arXiv:2305.05090v1 [cs.LG] 8 May 2023

表演式联邦学习：解决模型依赖性和异质分布偏移的方法

预印本

Kun Jin *密歇根大学 kunj@umich.edu

尹同心 *密歇根大学 tyin@umich.edu

陈中柱 *密歇根大学 zhongzhc@umich.edu

Zeyu Sun密歇根大学 zeyusun@umich.edu

Xueru Zhang
俄亥俄州立大学
zhang.12807@osu.edu

Yang Liu加利福尼亚大学圣克鲁兹分校 字节跳动人工智能实验室 yangliu@ucsc.edu

Mingyan Liu
密歇根大学
mingyan@umich.edu

2023年5月10日

摘要

我们考虑一个由多个客户端和一台服务器组成的联邦学习（FL）系统，其中客户端旨在利用各自的分布式数据协同学习一个共同的决策模型。不同于假设客户端数据静态不变的传统联邦学习框架，我们考虑这样一种场景：客户端的数据分布可能会受到所部署决策模型的重塑。在本文中，我们借鉴施为式预测 中的分布偏移映射思想，对这种依赖于模型的数据分布偏移进行形式化，并提出一种行为联邦学习框架。首先，我们给出唯一表现型稳定解存在的充要条件，并刻画其与表现型最优解之间的距离。随后，我们提出行为 FedAvg 算法，并证明在全参与和部分参与方案下，该算法均以 $\mathcal{O}(1/T)$ 的速率收敛至表现型稳定解。特别地，我们采用新颖的证明技术，展示客户端异质性如何影响收敛。数值结果验证了我们的分析，并为实际应用提供了有价值的洞见。

1 引言

传统学习问题通常假设数据分布是静态的。对于人脸识别等应用而言，这在很大程度上确实如此；在这一假设下设计算法，通常不会影响学习效果。然而，许多其他领域并非如此。在某些情况下，数据分布会自然演变并发生偏移，例如天气和气候数据；此时需要定期获取新数据并重新训练算法，以使其保持最新状态。在另一些情况下，分布偏移是学习结果本身造成的：人们会对算法施加于他们的决策作出反应。例如，如果具有某些口音的用户发现语音识别软件的错误率高得难以接受，因而停止使用该软件，这就会直接影响软件为训练下一代产品所收集的语音样本类型。另一个例子是“算法博弈”：用户通过诚实或不诚实的手段，试图改善关键特征，以便获得算法的有利决策（例如贷款审批或求职申请）。这同样会直接导致算法决策所依据的特征和标签分布发生变化。

*这些作者对本研究贡献相同。

后一类分布偏移尤其值得关注：已部署的模型本身会触发数据分布变化，并影响目标，这种现象称为施为式。在此类分布偏移存在的情况下进行预测，称为施为式预测 Perdomo 等人 [2020]。施为式预测的典型场景包括策略学习 Hardt 等人 [2016], Dong 等人 [2018], Milli 等人 [2019],Hu 等人 [2019], Braverman 和 Garg [2020], Chen 等人 [2020], Miller 等人 [2020],Shavit 等人 [2020], Haghtalab 等人 [2020], Kleinberg 和 Raghavan [2020], Zrnic 等人 [2021]。

施为式预测主要在集中式设置下得到研究，相关文献成果丰硕，包括收敛性分析方面的研究：Mendler-D¨unner 等人 [2020], Drusvyatskiy 和 Xiao [2020], Brown 等人 [2020], Li 和 Wai[2022], Wood 等人 [2022] ，以及算法开发方面的研究：Izzo 等人 [2021, 2022], Miller 等人 [2021], Ray 等人 [2022]。

在现代大规模机器学习中，与集中式学习相比，分布式学习能够提供更好的隐私保护，并更好地避免计算资源瓶颈，而联邦学习（FL）是其中最受欢迎的示例之一。在分布式设置中，由于数据异质性，分布偏移问题进一步加剧。具体而言，分布式数据源在性质上可能存在异质性，其各自的分布偏移也可能不同。以往在联邦学习系统中处理数据分布偏移的工作，例如 Guo 等人 [2021], Casado 等人 [2022], Rizk 等人 [2020], Hosseinalipour 等人 [2022], Zhu 等人 [2021a], Eichner 等人 [2019], Ding 等人 [2020],，通常不会考虑由模型在客户端引起的局部分布偏移。在本文中，我们提出了表演式联邦学习框架，以研究和处理联邦学习中的此类数据偏移。

将施为式预测中的现有结果扩展到去中心化联邦学习面临诸多挑战。以下列举几个：1) 数据异质性： 数据异质性本就是联邦学习中的主要难题之一；若将客户端分布偏移的差异纳入考量，应对数据异质性会面临额外挑战。2) 集中式 ⇄ 本地： 训练过程中，客户端会在特定步骤接收聚合模型，并以此为基础进行训练。尽管这种聚合能更好地拟合整体，但可能无法很好地适配每个客户端，从而导致更严重的分布偏移问题。3) 偏移异质性： 有些客户端可能对部署的决策更敏感，数据偏移也比其他客户端更剧烈，例如，由于策略学习中的操纵成本不同。

为此，我们正式提出表现型 *FedAvg*算法，即 P-FedAvg，并证明其收敛性。我们的主要发现如下。

- 我们证明了该算法所达到的表演稳定（PS）解的唯一性，并表明在温和条件下，该解是表演最优（PO）解的可证明近似。我们将在第 2.1 节中正式定义这两种解。更有趣的是，我们表明，稳定解具有其自身的博弈论解释，即最佳响应动态的不动点。
- 我们将在第 3.3 节证明，在与先前研究类似的温和假设下，P-FedAvg算法在全参与和部分参与方案中均收敛到自适应稳定解，并具有$\mathcal{O}(1/T)$收敛速率。$\mathcal{O}(1/T)$
- 在此过程中，我们还引入了一些新的证明技术：我们证明了在不需要有界梯度假设的情况下也能实现收敛。这一技术可以直接应用于传统联邦学习，而传统联邦学习是可表演性设定的一种特殊情况。

1.1 相关工作

联邦学习。 我们的工作与联邦学习（FL）相关领域的研究密切相关。尽管许多研究尝试通过约束梯度优化和知识蒸馏来解决联邦学习中的客户端异质性 Li 等人 [2020a], Karimireddy 等人 [2020], Wang 等人 [2020], Haddadpour 等人 [2021], Zhu 等人 [2021b], Li 和 Wang [2019], Lin 等人 [2020], ，但其中大多数仍假设数据是静态的，没有考虑分布偏移。据我们所知，近期只有少数研究考虑了联邦学习中的分布偏移 Guo 等人 [2021], Casado 等人[2022], Rizk 等人 [2020],Hosseinalipour 等人 [2022], Zhu 等人 [2021a], Eichner 等人 [2019], Ding 等人 [2020]。例如，Guo 等人 [2021] 研究了客户端随时间演化的联邦学习，其中将每个客户端的时间漂移建模为与时间无关的 加性噪声，该噪声具有零均值 和有界方差。Casado 等人 [2022] 提出了一种能够适应分布漂移的联邦学习算法；该算法在整个学习过程中监测模型预测的置信度分数，并假设当置信度分数显著下降时发生漂移。Rizk 等人 [2020] 也研究了动态联邦学习，并假设随时间演化的数据背后的真实模型遵循随机游走。Hosseinalipour 等人 [2022]研究了具有动态客户端的联邦学习，并通过连续两个时间步之间局部损失 的变化来建模漂移。Zhu 等人 [2021a], Eichner 等人 [2019], Ding 等人 [2020] 研究了客户端群体的周期性 分布偏移

在联邦学习中的情况；他们假设客户端具有块循环 结构，即来自两个不同时区的客户端交替参与训练。

施为式预测。 除了我们在引言中讨论的、侧重于施为式预测集中式设置的研究之外，近期，Li 等人 [2022] 形式化了多智能体/多参与者施为式预测，其中智能体试图学习共同决策规则，但对模型具有异质的分布偏移（响应），并研究了去中心化算法收敛到 PS 解的性质。去中心化施为式预测刻画了智能体/客户端对决策模型响应的异质性，并避免了为训练而集中收集数据。这项工作为我们构建表演式联邦学习框架提供了启发，而我们提出的 P-FedAvg 可被视为一种重要的算法扩展，支持不平衡数据、频率低得多的同步以及设备部分参与。Narang 等人 [2022] 提出了一种去中心化多参与者施为式预测框架，其中参与者会对竞争机构的行为作出反应。Raab 和 Liu [2021]提出了一种具有标签偏移的复制动态模型。

策略分类与回归。 正如 Perdomo 等人 [2020], 所讨论的，施为式预测可用于解决重复性的策略分类与回归问题。我们可以使用 Stackelberg 博弈对这些问题进行建模：决策者在第一阶段通过设计、发布并承诺采用一项决策规则采取行动；随后，智能体在第二阶段对该决策规则做出最佳响应，通过操纵自身特征来获得更理想的决策结果，而这种操纵可以通过分布偏移映射进行建模。传统的策略学习文献重点在于寻找 Stackelberg 均衡 Hardt 等人 [2016], Kleinberg 和 Raghavan[2020], Shavit 等人 [2020], Haghtalab 等人 [2020], , 即 PO 解，在该解中，决策者和智能体都知道彼此的效用；而施为式预测无论是否了解效用，都能在重复性的策略学习问题中找到 PS 解。

2 问题表述

本节将表述表演式联邦学习问题，定义学习目标，并介绍我们用于优化目标函数的表现型联邦学习算法。

为了帮助理解表演式联邦学习，我们首先回顾 Perdomo 等人 [2020]提出的施为式预测问题。考虑一个典型的损失最小化问题，其中数据分布会因模型参数而发生偏移，这种关系可表示为映射 $\mathcal{D}(\boldsymbol{\theta})$。因此，目标函数为

$$f(\boldsymbol{\theta}) := \mathbb{E}_{Z\sim\mathcal{D}(\boldsymbol{\theta})}[\ell(\boldsymbol{\theta}; Z)],$$

其中，ℓ表示损失函数。于是，施为最优（PO）解为 $\boldsymbol{\theta}^{PO} := \arg\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta})$。Perdomo 等人 [2020] 还引入了第二个相互解耦的目标函数，也称为施为稳定（PS）模型，˜它将决策参数（$\boldsymbol{\theta}$）与部署参数（$\boldsymbol{\theta}$）区分开来：

$$f(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}}) := \mathbb{E}_{Z\sim\mathcal{D}(\tilde{\boldsymbol{\theta}})}[\ell(\boldsymbol{\theta}; Z)].$$

最小化该目标函数，可以在部署参数所诱导的分布下实现最小风险，从而无需重新训练，因此更具实用性。PS 解定义为 $\boldsymbol{\theta}^{PS} := \arg\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta}; \boldsymbol{\theta}^{PS})$。Perdomo 等人 [2020] 证明，$\boldsymbol{\theta}^{PS} \neq \boldsymbol{\theta}^{PO}$一般而言成立。接下来，我们将考虑分布式设置，并介绍表演式联邦学习。

2.1 系统设置与目标

考虑一个包含 N 个客户端和一个服务器的系统，其中客户端具有特征分布 $\mathcal{D}_i(\boldsymbol{\theta})$，其支持集为$\mathcal{Z} \subseteq \mathbb{R}^M$，而 $\boldsymbol{\theta} \in \mathbb{R}^m$ 表示部署在第 i个客户端上的决策（模型）参数。我们考虑客户端具有异质分布 $\mathcal{D}_i(\boldsymbol{\theta}) \neq \mathcal{D}_j(\boldsymbol{\theta})$ 的一般情况，并且每个客户端代表总数据总体 $\sum_{i=1}^{N} p_i = 1$的 $p_i > 0$ 比例。

该系统旨在最小化所有智能体损失的加权平均值，其表演最优目标如下

$$\boldsymbol{\theta}^{PO} := \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^m} \sum_{i=1}^{N} p_i \mathbb{E}_{Z_i\sim\mathcal{D}_i(\boldsymbol{\theta})}[\ell(\boldsymbol{\theta}; Z_i)]. \tag{1}$$

我们注意到，当对所有 i 都有 $p_i = \frac{1}{N}$ 时，我们的目标可简化为 Li 等人 [2022] 中的目标。该目标通常可以用于建模系统中具有不同子群体的策略学习问题，其中每个客户端对应一个子群体。各子群体在某些属性上可能存在差异，因此对决策参数的响应也不同，例如行动成本不同 Milli 等人 [2019], Hu 等人 [2019], Braverman 和 Garg [2020], Zhang 等人 [2022], Jin 等人 [2022]。决策者对整个群体采用共同决策规则，并旨在最小化期望损失，而 p_i 表示各子群体的群体占比。相应地，解耦式/表演稳定目标为

$$f_i(\boldsymbol{\theta};\tilde{\boldsymbol{\theta}}) := \mathbb{E}_{Z_i\sim\mathcal{D}_i(\tilde{\boldsymbol{\theta}})}[\ell(\boldsymbol{\theta};Z_i)], f(\boldsymbol{\theta};\tilde{\boldsymbol{\theta}}) := \sum_{i=1}^{N} p_i f_i(\boldsymbol{\theta};\tilde{\boldsymbol{\theta}}),$$

其中，第一个参数表示客户端的决策参数，第二个参数表示部署参数；部署参数与 $\mathcal{D}_i(\cdot)$共同决定样本的分布。PS 解为

$$\begin{aligned}\boldsymbol{\theta}^{PS} :=& \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} p_i \mathbb{E}_{Z_i\sim\mathcal{D}_i(\boldsymbol{\theta}^{PS})}[\ell(\boldsymbol{\theta};Z_i)] \\ =& \arg\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta};\boldsymbol{\theta}^{PS}).\end{aligned} \tag{2}$$

请注意，这是一个以 $\boldsymbol{\theta}^{PS}$ 为不动点的不动点方程。

2.2 关键假设

我们作出以下假设，与 Li 等人 [2022],和 Perdomo 等人 [2020]的假设类似：

假设 **2.1** $\tilde{\boldsymbol{\theta}} \in \mathbb{R}^m$ $f\cdot\tilde{\boldsymbol{\theta}}$ μ $\boldsymbol{\theta}$ $f(\boldsymbol{\theta}';\tilde{\boldsymbol{\theta}}) \geq f(\boldsymbol{\theta};\tilde{\boldsymbol{\theta}}) + \langle\nabla f(\boldsymbol{\theta};\tilde{\boldsymbol{\theta}}), \boldsymbol{\theta}'-\boldsymbol{\theta}\rangle + \frac{\mu}{2}\|\boldsymbol{\theta}'-\boldsymbol{\theta}\|_2^2 \ \forall \boldsymbol{\theta}', \boldsymbol{\theta}\in\mathbb{R}^K$ (强凸性)。对于任意 ，(,) 在 上是 -强凸的，即 (;) (;) (;),,。

在假设 2.1 中，我们不要求每个 f_i 都是强凸的，只要求其加权平均 f 是强凸的。

假设 2.2 (光滑性)。 损失函数 $\ell(\boldsymbol{\theta};z)$ 是 L-光滑的，即，

$$\|\nabla\ell(\boldsymbol{\theta};\boldsymbol{z}) - \nabla\ell(\boldsymbol{\theta}';\boldsymbol{z}')\|_2 \leq L(\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|_2 + \|\boldsymbol{z}-\boldsymbol{z}'\|_2).$$

假设 2.3 (分布映射敏感性)。 对于任意 $i = 1,\ldots,n$，存在 $\epsilon_i > 0$ 使得

$$\mathcal{W}_1(\mathcal{D}_i(\boldsymbol{\theta}), \mathcal{D}_i(\boldsymbol{\theta}')) \leq \epsilon_i\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|_2, \ \forall\boldsymbol{\theta}', \boldsymbol{\theta}\in\mathbb{R}^m,$$

其中 $\mathcal{W}_1(\mathcal{D}, \mathcal{D}')$ 是分布 $\mathcal{D}$, $\mathcal{D}'$ 在 L_2 范数下的 1-Wasserstein 距离。

假设 2.2 和 2.3 共同保证了 $f_i(\cdot,\cdot)$ 的光滑性。这是 Drusvyatskiy 和 Xiao [2022] 中引理 2.1 的结果，并将在后续证明中使用。

引理 2.4 （∇f_i的连续性）。 在假设 *2.2* 和 *2.3* 下，对于任意 $\boldsymbol{\theta}_0$、$\boldsymbol{\theta}_1$、$\boldsymbol{\theta}$、$\hat{\boldsymbol{\theta}} \in \mathbb{R}^m$,

$$\|\nabla f_i(\boldsymbol{\theta}_0;\boldsymbol{\theta}) - \nabla f_i(\boldsymbol{\theta}_1;\hat{\boldsymbol{\theta}})\|_2 \leq L\|\boldsymbol{\theta}_0-\boldsymbol{\theta}_1\|_2 + L\epsilon_i\|\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\|_2.$$

我们引入以下专为去中心化施为预测提出的假设 Li 等人 [2022]。

假设 2.5 （随机梯度方差界）。 对于任意 $i = 1, \ldots, N$ 和 $\boldsymbol{\theta}\in\mathbb{R}^m$，存在 $\sigma \geq 0$ 使得

$$\mathbb{E}_{Z_i\sim\mathcal{D}_i(\boldsymbol{\theta})}\|\nabla\ell(\boldsymbol{\theta};Z_i) - \nabla f_i(\boldsymbol{\theta};\boldsymbol{\theta})\|_2^2 \leq \sigma^2(1+\|\boldsymbol{\theta}-\boldsymbol{\theta}^{PS}\|_2^2).$$

假设 **在 2.6 节（局部梯度界）**。 对于任意 $i = 1, \ldots..., N$和 $\boldsymbol{\theta}\in\mathbb{R}^m$，存在 $\varsigma \geq 0$，使得 at

$$\|\nabla f(\boldsymbol{\theta};\boldsymbol{\theta}) - \nabla f_i(\boldsymbol{\theta};\boldsymbol{\theta})\|_2^2 \leq \varsigma^2(1+\|\boldsymbol{\theta}-\boldsymbol{\theta}^{PS}\|_2^2).$$

这里我们详细阐述假设 2.6，并解释为何采用该假设，而非联邦学习中另一种常用假设 Li 等人 [2020b], which is

$$\mathbb{E}_{Z_i\sim\mathcal{D}_i(\boldsymbol{\theta})}[\|\nabla\ell(\boldsymbol{\theta};Z_i)\|_2^2]\le G^2. \tag{3}$$

首先，可以证明 (3) 蕴含假设 2.6，因此假设 2.6 弱于 (3)。为说明这一点：当 (3) 成立时，令 $\varsigma^2=4G^2$，则 $\|\nabla f(\boldsymbol{\theta};\boldsymbol{\theta})-\nabla f_i(\boldsymbol{\theta};\boldsymbol{\theta})\|_2^2\le 2\left\|\nabla f(\boldsymbol{\theta};\boldsymbol{\theta})\right\|_2^2+2\|\nabla f_i(\boldsymbol{\theta};\boldsymbol{\theta})\|_2^2\le 4G^2=\varsigma^2$。
我们进一步给出一个具体示例，其中（3）不成立，但假设 2.6 成立。

例 2.7. 假设我们有一个双客户端高斯均值估计问题 $\ell(\theta,Z)=\frac{1}{2}(\theta-Z)^2$ 其中 $\theta,Z\in\mathbb{R}$，$\mathcal{D}_1\ \theta=\mathcal{N}\ \frac{1}{2}\theta\ \sigma^2\quad\mathcal{D}_2\ \theta=\mathcal{N}\ -\frac{1}{2}\theta\ \sigma^2\qquad p_1=p_2=\frac{1}{2}\qquad-\mathbb{E}_{Z_1\sim\mathcal{D}_1\ \theta}\ [\|\nabla\ell\ \theta\ Z_1\ \|_2^2]=\mathbb{E}_{Z_1\sim\mathcal{D}_1\ \theta}\ [\ \theta-(\ )(\ ,),(\ )(\ ,),$ *and* 。那么 $(\ )(\ ;)(\ )(Z_1)^2]=\sigma^2+(\mathbb{E}_{Z_1\sim\mathcal{D}_1(\theta)}[\theta-Z_1])^2=\frac{1}{4}\theta^2+\sigma^2$ 并且 $\mathbb{E}_{Z_2\sim\mathcal{D}_2(\theta)}[\|\nabla\ell(\theta;Z_2)\|_2^2]=\frac{9}{4}\theta^2+\sigma^2$ 当 θ趋于无穷大时，它们都趋于无穷大。因此，式 *(3)* 不成立。另一方面，$\nabla f_1(\theta;\theta)=\frac{1}{8}\theta,\nabla f_2(\theta;\theta)=\frac{9}{8}\theta,$ *and* $\nabla f(\theta;\theta)=\frac{5}{8}\theta,\theta^{PS}=0$，那么通过取 $\varsigma=\frac{1}{2}$，我们可以验证假设 *2.6* 成立。

其次，(3) 也意味着假设 2.5：当 (3) 成立时，令 $\sigma^2=G^2$，即可得到 $\mathbb{E}[\|\nabla l(\boldsymbol{\theta};Z_i)-\nabla f_i(\boldsymbol{\theta},\ \boldsymbol{\theta})\|_2^2]\le\mathbb{E}[\|\nabla l(\boldsymbol{\theta};Z_i)\|_2^2]\le G^2=\sigma^2$。另一方面，假设 2.6 并不意味着假设 2.5。
事实证明，假设 2.6 能更好地刻画系统异质性，正如我们将展示异质性如何影响收敛（更多细节见定理 3.1、3.2 和 3.3）。

2.3 PS 解的性质

定义平均敏感度为 $\epsilon:=\sum_{i=1}^N p_i\epsilon_i$，并定义映射 $\Phi(\boldsymbol{\theta}):=\arg\min_{\boldsymbol{\theta}'\in\mathbb{R}^m} f(\boldsymbol{\theta}',\ \boldsymbol{\theta})$。由此，我们可以证明 PS 解的存在性和唯一性。

命题 2.8 ($\boldsymbol{\theta}^{PS}$的唯一性)。在假设 *2.1*、*2.2* 和 *2.3* 下，若 $\epsilon<\mu/L$，则 $\Phi(\cdot)$ 是一个压缩映射，且具有唯一不动点 $\boldsymbol{\theta}^{PS}=\Phi(\boldsymbol{\theta}^{PS})$；若 $\epsilon\ge\mu/L$，则存在一个实例，使得由 $\Phi(\cdot)$ 生成的任意序列都会发散。

命题 2.8 为 $\boldsymbol{\theta}^{PS}$ 的存在性确立了一个充分必要条件，与 Li 等人 [2022] 的结果类似。该条件仅取决于平均敏感度 ϵ，这意味着即使某些客户端不存在自适应稳定解，整个系统仍可能具有唯一的自适应稳定解 $\boldsymbol{\theta}^{PS}$。以下命题进一步从其与 $\boldsymbol{\theta}^{PO}$ 的距离角度验证 $\boldsymbol{\theta}^{PS}$ 的质量。

命题 2.9 (距离 $\|\boldsymbol{\theta}^{PO}-\boldsymbol{\theta}^{PS}\|_2$界)。在假设 *2.1* 和 *2.3* 下，设损失函数 $\ell(\boldsymbol{\theta};\ Z)$ 关于 Z 是 L_z-*Lipschitz* 的，则对于每个自适应稳定解 $\boldsymbol{\theta}^{PS}$ 和每个表现型最优解 $\boldsymbol{\theta}^{PO}$，有 $\|\boldsymbol{\theta}^{PS}-\boldsymbol{\theta}^{PO}\|_2\le\big(2L_z\epsilon\big)/\mu.$

命题 2.8 和 2.9 的证明见附录 A。

2.4 P-FedAvg算法

在 P-FedAvg 中，客户端每进行 E 次本地更新，就与服务器通信一次。将 $\mathcal{I}_E:=\{nE|n=1,2,\dots\}$ 记为聚合步骤的集合。接下来，我们将形式化所提出的 **P-FedAvg** 的全参与和部分参与方案。

全客户端参与。 所有客户端在每个聚合步骤都与服务器通信，并根据以下内容更新本地模型 $\boldsymbol{\theta}_i^{t+1}$：设 $Z_i^{t+1}\sim\mathcal{D}_i(\boldsymbol{\theta}_i^t)$，则

$$\begin{aligned}\boldsymbol{w}_i^{t+1}&=\boldsymbol{\theta}_i^t-\eta_t\nabla\ell(\boldsymbol{\theta}_i^t;Z_i^{t+1});\\ \boldsymbol{\theta}_i^{t+1}&=\begin{cases}\sum_{j=1}^N p_j\boldsymbol{w}_j^{t+1} & \text{if } t+1\in\mathcal{I}_E\\ \boldsymbol{w}_i^{t+1} & \text{o.w.}\end{cases}\end{aligned}$$

客户端部分参与。 这是一种更现实的设置，无需在每个聚合步骤中都等待所有客户端返回结果。在这种情况下，中心服务器仅在聚合步骤中收集最先响应的 $K<N$ 个客户端的输出。将第 t 步中最先响应的 $K<N$ 个客户端记为大小为 K 的集合 $\mathcal{S}_t:=\{i_1,\dots,i_K\}\in[N]$。令

$Z_i^{t+1} \sim \mathcal{D}_i(\boldsymbol{\theta}_i^t)$，则

$$\boldsymbol{w}_i^{t+1} = \boldsymbol{\theta}_i^t - \eta_t \nabla \ell(\boldsymbol{\theta}_i^t; Z_i^{t+1});$$
$$\boldsymbol{\theta}_i^{t+1} = \begin{cases} \begin{pmatrix} \text{samples } \mathcal{S}_{t+1}, \text{and} \\ \text{average } \left\{\boldsymbol{w}_{t+1}^k\right\}_{k\in\mathcal{S}_{t+1}} \end{pmatrix} & \text{if } t+1 \in \mathcal{I}_E \\ \boldsymbol{w}_i^{t+1} & \text{o.w.} \end{cases}$$

