabla_{\mathbf{v}} f^s_{k,m}(\mathbf{v}^{e,k}_{m,t}, \alpha^{e,k}_{m,t}; \mathbf{z}^{e,k}_{m,t}) \Bigg). \tag{26}$$

因此，我们得到

$$\begin{aligned}
③ &= \left\langle \frac{1}{K} \sum_{k=1}^{K} \frac{1}{M} \sum_{m \in \mathcal{G}^{e,k}} [\frac{1}{I} \sum_{t} \nabla_{\mathbf{v}} f^s_{k,m}(\mathbf{v}^{e,0}, \alpha^{e,0})] - \frac{1}{K} \sum_{k=1}^{K} \frac{1}{M} \sum_{m \in \mathcal{G}^{e,k}} [\frac{1}{I} \sum_{t} \nabla_{\mathbf{v}} F^s_{k,m}(\mathbf{v}^{e,k}_{m,t}, \alpha^{e,k}_{m,t}; \mathbf{z}^{e,k}_{m,t})], \mathbf{v}^{e+1,0} - \hat{\mathbf{v}}^{e+1,0} \right\rangle \\
&+ \left\langle \frac{1}{K} \sum_{k=1}^{K} \frac{1}{M} \sum_{m \in \mathcal{G}^{e,k}} [\frac{1}{I} \sum_{t} \nabla_{\mathbf{v}} f^s_{k,m}(\mathbf{v}^{e,0}, \alpha^{e,0})] - \frac{1}{K} \sum_{k=1}^{K} \frac{1}{M} \sum_{m \in \mathcal{G}^{e,k}} [\frac{1}{I} \sum_{t} \nabla_{\mathbf{v}} F^s_{k,m}(\mathbf{v}^{e,k}_{m,t}, \alpha^{e,k}_{m,t}; \mathbf{z}^{e,k}_{m,t})], \hat{\mathbf{v}}^{e+1,0} - \mathbf{v} \right\rangle \\
&= \eta K I \left\| \frac{1}{K} \sum_{k=1}^{K} \frac{1}{M} \sum_{m \in \mathcal{G}^{e,k}} \frac{1}{I} \sum_{t} [\nabla_{\mathbf{v}} f^s_{k,m}(\mathbf{v}^{e,0}, \alpha^{e,0}) - \nabla_{\mathbf{v}} F^s_{k,m}(\mathbf{v}^{e,k}_{m,t}, \alpha^{e,k}_{m,t}; z^{e,k}_{m,t})] \right\|^2 \\
&+ \left\langle \frac{1}{K} \sum_{k=1}^{K} \frac{1}{M} \sum_{m \in \mathcal{G}^{e,k}} [\frac{1}{I} \sum_{t} \nabla_{\mathbf{v}} f^s_{k,m}(\mathbf{v}^{e,0}, \alpha^{e,0})] - \frac{1}{K} \sum_{k=1}^{K} \frac{1}{M} \sum_{m \in \mathcal{G}^{e,k}} [\frac{1}{I} \sum_{t} \nabla_{\mathbf{v}} F^s_{k,m}(\mathbf{v}^{e,0}, \alpha^{e,0}; \mathbf{z}^{e,k}_{m,t})], \hat{\mathbf{v}}^{e+1,0} - \mathbf{v} \right\rangle \\
&+ \left\langle \frac{1}{K} \sum_{k=1}^{K} \frac{1}{M} \sum_{m \in \mathcal{G}^{e,k}} [\frac{1}{I} \sum_{t} \nabla_{\mathbf{v}} F^s_{k,m}(\mathbf{v}^{e,0}, \alpha^{e,0}; \mathbf{z}^{e,k}_{m,t})] - \frac{1}{K} \sum_{k=1}^{K} \frac{1}{M} \sum_{m \in \mathcal{G}^{e,k}} [\frac{1}{I} \sum_{t} \nabla_{\mathbf{v}} F^s_{k,m}(\mathbf{v}^{e,k}_{m,t}, \alpha^{e,k}_{m,t}; \mathbf{z}^{e,k}_{m,t})], \hat{\mathbf{v}}^{e+1,0} - \mathbf{v} \right\rangle,
\end{aligned} \tag{27}$$

其中

$$\begin{aligned}
&\left\langle \frac{1}{K} \sum_{k=1}^{K} \frac{1}{M} \sum_{m \in \mathcal{G}^{e,k}} [\frac{1}{I} \sum_{t} \nabla_{\mathbf{v}} F^s_{k,m}(\mathbf{v}^{e,0}, \alpha^{e,0}; \mathbf{z}^{e,k}_{m,t})] - \frac{1}{K} \sum_{k=1}^{K} \frac{1}{M} \sum_{m \in \mathcal{G}^{e,k}} [\frac{1}{I} \sum_{t} \nabla_{\mathbf{v}} F^s_{k,m}(\mathbf{v}^{e,k}_{m,t}, \alpha^{e,k}_{m,t}; \mathbf{z}^{e,k}_{m,t})], \hat{\mathbf{v}}^{e+1,0} - \mathbf{v} \right\rangle \\
&\leq 3\ell \frac{1}{K} \sum_{k=1}^{K} \frac{1}{M} \sum_{m \in \mathcal{G}^{e,k}} \frac{1}{I} \sum_{t} \|\mathbf{v}^{e,0} - \mathbf{v}^{e,k}_{m,t}\|^2 + \frac{\ell}{6} \|\hat{\mathbf{v}}^{e+1,0} - \mathbf{v}\|^2.
\end{aligned} \tag{28}$$

定义另一个辅助序列如下

$$\tilde{\mathbf{v}}^{e+1,0} = \tilde{\mathbf{v}}^{e,0} - \frac{\eta}{M} \sum_{k=1}^{K} \sum_{m \in \mathcal{G}^{e,k}} \sum_{t=1}^{I} \left( \nabla_{\mathbf{v}} F^s_{k,m}(\mathbf{v}^{e,0}, \alpha^{e,0}; z^{e,k}_{m,t}) - \nabla_{\mathbf{v}} f^s_{k,m}(\mathbf{v}^{e,0}, \alpha^{e,0}) \right), \text{ for } t > 0; \tilde{\mathbf{v}}_0 = \mathbf{v}_0. \tag{29}$$

记

$$\Theta_e(\mathbf{v}) = \left( \frac{1}{M} \sum_{k=1}^{K} \sum_{m \in \mathcal{G}^{e,k}} \sum_{t=1}^{I} \left( \nabla_{\mathbf{v}} f^s_{k,m}(\mathbf{v}^{e,0}, \alpha^{e,0}) - \nabla_{\mathbf{v}} F^s_{k,m}(\mathbf{v}^{e,0}, \alpha^{e,0}; z^{e,k}_{m,t}) \right) \right)^{\top} \mathbf{v} + \frac{1}{2\eta} \|\mathbf{v} - \tilde{\mathbf{v}}^{e,0}\|^2. \tag{30}$$

因此，对于辅助序列 $\tilde{\alpha}_t$，我们可以验证

$$\tilde{\mathbf{v}}^{e+1,0} = \arg\min_{\mathbf{v}} \Theta_e(\mathbf{v}). \tag{31}$$

由于 $\Theta_e(\mathbf{v})$ 是 $\frac{1}{\eta}$-强凸的，我们有

$$
\begin{aligned}
&\frac{1}{2\eta}\|\mathbf{v}-\tilde{\mathbf{v}}^{e+1,0}\|^2 \le \Theta_e(\mathbf{v})-\Theta_e(\tilde{\mathbf{v}}^{e+1,0}) \\
&=\left(\frac{1}{M}\sum_{k=1}^{K}\sum_{m\in\mathcal{G}^{e,k}}\sum_{t=1}^{I}\left(\nabla_{\mathbf{v}}f^s_{k,m}(\mathbf{v}^{e,0},\alpha^{e,0})-\nabla_{\mathbf{v}}F^s_{k,m}(\mathbf{v}^{e,0},\alpha^{e,0};z^{e,k}_{m,t})\right)\right)^{\top}\mathbf{v}+\frac{1}{2\eta}\|\mathbf{v}-\tilde{\mathbf{v}}^{e,0}\|^2 \\
&\quad-\left(\frac{1}{M}\sum_{k=1}^{K}\sum_{m\in\mathcal{G}^{e,k}}\sum_{t=1}^{I}\left(\nabla_{\mathbf{v}}f^s_{k,m}(\mathbf{v}^{e,0},\alpha^{e,0})-\nabla_{\mathbf{v}}F^s_{k,m}(\mathbf{v}^{e,0},\alpha^{e,0};z^{e,k}_{m,t})\right)\right)^{\top}\tilde{\mathbf{v}}^{e+1,0}-\frac{1}{2\eta}\|\tilde{\mathbf{v}}^{e+1,0}-\tilde{\mathbf{v}}^{e,0}\|^2 \\
&=\left(\frac{1}{M}\sum_{k=1}^{K}\sum_{m\in\mathcal{G}^{e,k}}\sum_{t=1}^{I}\left(\nabla_{\mathbf{v}}f^s_{k,m}(\mathbf{v}^{e,0},\alpha^{e,0})-\nabla_{\mathbf{v}}F^s_{k,m}(\mathbf{v}^{e,0},\alpha^{e,0};z^{e,k}_{m,t})\right)\right)^{\top}(\mathbf{v}-\tilde{\mathbf{v}}^{e,0})+\frac{1}{2\eta}\|\mathbf{v}-\tilde{\mathbf{v}}^{e,0}\|^2 \\
&\quad-\left(\frac{1}{M}\sum_{k=1}^{K}\sum_{m\in\mathcal{G}^{e,k}}\sum_{t=1}^{I}\left(\nabla_{\mathbf{v}}f^s_{k,m}(\mathbf{v}^{e,0},\alpha^{e,0})-\nabla_{\mathbf{v}}F^s_{k,m}(\mathbf{v}^{e,0},\alpha^{e,0};z^{e,k}_{m,t})\right)\right)^{\top}(\tilde{\mathbf{v}}^{e+1,0}-\tilde{\mathbf{v}}^{e,0})-\frac{1}{2\eta}\|\tilde{\mathbf{v}}^{e+1,0}-\tilde{\mathbf{v}}^{e,0}\|^2 \\
&\le\left(\frac{1}{M}\sum_{k=1}^{K}\sum_{m\in\mathcal{G}^{e,k}}\sum_{t=1}^{I}\left(\nabla_{\mathbf{v}}f^s_{k,m}(\mathbf{v}^{e,0},\alpha^{e,0})-\nabla_{\mathbf{v}}F^s_{k,m}(\mathbf{v}^{e,0},\alpha^{e,0};z^{e,k}_{m,t})\right)\right)^{\top}(\mathbf{v}-\tilde{\mathbf{v}}^{e,0})+\frac{1}{2\eta}\|\mathbf{v}-\tilde{\mathbf{v}}^{e,0}\|^2 \\
&\quad+\frac{\eta}{2}\left\|\frac{1}{M}\sum_{k=1}^{K}\sum_{m\in\mathcal{G}^{e,k}}\sum_{t=1}^{I}\left(\nabla_{\mathbf{v}}f^s_{k,m}(\mathbf{v}^{e,0},\alpha^{e,0})-\nabla_{\mathbf{v}}F^s_{k,m}(\mathbf{v}^{e,0},\alpha^{e,0};z^{e,k}_{m,t})\right)\right\|^2.
\end{aligned}
\tag{32}
$$

