数据流联邦学习

Othmane Marfoq Giovanni Neglia
Inria，蔚蓝海岸大学，Inria，蔚蓝海岸大学，埃森哲实验室，法国索菲亚·安蒂波利斯，法国索菲亚·安蒂波利斯，法国

Laetitia Kameni
埃森哲实验室，法国索菲亚·安蒂波利斯

Richard Vidal
埃森哲实验室，法国索菲亚·安蒂波利斯

arXiv:2301.01542v1 [cs.LG] 4 Jan 2023

摘要

联邦学习（FL）是一种有效的解决方案，能够在数据仍保留于本地的情况下，利用物联网设备和智能手机生成的海量数据训练机器学习模型。以往大多数关于联邦学习的研究都假设客户端使用的是训练开始前收集的静态数据集。这种方法可能效率低下，原因有二：1）它忽略了客户端在训练期间收集的新样本；2）它可能要求客户端先经历一个漫长的准备阶段，以收集足够的数据。此外，在设备总存储容量较小的场景中，基于静态数据集进行学习甚至可能无法实现。因此，有必要设计能够从数据流中学习的联邦算法。在本研究中，我们对数据流联邦学习问题进行了形式化并加以研究。我们提出了一种通用的联邦学习算法，通过适当的加权经验风险最小化从数据流中学习。我们的理论分析为如何配置该算法提供了见解，并且我们在广泛的机器学习任务上评估了其性能。

1 引言

联邦学习（McMahan 等人，2017年）通常涉及最小化一个目标函数，而该目标函数只能通过其梯度的无偏估计获得（Bottou 等人，2018年）。当客户端能够在每次迭代中采样新的数据点时，目标函数是期望风险；当客户端依赖固定数据集时，目标函数则是经验风险。

大多数关于联邦学习的早期研究，例如（McMa- han et al., 2017年；Koneˇcny et al., 2016年），都关注第二种情况，即最小化经验风险。这些研究假设每个客户端在训练开始前收集并存储所有样本。在许多情况下，在静态数据集上进行学习可能不是最优的（甚至是不可能的），因为：（1）训练期间收集的新样本会被忽略；（2）客户端的存储容量可能有限，无法存储大量数据样本。例如，传感器网络中的节点可能会持续收集新的测量数据，但只能将其中少量数据存储在本地内存中（De Francisci Morales et al., 2016年）。

我们的贡献。 本文提出并研究了从独立数据流中学习的问题。我们提出并从理论上分析了一种旨在实现这一目标的通用联邦算法。分析揭示了偏差与优化之间的权衡：通过控制较旧样本相对于较新样本的重要性，可以加快训练速度，但代价是学习得到的模型偏差更大；也可以降低偏差，但代价是训练时间更长。分析还为如何以最优方式配置我们的联邦算法提供了见解。我们通过涵盖广泛机器学习任务的仿真，展示了理论结果的相关性。具体而言，实验表明，将 FedAvg（McMahan 等人，2017年）扩展到数据流的“合理”方法可能会导致学习得到的模型表现不佳，而我们的配置规则则始终能带来近乎最优的性能。

论文结构。 本文其余部分安排如下。第 2 节回顾相关工作。第 3 节阐述数据流联邦学习问题。第 4 节介绍我们针对数据流提出的联邦学习算法，并给出其收敛结果。第 5 节研究一个具有实际意义的场景，并利用第 4 节的理论结果为我们的算法提供配置规则。最后，我们在第 6 节给出实验结果，并在第 7 节总结全文。

2 相关工作

自开创性研究（Koneˇcny 等人，2016年；McMahan 等人，2017年）提出以来，联邦学习作为一种具有发展前景的大规模分布式学习框架，日益受到关注，并已应用于广泛的任务，包括语言建模（Yang 等人，2018年）、自动语音识别（Gao 等人，2022）、医学影像（Courtiol 等人，2019；Silva 等人，2019），以及推荐系统（Yang 等人，2020年）。我们对数据流的关注，是我们与大多数联邦学习文献的一个关键区别；后者假设客户端拥有静态数据集。特别是，研究联邦学习算法在非独立同分布数据上的收敛性以及客户端部分参与（Li 等人，2019）、PAC 学习界（Mohri 等人，2019）、隐私保证（Wei 等人，2020）或对拜占庭故障的抵抗能力（Blanchard 等人，2017）的理论工作，都采用了这一假设，研究对象为非独立同分布数据和在客户端部分参与情况下的联邦学习算法收敛性、

独立同分布数据，以及客户端部分参与（Li 等人，2019）、PAC 学习界（Mohri 等人，2019）、隐私保证（Wei 等人，2020）或对拜占庭故障的抵抗能力（Blanchard 等人，2017）。

关于从数据流中学习的研究文献十分丰富，应用领域包括金融业（Zhu and Shasha, 2002）、网络监控（Babu and Widom, 2001）以及传感器网络（De Francisci Morales et al., 2016）。在这一领域，我们大致可以根据对数据过程的不同假设，将研究划分为三条主要路线。第一条关注这样一种情形：数据流中的样本独立地取自某个固定但未知的分布；这种情形可以通过随机逼近进行分析（Moulines and Bach, 2011）。第二条路线允许数据分布随时间变化，因此属于持续学习的范畴。在持续学习中，模型针对一系列任务进行训练，每项任务都可以对应不同的数据分布（Thrun, 1994; Kumar and Daum´e III, 2012; Ruvolo and Eaton, 2013; Kirkpatrick et al., 2017; Schwarz et al., 2018）。最后，第三条路线不对数据流作任何假设，并可将其视为由对手生成。这种情形可以在具有遗憾保证的在线学习框架下进行研究（Zinkevich, 2003）。本文假设每个客户端的数据均来自同一分布。与上述论文不同，我们研究的是从多个数据流中学习，且这些数据流的样本生成速率和客户端内存大小各不相同。

几乎没有研究对数据流联邦学习问题进行形式化并提供理论分析。据我们所知，唯有 Chen 等人（2020）、Yoon 等人（2021）以及 Odeyomi 和 Zaruba（2021）是例外。Chen 等人（2020）提出 ASO-Fed，这是一种异步联邦学习算法，旨在最小化基于客户端数据流聚合计算出的经验损失。他们的分析要求所有客户端具有相同的最优模型，并且任意时刻的更新 t 都与未来到达的新样本一致（更多详情见附录 A）。相比之下，我们论文中的理论分析适用于客户端本地数据分布存在统计异质性的情况，并考虑了因必须使用客户端当前存储的样本而产生的偏差。此外，我们还为算法提供了统计学习保证。Yoon 等人（2021）提出 FedWeIT，将基于正则化的算法扩展到持续学习的联邦学习场景。FedWeIT 的主要目标是在学习过程中减少不兼容任务之间的干扰，同时促进客户端之间的正向知识迁移，但没有提供泛化保证。Odeyomi 和 Zaruba（2021）研究了受整个时间范围内资源消耗量限制的在线联邦学习问题，并提出了一种基于在线镜像下降的算法

该算法具有遗憾保证。与我们的工作不同，Odeyomi 和 Zaruba （ 2021 ）以及 Yoon 等人（ 2021 ）都假设每个客户端只能使用最新的数据。我们的实验表明，重复利用仅占已收集样本 5% 的数据也可能带来显著收益。

从随时间变化的分布中进行联邦学习（Zhu 等人，2022；Eichner 等人，2019；Ding 等人，2020；Guo 等人，2021）与从数据流中学习相关，但并非同一问题。这些论文假设分布变化是由可用客户端集合的变化导致的（例如昼夜模式所致），但客户端的本地数据集不会变化。唯一的例外是 Guo 等人（2021），该研究能够刻画一种情形：客户端在训练期间持续收集数据，且不受存储限制。理论结果假设新数据来自客户端无关分布（见附录 A）。相比之下，我们的分析同时考虑了内存限制以及不同客户端本地数据分布之间的统计异质性。

最后，我们提及若干研究不同变体“在线联邦学习”问题的论文，这些研究大多聚焦于动态资源分配。近期综述（Dai 和 Meng，2022）讨论了其中许多论文。在这些论文中，Damaskinos 等人（2020）提出了 Fleet，这是一种位于边缘设备操作系统与机器学习应用之间的中间件，可用于在数据流上进行学习。该中间件的设计主要考虑设备的能量最小化。Jin 等人（2020）提出了一种在线算法，用于在每个通信轮次动态选择参与的客户端，并确定其本地梯度迭代次数，以在最终模型质量受到约束的条件下，最小化随时间累积的资源使用量。Zhou 等人（2020）研究了类似问题。他们考虑了丢弃新数据点或将其分配给资源更充足的客户端的可能性，并提出了一种基于李雅普诺夫优化（Neely，2010）的资源分配算法。Jin 等人（2020）和 Zhou 等人（2020）都没有考虑跨多个通信轮次重复使用样本的可能性。

3 问题 formulation

在本文中，我们用 $[M] \triangleq \{1, \ldots, M\}$表示不大于 M的正整数集合。我们考虑 $M > 0$ 个客户端；每个客户端对应一个可能不同的学习任务。我们为每个客户端 $m \in [M]$关联以下内容：1）域 $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$上的概率分布 $\mathcal{P}_m$，2）计数过程 $N_m^{(t)}$，$t \geq 0$，以及 3）容量为 $C_m > 0$的动态内存/缓存 $\mathcal{M}_m^{(t)}$，$t > 0$。在时间步 $t > 0$，客户端 $m \in [M]$ 会收到一个批次 $\mathcal{B}_m^{(t)} = \left\{ \boldsymbol{z}_m^{(t,i)} = \left(\boldsymbol{x}_m^{(t,i)}, y_m^{(t,i)} \right), \ i \in [b_m^{(t)}] \right\}$，其中包含 $b_m^{(t)} \triangleq N_m^{(t)} - N_m^{(t-1)}$个从 $\mathcal{P}_m$中独立同分布抽取的样本。客户端$m \in [M]$ 可以将其中一部分样本缓存在本地

内存中，但不能超过容量 C_m。不失一般性，我们假设 $1 \leq b_m^{(t)} \leq C_m$。我们考虑有限的时间范围 $T > 0$，并令 $N_m \triangleq N_{(T)m}$和 $\mathcal{S}_m \triangleq \bigcup_{t=1}^{T} \mathcal{B}_m^{(t)}$分别表示客户端 m 在该时间范围内收集的样本数量和样本集合。我们记作 $\mathcal{S}_m = \left\{\boldsymbol{z}_m^{(i)},\ i \in [N_m]\right\}$，其中我们任意排列了 $\mathcal{S}_m$中的元素。我们将 $\mathcal{I}_m^{(t)} \subset [N_m]$ 定义为内存 $\mathcal{M}_m^{(t)}$中样本的索引集合，即，当且仅当 $\boldsymbol{z}_m^{(j)} \in \mathcal{M}_m^{(t)}$时，$j \in \mathcal{I}_m^{(t)}$。最后，$\mathcal{S} \triangleq \bigcup_{m=1}^{M} \mathcal{S}_m$ 表示训练数据集（汇总各客户端在各时间步收集的数据），其大小为 $N \triangleq \sum_{m=1}^{M} N_m$。客户端 m 的数据集相对大小为 $n_m \triangleq N_m/N$。

令 $\mathcal{H}_\Theta = \left\{h_\theta \colon \mathcal{X} \mapsto \mathcal{Y}, \theta \in \Theta \subset \mathbb{R}^d\right\}$ 为一组参数化假设/模型，将 $\mathcal{X}$ 映射到 $\mathcal{Y}$，并令 $\ell \colon \Theta \times \mathcal{Z} \mapsto \mathbb{R}^+$ 为损失函数。我们将 $\mathcal{L}_{\mathcal{P}}(\theta) \triangleq \mathbb{E}_{\boldsymbol{z}\sim\mathcal{P}}\left[\ell(\theta; \boldsymbol{z})\right]$ 定义为假设$h_\theta \in \mathcal{H}_\Theta$ 在通用概率分布 $\mathcal{P}$（定义于$\mathcal{Z}$ 上）下的真实（期望）风险，并将 $\mathcal{L}_{\mathcal{S}}(\theta) = \frac{1}{|\mathcal{S}|}\sum_{(\boldsymbol{x},y)\in\mathcal{S}} \ell(\theta; \boldsymbol{z})$ 定义为模型（假设）$h_\theta \in \mathcal{H}_\Theta$ 在来自 $\mathcal{Z}$ 的样本组成的通用数据集 $\mathcal{S}$ 上的经验风险。在联邦学习中，客户端通常会协作来解决

$$\underset{\theta\in\Theta}{\text{minimize}}\, \mathcal{L}_{\mathcal{P}^{(\boldsymbol{\alpha})}}(\theta) = \sum_{m=1}^{M} \alpha_m \mathcal{L}_{\mathcal{P}_m}(\theta), \tag{1}$$

其中 $\mathcal{P}^{(\boldsymbol{\alpha})} \triangleq \sum_{m=1}^{M} \alpha_m \cdot \mathcal{P}_m$ 和 $\boldsymbol{\alpha} \triangleq (\alpha_m)_{1\leq m\leq M}$，且 $\alpha_m \geq 0$ 和 $\|\boldsymbol{\alpha}\|_1 = 1$。$\boldsymbol{\alpha}$的常见选择包括 $\alpha_m = n_m$ 和 $\alpha_m = \frac{1}{M}$。第一种选择对应于在聚合训练数据集 $\mathcal{S} = \bigcup_{m=1}^{M} \mathcal{S}_m$上最小化经验损失，从而赋予每个样本相同的重要性。第二种选择则通过赋予每个客户端相同的重要性，来实现客户端公平性。

在标准联邦学习中，本地数据集 $\{\mathcal{S}_m\}_{m\in[M]}$从训练开始时就已可用，因此考虑以下经验风险最小化问题，作为问题1的替代:

$$\underset{\theta\in\Theta}{\text{minimize}} \sum_{m=1}^{M} \alpha_m \cdot \mathcal{L}_{\mathcal{S}_m}(\theta). \tag{2}$$

我们的目标是设计一种可能具有随机性的算法 A，以联邦方式利用客户端数据流求解问题1，并考虑客户端的内存限制。

4 数据流的联邦学习元算法

从数据流中学习时，每个客户端只能访问当前存储在其本地内存中的样本。由于

算法 1： 从数据流中进行联邦学习的元算法**输入：** 本地训练轮次 E；小批量大小 K；本地学习率 $\eta > 0$；样本权重**输出：** $\bar{\theta}^{(T)} = \sum_{t=1}^{T} q^{(t)}\theta^{(t)}$ **1 对于** $t = 1 \ldots T$ **执行** **2 3 4 5 6 7 11 12 13 14** $\boldsymbol{\lambda} = \left\{\lambda_m^{(t,j)}; m \in [M], t \in [T], j \in \mathcal{I}_m^{(t)}\right\}$

```
,  ,
Server selects a subset 𝕊^(t) ⊆ [M] of clients;
for m ∈ 𝕊^(t) (in parallel) do
    θ_m^(t,1) ← θ^(t);
    Sample B_m^(t) = {z_m^(t,1), ... z_m^(t,b_m^(t))} ~ P_m^{b_m^(t)};
    M_m^(t) ← Update(M_m^(t−1), B_m^(t));
    for e = 1, ..., E do
        Sample min{K, |I_m^(t)|} indices ξ_m^(t,e)
          uniformly from I_m^(t) ;
        g_m^(t,e) ← (|I_m^(t)| / |ξ_m^(t,e)|) Σ_{j∈ξ_m^(t,e)} (λ_m^(t,j) / Σ_{j'∈I_m^(t)} λ_m^(t,j')) ·
          ∇ℓ(θ_m^(t,e); z_m^(t,j)) ;
        θ_m^(t,e+1) ← θ_m^(t,e) − η · g_m^(t,e);
    end
end
Δ^(t) ← Σ_{m=1}^M p_m^(t) · (θ_m^(t,E+1) − θ^(t)) ;
θ^(t+1) ← Π_Θ(θ^(t) + Δ^(t)) ;
end
```

由于每个客户端的存储容量有限，而且随时间到达的新样本数量会发生变化，样本在内存中停留的时间可能各不相同，因此在训练过程中被使用的次数也可能不同。为了有可能弥补这种异质性，我们允许样本在不同时间和不同客户端获得不同的权重。具体而言，我们用 $\lambda_m^{(t,j)} \geq 0$ 表示在时间 t 分配给存储在客户端 m 的内存中的样本 j 的权重（即 $j \in \mathcal{I}_m^{(t)}$），并用 $\boldsymbol{\lambda} \triangleq \left\{\lambda_m^{(t,j)};\ m \in [M], t \in [T], j \in \mathcal{I}_m^{(t)}\right\}$ 表示所有权重的集合。我们将时间步 $t \in [T]$ 时与客户端 m 的本地内存相关联的加权本地目标函数定义为

$$\mathcal{L}_{\mathcal{M}_m^{(t)}}^{(\boldsymbol{\lambda})}(\theta) \triangleq \frac{\sum_{j\in\mathcal{I}_m^{(t)}} \lambda_m^{(t,j)} \ell\left(\theta, \boldsymbol{z}_m^{(j)}\right)}{\sum_{j\in\mathcal{I}_m^{(t)}} \lambda_m^{(t,j)}}, \tag{3}$$

类似地，全局加权经验风险定义为

$$\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}(\theta) \triangleq \frac{\sum_{m=1}^{M}\sum_{t=1}^{T}\sum_{j\in\mathcal{I}_m^{(t)}} \lambda_m^{(t,j)} \cdot \ell\left(\theta; \boldsymbol{z}_m^{(j)}\right)}{\sum_{m=1}^{M}\sum_{t=1}^{T}\sum_{j\in\mathcal{I}_m^{(t)}} \lambda_m^{(t,j)}}. \tag{4}$$

此外，我们将客户端-m的聚合权重定义为

$$p_m^{(t)} \triangleq \frac{\sum_{j\in\mathcal{I}_m^{(t)}} \lambda_m^{(t,j)}}{\sum_{m'=1}^{M}\sum_{j\in\mathcal{I}_{m'}^{(t)}} \lambda_{m'}^{(t,j)}}, \tag{5}$$

and

$$q^{(t)} \triangleq \frac{\sum_{m=1}^{M}\sum_{j\in\mathcal{I}_m^{(t)}} \lambda_m^{(t,j)}}{\sum_{s=1}^{T}\sum_{m'=1}^{M}\sum_{j\in\mathcal{I}_{m'}^{(s)}} \lambda_{m'}^{(s,j)}}. \tag{6}$$