我们进一步考虑两种部分参与方案：

1.（**方案 I**）服务器通过以概率 p_1、$\cdots$、p_N 对索引 $k \in \{1, \cdots, N\}$ 进行有放回的独立同分布抽样，建立 $\mathcal{S}_{t+1}$，抽样 K 次。因此，$\mathcal{S}_{t+1}$ 是一个允许元素重复出现的多重集。随后服务器通过 $\boldsymbol{\theta}_i^{t+1} = \frac{1}{K}\sum_{k\in\mathcal{S}_{t+1}} \boldsymbol{w}_k^{t+1}$ 对参数求平均。该抽样方案最早由 Sahu 等人 [2018]提出，但理论分析最早由 Li 等人 [2020b]完成。 2.（**方案 II**）服务器从 $\mathcal{S}_{t+1}$ 中无放回抽样，且采用均匀抽样。因此，$\mathcal{S}_{t+1}$ 中的每个元素只出现一次。随后服务器通过 $\boldsymbol{\theta}_i^{t+1} = \sum_{k\in\mathcal{S}_{t+1}} p_k \frac{N}{K} \boldsymbol{w}_k^{t+1}$ 对参数求平均。请注意，当概率 $\{p_k\}$ 不相同时，无法保证 $\sum_{k\in\mathcal{S}_{t+1}} p_k \frac{N}{K} = 1$ Li 等人 [2020b]。−

P-FedAvg 的伪代码见附录**??**。

通信成本。 P-FedAvg 每隔 E 次迭代需要进行两轮通信、聚合和广播。因此，在时间步 T，系统完成 $2\lfloor T/E \rfloor$ 次通信。我们遵循 Li 等人[2020b] 的设定：服务器根据选定的方案进行聚合，并将聚合后的参数广播给所有客户端。

接下来，我们将证明，在上述假设下，P-FedAvg 具有 $\mathcal{O}(1/T)$ 收敛速率。作为补充，我们在附录 D 中证明，如果用 (3) 替换假设 2.6，P-FedAvg 也具有 $\mathcal{O}(1/T)$ 收敛速率。

3 收敛性分析

本节将展示，在第 2 节所作的假设下，P-FedAvg 以 $\mathcal{O}(1/T)$ 的速率收敛到唯一的 $\boldsymbol{\theta}^{PS}$；这些假设适用于上述所有方案。关键在于观察到，当学习率足够小且逐渐衰减时，E 步的效果类似于静态情形下使用较大学习率进行的一步更新，正如 Li 等人 [2020a] 在非表演式设定中所述。因此，在采用满足上述假设的适当采样和更新方案时，全局更新的行为类似于 Perdomo 等人 [2020] 中反复进行的表演式随机梯度下降。我们还表明，设备部分参与时，平均参数序列 $\{\boldsymbol{\theta}^t\}$ 的均值与全参与时相同，但方差更大；通过精心选择学习率，可以控制该方差。值得注意的是，客户端异质性在收敛性分析中起着关键作用，下面将对此作进一步阐述。

量化异质性。客户端异质性可以通过共识误差 $\sum_{i=1}^N p_i \|\boldsymbol{\theta}_i^t - \boldsymbol{\theta}^t\|_2^2$ 来量化。由于施为式预测的特性，共识误差是动态变化的。它取决于偏移映射 $\mathcal{D}_i$ 和决策参数。每次广播后，异质性都会导致异质分布偏移，进而造成各不相同的本地更新，最终产生共识误差。假设 2.6 中的 ς 值也是衡量异质性的一个良好指标。

接下来，我们将首先介绍全参与方案的收敛性分析，随后再将分析扩展到部分参与方案。由于表演性场景中的分析较为复杂，为了便于分析并清晰地呈现，我们定义以下常数。

与系统设计无关的常数。
$\epsilon := \sum_{i=1}^N p_i \epsilon_i, \quad \epsilon_{max} := \max_i \epsilon_i, \tilde{\mu} := \mu - (1+\delta)\epsilon L, \quad \delta > 0, c_1 := \left(L(1+\epsilon_{max})^2\right)/(2\delta\epsilon), c_2 := 4\left[\sigma^2 + L^2(1+\epsilon_{max})^2\right], c_3 := 6\left[2\sigma^2 + 3L^2(1+\epsilon_{\max})^2\right], c_4 := 16\sigma^2 + 12\varsigma^2 + (8\sigma^2 + 12\varsigma^2)/\mathbb{E}\|\boldsymbol{\theta}^0 - \boldsymbol{\theta}^{PS}\|_2^2, c_5 := (48\sigma^2 + 36\varsigma^2) \quad - \mathbb{E}\|\boldsymbol{\theta}^0 - \boldsymbol{\theta}^{PS}\|_2^2 + (24\sigma^2 + 36\varsigma^2).$

与系统设计相关的常数（例如 E, K）。
$c_6 := (2E + 3E + 1)\log(E+1), \tilde{\eta}_0 := \tilde{\mu}/\big(2\sigma^2 + (c_1c_3 + c_2/6)c_4c_6\big), \hat{\eta}_0 := \tilde{\mu}/\big(2\sigma^2 + (c_1c_3 + c_2/6)c_4(2E^2 - E)\log E\big),$
$B = 2\sigma^2 + 4c_1\eta_0 + 4c_2\eta_0^2 c_5 2E^2 - E \quad E \quad : (\hat{\ }\hat{\ }) ( ) \log, B_1 := 2\sigma^2 + (4c_1\tilde{\eta}_0 + 4c_2\tilde{\eta}_0^2 + 1/K)c_5c_6, B_2 := 2\sigma^2 + \big(4c_1\tilde{\eta}_0 + 4c_2\tilde{\eta}_0^2 + \frac{N-K}{K\ N-1()}\big)c_5c_6.$

3.1 全参与的收敛

定理 3.1 (全参与)**.** 考虑采用递减步长 $\eta_t = \frac{2}{\tilde{\mu}(t+\gamma)}$ 且全参与的 *P-FedAvg*，其中 $\gamma =$ 最大值 $\left\{\frac{2}{\tilde{\mu}}\hat{\eta}_0, E, \frac{2}{\tilde{\mu}}\sqrt{(4E^2+2E)c_3}\right\}$。在假设 *2.1*、*2.2*、*2.3*、*2.5*、*2.6* 下，以下结论成立

$$\mathbb{E}[\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2] \leq \frac{\upsilon}{\gamma + t}, \ \forall t$$

其中 $\upsilon =$ 最大值 $\left\{\frac{4B}{\tilde{\mu}^2}, \gamma\mathbb{E}\|\boldsymbol{\theta}^0 - \boldsymbol{\theta}^{PS}\|_2^2\right\}$。

定理 3.1 的证明见附录B。证明的关键在于，期望距离 $\mathbb{E}\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{PS}\|_2^2$ 和期望共识误差 $\sum_{i=1}^N p_i\mathbb{E}\|\boldsymbol{\theta}_i^{t-1} - \boldsymbol{\theta}^t\|_2^2$ 都取决于上一步中的期望距离 $\mathbb{E}\|\boldsymbol{\theta}^{t-1} - \boldsymbol{\theta}^{PS}\|_2^2$ 和期望共识误差 $\sum_{i=1}^N p_i\mathbb{E}\|\boldsymbol{\theta}_i^{t-1} - \boldsymbol{\theta}^{t-1}\|_2^2$。虽然我们可以为包含期望共识误差的期望距离建立下降引理，但无法为期望共识误差建立下降引理，因此无法像 Li 等人[2022]那样，为期望距离和期望共识误差建立联合下降引理。幸运的是，共识误差会在每个聚合步骤变为零，这使我们能够在步长足够小时，利用一种新颖的双重迭代技术，将每一步的期望共识误差控制在一个常数范围内。然后，将期望共识误差放宽为该常数，我们便可以在随机梯度下降分析中，为期望距离建立标准下降引理。

3.2 部分参与的收敛性

如第 2 节所述，部分参与方案在联邦学习 Li 等人 [2020b]中更加现实，也更受关注，因为它减轻了掉队者效应。

我们首先给出方案 I 的收敛结果。

定理 3.2 (部分参与，方案 I)。 考虑采用递减步长 $\eta_t = \frac{2}{\tilde{\mu}(t+\gamma)}$ 的部分参与 *P-FedAvg*（方案 *I*），其中 $\gamma =$ 最大 $\left\{\frac{2}{\tilde{\mu}\tilde{\eta}_0}, E, \frac{2}{\tilde{\mu}}\sqrt{(4E^2 + 10E + 6)c_3}\right\}$。在假设 *2.1*、*2.2*、*2.3*、*2.5*、*2.6* 下，以下结论成立

$$\mathbb{E}[\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2] \leq \frac{\upsilon}{\gamma + t}, \ \forall t$$

where $\upsilon = \max\left\{\frac{4B_1}{\tilde{\mu}^2}, \gamma\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2\right\}$.

随后，我们给出方案 II 的收敛结果。如第 2 节所述，我们需要概率 $p_i = \frac{1}{N}, \forall i$ 以确保 $\sum_{i\in\mathcal{S}_t} p_k \frac{N}{K} = 1$。

定理 3.3（部分参与，方案 II）。 考虑采用部分参与（方案 *II*）和递减步长 $\frac{2}{\tilde{\mu}(t+\gamma)}$ 的 *P-FedAvg*，其中 $\gamma = \max\left\{\frac{2}{\tilde{\mu}\tilde{\eta}_0}, E, \frac{2}{\tilde{\mu}}\sqrt{(4E^2 + 10E + 6)c_5}\right\}$。在假设 *2.1*、*2.2*、*2.3*、*2.5*、*2.6* 下，有如下结论

$$\mathbb{E}[\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2] \leq \frac{\upsilon}{\gamma + t}, \ \forall t$$

其中 $\upsilon =$ 最大值 $\left\{\frac{4B_2}{\tilde{\mu}^2}, \gamma\mathbb{E}\|\boldsymbol{\theta}^0 - \boldsymbol{\theta}^{PS}\|_2^2\right\}$。

定理 3.2 和 3.3 的证明见附录 C。除了与定理 3.1 类似的技术难点外，我们还需要界定 $\boldsymbol{\theta}^t$ 在聚合步骤中的方差。幸运的是，它可以由共识误差界定。然后，借鉴定理 3.1 证明中的类似技巧，我们可以在随机梯度下降分析中为期望距离建立一个标准的下降引理。

方案 II 要求 $p_i = \frac{1}{N}$，$\forall i$，这违背了联邦学习的不平衡特性。Li 等人 [2020b] 提出的一种解决方案是将局部目标函数缩放为 $g_i(\boldsymbol{\theta}; \boldsymbol{\theta}) = p_i N f_i(\boldsymbol{\theta}; \boldsymbol{\theta})$，然后将缩放后的局部目标函数取简单平均，作为全局目标函数

$$f(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}}) := \sum_{i=1}^{N} p_i f_i(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}}) = \frac{1}{N} \sum_{i=1}^{N} g_i(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}}).$$

由于缩放目标函数会改变相关性质，因此我们需要谨慎对待第 2 节中的假设。若将 L，μ，σ，ς 替换为 $L' := q_{max} L$，$\mu' := q_{min}\mu$，$\sigma' := \sqrt{q_{max}}\sigma$，$\varsigma' := \sqrt{q_{max}}\varsigma$，其中 $q_{max} := N \cdot \max_i p_i$，$q_{min} := N \cdot \min_i p_i$，收敛定理仍然成立。

3.3 算法与解的讨论

为方便起见，本小节中我们仅讨论聚合步骤，记作 $T \in \mathcal{I}_E$，那么在划分 E 时，我们可以直接使用 $\frac{T}{E}$。请注意，对于一般步骤 t，我们只需使用 $\lfloor \frac{t}{E} \rfloor$ 即可得到一个整数。

对 E的选择。我们关注达到 ϵ 准确度所需的总时间，以及这一总时间如何随E变化。我们利用定理 3.1、3.2 和 3.3 中的结果，并记 $T_\epsilon := \frac{\upsilon}{\epsilon} - \gamma$ 为保证达到 ϵ准确度所需的计算步骤数。为了将 T_ϵ 与所需的总时间联系起来，假设每个通信步骤的期望耗时是每个计算步骤期望耗时的 C 倍，那么达到 ϵ准确度所需的总时间与 $T_\epsilon + C \cdot \frac{T_\epsilon}{E}$ 呈线性关系。下面我们分别分析 E 对 $\frac{T_\epsilon}{E}$ 和 T_ϵ的影响，然后讨论在不同的 C 取值下如何选择最优的 E。

令 $B_0 := B$ 表示全参与条件下的定理 3.1，且 $\gamma_i (i = 0, 1, 2)$ 分别表示定理 3.1、3.2 和 3.3 中的 γ。那么在定理 3.1、3.2 和 3.3 中，T_ϵ 受 $\mathcal{O}\big(4B_i/\tilde{\mu}^2 + \gamma_i \mathbb{E}[\|\boldsymbol{\theta}^0 - \boldsymbol{\theta}^{PS}\|_2^2]\big)$ 支配，其中 $i = 0, 1, 2$。根据定义，我们知道 $B_i (i = 0, 1, 2)$ 相对于 E 几乎是一个常数，而 γ_i 为 $\mathcal{O}(E^2 \log E)$。这意味着，当 E 增长时，达到 ϵ-准确度所需的总更新步骤数 T_ϵ 将增加，而所需的聚合步骤数 $\frac{T_\epsilon}{E}$ 将先增加后减少。

现在我们考虑 $T_\epsilon + C \cdot \frac{T_\epsilon}{E}$，达到 ϵ-准确度所需的总时间。根据上述分析，我们知道其量级为 $\mathcal{O}(E^2 \log E) + C \cdot \mathcal{O}(E \log E) + C \cdot \mathcal{O}(\log E/E)$。当通信速度较快，即 C较小时，$\mathcal{O}(E^2 \log E)$ 是主导项，我们可以更多地关注计算迭代次数 T_ϵ，并且较小的 E 值更可取。然而，当 C较大时，$C \cdot \mathcal{O}(E \log E) + C \cdot \mathcal{O}(\log E/E)$ 就会成为主导项，此时我们应更多地关注通信轮次 $\frac{T_\epsilon}{E}$，并且中等大小的 E 值更可取。

选择 K。同样，T_ϵ 由 $\mathcal{O}\big(4B_i/\tilde{\mu}^2 + \gamma_i \mathbb{E}[\|\boldsymbol{\theta}^0 - \boldsymbol{\theta}^{PS}\|_2^2]\big)$ 主导，其中 $i = 1, 2$。接着，根据$B_i (i = 1, 2)$的公式，我们知道 T_ϵ 随着 K单调递减，但总通信时间会随着 K 增加而增加，这是由于掉队者效应更加严重。一般而言，正如定理 3.2 和 3.3 所示，收敛速率对 K的依赖较弱。我们在图 3(a) 中通过实验证实了这一现象。因此，我们可以将 $\frac{K}{N}$ 设为适当的小值，在保持收敛速率的同时减轻掉队者效应。

采样方案的选择。我们对第 2 节中的两种抽样方案进行形式化，并在定理 3.2 和定理 3.3 中展示它们的收敛性质。值得注意的是，方案 I 具有一项理想特性：它天然支持数据量不均衡的客户端。因此，如果服务器能够控制抽样过程，就应选择方案 I。

但正如 Li 等人 [2020b], 所讨论的，有时服务器可能无法控制采样过程，而只能直接使用最先 K 收到的结果进行更新。在这种情况下，如果来自每个客户端的接收时间是独立同分布随机变量，我们可以将这一过程视为从 N 中随机进行无放回抽样，选取 K 个样本。定理 3.3 展示了收敛性，而关于目标函数缩放的讨论则提供了相关指导，说明如何使系统适用于任意初始 $p_1, \ldots, p_N$ 值。然而，值得注意的是，当 $p_1, \ldots, p_N$ 高度不均匀时，相应的 L', σ', ς' 值将远大于 L, σ, ς中的对应值，而 μ' 则会远小于 μ。因此，根据η_0 和 $\tilde{\eta}_0$的公式，我们必须采用小得多的初始学习率，从而导致收敛速度变慢。然而，如此小的学习率可能导致模型完全无法训练。我们也在图7中通过实验展示了这一现象。

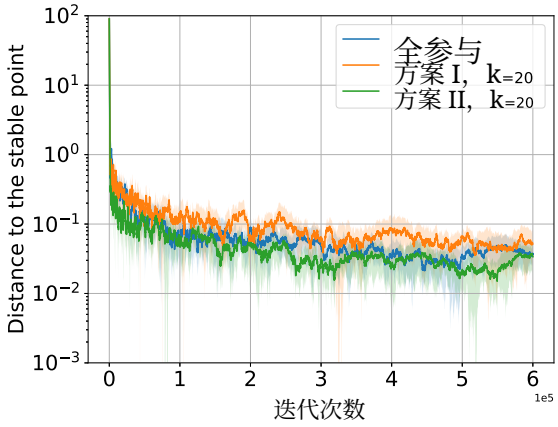


图 1：全参与、方案 I 和方案 II 下，自适应稳定解的距离与迭代次数的关系。

然而，一个有趣的观察是，当 $p_i = \frac{1}{N}$ 时，我们通过实验证明，在图 4 和图 5 中，方案 II 的表现略优于方案 I。

学习率衰减。即使客户端的数据是静态、独立同分布（IID）的，随机梯度下降（SGD）要收敛也必须采用学习率衰减。Li 等人 [2022] 在去中心化表演性预测中使用了这种衰减，并且在具有静态、非独立同分布客户端的 FedAvg 中证明了这种衰减的必要性。我们还通过实证结果表明，恒定学习率无法收敛，见图6。

$\boldsymbol{\theta}^{PS}$和 $\boldsymbol{\theta}^{PO}$。这里我们将更深入地讨论 $\boldsymbol{\theta}^{PS}$与 $\boldsymbol{\theta}^{PO}$解之间的关系。在策略学习设定中，Perdomo 等人 [2020] 证明了 $\boldsymbol{\theta}^{PO}$是 Stackelberg 均衡。值得注意的是，$\boldsymbol{\theta}^{PS}$并非只是对 $\boldsymbol{\theta}^{PO}$的近似，而是最佳响应动态（BRD）的一个自然收敛点。具体来说，当客户端和决策者都不了解其他人的效用时，逆向归纳法无法使用，采用 Stackelberg 均衡并不现实。在这种情况下，将其他人在前一时间步的策略视为常量，并据此优化自身策略，是一种理性的策略。这样的优化步骤就是最佳响应；在多轮序贯策略学习问题中，Zrnic 等人 [2021], 指出，最佳响应可以形成 BRD，而 $\boldsymbol{\theta}^{PS}$ 是**BRD 的收敛点**。尽管决策者的自然最佳响应步骤是风险最小化步骤，基于梯度的 P-FedAvg 也能找到相同的 $\boldsymbol{\theta}^{PS}$。关于$\boldsymbol{\theta}^{PS}$，另一个有趣的观察是，如果去除序贯决策的特性，那么 $\{\boldsymbol{\theta}^{PS}, \mathcal{D}_1(\boldsymbol{\theta}^{PS}) \ldots, \mathcal{D}_N(\boldsymbol{\theta}^{PS})\}$就是一个**纳什均衡**，因为没有参与者有动机单方面偏离。

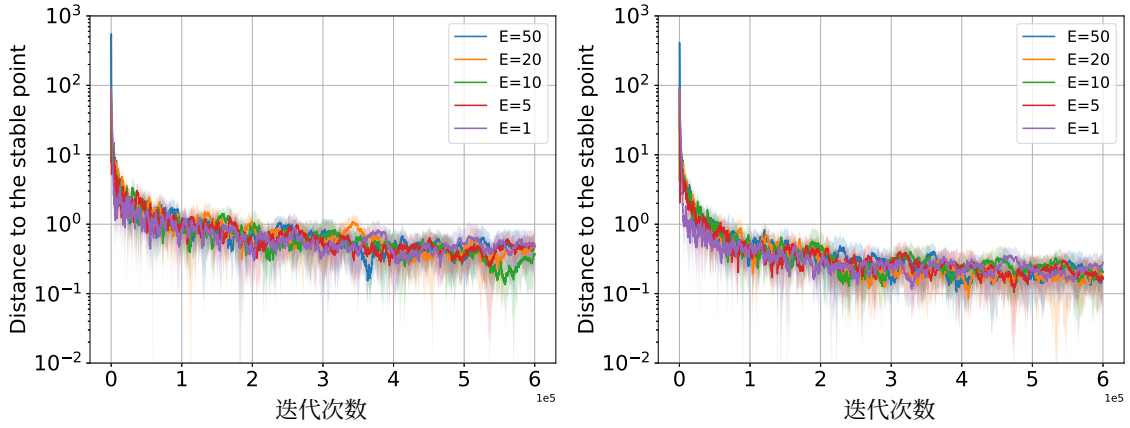
4 数值实验

4.1 加权高斯均值表现性预测

作为数值模拟，我们采用 P-FedAvg，在表现性效应下估计异质高斯数据的均值，并考察超参数、抽样方案和客户端异质性的影响。我们考虑 $N = 25$ 个客户端，其中第 i 个客户端最小化损失函数 $\ell(\theta; Z_i) := (\theta - Z)^2/2, \theta, \quad Z \in \mathbb{R}$，其数据为 $Z_i \sim \mathcal{D}_i(\theta) := \mathcal{N}(m_i + \epsilon_i\theta, \sigma^2)$。对于该损失函数，我们有 $\mu = 1, L = 1$。当 $\epsilon \in [0, 1)$ 时，PS 解为 $\theta^{PS} = \frac{\sum_{i=1}^N p_i m_i}{1-\epsilon}$；而当 $\epsilon \geq 1$ 时，θ^{PS} 不存在。将 m_i 的加权平均记为 $m = \sum_{i=1}^N p_i m_i$，方差记为 $\mathrm{Var}(m) = \sum_{i=1}^N p_i(m_i - m)^2$。在实验中，我们设定 $\epsilon = 0.9, m = 10$。

图 1 显示，在三种通信设置下，P-FedAvg 都会收敛到自适应稳定解：全参与，以及部分参与的两种方案。有趣的是，在本实验中，采用方案 II 的部分参与收敛得最快。尽管全参与方案在保证收敛所需迭代次数的上界方面最低，但我们的实验结果表明，三种方案的实际收敛行为非常相似，且对 K 的依赖较弱，尤其是在 $p_i = \frac{1}{N}$ 时。

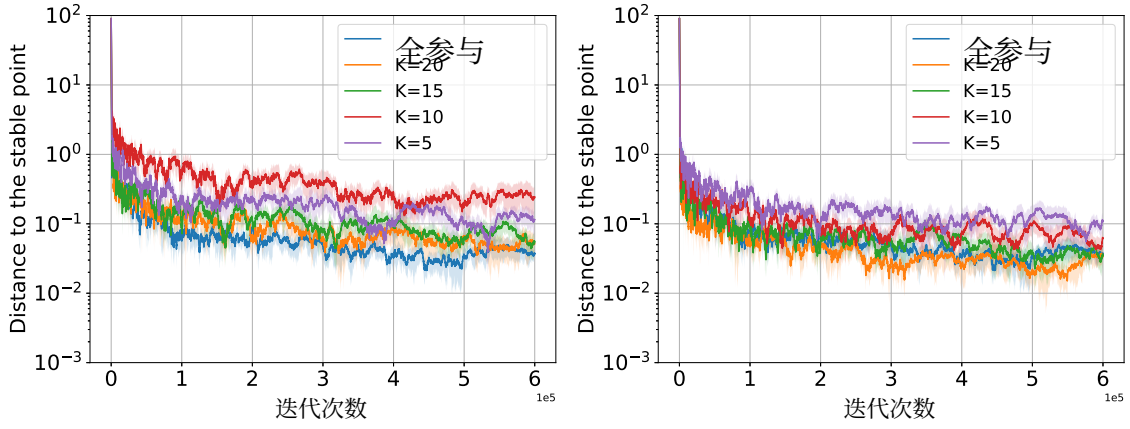
.....的影响 E。我们在同质系统下进行了一项实验，在多种 E 值下比较我们算法的性能。图 2 展示了两种抽样方案下的结果，其中 $K = 20$。稍大的 E 会带来更快的收敛。然而，过大的 E（实验中为 $E = 50$）也可能导致收敛变慢。这是因为在这种情况下，每次聚合时客户端之间的差异过大，从而导致效率低下。在现实场景中，随着通信成本变化，应谨慎选择 E。



(a) 方案 I　　(b) 方案 II

图 2：E 对方案 I 和方案 II 的影响。$K = 20$、Var(m) $= 0.6$、Var(ϵ) $= 0.1$ 适用于 (a) 和 (b)。对于 (b)，$p_i = \frac{1}{25}$。

影响 K。图 3 展示了 FedAvg 在不同 k 值下的收敛情况。对于方案 I，k 越大，收敛越快。而对于方案 II，随着 k 增大，收敛速率将先升高后降低。



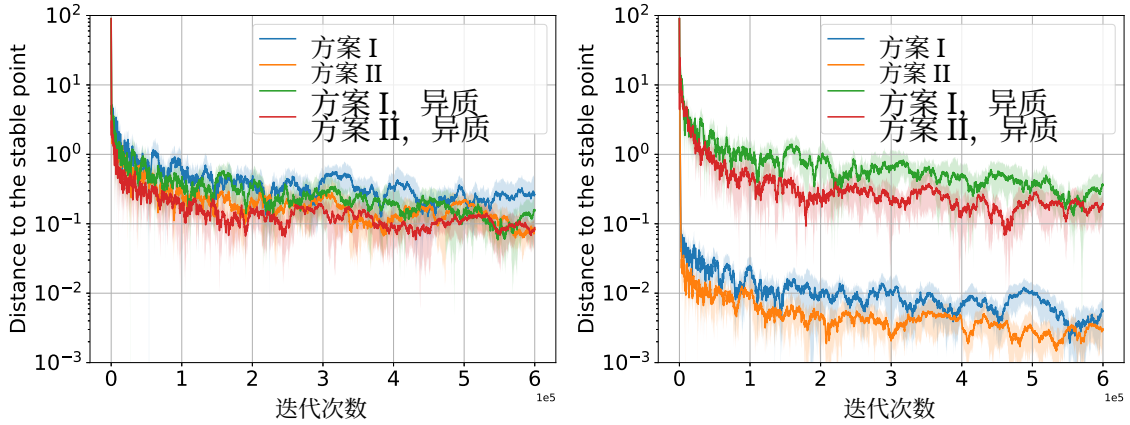
（a）方案 I　　（b）方案 II

图 3：K 对方案 I 和方案 II 的影响。对于 (a) 和 (b)，$E = 5$、Var(m) $= 0.6$、Var(ϵ) $= 0.1$。对于 (b)，$p_i = \frac{1}{25}$。