两边相乘 $1/KI$ 并将其与式（27）相加，可得

$$
\begin{aligned}
③\le&\eta KI\left\|\frac{1}{MKI}\sum_{k=1}^{K}\sum_{m\in\mathcal{G}^{e,k}}\sum_{t=1}^{I}\left(\nabla_{\mathbf{v}}f^s_{k}(\mathbf{v}^{e,0},\alpha^{e,0})-\nabla_{\mathbf{v}}F^s_{k}(\mathbf{v}^{e,k}_{m,t},\alpha^{e,k}_{m,t};z^{e,k}_{m,t})\right)\right\|^2 \\
&+\frac{\eta KI}{2}\left\|\frac{1}{MKI}\sum_{k=1}^{K}\sum_{m\in\mathcal{G}^{e,k}}\sum_{t=1}^{I}\left(\nabla_{\mathbf{v}}f^s_{k}(\mathbf{v}^{e,0},\alpha^{e,0})-\nabla_{\mathbf{v}}F^s_{k}(\mathbf{v}^{e,0},\alpha^{e,0};z^{e,k}_{m,t})\right)\right\|^2 \\
&+\frac{1}{2\eta KI}\|\mathbf{v}-\tilde{\mathbf{v}}^{e,0}\|^2-\frac{1}{2\eta KI}\|\mathbf{v}-\tilde{\mathbf{v}}^{e+1,0}\|^2 \\
&+\left\langle\frac{1}{MKI}\sum_{k=1}^{K}\sum_{m\in\mathcal{G}^{e,k}}\sum_{t=1}^{I}\left(\nabla_{\mathbf{v}}f^s_{k}(\mathbf{v}^{e,0},\alpha^{e,0})-\nabla_{\mathbf{v}}F^s_{k}(\mathbf{v}^{e,0},\alpha^{e,0};z^{e,k}_{m,t})\right),\hat{\mathbf{v}}^{e+1,0}-\tilde{\mathbf{v}}^{e,0}\right\rangle.
\end{aligned}
\tag{33}
$$

其中

$$
\mathbb{E}\left\langle\frac{1}{MKI}\sum_{k=1}^{K}\sum_{m\in\mathcal{G}^{e,k}}\sum_{t=1}^{I}\left(\nabla_{\mathbf{v}}f^s_{k}(\mathbf{v}^{e,0},\alpha^{e,0})-\nabla_{\mathbf{v}}F^s_{k}(\mathbf{v}^{e,0},\alpha^{e,0};z^{e,k}_{m,t})\right),\hat{\mathbf{v}}^{e+1,0}-\tilde{\mathbf{v}}^{e,0}\right\rangle=0. \tag{34}
$$

4◯可界定为

$$
④=-\frac{1}{\eta KI}\langle\mathbf{v}^{e+1,0}-\mathbf{v}^{e,0},\mathbf{v}^{e+1,0}-\mathbf{v}\rangle=\frac{1}{2\eta KI}(\|\mathbf{v}^{e,0}-\mathbf{v}\|^2-\|\mathbf{v}^{e,0}-\mathbf{v}^{e+1,0}\|^2-\|\mathbf{v}^{e+1,0}-\mathbf{v}\|^2). \tag{35}
$$

将（33）和（35）代入（25），并注意到 $\eta KI \le O(1)$，可得

$$
\begin{aligned}
&\mathbb{E}_{e,0}\left\langle \nabla_{\mathbf{v}} f(\mathbf{v}^{e,0},\alpha^{e,0}), \mathbf{v}^{e+1,0}-\mathbf{v}\right\rangle \\
&\le 3\eta KI\mathbb{E}\left\| \frac{1}{MKI}\sum_{k=1}^{K}\sum_{m\in\mathcal{G}^{e,k}}\sum_{t=1}^{I}\left(\nabla_{\mathbf{v}} F_k^s(\mathbf{v}^{e,0},\alpha^{e,0};z_{m,t}^{e,k})-\nabla_{\mathbf{v}} f_k^s(\mathbf{v}^{e,0},\alpha^{e,0})\right)\right\|^2 \\
&\quad+\frac{5\ell}{KMI}\sum_{k}\sum_{m\in\mathcal{G}^{e,k}}\sum_{t}\mathbb{E}(\|\mathbf{v}^{e,0}-\mathbf{v}_{m,t}^{e,k}\|^2+\|\alpha^{e,0}-\alpha_{m,t}^{e,k}\|^2) \\
&\quad+\frac{1}{2\eta KI}\mathbb{E}(\|\mathbf{v}^{e,0}-\mathbf{v}\|^2-\|\mathbf{v}^{e,0}-\mathbf{v}^{e+1,0}\|^2-\|\mathbf{v}^{e+1,0}-\mathbf{v}\|^2) \\
&\quad+\frac{1}{2\eta KI}(\|\mathbf{v}-\tilde{\mathbf{v}}^{e,0}\|^2-\|\mathbf{v}-\tilde{\mathbf{v}}^{e+1,0}\|^2)+\frac{\ell}{3}\mathbb{E}\|\hat{\mathbf{v}}^{e+1,0}-\mathbf{v}\|^2 \\
&\le \frac{6\tilde{\eta}K\sigma^2}{MKI}+\frac{6\ell}{KMI}\sum_{k}\sum_{m\in\mathcal{G}^{e,k}}\sum_{t}\mathbb{E}(\|\mathbf{v}^{e,0}-\mathbf{v}_{m,t}^{e,k}\|^2+\|\alpha^{e,0}-\alpha_{m,t}^{e,k}\|^2)+\frac{\ell}{3}\|\mathbf{v}^{e+1,0}-\mathbf{v}\|^2 \\
&\quad+\frac{1}{2\eta KI}\mathbb{E}(\|\mathbf{v}^{e,0}-\mathbf{v}\|^2-\|\mathbf{v}^{e,0}-\mathbf{v}^{e+1,0}\|^2-\|\mathbf{v}^{e+1,0}-\mathbf{v}\|^2)+\frac{1}{2\eta KI}(\|\mathbf{v}-\tilde{\mathbf{v}}^{e,0}\|^2-\|\mathbf{v}-\tilde{\mathbf{v}}^{e+1,0}\|^2),
\end{aligned}
$$

其中，最后一个不等式用到了

$$
\|\hat{\mathbf{v}}^{e+1,0}-\mathbf{v}\|^2 \le 2\|\hat{\mathbf{v}}^{e+1,0}-\mathbf{v}^{e+1,0}\|^2+2\|\mathbf{v}^{e+1,0}-\mathbf{v}\|^2, \tag{36}
$$

并且 $\|\hat{\mathbf{v}}^{e+1,0}-\mathbf{v}^{e+1,0}\|^2$ 与之前的 3◯ 类似地处理。 □

类似地， $A_2$ 可以界定为 $\frac{1}{K}$ $I$**引理 B.4**。定义 $\hat{\alpha}^{e+1,0}=\alpha^{e,0}+\eta\sum_{k=1}^{K}\frac{1}{M}\sum_{m\in\mathcal{G}^{e,k}}\sum_{t=1}^{I}\nabla_{\alpha} f_k^s(\mathbf{v}^{e,0},\alpha^{e,0})$

$$
\tilde{\alpha}^{e+1,0}=\tilde{\alpha}^{e,0}+\frac{\eta}{M}\sum_{k=1}^{K}\sum_{m\in\mathcal{G}^{e,k}}\sum_{t=1}^{I}\left(\nabla_{\alpha} F_k^s(\mathbf{v}^{e,0},\alpha^{e,0};z_{m,t}^{e,k})+\nabla_{\alpha} f_k^s(\mathbf{v}^{e,0},\alpha^{e,0})\right),\ \textit{for } t>0;\ \tilde{\alpha}_0=\alpha_0. \tag{37}
$$

$$
\begin{aligned}
&\mathbb{E}_{e,0}\left\langle \nabla_{\alpha} f(\mathbf{v}^{e,0},\alpha^{e,0}), \alpha^{e+1,0}-\alpha\right\rangle \\
&\le \frac{6\tilde{\eta}K\sigma^2}{MKI}+\frac{6\ell}{KMI}\sum_{k}\sum_{m\in\mathcal{G}^{e,k}}\sum_{t}\mathbb{E}(\|\mathbf{v}^{e,0}-\mathbf{v}_{m,t}^{e,k}\|^2+\|\alpha^{e,0}-\alpha_{m,t}^{e,k}\|^2)+\frac{\ell}{3}\mathbb{E}\|\alpha^{e+1,0}-\alpha\|^2 \\
&+\frac{1}{2\eta KI}\mathbb{E}(\|\alpha^{e,0}-\alpha\|^2-\|\alpha^{e,0}-\alpha^{e+1,0}\|^2-\|\alpha^{e+1,0}-\alpha\|^2)+\frac{1}{2\eta KI}\mathbb{E}(\|\alpha-\tilde{\alpha}^{e,0}\|^2-\|\alpha-\tilde{\alpha}^{e+1,0}\|^2).
\end{aligned}
$$