在这项工作中，我们考虑一种类似于原始 FedAvg（McMahan 等人，2017年）的元算法，以最小化加权经验风险（4）。算法 1 以迭代方式运行：在时间步 $t \in [T]$（也称为通信轮次），中央服务器将全局模型 $\theta^{(t)}$ 广播给一部分客户端（第 4 行）。随后，每个被选中的客户端（例如 m）都会收到一个新的数据批次（第 5 行），并用它更新客户端的本地内存 $\mathcal{M}_m^{(t)}$（第 6 行）。被选中的客户端执行 E 个本地随机梯度步（第 10 行），其中随机梯度 $\boldsymbol{g}_m^{(t,e)}$ 是使用至多 K 个样本计算得出的 $\nabla\mathcal{L}^{(\boldsymbol{\lambda})}_{\mathcal{M}_m^{(t)}}\left(\theta_m^{(t,e)}\right)$ 的无偏估计量（第 9 行）。经过 E 个本地步骤后，客户端将模型发回中央服务器进行聚合（第 13、14 行）。时间步 t 的更新也可以写为如下形式

$$\theta^{(t+1)} = \prod_{\Theta}\left(\theta^{(t)} - \eta\cdot\sum_{m=1}^{M} p_m^{(t)}\sum_{e=1}^{E}\boldsymbol{g}_m^{(t,e)}\right), \tag{7}$$

其中，$\Pi_\Theta(\cdot)$ 表示对集合 Θ进行投影。

请注意，算法 1 的输出取决于客户端处实际的样本到达序列、记忆更新规则以及权重 $\boldsymbol{\lambda}$。具体而言，记忆更新规则决定了在给定时间步可以考虑哪些样本，进而决定哪些权重可以非零。不过，为简便起见，我们将输出简单记为 $A^{(\boldsymbol{\lambda})}(\mathcal{S})$。

本文将分析范围限定在这样的情形：记忆更新规则和权重选择规则均为确定性的，且不依赖于内存中样本的特征或标签。更形式化地说，给定计数过程$N_{(t)\,m}$的一个特定实例，如果将 $\boldsymbol{z}_m^{(i)} = \left(\boldsymbol{x}_m^{(i)}, y_m^{(i)}\right)$ 替换为 $\boldsymbol{z}_m^{(i)} = \left(\tilde{\boldsymbol{x}}_m^{(i)}, \tilde{y}_m^{(i)}\right)$，且 $\tilde{\boldsymbol{x}}_m^{(i)} \neq \boldsymbol{x}_m^{(i)}$或 $\tilde{y}_m^{(i)} \neq y_m^{(i)}$，则样本 $\boldsymbol{z}_m^{(i)} \in \mathcal{S}_m$的权重 $\{\lambda_m^{(t,i)}\}_{t\in[T]}$保持不变。

对于给定的样本到达序列和记忆更新规则，算法的质量通过真实误差来评估

$$\epsilon_{\text{true}} \triangleq \mathbb{E}_{A^{(\boldsymbol{\lambda})},\mathcal{S}}\left[\mathcal{L}_{\mathcal{P}^{(\boldsymbol{\alpha})}}\left(A^{(\boldsymbol{\lambda})}(S)\right)\right] - \min_{\theta\in\Theta}\mathcal{L}_{\mathcal{P}^{(\boldsymbol{\alpha})}}(\theta), \tag{8}$$

其中，期望是针对算法 $A^{(\boldsymbol{\lambda})}$的潜在随机性取的，即客户端（第 2 行）和批次（第 8 行）的抽样过程，以及收集到的样本。

4.1 一般分析

式（8）中元算法的真实误差 ϵ_{true} 可作如下有界（证明见附录B.1）

$$\epsilon_{\text{true}} \leq \underbrace{\mathop{\mathbb{E}}_{\mathcal{S},A^{(\boldsymbol{\lambda})}}\left[\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}\left(A^{(\boldsymbol{\lambda})}\left(\mathcal{S}^{(T)}\right)\right) - \min_{\theta\in\Theta}\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}(\theta)\right]}_{\triangleq\epsilon_{\text{opt}}} + 2\underbrace{\mathop{\mathbb{E}}_{\mathcal{S}}\left[\sup_{\theta\in\Theta}\left|\mathcal{L}_{\mathcal{P}^{(\boldsymbol{\alpha})}}(\theta) - \mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}(\theta)\right|\right]}_{\triangleq\epsilon_{\text{gen}}}. \tag{9}$$

泛化误差 ϵ_{gen} 是数据集 $\mathcal{S}$的代表性 的期望值，即真实风险 $\mathcal{L}_{\mathcal{P}^{(\boldsymbol{\alpha})}}$ 与经验风险 $\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}$ 之间的最大距离。直观而言，泛化误差越小，我们就越能通过最小化 $\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}$ 来逼近 $\mathcal{L}_{\mathcal{P}^{(\boldsymbol{\alpha})}}$ 的最小值。

优化误差 ϵ_{opt} 衡量算法 1 对加权经验风险 $\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}$ 的极小化器的逼近程度。

在本节余下部分，我们首先给出泛化误差 ϵ_{gen}（定理 4.1）和优化误差 ϵ_{opt}（定理 4.3）的界，然后将二者结合起来，给出总体误差 ϵ_{true}（定理 4.4）的界。我们的结果依赖于以下假设：

假设 1。（有界损失）损失函数有界，即 $\forall\theta\in\Theta$, $\boldsymbol{z}\in\mathcal{Z}$, $\ell(\theta;\boldsymbol{z})\in[0,B]$。

假设 2。（有界域）我们假设 Θ 是凸、闭且有界的，其直径为 D。

假设 3。（凸性）对于所有 $\boldsymbol{z}\in\mathcal{Z}$，函数 $\theta\mapsto\ell(\theta;\boldsymbol{z})$ 在 $\mathbb{R}^d$ 上是凸的。

假设 4。（光滑性）对于所有 $\boldsymbol{z}\in\mathcal{Z}$，函数 $\theta\mapsto\ell(\theta;\boldsymbol{z})$ 在 $\mathbb{R}^d$ 上是 L-光滑的。

假设 1 是统计学习理论中的标准假设（例如，Mohri 等人（2018年）和 Shalev-Shwartz 和 Ben-David（2014））。假设 2–4 是分析（随机）梯度方法（例如，Bubeck 等人（2015）和 Bottou 等人（2018年））以及在线凸优化（Hazan，2019）时常用的假设。

注释1. 假设 *1* 和 *4* 意味着（可由附录 *B.2* 中的引理 *B.2* 推出）

$$\sigma_0^2 \triangleq \max_m \mathop{\mathbb{E}}_{\boldsymbol{z}\sim\mathcal{P}_m}\left[\sup_{\theta\in\Theta}\left\|\nabla\ell(\theta;\boldsymbol{z}) - \nabla\mathcal{L}_{\mathcal{P}_m}(\theta)\right\|^2\right] \tag{10}$$

$$\leq \left(2\cdot\sqrt{2LB}\right)^2, \tag{11}$$

并且（这可由附录 *B.2* 中的引理 *B.3* 得出）

$$\zeta \triangleq \max_{m,m'}\sup_{\theta\in\Theta}\left\|\nabla\mathcal{L}_{\mathcal{P}_{m'}}(\theta) - \nabla\mathcal{L}_{\mathcal{P}_m}(\theta)\right\| \tag{12}$$

$$\leq 2\cdot\sqrt{2LB}. \tag{13}$$

这些性质类似于随机梯度的有界方差假设，以及分析联邦学习算法时通常采用的客户端有界差异性假设（*Wang* 等，*2021*年*a*）。

4.2 界定泛化误差

定理 4.1（证明见附录 B.3）量化了泛化误差，尤其说明了加权经验风险 $\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}$ 与目标期望风险 $\mathcal{L}_{\mathcal{P}(\boldsymbol{\alpha})}$ 之间的差异，其中前者对应的极小化器为 $|\mathcal{L}_{\mathcal{P}(\boldsymbol{\alpha})}(\theta') - \mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}(\theta')|$，该定理对 $\theta' \in \arg\min_{\theta\in\Theta} \mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}(\theta)$给出了界。该界不同于经典统计学习结果（例如 Shalev-Shwartz 和 Ben-David，2014 年的结果），因为 $\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}$ 是加权经验风险，其期望值不一定与 $\mathcal{L}_{\mathcal{P}(\boldsymbol{\alpha})}$ 一致。回顾一下，与假设类 $\mathcal{H}$ 相关的标签差异度量了两个分布 $\mathcal{P}$ 和 $\mathcal{P}'$ 之间的距离，定义如下：$\text{disc}_{\mathcal{H}}(\mathcal{P}, \mathcal{P}') \triangleq \max_{h\in\mathcal{H}} |\mathcal{L}_{\mathcal{P}}(h) - \mathcal{L}_{\mathcal{P}'}(h)|$ (Mansour 等人，2020年)。

定理 4.1。 假设假设 *1* 成立，则使用带有权重 $\boldsymbol{\lambda}$的算法 *1* 时，可得

$$\epsilon_{gen} \leq \text{disc}_{\mathcal{H}}\left(\mathcal{P}^{(\boldsymbol{\alpha})}, \mathcal{P}^{(\boldsymbol{p})}\right) + \tilde{O}\left(\sqrt{\frac{\text{VCdim}\,(\mathcal{H})}{N_{\text{eff}}}}\right), \tag{14}$$

where $N_{eff} = \left(\sum_{m=1}^{M}\sum_{i=1}^{N_m} p_{m,i}^2\right)^{-1}$,

$$p_{m,i} = \frac{\sum_{t=1}^{T}\sum_{j\in\mathcal{I}_m^{(t)}} \mathbb{1}\{j=i\}\cdot\lambda_m^{(t,j)}}{\sum_{m'=1}^{M}\sum_{t=1}^{T}\sum_{j\in\mathcal{I}_{m'}^{(t)}}\lambda_{m'}^{(t,j)}}, \quad i\in N_m, \tag{15}$$

and $\boldsymbol{p} = \left(\sum_{i=1}^{N_m} p_{m,i}\right)_{1\leq m\leq M}$.

系数 $p_{m,\,i}$ 表示在整个训练期间，样本 i 相对于所有客户端收集的全部样本所具有的相对重要性；而 $p_m = \sum_{i=1}^{N_m} p_{m,\,i}$ 表示客户端 m 在训练期间所具有的相对重要性。注意，当 $p_m^{(t)}$随时间保持不变时，$p_m = \sum_{t=1}^{T} q^{(t)} p_m^{(t)}$和 $p_m^{(t)}$ 与相对重要性 p_m一致。

一般而言，我们应该赋予客户端的重要性（由式 (1) 中的 $\boldsymbol{\alpha}$ 量化）与训练期间实际赋予它们的重要性（由 $\boldsymbol{p}$量化）之间存在不一致性。右端的第一项通过差异刻画了目标分布 $\mathcal{P}^{(\boldsymbol{\alpha})}$ 与“有效分布” $\mathcal{P}^{(\boldsymbol{p})} = \sum_{m=1}^{M} p_m \mathcal{P}_m$ 之间的不匹配。

(14) 右侧的第二项在形式上类似于统计学习理论中的常见界，例如 (Shalev-Shwartz 和 Ben-David, 2014) 中的界，它们与假设类的 VC 维数和样本总数 N之比的平方根成正比。在我们的情形中，N_{eff} 起到有效样本数的作用，而引理 4.2（证明见附录 B.4）表明，正如预期的那样，N_{eff} 至多为 N，并且当每个样本具有相同的重要性时达到该值。

引理 4.2。 成立 $N_{\text{eff}} \leq N$，并且当每个样本具有相同的相对重要性时，即对于每个 i，$j\in[N_m]$，$p_{m,i} = p_{m,j}$，该界可取到。

当 $\boldsymbol{\alpha}$ 和 $\boldsymbol{p}$越接近、N_{eff} 越大时，泛化误差 ϵ_{gen} 越小。当 $\alpha_m = n_m$ （请记住 $n_m = N_m/N$）时，选择 $p_{m,i} = 1/N$ 可使该界最小，因为这样既能得到 $\boldsymbol{p} = \boldsymbol{n} = \boldsymbol{\alpha}$，也能得到 $N_{\text{eff}} = N$。

在我们的流式学习设定中，$p_{m,i} = 1/N$ 可以通过记忆更新规则和样本权重选择规则的不同组合来实现。例如，当客户端的内存仅包含当前轮次收到的样本（即算法 1 第 6 行中的 Update($\mathcal{M}_m^{(t-1)}, \mathcal{B}_m^{(t)}$) $= \mathcal{B}_m^{(t)}$），且内存中的所有样本权重均为 1（即对每个 $j\in\mathcal{I}_m^{(t)}$，$\lambda_m^{(t,j)} = 1$）时，便是这种情况。但如果记忆更新规则允许样本在内存中连续保留多个轮次（例如，客户端 m 上的样本 j 保留 $\tau_m^{(j)}$ 轮），且样本的权重与其连续保留的轮数成反比（即 $\lambda_m^{(t,j)} = 1/\tau_m^{(j)}$），也同样如此。下文将任何能够得到 $p_{m,i} = 1/N$ 的记忆更新规则与权重选择规则的组合称为均匀策略。

虽然均匀策略在 $\boldsymbol{\alpha} = \boldsymbol{n}$ 时能使泛化误差的上界 ϵ_{gen} 最小，但就优化误差 ϵ_{opt} 而言，它通常并非最优，正如我们将在下一节中证明的那样。

4.3 优化误差的界

我们给出了在所有客户端均参与（$\mathbb{S}^{(t)} = [M]$）且使用全批次（$K \geq |\mathcal{I}_m^{(t)}|$）时 ϵ_{opt} 的上界。在小批量梯度下，还会出现一个额外的渐消误差项。证明见附录 B.5。

定理 4.3。 假设假设 *1–4* 成立，序列 $\left(q^{(t)}\right)_t$ 非递增，且满足 $q^{(1)} = \mathcal{O}(1/T)$，并且 $\eta \propto 1/\sqrt{T}\cdot$最小值 $\{1, 1/\bar{\sigma}(\boldsymbol{\lambda})\}$。在客户端全参与（$\mathbb{S}^{(t)} = [M]$）且采用全批次（$K \geq |\mathcal{I}_m^{(t)}|$）的条件下，我们有

$$\epsilon_{opt} \leq \mathcal{O}\left(\bar{\sigma}(\boldsymbol{\lambda})\right) + \mathcal{O}\left(\frac{\bar{\sigma}(\boldsymbol{\lambda})}{\sqrt{T}}\right) + \mathcal{O}\left(\frac{1}{\sqrt{T}}\right), \tag{16}$$

其中，

$$\bar{\sigma}^2(\boldsymbol{\lambda}) \triangleq \sum_{t=1}^{T} q^{(t)} \times \mathbb{E}_{\mathcal{S}}\left[\sup_{\theta\in\Theta}\left\|\nabla\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}(\theta) - \sum_{m=1}^{M} p_m^{(t)}\nabla\mathcal{L}_{\mathcal{M}_m^{(t)}}^{(\boldsymbol{\lambda})}(\theta)\right\|^2\right]. \quad (17)$$

此外，存在一个数据到达过程和一个损失函数 ℓ，使得在采用先进先出内存更新规则时，[1] 对于任意权重选择 $\boldsymbol{\lambda}$，$\epsilon_{opt} = \Omega(\bar{\sigma}(\boldsymbol{\lambda}))$。

系数 $\bar{\sigma}^2(\boldsymbol{\lambda})$ 量化了第 t 轮更新中所考虑的梯度相对于全局目标函数 $\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}$ 的梯度的变化程度；正如定理 4.3 所示，当T 发散时，该系数会阻止优化误差趋于零。引理 B.4 根据随机梯度方差和客户端差异性，给出了 $\bar{\sigma}^2(\boldsymbol{\lambda})$ 的一般上界。

优化误差 ϵ_{opt} 越小，$\bar{\sigma}^2(\boldsymbol{\lambda})$ 就越接近零。在我们的流式学习场景中，如果内存从不更新（Update($\mathcal{M}_m^{(t-1)}$, $B_m^{(t)}$) $= \mathcal{M}_m^{(t-1)}$，$\forall t \geq 1$），且聚合权重始终保持不变（$p_m^{(t)} = p_m$，$\forall t \in [T]$），则可能得到 $\bar{\sigma}^2(\boldsymbol{\lambda})$ =0。事实上，容易验证，在这些条件下，$\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}(\theta) = \sum_{m=1}^{M} p_m^{(t)}\mathcal{L}_{\mathcal{M}_m^{(t)}}^{(\boldsymbol{\lambda})}(\theta)$（且它们等于$\sum_{m=1}^{M} p_m\mathcal{L}_{\mathcal{M}_m^{(0)}}^{(\boldsymbol{\lambda})}(\theta)$）。任何一组与时间无关的样本权重都会产生恒定的聚合权重；而在这些权重中，选择 $\lambda_m^{(t,j)} = 1$ 可降低泛化界 ϵ_{gen}。我们将这些内存更新和权重选择规则称为历史策略。

历史策略通过忽略训练期间收集的所有样本，使优化界最小化。这与均匀策略形成鲜明对比，后者对所有收集到的样本赋予相同的相对重要性。

4.4 主要结果

从上述讨论中可以明显看出，两个误差分量 ϵ_{gen} 和 ϵ_{opt} 之间存在权衡。每个时刻只考虑最新的样本，可以使 ϵ_{gen} 最小化；反之，忽略这些样本则可以使ϵ_{opt} 最小化。定理 4.4 结合定理 4.1 和定理 4.3，对这一权衡进行了正式量化，并给出了 ϵ_{true}的界。

定理 4.4. 在与定

[1]先进先出（First-In-First-Out）更新规则会从内存中淘汰最旧的样本，以存储最新的样本。

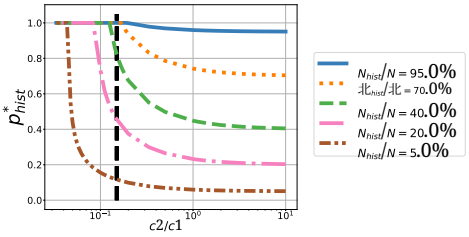


图 1：c_2/c_1 对不同 N_{hist}/N取值下历史客户端相对重要性 p^*_{hist} 的影响，其中$M = 50$ 且 $M_{\text{hist}} = 25$。虚线对应我们对 CIFAR-10 实验中 c_2/c_1 的估计（$\hat{c}_2/\hat{c}_1 = 0.15$）。