抽样方案的影响。图 1 还比较了不同的方案。我们可以看到，如果客户端的数据是均匀抽样的（$p_i = \frac{1}{N}$），那么方案 II 可实现更高的收敛速率，这与我们的理论结果一致，因为$B_1 > B_2$。

数据异质性和变化的异质性。在图 4 中，我们在数据异质性条件下测试了算法。具体而言，我们分别将 m 和 ϵ 设置为具有较大的方差。在这个示例中，m 主要反映数据异质性，而 ϵ 反映变化的异质性。该实验表明，我们的算法在一定程度的

异质性。比较我们算法在两幅图中的表现，可以看出，分布异质性的变化是表演式联邦学习中的主要影响因素。



(a) m_i 中的异质性　　(b) ϵ_i中的异质性

图 4：异质性对部分参与的两种方案的影响。 m_i 的影响见 (a)， ϵ_i 的影响见 (b)。(a) 中，异质性情形下 Var(m) $= 6$ ，同质性情形下为 0；Var(ϵ) $= 0.1$。(b) 中，异质性情形下 Var(ϵ) $= 0.6$ ，同质性情形下为 0.1；Var(m) $= 0.6$。

4.2 信用评分策略分类

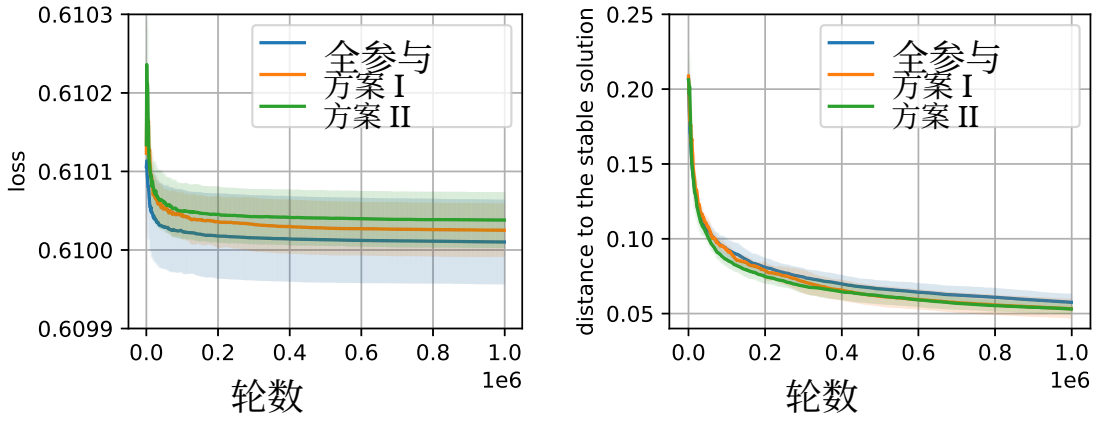
为了展示 P-FedAvg 在真实世界数据集上的表现，我们遵循 Perdomo 等人 [2020] ，使用相同的 Kaggle 数据集*，其中银行会预测贷款申请人是否信用可靠。特征由个人信息构成；如果该个人拖欠了贷款，则目标值为 1，否则为 0。我们采用与 Perdomo 等人 [2020] 相同的策略设置，申请人可以操纵以下特征：(1) 无担保信贷额度循环使用率，(2) 未结清信贷额度和贷款的数量，以及 (3) 房地产贷款或信贷额度的数量。第 i 个群体的操纵强度由 ϵ_i 控制。我们将训练集平均划分为 10 个子集，并分发给 10 个客户端，因此 $p_i = 0.1, \forall i$。10 个客户端的敏感度 ϵ_i 从 $[0.9, 1.1]$ 中独立均匀采样。在部分参与的设置下，我们设定 $K = 5$ 。我们训练一个逻辑回归二分类器。在 P-FedAvg 的每一轮中，我们在大小为 4 的随机小批量上执行 $E = 5$ 步梯度下降。关于批量大小影响的讨论见附录 E.2。

图5展示了随着部署轮数增加，损失函数与 PS 解之间的距离。均值和 1 个标准差误差线由使用不同随机种子进行的 5 次实验得出。与数值模拟类似，三种方案的实际收敛行为非常相似。

5 结论

在本研究中，我们借鉴了施为式预测中的分布偏移映射这一思想，研究数据存在偏移且偏移由决策参数诱导的联邦学习问题。我们对表演式联邦学习问题进行了形式化，并证明存在唯一的自适应稳定解；该解是服务器与客户端之间迭代更新过程的自然均衡。随后，我们形式化了 P-FedAvg算法，并证明全设备参与和设备部分参与方案均以 $\mathcal{O}(1/T)$ 的收敛速率收敛到自适应稳定解。我们还全面讨论了若干关键系统参数对收敛的影响，包括聚合间隔大小、部分参与时的抽样设备数量、抽样方案以及客户端之间的异质性。数值结果验证了我们的理论和讨论，并为表演式联邦学习在现实世界中的应用提供了有价值的见解。

*www.kaggle.com/competitions/GiveMeSomeCredit/data



(a) 训练损失　　(b) 到 $\boldsymbol{\theta}^{PS}$ 的距离

图5：损　　失（a）以及全参与、方案 I 和方案　　II.

参考文献

Mark Braverman 和 Sumegha Garg。随机性与噪声在策略性分类中的作用。发表于第一届负责任计算基础研讨会，2020年。Gavin Brown、Shlomi Hod 和 Iden Kalemaj。有状态世界中的施为式预测。*CoRR*，abs/2011.03885，2020年。网址：https://arxiv.org/abs/2011.03885。Fernando E Casado、Dylan Lema、Marcos F Criado、Roberto Iglesias、Carlos V Regueiro 和 Sen´en Barro。联邦学习与持续学习中的概念漂移检测和适应。多媒体工具与应用，81(3)：3397–3419，2022年。Yatong Chen、Jialu Wang 和 Yang Liu。线性分类中的策略性补救。*arXiv*预印本 *arXiv:2011.00355*，2020年。Yucheng Ding、Chaoyue Niu、Yikai Yan、Zhenzhe Zheng、Fan Wu、Guihai Chen、Shaojie Tang 和 Rongfei Jia。块循环数据上的分布式优化。*arXiv*预印本 *arXiv:2002.07454*，2020年。Jinshuo Dong、Aaron Roth、Zachary Schutzman、Bo Waggoner 和 Zhiwei Steven Wu。基于显示偏好的策略性分类。发表于*2018*年*ACM*经济学与计算大会论文集，第55–70页，2018年。Dmitriy Drusvyatskiy 和 Lin Xiao。决策依赖分布下的随机优化，2020年。网址：https://arxiv.org/abs/2011.11173。Dmitriy Drusvyatskiy 和 Lin Xiao。决策依赖分布下的随机优化。运筹学数学，2022年。Hubert Eichner、Tomer Koren、Brendan McMahan、Nathan Srebro 和 Kunal Talwar。半循环随机梯度下降。发表于国际机器学习大会，第1764–1773页。PMLR，2019年。Yongxin Guo、Tao Lin 和 Xiaoying Tang。迈向随时间演变异构数据上的联邦学习。*arXiv*预印本 *arXiv:2112.13246*，2021年。Farzin Haddadpour、Mohammad Mahdi Kamani、Aryan Mokhtari 和 Mehrdad Mahdavi。带压缩的联邦学习：统一分析与精确保证。发表于国际人工智能与统计大会，第2350–2358页。PMLR，2021年。Nika Haghtalab、Nicole Immorlica、Brendan Lucier 和 Jack Wang。利用激励感知评估机制最大化社会福利。第160–166页，2020年7月。doi：10.24963/ijcai.2020/23。Moritz Hardt、Nimrod Megiddo、Christos Papadimitriou 和 Mary Wootters。策略性分类。第111–122页，2016年1月。doi：10.1145/2840728.2840730。

Seyyedali Hosseinalipour、Su Wang、Nicolo Michelusi、Vaneet Aggarwal、Christopher G Brinton、David J Love 和 Mung Chiang。异构无线网络上的动态分布式模型训练的并行连续学习。*arXiv*预印本 *arXiv:2202.02947*，2022。Lily Hu、Nicole Immorlica 和 Jennifer Vaughan。策略性操纵的差异化影响。第 259–268 页，2019 年 1 月。doi: 10.1145/3287560.3287597。Zachary Izzo、Lexing Ying 和 James Zou。数据会对你的模型作出反应时，如何进行学习：表演式梯度下降。载于 Marina Meila 和 Tong Zhang 编，第 *38* 届国际机器学习大会论文集，第 139 卷《机器学习研究会议录》，第 4641–4650 页。PMLR，2021 年 7 月 18–24 日。URL https://proceedings.mlr.press/v139/izzo21a.html。Zachary Izzo、James Zou 和 Lexing Ying。数据逐渐对你的模型作出反应时，如何进行学习。载于 Gustau Camps-Valls、Francisco J. R. Ruiz 和 Isabel Valera 编，第 *25* 届国际人工智能与统计大会论文集，第 151 卷《机器学习研究会议录》，第 3998–4035 页。PMLR，2022 年 3 月 28–30 日。URL https://proceedings.mlr.press/v151/izzo22a.html。Kun Jin、Xueru Zhang、Mohammad Mahdi Khalili、Parinaz Naghizadeh 和 Mingyan Liu。策略分类与回归问题的激励机制。载于 David M. Pennock、Ilya Segal 和 Sven Seuken 编，*EC '22*：第 *23* 届 *ACM*经济学与计算大会，美国科罗拉多州博尔德，*2022* 年 *7* 月 *11–15* 日，第 760–790 页。ACM，2022。doi: 10.1145/3490486.3538300。URL https://doi.org/10.1145/3490486.3538300。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。载于国际机器学习大会，第 5132–5143 页。PMLR，2020。Jon Kleinberg 和 Manish Raghavan。分类器如何促使智能体策略性地投入努力？*ACM*经济学与计算汇刊，8：1–23，2020 年 11 月。doi: 10.1145/3417742。Daliang Li 和 Junpu Wang。FedMD：通过模型蒸馏实现异构联邦学习。*arXiv*预印本 *arXiv:1910.03581*，2019。Qiang Li 和 Hoi-To Wai。具有状态依赖性的施为式预测与随机逼近。载于 Gustau Camps- Valls、Francisco J. R. Ruiz 和 Isabel Valera 编，第 *25* 届国际人工智能与统计大会论文集，第 151 卷《机器学习研究会议录》，第 3164–3186 页。PMLR，2022 年 3 月 28–30 日。URL https://proceedings.mlr.press/v151/li22c.html。Qiang Li、Chung-Yiu Yau 和 Hoi-To Wai。采用贪婪部署和共识寻求智能体的多智能体表演式预测，2022。URL https://arxiv.org/abs/2209.03811。Tian Li、Anit Kumar Sahu、Manzil Zaheer、Maziar Sanjabi、Ameet Talwalkar 和 Virginia Smith。异构网络中的联邦优化。机器学习与系统论文集，2：429–450，2020a。Xiang Li、Kaixuan Huang、Wenhao Yang、Shusen Wang 和 Zhihua Zhang。FedAvg 在非独立同分布数据上的收敛性。载于国际学习表征会议，2020b。URL https://openreview.net/ forum?id=HJxNAnVtDS。Tao Lin、Lingjing Kong、Sebastian U Stich 和 Martin Jaggi。联邦学习中的鲁棒模型融合集成蒸馏。神经信息处理系统进展，33：2351–2363，2020。Celestine Mendler-D¨unner、Juan Perdomo、Tijana Zrnic 和 Moritz Hardt。用于施为式预测的随机优化。载于 H. Larochelle、M. Ranzato、R. Hadsell、M.F. Balcan 和 H. Lin 编，神经信息处理系统进展，第 33 卷，第 4929–4939 页。Curran Associates, Inc.，2020。URL https://proceedings.neurips.cc/paper/2020/file/33e75ff09dd601bbe69f351039152189-Paper.pdf。John Miller、Smitha Milli 和 Moritz Hardt。策略性分类实质上是一种因果建模。载于 Hal Daum´e III 和 Aarti Singh 编，第 *37* 届国际机器学习大会论文集，第 119 卷《机器学习研究会议录》，第 6917–6926 页。PMLR，2020 年 7 月 13–18 日。John P Miller、Juan C Perdomo 和 Tijana Zrnic。走出回音室：优化表演式风险。载于 Marina Meila 和 Tong Zhang 编，第 *38* 届国际机器学习大会论文集，第 139 卷《机器学习研究会议录》，第 7710–7720 页。PMLR，2021 年 7 月 18–24 日。URL https://proceedings.mlr.press/v139/miller21a.html。

Smitha Milli、John Miller、Anca Dragan 和 Moritz Hardt。策略分类的社会成本。第 230–239 页，2019 年 1 月。doi: 10.1145/3287560.3287576。Adhyyan Narang、Evan Faulkner、Dmitriy Drusvyatskiy、Maryam Fazel 和 Lillian J. Ratliff。多人施为预测：在决策依赖博弈中学习。*CoRR*，abs/2201.03398，2022。URL https://arxiv.org/ abs/2201.03398。Juan Perdomo、Tijana Zrnic、Celestine Mendler-D¨unner 和 Moritz Hardt。施为式预测。载于 Hal Daum´e III 和 Aarti Singh 编，第 *37* 届国际机器学习大会论文集，第 119 卷《机器学习研究会议录》，第 7599–7609 页。PMLR，2020 年 7 月 13–18 日。URL https: //proceedings.mlr.press/v119/perdomo20a.html。Reilly Raab 和 Yang Liu。非预期选择：持续存在的资格率差异与干预措施。神经信息处理系统进展，34：26053–26065，2021。Mitas Ray、Lillian J. Ratliff、Dmitriy Drusvyatskiy 和 Maryam Fazel。几何衰减动态环境中的决策依赖风险最小化。载于第三十六届 *AAAI*人工智能会议、*AAAI 2022*，第三十四届人工智能创新应用会议、*IAAI 2022*，第十二届人工智能教育进展研讨会、*EAAI 2022*线上活动，*2022* 年 *2* 月 *22* 日至 *3* 月 *1* 日，第 8081–8088 页。AAAI Press，2022。URL https://ojs.aaai.org/index.php/AAAI/article/view/20780。Elsa Rizk、Stefan Vlaski 和 Ali H. Sayed。动态联邦学习。*CoRR*，abs/2002.08782，2020。URL https://arxiv.org/abs/2002.08782。Anit Kumar Sahu、Tian Li、Maziar Sanjabi、Manzil Zaheer、Ameet Talwalkar 和 Virginia Smith。异构网络中联邦优化的收敛性。*CoRR*，abs/1812.06127，2018。URL http://arxiv.org/ abs/1812.06127。Yonadav Shavit、Benjamin Edelman 和 Brian Axelrod。因果策略线性回归，2020。Jianyu Wang、Qinghua Liu、Hao Liang、Gauri Joshi 和 H Vincent Poor。解决异构联邦优化中的目标不一致问题。神经信息处理系统进展，33：7611–7623，2020。Killian Wood、Gianluca Bianchin 和 Emiliano Dall'Anese。用于决策依赖分布随机优化的在线投影梯度下降。*IEEE*控制系统快报，6：1646–1651，2022。doi: 10.1109/LCSYS.2021.3124187。Xueru Zhang、Mohammad Mahdi Khalili、Kun Jin、Parinaz Naghizadeh 和 Mingyan Liu。公平性干预作为策略性操纵的（反）激励。载于 Kamalika Chaudhuri、Stefanie Jegelka、Le Song、Csaba Szepesvari、Gang Niu 和 Sivan Sabato 编，第 *39* 届国际机器学习大会论文集，第 162 卷《机器学习研究会议录》，第 26239–26264 页。PMLR，2022 年 7 月 17–23 日。URL https://proceedings.mlr.press/v162/zhang22l.html。Chen Zhu、Zheng Xu、Mingqing Chen、Jakub Koneˇcn`y、Andrew Hard 和 Tom Goldstein。昼行还是夜行？在周期性变化的分布中进行多分支网络联邦学习。载于国际学习表征会议，2021a。Zhuangdi Zhu、Junyuan Hong 和 Jiayu Zhou。异构联邦学习的无数据知识蒸馏。载于国际机器学习大会，第 12878–12889 页。PMLR，2021b。Tijana Zrnic、Eric Mazumdar、Shankar Sastry 和 Michael Jordan。策略性分类中谁主导、谁跟随？载于 M. Ranzato、A. Beygelzimer、Y. Dauphin、P.S. Liang 和 J. Wortman Vaughan 编，神经信息处理系统进展，第 34 卷，第 15257–15269 页。Curran Associates, Inc.，2021。URL https://proceedings.neurips.cc/paper/2021/file/ 812214fb8e7066bfa6e32c626c2c688b-Paper.pdf。

命题 2.8 和 2.9 的证明

命题 2.5.（$\boldsymbol{\theta}^{PS}$ 的唯一性）在假设 2.1、2.2 和 2.3 下，定义映射 Φ： $\mathbb{R}^m \mapsto \mathbb{R}^m$

$$\Phi(\boldsymbol{\theta}) := arg \min_{\boldsymbol{\theta}' \in \mathbb{R}^M} f(\boldsymbol{\theta}', \boldsymbol{\theta})$$

如果 $\epsilon := \sum_{i=1}^N p_i \epsilon_i < \mu/L$，那么 $\Phi(\cdot)$ 是一个压缩映射，且具有唯一不动点 $\boldsymbol{\theta}^{PS} = \Phi(\boldsymbol{\theta}^{PS})$。相反，如果 $\epsilon \geq \mu/L$，则存在一个实例，使得由 $\Phi(\cdot)$生成的任意序列都会发散。

证明。 该证明模拟了 Li 等人 [2022]中命题 1 的证明。

固定 $\boldsymbol{\theta}'$， $\boldsymbol{\theta} \in \mathbb{R}^m$，最优性条件意味着

$$\sum_{i=1}^N p_i \nabla f_i(\Phi(\boldsymbol{\theta}); \boldsymbol{\theta}) = \mathbf{0}, \ \sum_{i=1}^N p_i \nabla f_i(\Phi(\boldsymbol{\theta}'); \boldsymbol{\theta}') = \mathbf{0}$$

其中梯度是相对于 f_i的第一个参数而言的。因此我们有

$$\begin{aligned} 0 =& \langle \mathbf{0}, \Phi(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta}') \rangle \\ =& \langle \sum_{i=1}^N p_i \big(\nabla f_i(\Phi(\boldsymbol{\theta}); \boldsymbol{\theta}) - \nabla f_i(\Phi(\boldsymbol{\theta}'); \boldsymbol{\theta}') \big), \Phi(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta}') \rangle. \end{aligned}$$

重新整理上述方程，并在等式两边加上 $\sum_{i=1}^N p_i f_i(\Phi(\boldsymbol{\theta}) \ \boldsymbol{\theta}')$，得到

$$\begin{aligned} & \langle \sum_{i=1}^N p_i \big(\nabla f_i(\Phi(\boldsymbol{\theta}); \boldsymbol{\theta}') - \nabla f_i(\Phi(\boldsymbol{\theta}); \boldsymbol{\theta}) \big), \Phi(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta}') \rangle \\ =& \langle \sum_{i=1}^N p_i \big(\nabla f_i(\Phi(\boldsymbol{\theta}); \boldsymbol{\theta}') - \nabla f_i(\Phi(\boldsymbol{\theta}'); \boldsymbol{\theta}') \big), \Phi(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta}') \rangle. \end{aligned}$$

根据假设 2.1 中的强凸性，我们有

$$f(\Phi(\boldsymbol{\theta}); \boldsymbol{\theta}') \geq f(\Phi(\boldsymbol{\theta}'); \boldsymbol{\theta}') + \langle \nabla f(\Phi(\boldsymbol{\theta}'); \boldsymbol{\theta}'), \Phi(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta}') \rangle + \frac{\mu}{2} \|\Phi(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta}')\|_2^2,$$

$$f(\Phi(\boldsymbol{\theta})'; \boldsymbol{\theta}') \geq f(\Phi(\boldsymbol{\theta}); \boldsymbol{\theta}') + \langle \nabla f(\Phi(\boldsymbol{\theta}); \boldsymbol{\theta}'), \Phi(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta}') \rangle + \frac{\mu}{2} \|\Phi(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta}')\|_2^2,$$

从而

$$\langle \nabla f(\Phi(\boldsymbol{\theta}); \boldsymbol{\theta}') - \nabla f(\Phi(\boldsymbol{\theta}'); \boldsymbol{\theta}'), \Phi(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta}') \rangle \geq \mu \|\Phi(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta}')\|_2^2. \tag{4}$$

应用引理 2.4，我们得到

$$\sum_{i=1}^N p_i \langle \nabla f_i(\Phi(\boldsymbol{\theta}); \boldsymbol{\theta}') - \nabla f_i(\Phi(\boldsymbol{\theta}'); \boldsymbol{\theta}'), \Phi(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta}')) \rangle \leq \sum_{i=1}^N p_i L \epsilon_i \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2 \cdot \|\Phi(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta})\|_2. \tag{5}$$

结合（4）和（5），我们得到

$$\|\Phi(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta}')\|_2 \leq \frac{\sum_{i=1}^N p_i \epsilon_i L}{\mu} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2 = \frac{\bar{\epsilon} L}{\mu} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2. \tag{6}$$

因此，如果 $\bar{\epsilon} < \frac{L}{\mu}$, Φ 根据巴拿赫不动点定理，$(\cdot)$ 是一个压缩映射，并且存在唯一不动点 $\boldsymbol{\theta}^{PS}$.

为了展示当 $\epsilon \geq {}^{L}$ 时的发散$^{-}$ $_{\overline{\mu}}$，我们考虑以下示例，其中 $\theta \in \mathbb{R}$， ^{L} $_{\overline{\mu}} = 1$, $\neg\gamma : = \sum_{i=1}^N p_i \gamma_i \neq 0$, and

$$\ell(\theta; Z) = \frac{1}{2}(\theta - Z)^2, Z \sim \mathcal{D}_i(\theta) = \mathcal{N}(\gamma_i + \epsilon_i \theta, 1)$$

我们观察到

$$\begin{aligned} f_i(\theta';\theta) =& \mathbb{E}_{Z\sim\mathcal{D}_i(\theta)}[\frac{1}{2}(\theta - Z)^2] \\ =& \mathbb{E}_{\tilde{Z}\sim\mathcal{N}(0,1)}[\frac{1}{2}(\theta' - \gamma_i - \epsilon_i\theta - \tilde{Z})^2] \\ =& \frac{1}{2}(\theta' - \gamma_i - \epsilon_i\theta)^2 + \frac{1}{2}, \end{aligned}$$

$$\Phi(\theta) = \operatorname{argmin}_{\theta\in\mathbb{R}} \sum_{i=1}^{N} p_i(\theta' - \gamma_i - \epsilon_i\theta)^2 = \bar{\epsilon}\theta + \bar{\gamma},$$

因此，应用 $\Phi(\cdot)$ t 次后，我们得到

$$\Phi^t(\theta) = \bar{\epsilon}^t\theta + (1 + \bar{\epsilon} + \cdots + \bar{\epsilon}^{(t-1)})\bar{\gamma},$$

并且由于 $\bar{\epsilon} \geq \frac{L}{\mu} = 1, \bar{\gamma} \neq 0$，我们有 $\lim_{t\to\infty} \|\Phi^t(\theta)\|_2 = \infty$。 □

命题 2.6. 在假设 2.1 和 2.3 下，假设损失 $\ell(\boldsymbol{\theta}; Z)$ 关于 Z 是 L_z-Lipschitz 连续的，令 $\epsilon := \sum_{i=1}^{N} p_i\epsilon_i$，则对于每个自适应稳定解和每个表现型最优解 $\boldsymbol{\theta}^{PO}$，我们有

$$\|\boldsymbol{\theta}^{PS} - \boldsymbol{\theta}^{PO}\|_2 \leq \frac{2L_z\bar{\epsilon}}{\mu}.$$

证明。 此证明仿照 Perdomo 等人 [2020] 中定理 4.3 的证明。
首先，由 $\boldsymbol{\theta}^{PO}$ 的最优性可知，$f(\boldsymbol{\theta}^{PO}; \boldsymbol{\theta}^{PO}) \leq f(\boldsymbol{\theta}^{PS}; \boldsymbol{\theta}^{PS})$。根据假设 2.1 中的强凸性，可得

$$f(\boldsymbol{\theta}^{PO};\boldsymbol{\theta}^{PS}) \geq f(\boldsymbol{\theta}^{PS};\boldsymbol{\theta}^{PS}) + \langle \nabla f(\boldsymbol{\theta}^{PS};\boldsymbol{\theta}^{PS}),\, \boldsymbol{\theta}^{PO} - \boldsymbol{\theta}^{PS}\rangle + \frac{\mu}{2}\|\boldsymbol{\theta}^{PO} - \boldsymbol{\theta}^{PS}\|_2^2 \geq \frac{\mu}{2}\|\boldsymbol{\theta}^{PO} - \boldsymbol{\theta}^{PS}\|_2^2.$$

进一步根据假设 2.3，损失函数 $\ell(\boldsymbol{\theta}; Z)$ 关于 Z 是 L_z-Lipschitz 连续的，并利用 Kantorovich-Rubinstein 对偶性，可得

$$\begin{aligned} & f(\boldsymbol{\theta}^{PO};\boldsymbol{\theta}^{PS}) - f(\boldsymbol{\theta}^{PO};\boldsymbol{\theta}^{PO}) \\ =& \sum_{i=1}^{N} p_i\Big(\mathbb{E}_{Z_i\sim\mathcal{D}_i(\boldsymbol{\theta}^{PS})}[\ell(\boldsymbol{\theta}^{PO}; Z_i)] - \mathbb{E}_{Z_i\sim\mathcal{D}_i(\boldsymbol{\theta}^{PO})}[\ell(\boldsymbol{\theta}^{PO}; Z_i)]\Big) \\ \leq& \sum_{i=1}^{N} p_i L_z \mathcal{W}_1(\mathcal{D}_i(\boldsymbol{\theta}^{PS}), \mathcal{D}_i(\boldsymbol{\theta}^{PO})) \\ =& \sum_{i=1}^{N} p_i L_z\epsilon_i\|\boldsymbol{\theta}^{PO} - \boldsymbol{\theta}^{PS}\|_2 = L_z\bar{\epsilon}\|\boldsymbol{\theta}^{PO} - \boldsymbol{\theta}^{PS}\|_2. \end{aligned} \tag{7}$$