□

有了上述引理，我们可以给出单阶段中的收敛性。

### B.3 CyCp-Minimax 中单个阶段的收敛分析

我们有 以下引理用于界定每个第 $s$ 阶段中子问题的收敛性。

**引理 B.5.** （算法 *1* 的一次调用）设 $(\bar{\mathbf{v}}\ \bar{\alpha})$ 为算法 *1*的输出。假设假设条件 *4.1*成立。给定输入 $\mathbf{v}_0$，运行算法 *1* $\alpha_0$ 次迭代，其中 $T$、 $\gamma=2\ell$，且 $\eta\le\min(\frac{1}{3\ell+3\ell^2/\mu_2},\frac{1}{4\ell})$，则对任意 $\mathbf{v}$ 以及 $\alpha$

$$
\mathbb{E}[f^s(\bar{\mathbf{v}},\alpha)-f^s(\mathbf{v},\bar{\alpha})]\le\frac{1}{\eta T}\|\mathbf{v}_0-\mathbf{v}\|^2+\frac{1}{\eta T}(\alpha_0-\alpha)^2+\underbrace{\left(\frac{3\ell^2}{2\mu_2}+\frac{3\ell}{2}\right)36\eta^2I^2D^2}_{A_1}+\frac{3\eta\sigma^2}{K},
$$

其中 $\mu_2 = 2p(1-p)$ 是 $\alpha$中 $f(\mathbf{v}\ \alpha)$ 的强凹系数。

证明。 根据w的更新，我们得到

$$\begin{aligned}
&\mathbb{E}\|\mathbf{v}^{e+1,0} - \mathbf{v}^{e,0}\|^2 = \tilde{\eta}^2\mathbb{E}\left\|\frac{1}{MI}\sum_{k=1}^{K}\sum_{m\in\mathcal{G}^{e,k}}\sum_{t}(\nabla_{\mathbf{v}} f^s_{k,m}(\mathbf{v}^{e,k}_{m,t}, \alpha^{e,k}_{m,t}; \mathbf{z}^{e,k}_{m,t}))\right\|^2 \le \tilde{\eta}^2K^2G^2,\\
&\mathbb{E}\|\alpha^{e+1,0} - \alpha^{e,0}\|^2 \le \tilde{\eta}^2K^2G^2,\\
&\mathbb{E}\|\mathbf{v}^{e,k}_{m,t} - \mathbf{v}^{e,0}\|^2 \le \tilde{\eta}^2K^2G^2,\\
&\mathbb{E}\|\alpha^{e,k}_{m,t} - \alpha^{e,0}\|^2 \le \tilde{\eta}^2K^2G^2.
\end{aligned} \tag{38}$$

将引理 4.3 和引理 B.4 代入引理 B.3，并取期望，可得

$$\begin{aligned}
&\mathbb{E}[f^s(\bar{\mathbf{v}}, \alpha) - f^s(\mathbf{v}, \bar{\alpha})]\\
&\le \frac{1}{E}\sum_{e=1}^{E}\mathbb{E}\Bigg[\underbrace{\left(\frac{3\ell + 3\ell^2/\mu_2}{2} - \frac{1}{2\eta KI}\right)\|\mathbf{v}^{e,0} - \mathbf{v}^{e+1,0}\|^2 + \left(2\ell - \frac{1}{2\eta KI}\right)\|\alpha^{e+1,0} - \alpha^{e,0}\|^2}_{C_1}\\
&+ \underbrace{\left(\frac{1}{2\eta KI} - \frac{\mu_2}{3}\right)\|\alpha^{e,0} - \alpha\|^2 - \left(\frac{1}{2\eta KI} - \frac{\mu_2}{3}\right)(\alpha^{e+1,0} - \alpha)^2}_{C_2}\\
&+ \underbrace{\left(\frac{1}{2\eta KI} - \frac{\ell}{3}\right)\|\mathbf{v}^{e,0} - \mathbf{v}\|^2 - \left(\frac{1}{2\eta KI} - \frac{\ell}{3}\right)\|\mathbf{v}^{e+1,0} - \mathbf{v}\|^2}_{C_3}\\
&+ \underbrace{\frac{1}{2\eta KI}((\alpha - \tilde{\alpha}^{e,0})^2 - (\alpha - \tilde{\alpha}^{e+1,0})^2)}_{C_4} + \underbrace{\frac{1}{2\eta KI}(\|\mathbf{v} - \tilde{\mathbf{v}}^{e,0}\|^2 - \|\mathbf{v} - \tilde{\mathbf{v}}^{e+1,0}\|^2)}_{C_5}\\
&+ \underbrace{\left(\frac{3\ell^2}{2\mu_2} + \frac{3\ell}{2}\right)\frac{1}{K}\sum_{k=1}^{K}\frac{1}{M}\sum_{m\in\mathcal{G}^{e,k}}\frac{1}{I}\sum_{t=1}^{I}\|\mathbf{v}^{e,0} - \mathbf{v}^{e,k}_{m,t}\|^2 + \left(\frac{3\ell}{2} + \frac{3\ell^2}{2\mu_2}\right)\frac{1}{K}\sum_{k=1}^{K}\frac{1}{M}\sum_{m\in\mathcal{G}^{e,k}}\frac{1}{I}\sum_{t=1}^{I}(\alpha^{e,0} - \alpha^{e,k}_{m,t})^2}_{C_6}\\
&+ \frac{3\tilde{\eta}K\sigma^2}{MKI}
\end{aligned} \tag{39}$$

由于 $\eta \le \min(\frac{1}{3\ell+3\ell^2/\mu_2}, \frac{1}{4\ell})$，因此在式（39）的右侧， $C_1$ 可以消去。 $C_2$、 $C_3$、 $C_4$ 和 $C_5$ 将通过望远镜求和处理。 $C_6$ 可由式（38）界定。

进行望远镜求和可得

$$\begin{aligned}
&\mathbb{E}[f^s(\bar{\mathbf{v}}, \alpha) - f^s(\mathbf{v}, \bar{\alpha})\\
&\le \frac{1}{4\eta EKI}\|\mathbf{v}_0 - \mathbf{v}\|^2 + \frac{1}{4\eta EKI}\|\alpha_0 - \alpha\|^2 + \left(\frac{3\ell^2}{2\mu_2} + \frac{3\ell}{2}\right)36\eta^2I^2K^2G^2 + \frac{3\eta\sigma^2}{M}.
\end{aligned}$$

□

### B.4 定理4.5的主要证明4.5

证明。由于 $f(\mathbf{v}\ \alpha)$ 是 $\ell$-光滑的（因此关于 $\mathbf{v}$ 是 $\ell$-弱凸的），对于任意 $\alpha$，$\phi(\mathbf{v}) = \max_{\alpha'} f(\mathbf{v}\ \alpha')$ 也是 $\ell$-弱凸的。取 $\gamma = 2\ell$，我们有

$$\begin{aligned}
\phi(\mathbf{v}_{s-1}) &\geq \phi(\mathbf{v}_s) + \langle \partial\phi(\mathbf{v}_s), \mathbf{v}_{s-1} - \mathbf{v}_s\rangle - \frac{\ell}{2}\|\mathbf{v}_{s-1} - \mathbf{v}_s\|^2 \\
&= \phi(\mathbf{v}_s) + \langle \partial\phi(\mathbf{v}_s) + 2\ell(\mathbf{v}_s - \mathbf{v}_{s-1}), \mathbf{v}_{s-1} - \mathbf{v}_s\rangle + \frac{3\ell}{2}\|\mathbf{v}_{s-1} - \mathbf{v}_s\|^2 \\
&\overset{(a)}{=} \phi(\mathbf{v}_s) + \langle \partial\phi_s(\mathbf{v}_s), \mathbf{v}_{s-1} - \mathbf{v}_s\rangle + \frac{3\ell}{2}\|\mathbf{v}_{s-1} - \mathbf{v}_s\|^2 \\
&\overset{(b)}{=} \phi(\mathbf{v}_s) - \frac{1}{2\ell}\langle \partial\phi_s(\mathbf{v}_s), \partial\phi_s(\mathbf{v}_s) - \partial\phi(\mathbf{v}_s)\rangle + \frac{3}{8\ell}\|\partial\phi_s(\mathbf{v}_s) - \partial\phi(\mathbf{v}_s)\|^2 \\
&= \phi(\mathbf{v}_s) - \frac{1}{8\ell}\|\partial\phi_s(\mathbf{v}_s)\|^2 - \frac{1}{4\ell}\langle \partial\phi_s(\mathbf{v}_s), \partial\phi(\mathbf{v}_s)\rangle + \frac{3}{8\ell}\|\partial\phi(\mathbf{v}_s)\|^2,
\end{aligned} \tag{40}$$

其中，$(a)$ 和 $(b)$ 根据 $\phi_s(\mathbf{v})$ 的定义成立。

重新整理式（40）中的各项可得

$$\begin{aligned}
\phi(\mathbf{v}_s) - \phi(\mathbf{v}_{s-1}) &\leq \frac{1}{8\ell}\|\partial\phi_s(\mathbf{v}_s)\|^2 + \frac{1}{4\ell}\langle \partial\phi_s(\mathbf{v}_s), \partial\phi(\mathbf{v}_s)\rangle - \frac{3}{8\ell}\|\partial\phi(\mathbf{v}_s)\|^2 \\
&\overset{(a)}{\leq} \frac{1}{8\ell}\|\partial\phi_s(\mathbf{v}_s)\|^2 + \frac{1}{8\ell}(\|\partial\phi_s(\mathbf{v}_s)\|^2 + \|\partial\phi(\mathbf{v}_s)\|^2) - \frac{3}{8\ell}\|\phi(\mathbf{v}_s)\|^2 \\
&= \frac{1}{4\ell}\|\partial\phi_s(\mathbf{v}_s)\|^2 - \frac{1}{4\ell}\|\partial\phi(\mathbf{v}_s)\|^2 \\
&\overset{(b)}{\leq} \frac{1}{4\ell}\|\partial\phi_s(\mathbf{v}_s)\|^2 - \frac{\mu}{2\ell}(\phi(\mathbf{v}_s) - \phi(\mathbf{v}_*))
\end{aligned} \tag{41}$$

其中，$(a)$ 根据 $\langle \mathbf{a}, \mathbf{b}\rangle \leq \frac{1}{2}(\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2)$ 得出，而 $(b)$ 根据 $\phi(\mathbf{v})$ 的 $\mu$-PL 性质成立。因此，我们有

$$(4\ell + 2\mu)\,(\phi(\mathbf{v}_s) - \phi(\mathbf{v}_*)) - 4\ell(\phi(\mathbf{v}_{s-1}) - \phi(\mathbf{v}_*)) \leq \|\partial\phi_s(\mathbf{v}_s)\|^2. \tag{42}$$