注 *4.1* 和定理 *4.3*，

$$\epsilon_{true} \leq \mathcal{O}\left(\frac{1}{\sqrt{T}}\right) + \mathcal{O}\left(\bar{\sigma}(\boldsymbol{\lambda})\right) + 2\text{disc}_{\mathcal{H}}\left(\mathcal{P}^{(\boldsymbol{\alpha})}, \mathcal{P}^{(\boldsymbol{p})}\right) + \tilde{O}\left(\sqrt{\frac{\text{VCdim}(\mathcal{H})}{N_{\text{eff}}}}\right). \quad (18)$$

5 案例研究

在雾计算环境中，物联网设备、边缘服务器和云服务器可以共同参与训练机器学习模型（Bonomi 等，2012）。物联网设备不断生成新数据，但由于内存限制严格，可能无法永久存储这些数据。相反，边缘服务器可以提供规模更大的静态数据集（Hosseinalipour 等，2020；Wang 等，2021b）。受此场景启发，我们考虑两组客户端：M_{hist} 拥有“历史”数据集的客户端，这些数据集在训练期间不会发生变化；以及$M - M_{\text{hist}}$ 客户端，它们以恒定速率 $\{b_m > 0, m \in [\![M_{\text{hist}} + 1, M]\!]\}$ 收集“新鲜”样本，并且由于内存限制（即$C_m = b_m$），只存储最新的 b_m 个样本。[2] 我们分别将这两类称为历史客户端和新客户端。新客户端也可以用于描述这样一种情形：客户端仅在单个通信轮次内可用——详情见附录 C.1。

每个客户端中的所有样本被使用的次数相同（T，历史客户端和新客户端分别为 1）。由此可以证明，每个客户端（例如客户端 m）都应为其本地内存中当前可用的所有样本分配相同的权重，即 $\lambda_m^{(t,j)} = \lambda_m^{(t)}$。为简单起见，我们考虑平稳权重，即 $\lambda_m^{(t)} = \lambda_m$，并希望据此确定每个客户端的样本权重（$\lambda_m$）$_{m\in[M]}$，从而在 ϵ_{true} 方面获得最佳保证。[3] 换言之，我们希望确定客户端的

[2]请注意，我们实际上选择了先进先出作为记忆更新规则。

[3]将权重限制为平稳状态，即 $\lambda_m^{(t)} = \lambda_m$，可能并非最优。

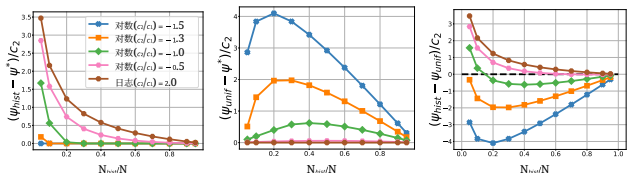


图 2：在 CIFAR-10 数据集上（$N = 5 \times 10^5$），当 $M = 50$ 且 $M_{\text{hist}} = 25$ 时，不同 c_2/c_1 取值下，$\psi_{\text{hist}} - \psi^*$（左）、$\psi_{\text{uniform}} - \psi^*$（中）和 $\psi_{\text{hist}} - \psi_{\text{uniform}}$（右）随 N_{hist}/N 变化的差异。

相对重要性值 $\boldsymbol{p} = (p_m)_{m\in[M]}$，其中 $p_m = \lambda_m N_m / \left(\sum_{m'=1}^{M} \lambda_{m'} N_{m'}\right)$。请注意，在这种设定下，聚合权重与相对重要性值相同（即 $p_m^{(t)} = p_m$）。推论 5.1′（附录 C）给出了此情形下 ϵ_{true} 关于 $\boldsymbol{p}$ 的界。为简单起见，我们在此给出 $\alpha_m = n_m$，$m \in [M]$ 情况下的界（本节余下部分均假设该条件成立）：

推论 5.1. 考虑有 M_{hist} 个历史客户端和 $M - M_{hist}$ 个新客户端的情形。假设定理 *4.4* 的相同条件成立，且 $\boldsymbol{\alpha} = \boldsymbol{n}$，并使用算法 *1*，客户端聚合权重为 $\boldsymbol{p} = (p_m)_{m\in[M]} \in \Delta^{M-1}$，则

$$\epsilon_{true} \leq \psi(\boldsymbol{p}; \boldsymbol{c}) \triangleq c_0 + c_1 \cdot \sqrt{\sum_{m=M_{hist}+1}^{M} p_m^2} + c_2 \cdot \sqrt{\sum_{m=1}^{M} \frac{p_m^2}{n_m}}, \tag{19}$$

其中 $\boldsymbol{c} = (c_0, c_1, c_2)$ 是与 $\boldsymbol{p}$ 无关的非负常数，定义如下：

$$c_0 = (C_1 + C_3) + \frac{C_2}{T} - 2 \cdot \max_{m,m'} \operatorname{disc}(\mathcal{P}_m, \mathcal{P}_{m'}) \tag{20}$$

$$c_1 = \sigma_0 \sqrt{M - M_0} \cdot \left(D + \frac{2}{\sqrt{T}}\right) \tag{21}$$

$$c_2 = 4 \cdot \sqrt{1 + \log\left(\frac{N}{\operatorname{VCdim}(\mathcal{H})}\right)} \cdot \sqrt{\frac{\operatorname{VCdim}(\mathcal{H})}{N}} + 2 \cdot \max_{m,m'} \operatorname{disc}(\mathcal{P}_m, \mathcal{P}_{m'}) \tag{22}$$

而 C_1、C_2 和 C_3 是定理 *4.3* 证明中定义的常数，σ_0 则在注释 *1* 中定义。

(19) 中的第二项刻画了梯度变异性（即 (18) 中的第二项），而 (19) 中的第三项则刻画了泛化误差的两种贡献，即分布差异和有效样本数量（分别对应 (19) 中的第三项和第四项）。具体而言，有 $\sum_{m=1}^{M} \frac{p_m^2}{n_m} \propto 1/N_{\text{eff}}$。

在单位单纯形上对 ψ 进行最小化是一个凸优化问题（证明见附录 C.4），因此可以通过例如投影梯度下降法高效求解。我们用 ψ^*、$\boldsymbol{p}^*$ 和 p^*_{hist} 分别表示 ψ 的最小值、其极小化器，以及分配给历史客户端的汇总相对重要性（$p^*_{\text{hist}} \triangleq \sum_{m=1}^{M_{\text{hist}}} p^*_m$）。

解 $\boldsymbol{p}^*$ 取决于 $\boldsymbol{n}$ 的值——尤其取决于历史样本的比例 N_{hist}/N（其中 $N_{\text{hist}} \triangleq \sum_{m=1}^{M_{\text{hist}}} N_m$）——以及比值 c_2/c_1。比值 c_2/c_1 仅取决于学习问题的内在性质（VCdim($\mathcal{H}$)、D、B 和 σ_0）以及样本总数 N（参见附录 C.3）。

图 1 展示了最优客户端的重要性值如何随比率 c_2/c_1 和历史样本占比 N_{hist}/N 变化（其他结果见图 4）。除具体数值外，还可以区分两种极端情况。当 $c_2/c_1 \gg 1$时，最优解对应于最小化 $\sum_{m=1}^{M} p_m^2/n_m$，即最大化有效样本数。此时，最优策略为均匀策略，历史客户端的重要性相对总量为 $p^*_{\text{hist}} = N_{\text{hist}}/N$。相反，当 $c_2/c_1 \ll 1$时，最优解对应于最小化 $\sum_{m>M_{\text{hist}}} p_m^2$，即梯度变异性。此时，历史策略为最优策略，对应于 $p^*_m = N_m/N_{\text{hist}} = \frac{N}{N_{\text{hist}}} n_m$，适用于 $m \in [M_{\text{hist}}]$ 和 $p^*_{\text{hist}} = 1$。

对于 c_2/c_1 的一般取值，分配客户端重要性值（或等价地，分配样本权重）的最优策略既不同于均匀策略，也不同于历史策略。因此，我们提出以下启发式方法，并将在下一节对其进行评估。在训练开始时，客户端协作地使用部分历史样本估计 c_2/c_1，即 $\hat{c}_2/\hat{c}_1 \approx \frac{B\mp\sqrt{d/N}}{GD\sqrt{M-M_{\text{hist}}}}$（详情见附录 C.6）。随后，选择客户端的重要性值，使式 (19) 中的界最小化，即 $\hat{\boldsymbol{p}}^* = \arg\min \psi(\cdot, \hat{\boldsymbol{c}})$。

除了为我们的元算法提供配置规则之外，我们的分析还使我们能够评估图 2 所示的均匀策略和历史策略等不同策略的性能如何取决于不同参数。下一节中的实验结果将证实这些理论预测。

6 实验结果

数据集和模型。 我们在五个联邦基准数据集上考察了不同的机器学习任务：图像分类（CIFAR-10 和 CIFAR-100（Krizhevsky，2009年））、手写字符识别（FEMNIST（Caldas 等人，2018年））、语言建模（莎士比亚（Caldas 等人，2018年；McMahan 等人，2017年）），以及在附录 D.1 中介绍的合成数据集上进行逻辑回归。表 1 总结了数据集、模型和客户端总数。关于数据集、模型和超参数选择的详细信息

表 1：数据集和模型。

数据集	客户端	总样本数	模型
合成	11	200	线性模型
CIFAR-10 / 100	50	50，000	2 卷积神经网络 + 2 全连接
FEMNIST	3,597	817,851	2 FC
莎士比亚	916	3,436,096	堆叠式长短期记忆网络

见附录 D。代码可在 https://github.com/omarfoq/streaming-fl 获取。

到达过程。 对于合成数据集和 CIFAR-10/100，我们采用常见策略在客户端之间划分数据集，并按照第 5 节中的做法，使用 $M_{\text{hist}} = 10$和 $M_{\text{hist}} = 25$将客户端分为两组。对于 FEMNIST 和莎士比亚数据集，我们采用其自然划分，并设置 M_{hist}，使其分别为 $M_{\text{hist}}/M = 5\%$、20% 和 50%，但允许新客户端参与训练几个轮次。这两个数据集的实验结果表明，我们的分析对于偏离第 5 节所考虑设定的情况仍然稳健。详情见附录 D.3。

基线方法。 我们将选择客户端重要性值的策略（见第 5 节）与三种基线方法进行了比较：上述均匀策略和历史策略，以及仅考虑新客户端的新数据策略。我们观察到，在样本到达过程和 $\boldsymbol{\alpha} = \boldsymbol{n}$下，经典 FedAvg 聚合规则（McMahan 等人，2017年）有两种自然的扩展方式：将每个客户端的聚合权重设置为与（1）该客户端在整个时间范围内收集的样本数量，或（2）该客户端当前内存中的样本数量成正比。第一种聚合规则与均匀策略一致；在我们考虑的所有设置中，第二种规则都会使新客户端的聚合权重非常小，因此实际上与历史策略无法区分。有趣的是，这两种规则通常都不是最优的，这凸显了本研究及所提出策略的实践价值。

主要结果。 表 2 报告了不同策略在$N_{\text{hist}}/N = 20\%$时的测试准确率，以及在网格 $\{0, 0.2, 0.5, 0.8, 1.0\}$ 中选择 $p_{\text{hist}} = \sum_{m=1}^{M_{\text{hist}}} p_m$ 的值所获得的最优测试准确率。我们的观察结果也得到了 N_{hist}/N其他取值下实验结果的证实（见附录 E 中的表 4 和表 5）。首先，只使用新数据（如新鲜策略所做的那样）从来都不是最优选择，即使历史数据仅占数据集总量的 5%（表 4）。其次，两种“合理”的 FedAvg 扩展方式都无法始终取得良好的准确率：历史策略在合成数据集上表现不佳，而均匀策略在 FEMNIST 和莎士比亚数据集上表现不佳。相反，我们的方法始终至少与最佳基线方法表现相当，而且通常能够取得接近（估计）最优值的测试准确率。在具体而言，在合成数据集上，它正确地将权重设置为均匀策略，在 FEMNIST 和莎士比亚数据集上则设置为历史策略。我们发现，我们的分析还有助于解释这一反直觉的结论：在 FEMNIST 和莎士比亚数据集上，忽略新收集的样本是有益的（即使在 $N_{\text{hist}}/N = 5\%$时也是如此，见表 4）。我们的策略正确地设置了 $\hat{p}^*_{\text{hist}} = 1$，因为它估计出，对于这两个数据集，参数数量与训练数据集总规模的比值（d/N）远小于梯度范数（G）——数值结果见附录 D.4。这一信息表明，我们可以使用原始数据集的一个小子集，在所选假设类中确定一个良好的模型；尤其是，我们可以仅依赖历史数据，从而避免新样本可能引入的噪声。

图 3 展示了不同比率 N_{hist}/N 下， $\boldsymbol{p}$ 对 CIFAR-10 测试准确率的影响——其他数据集的类似图表见附录 E。结果证实，就最终测试准确率而言，性能与模型对图 2 中所示界限 ψ 的预测相符。首先，图 3 显示，历史策略与最优分配之间的性能差距会随着 N_{hist}/N 增大而缩小（如图 2（左）所预测）：当 N_{hist}/N 分别为 5%、20% 和 50% 时，差距依次为 15.5 ± 0.30、7.9 ± 1.17 和 5.3 ± 2.8 个百分点。其次，图 3 证实，随着 N_{hist}/N 增大，均匀策略与最优分配之间的性能差距先增大后减小（如图 2（中）所示）：当 N_{hist}/N 分别为 5%、20% 和 50% 时，差距依次为3.0 ± 0.57、6.2 ± 0.55 和 4.3 ± 0.35 个百分点。最后，图 3 显示，均匀策略和历史策略的相对排名会发生变化：当 N_{hist}/N 较小时，均匀策略更优；当其较大时，历史策略则略胜一筹。同样，这一现象也符合我们的分析。事实上，在本实验中，我们估计比率 c_2/c_1 为$\hat{c}_2/\hat{c}_1 \approx 0.15 \in [10^{-1.3}, 10^{-0.5}]$，对应于图 2（右）中 $\psi_{\text{hist}} - \psi_{\text{unif}}$ 符号发生变化的情形。

7 结论

本文形式化了数据流联邦学习问题，并指出了一种新的异质性来源：本地数据集会随时间变化。我们提出了一种适用于这一场景的通用联邦算法，并研究了其理论保证。我们的分析揭示了一种新的偏差-优化权衡，该权衡由较旧样本相较于较新样本的相对重要性所决定，并由此得出如何在算法中配置这种重要性的实用指南。实验表明，在存在数据流的情况下，我们的配置规则优于扩展常用 FedAvg 聚合规则的自然方法。此外，尽管理论假设与相应的性能指标并不匹配，实验结果仍验证了其他理论结论，

表 2： 当 $N_{\text{hist}}/N = 20\%$时，不同数据集下各客户端的平均测试准确率。

数据集	$\hat{c}_2/\hat{c}_1$	$\hat{p}^*_{\text{历史}}$	测试准确率				
			新鲜	历史	均匀策略	OURS	最优
合成	0.092	0.20	84.7 ± 1.44	77.3 ± 3.15	$\mathbf{85.5} \pm 1.60$	$\mathbf{85.5} \pm 1.60$	85.5 ± 1.60
CIFAR-10	0.150	0.45	59.6 ± 0.94	59.8 ± 2.16	61.5 ± 0.63	$\mathbf{66.9} \pm 0.81$	67.7 ± 0.91
CIFAR-100	0.284	0.32	22.4 ± 0.57	22.6 ± 0.50	25.3 ± 0.43	$\mathbf{28.5} \pm 0.57$	31.5 ± 0.25
FEMNIST	0.001	1.00	53.3 ± 1.85	$\mathbf{66.1} \pm 0.20$	55.4 ± 0.80	$\mathbf{66.1} \pm 0.20$	66.1 ± 0.80
莎士比亚	0.064	1.00	38.4 ± 0.43	$\mathbf{49.0} \pm 0.26$	39.3 ± 0.38	$\mathbf{49.0} \pm 0.26$	49.0 ± 0.26

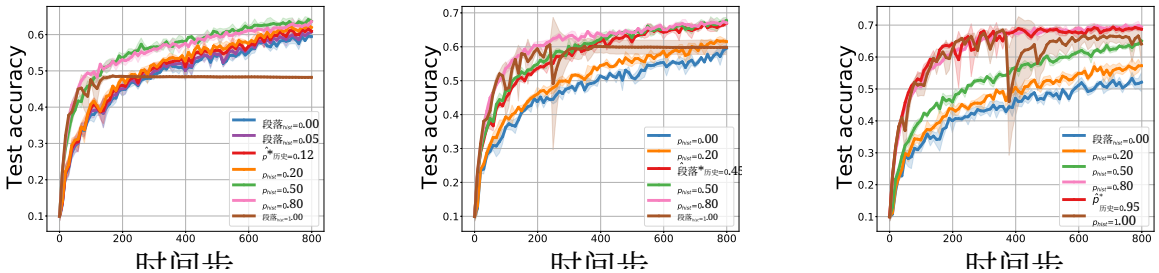


图 3：使用不同的 p_{hist} 时 CIFAR-10（左）的测试准确率变化， N_{hist}/N =分别为 5%（左）、20%（中）和 50%（右）。设置 $p_{\text{hist}} = N_{\text{hist}}/N$ 对应于均匀策略。

指标（例如，测试准确率与损失上界的关系）。

据我们所知，这项工作首次提出了数据流联邦学习问题。它指出了新的挑战，并且——我们认为——为进一步研究奠定了基础。例如，我们的部分结果仅适用于一种重要但仍相当特定的场景：一些客户端拥有静态数据集，而另一些客户端则在每一步处理新样本。在这种场景下，不同客户端使用样本的次数不同，但在同一客户端内，每个样本的使用次数完全相同，这简化了分析。但如果样本可用性的异质性也出现在单个客户端层面，会发生什么？不同的内存更新规则会如何影响这种异质性？我们又该如何设计此类策略，以尽量降低最终模型的总误差？最后，如果本地数据分布随时间变化，我们的结果会有何变化？

致谢

本研究部分由 Groupe La Poste（Inria 基金会的赞助方）在 FedMalin Inria 挑战赛框架下资助，部分由法国政府通过法国国家研究署（ANR）管理的“未来投资”项目“3IA Côte d’Azur”资助，项目参考编号为 ANR-19-P3IA- 0002。作者感谢 Universit´e Côte d’Azur 的 OPAL 基础设施提供计算资源和技术支持。

参考文献

Brendan McMahan、Eider Moore、Daniel Ramage、Seth Hampson 和 Blaise Aguera y Arcas。《基于去中心化数据的深度网络通信高效学习》。载于人工智能与统计，第 1273–1282 页。PMLR，2017年。