其中，不等式是最优传输理论中的一个众所周知的结论。由方程 7，$L_z\epsilon\epsilon_i\|\boldsymbol{\theta}^{PO} - \boldsymbol{\theta}^{PS}\|_2 \geq f(\boldsymbol{\theta}^{PO}, \boldsymbol{\theta}^{PS}) - f(\boldsymbol{\theta}^{PO}, \boldsymbol{\theta}^{PO}) \geq f(\boldsymbol{\theta}^{PO}, \boldsymbol{\theta}^{PS}) - f(\boldsymbol{\theta}^{PS}, \boldsymbol{\theta}^{PS}) \geq \frac{\mu}{2}\|\boldsymbol{\theta}^{PO} - \boldsymbol{\theta}^{PS}\|_2^2$ 可得 $\|\boldsymbol{\theta}^{PO} - \boldsymbol{\theta}^{PS}\|_2 \leq \frac{2L_z\epsilon\epsilon_i}{\mu}$，这意味着 。

□

B 定理 3.1 的证明

B.1 补充符号说明

在我们的分析中，为方便起见，我们将按照 Li 等人 [2020b] 的做法，定义另外两个序列为 $\boldsymbol{w}^t := \sum_{i=1}^{N} p_i\boldsymbol{w}_i^t$ 和 $\boldsymbol{\theta}^t := \sum_{i=1}^{N} p_i\boldsymbol{\theta}_i^t$。注意，$\boldsymbol{w}^t$ 源自对 $\boldsymbol{\theta}^t$ 执行一步随机梯度下降。当 $t+1 \notin \mathcal{I}_E$ 时，$\boldsymbol{w}^t$ 和 $\boldsymbol{\theta}^t$ 均无法访问。当 $t+1 \in \mathcal{I}_E$ 时，我们可以得到 $\boldsymbol{\theta}^t$。此外，我们还定义 $\boldsymbol{g}_t := \sum_{i=1}^{N} p_i\nabla f_i(\boldsymbol{\theta}_i^t;\boldsymbol{\theta}_i^t)$ $\boldsymbol{g}_t := \sum_{i=1}^{N} p_i\nabla\ell(\boldsymbol{\theta}_i^t; Z_i^{t+1})$，其中 $Z_i^{t+1} \sim \mathcal{D}_i(\boldsymbol{\theta}_i^t)$。显然，在全参与的情况下，$\boldsymbol{w}^{t+1} = \boldsymbol{\theta}^t - \eta_t\boldsymbol{g}_t$ 和 $\mathbb{E}\boldsymbol{g}_t = \boldsymbol{g}_t$。显然，对于任意 t，都有 $\boldsymbol{\theta}^t = \boldsymbol{w}^t$。

B.2 关键引理

为清晰起见，我们将给出若干引理，以建立我们的主要定理。具体而言，我们将给出关于 $\mathbb{E}[\|\boldsymbol{w}^t-\boldsymbol{\theta}^{PS}\|_2^2]$ 的下降引理以及 $\sum_{i=1}^N p_i\mathbb{E}\|\boldsymbol{\theta}_i^t-\boldsymbol{\theta}^t\|_2^2$ 的上界；二者共同给出了随机梯度下降分析中 $\mathbb{E}[\|\boldsymbol{\theta}^{t+1}-\boldsymbol{\theta}^{PS}\|_2^2]$ 的标准下降引理，并导出 $\mathcal{O}(\frac{1}{t})$ 收敛性。引理的证明将留待后文。

在下述引理中，我们旨在为 $\mathbb{E}[\|\boldsymbol{\theta}^{t+1}-\boldsymbol{\theta}^{PS}\|_2^2]$ 建立一个上界。由于在全参与情况下 $\boldsymbol{\theta}^{t+1}=\boldsymbol{w}^{t+1}$，这等价于为 $\mathbb{E}[\|\boldsymbol{w}^{t+1}-\boldsymbol{\theta}^{PS}\|_2^2]$ 建立一个上界。

引理 B.1。（下降引理）在假设 *2.1*、*2.2*、*2.3*、*2.5* 下，且在全参与情况下

$$\mathbb{E}[\|\overline{\boldsymbol{\theta}}^{t+1}-\boldsymbol{\theta}^{PS}\|_2^2]=\mathbb{E}[\|\overline{\boldsymbol{w}}^{t+1}-\boldsymbol{\theta}^{PS}\|_2^2]\leq(1-\tilde{\mu}\eta_t)\mathbb{E}\|\overline{\boldsymbol{\theta}}^t-\boldsymbol{\theta}^{PS}\|_2^2+2\sigma^2\eta_t^2+(c_1\eta_t+c_2\eta_t^2)\sum_{i=1}^N p_i\mathbb{E}\|\boldsymbol{\theta}_i^t-\overline{\boldsymbol{\theta}}^t\|_2^2$$

对于任意 t，其中 $\epsilon_{max}:=\max_i\epsilon_i$，$\epsilon:=\sum_{i=1}^N p_i\epsilon_i$，$c_1:=\frac{L(1+\epsilon_{max})2\delta\epsilon}{}^2$，$c_2:=4[\sigma^2+L^2(1+\epsilon_{max})^2]$，$\tilde{\mu}:=\mu-(1+\delta)\epsilon L$。

现在我们将为 $\sum_{i=1}^N p_i\mathbb{E}\|\boldsymbol{\theta}_i^t-\boldsymbol{\theta}^t\|_2^2$ 建立一个上界。注意，如果 $t\in\mathcal{I}_E$，即同步步骤，则对于任意 $i\in[N]$，我们都有 $\boldsymbol{\theta}_i^t\equiv\boldsymbol{\theta}^t$，这意味着 $\sum_{i=1}^N p_i\mathbb{E}\|\boldsymbol{\theta}_i^t-\boldsymbol{\theta}^t\|_2^2=0$。如果 $t\notin\mathcal{I}_E$，则下面的引理给出了 $\sum_{i=1}^N p_i\mathbb{E}\|\boldsymbol{\theta}_i^t-\boldsymbol{\theta}^t\|_2^2$ 的一个上界。

引理 B.2.（共识误差）在假设 *2.1*、*2.2*、*2.3*、*2.5*、*2.6* 下，若 $\{\eta_t\}$ 单调不增，$\eta_t\leq 2\eta_{t+E}$、$t\notin\mathcal{I}_E$、$\eta_t^2\leq 1/\big(2c_3(t+1-t_0)(1+2(t+1-t_0))\big)$，且

$$\eta_0\leq\hat{\eta}_0:=\frac{\tilde{\mu}\mathbb{E}\|\overline{\boldsymbol{\theta}}^0-\boldsymbol{\theta}^{PS}\|_2^2}{2\sigma^2+(c_1c_3+c_2/6)(2E^2-E)\log E\big((16\sigma^2+12\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^0-\boldsymbol{\theta}^{PS}\|_2^2+(8\sigma^2+12\varsigma^2)\big)},$$

那么，在全参与的情况下，我们有

$$\sum_{i=1}^N p_i\mathbb{E}\|\boldsymbol{\theta}_i^t-\overline{\boldsymbol{\theta}}^t\|_2^2\leq 4\eta_t^2(2E^2-E)\log E(48\sigma^2+36\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^0-\boldsymbol{\theta}^{PS}\|_2^2+4\eta_t^2(2E^2-E)\log E(24\sigma^2+36\varsigma^2).$$

其中，对于任意 t，$\mu-(1+\delta)\epsilon L$ $c_3=12\sigma^2+18L^2(1+\epsilon_{\max})^2$ $\epsilon_{max}=\max_i\epsilon_i$ ϵ c_1 c_2 $=$：最大值、：$=\sum_{i=1}^N p_i\epsilon_i$、：$=\frac{L(1+\epsilon_{max})2\delta\epsilon}{}^2$、：$=4[\sigma^2+L^2(1+\epsilon_{max})^2]$，$\tilde{\mu}$：（）、：（）。

（需要注意的是，$4\eta_t^2$、$(48\sigma^2+36\varsigma^2)$ 和 $(24\sigma^2+36\varsigma^2)$ 来自多次应用 $\eta_{t-1}\leq 2\eta_t$，通过谨慎选择步长，实际常数可能会小得多。）

下面的引理在建立定理 3.1 中 $\mathcal{O}(\frac{1}{T})$ 的收敛性时，在技术条件下为随机梯度下降分析提供了一个标准下降引理。

引理 B.3. 在假设 *2.1*、*2.2*、*2.3*、*2.5*、*2.6* 下，若 $\{\eta_t\}$ 单调不增，$\eta_t\leq 2\eta_{t+E}$, $\eta_t^2\leq 1/\big(2c_3(t+1-t_0)(1+2(t+1-t_0))\big)$，且

$$\eta_0\leq\hat{\eta}_0:=\frac{\tilde{\mu}\mathbb{E}\|\overline{\boldsymbol{\theta}}^0-\boldsymbol{\theta}^{PS}\|_2^2}{2\sigma^2+(c_1c_3+c_2/6)(2E^2-E)\log E\big((16\sigma^2+12\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^0-\boldsymbol{\theta}^{PS}\|_2^2+(8\sigma^2+12\varsigma^2)\big)},$$

则在全参与的情况下，我们有

$$\mathbb{E}[\|\overline{\boldsymbol{\theta}}^{t+1}-\boldsymbol{\theta}^{PS}\|_2^2]\leq(1-\tilde{\mu}\eta_t)\mathbb{E}\|\overline{\boldsymbol{\theta}}^t-\boldsymbol{\theta}^{PS}\|_2^2+B\eta_t^2$$

其中，对于任意 t，其中 $\epsilon_{max}:=\max_i\epsilon_i$, $\epsilon:=\sum_{i=1}^N p_i\epsilon_i$, $c_1:=\frac{L(1+\epsilon_{max})2\delta\epsilon}{}^2$, $c_2:=4[\sigma^2+L^2(1+\epsilon_{max})^2]$, $\tilde{\mu}:=\mu-(1+\delta)\epsilon L$, $c_3:=12\sigma^2+18L^2(1+\epsilon_{\max})^2$, 以及 $B:=2\sigma^2+(4c_1\hat{}\eta_0+4c_2\hat{}\eta_0^2)(2E^2-E)\log E\big((48\sigma^2+36\varsigma^2)\mathbb{E}\|\boldsymbol{\theta}^0-\boldsymbol{\theta}^{PS}\|_2^2+(24\sigma^2+36\varsigma^2)\big)$。

B.3 定理3.1的证明补充

我们在此重述所有常数的定义：

与系统设计无关的常数。

$\epsilon_{max} :=$ 最大值$_i\ \epsilon_i$,

$\bar{\epsilon} := \sum_{i=1}^{N} p_i \epsilon_i$,

$\tilde{\mu} := \mu - (1+\delta)\bar{\epsilon}L$,

$c_1 := \left(L(1+\epsilon_{max})^2\right)/(2\delta\epsilon)$,

$c_2 := 4\left[\sigma^2 + L^2(1+\epsilon_{max})^2\right]$,

$c_3 := 6\left[2\sigma^2 + 3L^2(1+\epsilon_{\max})^2\right]$,

$c_4 := 16\sigma^2 + 12\varsigma^2 + (8\sigma^2 + 12\varsigma^2)/\mathbb{E}\|\boldsymbol{\theta}^0 - \boldsymbol{\theta}^{PS}\|_2^2$,

$c_5 := (48\sigma^2 + 36\varsigma^2)\mathbb{E}\|\boldsymbol{\theta}^0 - \boldsymbol{\theta}^{PS}\|_2^2 + (24\sigma^2 + 36\varsigma^2)$

与系统设计相关的常数 y计　　gn (e.g., E, K).

$\hat{}\eta_0 := \tilde{\mu}/\left(2\sigma^2 + (c_1c_3 + c_2/6)c_4(2E^2 - E)\right.$ 对数 $\left.E\right)$,

$B := 2\sigma^2 + (4c_1\hat{}\eta_0 + 4c_2\hat{}\eta_0^2)c_5(2E^2 - E)\log\ E$,

$c_6 := (2E^2 + 3E + 1)\log(E+1)$,$\tilde{\eta}_0$:

$= \tilde{\mu}/\left(2\sigma^2 + (c_1c_3 + c_2/6)c_4c_6\right)$,$B_1 := 2\sigma^2 +$ (

$4c_1\tilde{\eta}_0 + 4c_2\tilde{\eta}_0^2 + 1/K)c_5c_6$,$B_2$:

$= 2\sigma^2 + \left(4c_1\tilde{\eta}_0 + 4c_2\tilde{\eta}_0^2 + \frac{N-K}{KN}_{(N-1)}\right)c_5c_6$.

我们不直接证明定理 3.1，而是证明一个更一般的收敛结果：假设满足关于步长的一些条件。然后我们将说明，定理 3.1 中给出的步长满足这些条件。

定理 B.4. 在假设 *2.1*、*2.2*、*2.3*、*2.5*、*2.6* 下，对于递减步长 $\eta_t = \frac{\beta}{t+\gamma}$，其中 $\beta > \frac{1}{\tilde{\mu}}$、$\gamma > 0$ 满足 $\eta_0 \leq \hat{}\eta_0$、$\eta_t \leq 2\eta_{t+E}$，且 $\eta_t^2 \leq 1/\left(2c_3(t+1-t_0)(1+2(t+1-t_0))\right)$，则在全参与情况下，对于任意 t

$$\mathbb{E}[\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2] \leq \frac{\upsilon}{\gamma + t}$$

其中 $\upsilon =$最大值 $\left\{\frac{4B}{\tilde{\mu}^2}\right.$、 $\left.\gamma\mathbb{E}[\|\boldsymbol{\theta}^0 - \boldsymbol{\theta}^{PS}\|_2^2]\right\}$。

证明。 设 $\Delta_t := \mathbb{E}[\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{PS}\|_2^2]$，则根据引理 B.3，我们有

$$\Delta_{t+1} \leq (1 - \tilde{\mu}\eta_t)\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 + B\eta_t^2.$$

对于递减步长 $\eta_t = \frac{\beta}{t+\gamma}$，其中 $\beta > \frac{1}{\tilde{\mu}}$，$\gamma > 0$，使得 $\eta_t^2 \leq 1/\left(2c_3(t+1-t_0)(1+2(t+1-t_0))\right)$，$\eta_0 \leq \hat{}\eta_0$，且 $\eta_t \leq 2\eta_{t+E}$，我们将通过归纳法证明 $\Delta_t \leq \frac{\upsilon}{\gamma+t}$，其中 $\upsilon = \max\left\{\frac{\beta^2 B}{\beta\tilde{\mu}-1}, \gamma\Delta_0\right\} = \max\left\{\frac{4B}{\tilde{\mu}^2}, \gamma\Delta_0\right\}$。

首先，根据 υ的定义，$\Delta_0 \leq \frac{\upsilon}{\gamma}$。假设对于某个 $0 \leq t$， $\Delta_t \leq \frac{\upsilon}{\gamma+t}$，那么

$$\begin{aligned}\Delta_{t+1} &\leq (1-\eta_t\tilde{\mu})\,\Delta_t + \eta_t^2 B \\ &\leq \left(1-\frac{\beta\tilde{\mu}}{t+\gamma}\right)\frac{v}{t+\gamma} + \frac{\beta^2 B}{(t+\gamma)^2} \\ &= \frac{t+\gamma-1}{(t+\gamma)^2}v + \left[\frac{\beta^2 B}{(t+\gamma)^2} - \frac{\beta\tilde{\mu}-1}{(t+\gamma)^2}v\right] \\ &\leq \frac{v}{t+\gamma+1}.\end{aligned}$$

具体而言，如果我们选择 $\beta = \frac{2}{\tilde{\mu}}, \gamma = \max\{\frac{2}{\tilde{\mu}\hat{\eta}_0}, E, \frac{2}{\tilde{\mu}}\sqrt{2E(2E+1)(12\sigma^2+18L^2(1+\epsilon_{\max})^2)}\}$，那么我们有

$$\eta_0 = \frac{\beta}{\gamma} \leq \frac{2}{\tilde{\mu}\frac{2}{\tilde{\mu}\hat{\eta}_0}} = \hat{\eta}_0$$

and

$$\eta_t - 2\eta_{t+E} = \frac{\beta}{\gamma+t} - \frac{2\beta}{\gamma+t+E} = \frac{\beta(E-\gamma-t)}{(\gamma+t)(\gamma+t+E)} \leq \frac{\beta(E-\gamma)}{(\gamma+t)(\gamma+t+E)} \leq 0.$$

为了证明对于任意 t， $\eta_t^2 \leq 1/\big(2c_3(t+1-t_0)(1+2(t+1-t_0))\big)$，只需证明对于 $0 \leq t \leq E-1$，因为 $\{\eta_t\}$，即 $t_0 = 0$，是非递增的，而 $t+1-t_0$以 E为周期。当 $t_0 = 0$时，我们需要证明对于 $0 \leq t \leq E-1$，$\eta_t^2 \leq 1/\big(2c_3(t+1)(1+2(t+1))\big)$，这在以下条件满足时成立

$$\begin{aligned}&\max_{0\leq t\leq E-1}\eta_t \leq \min_{0\leq t\leq E-1}\sqrt{1/\big(2c_3(t+1)(1+2(t+1))\big)} \\ \iff &\eta_0 \leq \sqrt{\frac{1}{2E(2E+1)c_3}} \\ \iff &\gamma \geq \beta\sqrt{2E(2E+1)c_3} = \frac{2}{\tilde{\mu}}\sqrt{2E(2E+1)c_3} = \frac{2}{\tilde{\mu}}\sqrt{2E(2E+1)(12\sigma^2+18L^2(1+\epsilon_{\max})^2)}.\end{aligned}$$

□

B.4 关键引理的延后证明

引理 *B.1* 的证明。此证明由 Li 等人 [2022] 的引理 3 推得。我们首先将 $\|\boldsymbol{w}^{t+1} - \boldsymbol{\theta}^{PS}\|_2^2$ 分解为

$$\mathbb{E}\|\overline{\boldsymbol{w}}^{t+1} - \boldsymbol{\theta}^{PS}\|_2^2 = \mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \eta_t\boldsymbol{g}_t - \boldsymbol{\theta}^{PS}\|_2^2 = \mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 - 2\eta_t\mathbb{E}\langle\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS},\, \boldsymbol{g}_t\rangle + \eta_t^2\mathbb{E}\|\boldsymbol{g}_t\|_2^2. \tag{8}$$

接下来我们提出一个向上的 的界。根据 $\boldsymbol{\theta}^{PS}$ 的定义，我们有 $\sum_{i=1}^N p_i\nabla f_i(\boldsymbol{\theta}^{PS};\, \boldsymbol{\theta}^{PS}) = \mathbf{0}$，因此 s

$$\begin{aligned}\mathbb{E}\|\boldsymbol{g}_t\|_2^2 &= \mathbb{E}\|\sum_{i=1}^N p_i[\nabla\ell(\boldsymbol{\theta}_i^t; Z_i^{t+1}) - \nabla f_i(\boldsymbol{\theta}_i^t;\boldsymbol{\theta}_i^t) + \nabla f_i(\boldsymbol{\theta}_i^t;\boldsymbol{\theta}_i^t) - \nabla f_i(\boldsymbol{\theta}^{PS};\boldsymbol{\theta}^{PS})]\|_2^2 \\ &\leq 2\mathbb{E}\|\sum_{i=1}^N \nabla\ell(\boldsymbol{\theta}_i^t; Z_i^{t+1}) - \nabla f_i(\boldsymbol{\theta}_i^t;\boldsymbol{\theta}_i^t)\|_2^2 + 2\mathbb{E}\|\sum_{i=1}^N p_i[\nabla f_i(\boldsymbol{\theta}_i^t;\boldsymbol{\theta}_i^t) - \nabla f_i(\boldsymbol{\theta}^{PS};\boldsymbol{\theta}^{PS})]\|_2^2 \\ &\leq 2\sum_{i=1}^N p_i\mathbb{E}[\|\nabla\ell(\boldsymbol{\theta}_i^t; Z_i^{t+1}) - \nabla f_i(\boldsymbol{\theta}_i^t;\boldsymbol{\theta}_i^t)\|_2^2] + 2\sum_{i=1}^N p_i\mathbb{E}[\|\nabla f_i(\boldsymbol{\theta}_i^t;\boldsymbol{\theta}_i^t) - \nabla f_i(\boldsymbol{\theta}^{PS};\boldsymbol{\theta}^{PS})\|_2^2] \\ &\leq 2\sum_{i=1}^N p_i\sigma^2\left(1+\mathbb{E}\|\boldsymbol{\theta}_i^t - \boldsymbol{\theta}^{PS}\|_2^2\right) + 2\sum_{i=1}^N p_i L^2(1+\epsilon_i)^2\mathbb{E}\|\boldsymbol{\theta}_i^t - \boldsymbol{\theta}^{PS}\|_2^2\end{aligned}$$

其中，第二个不等式源于二范数的凸性，最后一个不等式则源于假设 2.5 和引理 2.4。由于 $\|\boldsymbol{\theta}_i^t - \boldsymbol{\theta}^{PS}\|_2^2 \leq 2\|\boldsymbol{\theta}_i^t - \boldsymbol{\theta}^t\|_2^2 + 2\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{PS}\|_2^2$和 $\bar{\epsilon_i} \leq \epsilon_{max}$，我们有

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{g}_t\|_2^2] &\leq 2\sigma^2 + 4[\sigma^2 + L^2(1+\epsilon_{max})^2]\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 + 4[\sigma^2 + L^2(1+\epsilon_{max})^2]\sum_{i=1}^N p_i \mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2 \\ &= 2\sigma^2 + c_2\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 + c_2\sum_{i=1}^N p_i \mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2. \end{aligned} \tag{9}$$

接下来，我们将重点为 $\mathbb{E}[\langle \boldsymbol{\theta}^t - \boldsymbol{\theta}^{PS},\ \boldsymbol{g}_t\rangle]$ 建立一个下界。根据全期望公式和 $\sum_{i=1}^N p_i \nabla f_i(\boldsymbol{\theta}^{PS};\ \boldsymbol{\theta}^{PS}) = \mathbf{0}$，我们有

$$\begin{aligned}\mathbb{E}\langle \overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS},\ \boldsymbol{g}_t\rangle &= \mathbb{E}\big[\mathbb{E}_t\langle \overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS},\ \boldsymbol{g}_t\rangle\big] \\ &= \mathbb{E}[\langle \overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS},\ \overline{\boldsymbol{g}}_t\rangle] \\ &= \mathbb{E}\Bigg[\underbrace{\sum_{i=1}^N p_i \big\langle \overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS},\ \nabla f_i(\boldsymbol{\theta}_i^t;\boldsymbol{\theta}_i^t) - \nabla f_i(\overline{\boldsymbol{\theta}}^t;\boldsymbol{\theta}^{PS})\big\rangle}_{A} \\ &\qquad + \underbrace{\sum_{i=1}^N p_i \big\langle \overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS},\ \nabla f_i(\overline{\boldsymbol{\theta}}^t;\boldsymbol{\theta}^{PS}) - \nabla f_i(\boldsymbol{\theta}^{PS};\boldsymbol{\theta}^{PS})\big\rangle}_{B}\Bigg].\end{aligned}$$

一方面，应用柯西-施瓦茨不等式和引理 2.4，我们有

$$\begin{aligned}A \geq\ & -\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2 \sum_{i=1}^N p_i\big(L\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2 + L\epsilon_i\|\boldsymbol{\theta}_i^t - \boldsymbol{\theta}^{PS}\|_2\big) \\ \geq\ & -\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2 \sum_{i=1}^N p_i\big(L(1+\epsilon_i)\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2 + L\epsilon_i\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2\big) \\ \geq\ & -L\bar{\epsilon}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 - L(1+\epsilon_{max})\sum_{i=1}^N p_i\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2.\end{aligned}$$

另一方面，根据假设 2.1 中的强凸性，我们有 $B \geq \mu\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{PS}\|_2^2$。因此，对于任意 $\alpha > 0$，利用 A，B 的下界以及杨氏不等式可得，

$$\begin{aligned}&\mathbb{E}[\langle \overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS},\ \boldsymbol{g}_t\rangle] \\ \geq\ & (\mu - L\bar{\epsilon})\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 - L(1+\epsilon_{max})\sum_{i=1}^N p_i\mathbb{E}\big[\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2\big] \\ \geq\ & \big(\mu - L\bar{\epsilon} - \frac{\alpha}{2}L(1+\epsilon_{max})\big)\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 - \frac{L(1+\epsilon_{max})}{2\alpha}\sum_{i=1}^N p_i\mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2 \\ \geq\ & (\mu - (1+\delta)L\bar{\epsilon})\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 - \frac{L(1+\epsilon_{max})^2}{4\delta\bar{\epsilon}}\sum_{i=1}^N p_i\mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2\end{aligned} \tag{10}$$

其中，最后一行我们设定了 $\alpha := \frac{2\delta\epsilon}{1+\epsilon_{max}}$。

回顾一下，我们记

$$c_1 := \frac{L(1+\epsilon_{max})^2}{2\delta\bar{\epsilon}},\ \ c_2 := 4[\sigma^2 + L^2(1+\epsilon_{max})^2],\ \ \tilde{\mu} := \mu - (1+\delta)\bar{\epsilon}L.$$