由于 $\gamma = 2\ell$，$f^s(\mathbf{v}\ \alpha)$ 在 $\ell$-强凸于 $\mathbf{v}$ 且 $\mu_2 = 2p(1-p)$ 关于 $\alpha$ 强凹。将引理B.1应用于 $f^s$，我们可知

$$\frac{\ell}{4}\|\hat{\mathbf{v}}_s(\alpha_s) - \mathbf{v}_0^s\|^2 + \frac{\mu_2}{4}\|\hat{\alpha}_s(\mathbf{v}_s) - \alpha_0^s\|^2 \leq \mathrm{Gap}_s(\mathbf{v}_0^s, \alpha_0^s) + \mathrm{Gap}_s(\mathbf{v}_s, \alpha_s). \tag{43}$$

根据 $\eta_s = \eta_0 \exp\left(-(s-1)\frac{2\mu}{c+2\mu}\right)$ 和 $T_s = E_s K I_s = \eta_0^{212}\ell \min\{\ ,\mu_2\} \exp\left((s-1)\frac{2\mu}{c+2\mu}\right)$ 的设定，我们注意到 $\frac{1}{\eta_s T_s} \leq \min\{\ell,\mu_2\}$ 212。令 $I_s$ 满足 $\left(\frac{3\ell^2}{2\mu_2} + \frac{3\ell}{2}\right) 36\eta_s^2 I_s^2 K^2 G^2 \leq \frac{\eta_s\sigma^2}{M}$，其中 $I_s$ 的具体取值将在后文给出。应用引理 B.5，并取 $\hat{}\mathbf{v}_s(\alpha_s) = \arg\min_{\mathbf{v}'} f^s(\mathbf{v}', \alpha_s)$ 和 $\hat{}\alpha_s(\mathbf{v}_s) = \arg\max_{\alpha'} f^s(\mathbf{v}_s, \alpha')$，可得

$$\begin{aligned}
\mathbb{E}[\mathrm{Gap}_s(\mathbf{v}_s, \alpha_s)] &\leq \frac{4\eta_s\sigma^2}{M} + \frac{1}{53}\mathbb{E}\left[\frac{\ell}{4}\|\hat{\mathbf{v}}_s(\alpha_s) - \mathbf{v}_0^s\|^2 + \frac{\mu_2}{4}\|\hat{\alpha}_s(\mathbf{v}_s) - \alpha_0^s\|^2\right] \\
&\leq \frac{4\eta_s\sigma^2}{M} + \frac{1}{53}\mathbb{E}\left[\mathrm{Gap}_s(\mathbf{v}_0^s, \alpha_0^s) + \mathrm{Gap}_s(\mathbf{v}_s, \alpha_s)\right].
\end{aligned} \tag{44}$$

由于 $\phi(\mathbf{v})$ 是 $L$-光滑的，且 $\gamma = 2\ell$，因此 $\phi_s(\mathbf{v})$ 是 $\hat{}L = (L + 2\ell)$-光滑的。根据（Nesterov，2004年）的定理 2.1.5，可得

$$\begin{aligned}
&\mathbb{E}[\|\partial\phi_s(\mathbf{v}_s)\|^2] \le 2\hat{L}\mathbb{E}(\phi_s(\mathbf{v}_s) - \min_{x\in\mathbb{R}^d}\phi_s(\mathbf{v})) \le 2\hat{L}\mathbb{E}[\mathrm{Gap}_s(\mathbf{v}_s,\alpha_s)] \\
&= 2\hat{L}\mathbb{E}[4\mathrm{Gap}_s(\mathbf{v}_s,\alpha_s) - 3\mathrm{Gap}_s(\mathbf{v}_s,\alpha_s)] \\
&\le 2\hat{L}\mathbb{E}\left[4\left(\frac{4\eta_s\sigma^2}{M} + \frac{1}{53}\left(\mathrm{Gap}_s(\mathbf{v}_0^s,\alpha_0^s) + \mathrm{Gap}_s(\mathbf{v}_s,\alpha_s)\right)\right) - 3\mathrm{Gap}_s(\mathbf{v}_s,\alpha_s)\right] \\
&= 2\hat{L}\mathbb{E}\left[\frac{16\eta_s\sigma^2}{M} + \frac{4}{53}\mathrm{Gap}_s(\mathbf{v}_0^s,\alpha_0^s) - \frac{155}{53}\mathrm{Gap}_s(\mathbf{v}_s,\alpha_s)\right].
\end{aligned} \tag{45}$$

应用引理 B.2 于（45），可得

$$\begin{aligned}
\mathbb{E}[\|\partial\phi_s(\mathbf{v}_s)\|^2] &\le 2\hat{L}\mathbb{E}\Bigg[\frac{16\eta_s\sigma^2}{M} + \frac{4}{53}\mathrm{Gap}_s(\mathbf{v}_0^s,\alpha_0^s) \\
&\qquad - \frac{155}{53}\left(\frac{3}{50}\mathrm{Gap}_{s+1}(\mathbf{v}_0^{s+1},\alpha_0^{s+1}) + \frac{4}{5}(\phi(\mathbf{v}_0^{s+1}) - \phi(\mathbf{v}_0^s))\right)\Bigg] \\
&= 2\hat{L}\mathbb{E}\left[\frac{16\eta_s\sigma^2}{M} + \frac{4}{53}\mathrm{Gap}_s(\mathbf{v}_0^s,\alpha_0^s) - \frac{93}{530}\mathrm{Gap}_{s+1}(\mathbf{v}_0^{s+1},\alpha_0^{s+1}) - \frac{124}{53}(\phi(\mathbf{v}_0^{s+1}) - \phi(\mathbf{v}_0^s))\right].
\end{aligned} \tag{46}$$

将其与（42结合，重新整理各项，并定义常数 $c = 4\ell + \frac{248}{53}\hat{}L \in O(L+\ell)$，可得 —

$$\begin{aligned}
&(c+2\mu)\,\mathbb{E}[\phi(\mathbf{v}_0^{s+1}) - \phi(\mathbf{v}_*)] + \frac{93}{265}\hat{L}\mathbb{E}[\mathrm{Gap}_{s+1}(\mathbf{v}_0^{s+1},\alpha_0^{s+1})] \\
&\le \left(4\ell + \frac{248}{53}\hat{L}\right)\mathbb{E}[\phi(\mathbf{v}_0^s) - \phi(\mathbf{v}_*)] + \frac{8\hat{L}}{53}\mathbb{E}[\mathrm{Gap}_s(\mathbf{v}_0^s,\alpha_0^s)] + \frac{32\eta_s\hat{L}\sigma^2}{M} \\
&\le c\mathbb{E}\left[\phi(\mathbf{v}_0^s) - \phi(\mathbf{v}_*) + \frac{8\hat{L}}{53c}\mathrm{Gap}_s(\mathbf{v}_0^s,\alpha_0^s)\right] + \frac{32\eta_s\hat{L}\sigma^2}{M}.
\end{aligned} \tag{47}$$

$\hat{}$ 利用 $L \ge \mu$ 这一事实，

$$(c+2\mu)\frac{8\hat{L}}{53c} = \left(4\ell + \frac{248}{53}\hat{L} + 2\mu\right)\frac{8\hat{L}}{53(4\ell + \frac{248}{53}\hat{L})} \le \frac{8\hat{L}}{53} + \frac{16\mu\hat{L}}{248\hat{L}} \le \frac{93}{265}\hat{L}. \tag{48}$$

那么，可得

$$\begin{aligned}
&(c+2\mu)\mathbb{E}\left[\phi(\mathbf{v}_0^{s+1}) - \phi(\mathbf{v}_*) + \frac{8\hat{L}}{53c}\mathrm{Gap}_{s+1}(\mathbf{v}_0^{s+1},\alpha_0^{s+1})\right] \\
&\le c\mathbb{E}\left[\phi(\mathbf{v}_0^s) - \phi(\mathbf{v}_*) + \frac{8\hat{L}}{53c}\mathrm{Gap}_s(\mathbf{v}_0^s,\alpha_0^s)\right] + \frac{32\eta_s\hat{L}\sigma^2}{M}.
\end{aligned} \tag{49}$$

$\hat{}$ 定义 $\Delta_s = \phi(\mathbf{v}_0^s) - \phi(\mathbf{v}_*) + \frac{8L}{53c}\mathrm{Gap}_s(\mathbf{v}_0^s,\alpha_0^s)$，则

$$\mathbb{E}[\Delta_{s+1}] \le \frac{c}{c+2\mu}\mathbb{E}[\Delta_s] + \frac{32\eta_s\hat{L}\sigma^2}{(c+2\mu)M} \tag{50}$$

递归地应用这个不等式，可得

$$E[\Delta_{S+1}] \le \left(\frac{c}{c+2\mu}\right)^S E[\Delta_1] + \frac{32\hat{L}\sigma^2}{(c+2\mu)M}\sum_{s=1}^{S}\left(\eta_s\left(\frac{c}{c+2\mu}\right)^{S-s}\right). \tag{51}$$

根据定义，

$$\begin{aligned}\Delta_1 &= \phi(\mathbf{v}_0^1) - \phi(\mathbf{v}^*) + \frac{8\hat{L}}{53c}\widehat{Gap}_1(\mathbf{v}_0^1, \alpha_0^1) \\ &= \phi(\mathbf{v}_0) - \phi(\mathbf{v}^*) + \frac{8\hat{L}}{53c}\left(f(\mathbf{v}_0, \hat{\alpha}_1(\mathbf{v}_0)) + \frac{\gamma}{2}\|\mathbf{v}_0 - \mathbf{v}_0\|^2 - f(\hat{\mathbf{v}}_1(\alpha_0), \alpha_0) - \frac{\gamma}{2}\|\hat{\mathbf{v}}_1(\alpha_0) - \mathbf{v}_0\|^2\right) \\ &\leq \epsilon_0 + \frac{8\hat{L}}{53c}\left(f(\mathbf{v}_0, \hat{\alpha}_1(\mathbf{v}_0)) - f(\hat{\mathbf{v}}(\alpha_0), \alpha_0)\right) \leq 2\epsilon_0.\end{aligned} \tag{52}$$