L´eon Bottou、Frank E Curtis 和 Jorge Nocedal。面向大规模机器学习的优化方法。*SIAM Review*，60(2):223–311，2018年。

Jakub Kone˘cny、H Brendan McMahan、Felix X Yu、Pe- ter Richt´arik、Ananda Theertha Suresh 和 Dave Ba- con。联邦学习：提升通信效率的策略。*arXiv* 预印本 *arXiv:1610.05492*，2016年。

Gianmarco De Francisci Morales、Albert Bifet、Lati- fur Khan、Joao Gama 和 Wei Fan。物联网大数据流挖掘。载于第 *22* 届 *ACM SIGKDD* 国际知识发现与数据挖掘会议论文集，KDD ’16，第 2119–2120 页，纽约，纽约州，美国，2016年。计算机协会。ISBN 9781450342322。doi: 10.1145/ 2939672.2945385。URL https://doi.org/10. 1145/2939672.2945385。

Timothy Yang、Galen Andrew、Hubert Eichner、Haicheng Sun、Wei Li、Nicholas Kong、Daniel Ra mage 和 Franc¸oise Beaufays。应用联邦学习：改进 Google 键盘的查询建议。*arXiv* 预印本 *arXiv:1812.02903*，2018年。

Yan Gao、Titouan Parcollet、Salah Zaiem、Javier Fernandez-Marques、Pedro PB de Gusmao、Daniel J

Beutel 和 Nicholas D Lane。基于联邦声学模型的端到端语音识别。载于*ICASSP 2022-2022 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*，第 7227–7231 页。IEEE，2022。

Pierre Courtiol、Charles Maussion、Matahi Moarii、Elodie Pronier、Samuel Pilcer、Meriem Sefta、Pierre Manceron、Sylvain Toldo、Mikhail Zaslavskiy、Nolwenn Le Stang 等。基于深度学习的间皮瘤分类可改善患者预后的预测。《自然-医学》，25(10):1519–1525，2019。

Santiago Silva、Boris A Gutman、Eduardo Romero、Paul M Thompson、Andre Altmann 和 Marco Lorenzi。在分布式医学数据库中进行联邦学习：大规模皮质下脑数据的荟萃分析。载于*2019*年第*16*届*IEEE*生物医学成像国际研讨会（*ISBI 2019*），第270–274页。IEEE，2019。Liu Yang、Ben Tan、Vincent W Zheng、Kai Chen 和 Qiang Yang。联邦推荐系统。载于联邦学习，第225–239页。施普林格，2020。Xiang Li、Kaixuan Huang、Wenhao Yang、Shusen Wang 和 Zhihua Zhang。论非独立同分布数据上的FedAvg收敛。载于国际学习表征会议，2019。Mehryar Mohri、Gary Sivek 和 Ananda Theertha Suresh。无关型联邦学习。载于国际机器学习会议，第4615–4625页。PMLR，2019。Kang Wei、Jun Li、Ming Ding、Chuan Ma、Howard H Yang、Farhad Farokhi、Shi Jin、Tony QS Quek 和 H Vincent Poor。采用差分隐私的联邦学习：算法与性能分析。《*IEEE*信息取证与安全汇刊》，15:3454–3469，2020。Peva Blanchard、El Mahdi El Mhamdi、Rachid Guerraoui 和 Julien Stainer。对抗环境下的机器学习：拜占庭容错梯度下降。载于 I. Guyon、U. Von Luxburg、S. Bengio、H. Wallach、R. Fergus、S. Vishvanathan 和 R. Garnett 编，神经信息处理系统进展，第30卷。Curran Associates, Inc.，2017年。网址：https://proceedings. neurips.cc/paper/2017/file/f4b9ec30ad9f68f89b29639786cb62ef-Paper

. pdf. Yunyue Zhu 和 Dennis Shasha。Statstream：实时统计监测数千条数据流。载于 *VLDB*，2002 年。Shivnath Babu 和 Jennifer Widom。数据流上的连续查询。*SIGMODRec.*，30(3):109–120，2001 年 9 月。ISSN 0163-5808。doi: 10.1145/603867. 603884。URL https://doi.org/10.1145/ 603867.603884。

Eric Moulines 和 Francis Bach。机器学习随机逼近算法的非渐近分析。神经信息处理系统进展，24，2011年。S. Thrun。移动机器人控制的终身学习视角。载于*IEEE/RSJ*智能机器人与系统国际会议论文集（*IROS'94*），第1卷，第23–30页，第1卷，1994年。doi：10.1109/IROS. 1994.407413。Abhishek Kumar 和 Hal Daum´e III。多任务学习中的任务分组与重叠学习。载于第*29*届国际机器学习会议论文集，第1723–1730页，2012年。Paul Ruvolo 和 Eric Eaton。ELLA：一种高效的终身学习算法。载于 Sanjoy Dasgupta 和 David McAllester 编，第*30*届国际机器学习会议，机器学习研究论文集第28卷，第507–515页，美国佐治亚州亚特兰大，2013年6月。PMLR。网址：https://proceedings.mlr.press/v28/ruvolo13.html。James Kirkpatrick、Razvan Pascanu、Neil Rabinowitz、Joel Veness、Guillaume Desjardins、Andrei A Rusu、Kieran Milan、John Quan、Tiago Ramalho、Agnieszka Grabska-Barwinska 等。克服神经网络中的灾难性遗忘。美国国家科学院院刊，114(13):3521–3526，2017年。Jonathan Schwarz、Wojciech Czarnecki、Jelena Luketina、Agnieszka Grabska-Barwinska、Yee Whye Teh、Razvan Pascanu 和 Raia Hadsell。Progress & compress：一种可扩展的持续学习框架。载于国际机器学习会议，第4528–4537页。PMLR，2018年。Martin Zinkevich。在线凸规划与广义无穷小梯度上升。载于 Tom Fawcett 和 Nina Mishra 编，*ICML*，第928–936页。AAAI出版社，2003年。ISBN 1-57735-189-4。网址：http://dblp.uni-trier.de/db/conf/icml/icml2003.html#Zinkevich03。Yujing Chen、Yue Ning、Martin Slawski 和 Huzefa Rangwala。面向边缘设备的非独立同分布数据异步在线联邦学习。载于*2020*年 *IEEE*大数据国际会议（大数据），第15–24页。IEEE，2020年。Jaehong Yoon、Wonyong Jeong、Giwoong Lee、Eunho Yang 和 Sung Ju Hwang。基于加权客户端间迁移的联邦持续学习。载于国际机器学习会议，第12073–12086页。PMLR，2021年。Olusola Odeyomi 和 Gergely Zaruba。使用在线镜像下降的、具有长期约束的差分隐私联邦学习。载于2021年 IEEE国际

信息论研讨会（*ISIT*），第1308–1313页，2021年。doi：10.1109/ISIT45174.2021.9518177。

Chen Zhu、Zheng Xu、Mingqing Chen、Jakub Koneˇcn´y、Andrew Hard 和 Tom Goldstein。《昼行还是夜行？基于周期性变化分布的多分支网络联邦学习》。发表于国际学习表征会议，2022。网址：https: //openreview.net/forum?id=E4EE_ohFGz。

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会议，第1764–1773页。PMLR，2019。

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》。《style id='1'>arXiv 预印本，arXiv:2002.07454，2020年。

Yongxin Guo、Tao Lin 和 Xiaoying Tang。面向随时间演变的异构数据的联邦学习。*arXiv* 预印本 *arXiv:2112.13246*，2021年。

Shuang Dai 和 Fanlin Meng。应对机器学习中的现代实际挑战：在线联邦学习与迁移学习综述。*arXiv* 预印本 *arXiv:2202.03070*，2022。

Georgios Damaskinos、Rachid Guerraoui、Anne-Marie Kermarrec、Vlad Nitu、Rhicheek Patra 和 Franc.ois Ta¨ıani。Fleet：通过陈旧性感知和性能预测实现在线联邦学习。发表于*ACM/IFIP*中间件会议，2020年。

Yibo Jin、Lei Jiao、Zhuzhong Qian、Sheng Zhang、San- glu Lu 和 Xiaoliang Wang。联邦学习中资源高效且保持收敛性的在线参与者选择。*2020*年 *IEEE* 第 *40* 届分布式计算系统国际会议 *(ICDCS)*，第606–616页，2020年。

Zhi Zhou、Song Yang、Lingjun Pu 和 Shuai Yu。CEFL：面向边缘节点的高效联邦学习在线准入控制、数据调度与精度调优。*IEEE*物联网期刊，7：9341–9356，2020年。

Michael JNeely。随机网络优化及其在通信与排队系统中的应用。通信网络综合讲座，3(1)：1–211，2010。

Mehryar Mohri、Afshin Rostamizadeh 和 Ameet Talwalkar。《机器学习基础》。自适应计算与机器学习。MIT出版社，美国马萨诸塞州剑桥，第二版，2018年。ISBN 978-0-262-03940-6。

Shai Shalev-Shwartz 和 Shai Ben-David。《理解机器学习：从理论到算法》。剑桥大学出版社，2014年。

S´ebastien Bubeck 等人。凸优化：算法与复杂度。《机器学习基础与趋势》，8(3-4):231–357，2015年。

Elad Hazan。在线凸优化导论。*arXiv*预印本 *arXiv:1909.05207*，2019年。Jianyu Wang、Zachary Charles、Zheng Xu、Gauri Joshi、H Brendan McMahan、Maruan Al-Shedivat、Galen Andrew、Salman Avestimehr、Katharine Daly、Deepesh Data 等人。联邦优化指南。*arXiv*预印本 *arXiv:2107.06917*，2021a。

Yishay Mansour、Mehryar Mohri、Jae Ro 和 Ananda Theertha Suresh。用于个性化的三种方法及其在联邦学习中的应用。*arXiv* 预印本 *arXiv:2002.10619*，2020年。

Flavio Bonomi、Rodolfo Milito、Jiang Zhu 和 Sateesh Addepalli。《雾计算及其在物联网中的作用》。载于首届移动云计算研讨会论文集，MCC '12，第13–16页，美国纽约州纽约，2012年。计算机协会。ISBN 9781450315197。doi: 10.1145/2342509.2342513。URLhttps://doi. org/10.1145/2342509.2342513。

Seyyedali Hosseinalipour、Christopher G. Brinton、Vaneet Aggarwal、Huaiyu Dai 和 Mung Chiang。《从联邦学习到雾学习：异构无线网络上的分布式机器学习》。《style id='1'>IEEE通信杂志》，58(12):41–47，2020年12月。ISSN 1558-1896。doi: 10.1109/MCOM.001.2000410。会议名称：《IEEE通信杂志》。

Su Wang、Yichen Ruan、Yuwei Tu、Satyavrat Wagle、Christopher G. Brinton 和 Carlee Joe-Wong。面向雾计算的网络感知分布式学习优化。《IEEE/ACM网络汇刊》，29(5):2019–2032，2021年10月b。ISSN 1558-2566。doi: 10.1109/TNET.2021.3075432。会议名称：《IEEE/ACM网络汇刊》。

Alex Krizhevsky。从小型图像中学习多层特征。技术报告，2009年。

Sebastian Caldas、Sai Meher Karthik Duddu、Peter Wu、Tian Li、Jakub Koneˇcn`y、H Brendan McMahan、Virginia Smith 和 Ameet Talwalkar。Leaf：联邦设置的基准。*arXiv* 预印本 *arXiv:1812.01097*，2018年。

Othmane Marfoq、Giovanni Neglia、Richard Vidal 和 Laetitia Kameni。通过本地记忆实现个性化联邦学习。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，机器学习研究论文集第162卷，第15070–15092页。PMLR，2022年7月17日至23日。网址：https://proceedings.mlr.press/v162/marfoq22a.html。

Hongyi Wang、Mikhail Yurochkin、Yuekai Sun、Dimitris Papailiopoulos 和 Yasaman Khazaeni。采用匹配平均的联邦学习。载于国际学习表征会议，2020年。网址：https://openreview.net/forum?id=BkluqlSFDS。

Wei Li 和 Andrew McCallum。Pachinko 分配：主题相关性的有向无环图结构混合模型。载于第*23*届国际机器学习会议论文集，ICML '06，第577–584页，美国纽约州纽约市，2006年。计算机协会。ISBN 1595933832。doi：10.1145/1143844. 1143917。URL https://doi.org/10.1145/ 1143844.1143917。

Sashank J. Reddi、Zachary Charles、Manzil Zaheer、Zachary Garrett、Keith Rush、Jakub Konečnỳ、Sanjiv Kumar 和 Hugh Brendan McMahan。《自适应联邦优化》。载于国际学习表征会议，2021年。网址：https://openreview.net/forum?id=LkFG3lB13U5。

相关工作

本节将进一步介绍一些相关工作的详细信息。

Chen 等人（2020年）提出了 ASO-Fed，这是一种异步联邦学习算法，旨在最小化基于客户端数据流聚合计算得出的经验损失。尽管论文陈述了一些收敛结果，但其意义和适用性值得质疑，因为分析要求所有客户端具有相同的最优模型，并且任意时刻 t 的更新都与未来到达的新样本保持一致。事实上，论文提到客户端在训练期间可以接收新样本（见图 2），但同时还要求在任意时刻 t，对于任意客户端 k，更新 $\nabla \zeta_k(w)$ 的期望值在全局经验损失 F 的梯度方向上具有非零分量，而该损失取决于之后时间 t 到达的样本（见假设 1）。此外，有界梯度差异假设意味着，F (F 被假设为强凸) 的极小化器也是每个局部目标函数 f_k 的驻点（考虑 $\beta = 0$ 和 $\lambda = 0$）。与此相反，我们论文中的理论分析适用于客户端本地数据分布存在统计异质性的情况，并将使用客户端当前存储的样本所带来的偏差纳入考量。此外，我们还为算法提供了统计学习保证。

(Guo 等人，2021年) 中考虑的模型能够刻画这样一种场景：客户端在训练期间持续收集数据，且不受存储限制。事实上，客户端会跟踪 (Guo 等人，2021年，式 (2)) 中的动态目标函数，该函数取决于截至当前时刻收到的数据样本。理论结果假设新数据是从客户端无关分布中抽取的。(Guo 等人，2021年，式 (5)) 展示了这一点，该式要求基于新数据样本计算出的局部梯度是全局目标函数梯度的无偏估计量。相比之下，我们的分析同时考虑了内存限制以及不同客户端本地数据分布之间的统计异质性。

B 证明

我们提醒读者，所有结果均基于以下假设：

Assu假设 1. *(有界损失)* 损失函数有界，即 $\forall\theta \in \Theta$, $\boldsymbol{z} \in \mathcal{Z}$, $\ell(\theta;\boldsymbol{z}) \in [0,B]$

假设 2. *(有界域)* 我们假设 Θ 是凸、闭且有界的；用 D 表示其直径，即 $\forall\theta,\theta' \in \Theta$, $\|\theta-\theta'\| \leq D$。

假设 3. *(凸性)* 对所有 $\boldsymbol{z} \in \mathcal{Z}$，函数 $\theta \mapsto \ell(\theta;\boldsymbol{z})$ 在 $\mathbb{R}^d$上是凸的。

假设 4. *(光滑性)* 对所有 $\boldsymbol{z} \in \mathcal{Z}$，函数 $\theta \mapsto \ell(\theta;\boldsymbol{z})$ 在 $\mathbb{R}^d$上是 L-光滑的。

在下文中，我们用 Δ^{D-1} 表示维度为 $D-1$ 的单位单纯形，即 $\Delta^{D-1} = \left\{\boldsymbol{f} \in \mathbb{R}_+^D, \sum_{i=1}^D f_i = 1\right\}$

B.1 式(9)的证明

$$\begin{aligned}\epsilon_{\text{true}} = &\mathop{\mathbb{E}}_{\mathcal{S},A^{(\boldsymbol{\lambda})}}\left[\mathcal{L}_{\mathcal{P}^{(\boldsymbol{\alpha})}}\left(A^{(\boldsymbol{\lambda})}(\mathcal{S})\right) - \mathcal{L}_S^{(\boldsymbol{\lambda})}\left(A^{(\boldsymbol{\lambda})}(\mathcal{S})\right)\right] + \mathop{\mathbb{E}}_{\mathcal{S},A^{(\boldsymbol{\lambda})}}\left[\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}\left(A^{(\boldsymbol{\lambda})}(\mathcal{S})\right) - \min_{\theta\in\Theta}\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}(\theta)\right] \\ &+ \mathop{\mathbb{E}}_{\mathcal{S}}\left[\min_{\theta\in\Theta}\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}(\theta)\right] - \min_{\theta\in\Theta}\mathcal{L}_{\mathcal{P}^{(\boldsymbol{\alpha})}}(\theta) \qquad (23)\\ \leq &\underbrace{2\mathop{\mathbb{E}}_{\mathcal{S}}\left[\sup_{\theta\in\Theta}\left|\mathcal{L}_{\mathcal{P}^{(\boldsymbol{\alpha})}}(\theta) - \mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}(\theta)\right|\right]}_{\triangleq\epsilon_{\text{gen}}} + \underbrace{\mathop{\mathbb{E}}_{\mathcal{S},A^{(\boldsymbol{\lambda})}}\left[\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}\left(A^{(\boldsymbol{\lambda})}(\mathcal{S})\right) - \min_{\theta\in\Theta}\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}(\theta)\right]}_{\triangleq\epsilon_{\text{opt}}}, \qquad (24)\end{aligned}$$

其中我们利用了以下事实：$\min_{x\in X} f(x) - \min_{x\in X} g(x) \leq \sup_{x\in X}|f(x) - g(x)|$。

B.2 性质

引理 B.1. 设 f 是一个取值于 $[0,B]$ 的 L-光滑函数，则 $\|\nabla f\| \leq \sqrt{2LB}$。

证明. 设 $\theta \in \Theta$，则根据 f 关于 $\theta' = \theta - \frac{1}{L}\nabla f(\theta)$ 的 L-光滑性定义，我们有 –

$$\begin{aligned}f(\theta') = f\left(\theta - \frac{1}{L}\nabla f(\theta)\right) &\leq f(\theta) - \frac{1}{L}\langle\nabla f(\theta), \nabla f(\theta)\rangle + \frac{L}{2}\left\|\frac{1}{L}\nabla f(\theta)\right\|^2 \qquad (25)\\ &= f(\theta) - \frac{1}{2L}\|\nabla f(\theta)\|^2. \qquad (26)\end{aligned}$$

由此可得，

$$\|\nabla f(\theta)\|^2 \leq 2L(f(\theta) - f(\theta')) \leq 2LB. \qquad (27)$$

□

引理 B.2. 假设条件 *1* 和 *4* 成立。对于所有

$$\sup_{\theta\in\Theta}\|\nabla\ell(\theta;\boldsymbol{z}) - \nabla\mathcal{L}_{\mathcal{P}_m}(\theta)\|^2 \leq \left(2\sqrt{2LB}\right)^2 \qquad (28)$$

.