结合式（8）、（9）和（10），可得

$$
\begin{aligned}
&\mathbb{E}\big[\|\overline{\boldsymbol{w}}^{t+1}-\boldsymbol{\theta}^{PS}\|_2^2\big]\\
\leq\ &\mathbb{E}\|\overline{\boldsymbol{\theta}}^{t}-\boldsymbol{\theta}^{PS}\|_2^2\\
&-2\eta_t\left[(\mu-(1+\delta)L\bar{\epsilon})\mathbb{E}\|\overline{\boldsymbol{\theta}}^{t}-\boldsymbol{\theta}^{PS}\|_2^2-\frac{L(1+\epsilon_{max})}{4\delta\bar{\epsilon}}\sum_{i=1}^{N}p_i\mathbb{E}\|\boldsymbol{\theta}_i^t-\overline{\boldsymbol{\theta}}^{t}\|_2^2\right]\\
&+\eta_t^2\left[2\sigma^2+c_2\mathbb{E}\|\overline{\boldsymbol{\theta}}^{t}-\boldsymbol{\theta}^{PS}\|_2^2+c_2\sum_{i=1}^{N}p_i\mathbb{E}\|\boldsymbol{\theta}_i^t-\overline{\boldsymbol{\theta}}^{t}\|_2^2\right]\\
=\ &(1-2\tilde{\mu}\eta_t+c_2\eta_t^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^{t}-\boldsymbol{\theta}^{PS}\|_2^2+(c_1\eta_t+c_2\eta_t^2)\sum_{i=1}^{N}p_i\mathbb{E}\|\boldsymbol{\theta}_i^t-\overline{\boldsymbol{\theta}}^{t}\|_2^2+2\sigma^2\eta_t^2\\
\leq\ &(1-\tilde{\mu}\eta_t)\mathbb{E}\|\overline{\boldsymbol{\theta}}^{t}-\boldsymbol{\theta}^{PS}\|_2^2+(c_1\eta_t+c_2\eta_t^2)\sum_{i=1}^{N}p_i\mathbb{E}\|\boldsymbol{\theta}_i^t-\overline{\boldsymbol{\theta}}^{t}\|_2^2+2\sigma^2\eta_t^2
\end{aligned}
$$

其中，最后一个不等式是通过观察条件 $\eta_t\leq\tilde{\mu}/c_2$得到的。 □

引理 *B.2* 的证明。在此证明中，为方便起见，我们将针对 $t+1$ 进行讨论，其中假设 $t+1\notin\mathcal{I}_E$，最后再转换回 t。首先，根据更新规则，我们有

$$
\boldsymbol{\theta}_i^{t+1}-\overline{\boldsymbol{\theta}}^{t+1}=\boldsymbol{\theta}_i^{t}-\overline{\boldsymbol{\theta}}^{t}-\eta_t(\nabla\ell(\boldsymbol{\theta}_i^t;Z_i^{t+1})-g_t).
$$

根据杨氏不等式，我们有

$$
\begin{aligned}
\sum_{i=1}^{N}p_i\mathbb{E}\|\boldsymbol{\theta}_i^{t+1}-\overline{\boldsymbol{\theta}}^{t+1}\|_2^2&=\sum_{i=1}^{N}p_i\mathbb{E}\|\boldsymbol{\theta}_i^{t}-\overline{\boldsymbol{\theta}}^{t}-\eta_t(\nabla\ell(\boldsymbol{\theta}_i^t;Z_i^{t+1})-g_t)\|_2^2\\
&\leq(1+\alpha_t)\sum_{i=1}^{N}p_i\mathbb{E}\|\boldsymbol{\theta}_i^{t}-\overline{\boldsymbol{\theta}}^{t}\|_2^2+\eta_t^2(1+\alpha_t^{-1})\underbrace{\sum_{i=1}^{N}p_i\mathbb{E}\|\nabla\ell(\boldsymbol{\theta}_i^t;Z_i^{t+1})-g_t\|_2^2}_{B}
\end{aligned}
\tag{11}
$$

其中 $\alpha_t >$ 0 是一个自由选择的参数。接下来，我们将为 B建立一个上界。注意到

$$
\begin{aligned}
B &= \sum_{i=1}^{N} p_i \mathbb{E}\|\nabla \ell(\boldsymbol{\theta}_i^t; Z_i^{t+1}) - g_t\|_2^2 \\
&= \mathbb{E}\Big[\sum_{i=1}^{N} p_i \|\nabla \ell(\boldsymbol{\theta}_i^t; Z_i^{t+1}) - \sum_{j=1}^{N} p_j \nabla \ell(\boldsymbol{\theta}_j^t; Z_j^{t+1})\|_2^2\Big] \\
&= \mathbb{E}\Big[\sum_{i=1}^{N} p_i \|\nabla \ell(\boldsymbol{\theta}_i^t; Z_i^{t+1}) - \nabla f_i\left(\boldsymbol{\theta}_i^t, \boldsymbol{\theta}_i^t\right) + \nabla f_i\left(\boldsymbol{\theta}_i^t, \boldsymbol{\theta}_i^t\right) - \sum_{j=1}^{N} p_j \nabla f_j(\boldsymbol{\theta}_j^t; \boldsymbol{\theta}_j^t) \\
&\quad + \sum_{j=1}^{N} p_j \nabla f_j(\boldsymbol{\theta}_j^t; \boldsymbol{\theta}_j^t) - \sum_{j=1}^{N} p_j \nabla \ell(\boldsymbol{\theta}_j^t; Z_j^{t+1})\|_2^2\Big] \\
&\leq 3\mathbb{E}\Big[\sum_{i=1}^{N} p_i \|\nabla \ell\left(\boldsymbol{\theta}_i^t; Z_i^{t+1}\right) - \nabla f_i\left(\boldsymbol{\theta}_i^t, \boldsymbol{\theta}_i^t\right)\|_2^2\Big] + 3\mathbb{E}\Big[\sum_{i=1}^{N} p_i \|\nabla f_i\left(\boldsymbol{\theta}_i^t, \boldsymbol{\theta}_i^t\right) - \sum_{j=1}^{N} p_j \nabla f_j\left(\boldsymbol{\theta}_j^t, \boldsymbol{\theta}_j^t\right)\|_2^2\Big] \\
&\quad + 3\mathbb{E}\Big[\sum_{i=1}^{N} p_i \| \sum_{j=1}^{N} p_j \nabla f_j(\boldsymbol{\theta}_j^t; \boldsymbol{\theta}_j^t) - \sum_{j=1}^{N} p_j \nabla \ell(\boldsymbol{\theta}_j^t; Z_j^{t+1})\|_2^2\Big] \\
&\leq 3\mathbb{E}\Big[\sum_{i=1}^{N} p_i \|\nabla \ell\left(\boldsymbol{\theta}_i^t; Z_i^{t+1}\right) - \nabla f_i\left(\boldsymbol{\theta}_i^t, \boldsymbol{\theta}_i^t\right)\|_2^2\Big] + 3\mathbb{E}\Big[\sum_{i=1}^{N} p_i \|\nabla f_i\left(\boldsymbol{\theta}_i^t, \boldsymbol{\theta}_i^t\right) - \sum_{j=1}^{N} p_j \nabla f_j\left(\boldsymbol{\theta}_j^t, \boldsymbol{\theta}_j^t\right)\|_2^2\Big] \\
&\quad + 3\mathbb{E}\Big[\sum_{i=1}^{N} p_i \sum_{j=1}^{N} p_j \|\nabla f_j(\boldsymbol{\theta}_j^t; \boldsymbol{\theta}_j^t) - \nabla \ell(\boldsymbol{\theta}_j^t; Z_j^{t+1})\|_2^2\Big] \\
&= 3\mathbb{E}\Big[\sum_{i=1}^{N} p_i \|\nabla \ell\left(\boldsymbol{\theta}_i^t; Z_i^{t+1}\right) - \nabla f_i\left(\boldsymbol{\theta}_i^t, \boldsymbol{\theta}_i^t\right)\|_2^2\Big] + 3\mathbb{E}\Big[\sum_{i=1}^{N} p_i \|\nabla f_i\left(\boldsymbol{\theta}_i^t, \boldsymbol{\theta}_i^t\right) - \sum_{j=1}^{N} p_j \nabla f_j\left(\boldsymbol{\theta}_j^t, \boldsymbol{\theta}_j^t\right)\|_2^2\Big] \\
&\quad + 3\mathbb{E}\Big[\sum_{j=1}^{N} p_j \|\nabla f_j(\boldsymbol{\theta}_j^t; \boldsymbol{\theta}_j^t) - \nabla \ell(\boldsymbol{\theta}_j^t; Z_j^{t+1})\|_2^2\Big] \\
&\leq 6\sigma^2\Big(1 + \mathbb{E}\Big[\sum_{i=1}^{N} p_i \|\boldsymbol{\theta}_i^t - \boldsymbol{\theta}^{PS}\|_2^2\Big]\Big) \\
&\quad + 3\mathbb{E}\Big[\sum_{i=1}^{N} p_i \|\nabla f_i\left(\boldsymbol{\theta}_i^t, \boldsymbol{\theta}_i^t\right) - \sum_{j=1}^{N} p_j \nabla f_j\left(\boldsymbol{\theta}_j^t, \boldsymbol{\theta}_j^t\right)\|_2^2\Big]
\end{aligned}
$$

其中最后一个不等式由假设 2.5 得出。另一方面，我们有

$$
\begin{aligned}
&3\mathbb{E}\Big[\sum_{i=1}^{N} p_i \big\|\nabla f_i(\boldsymbol{\theta}_i^t, \boldsymbol{\theta}_i^t) - \sum_{j=1}^{N} p_j \nabla f_j(\boldsymbol{\theta}_j^t, \boldsymbol{\theta}_j^t)\big\|_2^2\Big] \\
&= 3\sum_{i=1}^{N} p_i \mathbb{E}\big\|\nabla f_i(\boldsymbol{\theta}_i^t, \boldsymbol{\theta}_i^t) - \nabla f_i(\overline{\boldsymbol{\theta}}^t, \overline{\boldsymbol{\theta}}^t) + \nabla f_i(\overline{\boldsymbol{\theta}}^t, \overline{\boldsymbol{\theta}}^t) - \sum_{j=1}^{N} p_j \nabla f_j(\overline{\boldsymbol{\theta}}^t, \overline{\boldsymbol{\theta}}^t) - \sum_{j=1}^{N} p_j \big(\nabla f_j(\boldsymbol{\theta}_j^t, \boldsymbol{\theta}_j^t) - \nabla f_j(\overline{\boldsymbol{\theta}}^t, \overline{\boldsymbol{\theta}}^t)\big)\big\|_2^2 \\
&\leq \underbrace{9\sum_{i=1}^{N} p_i \mathbb{E}\|\nabla f_i(\boldsymbol{\theta}_i^t, \boldsymbol{\theta}_i^t) - \nabla f_i(\overline{\boldsymbol{\theta}}^t, \overline{\boldsymbol{\theta}}^t)\|_2^2}_{B_1} + \underbrace{9\sum_{i=1}^{N} p_i \mathbb{E}\|\nabla f_i(\overline{\boldsymbol{\theta}}^t, \overline{\boldsymbol{\theta}}^t) - \sum_{j=1}^{N} p_j \nabla f_j(\overline{\boldsymbol{\theta}}^t, \overline{\boldsymbol{\theta}}^t)\|_2^2}_{B_2} \\
&\quad + \underbrace{9\sum_{i=1}^{N} p_i \mathbb{E}\|\sum_{j=1}^{N} p_j \big(\nabla f_j(\boldsymbol{\theta}_j^t, \boldsymbol{\theta}_j^t) - \nabla f_j(\overline{\boldsymbol{\theta}}^t, \overline{\boldsymbol{\theta}}^t)\big)\|_2^2}_{B_3}.
\end{aligned}
$$

根据引理 2.4，我们有

$$B_1 \leq 9\sum_{i=1}^{N} p_i L^2(1+\epsilon_i)^2\mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2 \leq 9\sum_{i=1}^{N} p_i L^2(1+\epsilon_{max})^2\mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2.$$

根据假设 2.6，我们有

$$\begin{aligned} B_2 &= 9\sum_{i=1}^{N} p_i\mathbb{E}\big\|\nabla f_i(\overline{\boldsymbol{\theta}}^t, \overline{\boldsymbol{\theta}}^t) - \nabla f(\overline{\boldsymbol{\theta}}^t, \overline{\boldsymbol{\theta}}^t)\big\|_2^2 \\ &\leq 9\sum_{i=1}^{N} p_i\varsigma^2(1+\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2) \\ &= 9\varsigma^2 + 9\varsigma^2\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2. \end{aligned}$$

根据引理 2.4，我们有

$$B_3 \leq 9\sum_{i=1}^{N} p_i \sum_{j=1}^{N} p_j\mathbb{E}\|\nabla f_j(\boldsymbol{\theta}_j^t, \boldsymbol{\theta}_j^t) - \nabla f_j(\overline{\boldsymbol{\theta}}^t, \overline{\boldsymbol{\theta}}^t)\|_2^2 \leq 9\sum_{i=1}^{N} p_i L^2(1+\epsilon_{max})^2\mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2.$$

因此，

$$B_1 + B_2 + B_3 \leq 18\sum_{i=1}^{N} p_i L^2(1+\epsilon_{max})^2\mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2 + 9\varsigma^2 + 9\varsigma^2\sum_{i=1}^{N} p_i\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2,$$

这导致

$$\begin{aligned} B &\leq 6\sigma^2\bigg(1+\mathbb{E}\Big[\sum_{i=1}^{N} p_i\|\boldsymbol{\theta}_i^t - \boldsymbol{\theta}^{PS}\|_2^2\Big]\bigg) + 18\sum_{i=1}^{N} p_i L^2(1+\epsilon_{max})^2\mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2 + \\ &\quad 9\varsigma^2 + 9\varsigma^2\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 \\ &\leq 6\sigma^2\bigg(1+2\sum_{i=1}^{N} p_i\mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2 + 2\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 + 18\sum_{i=1}^{N} p_i L^2(1+\epsilon_{max})^2\mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2 + \\ &\quad 9\varsigma^2 + 9\varsigma^2\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2\bigg) \\ &= 6\sigma^2 + 9\varsigma^2 + \big(12\sigma^2 + 18L^2(1+\epsilon_{\max})^2\big)\sum_{i=1}^{N} p_i\mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2 + (12\sigma^2 + 9\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2. \end{aligned}$$

将该公式代入 (11)，可得

$$\begin{aligned} &\sum_{i=1}^{N} p_i\mathbb{E}\|\boldsymbol{\theta}_i^{t+1} - \overline{\boldsymbol{\theta}}^{t+1}\|_2^2 \\ &\leq (1+\alpha_t)\sum_{i=1}^{N} p_i\mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2 \\ &\quad + \eta_t^2(1+\alpha_t^{-1})\bigg(6\sigma^2 + 9\varsigma^2 + \big(12\sigma^2 + 18L^2(1+\epsilon_{\max})^2\big)\sum_{i=1}^{N} p_i\mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2 + (12\sigma^2 + 9\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2\bigg) \\ &= \big(1+\alpha_t + \eta_t^2(1+\alpha_t^{-1})\big(12\sigma^2 + 18L^2(1+\epsilon_{\max})^2\big)\big)\sum_{i=1}^{N} p_i\mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2 + \eta_t^2(1+\alpha_t^{-1})(12\sigma^2 + 9\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 \\ &\quad + \eta_t^2(1+\alpha_t^{-1})(6\sigma^2 + 9\varsigma^2) \end{aligned}$$

其中， $\alpha_t > 0$ 是一个任意选取的参数。令 $t_0 := \max\{s \mid s < t+1, s \in \mathcal{I}_E\}$ ，且 $c_3 := 12\sigma^2 + 18L^2(1+\epsilon_{\max})^2$。那么，如果满足以下条件，我们选择 $\alpha_t = \frac{1}{(t+1-t_0)}$：

$$\begin{aligned} &\eta_t^2(1+\alpha_t^{-1})\big(12\sigma^2 + 18L^2(1+\epsilon_{\max})^2\big) = \eta_t^2(1+\alpha_t^{-1})c_3 \le \frac{1}{2(t+1-t_0)} \\ \iff &\eta_t^2 \le \frac{1}{2c_3(t+1-t_0)\big(1+2(t+1-t_0)\big)}, \end{aligned} \tag{12}$$

然后注意到 $1+\alpha_t^{-1} = 1+2(t+1-t_0) \le 2E-1$,

$$\begin{aligned} &\sum_{i=1}^N p_i \mathbb{E}\|\boldsymbol{\theta}_i^{t+1} - \overline{\boldsymbol{\theta}}^{t+1}\|_2^2 \\ \le &\frac{t+2-t_0}{t+1-t_0}\sum_{i=1}^N p_i \mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2 + \eta_t^2(2E-1)(12\sigma^2+9\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 + \eta_t^2(2E-1)(6\sigma^2+9\varsigma^2). \end{aligned}$$

继续上述展开，直到 t_0 并利用 $\eta_s \le \eta_{t_0} \le 2\eta_{t_0+E} \le 2\eta_t$ ，我们得到

$$\begin{aligned} &\sum_{i=1}^N p_i \mathbb{E}\|\boldsymbol{\theta}_i^{t+1} - \overline{\boldsymbol{\theta}}^{t+1}\|_2^2 \\ \le &\frac{t+2-t_0}{t_0+1-t_0}\sum_{i=1}^N p_i \mathbb{E}\|\boldsymbol{\theta}_i^{t_0} - \overline{\boldsymbol{\theta}}^{t_0}\|_2^2 + \sum_{s=t_0}^t \frac{t+2-t_0}{s+2-t_0}\eta_s^2(2E-1)(12\sigma^2+9\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^s - \boldsymbol{\theta}^{PS}\|_2^2 \\ &+ \sum_{s=t_0}^t \frac{t+2-t_0}{s+2-t_0}\eta_s^2(2E-1)(6\sigma^2+9\varsigma^2) \\ = &\sum_{s=0}^{t-t_0} \frac{t+2-t_0}{s+2}\eta_s^2(2E-1)(12\sigma^2+9\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^s - \boldsymbol{\theta}^{PS}\|_2^2 + \sum_{s=0}^{t-t_0} \frac{t+2-t_0}{s+2}\eta_s^2(2E-1)(6\sigma^2+9\varsigma^2) \\ \le &\sum_{s=0}^{t-t_0} \frac{t+2-t_0}{s+2}\eta_t^2(2E-1)(48\sigma^2+36\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^s - \boldsymbol{\theta}^{PS}\|_2^2 + \sum_{s=0}^{t-t_0} \frac{t+2-t_0}{s+2}\eta_t^2(2E-1)(24\sigma^2+36\varsigma^2). \end{aligned} \tag{13}$$

根据上述公式和引理 B.1，我们现在证明，如果 η_0 足够小，那么对于任意 t，都有 $\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 \le \mathbb{E}\|\boldsymbol{\theta}^0 - \boldsymbol{\theta}^{PS}\|_2^2$。我们首先推导出以下不等式，稍后将用到。注意，对于任意 t，其中 $t_0 := \max\{s \mid s < t+1, s \in \mathcal{I}_E\}$，都有

$$\sum_{s=0}^{t-t_0} \frac{t+2-t_0}{s+2} = (t+2-t_0)\big(\frac{1}{2} + \ldots + \frac{1}{t-t_0+2}\big) \le (t+2-t_0)\log(t+2-t_0) \le E\log E. \tag{14}$$

我们通过归纳法证明 $\mathbb{E}\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{PS}\|_2^2 \le \mathbb{E}\|\boldsymbol{\theta}^0 - \boldsymbol{\theta}^{PS}\|_2^2$。

首先，对于 $t=0$，该不等式显然成立。假设对于 $0 \leq s \leq t$，其中 $t \leq E-1$，该不等式成立。那么根据引理 B.1 和 (13)，我们有

$$
\begin{aligned}
&\mathbb{E}[\|\overline{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^{PS}\|_2^2] \\
&= \mathbb{E}[\|\overline{\boldsymbol{w}}^{t+1} - \boldsymbol{\theta}^{PS}\|_2^2] \\
&\leq (1-\tilde{\mu}\eta_t)\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2 + 2\sigma^2\eta_t^2 + (c_1\eta_t + c_2\eta_t^2)\Bigg(\sum_{s=0}^{t-t_0-1} \frac{t+1-t_0}{s+2}\eta_{t-1}^2(2E-1)(48\sigma^2+36\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2 \\
&\quad + \sum_{s=0}^{t-t_0-1} \frac{t+1-t_0}{s+2}\eta_{t-1}^2(2E-1)(24\sigma^2+36\varsigma^2)\Bigg) \\
&= (1-\tilde{\mu}\eta_t)\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2 + 2\sigma^2\eta_t^2 \\
&\quad + (c_1\eta_t + c_2\eta_t^2)\eta_{t-1}^2(2E^2-E)\log E\Bigg((48\sigma^2+36\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2 + (24\sigma^2+36\varsigma^2)\Bigg).
\end{aligned}
$$

根据 (12)，我们需要 ~~η_0~~ $\leq \frac{1}{6c_3}$，而 $\eta_{t-1} \leq 2\eta_t$ 意味着

$$
(c_1\eta_t + c_2\eta_t^2)\eta_{t-1}^2 \leq \eta_t^2(c_1 + c_2(6c_3)^{-1})\frac{c_3}{3}.
$$

因此，我们有

$$
\begin{aligned}
&\mathbb{E}[\|\overline{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^{PS}\|_2^2] \\
&\leq (1-\tilde{\mu}\eta_t)\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2 + 2\sigma^2\eta_t^2 \\
&\quad + \eta_t^2(c_1c_3 + c_2/6)(2E^2-E)\log E\Bigg((16\sigma^2+12\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2 + (8\sigma^2+12\varsigma^2)\Bigg)
\end{aligned}
$$

其右端不大于 $\mathbb{E}\|\boldsymbol{\theta}^0 - \boldsymbol{\theta}^{PS}\|_2^2$ 若

$$
\begin{aligned}
\eta_t &\leq \eta_0 \\
&\leq \frac{\tilde{\mu}\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2}{2\sigma^2 + (c_1c_3 + c_2/6)(2E^2-E)\log E\big((16\sigma^2+12\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2 + (8\sigma^2+12\varsigma^2)\big)} = \hat{\eta}_0. \qquad (15)
\end{aligned}
$$

因此，我们已经证明，对于任意 $0 \leq t \leq E$，如果（15）成立，则 $\mathbb{E}\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{PS}\|_2^2 \leq \mathbb{E}\|\boldsymbol{\theta}^0 - \boldsymbol{\theta}^{PS}\|_2^2$。相同的证明技巧可以推广到任意 $nE \leq t \leq (n+1)E$，其中 $n \in \mathbb{N}_+$，因此对于任意 t，如果（15）成立，则$\mathbb{E}\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{PS}\|_2^2 \leq \mathbb{E}\|\boldsymbol{\theta}^0 - \boldsymbol{\theta}^{PS}\|_2^2$。

因此，在满足 (12)、(15) 和 $\eta_{t-1} \leq 2\eta_t$的条件下，根据 (11) 和 (13)，若 $t \notin \mathcal{I}_E$，则有

$$
\begin{aligned}
\sum_{i=1}^N p_i\mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2 &\leq \eta_{t-1}^2(2E^2-E)\log E(48\sigma^2+36\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2 + \eta_{t-1}^2(2E^2-E)\log E(24\sigma^2+36\varsigma^2) \\
&\leq 4\eta_t^2(2E^2-E)\log E(48\sigma^2+36\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2 + 4\eta_t^2(2E^2-E)\log E(24\sigma^2+36\varsigma^2).
\end{aligned}
$$

□

引理 *B.3* 的证明。我们分两种情况讨论，$t \in \mathcal{I}_E$ 和 $t \notin \mathcal{I}_E$。如果 $t \in \mathcal{I}_E$，则有 $\boldsymbol{\theta}_i^t = \boldsymbol{\theta}^t$，根据引理 B.1，

$$
\begin{aligned}
&\mathbb{E}[\|\overline{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^{PS}\|_2^2] \\
&\leq (1-\tilde{\mu}\eta_t)\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 + 2\sigma^2\eta_t^2 + (c_1\eta_t + c_2\eta_t^2)\sum_{i=1}^N p_i\mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2 \\
&= (1-\tilde{\mu}\eta_t)\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 + 2\sigma^2\eta_t^2 \\
&\leq (1-\tilde{\mu}\eta_t)\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 + B\eta_t^2.
\end{aligned}
$$

如果 $t \notin \mathcal{I}_E$，结合引理 B.1 和引理 B.2，可得

$$
\begin{aligned}
&\mathbb{E}[\|\overline{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^{PS}\|_2^2] \\
&\leq (1-\tilde{\mu}\eta_t)\mathbb{E}\|\overline{\boldsymbol{\theta}}^{t} - \boldsymbol{\theta}^{PS}\|_2^2 + 2\sigma^2\eta_t^2 + (c_1\eta_t + c_2\eta_t^2)\sum_{i=1}^N p_i \mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2 \\
&\leq (1-\tilde{\mu}\eta_t)\mathbb{E}\|\overline{\boldsymbol{\theta}}^{t} - \boldsymbol{\theta}^{PS}\|_2^2 + 2\sigma^2\eta_t^2 \\
&\quad + (4c_1\eta_t + 4c_2\eta_t^2)\left(\eta_t^2(2E^2-E)\log E(48\sigma^2+36\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2 + \eta_t^2(2E^2-E)\log E(24\sigma^2+36\varsigma^2)\right) \\
&\leq (1-\tilde{\mu}\eta_t)\mathbb{E}\|\overline{\boldsymbol{\theta}}^{t} - \boldsymbol{\theta}^{PS}\|_2^2 + 2B\eta_t^2.
\end{aligned}
$$

□

C 定理 3.2 和定理 3.3 的证明

C.1 补充符号

与附录B类似，为方便分析，我们将定义两个额外的序列为$\boldsymbol{w}^t := \sum_{i=1}^N p_i \boldsymbol{w}_i^t$ 和 $\boldsymbol{\theta}^t := \sum_{i=1}^N p_i \boldsymbol{\theta}_i^t$，沿用 Li 等大 [2020b] 的做法。注意，$\boldsymbol{w}^t$ 是从 $\boldsymbol{\theta}^t$执行一步随机梯度下降得到的。当 $t+1 \notin \mathcal{I}_E$时，无法获取 $\boldsymbol{w}^t$ 和 $\boldsymbol{\theta}^t$。当 $t+1 \in \mathcal{I}_E$时，我们可以得到 $\boldsymbol{\theta}^t$。此外，我们还定义 $\boldsymbol{g}_t := \sum_{i=1}^N p_i \nabla f_i(\boldsymbol{\theta}_i^t; \boldsymbol{\theta}_i^t)$ $\overline{\boldsymbol{g}}_t := \sum_{i=1}^N p_i \overline{\nabla} \ell(\boldsymbol{\theta}_i^t;\, Z_i^{t+1})$，其中 $Z_i^{t+1} \sim \overline{\mathcal{D}}_i(\boldsymbol{\theta}_i^t)$。显然，在全参与的情况下，$\boldsymbol{w}^{t+1} = \boldsymbol{w}^t - \eta_t \boldsymbol{g}_t$ 和 $\overline{\mathbb{E}}\boldsymbol{g}_t = \overline{\boldsymbol{g}}_t$。请注意，此时对于任意 t，我们都没有 $\overline{\boldsymbol{\theta}^t} = \boldsymbol{w}^t$。但我们将在后文证明，关于 $\mathcal{S}_t$的选择，它们在期望意义下相等。