利用不等式 $1 - x \leq \exp(-x)$，我们有

$$\begin{aligned}\mathbb{E}[\Delta_{S+1}] &\leq \exp\left(\frac{-2\mu S}{c+2\mu}\right)\mathbb{E}[\Delta_1] + \frac{32\eta_0\hat{L}\sigma^2}{(c+2\mu)M}\sum_{s=1}^{S}\exp\left(-\frac{2\mu S}{c+2\mu}\right) \\ &\leq 2\epsilon_0\exp\left(\frac{-2\mu S}{c+2\mu}\right) + \frac{32\eta_0\hat{L}\sigma^2}{(c+2\mu)M}S\exp\left(-\frac{2\mu S}{(c+2\mu)}\right).\end{aligned}$$

要使其小于 $\epsilon$，只需令

$$\begin{aligned}2\epsilon_0\exp\left(\frac{-2\mu S}{c+2\mu}\right) &\leq \frac{\epsilon}{2}, \\ \frac{32\eta_0\hat{L}\sigma^2}{(c+2\mu)M}S\exp\left(-\frac{2\mu S}{c+2\mu}\right) &\leq \frac{\epsilon}{2}.\end{aligned} \tag{53}$$

令 $S$为满足 $\exp\left(\frac{-2\mu S}{c+2\mu}\right) \leq \min\{\frac{\epsilon}{4\epsilon_0}, \frac{(c+2\mu\ M\epsilon)}{64\eta_0 LS\sigma^2}\}$的最小值。我们可以设 $S = \max\left\{\frac{c+2\mu}{2\mu}\log\frac{4\epsilon_0}{\epsilon}, \frac{c+2\mu}{2\mu}\log\frac{64\eta_0\hat{L}S\sigma^2(c+2\mu)M\epsilon}{}\right\}$。

因此，总迭代复杂度为

$$\begin{aligned}\sum_{s=1}^{S}T_s &\leq O\left(\frac{424}{\eta_0\min\{\ell,\mu_2\}}\sum_{s=1}^{S}\exp\left((s-1)\frac{2\mu}{c+2\mu}\right)\right) \\ &\leq O\left(\frac{1}{\eta_0\min\{\ell,\mu_2\}}\frac{\exp(S\frac{2\mu}{c+2\mu})-1}{\exp(\frac{2\mu}{c+2\mu})-1}\right) \\ &\overset{(a)}{\leq} \widetilde{O}\left(\frac{c}{\eta_0\mu\min\{\ell,\mu_2\}}\max\left\{\frac{\epsilon_0}{\epsilon}, \frac{\eta_0\hat{L}S\sigma^2}{(c+2\mu)M\epsilon}\right\}\right) \\ &\leq \widetilde{O}\left(\max\left\{\frac{(L+\ell)\epsilon_0}{\eta_0\mu\min\{\ell,\mu_2\}\epsilon}, \frac{(L+\ell)^2\sigma^2}{\mu^2\min\{\ell,\mu_2\}M\epsilon}\right\}\right) \\ &\leq \widetilde{O}\left(\max\left\{\frac{1}{\mu_1\mu_2^2\epsilon}, \frac{1}{\mu_1^2\mu_2^3 M\epsilon}\right\}\right),\end{aligned} \tag{54}$$

其中，$(a)$采用 $S$的设置，且 $\exp(x) - 1 \geq x$，而 $O$省略了对数因子。

$\eta_s = \eta_0\exp(-(s-1)\frac{2\mu}{c+2\mu}), T_s = \frac{212}{\eta_0\mu_2}\exp\left((s-1)\frac{2\mu}{c+2\mu}\right)$.

接下来，我们将分析通信成本。

为保证我们在上述证明中使用的 $\left(\frac{3\ell^2}{2\mu_2} + \frac{3\ell}{2}\right)36\eta^2 I_s^2 K^2 G^2 \leq \frac{\eta_s\sigma^2}{M}$，需要取 $I_s = O(\frac{1}{K\sqrt{\eta_s M}})$。若 $\frac{1}{K\sqrt{\eta_0 M}} \leq O(1)$，对于 $s \leq S_2 := O(\frac{c+2\mu}{2\mu}\log(M\eta_0))$，则取 $I_s = 1$，相应地有 $E_s = T_s/KI_s = T_s/K$。对于 $s > S_2$，$I_s = O(\frac{\exp((s-1)\frac{\mu}{c+2\mu})}{K(\eta_0 M)^{1/2}})$，相应地有 $E_s = T_s/KI_s = O(KM^{1/2}\exp((s-1)\frac{\mu}{c+2\mu}))$。

我们有

$$\sum_{s=1}^{S_2} T_s = \sum_{s=1}^{S_2} O\left(\frac{212}{\eta_0}\exp\left((s-1)\frac{2\mu}{c+2\mu}\right)\right) = \widetilde{O}\left(\frac{M}{\mu}\right). \tag{55}$$

因此，通信复杂度可界定为

$$\begin{aligned}
&\sum_{s=1}^{S_2} T_s + \sum_{s=S_2+1}^{S} \frac{T_s}{I_s} = \widetilde{O}\left(\frac{K}{\mu} + \sqrt{M}K\exp\left(\frac{(s-1)\frac{2\mu}{c+2\mu}}{2}\right)\right) \\
&\leq \widetilde{O}(\frac{M}{\mu} + \sqrt{M}K\frac{\exp\left(\frac{S}{2}\frac{2\mu}{c+2\mu}\right)-1}{\exp\frac{\mu}{c+2\mu}-1}) \leq O\left(\frac{M}{\mu} + \frac{K}{\mu^{3/2}\epsilon^{1/2}}\right).
\end{aligned} \tag{56}$$

□

## C CyCp-FedX 分析

在本节中，我们给出定理5.2的分析。

证明。我们记

$$\begin{aligned}
G_1(\mathbf{w},\mathbf{z},\mathbf{w}',\mathbf{z}') &= \nabla_1\psi(h(\mathbf{w};\mathbf{z}),h(\mathbf{w};\mathbf{z}'))^\top\nabla h(\mathbf{w};\mathbf{z}) \\
G_2(\mathbf{w};\mathbf{z},\mathbf{w}';\mathbf{z}') &= \nabla_2\psi(h(\mathbf{w};\mathbf{z}),h(\mathbf{w};\mathbf{z}'))^\top\nabla h(\mathbf{w};\mathbf{z}') \\
G^{e,k}_{m,t,1} &= \nabla_1\psi(h(\mathbf{w}^{e,k}_{m,t},\mathbf{z}^{e,k}_{m,t,1}),h_{2,\xi})\nabla h(\mathbf{w}^{e,k}_{m,t},\mathbf{z}^{e,k}_{m,t,1}) \\
G^{e,k}_{m,t,2} &= \nabla_2\psi(h_{1,\xi},h(\mathbf{w}^{e,k}_{m,t},\mathbf{z}^{e,k}_{m,t,2}))\nabla h(\mathbf{w}^{e,k}_{m,t},\mathbf{z}^{e,k}_{m,t,2})
\end{aligned} \tag{57}$$

对于 $\tilde{\eta} = \eta I$,

$$\begin{aligned}
&F(\mathbf{w}^{e+1,0}) - F(\mathbf{w}^{e,0}) \leq \nabla F(\mathbf{w}^{e,0})^\top(\mathbf{w}^{e+1,0}-\mathbf{w}^{e,0}) + \frac{L}{2}\|\mathbf{w}^{e+1,0}-\mathbf{w}^{e,0}\|^2 \\
&= -\tilde{\eta}\nabla F(\mathbf{w}^{e,0})^\top\frac{1}{MI}\sum_{k=1}^{K}\sum_{m\in\mathcal{G}^{e,k}}\sum_{t=1}^{I}(G^{e,k}_{m,t,1}+G^{e,k}_{m,t,2}) + \frac{L}{2}\|\mathbf{w}^{e+1,0}-\mathbf{w}^{e,0}\|^2 \\
&= -\tilde{\eta}(\nabla F(\mathbf{w}^{e,0}) - \nabla F(\mathbf{w}^{e-1,0}) + \nabla F(\mathbf{w}^{e-1,0}))^\top\frac{1}{MI}\sum_{k=1}^{K}\sum_{m\in\mathcal{G}^{e,k}}\sum_{t=1}^{I}(G^{e,k}_{m,t,1}+G^{e,k}_{m,t,2}) + \frac{L}{2}\|\mathbf{w}^{e+1,0}-\mathbf{w}^{e,0}\|^2 \\
&\leq \frac{1}{2L}\|\nabla F(\mathbf{w}^{e,0}) - \nabla F(\mathbf{w}^{e-1,0})\|^2 + 2\tilde{\eta}^2L\|\frac{1}{MI}\sum_{k=1}^{K}\sum_{m\in\mathcal{G}^{e,k}}\sum_{t=1}^{I}(G^{e,k}_{m,t,1}+G^{e,k}_{m,t,2})\|^2 \\
&\quad - \tilde{\eta}\nabla F(\mathbf{w}^{e-1,0})^\top\left(\frac{1}{MI}\sum_{k=1}^{K}\sum_{m\in\mathcal{G}^{e,k}}\sum_{t=1}^{I}(G^{e,k}_{m,t,1}+G^{e,k}_{m,t,2})\right) + \frac{L}{2}\|\mathbf{w}^{e+1,0}-\mathbf{w}^{e,0}\|^2.
\end{aligned} \tag{58}$$