证明。令 $\boldsymbol{z} \in \mathcal{Z}$ 和 $m \in [M]$。$\ell(\cdot, \boldsymbol{z})$ 和 $\mathcal{L}_{\mathcal{P}_m}$ 都是 L-光滑的，并且在 $[0,B]$ 内有界。

对于 $\theta \in \Theta$，我们有

$$\begin{aligned}\|\nabla\ell(\theta;\boldsymbol{z}) - \nabla\mathcal{L}_{\mathcal{P}_m}(\theta)\|^2 &\leq 2\|\nabla\ell(\theta;\boldsymbol{z})\|^2 + 2\|\nabla\mathcal{L}_{\mathcal{P}_m}(\theta)\|^2 \qquad (29)\\ &\leq 2\cdot 2LB + 2\cdot 2LB \qquad (30)\\ &= 8LB = \left(2\sqrt{2LB}\right)^2, \qquad (31)\end{aligned}$$

其中，我们使用引理 B.1 得到了最后一个不等式。 □

引理 B.3. 假设条件 *1* 和 *4* 成立。对于所有 $z \in \mathcal{Z}$，我们有

$$\max_{m,m'} \sup_{\theta \in \Theta} \left\| \nabla \mathcal{L}_{\mathcal{P}_{m'}}(\theta) - \nabla \mathcal{L}_{\mathcal{P}_m}(\theta) \right\| \leq 2\sqrt{2LB}. \tag{32}$$

.

证明。 利用三角不等式和引理 B.1 可得证明。 □

B.3 定理 4.1 的证明

定理 4.1。 假设假设 *1* 成立，使用权重 $\boldsymbol{\lambda}$的算法 *1*，可得

$$\epsilon_{gen} \leq \operatorname{disc}_{\mathcal{H}}\left(\mathcal{P}^{(\boldsymbol{\alpha})}, \mathcal{P}^{(\boldsymbol{p})}\right) + \tilde{O}\left(\sqrt{\frac{\operatorname{VCdim}(\mathcal{H})}{N_{\text{eff}}}}\right),$$

其中 $N_{eff} = \left(\sum_{m=1}^{M} \sum_{i=1}^{N_m} p_{m,i}^2\right)^{-1}$，

$$p_{m,i} = \frac{\sum_{t=1}^{T} \sum_{j \in \mathcal{I}_m^{(t)}} \mathbb{1}\{j = i\} \cdot \lambda_m^{(t,j)}}{\sum_{m'=1}^{M} \sum_{t=1}^{T} \sum_{j \in \mathcal{I}_{m'}^{(t)}} \lambda_{m'}^{(t,j)}}, \quad i \in N_m,$$

且 $\boldsymbol{p} = \left(\sum_{i=1}^{N_m} p_{m,i}\right)_{1 \leq m \leq M}$。

证明。 对于客户端 $m \in [M]$，请注意，$p_m \triangleq \sum_{i=1}^{N_m} p_{m,\,i}$ 是客户端 m 相对于其他客户端的相对重要性。我们定义

$$\mathcal{L}_{\mathcal{S},\boldsymbol{p}} = \sum_{m=1}^{M} \sum_{i=1}^{N_m} p_{m,i} \cdot \ell(\cdot; \boldsymbol{z}_m^{(i)}). \tag{33}$$

注意，对于任意 $\theta \in \Theta$，$\mathcal{L}_{\mathcal{S},\,\boldsymbol{p}} = \mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}$，且 $\mathbb{E}_{\mathcal{S}}\left[\mathcal{L}_{\mathcal{S},\boldsymbol{p}}(\theta)\right] = \sum_m p_m \mathcal{L}_{\mathcal{P}_m}(\theta) = \mathcal{L}_{P^{(\boldsymbol{p})}}(\theta)$，其中 $\mathcal{P}^{(\boldsymbol{p})} = \sum_m p_m \mathcal{P}_m$。我们有

$$\begin{aligned}
\epsilon_{\text{gen}} &= \mathop{\mathbb{E}}_{\mathcal{S}}\left[\sup_{h \in \mathcal{H}} \left|\mathcal{L}_{\mathcal{P}^{(\boldsymbol{\alpha})}}(h) - \mathcal{L}_{\mathcal{S},\boldsymbol{p}}(h)\right|\right] && (34)\\
&= \mathop{\mathbb{E}}_{\mathcal{S}}\left[\sup_{h \in \mathcal{H}} \left|\mathcal{L}_{\mathcal{P}^{(\boldsymbol{\alpha})}}(h) - \mathcal{L}_{\mathcal{P}^{(\boldsymbol{p})}}(h) + \mathcal{L}_{\mathcal{P}^{(\boldsymbol{p})}}(h) - \mathcal{L}_{\mathcal{S},\boldsymbol{p}}(h)\right|\right] && (35)\\
&\leq \mathop{\mathbb{E}}_{\mathcal{S}}\left[\sup_{h \in \mathcal{H}} \left|\mathcal{L}_{\mathcal{P}^{(\boldsymbol{\alpha})}}(h) - \mathcal{L}_{\mathcal{P}^{(\boldsymbol{p})}}(h)\right|\right] + \mathop{\mathbb{E}}_{\mathcal{S}}\left[\sup_{h \in \mathcal{H}} \left|\mathcal{L}_{\mathcal{P}^{(\boldsymbol{p})}}(h) - \mathcal{L}_{\mathcal{S},\boldsymbol{p}}(h)\right|\right] && (36)\\
&\leq \operatorname{disc}_{\mathcal{H}}\left(\mathcal{P}^{(\boldsymbol{\alpha})}, \mathcal{P}^{(\boldsymbol{p})}\right) + \mathop{\mathbb{E}}_{\mathcal{S}}\left[\sup_{h \in \mathcal{H}} \left|\mathcal{L}_{\mathcal{P}^{(\boldsymbol{p})}}(h) - \mathcal{L}_{\mathcal{S},\boldsymbol{p}}(h)\right|\right]. && (37)
\end{aligned}$$

现在对式（37）右侧的第二项进行界定。注意，对于 $h \in \mathcal{H}$，我们可以写成 $\mathcal{L}_{\mathcal{P}^{(\boldsymbol{p})}}(h) = \mathbb{E}_{\mathcal{S}'}\left[\mathcal{L}_{\mathcal{S}',\boldsymbol{p}}(h)\right]$，其中 $\mathcal{S}' = \bigcup_{m=1}^{M} \mathcal{S}'_m$，且 $\mathcal{S}'_m \sim \mathcal{P}_m^{N_m}$ 是一个由 $\mathcal{P}_m$ 中独立同分布抽取的 N_m 个样本组成的数据集，使得 $\mathcal{S}_m = \left\{z_m^{(i)},\ i \in [N_m]\right\}$ 且 $\mathcal{S}'_m = \left\{z'^{(i)}_m,\ i \in [N_m]\right\}$。根据三角不等式，可得

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathcal{S}}\left[\sup_{h \in \mathcal{H}} \left|\mathcal{L}_{\mathcal{P}^{(\boldsymbol{p})}}(h) - \mathcal{L}_{\mathcal{S},\boldsymbol{p}}(h)\right|\right]\\
&\quad\leq \mathop{\mathbb{E}}_{\mathcal{S},\mathcal{S}'}\left[\sup_{h \in \mathcal{H}} \left|\mathcal{L}_{\mathcal{S}',\boldsymbol{p}}(h) - \mathcal{L}_{\mathcal{S},\boldsymbol{p}}(h)\right|\right] && (38)\\
&\quad= \mathop{\mathbb{E}}_{\mathcal{S},\mathcal{S}'}\left[\sup_{h \in \mathcal{H}} \left|\sum_{m=1}^{M} \sum_{i=1}^{N_m} p_{m,i}\left(\ell(h; z_m^{(i)}) - \ell(h; z'^{(i)}_m)\right)\right|\right] && (39)\\
&\quad= \mathop{\mathbb{E}}_{\mathcal{S},\mathcal{S}'} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{h \in \mathcal{H}} \left|\sum_{m=1}^{M} \sum_{i=1}^{N_m} \sigma_m^{(i)} \cdot p_{m,i}\left(\ell(h; z_m^{(i)}) - \ell(h; z'^{(i)}_m)\right)\right|\right], && (40)
\end{aligned}$$

其中 $\sigma_m^{(i)}$， $m \in [M]$， $i \in [N_m]$ 是一个从 $\{\pm 1\}$ 上的均匀分布中抽取的随机变量。固定 $\mathcal{S}$ 和 $\mathcal{S}'$，并令 C 为出现在 $\mathcal{S}$ 和 $\mathcal{S}'$ 中的实例， $\mathcal{H}_C$ 为 $\mathcal{H}$ 在 C 上的限制，定义见（ Shalev-Shwartz 和 Ben-David，2014，定义 6.2 ）。由此可得

$$\begin{aligned}&\mathop{\mathbb{E}}_{\mathcal{S}}\left[\sup_{h\in\mathcal{H}}\left|\mathcal{L}_{\mathcal{P}^{(\boldsymbol{p})}}(h)-\mathcal{L}_{\mathcal{S},\boldsymbol{p}}(h)\right|\right]\\&\qquad\leq\mathop{\mathbb{E}}_{\mathcal{S}',\mathcal{S}'}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}_C}\left|\sum_{m=1}^{M}\sum_{i=1}^{N_m}\sigma_m^{(i)}\cdot p_{m,i}\left(\ell(h;z_m^{(i)})-\ell(h;z'^{(i)}_m)\right)\right|\right].\end{aligned}\tag{41}$$

固定某个 $h \in \mathcal{H}_C$，并记作 $\gamma_m^{(i)} = \sigma_m^{(i)} \cdot p_{m,\,i}\left(\ell(h;\, z_m^{(i)}) - \ell(h;\, z'^{(i)}_m)\right)$，其中 $m \in [M]$ 且 $i \in [N_m]$。我们有 $\mathbb{E}\left[\gamma_m^{(i)}\right] = 0$，根据假设 1，有 $\gamma_m^{(i)} \in [-p_{m,i}\cdot B, p_{m,i}\cdot B]$。由于随机变量 $\left\{\gamma_m^{(i)},\ \ m\in[M],\ i\in[N_m]\right\}$ 相互独立，利用 Hoeffding不等式可知，对于所有 $\rho \geq 0$，有

$$\mathbb{P}\left[\left|\sum_{m=1}^{M}\sum_{i=1}^{N_m}\sigma_m^{(i)}\cdot p_{m,i}\left(\ell(h;z_m^{(i)})-\ell(h;z'^{(i)}_m)\right)\right|\geq\rho\right]\leq 2\exp\left(-2B^2N_{\text{eff}}\rho^2\right),\tag{42}$$

其中 $N_{\text{eff}} = \left(\sum_{m=1}^{M}\sum_{i=1}^{N_m}(p_{m,\,i})^2\right)^{-1}$。对 $h \in \mathcal{H}_C$ 应用并集界，并使用（Shalev-Shwartz 和 Ben-David，2014，引理 A.4），可得

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}_C}\left|\sum_{m=1}^{M}\sum_{i=1}^{N_m}\sigma_m^{(i)}\cdot p_{m,i}\left(\ell(h;z_m^{(i)})-\ell(h;z'^{(i)}_m)\right)\right|\right]\leq\frac{4+\sqrt{\log\left(|\mathcal{H}_C|\right)}}{\sqrt{2N_{\text{eff}}}B}.\tag{43}$$

由此可得，

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}_C}\left|\sum_{m=1}^{M}\sum_{i=1}^{N_m}\sigma_m^{(i)}\cdot p_{m,i}\left(\ell(h;z_m^{(i)})-\ell(h;z'^{(i)}_m)\right)\right|\right]\leq\frac{4+\sqrt{\log\left(\tau_{\mathcal{H}}\left(N^{(T)}\right)\right)}}{\sqrt{2N_{\text{eff}}B}},\tag{44}$$

其中 $\tau_{\mathcal{H}}$ 是 $\mathcal{H}$ 的增长函数，定义见 (Shalev-Shwartz 和 Ben-David, 2014, Definition 6.9)。利用 Sauer引理 (Shalev-Shwartz 和 Ben-David, 2014, Lemma 6.10)，并遵循 (Marfoq etal.,2022, LemmaA.1) 证明中的相同步骤，可得

$$\mathop{\mathbb{E}}_{\mathcal{S}}\left[\sup_{h\in\mathcal{H}}\left|\mathcal{L}_{\mathcal{P}^{(\boldsymbol{p})}}(h)-\mathcal{L}_{\mathcal{S},\boldsymbol{p}}(h)\right|\right]\leq 2\sqrt{\frac{\text{VCdim}\left(\mathcal{H}\right)}{N_{\text{eff}}}}\cdot\sqrt{1+\log\left(\frac{N}{\text{VCdim}\left(\mathcal{H}\right)}\right)},\tag{45}$$

其中 δ_1 和 δ_2 为非负常数。因此，

$$\mathop{\mathbb{E}}_{\mathcal{S}}\left[\sup_{h\in\mathcal{H}}\left|\mathcal{L}_{\mathcal{P}^{(\boldsymbol{p})}}(h)-\mathcal{L}_{\mathcal{S},\boldsymbol{p}}(h)\right|\right]\leq\tilde{O}\left(\sqrt{\frac{\text{VCdim}\left(\mathcal{H}\right)}{N_{\text{eff}}}}\right),\tag{46}$$

因此，

$$\epsilon_{\text{gen}}\leq\tilde{O}\left(\sqrt{\frac{\text{VCdim}\left(\mathcal{H}\right)}{N_{\text{eff}}}}\right)+\texttt{disc}_{\mathcal{H}}\left(\mathcal{P}^{(\boldsymbol{\alpha})},\mathcal{P}^{(\boldsymbol{p})}\right).\tag{47}$$

□

B.4 引理 4.2 的证明

引理 4.2。 采用与定理 *4.1* 相同的记号， $N_{\text{eff}} \leq N$，且当 $\boldsymbol{p}$ 为均匀策略时，此界可取到。

证明。 我们回顾一下

$$N_{\text{eff}} = \left(\sum_{m=1}^{M}\sum_{i=1}^{N_m}\left(p_{m,i}\right)^2\right)^{-1}.\tag{48}$$

令 $\boldsymbol{u} \in \Delta^N$ 为将所有值 $p_{m,i}$ （其中 $m \in [M]$ 且 $i \in [N_m]$）拼接而成的向量。由此可得

$$N_{\text{eff}} = \left(\sum_{n=1}^{N} u_n^2 \right)^{-1} = \|\boldsymbol{u}\|_2^{-2}. \tag{49}$$

令 $\boldsymbol{u}^* \triangleq \mathbf{1}/N$，显然 可知 $\boldsymbol{u}^* \in \Delta^N$，且 $\|\boldsymbol{u}^*\|_2^2 = 1/N$。令 $\boldsymbol{u} \in \Delta^N$，根据柯西-施瓦茨不等式，我们有

$$1 = \sum_{n=1}^{N} u_n = \sum_{n=1}^{N} (u_n \times 1) \leq \sqrt{\sum_{n=1}^{N} u_n^2} \cdot \sqrt{\sum_{n=1}^{N} 1} = \|\boldsymbol{u}\|_2 \cdot \sqrt{N}. \tag{50}$$

因此， $\|\boldsymbol{u}\|_2^{-2} \leq N$，证毕。

□

B.5 定理 4.3 的证明

定理 4.3。 假设假设 *1–4* 成立，序列 $\left(q^{(t)}\right)_t$ 单调不增，且满足 $q^{(1)} = \mathcal{O}(1/T)$，并且 $\eta \propto 1/\sqrt{T} \cdot \min\{1, 1/\bar{\sigma}(\boldsymbol{\lambda})\}$。在客户端全参与 （$\mathbb{S}^{(t)} = [M]$）且使用全批次 （$K \geq |\mathcal{I}_m^{(t)}|$） 的情况下，我们有

$$\epsilon_{opt} \leq \mathcal{O}\Big(\bar{\sigma}(\boldsymbol{\lambda})\Big) + \mathcal{O}\Big(\frac{\bar{\sigma}(\boldsymbol{\lambda})}{\sqrt{T}}\Big) + \mathcal{O}\left(\frac{1}{\sqrt{T}}\right),$$

其中，

$$\bar{\sigma}^2(\boldsymbol{\lambda}) \triangleq \sum_{t=1}^{T} q^{(t)} \times \mathop{\mathbb{E}}_{\mathcal{S}} \left[\sup_{\theta \in \Theta} \left\| \nabla \mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}(\theta) - \sum_{m=1}^{M} p_m^{(t)} \nabla \mathcal{L}_{\mathcal{M}_m^{(t)}}^{(\boldsymbol{\lambda})}(\theta) \right\|^2 \right].$$

此外，存在一个数据到达过程和一个损失函数 ℓ，使得在采用先进先出内存更新规则时，对于任意权重选择 $\boldsymbol{\lambda}$, $\epsilon_{opt} = \Omega(\bar{\sigma}(\boldsymbol{\lambda}))$。

证明。 我们提醒

$$p_m^{(t)} = \frac{\sum_{j \in \mathcal{I}_m^{(t)}} \lambda_m^{(t,j)}}{\sum_{m'=1}^{M} \sum_{j \in \mathcal{I}_{m'}^{(t)}} \lambda_{m'}^{(t,j)}}, \tag{51}$$

and

$$q^{(t)} = \frac{\sum_{m=1}^{M} \sum_{j \in \mathcal{I}_m^{(t)}} \lambda_m^{(t,j)}}{\sum_{s=1}^{T} \sum_{m=1}^{M} \sum_{j \in \mathcal{I}_{m'}^{(s)}} \lambda_{m'}^{(s,j)}}. \tag{52}$$

为便于记号表示，我们引入以下定义在 Θ上的函数;

$$f_m^{(t)} \triangleq \mathcal{L}_{\mathcal{M}_m^{(t)}}^{(\boldsymbol{\lambda})}, \tag{53}$$

$$F^{(t)} \triangleq \sum_{m=1}^{M} p_m^{(t)} \cdot \mathcal{L}_{\mathcal{M}_m^{(t)}}^{(\boldsymbol{\lambda})} = \sum_{m=1}^{M} p_m^{(t)} \cdot f_m^{(t)}, \tag{54}$$

$$F \triangleq \mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})} = \sum_{t=1}^{T} q^{(t)} \cdot F^{(t)}. \tag{55}$$