具体而言，在我们的分析中存在两类随机性，一类来自随机梯度，另一类来自设备的随机抽样。附录B中的所有分析仅涉及前者。为加以区分，我们用 $\mathbb{E}_{\mathcal{S}_t}$ 表示后者。

C.2 关键引理

本小节给出证明收敛所需的关键引理，并将证明留待后文。我们首先证明抽样方案 I 和 II 是无偏的。

引理 C.1. *Li 等人 [2020b]（无偏抽样方案）。若 $t+1 \in \mathcal{I}_E$，对于方案 I 和方案 II，我们有*

$$\mathbb{E}_{\mathcal{S}_t}\left[\overline{\boldsymbol{\theta}}^{t+1}\right] = \overline{\boldsymbol{w}}^{t+1}.$$

与引理 B.1 类似，我们将为 $\mathbb{E}[\|\boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^{PS}\|_2^2]$建立一个上界。当 $t+1 \notin \mathcal{I}_E$时，对于两种方案，我们都有 $\boldsymbol{\theta}^{t+1} = \boldsymbol{w}^{t+1}$，因此这等价于为 $\mathbb{E}[\|\boldsymbol{w}^{t+1} - \boldsymbol{\theta}^{PS}\|_2^2]$建立一个上界。然而，当 $t+1 \in \mathcal{I}_E$时，我们只有 $\mathbb{E}_{\mathcal{S}_t}\left[\boldsymbol{\theta}^{t+1}\right] = \boldsymbol{w}^{t+1}$，因此需要采用其他求上界的策略。

引理 C.2. *在假设 2.1、2.2、2.3、2.5 下，对于方案 I 和 II：*

1. 如果 $t+1 \notin \mathcal{I}_E$，

$$\mathbb{E}[\|\overline{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^{PS}\|_2^2] = \mathbb{E}[\|\overline{\boldsymbol{w}}^{t+1} - \boldsymbol{\theta}^{PS}\|_2^2] \leq (1-\tilde{\mu}\eta_t)\mathbb{E}\|\overline{\boldsymbol{\theta}}^{t} - \boldsymbol{\theta}^{PS}\|_2^2 + 2\sigma^2\eta_t^2 + (c_1\eta_t + c_2\eta_t^2)\sum_{i=1}^N p_i \mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2.$$

2. 若 $t+1 \in \mathcal{I}_E$：对于方案 I，

$$
\begin{aligned}
&\mathbb{E}[\|\overline{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^{PS}\|_2^2] \\
&\leq \frac{1}{K}\sum_{k=1}^N p_k \mathbb{E}\|\boldsymbol{w}_k^{t+1} - \overline{\boldsymbol{w}}^{t+1}\|_2^2 + (1-\tilde{\mu}\eta_t)\mathbb{E}\|\overline{\boldsymbol{\theta}}^{t} - \boldsymbol{\theta}^{PS}\|_2^2 + 2\sigma^2\eta_t^2 + (c_1\eta_t + c_2\eta_t^2)\sum_{i=1}^N p_i \mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2,
\end{aligned}
$$

而对于方案 *II*，

$$\begin{aligned}&\mathbb{E}[\|\overline{\boldsymbol{\theta}}^{t+1}-\boldsymbol{\theta}^{PS}\|_2^2]\\&\leq\frac{N}{K(N-1)}\left(1-\frac{K}{N}\right)\sum_{k=1}^{N}p_k\mathbb{E}\|\boldsymbol{w}_k^{t+1}-\overline{\boldsymbol{w}}^{t+1}\|_2^2+(1-\tilde{\mu}\eta_t)\mathbb{E}\|\overline{\boldsymbol{\theta}}^{t}-\boldsymbol{\theta}^{PS}\|_2^2+2\sigma^2\eta_t^2+(c_1\eta_t+c_2\eta_t^2)\sum_{i=1}^{N}p_i\mathbb{E}\|\boldsymbol{\theta}_i^t-\overline{\boldsymbol{\theta}}^{t}\|_2^2.\end{aligned}$$

其中 $\epsilon_{max}:=$最大值，$\epsilon:=\sum_{i=1}^{N}p_i\epsilon_i$，$c_1:=\frac{L_{(1+\epsilon_{max})}}{2\delta\epsilon}^2$，$c_2:=4[\sigma^2+L^2(1+\epsilon_{max})^2]$，$\tilde{\mu}:=\mu-(1+\delta)\epsilon L_{\circ}$-

为了真正给出如随机梯度下降分析中的下降引理，我们必须分别界定 $\sum_{i=1}^{N}p_i\mathbb{E}\|\boldsymbol{\theta}_i^t-\boldsymbol{\theta}^t\|_2^2$ 和 $t\notin\mathcal{I}_E$ 以及 $\sum_{i=1}^{N}p_i\mathbb{E}\|\boldsymbol{w}_i^t=\boldsymbol{w}^t\|_2^2$ 和 $t\in\mathcal{I}_E$，具体如下引理所述。

引理 C.3。在假设 *2.1*、*2.2*、*2.3*、*2.5*、*2.6* 下，若 $\{\eta_t\}$ 非递增，$\eta_t\leq 2\eta_{t+E}$, $t\notin\mathcal{I}_E$, $\eta_t^2\leq 1/\big(2c_3($ $t+1-t_0)(1+2(t+1-t_0))\big)$，且

$$\eta_0\leq\tilde{\eta}_0:=\frac{\tilde{\mu}\mathbb{E}\|\overline{\boldsymbol{\theta}}^{0}-\boldsymbol{\theta}^{PS}\|_2^2}{2\sigma^2+(c_1c_3+c_2/6)(2E^2+3E+1)\log(E+1)\big((16\sigma^2+12\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^{0}-\boldsymbol{\theta}^{PS}\|_2^2+(8\sigma^2+12\varsigma^2)\big)},$$

then

1. 对于方案 *I*，

$$\begin{aligned}&\mathbb{E}[\|\overline{\boldsymbol{\theta}}^{t+1}-\boldsymbol{\theta}^{PS}\|_2^2]\\&\leq(1-\tilde{\mu}\eta_t)\mathbb{E}\|\overline{\boldsymbol{\theta}}^{t}-\boldsymbol{\theta}^{PS}\|_2^2+2\sigma^2\eta_t^2\\&\quad+(4c_1\eta_t+4c_2\eta_t^2+K^{-1})(2E^2+3E+1)\log(E+1)\eta_t^2\left((48\sigma^2+36\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^{0}-\boldsymbol{\theta}^{PS}\|_2^2+(24\sigma^2+36\varsigma^2)\right),\end{aligned}$$

2. 对于方案 *II*，

$$\begin{aligned}&\mathbb{E}[\|\overline{\boldsymbol{\theta}}^{t+1}-\boldsymbol{\theta}^{PS}\|_2^2]\\&\leq(1-\tilde{\mu}\eta_t)\mathbb{E}\|\overline{\boldsymbol{\theta}}^{t}-\boldsymbol{\theta}^{PS}\|_2^2+2\sigma^2\eta_t^2\\&\quad+\left(4c_1\eta_t+4c_2\eta_t^2+\frac{N-K}{K(N-1)}\right)(2E^2+3E+1)\log(E+1)\eta_t^2\left((48\sigma^2+36\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^{0}-\boldsymbol{\theta}^{PS}\|_2^2+(24\sigma^2+36\varsigma^2)\right).\end{aligned}$$

其中，对于任意 t，其中 $\mu-1+\delta\ \epsilon L\ c_3=12\sigma^2+18L^2\ 1+\epsilon_{\max}{}^2$ ：

$\max,:=\sum_{i=1}^{N}p_i\epsilon_i,:=\frac{L_{(1+\epsilon_{max})}}{2\delta\epsilon}^2,:=4[\sigma^2+L^2(1+\epsilon_{max})^2],\tilde{\mu}:(),:()_{\circ}$

（需要注意的是，$4c_1\eta_t+4c_2\eta_t^2$、$(48\sigma^2+36\varsigma^2)$ 和 $(24\sigma^2+36\varsigma^2)$ 是多次应用$\eta_{t-1}\leq 2\eta_t$ 所得，而且通过谨慎选择步长，实际常数可以小得多。）

下面的引理给出了随机梯度下降分析中的标准下降引理，并附有技术条件，用于建立定理 3.2 和 3.3 中的 $\mathcal{O}(\frac{1}{T})$ 收敛。

引理 C.4。在假设 *2.1*、*2.2*、*2.3*、*2.5*、*2.6* 下，若 $\{\eta_t\}$ 单调不增，$\eta_t\leq 2\eta_{t+E}$, $t\notin\mathcal{I}_E$, $\eta_t^2\leq 1/\big(2c_3$ $(t+1-t_0)$ $(1+2\ (t+1-t_0)\ )$，且

$$\eta_0\leq\tilde{\eta}_0:=\frac{\tilde{\mu}\mathbb{E}\|\overline{\boldsymbol{\theta}}^{0}-\boldsymbol{\theta}^{PS}\|_2^2}{2\sigma^2+(c_1c_3+c_2/6)(2E^2+3E+1)\log(E+1)\big((16\sigma^2+12\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^{0}-\boldsymbol{\theta}^{PS}\|_2^2+(8\sigma^2+12\varsigma^2)\big)},$$

then

1. 对于方案 *I*，

$$\mathbb{E}[\|\overline{\boldsymbol{\theta}}^{t+1}-\boldsymbol{\theta}^{PS}\|_2^2]\leq(1-\tilde{\mu}\eta_t)\mathbb{E}\|\overline{\boldsymbol{\theta}}^{t}-\boldsymbol{\theta}^{PS}\|_2^2+B_1\eta_t^2,$$

with

$$B_1:=2\sigma^2+(4c_1\eta_t+4c_2\eta_t^2+K^{-1})(2E^2+3E+1)\log(E+1)\left((48\sigma^2+36\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^{0}-\boldsymbol{\theta}^{PS}\|_2^2+(24\sigma^2+36\varsigma^2)\right),$$

2. 对于方案 *II*,

$$\mathbb{E}[\|\overline{\boldsymbol{\theta}}^{t+1}-\boldsymbol{\theta}^{PS}\|_2^2] \le (1-\tilde{\mu}\eta_t)\mathbb{E}\|\overline{\boldsymbol{\theta}}^{t}-\boldsymbol{\theta}^{PS}\|_2^2 + B_2\eta_t^2$$

with

$$B_2 := 2\sigma^2 + \left(4c_1\eta_t + 4c_2\eta_t^2 + \frac{N-K}{K(N-1)}\right)(2E^2+3E+1)\log(E+1)\Big((48\sigma^2+36\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^{0}-\boldsymbol{\theta}^{PS}\|_2^2 + (24\sigma^2+36\varsigma^2)\Big)$$

其中，对于任意 t，其中 $\mu - 1+\delta\ \epsilon L\ c_3 = 12\sigma^2 + 18L^2\ 1+\epsilon_{\max}{}^2$ ：

最大值,: $= \sum_{i=1}^N p_i\epsilon_b$: $= {}^{L}{}_{(1+\epsilon_{max})_{2\delta\epsilon}}{}^{2}$,: $= 4[\sigma^2 + L^2(1+\epsilon_{max})^2]$, $\tilde{\mu}$:(),:()。

C.3 定理 3.2 和 3.3 的证明补充

我们在此重述所有常数的定义：

与系统设计无关的常数。

$\epsilon_{max} := \max_i \epsilon_i$,

$\overline{\epsilon} := \sum_{i=1}^N p_i\epsilon_i$,

$\tilde{\mu} := \mu - (1+\delta)\overline{\epsilon}L$,

$c_1 := \big(L(1+\epsilon_{max})^2\big)/(2\delta\overline{\epsilon})$,

$c_2 := 4\big[\sigma^2 + L^2(1+\epsilon_{max})^2\big]$,

$c_3 := 6\big[2\sigma^2 + 3L^2(1+\epsilon_{\max})^2\big]$,

$c_4 := 16\sigma^2 + 12\varsigma^2 + (8\sigma^2+12\varsigma^2)/\mathbb{E}\|\overline{\boldsymbol{\theta}}^{0}-\boldsymbol{\theta}^{PS}\|_2^2$,

$c_5 := (48\sigma^2+36\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^{0}-\boldsymbol{\theta}^{PS}\|_2^2 + (24\sigma^2+36\varsigma^2)$.

与系统设计相关的常数（例如， E, K）。

$\hat{}\eta_0 := \tilde{\mu}/\big(2\sigma^2 + (c_1c_3 + c_2/6)c_4(2E^2-E)$ 对数 E),

B:$= 2\sigma^2 + (4c_1\hat{}\eta_0 + 4c_2\hat{}\eta_0^2)c_5(2E^2-E)$ 对数

E,c_6 :$= (2E^2+3E+1)$对数$(E+1)$,$\tilde{\eta}_0$:

$= \tilde{\mu}/\big(2\sigma^2 + (c_1c_3 + c_2/6)c_4c_6\big)$,$B_1$:$= 2\sigma^2 +$(

$4c_1\tilde{\eta}_0 + 4c_2\tilde{\eta}_0^2 + 1/K)c_5c_6$,$B_2$:

$= 2\sigma^2 + \big(4c_1\tilde{\eta}_0 + 4c_2\tilde{\eta}_0^2 + {}^{N-K}_{K\ N-1()}\big)c_5c_6$.

我们不直接证明定理 3.2 和 3.3，而是假设满足关于步长的某些条件，证明一个更一般版本的收敛结果。然后我们将证明定理 3.2 和 3.3 中给出的步长满足这些条件。引理的证明将推迟

定理 C.5。 在假设 *2.1*、*2.2*、*2.3*、*2.5*、*2.6* 下，对于递减步长 $\eta_t = \frac{\beta}{t+\gamma}$，其中 $\beta > \frac{1}{\tilde{\mu}}$、$\gamma > 0$ 满足 $\eta_0 \le \tilde{\eta}_0$、$\eta_t \le 2\eta_{t+E}$，且 $\eta_t^2 \le 1/\big(2c_3\ (t+1-t_0)\ \ (1+2\ (t+1-t_0)\ )\big)$，则

1. 对于方案 *I*,

$$\mathbb{E}[\|\overline{\boldsymbol{\theta}}^{t+1}-\boldsymbol{\theta}^{PS}\|_2^2] \le \frac{\upsilon}{\gamma+t},$$

其中 υ =最大值 $\left\{\frac{4B_1}{\tilde{\mu}^2}\text{、}\ \gamma\mathbb{E}[\|\boldsymbol{\theta}^{0}-\boldsymbol{\theta}^{PS}\|_2^2]\right\}$;

2. 对于方案 *II*,

$$\mathbb{E}[\|\overline{\boldsymbol{\theta}}^{t+1}-\boldsymbol{\theta}^{PS}\|_2^2] \le \frac{\upsilon}{\gamma+t}$$

其中 υ =最大化 $\left\{\frac{4B_2}{\tilde{\mu}^2}, \gamma\mathbb{E}[\|\boldsymbol{\theta}^{0}-\boldsymbol{\theta}^{PS}\|_2^2]\right\}$。

证明。 我们给出方案 I 的证明，方案 II 的证明完全相同。 .
设 $\Delta_t := \mathbb{E}[\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{PS}\|_2^2]$，则由引理 C.4 可得

$$\Delta_{t+1} \leq (1-\tilde{\mu}\eta_t)\mathbb{E}\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\|_2^2 + B_1\eta_t^2.$$

对于递减步长 $\eta_t = \frac{\beta}{t+\gamma}$，其中 $\beta > \frac{1}{\tilde{\mu}}$，$\gamma > 0$，使得 $\eta_t^2 \leq 1/\big(2c_3(t+1-t_0)(1+2(t+1-t_0))\big)$，$\eta_0 \leq \hat{}\eta_0$，且 $\eta_t \leq 2\eta_{t+E}$，我们将通过归纳法证明 $\Delta_t \leq \frac{v}{\gamma+t}$，其中 $v = \max\left\{\frac{\beta^2 B_1}{\beta\tilde{\mu}-1},\ \gamma\Delta_0\right\} = \max\left\{\frac{4B_1}{\tilde{\mu}^2},\ \gamma\Delta_0\right\}$。

首先，$\Delta_0 \leq \frac{v}{\gamma}$ 根据 v的定义。假设对于某些 $0 \leq t$，$\Delta_t \leq \frac{v}{\gamma+t}$，则 —

$$\begin{aligned}
\Delta_{t+1} &\leq (1-\eta_t\tilde{\mu})\,\Delta_t + \eta_t^2 B_1 \\
&\leq \left(1 - \frac{\beta\tilde{\mu}}{t+\gamma}\right)\frac{v}{t+\gamma} + \frac{\beta^2 B_1}{(t+\gamma)^2} \\
&= \frac{t+\gamma-1}{(t+\gamma)^2}v + \left[\frac{\beta^2 B_1}{(t+\gamma)^2} - \frac{\beta\tilde{\mu}-1}{(t+\gamma)^2}v\right] \\
&\leq \frac{v}{t+\gamma+1}.
\end{aligned}$$

具体而言，如果我们选择 $\beta = \frac{2}{\tilde{\mu}}$，$\gamma = \max\{\frac{2}{\tilde{\mu}\tilde{\eta}_0} - E, \frac{2}{\tilde{\mu}}\sqrt{(4E^2+10E+6)(12\sigma^2+18L^2(1+\epsilon_{\max})^2)}\}$，则有

$$\eta_0 = \frac{\beta}{\gamma} \leq \frac{2}{\tilde{\mu}\frac{2}{\tilde{\mu}\tilde{\eta}_0}} = \hat{\eta}_0$$

and

$$\eta_t - 2\eta_{t+E} = \frac{\beta}{\gamma+t} - \frac{2\beta}{\gamma+t+E} = \frac{\beta(E-\gamma-t)}{(\gamma+t)(\gamma+t+E)} \leq \frac{\beta(E-\gamma)}{(\gamma+t)(\gamma+t+E)} \leq 0.$$

要证明对于任意 t，$\eta_t^2 \leq 1/\big(2c_3(t+1-t_0)(1+2(t+1-t_0))\big)$，只需证明对于 $0 \leq t \leq E$，因为$\{\eta_t\}$，即 $t_0 = 0$是非递增的，且 $t+1-t_0$ 的周期为 E。当 $t_0 = 0$时，我们需要证明对于 $0 \leq t \leq E$，$\eta_t^2 \leq 1/\big(2c_3(t+1)(1+2(t+1))\big)$，而这在以下条件成立时即可满足

$$\begin{aligned}
&\max_{0\leq t\leq E}\eta_t \leq \min_{0\leq t\leq E}\sqrt{1/\big(2c_3(t+1)(1+2(t+1))\big)} \\
\iff &\eta_0 \leq \sqrt{\frac{1}{(4E^2+10E+6)c_3}} \\
\iff &\gamma \geq \beta\sqrt{(4E^2+10E+6)c_3} = \frac{2}{\tilde{\mu}}\sqrt{(4E^2+10E+6)c_3} = \frac{2}{\tilde{\mu}}\sqrt{(4E^2+10E+6)(12\sigma^2+18L^2(1+\epsilon_{\max})^2)}.
\end{aligned}$$

□

C.4 关键引理的延后证明

C.1 的证明。令 $\{x_i\}_{i=1}^N$ 表示任意固定的确定性序列。我们通过以下过程，使用 $|\mathcal{S}_t| = K$ 对 $\mathcal{S}_t$ 进行多重集抽样：每次抽样时，以概率 q_k 抽取 x_k。请注意，两次抽样不一定相互独立。我们只要求每次抽样的抽样分布相同。令 $\mathcal{S}_t = \{i_1, \ldots, i_K\} \subset [N]$（如果是有放回抽样，某些 i_k 可能取值相同）。那么

$$\mathbb{E}_{\mathcal{S}_t}\sum_{k\in\mathcal{S}_t} x_k = \mathbb{E}_{\mathcal{S}_t}\sum_{k=1}^K x_{i_k} = K\mathbb{E}_{\mathcal{S}_t}x_{i_1} = K\sum_{k=1}^K q_k x_k.$$

对于 Sche 对于方案 I，$q_k = p_k$；对于方案 II，$q_k = \frac{1}{N}$，将这些值代入上式即可证明该引理。 a. □

引理 *C.2* 的证明。当 $t+1 \notin \mathcal{I}_E$ 时，由于两种方案的 $\overline{\boldsymbol{\theta}}^{t+1} = \boldsymbol{w}^{t+1}$，根据引理 B.1，我们得出结论。当 $t+1 \in \mathcal{I}_E$ 时，我们有

$$\mathbb{E}\|\overline{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^{PS}\|_2^2 = \mathbb{E}\|\overline{\boldsymbol{\theta}}^{t+1} - \overline{\boldsymbol{w}}^{t+1}\|_2^2 + \mathbb{E}\|\overline{\boldsymbol{w}}^{t+1} - \boldsymbol{\theta}^{PS}\|_2^2 + 2\mathbb{E}\langle\overline{\boldsymbol{\theta}}^{t+1} - \overline{\boldsymbol{w}}^{t+1}, \overline{\boldsymbol{w}}^{t+1} - \boldsymbol{\theta}^{PS}\rangle.$$

根据引理 C.1 和全期望公式，我们有

$$\mathbb{E}\langle\overline{\boldsymbol{\theta}}^{t+1} - \overline{\boldsymbol{w}}^{t+1}, \overline{\boldsymbol{w}}^{t+1} - \boldsymbol{\theta}^{PS}\rangle = \mathbb{E}\big[\mathbb{E}_{\mathcal{S}_{t+1}}\langle\overline{\boldsymbol{\theta}}^{t+1} - \overline{\boldsymbol{w}}^{t+1}, \overline{\boldsymbol{w}}^{t+1} - \boldsymbol{\theta}^{PS}\rangle\big] = 0.$$

接下来我们将重点研究在两种抽样方案下对 $\mathbb{E}[\|\boldsymbol{\theta}^{t+1} - \boldsymbol{w}^{t+1}\|_2^2]$ 求上界。

记作 $\mathcal{S}_{t+1} = \{i_1, \ldots, i_K\}$，然后对于方案 I，$\overline{\boldsymbol{\theta}}^{t+1} \equiv \frac{1}{K}\sum_{l=1}^K \boldsymbol{w}_{i_l}^{t+1}$。因此根据全期望公式，我们有

$$\begin{aligned}
\mathbb{E}\|\overline{\boldsymbol{\theta}}^{t+1} - \overline{\boldsymbol{w}}^{t+1}\|_2^2 &= \mathbb{E}\big[\mathbb{E}_{\mathcal{S}_{t+1}}\|\overline{\boldsymbol{\theta}}^{t+1} - \overline{\boldsymbol{w}}^{t+1}\|_2^2\big] \\
&= \mathbb{E}\big[\mathbb{E}_{\mathcal{S}_{t+1}}\|\frac{1}{K}\sum_{l=1}^K \boldsymbol{w}_{i_l}^{t+1} - \overline{\boldsymbol{w}}^{t+1}\|_2^2\big] \\
&\le \mathbb{E}\big[\mathbb{E}_{\mathcal{S}_{t+1}}\frac{1}{K^2}\sum_{l=1}^K \|\boldsymbol{w}_{i_l}^{t+1} - \overline{\boldsymbol{w}}^{t+1}\|_2^2\big] \\
&= \frac{1}{K}\sum_{k=1}^N p_k \mathbb{E}\|\boldsymbol{w}_k^{t+1} - \overline{\boldsymbol{w}}^{t+1}\|_2^2.
\end{aligned}$$

同样，对于 $\overline{\boldsymbol{\theta}}^{t+1} = \frac{1}{K}\sum_{l=1}^K \boldsymbol{w}_{i_l}^{t+1}$，对于方案 II，根据全期望公式，我们有

$$\begin{aligned}
\mathbb{E}\|\overline{\boldsymbol{\theta}}^{t+1} - \overline{\boldsymbol{w}}^{t+1}\|_2^2 &= \mathbb{E}\left[\mathbb{E}_{\mathcal{S}_{t+1}}\|\overline{\boldsymbol{\theta}}^{t+1} - \overline{\boldsymbol{w}}^{t+1}\|_2^2\right] \\
&= \mathbb{E}\left[\mathbb{E}_{\mathcal{S}_{t+1}}\left[\|\frac{1}{K}\sum_{l=1}^K \boldsymbol{w}_{i_l}^{t+1} - \overline{\boldsymbol{w}}^{t+1}\|_2^2\right]\right] \\
&= \mathbb{E}\left[\frac{1}{K^2}\mathbb{E}_{\mathcal{S}_{t+1}}\left[\|\sum_{i=1}^N \mathbf{1}\{i \in \mathcal{S}_{t+1}\}(\boldsymbol{w}_i^{t+1} - \overline{\boldsymbol{w}}^{t+1})\|_2^2\right]\right] \\
&\le \frac{1}{K^2}\mathbb{E}\left[\sum_{i=1}^N \mathbb{P}(i \in \mathcal{S}_{t+1})\|\boldsymbol{w}_i^{t+1} - \overline{\boldsymbol{w}}^{t+1}\|_2^2 + \sum_{j\neq i}\mathbb{P}(i,j \in \mathcal{S}_{t+1})\left\langle\boldsymbol{w}_i^{t+1} - \overline{\boldsymbol{w}}^{t+1},\ \boldsymbol{w}_j^{t+1} - \overline{\boldsymbol{w}}^{t+1}\right\rangle\right] \\
&= \frac{1}{KN}\sum_{i=1}^N \mathbb{E}\|\boldsymbol{w}_i^{t+1} - \overline{\boldsymbol{w}}^{t+1}\|_2^2 + \frac{K-1}{KN(N-1)}\sum_{i\neq j}\mathbb{E}\left\langle\boldsymbol{w}_i^{t+1} - \overline{\boldsymbol{w}}^{t+1},\ \boldsymbol{w}_j^{t+1} - \overline{\boldsymbol{w}}^{t+1}\right\rangle \\
&= \frac{N}{K(N-1)}\left(1 - \frac{K}{N}\right)\sum_{i=1}^N p_i \mathbb{E}\|\boldsymbol{w}_i^{t+1} - \overline{\boldsymbol{w}}^{t+1}\|_2^2
\end{aligned}$$