将 $k'$、$k''$、$m'$、$m''$、$t'$、$t''$ 记为对应于在$k$、 $m$、 $t$时使用被动部件的索引的随机变量,

$$
\begin{aligned}
&-\mathbb{E}\left[\tilde{\eta}\nabla F(\mathbf{w}^{e-1,0})^\top\left(\frac{1}{MI}\sum_{k=1}^{K}\sum_{m\in\mathcal{G}^{e,k}}\sum_{t=1}^{I}(G_{m,t,1}^{e,k}+G_{m,t,2}^{e,k})\right)\right]\\
&=-\mathbb{E}\Bigg[\tilde{\eta}\nabla F(\mathbf{w}^{e-1,0})^\top\frac{1}{MI}\sum_{k=1}^{K}\sum_{m\in\mathcal{G}^{e,k}}\sum_{t=1}^{I}\Bigg(G_1(\mathbf{w}_{m,t}^{e,k},\mathbf{z}_{m,t,1}^{e,k},\mathbf{w}_{t'}^{e-1,m'},\mathbf{z}_{m',t',2}^{e-1,k'})+G_2(\mathbf{w}_{t''}^{e-1,c''},\mathbf{z}_{m'',t'',1}^{e-1,k''},\mathbf{w}_{m,t}^{e,k},\mathbf{z}_{m,t,2}^{e,k})\\
&\qquad\qquad-G_1(\mathbf{w}^{e-1,0},\mathbf{z}_{m,t,1}^{e,k},\mathbf{w}^{e-1,0},\mathbf{z}_{m',t',2}^{e-1,k'})-G_2(\mathbf{w}^{e-1,0},\mathbf{z}_{m'',t'',1}^{e-1,k''},\mathbf{w}^{e-1,0},\mathbf{z}_{m,t,2}^{e,k})\\
&\qquad\qquad+G_1(\mathbf{w}^{e-1,0},\mathbf{z}_{m,t,1}^{e,k},\mathbf{w}^{e-1,0},\mathbf{z}_{m',t',2}^{e-1,k'})+G_2(\mathbf{w}^{e-1,0},\mathbf{z}_{m'',t'',1}^{e-1,k''},\mathbf{w}^{e-1,0},\mathbf{z}_{m,t,2}^{e,k})\Bigg)\Bigg]\\
&\overset{(a)}{\leq}4\tilde{\eta}KL^2\frac{1}{K}\sum_{k=1}^{K}\frac{1}{MI}\sum_{m\in\mathcal{G}^{e,k}}\sum_{t=1}^{I}\mathbb{E}(2\|\mathbf{w}_{m,t}^{e,k}-\mathbf{w}^{e-1,0}\|^2)+4\tilde{\eta}KL^2\frac{1}{K}\frac{1}{MI}\sum_{k'}\sum_{m'\in\mathcal{S}^{e-1,k'}}\sum_{t'}(\|\mathbf{w}_{m',t'}^{e-1,k'}-\mathbf{w}^{e-1,0}\|^2)\\
&\quad+\frac{\tilde{\eta}K}{4}\mathbb{E}\|\nabla F(\mathbf{w}^{e-1,0})\|^2-\mathbb{E}\left[\tilde{\eta}K\nabla F(\mathbf{w}^{e-1,0})^\top\left(\frac{1}{K}\sum_{k\in[K]}\nabla F_k(\mathbf{w}^{e-1,0})\right)\right]\\
&\leq 16\tilde{\eta}KL^2\|\mathbf{w}^{e,0}-\mathbf{w}^{e-1,0}\|^2+16\tilde{\eta}KL^2\frac{1}{KMI}\sum_{k}\sum_{m\in\mathcal{G}^{e,k}}\sum_{t}\|\mathbf{w}_{m,t}^{e,k}-\mathbf{w}^{e,0}\|^2\\
&\quad+8\tilde{\eta}KL^2\frac{1}{KMI}\sum_{k'}\sum_{m'\in\mathcal{S}^{e-1,k'}}\sum_{t}\|\mathbf{w}_{m',t'}^{e-1,k'}-\mathbf{w}^{e-1,0}\|^2-\frac{\tilde{\eta}K}{2}\mathbb{E}\|\nabla F(\mathbf{w}^{e-1,0})\|^2,
\end{aligned}
\tag{59}
$$

其中 (a) 成立是因为

$$
\mathbb{E}\left[G_1(\mathbf{w}^{e-1,0},\mathbf{z}_{m,t,1}^{e,k},\mathbf{w}^{e-1,0},\mathbf{z}_{m',t',2}^{e-1,k'})+G_2(\mathbf{w}^{e-1,0},\mathbf{z}_{m'',t'',1}^{e-1,k''},\mathbf{w}^{e-1,0},\mathbf{z}_{m,t,2}^{e,k})\right)\right]=\nabla F(\mathbf{w}^{e-1,0}).
\tag{60}
$$

通过对$\mathbf{w}$进行更新，我们得到

$$
\begin{aligned}
&\mathbb{E}\|\mathbf{w}^{e+1,0}-\mathbf{w}^{e,0}\|^2=\tilde{\eta}^2\left\|\frac{1}{MI}\sum_{k=1}^{K}\sum_{m\in\mathcal{G}^{e,k}}\sum_{t}(G_{m,t,1}^{e,k}+G_{m,t,2}^{e,k})\right\|^2\\
&=\tilde{\eta}^2\mathbb{E}\left\|\frac{1}{MI}\sum_{k=1}^{K}\sum_{m\in\mathcal{G}^{e,k}}\sum_{t}(G_1(\mathbf{w}_{m,t}^{e,k},\mathbf{z}_{m,t,1}^{e,k},\mathbf{w}_{m',t'}^{e-1,k'},\mathbf{z}_{m',t',2}^{e-1,k'})+G_2(\mathbf{w}_{m'',t''}^{e-1,k''},\mathbf{z}_{m'',t'',1}^{e-1,k''},\mathbf{w}_{m,t}^{e,k},\mathbf{z}_{m,t,2}^{e,k}))\right\|^2\\
&\leq 3\tilde{\eta}^2\mathbb{E}\Bigg\|\frac{1}{MI}\sum_{k=1}^{K}\sum_{m\in\mathcal{G}^{e,k}}\sum_{t}(G_1(\mathbf{w}_{m,t}^{e,k},\mathbf{z}_{m,t,1}^{e,k},\mathbf{w}_{m',t'}^{e-1,k'},\mathbf{z}_{m',t',2}^{e-1,k'})+G_2(\mathbf{w}_{m'',t''}^{e-1,k''},\mathbf{z}_{m'',t'',1}^{e-1,k''},\mathbf{w}_{m,t}^{e,k},\mathbf{z}_{m,t,2}^{e,k}))\\
&\qquad-\frac{1}{MI}\sum_{k=1}^{K}\sum_{m\in\mathcal{G}^{e,k}}\sum_{t}(G_1(\mathbf{w}^{e-1,0},\mathbf{z}_{m,t,1}^{e,k},\mathbf{w}^{e-1,0},\mathbf{z}_{m',t',2}^{e-1,k'})+G_2(\mathbf{w}^{e-1,0},\mathbf{z}_{m'',t'',1}^{e-1,k''},\mathbf{w}^{e-1,0},\mathbf{z}_{m,t,2}^{e,k}))\Bigg\|^2\\
&+3\tilde{\eta}^2\mathbb{E}\left\|\frac{1}{MI}\sum_{k=1}^{K}\sum_{m\in\mathcal{G}^{e,k}}\sum_{t}(G_1(\mathbf{w}^{e-1,0},\mathbf{z}_{m,t,1}^{e,k},\mathbf{w}^{e-1,0},\mathbf{z}_{t',2}^{e-1,k'})+G_2(\mathbf{w}^{e-1,0},\mathbf{z}_{m'',t'',1}^{e-1,k''},\mathbf{w}^{e-1,0},\mathbf{z}_{m,t,2}^{e,k}))-\nabla F(\mathbf{w}^{e-1,0})\right\|^2\\
&+3\tilde{\eta}^2K^2\mathbb{E}\|\nabla F(\mathbf{w}^{e-1,0})\|^2,
\end{aligned}
\tag{61}
$$

由此可得

$$
\begin{aligned}
&\mathbb{E}\|\mathbf{w}^{e+1,0}-\mathbf{w}^{e,0}\|^2\\
&\le 6\tilde{\eta}^2K^2\frac{\tilde{L}^2}{KMI}\sum_{k=1}^{K}\sum_{m\in\mathcal{G}^{e,k}}\sum_{t}\mathbb{E}\|\mathbf{w}_{m,t}^{e,k}-\mathbf{w}^{e,0}\|^2+6\tilde{\eta}^2K^2\frac{\tilde{L}^2}{KMI}\sum_{k=1}^{K}\sum_{m\in\mathcal{G}^{e,k}}\sum_{t}\mathbb{E}\|\mathbf{w}_{m',t'}^{e-1,k'}-\mathbf{w}^{e-1,0}\|^2\\
&+6\tilde{\eta}^2K^2\frac{\tilde{L}^2}{KMI}\sum_{k=1}^{K}\sum_{m\in\mathcal{G}^{e,k}}\sum_{t}\mathbb{E}\|\mathbf{w}^{e,0}-\mathbf{w}^{e-1,0}\|^2\\
&+3\tilde{\eta}^2K^2\mathbb{E}\left\|\frac{1}{KMI}\sum_{k=1}^{K}\sum_{m\in\mathcal{G}^{e,k}}\sum_{t}(G_1(\mathbf{w}^{e-1,0},\mathbf{z}_{m,t,1}^{e,k},\mathbf{w}^{e-1,0},\mathbf{z}_{m',t',2}^{e-1,k'})+G_2(\mathbf{w}^{e-1,0},\mathbf{z}_{m'',t'',1}^{e-1,k''},\mathbf{w}^{e,0},\mathbf{z}_{m,t,2}^{e,k}))-\nabla F_k(\mathbf{w}^{e-1,0})\right\|^2\\
&+3\tilde{\eta}^2K^2\mathbb{E}\|\nabla F(\mathbf{w}^{e-1,0})\|^2.
\end{aligned}
\tag{62}
$$