请注意，这种记法隐藏了函数 $f_m^{(t)}$、 $F_{(t)}$和 F 对样本 $\mathcal{S}$及参数 $\boldsymbol{\lambda}$的依赖。在此证明中，我们只用 $\mathbb{E}$ 表示样本 $\mathcal{S}$的期望，例如 $\mathbb{E}[\nabla F(\theta)] = \mathbb{E}_{\mathcal{S}}\left[\nabla \mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}(\theta)\right]$。

我们提醒一下，

$$\Delta^{(t)} = \sum_{m=1}^{M} p_m^{(t)} \cdot \left(\theta_m^{(t,E+1)} - \theta^{(t)}\right) = -\eta \cdot \sum_{e=1}^{E} \sum_{m=1}^{M} p_m^{(t)} \cdot \nabla f_m^{(t)} \left(\theta_m^{(t,e)}\right). \tag{56}$$

我们定义 $\tilde{\eta} \triangleq \eta E > 0$ 和 $\tilde{\nabla}^{(t)} \triangleq -\frac{\Delta^{(t)}}{\tilde{\eta}} \in \mathbb{R}^d$。系数 $\tilde{\eta}$ 和向量 $\tilde{\nabla}^{(t)}$ 分别可视为全局迭代 $t \in [T]$中使用的有效学习率和伪梯度 （Wang 等人，2021a，第 2 节）。采用这组符号，算法 1 的更新规则可概括为

$$\tilde{\nabla}^{(t)} = \frac{1}{E} \sum_{e=1}^{E} \sum_{m=1}^{M} p_m^{(t)} \cdot \nabla f_m^{(t)} \left(\theta_m^{(t,e)}\right) \tag{57}$$

$$\theta^{(t+1)} = \prod_{\Theta} \left(\theta^{(t)} - \tilde{\eta} \cdot \tilde{\nabla}^{(t)}\right) \tag{58}$$

根据假设 3–4，函数 $f_m^{(t)}$、 $F_{(t)}$和 F作为有界、凸且 L-光滑函数的凸组合，均为有界、凸且 L-光滑。

令 θ^*为 F在 Θ上的极小化器，且 $F^* \triangleq F(\theta^*)$（注意， θ^*和 F^*取决于 $\mathcal{S}$）。根据 F的凸性，我们有

$$-\left\langle \nabla F(\theta), \theta - \theta^* \right\rangle \leq -\left(F(\theta) - F^*\right). \tag{59}$$

引理 B.1 和詹森不等式可推出

$$\max\left\{\left\|\nabla f_m^{(t,e)}(\theta)\right\|, \left\|\nabla F^{(t)}(\theta)\right\|, \left\|\nabla F(\theta)\right\|, \left\|\tilde{\nabla}^{(t)}\right\|\right\} \leq G, \tag{60}$$

其中 $G \triangleq \sqrt{2LB}$。

为　方便起见，我们用以下方式量化当前函数与全局函数的梯度之间的方差

$$\sigma_t = \sup_{\theta \in \Theta} \left\|\nabla F(\theta) - \nabla F^{(t)}(\theta)\right\|. \tag{61}$$

我们定义 $\sigma^2(\boldsymbol{\lambda}) \triangleq \sum_{t=1}^{T} q^{(t)} \sigma_t^2$。因此， $\bar{\sigma}^2(\boldsymbol{\lambda}) = \mathbb{E}\left[\sigma^2(\boldsymbol{\lambda})\right]$。

证明的思路是界定伪梯度 $\tilde{\nabla}^{(t)}$ 与正确梯度 $\nabla F\left(\theta^{(t)}\right)$之间的距离；在迭代$t > 0$时本应使用正确梯度。可以写成

$$\mathbb{E}\left[\left\|\theta^{(t+1)} - \theta^*\right\|^2\right] = \mathbb{E}\left[\left\|\prod_{\Theta}\left(\theta^{(t)} - \tilde{\eta}\tilde{\nabla}\right) - \theta^*\right\|^2\right] \tag{62}$$

$$\leq \mathbb{E}\left[\left\|\theta^{(t)} - \tilde{\eta}\tilde{\nabla} - \theta^*\right\|^2\right] \tag{63}$$

$$= \mathbb{E}\left[\left\|\theta^{(t)} - \tilde{\eta}\nabla F\left(\theta^{(t)}\right) - \theta^* + \tilde{\eta}\left(\nabla F\left(\theta^t\right) - \tilde{\nabla}^{(t)}\right)\right\|^2\right] \tag{64}$$

$$= \mathbb{E}\left[\underbrace{\left\|\theta^{(t)} - \tilde{\eta}\nabla F\left(\theta^{(t)}\right) - \theta^*\right\|^2}_{\triangleq T_1}\right] + \tilde{\eta}^2\,\mathbb{E}\left[\underbrace{\left\|\nabla F\left(\theta^{(t)}\right) - \tilde{\nabla}^{(t)}\right\|^2}_{\triangleq T_2}\right]$$
$$+ 2\tilde{\eta}\,\mathbb{E}\left[\underbrace{\left\langle \nabla F\left(\theta^{(t)}\right) - \tilde{\nabla}^{(t)}, \theta^{(t)} - \tilde{\eta}\nabla F\left(\theta^{(t)}\right) - \theta^* \right\rangle}_{\triangleq T_3}\right]. \tag{65}$$

界定 T_1。我们有：

$$T_1 = \left\|\theta^{(t)} - \tilde{\eta}\nabla F\left(\theta^{(t)}\right) - \theta^*\right\|^2 \tag{66}$$

$$= \left\|\theta^{(t)} - \theta^*\right\|^2 + \tilde{\eta}^2 \left\|\nabla F\left(\theta^{(t)}\right)\right\|^2 - 2\tilde{\eta} \cdot \left\langle \nabla F\left(\theta^{(t)}\right), \theta^{(t)} - \theta^* \right\rangle \tag{67}$$

$$\leq \left\|\theta^{(t)} - \theta^*\right\|^2 + \tilde{\eta}^2 G^2 - 2\tilde{\eta}\left(F\left(\theta^{(t)}\right) - F^*\right), \tag{68}$$

其中，我们利用 (59) 和 (60) 得到了最后一个不等式。

界 T_2**.** 令 $\alpha > 0$，则有，

$$T_2 = \left\| \nabla F\left(\theta^t\right) - \tilde{\nabla}^{(t)} \right\|^2 \tag{69}$$

$$= \left\| \nabla F\left(\theta^{(t)}\right) - \sum_{m=1}^{M} p_m^{(t)} \nabla f_m^{(t)}\left(\theta^{(t)}\right) + \sum_{m=1}^{M} p_m^{(t)} \nabla f_m^{(t)}\left(\theta^{(t)}\right) - \tilde{\nabla}^{(t)} \right\|^2 \tag{70}$$

$$\leq (1+\alpha) \left\| \nabla F\left(\theta^{(t)}\right) - \nabla F^{(t)}\left(\theta^{(t)}\right) \right\|^2 + (1+\alpha^{-1}) \left\| \sum_{m=1}^{M} p_m^{(t)} \nabla f_m^{(t)}\left(\theta^{(t)}\right) - \tilde{\nabla}^{(t)} \right\|^2, \tag{71}$$

其中我们用到了以下事实：对于任意两个向量 $\boldsymbol{a}$，$\boldsymbol{b} \in \mathbb{R}^d$和一个系数 $\alpha > 0$，有 $\|\boldsymbol{a}+\boldsymbol{b}\|^2 \leq (1+\alpha) \|\boldsymbol{a}\|^2 + (1+\alpha^{-1}) \|\boldsymbol{b}\|^2$，其中取值为 $\boldsymbol{a} = \nabla F\left(\theta^{(t)}\right) - \nabla F_{(t)}\left(\theta^{(t)}\right)$，且 $\boldsymbol{b} = \sum_{m=1}^{M} p_m^{(t)} \nabla f_m^{(t)}\left(\theta^{(t)}\right) - \tilde{\nabla}^{(t)}$。

我们提醒读者，

$$\tilde{\nabla} = -\frac{\Delta^{(t)}}{\eta E} = \sum_{e=1}^{E} \sum_{m=1}^{M} \frac{p_m^{(t)}}{E} \boldsymbol{g}_m^{(t,e)} = \sum_{e=1}^{E} \sum_{m=1}^{M} \frac{p_m^{(t)}}{E} \nabla f_m^{(t)}\left(\theta_m^{(t,e)}\right). \tag{72}$$

因此，

$$\left\| \sum_{m=1}^{M} p_m^{(t)} \nabla f_m^{(t)}\left(\theta^{(t)}\right) - \tilde{\nabla}^{(t)} \right\|^2 = \left\| \sum_{e=1}^{E} \sum_{m=1}^{M} \frac{p_m^{(t)}}{E} \left(\nabla f_m^{(t)}\left(\theta^{(t)}\right) - \nabla f_m^{(t)}\left(\theta^{(t,e)}\right) \right) \right\|^2 \tag{73}$$

$$\leq \sum_{e=1}^{E} \sum_{m=1}^{M} \frac{p_m^{(t)}}{E} \left\| \nabla f_m^{(t)}\left(\theta^{(t)}\right) - \nabla f_m^{(t)}\left(\theta_m^{(t,e)}\right) \right\|^2 \tag{74}$$

$$= \sum_{e=1}^{E} \sum_{m=1}^{M} \frac{p_m^{(t)}}{E} \left\| \nabla f_m^{(t)}\left(\theta_m^{(t,1)}\right) - \nabla f_m^{(t)}\left(\theta_m^{(t,e)}\right) \right\|^2 \tag{75}$$

$$\leq L^2 \sum_{e=1}^{E} \sum_{m=1}^{M} \frac{p_m^{(t)}}{E} \left\| \theta_m^{(t,1)} - \theta_m^{(t,e)} \right\|^2 \tag{76}$$

$$= L^2 \sum_{e=1}^{E} \sum_{m=1}^{M} \frac{p_m^{(t)}}{E} \left\| \sum_{e'=1}^{e-1} \theta_m^{(t,e')} - \theta_m^{(t,e'+1)} \right\|^2 \tag{77}$$

$$= \frac{\tilde{\eta}^2 L^2}{E^3} \sum_{m=1}^{M} p_m^{(t)} \sum_{e=1}^{E} \left\| \sum_{e'=1}^{e-1} \nabla f_m^{(t)}\left(\theta_m^{(t,e')}\right) \right\|^2 \tag{78}$$

$$\leq \frac{\tilde{\eta}^2 L^2}{E^3} \sum_{m=1}^{M} p_m^{(t)} \sum_{e=1}^{E} (e-1) \sum_{e'=1}^{e-1} \left\| \nabla f_m^{(t)}\left(\theta_m^{(t,e')}\right) \right\|^2 \tag{79}$$

$$\leq \frac{\tilde{\eta}^2 L^2 G^2}{E^3} \sum_{e=1}^{E} (e-1)^2 \tag{80}$$

$$\leq 2\tilde{\eta}^2 L^2 G^2 (1 - E^{-1}), \tag{81}$$

其中，我们使用詹森不等式得到 (74) 和 (79)，使用 $f_m^{(t)}$ 的 L-光滑性得到 (76)，并使用 (60) 得到 (80)。将 (81) 代入 (71)，并使用在 (61) 中定义的 σ_t，可得

$$T_2 \leq (1+\alpha)\, \sigma_t^2 + 2\left(1+\alpha^{-1}\right) \tilde{\eta}^2 L^2 G^2 (1 - E^{-1}). \tag{82}$$

采用特定选择 $\alpha = \frac{\tilde{\eta} L G}{\sigma_t} \cdot \sqrt{2(1 - E^{-1})}$，可得

$$T_2 \leq \left(\sigma_t + \tilde{\eta} L G \sqrt{2\left(1 - E^{-1}\right)} \right)^2 \leq 2\sigma_t^2 + 4\tilde{\eta}^2 L^2 G^2 \left(1 - E^{-1}\right) \tag{83}$$

我们的界 ((83)) 表明，正如预期的那样，衡量真实梯度 $\nabla F\left(\theta^{(t)}\right)$ 与伪梯度 $\tilde{\nabla}^{(t)}$之间偏差的项 T_2，在 $E = 1$ 和 $\sigma_t = 0$时等于零。这种情形恰好对应于梯度下降的集中式版本。

界 T_3。我们有

$$T_3 = \left\langle \nabla F\left(\theta^{(t)}\right) - \tilde{\nabla}^{(t)}, \theta^{(t)} - \tilde{\eta}\nabla F\left(\theta^{(t)}\right) - \theta^* \right\rangle \tag{84}$$

$$= \left\langle \nabla F\left(\theta^{(t)}\right) - \nabla F^{(t)}\left(\theta^{(t)}\right), \theta^{(t)} - \theta^* \right\rangle + \left\langle \nabla F^{(t)}\left(\theta^{(t)}\right) - \tilde{\nabla}^{(t)}, \theta^{(t)} - \theta^* \right\rangle - \tilde{\eta}\left\langle \nabla F\left(\theta^{(t)}\right) - \tilde{\nabla}^{(t)}, \nabla F\left(\theta^{(t)}\right) \right\rangle. \tag{85}$$

我们回想起 Θ 是有界的，且 D 是其直径。根据柯西-施瓦茨不等式，我们有

$$\left\langle \nabla F^{(t)}\left(\theta^{(t)}\right) - \tilde{\nabla}^{(t)}, \theta^{(t)} - \theta^* \right\rangle \leq \left\| \nabla F^{(t)}\left(\theta^{(t)}\right) - \tilde{\nabla}^{(t)} \right\| \cdot \left\| \theta^{(t)} - \theta^* \right\| \tag{86}$$

$$= \left\| \sum_{m=1}^{M} p_m^{(t)} \nabla f_m^{(t)}\left(\theta^{(t)}\right) - \tilde{\nabla}^{(t)} \right\| \cdot \left\| \theta^{(t)} - \theta^* \right\| \tag{87}$$

$$\leq \tilde{\eta} LDG\sqrt{2\left(1 - E^{-1}\right)}, \tag{88}$$

这里使用（81）得到了最后一个不等式。再次使用柯西-施瓦茨不等式，并利用梯度有界这一事实（（60）），我们有

$$-\tilde{\eta}\left\langle \nabla F\left(\theta^{(t)}\right) - \tilde{\nabla}^{(t)}, \nabla F\left(\theta^{(t)}\right) \right\rangle \leq \tilde{\eta}\left\| \nabla F\left(\theta^{(t)}\right) - \tilde{\nabla}^{(t)} \right\| \cdot \left\| \nabla F\left(\theta^{(t)}\right) \right\| \leq 2\tilde{\eta} \cdot G^2. \tag{89}$$

最后，使用柯西-施瓦茨不等式以及 Θ的有界性，我们有

$$\left\langle \nabla F\left(\theta^{(t)}\right) - \nabla F^{(t)}\left(\theta^{(t)}\right), \theta^{(t)} - \theta^* \right\rangle \leq \sigma^{(t)} \cdot D. \tag{90}$$

将（88）、（89）和（90）代入（85），我们有

$$T_3 \leq \sigma^{(t)} \cdot D + \tilde{\eta} G\left(2G + LD\sqrt{2\left(1 - E^{-1}\right)}\right) \tag{91}$$

界 ϵ_{opt}。将（68）、（83）和（91）代入（65），可得

$$\mathbb{E}\left[\left\|\theta^{(t+1)} - \theta^*\right\|^2\right] = \mathbb{E}\left[\left\|\theta^{(t)} - \theta^*\right\|^2\right] - 2\tilde{\eta} \cdot \mathbb{E}\left[F\left(\theta^{(t)}\right) - F^*\right] + 2\tilde{\eta} \cdot \bar{\sigma}^{(t)} D + \tilde{\eta}^2 \cdot \left(2\bar{\sigma}_t^2 + G\left(5G + 2LD\sqrt{2\left(1 - E^{-1}\right)}\right)\right) + 4\tilde{\eta}^4 \cdot L^2 G^2\left(1 - E^{-1}\right), \tag{92}$$

其中 $\bar{\sigma}_t^2 = \mathbb{E}\left[\sigma_t^2\right] = \mathbb{E}\left[\text{上确界}_{\theta \in \Theta} \left\|\nabla F(\theta) - \nabla F^{(t)}(\theta)\right\|^2\right]$。

序列 $\left(q^{(t)}\right)_t$ 是非递增的，即对于 $t \in [T]$ $q^{(t+1)} \leq q^{(t)}$。由（92）可知，对于 $t > 0$，有

$$q^{(t+1)}\,\mathbb{E}\left[\left\|\theta^{(t+1)} - \theta^*\right\|^2\right] \leq q^{(t)}\,\mathbb{E}\left[\left\|\theta^{(t+1)} - \theta^*\right\|^2\right] \tag{93}$$

$$\leq q^{(t)}\,\mathbb{E}\left[\left\|\theta^{(t)} - \theta^*\right\|^2\right] - 2\tilde{\eta} q^{(t)}\,\mathbb{E}\left[F\left(\theta^{(t)}\right) - F^*\right] + 2\tilde{\eta} \cdot q^{(t)} \bar{\sigma}^{(t)} D + 2\tilde{\eta}^2 \cdot q^{(t)} \bar{\sigma}_t^2 + 2\tilde{\eta}^2 q^{(t)} \cdot C_1 + 2\tilde{\eta}^4 q^{(t)} \cdot C_2, \tag{94}$$

其中 $C_1 = G\left(\frac{5}{2}G + LD\sqrt{2\left(1 - E^{-1}\right)}\right)$，且 $C_2 = 2L^2G^2\left(1 - E^{-1}\right)$。重新排列各项并对 $t \in \{1, \ldots, T\}$ 求和，可得

$$\sum_{t=1}^{T} q^{(t)}\,\mathbb{E}\left[F\left(\theta^{(t)}\right) - F^*\right] \leq \left(\sum_{t=1}^{T} q^{(t)} \bar{\sigma}_t\right) \cdot D + Tq^{(1)} \cdot \frac{D^2}{2\tilde{\eta}T} + \tilde{\eta} \cdot \left(\sum_{t=1}^{T} q^{(t)} \bar{\sigma}_t^2\right) + \tilde{\eta} \cdot \left(C_1 + \tilde{\eta}^2 C_2\right) \tag{95}$$