其中我们使用了以下等式：(1) 对所有 $i \neq j$，$\mathbb{P}(i \in S_{t+1}) = \frac{K}{N}$ 和 $\mathbb{P}(i,j \in S_{t+1}) = \frac{K}{(K-1)N(N-1)}$；以及 (2)$\sum_{i=1}^N \left\|\boldsymbol{w}_i^t - \boldsymbol{w}^t\right\|^2 + \sum_{i\neq j}\langle\boldsymbol{w}_i^{t+1} - \boldsymbol{w}^{t+1}, \boldsymbol{w}_j^{t+1} - \boldsymbol{w}^{t+1}\rangle = 0$。

结论由上述讨论得出。 □

引理 *C.4* 的证明。为方便起见，在本证明中，我们将针对 $t+1$ 进行讨论，其中假设 $t+1 \notin \mathcal{I}_E$，最后再转换回 t。首先，根据更新规则，当 $t+1 \notin \mathcal{I}_E$ 时，我们有

$$\boldsymbol{\theta}_i^{t+1} - \overline{\boldsymbol{\theta}}^{t+1} = \boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t - \eta_t(\nabla\ell(\boldsymbol{\theta}_i^t; Z_i^{t+1}) - g_t)$$

且当 $t+1 \in \mathcal{I}_E$，

$$\boldsymbol{w}_i^{t+1} - \overline{\boldsymbol{w}}^{t+1} = \boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t - \eta_t(\nabla\ell(\boldsymbol{\theta}_i^t; Z_i^{t+1}) - g_t).$$

然后，采用引理 B.2 中的相同方法，令 $t_0 := \max\{s \mid s < t+1, s \in \mathcal{I}_E\}$ 且 $c_3 := 12\sigma^2 + 18L^2(1+\epsilon_{\max})^2$，如果 $\eta_t^2 \leq \frac{1}{2c_3\ t+1-t_0\ \ ()\left(1+2(t+1-t_0)\right)}$，我们将得到：如果 $t+1 \notin \mathcal{I}_E$，

$$\begin{aligned}
&\sum_{i=1}^{N} p_i \mathbb{E}\|\boldsymbol{\theta}_i^{t+1} - \overline{\boldsymbol{\theta}}^{t+1}\|_2^2 \\
&\leq \sum_{s=0}^{t-t_0} \frac{t+2-t_0}{s+2} \eta_t^2 (2E+1)(48\sigma^2 + 36\varsigma^2) \mathbb{E}\|\overline{\boldsymbol{\theta}}^s - \boldsymbol{\theta}^{PS}\|_2^2 + \sum_{s=0}^{t-t_0} \frac{t+2-t_0}{s+2} \eta_t^2 (2E+1)(24\sigma^2 + 36\varsigma^2)
\end{aligned}$$

且如果 $t+1 \in \mathcal{I}_E$，

$$\begin{aligned}
&\sum_{i=1}^{N} p_i \mathbb{E}\|\boldsymbol{w}_i^{t+1} - \overline{\boldsymbol{w}}^{t+1}\|_2^2 \\
&\leq \sum_{s=0}^{t-t_0} \frac{t+2-t_0}{s+2} \eta_t^2 (2E+1)(48\sigma^2 + 36\varsigma^2) \mathbb{E}\|\overline{\boldsymbol{\theta}}^s - \boldsymbol{\theta}^{PS}\|_2^2 + \sum_{s=0}^{t-t_0} \frac{t+2-t_0}{s+2} \eta_t^2 (2E+1)(24\sigma^2 + 36\varsigma^2).
\end{aligned}$$

结合上述公式和引理 C.2，我们现在证明，若 η_0 足够小，则对于任意 t，都有 $\mathbb{E}\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{PS}\|_2^2 \leq \mathbb{E}\|\boldsymbol{\theta}^0 - \boldsymbol{\theta}^{PS}\|_2^2$。我们先推导下面的不等式，稍后将用到它。注意，对于任意满足 t_0 的 t：$= \max\{s \mid s < t+1, s \in \mathcal{I}_E\}$，都有

$$\sum_{s=0}^{t-t_0} \frac{t+2-t_0}{s+2} = (t+2-t_0)(\frac{1}{2} + \ldots + \frac{1}{t-t_0+2}) \leq (t+2-t_0)\log(t+2-t_0) \leq (E+1)\log(E+1).$$

接着，按照引理 B.2 中相同的归纳法，我们有：如果

$$\begin{aligned}
\eta_t &\leq \eta_0 \\
&\leq \frac{\tilde{\mu}\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2}{2\sigma^2 + (c_1c_3 + c_2/6)(2E^2+3E+1)\log(E+1)\big((16\sigma^2+12\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2 + (8\sigma^2+12\varsigma^2)\big)} = \tilde{\eta}_0,
\end{aligned}$$

那么对于任意 t，都有 $\mathbb{E}\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{PS}\|_2^2 \leq \mathbb{E}\|\boldsymbol{\theta}^0 - \boldsymbol{\theta}^{PS}\|_2^2$。

在所有这些条件下，如果 $t \notin \mathcal{I}_E$，则有

$$\begin{aligned}
\sum_{i=1}^{N} p_i \mathbb{E}\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2 &\leq \eta_{t-1}^2 (2E^2+3E+1)\log(E+1)(48\sigma^2+36\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2 \\
&\quad + \eta_{t-1}^2 (2E^2+3E+1)\log(E+1)(24\sigma^2+36\varsigma^2) \\
&\leq 4\eta_t^2 (2E^2+3E+1)\log(E+1)(48\sigma^2+36\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2 \\
&\quad + 4\eta_t^2 (2E^2+3E+1)\log(E+1)(24\sigma^2+36\varsigma^2),
\end{aligned}$$

并且，如果 $t+1 \in \mathcal{I}_E$，则有

$$\begin{aligned}
\sum_{i=1}^{N} p_i \mathbb{E}\|\boldsymbol{w}_i^{t+1} - \overline{\boldsymbol{w}}^{t+1}\|_2^2 &\leq \eta_{t-1}^2 (2E^2+3E+1)\log(E+1)(48\sigma^2+36\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\|_2^2 \\
&\quad + \eta_{t-1}^2 (2E^2+3E+1)\log(E+1)(24\sigma^2+36\varsigma^2).
\end{aligned}$$

注意，在引理 C.2 中，$t \notin \mathcal{I}_E$ 的不等式比 $t \in \mathcal{I}_E$ 的不等式更宽松。因此，我们可以对所有 t 应用 $t \in \mathcal{I}_E$ 的不等式。将此不等式与上述公式结合可得：

1. 对于方案 I,

$$\begin{aligned}
&\mathbb{E}[\|\overline{\boldsymbol{\theta}}^{t+1}-\boldsymbol{\theta}^{PS}\|_2^2]\\
&\leq \frac{1}{K}\sum_{k=1}^{N} p_k\mathbb{E}\|\boldsymbol{w}_k^{t+1}-\overline{\boldsymbol{w}}^{t+1}\|_2^2+(1-\tilde{\mu}\eta_t)\mathbb{E}\|\overline{\boldsymbol{\theta}}^{t}-\boldsymbol{\theta}^{PS}\|_2^2+2\sigma^2\eta_t^2+(c_1\eta_t+c_2\eta_t^2)\sum_{i=1}^{N}p_i\mathbb{E}\|\boldsymbol{\theta}_i^t-\overline{\boldsymbol{\theta}}^{t}\|_2^2\\
&\leq (1-\tilde{\mu}\eta_t)\mathbb{E}\|\overline{\boldsymbol{\theta}}^{t}-\boldsymbol{\theta}^{PS}\|_2^2+2\sigma^2\eta_t^2\\
&\quad+(4c_1\eta_t+4c_2\eta_t^2+K^{-1})(2E^2+3E+1)\log(E+1)\eta_t^2\bigg((48\sigma^2+36\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^{0}-\boldsymbol{\theta}^{PS}\|_2^2+(24\sigma^2+36\varsigma^2)\bigg),
\end{aligned}$$

2. 对于方案 II,

$$\begin{aligned}
&\mathbb{E}[\|\overline{\boldsymbol{\theta}}^{t+1}-\boldsymbol{\theta}^{PS}\|_2^2]\\
&\leq \frac{N-K}{KN(N-1)}\sum_{k=1}^{N} p_k\mathbb{E}\|\boldsymbol{w}_k^{t+1}-\overline{\boldsymbol{w}}^{t+1}\|_2^2+(1-\tilde{\mu}\eta_t)\mathbb{E}\|\overline{\boldsymbol{\theta}}^{t}-\boldsymbol{\theta}^{PS}\|_2^2+2\sigma^2\eta_t^2+(c_1\eta_t+c_2\eta_t^2)\sum_{i=1}^{N}p_i\mathbb{E}\|\boldsymbol{\theta}_i^t-\overline{\boldsymbol{\theta}}^{t}\|_2^2\\
&\leq (1-\tilde{\mu}\eta_t)\mathbb{E}\|\overline{\boldsymbol{\theta}}^{t}-\boldsymbol{\theta}^{PS}\|_2^2+2\sigma^2\eta_t^2\\
&\quad+\left(4c_1\eta_t+4c_2\eta_t^2+\frac{N-K}{K(N-1)}\right)(2E^2+3E+1)\log(E+1)\eta_t^2\bigg((48\sigma^2+36\varsigma^2)\mathbb{E}\|\overline{\boldsymbol{\theta}}^{0}-\boldsymbol{\theta}^{PS}\|_2^2+(24\sigma^2+36\varsigma^2)\bigg).
\end{aligned}$$

□

引理 *C.4* 的证明。结论可直接由引理 C.3 得出。 □

D 替代假设 (3) 下的收敛证明

假设 D.1。假设以下条件成立

$$\mathbb{E}_{Z_i\sim\mathcal{D}_i(\boldsymbol{\theta})}[\|\nabla\ell(\boldsymbol{\theta};Z_i)\|_2^2]\leq G^2. \tag{16}$$

引理 D.2。 （参数发散性的 $g^{ence\,o}\ fp$参数，即共识误差界）
当 $E>1$ 时，在假设 *2.1*、*2.2*、*2.3*、*2.5* 下，且如果 η_t 为非递增函数，并且对所有 $t\geq 0$ 都有 $\eta_t\leq 2\eta_{t+E}$，则有

$$\mathbb{E}\bigg[\sum_{i=1}^{N}p_i\|\boldsymbol{\theta}_i^t-\overline{\boldsymbol{\theta}}^{t}\|_2^2\bigg]\leq 4(E-1)^2\eta_t^2G^2 \tag{17}$$

证明。 FedAvg 每隔 E 步进行一次通信，因此对于任意 $t\geq 0$，存在一个 $t_0\leq t$， $t_0\in\mathcal{I}_E$，使得 $t-t_0\leq E-1$ 且 $\boldsymbol{\theta}_i^{t_0}=\overline{\boldsymbol{\theta}}^{t_0}$， $\forall i$。此外，我们利用以下事实：对于所有 $t-t_0\leq E-1$， $\eta_{t_0}\leq 2\eta_t$，于是

$$\mathbb{E}\bigg[\sum_{i=1}^{N}p_i\|\boldsymbol{\theta}_i^t-\overline{\boldsymbol{\theta}}^{t}\|_2^2\bigg]=\mathbb{E}\bigg[\sum_{i=1}^{N}p_i\|(\boldsymbol{\theta}_i^t-\overline{\boldsymbol{\theta}}^{t_0})-(\overline{\boldsymbol{\theta}}^{t}-\overline{\boldsymbol{\theta}}^{t_0})\|_2^2\bigg]\leq\mathbb{E}\bigg[\sum_{i=1}^{N}p_i\|\boldsymbol{\theta}_i^t-\overline{\boldsymbol{\theta}}^{t_0}\|_2^2\bigg], \tag{18}$$

因为 $\mathbb{E}\|X-\mathbb{E}X\|_2^2\leq\mathbb{E}\|X\|_2^2$ 其中 $X=\boldsymbol{\theta}_i^t-\overline{\boldsymbol{\theta}}^{t_0}$。利用詹森不等式，我们进一步得到

$$\|\boldsymbol{\theta}_i^t-\overline{\boldsymbol{\theta}}^{t_0}\|_2^2=\bigg\|\sum_{s=t_0}^{t-1}\eta_s\nabla\ell(\boldsymbol{\theta}_i^s;Z_i^{s+1})\bigg\|_2^2\leq(t-t_0)\sum_{s=t_0}^{t-1}\eta_s^2\big\|\nabla\ell(\boldsymbol{\theta}_i^s;Z_i^{s+1})\big\|_2^2, \tag{19}$$

$$\mathbb{E}\bigg[\sum_{i=1}^{N}p_i\|\boldsymbol{\theta}_i^t-\overline{\boldsymbol{\theta}}^{t_0}\|_2^2\bigg]=\mathbb{E}\bigg[\bigg\|\sum_{i=1}^{N}p_i\sum_{s=t_0}^{t-1}\eta_s^2\nabla\ell(\boldsymbol{\theta}_i^s;Z_i^{s+1})\bigg\|_2^2\bigg]\leq(t-t_0)\sum_{s=t_0}^{t-1}\eta_s^2\sum_{i=1}^{N}p_i\mathbb{E}\bigg[\big\|\nabla\ell(\boldsymbol{\theta}_i^s;Z_i^{s+1})\big\|_2^2\bigg], \tag{20}$$

$$\mathbb{E}\Big[\sum_{i=1}^{N} p_i \|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^{t_0}\|_2\Big] = \mathbb{E}\Big[\Big\|\sum_{i=1}^{N} p_i \sum_{s=t_0}^{t-1} \eta_s \nabla \ell(\boldsymbol{\theta}_i^s; Z_i^{s+1})\Big\|_2\Big] \leq \sum_{s=t_0}^{t-1} \eta_s \sum_{i=1}^{N} p_i \mathbb{E}\Big[\big\|\nabla \ell(\boldsymbol{\theta}_i^s; Z_i^{s+1})\big\|_2\Big], \tag{21}$$

其中我们使用了 $\eta_s \leq \eta_{t_0}$。因此，根据**A5**，我们得到

$$\begin{aligned}
\mathbb{E}\Big[\sum_{i=1}^{N} p_i \|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\|_2^2\Big] &\leq \sum_{i=1}^{N} p_i \mathbb{E}\Big[\sum_{s=t_0}^{t-1} (E-1)\eta_s^2 \|\nabla \ell(\boldsymbol{\theta}_i^s; Z_i^{s+1})\|_2^2\Big] \\
&\leq \sum_{i=1}^{N} p_i \Big[\sum_{s=t_0}^{t-1} (E-1)\eta_s^2 G^2\Big] \\
&\leq \sum_{i=1}^{N} p_i (E-1)^2 \eta_{t_0}^2 G^2 \\
&\leq 4(E-1)^2 \eta_t^2 G^2
\end{aligned} \tag{22}$$

因为 $\eta_s \leq \eta_{t_0} \leq 2\eta_{t_0+E} \leq 2\eta_t$ 在最后两个不等式中。 □

引理 D.3. *Li* 等人 *[2022]* 考虑一个非负且单调不增的步长序列 $\{\eta_t\}_{t\geq 1}$。令 $a > 0, p \in \mathbb{Z}_+$和 $\eta_1 < 2/a$。若对任意 $t \geq 1$，均有 $\eta_t^p / \eta_{t+1}^p \leq 1 + (a/2)\eta_{t+1}^p$，则

$$\sum_{j=1}^{t} \eta_j^{p+1} \prod_{\ell=j+1}^{t} (1 - \eta_\ell a) \leq \frac{2}{a} \eta_t^p, \quad \forall t \geq 1 \tag{23}$$

引理 D.4. 在假设 *2.1*、*2.2*、*2.3*、*D.1* 以及以下条件下： $\eta_t \leq \tilde{\mu}/c_2$，，$\eta_t \leq \eta_{t_0} \leq 2\eta_t$ 其中$t_0 =$最大值 $_s\{s \in \mathbb{N} | Es \leq t\}$, $\eta_{t+1} < \eta_t$对任意 $t \geq 0$, $\eta_1 < \frac{2}{\tilde{\mu}}$以及 $\eta_t^q / \eta_{t+1}^q \leq 1 + (\tilde{\mu}/2)\eta_{t+1}^q$对任意 $t \geq 0$和 $q = 1,2,3$。

$$\mathbb{E}[\|\overline{\boldsymbol{w}}^{t+1} - \boldsymbol{\theta}^{PS}\|_2^2] \leq \prod_{i=0}^{t} (1 - \tilde{\mu}\eta_i) \big\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\big\|_2^2 + \frac{2c_2c_7}{\tilde{\mu}} \eta_t^3 + \frac{2c_1c_7}{\tilde{\mu}} \eta_t^2 + \frac{4\sigma^2}{\tilde{\mu}} \eta_t, \tag{24}$$

其中 $c_7 := 4\,(E-1)\,G^2$。

证明。 由引理 B.1 可得

$$\begin{aligned}
\mathbb{E}\big[\|\overline{\boldsymbol{w}}^{t+1} - \boldsymbol{\theta}^{PS}\|_2^2\big] &\leq (1 - \tilde{\mu}\eta_t)\mathbb{E}\Big[\big\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\big\|_2^2\Big] + (c_1\eta_t + c_2\eta_t^2)\mathbb{E}\Big[\sum_{i=1}^{N} p_i \big\|\boldsymbol{\theta}_i^t - \overline{\boldsymbol{\theta}}^t\big\|_2^2\Big] + 2\sigma^2\eta_t^2 \\
&\leq (1 - \tilde{\mu}\eta_t)\mathbb{E}\Big[\big\|\overline{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^{PS}\big\|_2^2\Big] + c_2c_7\eta_t^4 + c_1c_7\eta_t^3 + 2\sigma^2\eta_t^2 \\
&= \prod_{i=0}^{t} (1 - \tilde{\mu}\eta_i) \big\|\overline{\boldsymbol{\theta}}^0 - \boldsymbol{\theta}^{PS}\big\|_2^2 + \sum_{s=1}^{t} \prod_{i=s+1}^{t} (1 - \tilde{\mu}\eta_i)\Big(c_2c_7\eta_s^4 + c_1c_7\eta_s^3 + 2\sigma^2\eta_s^2\Big).
\end{aligned} \tag{25}$$

第二个不等式成立是因为引理 D.2。利用引理 D.3,

$$\sum_{s=1}^{t} \prod_{i=s+1}^{t} (1 - \tilde{\mu}\eta_i)\Big(c_2c_7\eta_s^4 + c_1c_7\eta_s^3 + 2\sigma^2\eta_s^2\Big) \leq \frac{2c_2c_7}{\tilde{\mu}} \eta_t^3 + \frac{2c_1c_7}{\tilde{\mu}} \eta_t^2 + \frac{4\sigma^2}{\tilde{\mu}} \eta_t. \tag{26}$$

□

定理 D.5。（全参与收敛定理，替代假设）在假设 *2.1*、*2.2*、*2.3*、*2.5* 下，全参与方案的收敛速率为 $\mathcal{O}(\frac{1}{T})$，即，记 $\Delta_t := \mathbb{E}[\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{PS}\|_2^2]$，则对于某个 $\gamma > 0$，

$$\Delta_t \leq \frac{\upsilon}{\gamma + t}, \tag{27}$$

其中 $\upsilon :=$ 最大值$\{\frac{c_2c_7\beta^4\gamma^{-2}+c_1c_7\beta^3\gamma^{-1}+2\sigma^2\beta^2}{\beta\mu-1}, (\gamma+1)\Delta_1\}$。

证明。我们将针对部分参与算法给出证明，全参与情形的证明类似。

对于递减步长 $\eta_t = \frac{\beta}{t+\gamma}$，其中 $\beta > \frac{1}{\tilde{\mu}}$ 且 $\gamma > 0$，满足 $\eta_1 \leq$最小值$\{\frac{1}{\tilde{\mu}}, \frac{1}{4L}\} = \frac{1}{4L}$ 和 $\eta_t \leq 2\eta_{t+E}$ 我们将证明 $\Delta_t := \mathbb{E}[\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{PS}\|_2^2] \leq \frac{\upsilon}{\gamma+t}$，其中 $\upsilon :=$ max$\{\frac{c_2c_7\beta^4\gamma^{-2}+c_1c_7\beta^3\gamma^{-1}+2\sigma^2\beta^2}{\beta\mu-1}, (\gamma+1)\Delta_1\}$。我们通过归纳法证明。首先，根据 υ 的定义可知，该结论对 $t = 1$ 成立。假设该结论对某个 t 成立，即 $\eta_t = \frac{\beta}{t+\gamma}$，则根据引理 D.4 可得

$$\begin{aligned}
\Delta_{t+1} &\leq (1 - \eta_t\tilde{\mu})\Delta_t + c_2c_7\eta_t^4 + c_1c_7\eta_t^3 + 2\sigma^2\eta_t^2 \\
&\leq (1 - \frac{\beta\tilde{\mu}}{t+\gamma})\frac{\upsilon}{t+\gamma} + \frac{c_2c_7\beta^4}{(t+\gamma)^4} + \frac{c_1c_7\beta^3}{(t+\gamma)^3} + \frac{2\sigma^2\beta^2}{(t+\gamma)^2} \\
&= \frac{t+\gamma-1}{(t+\gamma)^2}\upsilon + \left[\frac{c_2c_7\beta^4}{(t+\gamma)^4} + \frac{c_1c_7\beta^3}{(t+\gamma)^3} + \frac{2\sigma^2\beta^2}{(t+\gamma)^2} - \frac{\beta\mu-1}{(t+\gamma)^2}\upsilon\right] \\
&\leq \frac{t+\gamma-1}{(t+\gamma)^2}\upsilon + \left[\frac{c_2c_7\beta^4}{(t+\gamma)^2\gamma^2} + \frac{c_1c_7\beta^3}{(t+\gamma)^2\gamma} + \frac{2\sigma^2\beta^2}{(t+\gamma)^2} - \frac{\beta\mu-1}{(t+\gamma)^2}\upsilon\right] \\
&\leq \frac{\upsilon}{t+\gamma+1}
\end{aligned} \tag{28}$$

其中 $\tilde{\mu}, c_1, c_2c_3, c_7$ 的定义与前面的证明相同，因此给出了 $\mathcal{O}(1/T)$ 收敛速率 . □

引理 D.6.（部分参与时差异 $\boldsymbol{w}^{t+1} - \boldsymbol{\theta}^{t+1}$ 的界）假设假设 *2.1*、*2.2*、*2.3* 和 *2.5* 成立。对于 $t+1 \in \mathcal{I}_E$，假设 η_t 非递增，且对于所有 t，$\eta_t \leq 2\eta_{t+E}$，则有以下结果

1. 对于方案 *I*，期望差 $\boldsymbol{w}^{t+1} - \boldsymbol{\theta}^{t+1}$ 有如下界

$$\mathbb{E}_{\mathcal{S}_t}\|\overline{\boldsymbol{\theta}}^{t+1} - \overline{\boldsymbol{w}}^{t+1}\|_2^2 \leq \frac{4}{K}\eta_t^2E^2G^2. \tag{29}$$

2. 对于方案 *II*，假设 $p_1 = p_2 = \cdots = p_N = \frac{1}{N}$，期望差 $\boldsymbol{w}^{t+1} - \boldsymbol{\theta}^{t+1}$ 有如下界

$$\mathbb{E}_{\mathcal{S}_t}\|\overline{\boldsymbol{\theta}}^{t+1} - \overline{\boldsymbol{w}}^{t+1}\|_2^2 \leq \frac{4(N-K)}{K(N-1)}\eta_t^2E^2G^2 \tag{30}$$

证明。我们证明方案 I 的界如下。由于 $\boldsymbol{\theta}^{t+1} = \frac{1}{K}\sum_{l=1}^K \boldsymbol{w}_{i_l}^t$，对 $\mathcal{S}_{t+1}$ 取期望可得

$$\mathbb{E}_{\mathcal{S}_t}\|\overline{\boldsymbol{\theta}}^{t+1} - \overline{\boldsymbol{w}}^{t+1}\|_2^2 = \mathbb{E}_{\mathcal{S}_t}\frac{1}{K^2}\sum_{l=1}^K\|\boldsymbol{w}_{i_l}^{t+1} - \overline{\boldsymbol{w}}^{t+1}\|_2^2 = \frac{1}{K}\sum_{k=1}^N p_k\|\boldsymbol{w}_k^{t+1} - \overline{\boldsymbol{w}}^{t+1}\|_2^2 \tag{31}$$

我们注意到，由于 $t+1 \in \mathcal{I}_E$，可知时间 $t_0 = t - E + 1 \in \mathcal{I}_E$ 是通信时间，这意味着 $\{\boldsymbol{\theta}_{t_0}^k\}$ 相同。接着

$$\sum_{k=1}^N p_k\|\boldsymbol{w}_k^{t+1} - \overline{\boldsymbol{w}}^{t+1}\|_2^2 = \sum_{i=1}^N p_k\|(\boldsymbol{w}_k^{t+1} - \overline{\boldsymbol{\theta}}^{t_0}) - (\overline{\boldsymbol{w}}^{t+1} - \overline{\boldsymbol{\theta}}^{t_0})\|_2^2 \leq \sum_{i=1}^N p_k\|\boldsymbol{w}_k^{t+1} - \overline{\boldsymbol{\theta}}^{t_0}\|_2^2 \tag{32}$$

与引理 D.2 类似，最后一个不等式是由 $\mathbb{E}\|X-\mathbb{E}X\|_2^2 \le \mathbb{E}\|X\|_2^2$ 得出的，其中 $X = \boldsymbol{w}_k^{t+1} - \boldsymbol{\theta}^{t_0}$，且 $\sum_{k=1}^N p_k(\boldsymbol{w}_k^{t+1} - \boldsymbol{\theta}^{t_0}) = \boldsymbol{w}^{t+1} - \boldsymbol{\theta}^{t_0}$。类似地，我们有