因此,

$$
\begin{aligned}
&\frac{1}{E}\sum_{e=1}^{E}\mathbb{E}\|\mathbf{w}^{e+1,0}-\mathbf{w}^{e,0}\|^2\\
&\le\frac{1}{E}\sum_{e=1}^{E}\Bigg[6\tilde{\eta}^2K^2\frac{\tilde{L}^2}{KMI}\sum_{k=1}^{K}\sum_{m\in\mathcal{G}^{e,k}}\sum_{t}\mathbb{E}\|\mathbf{w}_{m,t}^{e,k}-\mathbf{w}^{e,0}\|^2+6\tilde{\eta}^2K^2\frac{\tilde{L}^2}{KMI}\sum_{k=1}^{K}\sum_{m\in\mathcal{G}^{e,k}}\sum_{t}\mathbb{E}\|\mathbf{w}_{m,t}^{e-1,k}-\mathbf{w}^{e-1,0}\|^2\\
&\quad+6\tilde{\eta}^2K^2\frac{\tilde{L}^2}{KMI}\sum_{k=1}^{K}\sum_{m\in\mathcal{G}^{e,k}}\sum_{t}\mathbb{E}\|\mathbf{w}_{m',t'}^{e-1,k'}-\mathbf{w}^{e-1,0}\|^2+6\tilde{\eta}^2K^2\frac{\sigma^2}{KMI}+3\tilde{\eta}^2K^2\mathbb{E}\|\nabla F(\mathbf{w}^{e-1,0})\|^2\Bigg].
\end{aligned}
\tag{63}
$$

$$
\begin{aligned}
&\|\mathbf{w}_{m,t}^{e,k}-\mathbf{w}^{e,0}\|^2=\tilde{\eta}^2\left\|\frac{1}{I}\sum_{\tau=1}^{k}\sum_{\nu=1}^{t_\tau}(G_{m,\nu,1}^{e,\tau}+G_{m,\nu,2}^{e,\tau})\right\|^2\\
&=\tilde{\eta}^2\mathbb{E}\left\|\frac{1}{I}\sum_{\tau=1}^{k}\sum_{\nu=1}^{t_\tau}(G_1(\mathbf{w}_{m,t}^{e,\tau},\mathbf{z}_{m,t,1}^{e,\tau},\mathbf{w}_{m',t'}^{e-1,\tau'},\mathbf{z}_{m',t',2}^{e-1,\tau'})+G_2(\mathbf{w}_{m'',t''}^{e-1,\tau''},\mathbf{z}_{m'',t'',1}^{e-1,\tau''},\mathbf{w}_{m,t}^{e,\tau},\mathbf{z}_{m,t,2}^{e,\tau}))\right\|^2\\
&\le 4\tilde{\eta}^2\mathbb{E}\Bigg\|\frac{1}{I}\sum_{\tau=1}^{k}\sum_{\nu=1}^{t_\tau}(G_1(\mathbf{w}_{m,t}^{e,\tau},\mathbf{z}_{m,t,1}^{e,\tau},\mathbf{w}_{m',t'}^{e-1,\tau'},\mathbf{z}_{m',t',2}^{e-1,\tau'})+G_2(\mathbf{w}_{m'',t''}^{e-1,\tau''},\mathbf{z}_{m'',t'',1}^{e-1,\tau''},\mathbf{w}_{m,t}^{e,\tau},\mathbf{z}_{m,t,2}^{e,\tau}))\\
&\qquad-\frac{1}{I}\sum_{\tau=1}^{k}\sum_{\nu=1}^{t_\tau}(G_1(\mathbf{w}^{e-1,0},\mathbf{z}_{m,t,1}^{e,\tau},\mathbf{w}^{e-1,0},\mathbf{z}_{m',t',2}^{e-1,\tau'})+G_2(\mathbf{w}^{e-1,0},\mathbf{z}_{m'',t'',1}^{e-1,\tau''},\mathbf{w}^{e-1,0},\mathbf{z}_{m,t,2}^{e,\tau}))\Bigg\|^2\\
&+4\tilde{\eta}^2\mathbb{E}\left\|\frac{1}{I}\left[\sum_{\tau=1}^{k}\sum_{\nu=1}^{t_\tau}(G_1(\mathbf{w}^{e-1,0},\mathbf{z}_{m,t,1}^{e,\tau},\mathbf{w}^{e-1,0},\mathbf{z}_{t',2}^{e-1,\tau'})+G_2(\mathbf{w}^{e-1,0},\mathbf{z}_{m'',t'',1}^{e-1,\tau''},\mathbf{w}^{e-1,0},\mathbf{z}_{m,t,2}^{e,\tau}))-\nabla F_\tau(\mathbf{w}^{e-1,0})\right]\right\|^2\\
&+4\tilde{\eta}^2\mathbb{E}\|\frac{1}{I}\sum_{\tau=1}^{k}\sum_{\nu=1}^{t_\tau}(\nabla F_\tau(\mathbf{w}^{e-1,0})-\nabla F(\mathbf{w}^{e-1,0}))\|^2+4\tilde{\eta}^2\mathbb{E}\|\frac{1}{I}\sum_{\tau=1}^{k}\sum_{\nu=1}^{t_\tau}\nabla F(\mathbf{w}^{e-1,0})\|^2\\
&\le 8\tilde{\eta}^2\tilde{L}^2\frac{K}{I}\sum_{\tau=1}^{k}\sum_{\nu=1}^{t_\tau}\|\mathbf{w}_{m,t}^{e,\tau}-\mathbf{w}^{e,0}\|^2+8\tilde{\eta}^2\tilde{L}^2\frac{K}{I}\sum_{\tau=1}^{k}\sum_{\nu=1}^{t_\tau}\|\mathbf{w}^{e,0}-\mathbf{w}^{e-1,0}\|^2\\
&\quad+4\tilde{\eta}^2k^2\frac{\sigma^2}{kI}+4\tilde{\eta}^2k^2D^2+4\tilde{\eta}^2k^2\mathbb{E}\|\nabla F(\mathbf{w}^{e-1,0})\|^2.
\end{aligned}
\tag{64}
$$

$$\frac{1}{EKMI}\sum_{e}\sum_{k}\sum_{m\in\mathcal{G}^{e,k}}\sum_{t}\|\mathbf{w}_{m,t}^{e,k}-\mathbf{w}^{e,0}\|^2$$
$$\leq 8\tilde{\eta}^2\tilde{L}^2\frac{K}{I}\frac{1}{EKMI}\sum_{e=1}^{E}\sum_{\tau=1}^{K}\sum_{m\in\mathcal{S}^{e,\tau}}\sum_{t}KI\|\mathbf{w}_{m,t}^{e,\tau}-\mathbf{w}^{e,0}\|^2+8\tilde{\eta}^2\tilde{L}^2\frac{K}{I}\frac{1}{EKMI}\sum_{e=1}^{E}\sum_{\tau=1}^{K}\sum_{m\in\mathcal{S}^{e,\tau}}\sum_{t}KI\|\mathbf{w}^{e,0}-\mathbf{w}^{e-1,0}\|^2$$
$$+4\tilde{\eta}^2K\frac{\sigma^2}{I}+4\tilde{\eta}^2K^2D^2+4\tilde{\eta}^2K^2\frac{1}{E}\sum_{e=1}^{E}\mathbb{E}\|\nabla F(\mathbf{w}^{e-1,0})\|^2. \tag{65}$$

$$\frac{1}{EKMI}\sum_{e}\sum_{k}\sum_{m\in\mathcal{G}^{e,k}}\sum_{t}\|\mathbf{w}_{m,t}^{e,k}-\mathbf{w}^{e,0}\|^2$$
$$\leq 16\tilde{\eta}^2\tilde{L}^2K^2\frac{1}{E}\sum_{e=1}^{E}\|\mathbf{w}^{e,0}-\mathbf{w}^{e-1,0}\|^2+8\tilde{\eta}^2K\frac{\sigma^2}{I}+8\tilde{\eta}^2K^2\alpha^2+8\tilde{\eta}^2K^2\mathbb{E}\|\nabla F(\mathbf{w}^{e-1,0})\|^2. \tag{66}$$

˜使用 $\tilde{\eta}LK\leq O(1)$,

$$\frac{1}{EKMI}\sum_{e}\sum_{k}\sum_{m\in\mathcal{G}^{e,k}}\sum_{t}\|\mathbf{w}_{m,t}^{e,k}-\mathbf{w}^{e,0}\|^2\leq 16\tilde{\eta}^2K\frac{\sigma^2}{I}+16\tilde{\eta}^2K^2\alpha^2+16\tilde{\eta}^2K^2\|\nabla F(\mathbf{w}^{e-1,0})\|^2 \tag{67}$$

and

$$\frac{1}{E}\sum_{e=1}^{E}\mathbb{E}\|\mathbf{w}^{e+1}-\mathbf{w}^{e}\|^2\leq 16\tilde{\eta}^2K\frac{\sigma^2}{MI}+16\tilde{\eta}^2K^2\|\nabla F(\mathbf{w}^{e-1,0})\|^2. \tag{68}$$

将这些界代入（58）和（59）中,

$$F(\mathbf{w}^{e+1,0})-F(\mathbf{w}^{e,0})\leq\tilde{L}^2\|\mathbf{w}^{e,0}-\mathbf{w}^{e-1,0}\|^2+\tilde{\eta}^2L^2K\|\mathbf{w}^{e,0}-\mathbf{w}^{e-1,0}\|^2$$
$$+16\tilde{\eta}^2\tilde{L}^2\frac{1}{MI}\sum_{m\in\mathcal{G}^{e,k}}\sum_{t}\|\mathbf{w}_{m,t}^{e,k}-\mathbf{w}^{e,0}\|^2-\frac{\tilde{\eta}K}{2}\mathbb{E}\|\nabla F(\mathbf{w}^{e-1,0})\|^2. \tag{69}$$

$$\frac{1}{E}\sum_{e=1}^{E}\mathbb{E}\|\nabla F(\mathbf{w}^{e-1})\|^2\leq O\left(\frac{F(\mathbf{w}^{e,0})-F(\mathbf{w}_*)}{\tilde{\eta}KE}+\tilde{\eta}^2K\frac{\sigma^2}{I}+\tilde{\eta}^2K^2D^2+24\tilde{\eta}K\frac{\sigma^2}{KMI}\right). \tag{70}$$

由于 $\tilde{\eta}=\eta I$。设 $\eta=O(\epsilon^2M)$, $I=O(\frac{1}{MK\epsilon})$, $E=\frac{1}{\epsilon^3})$, 迭代复杂度为 $EKI=O(\frac{1}{M\epsilon^4})$,—通信成本为 $EK=O(\frac{K}{\epsilon^3})$。 — □

## D CyCp-FedX 的 PL条件分析

下面给出定理5.3的证明。

证明。 利用 PL条件的性质，我们有

$$F(\mathbf{w}^{s+1,0})-F(\mathbf{w}_*)\leq\frac{1}{2\mu}\|\nabla F(\mathbf{w}^{s+1,0})\|^2 \tag{71}$$