提醒一下， $\bar{\sigma}^2(\boldsymbol{\lambda}) = \sum_{t=1}^{T} q^{(t)} \bar{\sigma}_t^2$。利用函数 $\sqrt{\cdot}$ 的凹性，可得 $\bar{\sigma}(\boldsymbol{\lambda}) \geq \sum_{t=1}^{T} q^{(t)} \bar{\sigma}_t$。因此

$$\mathbb{E}\left[F\left(\bar{\theta}^{(t)}\right) - F^*\right] \leq \bar{\sigma}\left(\boldsymbol{\lambda}\right) \cdot D + T q^{(1)} \cdot \frac{D^2}{2\tilde{\eta} T} + \tilde{\eta} \cdot \bar{\sigma}^2\left(\boldsymbol{\lambda}\right) + \tilde{\eta} C_1 + \tilde{\eta}^3 C_2. \tag{96}$$

最终结果通过使用 $\mathcal{O}\left(T q^{(1)}\right) = 1$ 得到。我们有

$$\mathbb{E}\left[F\left(\bar{\theta}^{(t)}\right) - F^*\right] \leq \bar{\sigma}\left(\boldsymbol{\lambda}\right) \cdot D + \frac{\bar{\sigma}\left(\boldsymbol{\lambda}\right)}{\sqrt{T}} + \frac{C_1 + C_3}{\sqrt{T}} + \frac{C_2}{\sqrt{T^3}}, \tag{97}$$

其中， C_3 是一个与 D^2 成正比的常数。

下界。在本证明的其余部分中，我们用 $\boldsymbol{\theta}$ 表示模型参数，用 θ_1 和 θ_2 表示其分量。

我们人为地构造一个简单问题和一个特定的到达过程，使得算法 1 在采用 $M = 1$、$C_1 = 1$、先进先出更新规则和 $\eta = \Omega\left(1/\sqrt{T}\right)$ 时，其输出满足 $\lim_{T\to\infty} F\left(\bar{\boldsymbol{\theta}}^{(T)}\right) - F^* \geq c \cdot \bar{\sigma}^2(\lambda)$，其中 $c > 0$ 是一个常数。我们考虑一种设置，其中 $\Theta = [-1, 1]^2$、 $\mathcal{Z} = \{1, 2\}$，并且损失函数定义在 $\boldsymbol{\theta} \in \Theta$ 上，满足

$$\ell(\boldsymbol{\theta}; 1) \triangleq (\theta_1 + 1)^2 + \frac{1}{2}(\theta_1 + \theta_2 + 1)^2, \tag{98}$$

and

$$\ell\left(\boldsymbol{\theta}; 2\right) \triangleq \frac{1}{2}\left(\theta_1 - 1\right)^2 + \frac{1}{2}(\theta_1 + \theta_2 - 1)^2. \tag{99}$$

我们观察到， $\ell(;1)$（分别为 $\ell(;2)$）的极小化器是 $\boldsymbol{\theta}_1^* = (-1,\ 0)$（分别为 $\boldsymbol{\theta}_2^* = (1,\ 0)$）。

对于时间范围 T，我们考虑到达过程：在时间步 $t_1 = 1$，从 $\mathcal{Z}$ 中均匀随机抽取一个样本，记为 $\boldsymbol{z}_1$；在时间步 $t_2 = T/2$，从 $\mathcal{Z}$ 中均匀随机抽取第二个样本 $\boldsymbol{z}_2$。我们定义 $q \triangleq \sum_{t=1}^{T/2} q^{(t)}$。由于 $\left(q^{(t)}\right)_{t\geq 1}$ 是非增的，因此 $q \geq 1/2$。我们指出，在此设定下，算法 1 的轨迹仅由 $\boldsymbol{z}_1$ 和 $\boldsymbol{z}_2$的取值决定；也就是说，序列 $\left(\boldsymbol{\theta}^{(t)}\right)_{t\geq 1}$ 所取的值仅由 $\boldsymbol{z}_1$ 和 $\boldsymbol{z}_2$的取值决定。

我们有

$$\begin{aligned}
\epsilon_{\text{opt}} &= \mathbb{E}_{\mathcal{S}}\left[\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}\left(\bar{\boldsymbol{\theta}}^{(T)}\right) - \text{最小值}_{\boldsymbol{\theta}\in\Theta}\, \mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}(\boldsymbol{\theta})\right] && (100)\\
&= \frac{1}{2}\mathbb{E}_{\mathcal{S}}\left[\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}\left(\bar{\boldsymbol{\theta}}^{(T)}\right) - \min_{\boldsymbol{\theta}\in\Theta} \mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}(\boldsymbol{\theta}) \,|\mathcal{S} = \{1,2\}\right] + \frac{1}{4}\mathbb{E}_{\mathcal{S}}\left[\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}\left(\bar{\boldsymbol{\theta}}^{(T)}\right) - \min_{\boldsymbol{\theta}\in\Theta} \mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}(\boldsymbol{\theta}) \,|\mathcal{S} = \{1\}\right] && (101)\\
&\quad + \frac{1}{4}\mathbb{E}_{\mathcal{S}}\left[\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}\left(\bar{\boldsymbol{\theta}}^{(T)}\right) - \min_{\boldsymbol{\theta}\in\Theta} \mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}(\boldsymbol{\theta}) \,|\mathcal{S} = \{2\}\right] && (102)\\
&\geq \frac{1}{2}\mathbb{E}_{\mathcal{S}}\left[\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}\left(\bar{\boldsymbol{\theta}}^{(T)}\right) - \min_{\boldsymbol{\theta}\in\Theta} \mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}(\boldsymbol{\theta}) \,|\mathcal{S} = \{1,2\}\right], && (103)
\end{aligned}$$

and

$$\begin{aligned}
\bar{\sigma}^2(\boldsymbol{\lambda}) &= q\left(1-q\right)\mathbb{E}_{\mathcal{S}}\left[\max_{\boldsymbol{\theta}\in\Theta} \|\nabla\ell(\boldsymbol{\theta}; \boldsymbol{z}_1) - \nabla\ell(\boldsymbol{\theta}; \boldsymbol{z}_2)\|^2\right] && (104)\\
&\leq \frac{q(1-q)}{2} \cdot \max_{\boldsymbol{\theta}\in\Theta} \|\nabla\ell(\boldsymbol{\theta}; 1) - \nabla\ell(\boldsymbol{\theta}; 2)\|^2 && (105)\\
&\leq 20 \cdot q\left(1-q\right)。 && (106)
\end{aligned}$$

我们考虑 $\boldsymbol{z}_1 = 1$且 $\boldsymbol{z}_2 = 2$ 的情况。因此

$$\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}(\boldsymbol{\theta}) = q \cdot \ell(\boldsymbol{\theta}; 1) + (1-q) \cdot \ell(\boldsymbol{\theta}; 2). \tag{107}$$

设 $\boldsymbol{\theta}^*$ 是 $\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}$ 的一个极小化器，则

$$\theta_1^* = \frac{1-3q}{1+q} \qquad \text{and} \qquad \theta_2^* = 1 - 2q - \frac{1-3q}{1+q}. \tag{108}$$

此外，可以证明

$$\min_{\theta\in[-1,1]} \mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}\left((\theta,0)\right) - \min_{\boldsymbol{\theta}\in\Theta} \mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}\left(\boldsymbol{\theta}\right) \geq 6\cdot q(1-q) \tag{109}$$

Fo对于 $\epsilon>0$，存在 $E\geq 1$ 和 $T_0\geq 1$，使得对于任意 $T\geq T_0$，都有 $\left|\bar{\theta}_2^{(T)}\right|\leq\epsilon$。因此，

$$\begin{aligned}
\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}\left(\bar{\boldsymbol{\theta}}^{(T)}\right) - \min_{\boldsymbol{\theta}\in\Theta} \mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}\left(\boldsymbol{\theta}\right) &\sim_{\epsilon\to 0} \mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}\left(\left(\theta_1^{(T)},0\right)\right) - \min_{\boldsymbol{\theta}\in\Theta} \mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}\left(\boldsymbol{\theta}\right) &(110)\\
&\geq \min_{\theta\in[-1,1]} \mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}\left((\theta,0)\right) - \min_{\boldsymbol{\theta}\in\Theta} \mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}\left(\boldsymbol{\theta}\right) &(111)\\
&\geq 6\cdot q(1-q) &(112)\\
&= \frac{3}{10}\bar{\sigma}^2\left(\boldsymbol{\lambda}\right) &(113)
\end{aligned}$$

当 $\boldsymbol{z}_1=2$且 $\boldsymbol{z}_2=1$时，同样成立。由此可得

$$\epsilon_{\text{opt}} \geq \frac{3}{20}\bar{\sigma}^2\left(\boldsymbol{\lambda}\right). \tag{114}$$

□

B.6 界 $\bar{\sigma}^2(\lambda)$

我们根据注释 1 回顾一下，

$$\sigma_0^2 \triangleq \max_m \mathop{\mathbb{E}}_{\boldsymbol{z}\sim\mathcal{P}_m}\left[\sup_{\theta\in\Theta}\left\|\nabla\ell(\theta;\boldsymbol{z}) - \nabla\mathcal{L}_{\mathcal{P}_m}\left(\theta\right)\right\|^2\right], \tag{115}$$

and

$$\zeta \triangleq \max_{m,m'}\sup_{\theta\in\Theta}\left\|\nabla\mathcal{L}_{\mathcal{P}_{m'}}\left(\theta\right) - \nabla\mathcal{L}_{\mathcal{P}_m}\left(\theta\right)\right\|. \tag{116}$$

引理 B.4. 对于任意记忆更新规则以及任意记忆参数选择 $\boldsymbol{\lambda}$，都有

$$\bar{\sigma}^2\left(\boldsymbol{\lambda}\right) = \mathcal{O}\left(\sigma_0^2 + \zeta^2\cdot\sum_{t=1}^{T} q^{(t)}\sum_{m=1}^{M}\left(p_m - p_m^{(t)}\right)^2\right). \tag{117}$$

证明。我们提醒读者，

$$\bar{\sigma}^2\left(\boldsymbol{\lambda}\right) = \sum_{t=1}^{T} q^{(t)}\mathop{\mathbb{E}}_{\mathcal{S}}\left[\sup_{\theta\in\Theta}\left\|\nabla\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}\left(\theta\right) - \sum_{m=1}^{M} p_m^{(t)}\nabla\mathcal{L}_{\mathcal{M}_m^{(t)}}^{(\boldsymbol{\lambda})}\left(\theta\right)\right\|^2\right], \tag{118}$$

并且，对于 $m\in[M]$，我们定义

$$\mathcal{L}_{\mathcal{S}_m}^{(\boldsymbol{\lambda})}\left(\cdot\right) \triangleq \frac{\sum_{t=1}^{T}\sum_{j\in\mathcal{I}_m^{(t)}}\lambda_m^{(t,j)}\ell\left(\cdot,\boldsymbol{z}_m^{(j)}\right)}{\sum_{s=1}^{T}\sum_{i\in\mathcal{I}_m^{(s)}}\lambda_m^{(s,i)}}, \tag{119}$$

并且我们提醒读者（参见定理 4.1），

$$p_m = \frac{\sum_{t=1}^{T}\sum_{j\in\mathcal{I}_m^{(t)}}\lambda_m^{(t,j)}}{\sum_{m'=1}^{M}\sum_{s=1}^{T}\sum_{i\in\mathcal{I}_m^{(s)}}\lambda_m^{(s,i)}}. \tag{120}$$

$\mathcal{L}_{\mathcal{S}_m}^{(\boldsymbol{\lambda})}$ 和 p_m 分别表示客户端 m的加权经验风险以及客户端 m 的相对重要性。我们指出，

$$\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})} = \sum_{m=1}^{M} p_m\mathcal{L}_{\mathcal{S}_m}^{(\boldsymbol{\lambda})}, \tag{121}$$

and

$$p_m = \sum_{t=1}^{T} q^{(t)} p_m^{(t)}. \tag{122}$$

对于 $t \in [T]$ 和 $\theta \in \Theta$，有

$$\begin{aligned}
&\left\| \nabla \mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}(\theta) - \sum_{m=1}^{M} p_m^{(t)} \nabla \mathcal{L}_{\mathcal{M}_m^{(t)}}^{(\boldsymbol{\lambda})}(\theta) \right\|^2 \\
&\quad = \left\| \nabla \mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}(\theta) - \sum_{m=1}^{M} p_m^{(t)} \nabla \mathcal{L}_{\mathcal{S}_m}^{(\boldsymbol{\lambda})}(\theta) + \sum_{m=1}^{M} p_m^{(t)} \nabla \mathcal{L}_{\mathcal{S}_m}^{(\boldsymbol{\lambda})}(\theta) - \sum_{m=1}^{M} p_m^{(t)} \nabla \mathcal{L}_{\mathcal{M}_m^{(t)}}^{(\boldsymbol{\lambda})}(\theta) \right\|^2 && (123) \\
&\quad \le 2 \left\| \nabla \mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}(\theta) - \sum_{m=1}^{M} p_m^{(t)} \nabla \mathcal{L}_{\mathcal{S}_m}^{(\boldsymbol{\lambda})}(\theta) \right\|^2 + 2 \left\| \sum_{m=1}^{M} p_m^{(t)} \nabla \mathcal{L}_{\mathcal{S}_m}^{(\boldsymbol{\lambda})}(\theta) - \sum_{m=1}^{M} p_m^{(t)} \nabla \mathcal{L}_{\mathcal{M}_m^{(t)}}^{(\boldsymbol{\lambda})}(\theta) \right\|^2 && (124) \\
&\quad = 2 \underbrace{\left\| \sum_{m=1}^{M} p_m^{(t)} \left(\nabla \mathcal{L}_{\mathcal{S}_m}^{(\boldsymbol{\lambda})}(\theta) - \nabla \mathcal{L}_{\mathcal{M}_m^{(t)}}^{(\boldsymbol{\lambda})}(\theta) \right) \right\|^2}_{\triangleq T_1} + 2 \underbrace{\left\| \sum_{m=1}^{M} \left(p_m - p_m^{(t)} \right) \cdot \nabla \mathcal{L}_{\mathcal{S}_m}^{(\boldsymbol{\lambda})}(\theta) \right\|^2}_{\triangleq T_2}. && (125)
\end{aligned}$$

界 T_1。我们有

$$\begin{aligned}
T_1 &= \left\| \sum_{m=1}^{M} p_m^{(t)} \left(\nabla \mathcal{L}_{\mathcal{S}_m}^{(\boldsymbol{\lambda})}(\theta) - \nabla \mathcal{L}_{\mathcal{M}_m^{(t)}}^{(\boldsymbol{\lambda})}(\theta) \right) \right\|^2 && (126) \\
&\le \sum_{m=1}^{M} p_m^{(t)} \left\| \nabla \mathcal{L}_{\mathcal{S}_m}^{(\boldsymbol{\lambda})}(\theta) - \nabla \mathcal{L}_{\mathcal{M}_m^{(t)}}^{(\boldsymbol{\lambda})}(\theta) \right\|^2 && (127) \\
&= \sum_{m=1}^{M} p_m^{(t)} \left\| \nabla \mathcal{L}_{\mathcal{S}_m}^{(\boldsymbol{\lambda})}(\theta) - \nabla \mathcal{L}_{\mathcal{P}_m}(\theta) + \nabla \mathcal{L}_{\mathcal{P}_m}(\theta) - \nabla \mathcal{L}_{\mathcal{M}_m^{(t)}}^{(\boldsymbol{\lambda})}(\theta) \right\|^2 && (128) \\
&\le 2 \sum_{m=1}^{M} p_m^{(t)} \left\| \nabla \mathcal{L}_{\mathcal{S}_m}^{(\boldsymbol{\lambda})}(\theta) - \nabla \mathcal{L}_{\mathcal{P}_m}(\theta) \right\|^2 + 2 \sum_{m=1}^{M} p_m^{(t)} \left\| \nabla \mathcal{L}_{\mathcal{P}_m}(\theta) - \nabla \mathcal{L}_{\mathcal{M}_m^{(t)}}^{(\boldsymbol{\lambda})}(\theta) \right\|^2. && (129)
\end{aligned}$$

有界 T_2。对于 $m' \in [m]$，我们有

$$\begin{aligned}
T_2 &= \left\| \sum_{m=1}^{M} \left(p_m - p_m^{(t)} \right) \cdot \nabla \mathcal{L}_{\mathcal{S}_m}^{(\boldsymbol{\lambda})}(\theta) \right\|^2 && (130) \\
&= \left\| \sum_{m=1}^{M} \left(p_m - p_m^{(t)} \right) \cdot \left(\nabla \mathcal{L}_{\mathcal{S}_m}^{(\boldsymbol{\lambda})}(\theta) - \nabla \mathcal{L}_{\mathcal{S}_{m'}}^{(\boldsymbol{\lambda})}(\theta) \right) \right\|^2 && (131) \\
&\le \sum_{m=1}^{M} \left(p_m - p_m^{(t)} \right)^2 \cdot \sum_{m=1}^{M} \left\| \nabla \mathcal{L}_{\mathcal{S}_m}^{(\boldsymbol{\lambda})}(\theta) - \nabla \mathcal{L}_{\mathcal{S}_{m'}}^{(\boldsymbol{\lambda})}(\theta) \right\|^2 && (132) \\
&= \sum_{m=1}^{M} \left(p_m - p_m^{(t)} \right)^2 \cdot \sum_{m=1}^{M} \left\| \nabla \mathcal{L}_{\mathcal{S}_m}^{(\boldsymbol{\lambda})}(\theta) - \nabla \mathcal{L}_{\mathcal{P}_m}(\theta) + \nabla \mathcal{L}_{\mathcal{P}_m}(\theta) - \nabla \mathcal{L}_{\mathcal{P}_{m'}}(\theta) + \nabla \mathcal{L}_{\mathcal{P}_{m'}}(\theta) - \nabla \mathcal{L}_{\mathcal{S}_{m'}}^{(\boldsymbol{\lambda})}(\theta) \right\|^2 && (133) \\
&\le 3 \sum_{m=1}^{M} \left(p_m - p_m^{(t)} \right)^2 \cdot \left(\sum_{m=1}^{M} \left\| \nabla \mathcal{L}_{\mathcal{S}_m}^{(\boldsymbol{\lambda})}(\theta) - \nabla \mathcal{L}_{\mathcal{P}_m}(\theta) \right\|^2 + \left\| \nabla \mathcal{L}_{\mathcal{S}_{m'}}^{(\boldsymbol{\lambda})}(\theta) - \nabla \mathcal{L}_{\mathcal{P}_{m'}}(\theta) \right\|^2 \right) \\
&\quad + 3 \sum_{m=1}^{M} \left(p_m - p_m^{(t)} \right)^2 \cdot \sum_{m=1}^{M} \left\| \nabla \mathcal{L}_{\mathcal{P}_m}(\theta) - \nabla \mathcal{L}_{\mathcal{P}_{m'}}(\theta) \right\|^2. && (134) \\
&\le 3 \sum_{m=1}^{M} \left(p_m - p_m^{(t)} \right)^2 \cdot \left(\sum_{m=1}^{M} \left\| \nabla \mathcal{L}_{\mathcal{S}_m}^{(\boldsymbol{\lambda})}(\theta) - \nabla \mathcal{L}_{\mathcal{P}_m}(\theta) \right\|^2 + \left\| \nabla \mathcal{L}_{\mathcal{S}_{m'}}^{(\boldsymbol{\lambda})}(\theta) - \nabla \mathcal{L}_{\mathcal{P}_{m'}}(\theta) \right\|^2 \right)
\end{aligned}$$

$$+ 3M\zeta^2 \sum_{m=1}^{M} \left(p_m - p_m^{(t)}\right)^2. \tag{135}$$

我们观察到

$$\nabla \mathcal{L}_{\mathcal{S}_m}^{(\boldsymbol{\lambda})}(\theta) = \sum_{i=1}^{N_m} \tilde{p}_{m,i} \nabla \ell(\theta; \boldsymbol{z}_m^{(i)}), \tag{136}$$