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}_t}\big[\|\overline{\boldsymbol{\theta}}^{t+1} - \overline{\boldsymbol{w}}^{t+1}\|_2^2\big] &\le \frac{1}{K}\sum_{k=1}^N p_k \mathbb{E}\big[\|\boldsymbol{w}_k^{t+1} - \overline{\boldsymbol{\theta}}^{t_0}\|_2^2\big] \\
&= \frac{1}{K}\sum_{k=1}^N p_k \mathbb{E}\big[\|\boldsymbol{w}_k^{t+1} - \boldsymbol{\theta}_k^{t_0}\|_2^2\big] \\
&= \frac{1}{K}\sum_{k=1}^N p_k \mathbb{E}\Big[\big\|\sum_{s=t_0}^{t}\eta_s \nabla l(\boldsymbol{\theta}_k^s; Z_k^{s+1})\big\|_2^2\Big] \\
&\le \frac{1}{K}\sum_{k=1}^N p_k E\sum_{s=t_0}^{t}\mathbb{E}\big[\|\eta_s \nabla l(\boldsymbol{\theta}_k^s; Z_k^{s+1})\|_2^2\big] \\
&\le \frac{1}{K}E^2\eta_{t_0}^2 G^2 \le \frac{4}{K}\eta_t^2 E^2 G^2. \qquad (33)
\end{aligned}$$

接下来，我们证明方案 II 的界。由于 $\boldsymbol{\theta}^{t+1} = \frac{1}{K}\sum_{l=1}^K \boldsymbol{w}_{i_l}^{t+1}$，我们有

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S}_t}\big[\|\overline{\boldsymbol{\theta}}^{t+1} - \overline{\boldsymbol{w}}^{t+1}\|_2^2\big] \\
&= \mathbb{E}_{\mathcal{S}_t}\Big[\big\|\frac{1}{K}\sum_{l=1}^K \boldsymbol{w}_{i_l}^{t+1} - \overline{\boldsymbol{w}}^{t+1}\big\|_2^2\Big] \\
&= \frac{1}{K^2}\mathbb{E}_{\mathcal{S}_t}\Big[\big\|\sum_{i=1}^N \mathbf{1}\{i\in\mathcal{S}_t\}(\boldsymbol{w}_i^{t+1} - \overline{\boldsymbol{w}}^{t+1})\big\|_2^2\Big] \\
&= \frac{1}{K^2}\Big[\sum_{i=1}^N \mathbb{P}(i\in\mathcal{S}_t)\|\boldsymbol{w}_i^{t+1} - \overline{\boldsymbol{w}}^{t+1}\|_2^2 + \sum_{j\neq i}\mathbb{P}(i,j\in\mathcal{S}_t)\big\langle \boldsymbol{w}_i^{t+1} - \overline{\boldsymbol{w}}^{t+1},\ \boldsymbol{w}_j^{t+1} - \overline{\boldsymbol{w}}^{t+1}\big\rangle\Big] \\
&= \frac{1}{KN}\sum_{i=1}^N \|\boldsymbol{w}_i^{t+1} - \overline{\boldsymbol{w}}^{t+1}\|_2^2 + \frac{K-1}{KN(N-1)}\sum_{i\neq j}\big\langle \boldsymbol{w}_i^{t+1} - \overline{\boldsymbol{w}}^{t+1},\ \boldsymbol{w}_j^{t+1} - \overline{\boldsymbol{w}}^{t+1}\big\rangle \\
&= \frac{1}{K(N-1)}(1-\frac{K}{N})\sum_{i=1}^N \|\boldsymbol{w}_i^{t+1} - \overline{\boldsymbol{w}}^{t+1}\|_2^2. \qquad (34)
\end{aligned}$$

注意，倒数第二个等式成立是因为 $\mathbb{P}(i\in\mathcal{S}_t) = \frac{K}{N}$ 和 $\mathbb{P}(i,\ j\in\mathcal{S}_t) = \frac{K(K-1)}{N(N-1)}$；最后一个等式成立是因为

$$\begin{aligned}
&\sum_{i=1}^N \|\boldsymbol{w}_i^{t+1} - \overline{\boldsymbol{w}}^{t+1}\|_2^2 + \sum_{i\neq j}\big\langle \boldsymbol{w}_i^{t+1} - \overline{\boldsymbol{w}}^{t+1},\ \boldsymbol{w}_j^{t+1} - \overline{\boldsymbol{w}}^{t+1}\big\rangle \\
&= \sum_{i=1}^N \Big\langle \boldsymbol{w}_i^{t+1} - \overline{\boldsymbol{w}}^{t+1},\ \Big(\sum_{j=1}^N \boldsymbol{w}_j^{t+1}\Big) - N\overline{\boldsymbol{w}}^{t+1}\Big\rangle = 0.
\end{aligned}$$

回顾一下

$$\sum_{k=1}^N p_k\|\boldsymbol{w}_k^{t+1} - \overline{\boldsymbol{w}}^{t+1}\|_2^2 \le \sum_{i=1}^N p_k\|\boldsymbol{w}_k^{t+1} - \overline{\boldsymbol{\theta}}^{t_0}\|_2^2,$$

我们得到

$$\begin{aligned}\mathbb{E}\big[\|\overline{\boldsymbol{\theta}}^{t+1}-\overline{\boldsymbol{w}}^{t+1}\|_2^2\big] &= \frac{1}{K(N-1)}(1-\frac{K}{N})\mathbb{E}\bigg[\sum_{i=1}^{N}\|\boldsymbol{w}_i^{t+1}-\overline{\boldsymbol{w}}^{t+1}\|_2^2\bigg]\\ &\leq \frac{N}{K(N-1)}(1-\frac{K}{N})\mathbb{E}\bigg[\sum_{i=1}^{N}\frac{1}{N}\|\boldsymbol{w}_i^{t+1}-\overline{\boldsymbol{\theta}}^{t_0}\|_2^2\bigg]\\ &\leq \frac{N}{K(N-1)}(1-\frac{K}{N})4\eta_t^2E^2G^2=\frac{4(N-K)}{K(N-1)}\eta_t^2E^2G^2 \end{aligned} \tag{35}$$

最后一个不等式可见引理 D.2 的证明中的 (22)。 □

引理 D.7。 在假设 *2.1*、*2.2*、*2.3*、*2.5* 以及条件 $\eta_t\leq\tilde{\mu}/c_2$、, $\eta_t\leq\eta_{t_0}\leq 2\eta_t$ *where*$t_0=$最大$_s\{s\in\mathbb{N}|Es\leq t\}$, $\eta_{t+1}<\eta_t$*for any* $t\geq 0$, $\eta_1<\frac{2}{\tilde{\mu}}$ *and* $\eta_t^q/\eta_{t+1}^q\leq 1+(\tilde{\mu}/2)\eta_{t+1}^q-$*for any* $t\geq 0$ *and* $q=1,2,3$ 下，我们有

$$\mathbb{E}_{\mathcal{S}_t}[||\overline{\boldsymbol{\theta}}^{t+1}-\boldsymbol{\theta}^{PS}||_2^2]\leq\prod_{i=0}^{t}(1-\tilde{\mu}\eta_i)\big|\big|\overline{\boldsymbol{\theta}}^{0}-\boldsymbol{\theta}^{PS}\big|\big|_2^2+\frac{2c_2c_3}{\tilde{\mu}}\eta_t^3+\frac{2c_1c_3}{\tilde{\mu}}\eta_t^2+\frac{2c_8}{\tilde{\mu}}\eta_t, \tag{36}$$

(c_8 对于方案 *I*，在方案 *II* 中将 c_8 替换为 c_9 *)*，其中我们定义 $c_8:=2\sigma^2+\frac{4}{K}E^2G^2$ 在方案 *I* 中，以及 $c_9:=2\sigma^2+\frac{4(N-K)}{K(N-1)}E^2G^2$ 在方案 *II* 中。

证明。 注意到

$$\begin{aligned}&||\overline{\boldsymbol{\theta}}^{t+1}-\boldsymbol{\theta}^{PS}||_2^2\\ =&||\overline{\boldsymbol{\theta}}^{t+1}-\overline{\boldsymbol{w}}^{t+1}+\overline{\boldsymbol{w}}^{t+1}-\boldsymbol{\theta}^{PS}||_2^2\\ =&\underbrace{||\overline{\boldsymbol{\theta}}^{t+1}-\overline{\boldsymbol{w}}^{t+1}||_2^2}_{T_1}+\underbrace{||\overline{\boldsymbol{w}}^{t+1}-\boldsymbol{\theta}^{PS}||_2^2}_{T_2}+\underbrace{2\langle\overline{\boldsymbol{w}}^{t+1}-\overline{\boldsymbol{\theta}}^{t+1},\overline{\boldsymbol{\theta}}^{t+1}-\boldsymbol{\theta}^{PS}\rangle}_{T_3}\end{aligned} \tag{37}$$

当期望是关于 $\mathcal{S}_{t+1}$取得时，最后一项 T_3 由于引理 C.1而消失。
如果 $t+1\notin\mathcal{I}_E$，则当 $t+1$ 不是通信步骤时，根据定义， $\boldsymbol{\theta}^{t+1}=\boldsymbol{w}^{t+1}$，因此 T_1 消失。对于项 T_2，不难看出，我们可以使用引理 B.1 推导出其单步界（并在引理 D.4 中使用（25）），从而得到

$$\begin{aligned}\mathbb{E}\big[||\overline{\boldsymbol{\theta}}^{t+1}-\boldsymbol{\theta}^{PS}||_2^2\big]&=\mathbb{E}\big[||\overline{\boldsymbol{w}}^{t+1}-\boldsymbol{\theta}^{PS}||_2^2\big]\\ &\leq(1-\tilde{\mu}\eta_t)\mathbb{E}\big|\big|\overline{\boldsymbol{\theta}}^{t}-\boldsymbol{\theta}^{PS}\big|\big|_2^2+c_2c_7\eta_t^4+c_1c_7\eta_t^3+2\sigma^2\eta_t^2,\end{aligned} \tag{38}$$

并且我们回顾 $c_1:=\frac{L^2(1+\epsilon_{max})}{2\delta\epsilon}^2$， $c_2:=4[\sigma^2+L^2(1+\epsilon_{max})^2]$， $c3:=4(E-1)^2G^2$,- $\tilde{\mu}:=\mu-(1+\delta)\epsilon L$。如果 $t+1\in\mathcal{I}_E$，则根据引理 D.6 可得以下结果，

$$\begin{aligned}&\mathbb{E}\big[||\overline{\boldsymbol{\theta}}^{t+1}-\boldsymbol{\theta}^{PS}||_2^2\big]\\ =&\mathbb{E}\big[||\overline{\boldsymbol{\theta}}^{t+1}-\overline{\boldsymbol{w}}^{t+1}||_2^2\big]+\mathbb{E}\big[||\overline{\boldsymbol{w}}^{t+1}-\boldsymbol{\theta}^{PS}||_2^2\big]\\ \leq&(1-\tilde{\mu}\eta_t)\mathbb{E}\big|\big|\overline{\boldsymbol{\theta}}^{t}-\boldsymbol{\theta}^{PS}\big|\big|_2^2+c_2c_7\eta_t^4+c_1c_7\eta_t^3+c_8\eta_t^2,\end{aligned} \tag{39}$$

wh在此回顾 $c_8:=2\sigma^2+\frac{4}{K}E^2G^2$ 在方案 I 中，以及 $c_9:=2\sigma^2+\frac{4(N-K)}{K(N-1)}E^2G^2$ 在方案 II 中。

（26）与（39）之间的唯一差异在于（$c_8 - 2\sigma^2$）η_t^2。因此，我们可以采用类似的技术来证明收敛性，

$$\begin{aligned}\mathbb{E}[||\overline{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^{PS}||_2^2] &\leq (1-\tilde{\mu}\eta_t)\mathbb{E}\big|\big|\overline{\boldsymbol{\theta}}^{t} - \boldsymbol{\theta}^{PS}\big|\big|_2^2 + c_2c_7\eta_t^4 + c_1c_7\eta_t^3 + c_8\eta_t^2 \\ &= \prod_{i=0}^{t}(1-\tilde{\mu}\eta_i)\big|\big|\overline{\boldsymbol{\theta}}^{0} - \boldsymbol{\theta}^{PS}\big|\big|_2^2 \\ &\quad + \sum_{s=1}^{t}\prod_{i=s+1}^{t}(1-\tilde{\mu}\eta_i)\Big(c_2c_7\eta_s^4 + c_1c_7\eta_s^3 + c_8\eta_s^2\Big) \\ &\leq \prod_{i=0}^{t}(1-\tilde{\mu}\eta_i)\big|\big|\overline{\boldsymbol{\theta}}^{0} - \boldsymbol{\theta}^{PS}\big|\big|_2^2 + \frac{2c_2c_7}{\tilde{\mu}}\eta_t^3 + \frac{2c_1c_7}{\tilde{\mu}}\eta_t^2 + \frac{2c_8}{\tilde{\mu}}\eta_t\end{aligned}$$

（对于方案 I，将 c_8 替换为 c_9；对于方案 II，则将 c_8 替换为 c_9。）

□

定理 D.8.（全参与收敛定理，替代假设）在假设 *2.1*、*2.2*、*2.3*、*2.5* 下，全参与方案的收敛速率为 $\mathcal{O}(\frac{1}{T})$，即令 $\Delta_t := \mathbb{E}[\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{PS}\|_2^2]$，则存在某个 $\gamma > 0$，

$$\Delta_t \leq \frac{\upsilon}{\gamma + t}, \tag{40}$$

其中 $\upsilon := \max\{\frac{c_2c_7\beta^4\gamma^{-2} + c_1c_7\beta^3\gamma^{-1} + c_8\beta^2}{\beta\mu - 1}, (\gamma+1)\Delta_1\}$（对于方案 *I*，将 c_8 替换为 c_9；对于方案 *II*，则将 c_8 替换为 c_9。）

证明。我们将针对部分参与算法进行证明；全参与情形的证明类似。

对于递减步长 $\eta_t = \frac{\beta}{t+\gamma}$，其中 ~~$\beta > \frac{1}{\tilde{\mu}}$~~、$\gamma > 0$ 满足 $\eta_1 \leq \min\{\frac{1}{\tilde{\mu}}, \frac{1}{4L}\} = \frac{1}{4L}$，且 ~~$\eta_t \leq 2\eta_{t+E}$~~。我们将证明 $\triangle_t := \mathbb{E}[||\boldsymbol{\theta}^t - \boldsymbol{\theta}^{PS}||_2^2] \leq \frac{\upsilon}{\gamma+t}$，其中 $\upsilon := \max\{\frac{c_2c_7\beta^4\gamma^{-2} + c_1c_7\beta^3\gamma^{-1} + c_8\beta^2}{\beta\mu - 1}, (\gamma+1)\triangle_1\}$。我们通过归纳法证明这一点。首先，$\upsilon$ 的定义确保该结论对 $t = 1$成立。假设该结论对某个 t成立，即 $\eta_t = \frac{\beta}{t+\gamma}$，则可推出

$$\begin{aligned}\triangle_{t+1} &\leq (1-\eta_t\tilde{\mu})\triangle_t + c_2c_7\eta_t^4 + c_1c_7\eta_t^3 + c_8\eta_t^2 \\ &\leq (1 - \frac{\beta\tilde{\mu}}{t+\gamma})\frac{\upsilon}{t+\gamma} + \frac{c_2c_7\beta^4}{(t+\gamma)^4} + \frac{c_1c_7\beta^3}{(t+\gamma)^3} + \frac{c_8\beta^2}{(t+\gamma)^2} \\ &= \frac{t+\gamma-1}{(t+\gamma)^2}\upsilon + \left[\frac{c_2c_7\beta^4}{(t+\gamma)^4} + \frac{c_1c_7\beta^3}{(t+\gamma)^3} + \frac{c_8\beta^2}{(t+\gamma)^2} - \frac{\beta\mu-1}{(t+\gamma)^2}\upsilon\right] \\ &\leq \frac{t+\gamma-1}{(t+\gamma)^2}\upsilon + \left[\frac{c_2c_7\beta^4}{(t+\gamma)^2\gamma^2} + \frac{c_1c_7\beta^3}{(t+\gamma)^2\gamma} + \frac{c_8\beta^2}{(t+\gamma)^2} - \frac{\beta\mu-1}{(t+\gamma)^2}\upsilon\right] \\ &\leq \frac{\upsilon}{t+\gamma+1}\end{aligned} \tag{41}$$

（c_8 对于方案 I，将 c_8 替换为方案 II 中的 c_9），其中，$\tilde{\mu}$至 c_9 的定义与前文证明中的相同，因此可得 $\mathcal{O}$（$1/T$）收敛速率。

□

E 实验

E.1 数值模拟

E.1.1 学习率衰减（图6）

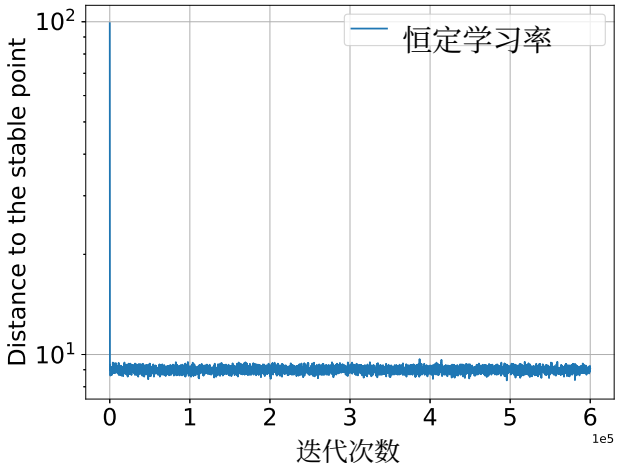


图6：ExpGaussian 上的恒定学习率。采用全参与，$E = 10$。学习率设为 0.02。

E.1.2 较低学习率下的方案 II（图7）

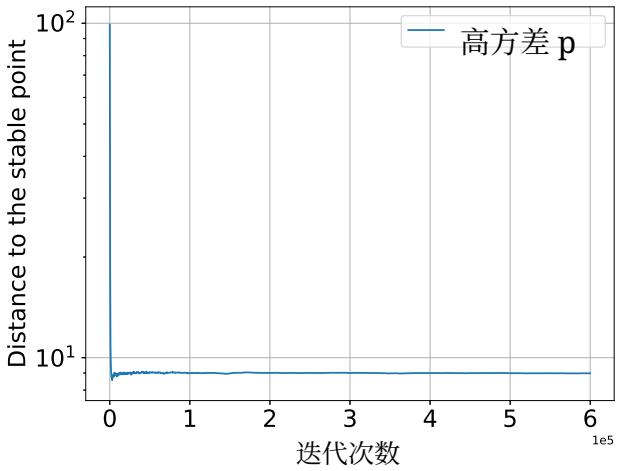
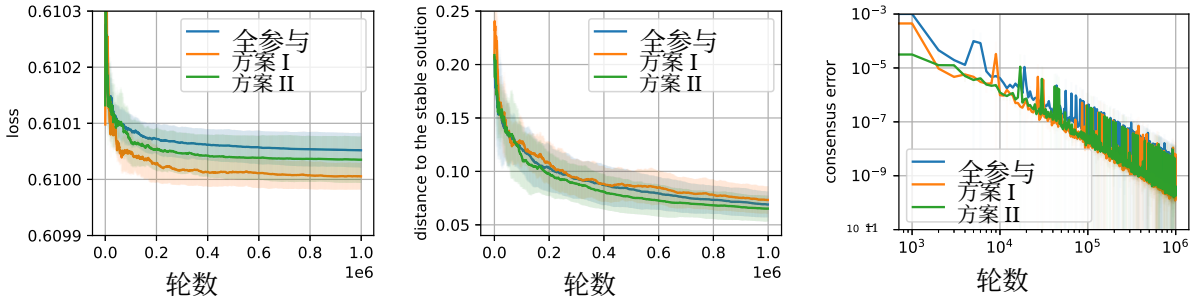


图7：ExpGaussian 上的无放回抽样。p_i 的方差设为 0.01。学习率设为 $\frac{5}{t+10000}$。

E.2 信用评分策略性分类

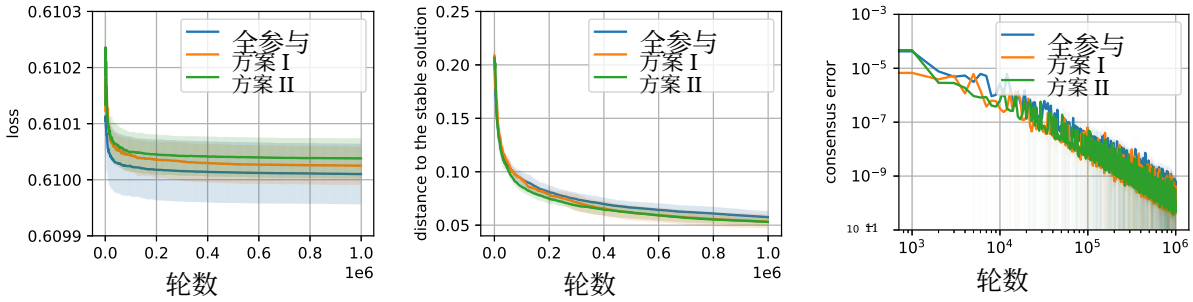
图8展示了使用批量大小为 1 的客户端梯度下降时，全参与、方案 I 和方案 II 相对于 PS 解的损失和距离。图9和图10分别展示了批量大小为 4 和 16 时的相同结果。为便于比较，y 轴的刻度设置为相同。增大批量大小可以提高三种方案的收敛速度，尤其是两种部分参与方案；当批量大小为 16 时，两者的收敛速度都与全参与一样快。

为研究批量大小如何影响收敛性，我们使用 $\boldsymbol{\theta}^{PS}$（使表演性目标函数最小化的解）初始化 P-FedAvg。由于小批量随机下降具有随机性，我们预计该参数将



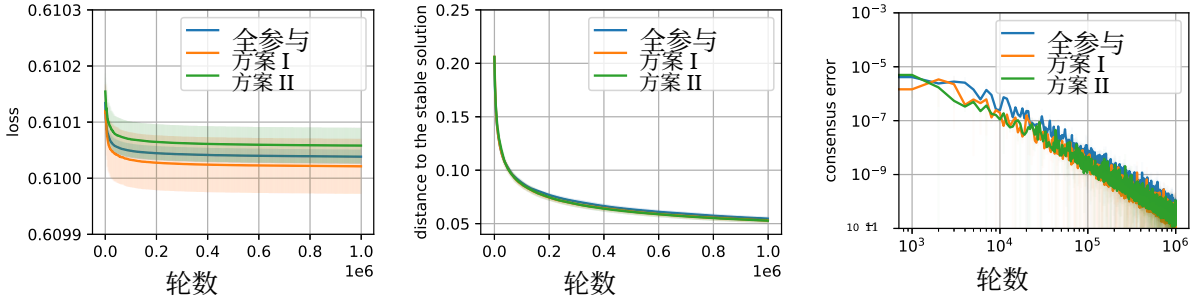
(a) (b) (c)

图8：采用批量大小为 1 的客户端梯度下降时，全参与、方案 I 和方案 II 对应的 PS 解的损失 (a) 和距离 (b)。



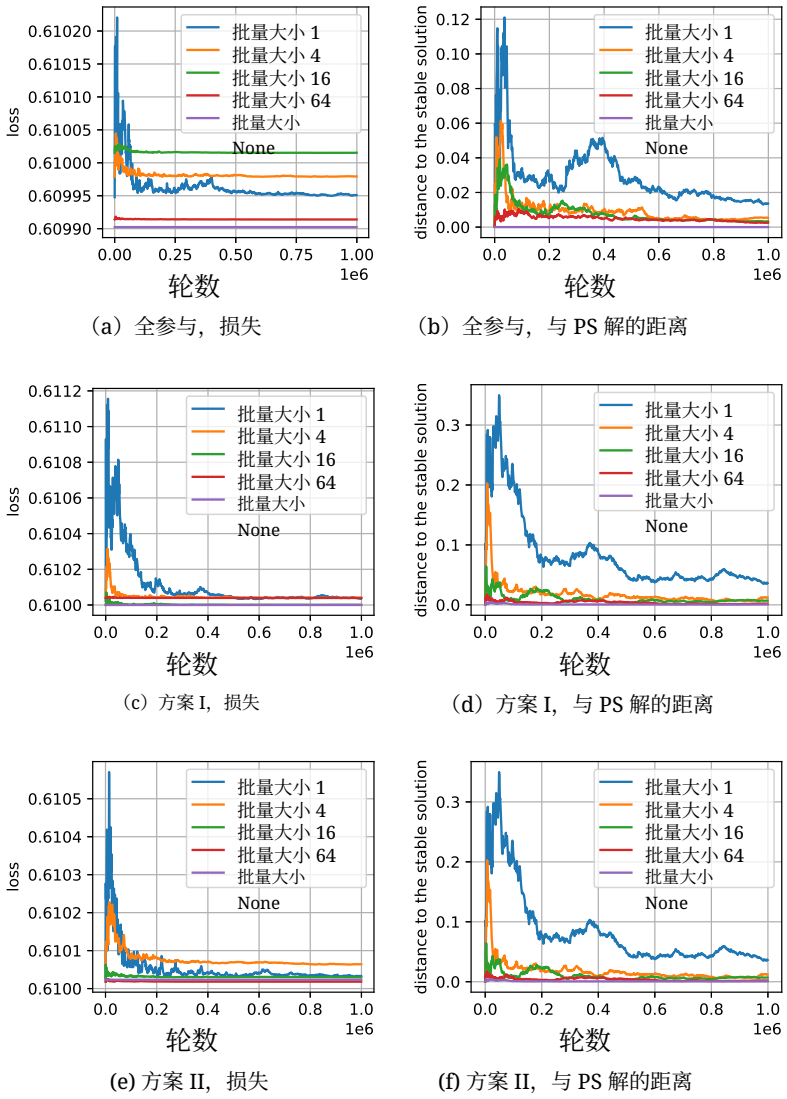
(a) (b) (c)

图10：与图8相同，但使用批量大小 4。



(a) (b) (c)

图10：与图8相同，但使用批量大小 16。



（a）全参与，损失　（b）全参与，与 PS 解的距离

（c）方案 I，损失　（d）方案 I，与 PS 解的距离

(e) 方案 II，损失　(f) 方案 II，与 PS 解的距离

图 11：使用 $\boldsymbol{\theta}^{PS}$ 初始化的 P-FedAvg 的损失函数及其与 $\boldsymbol{\theta}^{PS}$ 的距离。

随着算法以递减的步长继续运行，偏离 $\boldsymbol{\theta}^{PS}$ 并逐渐稳定回到 $\boldsymbol{\theta}^{PS}$。从图 11 (b)、(c) 和 (f) 可以看出，批量大小大于 1 时确实如此。这也是我们选择大于 1 的批量大小的原因。