为确保 $F(\mathbf{w}^{s+1,0})-F(\mathbf{w}_*) \leq \epsilon_s, \|\nabla F(\mathbf{w}^{s+1,0})\|^2 \leq \mu\epsilon_s$。令 $\epsilon_0 = \max(F(\mathbf{w}^{0,0})-F(\mathbf{w}_*)\ \|\nabla F(\mathbf{w}^{0,0})\|^2/\mu)$。我们需要

$$\begin{aligned} \frac{F(\mathbf{w}^{s,0}) - F(\mathbf{w}_*)}{\tilde{\eta}KE} &\leq \frac{\mu\epsilon_s}{3} \\ \tilde{\eta}^2K^2\alpha^2 &\leq \frac{\mu\epsilon_s}{3} \\ \tilde{\eta}K\frac{\sigma^2}{KMI} &\leq \frac{\mu\epsilon_s}{3}, \end{aligned} \tag{72}$$

其中

(73)

令 $\eta_s = O(\frac{\mu\epsilon_s M}{\sigma^2})$、 $I_s = O(\frac{\sigma^2}{MK\sqrt{\mu\epsilon_s}})$、 $E_s = (\frac{1}{\mu^{3/2}\epsilon_s^{1/2}})$。每个阶段的迭代次数为 $I_sE_s = \frac{\sigma^2}{MK\mu^2\epsilon_s}$。为确保 $\epsilon_S \leq \epsilon$，阶段总数为 $\log(\mu\epsilon)$。因此，总迭代复杂度为

$$\sum_{s=1}^{S} I_sE_s = O(\sum_{s=1}^{S}\frac{1}{MK\mu^2\epsilon_s}) = O(\frac{1}{MK\mu^2\epsilon}). \tag{74}$$

总通信复杂度为

$$\sum_{s=1}^{S} E_sK = O(\sum_{s=1}^{S}\frac{1}{\mu^{3/2}\epsilon_s^{1/2}}K) = O(\frac{K}{\mu^{3/2}\epsilon^{1/2}}). \tag{75}$$

□

## E 数据统计

| 数据集 | 拆分 | 正样本 | 负样本 | 正样本 % |
|---|---|---|---|---|
| CIFAR-10 | 训练集 | 224 | 40,505 | 0.55% |
| | 验证集 | 505 | 4,495 | 10.10% |
| | Test | 1,000 | 9,000 | 10.00% |
| CIFAR-100 | 训练集 | 227 | 44,546 | 0.51% |
| | 验证集 | 46 | 4,954 | 0.92% |
| | Test | 100 | 9,900 | 10.00% |
| ChestMNIST | 训练集 | 1,994 | 74,480 | 2.61% |
| | 验证集 | 625 | 10,594 | 5.57% |
| | Test | 1,133 | 21,300 | 5.05% |
| 保险 | 训练集 | 2,863 | 4,028,889 | 0.07% |
| | 验证集 | 265 | 1,059,078 | 0.03% |
| | Test | 132 | 1,085,053 | 0.01% |

表7：数据统计（未翻转）。

## F 更多实验结果

本节通过消融研究，测试算法在不同设置下的性能。

在表 8中，我们展示了客户端数量不同时的实验结果。我们的方法仍然具有优势。

表 9 报告了三轮独立随机数据移除的结果，每轮使用不同随机种子运行三次。在所有重复实验中，我们的方法均始终优于基线方法。

表8：客户端数量消融（$dir$=0.1， $flip$=0）

| | CIFAR-10 | | CIFAR-100 | |
|---|---|---|---|---|
| | 50个客户端 | 200个客户端 | 50个客户端 | 200个客户端 |
| CyCp-FedAVG | $0.7227 \pm 0.0015$ | $0.7846 \pm 0.0011$ | $0.8817 \pm 0.0020$ | $0.8920 \pm 0.0018$ |
| A-FedAVG | $0.7418 \pm 0.0026$ | $0.7923 \pm 0.0016$ | $0.8756 \pm 0.0025$ | $0.8973 \pm 0.0012$ |
| A-SCAFFOLD | $0.7450 \pm 0.0024$ | $0.7940 \pm 0.0013$ | $0.8866 \pm 0.0017$ | $0.9012 \pm 0.0015$ |
| RS-Minimax | $0.7572 \pm 0.0018$ | $0.8012 \pm 0.0006$ | $0.8994 \pm 0.0013$ | $0.9183 \pm 0.0013$ |
| RS-Pairwise | $0.7721 \pm 0.0014$ | $0.8225 \pm 0.0010$ | $0.9188 \pm 0.0012$ | $\mathbf{0.9422 \pm 0.0021}$ |
| CyCp-Minimax | $\mathbf{0.8209 \pm 0.0012}$ | $\mathbf{0.8478 \pm 0.0016}$ | $\mathbf{0.9275 \pm 0.0022}$ | $0.9365 \pm 0.0016$ |
| CyCp-Pairwise | $\mathbf{0.8367 \pm 0.0013}$ | $\mathbf{0.8448 \pm 0.0017}$ | $\mathbf{0.9647 \pm 0.0015}$ | $\mathbf{0.9653 \pm 0.0019}$ |

Ta表9：随机移除数据的重复实验。（Cifar-10/100 的 $dir$=0.5 ）

| | Cifar-10 | Cifar-100 |
|---|---|---|
| CyCp-FedAVG | $0.8083 \pm 0.0175$ | $0.8815 \pm 0.0073$ |
| A-FedAVG | $0.8123 \pm 0.0037$ | $0.9006 \pm 0.0052$ |
| A-SCAFFOLD | $0.8208 \pm 0.0041$ | $0.9016 \pm 0.0039$ |
| RS-Minimax | $0.8225 \pm 0.0028$ | $0.9342 \pm 0.0044$ |
| RS-Pairwise | $0.8303 \pm 0.0019$ | $\mathbf{0.9541 \pm 0.0037}$ |
| CyCp-Minimax | $\mathbf{0.8460 \pm 0.0060}$ | $0.9422 \pm 0.0029$ |
| CyCp-Pairwise | $\mathbf{0.8503 \pm 0.0048}$ | $\mathbf{0.9730 \pm 0.0055}$ |

在图 2中，我们将数据划分为 10 个客户端组，每组包含 10 个客户端。通过改变 $M$（即同时参与的客户端数量），我们观察到， $M$ 值越大，收敛越快，这证实了预期的加速效果。

在图3中，我们将客户端（$N = 100$）随机划分为不同数量的组。每轮采样一个客户端参与。我们观察到，较大的 $K$ 值通常会带来更快的收敛，而 $K = 1$ 则基本等同于随机采样基线。在现实世界部署中，自然形成的组（例如按地理位置、设备类型或可用性模式划分）可能具有内部相似的数据分布。将这类组拆分成更小的组，可能会导致模型在充分探索其他分布之前，就对狭窄的数据分布过拟合。因此， $K$ 的最优选择取决于现实世界异质性及分组结构。进一步系统地研究 $K$与真实数据分布之间的相互作用，是未来工作的重要方向。

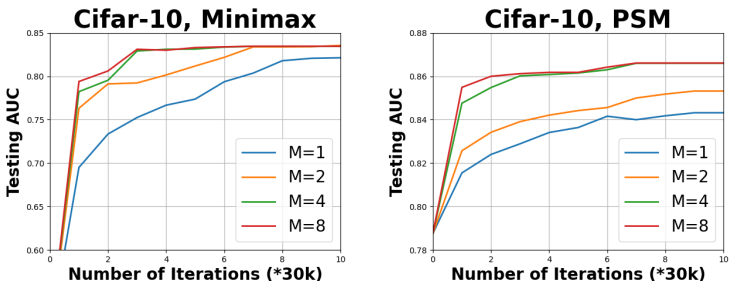
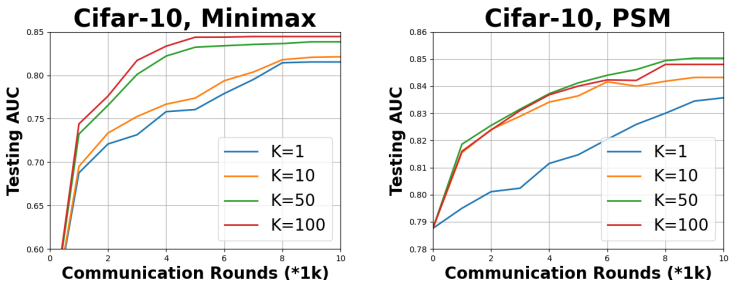



图2：消融研究：同时参与的客户端数量的影响 $M$（$dir = 0.5$， $flip$ =0）

## G 适用于一大类 AUC 最大化公式

表10列出了多种 AUC 一致损失及其与我们算法的兼容性。




Fig图3：消融研究：客户端组数量的影响 $K$（$dir = 0.5$， $flip = 0$ ）

表10：对不同 AUC 最大化公式的适用性（$m$表示作为超参数的间隔常数， $\tau$ 表示缩放超参数）。

| Loss | 公式 | 是否适用? |
|---|---|---|
| 极小极大损失（Ying 等人，2016年） | (3) | 是（算法1，2） |
| 成对平方损失（Gao 等人，2013） | (1) 其中 $\psi(a, b) = (m - (a-b))^2$ | 是（算法3，4） |
| 成对平方合页损失（Zhao等人， 2011b） | (1) 其中 $\psi(a, b) = (m - (a-b))_+^2$ | 是（算法3，4） |
| 成对逻辑损失（Gao和Zhou，2015） | (1) 使用 $\psi(a, b) = \log(1 + \exp(-s(a-b)))$ | 是（算法3，4） |
| 成对 Sigmoid 损失（Calders和Jaroszewicz， 2007） | (1) 使用 $\psi(a, b) = (1 + \exp(s(a-b)))^{-1}$ | 是（算法3、4） |
| 成对屏障合页损失（Charoenphakdee等人，2019b） | (1) 其中 $\psi(a, b) = \max(m - \tau(m+t),$ $\max(\tau(t-m)\ \ m-t))$, 其中 $t = a - b$ | 是/否[1]（算法 3、4） |
| q-范数合页损失 | (1) 使用 $\psi(a, b) = (m - (a-b))_+^q (q > 1)$ | 是（算法 3、4） |

[1]对于成对屏障合页损失，我们的算法仍然可以直接适用。然而，由于该替代损失不满足光滑性假设，从理论上讲，它不具备线性加速保证，这是非光滑联邦学习目标的常见局限（Yuan 等人，2021a）。