其中，对于 $i \in N_m$,

$$\tilde{p}_{m,i} = \frac{\sum_{t=1}^{T} \sum_{j \in \mathcal{I}_m} \mathbb{1}\{j = i\} \cdot \lambda_m^{(t,j)}}{\sum_{t=1}^{T} \sum_{j \in \mathcal{I}_m^{(t)}} \lambda_m^{(t,j)}}. \tag{137}$$

因此,

$$\mathop{\mathbb{E}}_{\mathcal{S}} \left[\left\| \nabla \mathcal{L}_{\mathcal{S}_m}^{(\boldsymbol{\lambda})}(\theta) - \nabla \mathcal{L}_{\mathcal{P}_m}(\theta) \right\|^2 \right] = \mathop{\mathbb{E}}_{\mathcal{S}_m} \left[\left\| \nabla \mathcal{L}_{\mathcal{S}_m}^{(\boldsymbol{\lambda})}(\theta) - \nabla \mathcal{L}_{\mathcal{P}_m}(\theta) \right\|^2 \right] \tag{138}$$

$$= \mathop{\mathbb{E}}_{\mathcal{S}_m} \left[\left\| \sum_{i=1}^{N_m} \tilde{p}_{m,i} \nabla \ell(\theta; \boldsymbol{z}_m^{(i)}) - \nabla \mathcal{L}_{\mathcal{P}_m}(\theta) \right\|^2 \right] \tag{139}$$

$$= \mathop{\mathbb{E}}_{\mathcal{S}_m} \left[\left\| \sum_{i=1}^{N_m} \tilde{p}_{m,i} \left(\nabla \ell(\theta; \boldsymbol{z}_m^{(i)}) - \nabla \mathcal{L}_{\mathcal{P}_m}(\theta) \right) \right\|^2 \right] \tag{140}$$

$$\leq \sum_{i=1}^{N_m} \tilde{p}_{m,i} \mathop{\mathbb{E}}_{\mathcal{S}_m} \left[\left\| \nabla \ell(\theta; \boldsymbol{z}_m^{(i)}) - \nabla \mathcal{L}_{\mathcal{P}_m}(\theta) \right\|^2 \right] \tag{141}$$

$$= \sum_{i=1}^{N_m} \tilde{p}_{m,i} \mathop{\mathbb{E}}_{\boldsymbol{z}_m^{(i)}} \left[\left\| \nabla \ell(\theta; \boldsymbol{z}_m^{(i)}) - \nabla \mathcal{L}_{\mathcal{P}_m}(\theta) \right\|^2 \right] \tag{142}$$

$$\leq \sum_{i=1}^{N_m} \tilde{p}_{m,i} \sigma_0^2 \tag{143}$$

$$= \sigma_0^2. \tag{144}$$

同理，我们可以证明

$$\mathop{\mathbb{E}}_{\mathcal{S}} \left\| \nabla \mathcal{L}_{\mathcal{P}_m}(\theta) - \nabla \mathcal{L}_{\mathcal{M}_m^{(t)}}^{(\boldsymbol{\lambda})}(\theta) \right\|^2 \leq \sigma_0^2. \tag{145}$$

最后，结合 (125)、(129)、(135)、(144) 和 (145)，即可得出结论。

□

B.7 定理 4.4 的证明

定理 4.4。 在与定理 *4.1* 和定理 *4.3* 相同的假设下，

$$\epsilon_{true} \leq \mathcal{O}\left(\frac{1}{\sqrt{T}}\right) + \mathcal{O}\big(\bar{\sigma}(\boldsymbol{\lambda})\big) + 2\mathrm{disc}_{\mathcal{H}}\left(\mathcal{P}^{(\boldsymbol{\alpha})}, \mathcal{P}^{(\boldsymbol{p})}\right) + \tilde{O}\left(\sqrt{\frac{\mathrm{VCdim}(\mathcal{H})}{N_{\mathrm{eff}}}}\right).$$

证明。 利用式（9），该结果可直接由定理 4.1 和定理 4.3 推得。 □

C 案例研究

C.1 客户端间歇性可用

在第 5 节中，我们考虑了包含两组客户端的情形：M_{hist} 拥有“历史”数据集的客户端，其数据集在训练期间不会改变；以及 $M - M_{\text{hist}}$ 以恒定速率收集“新鲜”样本的客户端$\{b_m > 0, m \in [\![M_{\text{hist}} + 1, M]\!]\}$，由于内存限制，它们只存储最近的 b_m 个样本（即 $C_m = b_m$）。新客户端也可以表示客户端仅在单个通信轮次内可用的情形：此时会有 M_{hist} 个“永久”客户端，它们始终可用且在训练期间不会改变；以及 $M - M_{\text{hist}}$ 个“间歇性”客户端，每个客户端仅在一个或几个连续通信轮次内可用。

在第 5 节的设置中，每个客户端对其内存中的所有样本赋予相同的权重，且该权重不随时间变化；令 λ_m 为客户端 $m \in [M]$ 赋予其内存中当前样本的权重，即$\lambda_m^{(t,j)} = \lambda_m$ 对每个 $t \in [T]$ 和 $j \in \mathcal{I}_m^{(t)}$均成立。

提醒一下，客户端 $m \in [M]$ 收集的样本总数为 N_m。对于一个新鲜客户端，设其为 $m > M_{\text{hist}}$，$N_m = b_m T$。

C.2 一般情形

推论 5。1′. 考虑有 M_{hist} 个历史客户端和 $M - M_{hist}$ 个新客户端的情形。假设满足定理 ***4.4*** 中的相同假设，且算法 ***1*** 使用客户端聚合权重 $\boldsymbol{p} = (p_m)^{m\in[M]} \in \Delta^{M-1}$，则

$$\epsilon_{true} \leq \frac{(C_1 + C_3)}{\sqrt{T}} + \frac{C_2}{\sqrt{T^3}} + \left(D + \frac{2}{\sqrt{T}}\right)\sigma_0\sqrt{M - M_{hist}}\sqrt{\sum_{m=M_{hist}+1}^{M} p_m^2} + 2 \cdot \max_{m,m'} \operatorname{disc}\left(\mathcal{P}_m, \mathcal{P}_{m'}\right) \cdot \|\boldsymbol{\alpha} - \boldsymbol{p}\|_1$$
$$+ 4 \cdot \sqrt{1 + \log\left(\frac{N}{\operatorname{VCdim}(\mathcal{H})}\right)} \cdot \sqrt{\frac{\operatorname{VCdim}(\mathcal{H})}{N}} \cdot \sqrt{\sum_{m=1}^{M} \frac{p_m^2}{n_m}}, \tag{146}$$

其中 $C_1\, C_2$ 和 C_3 是在定理 ***4.3*** 的证明中定义的常数，而 σ_0 是在注释 ***1*** 中定义的。

证明。我们提醒读者

$$p_{m,i} = \frac{\sum_{t=1}^{T}\sum_{j\in\mathcal{I}_m^{(t)}} \mathbb{1}\{j = i\} \cdot \lambda_m^{(t,j)}}{\sum_{m'=1}^{M}\sum_{t=1}^{T}\sum_{j\in\mathcal{I}_{m'}^{(t)}} \lambda_{m'}^{(t,j)}}, \qquad i \in N_m^{(T)}, \tag{147}$$

and

$$p_m^{(t)} = \frac{\sum_{j\in\mathcal{I}_m^{(t)}} \lambda_m^{(t,j)}}{\sum_{m'=1}^{M}\sum_{j\in\mathcal{I}_{m'}^{(t)}} \lambda_{m'}^{(t,j)}}, \qquad t \in [T]. \tag{148}$$

代入 $\lambda_m^{(t,j)} = \lambda_m$，可得

$$p_{m,i} = \frac{\lambda_m \cdot \sum_{t=1}^{T}\sum_{j\in\mathcal{I}_m^{(t)}} \mathbb{1}\{j = i\}}{\sum_{m'=1}^{M} \lambda_{m'} \sum_{t=1}^{T} \left|\mathcal{I}_{m'}^{(t)}\right|}, \tag{149}$$

and,

$$p_m^{(t)} = \frac{\lambda_m \left|\mathcal{I}_m^{(t)}\right|}{\sum_{m'=1}^{M} \lambda_{m'} \left|\mathcal{I}_{m'}^{(t)}\right|}. \tag{150}$$

在推论 5.1′ 的设定下，我们有

$$\mathcal{I}_m^{(t)} = \begin{cases} \{1, \ldots, N_m\} & , \quad m \in \{1, \ldots, M_{\text{hist}}\} \\ \{(t-1)\cdot b_m + 1, \ldots, t \cdot b_m - 1\} & , \quad m \in \{M_{\text{hist}} + 1, \ldots, M\}. \end{cases} \tag{151}$$

因此,

$$p_m^{(t)} = \frac{N_m\lambda_m \cdot \mathbb{1}\left\{m \in [\![1, M_{\text{hist}}]\!]\right\} + b_m\lambda_m \cdot \mathbb{1}\left\{m \in [\![M_{\text{hist}}+1, M]\!]\right\}}{\sum_{m'=1}^{M_{\text{hist}}} N_{m'}\lambda_{m'} + \sum_{m'=M_{\text{hist}}+1}^{M} b_{m'}\lambda_{m'}}, \tag{152}$$

and

$$p_{m,i} = \frac{\lambda_m T \cdot \mathbb{1}\left\{m \in [\![1, M_{\text{hist}}]\!]\right\} + \lambda_m \cdot \mathbb{1}\left\{m \in [\![M_{\text{hist}}+1, M]\!]\right\}}{\sum_{m'=1}^{M} N_{m'}\lambda_{m'}}. \tag{153}$$

因此，对于每个样本 $i \in [N_m]$，$p_{m,\ i} = \frac{p_m}{N_m}$。

界 $\text{disc}_{\mathcal{H}}\left(\mathcal{P}^{(\boldsymbol{\alpha})},\ \mathcal{P}^{(\boldsymbol{p})}\right)$ 令 $m' \in [M]$，我们有

$$\text{disc}_{\mathcal{H}}\left(\mathcal{P}^{(\boldsymbol{\alpha})}, \mathcal{P}^{(\boldsymbol{p})}\right) = \sup_{\theta\in\Theta}\left|\sum_{m=1}^{M}(\alpha_m - p_m)\cdot\mathcal{L}_{\mathcal{P}_m}(\theta)\right| \tag{154}$$

$$= \sup_{\theta\in\Theta}\left|\sum_{m=1}^{M}(\alpha_m - p_m)\cdot\left(\mathcal{L}_{\mathcal{P}_m}(\theta) - \mathcal{L}_{\mathcal{P}_{m'}}(\theta)\right)\right|, \tag{155}$$

其中，最后一个等式由以下事实得出：$\sum_{m=1}^{M}\alpha_m = \sum_{m=1}^{M} p_m = 1$。对于所有 $m \in [M]$，我们有

$$(\alpha_m - p_m)\cdot\left(\mathcal{L}_{\mathcal{P}_m}(\theta) - \mathcal{L}_{\mathcal{P}_{m'}}(\theta)\right) \le |\alpha_m - p_m|\cdot\left|\mathcal{L}_{\mathcal{P}_m}(\theta) - \mathcal{L}_{\mathcal{P}_{m'}}(\theta)\right| \tag{156}$$

$$\le |\alpha_m - p_m|\cdot\sup_{\theta\in\Theta}\left|\mathcal{L}_{\mathcal{P}_m}(\theta) - \mathcal{L}_{\mathcal{P}_{m'}}(\theta)\right| \tag{157}$$

$$= |\alpha_m - p_m|\cdot\text{disc}_{\mathcal{H}}(\mathcal{P}_m, \mathcal{P}_{m'}) \tag{158}$$

$$\le |\alpha_m - p_m|\max_{m,m'}\text{disc}_{\mathcal{H}}(\mathcal{P}_m, \mathcal{P}_{m'}). \tag{159}$$

结合（155）和（159），我们有

$$\text{disc}_{\mathcal{H}}\left(\mathcal{P}^{(\boldsymbol{\alpha})}, \mathcal{P}^{(\boldsymbol{p})}\right) \le \sum_{m=1}^{M}|\alpha_m - p_m|\cdot\max_{m,m'}\text{disc}_{\mathcal{H}}(\mathcal{P}_m, \mathcal{P}_{m'}) \tag{160}$$

$$= \|\boldsymbol{\alpha} - \boldsymbol{p}\|_1\cdot\max_{m,m'}\text{disc}_{\mathcal{H}}(\mathcal{P}_m, \mathcal{P}_{m'}). \tag{161}$$

计算 N^{-1}有效 我们有 N^{-1}有效 $= \sum_{m=1}^{M}\sum_{i=1}^{N_m}\left(\frac{p_m}{N_m}\right)^2 = \sum_{m=1}^{M}\frac{p_m^2}{N_m} = \frac{1}{N}\sum_{m=1}^{M}\frac{p_m^2}{n_m}$。

界 $\bar{\sigma}(\boldsymbol{\lambda})$ 我们有

$$\bar{\sigma}^2(\boldsymbol{\lambda}) = \sum_{t=1}^{T} q^{(t)}\mathop{\mathbb{E}}_{\mathcal{S}}\left[\sup_{\theta\in\Theta}\left\|\nabla\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}(\theta) - \sum_{m=1}^{M} p_m^{(t)}\nabla\mathcal{L}_{\mathcal{M}_m^{(t)}}^{(\boldsymbol{\lambda})}(\theta)\right\|^2\right]. \tag{162}$$

在推论 5.1′、$q^{(t)} = 1/T$ 和 $p_m^{(t)} = p_m$ 的设定下，因此

$$\bar{\sigma}^2(\boldsymbol{\lambda}) = \frac{1}{T}\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\mathcal{S}}\left[\sup_{\theta\in\Theta}\left\|\nabla\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}(\theta) - \sum_{m=1}^{M} p_m\nabla\mathcal{L}_{\mathcal{M}_m^{(t)}}(\theta)\right\|^2\right], \tag{163}$$

其中 $\mathcal{L}_{\mathcal{M}_m^{(t)}} = \sum_{j\in\mathcal{I}_m^{(t)}}\ell\left(\cdot, \boldsymbol{z}_m^{(j)}\right)/\left|\mathcal{I}_m^{(t)}\right|$。此外，很容易验证，在此设定下，

$$\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})} = \frac{1}{T}\sum_{t=1}^{T}\sum_{m=1}^{M} p_m\cdot\mathcal{L}_{\mathcal{M}_m^{(t)}}. \tag{164}$$

此外，$\mathcal{M}_m^{(t)} = \mathcal{M}_m^{(1)}$对于 $m \in [M_{\text{hist}}]$，因此对于 $\theta \in \Theta$,

$$\nabla\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}(\theta) - \sum_{m=1}^{M} p_m\nabla\mathcal{L}_{\mathcal{M}_m^{(t)}}(\theta) = \sum_{m=M_{\text{hist}}+1}^{M} p_m\cdot\frac{1}{T}\sum_{s=1}^{T}\left(\nabla\mathcal{L}_{\mathcal{M}_m^{(s)}}(\theta) - \nabla\mathcal{L}_{\mathcal{M}_m^{(t)}}(\theta)\right). \tag{165}$$

由此可得,

$$\left\|\nabla\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}(\theta)-\sum_{m=1}^{M}p_m\nabla\mathcal{L}_{\mathcal{M}_m^{(t)}}(\theta)\right\|^2=\left\|\sum_{m=M_{\text{hist}}+1}^{M}p_m\cdot\frac{1}{T}\sum_{s=1}^{T}\left(\nabla\mathcal{L}_{\mathcal{M}_m^{(s)}}(\theta)-\nabla\mathcal{L}_{\mathcal{M}_m^{(t)}}(\theta)\right)\right\|^2 \tag{166}$$

$$\leq(M-M_{\text{hist}})\sum_{m=M_{\text{hist}}+1}^{M}p_m^2\left\|\frac{1}{T}\sum_{s=1}^{T}\left(\nabla\mathcal{L}_{\mathcal{M}_m^{(s)}}(\theta)-\nabla\mathcal{L}_{\mathcal{M}_m^{(t)}}(\theta)\right)\right\|^2 \tag{167}$$

$$\leq(M-M_{\text{hist}})\sum_{m=M_{\text{hist}}+1}^{M}\frac{p_m^2}{T}\sum_{t=1}^{T}\left\|\nabla\mathcal{L}_{\mathcal{M}_m^{(s)}}(\theta)-\nabla\mathcal{L}_{\mathcal{M}_m^{(t)}}(\theta)\right\|^2. \tag{168}$$

对于新客户端，即对于 $m>M_0$，我们有 $\mathcal{L}_{\mathcal{M}_m^{(t)}}(\theta)=\sum_{i=1}^{b_m}\ell(\theta,\ z_m^{(t,i)})/b_m$，因此

$$\mathbb{E}_{\mathcal{S}}\left\|\nabla\mathcal{L}_{\mathcal{M}_m^{(s)}}(\theta)-\nabla\mathcal{L}_{\mathcal{M}_m^{(t)}}(\theta)\right\|^2\leq\mathbb{E}_{\mathcal{S}}\left\|\frac{1}{b_m}\sum_{i=1}^{b_m}\nabla\ell\left(\theta;z_m^{(t,i)}\right)-\nabla\ell\left(\theta;z_m^{(s,i)}\right)\right\|^2 \tag{169}$$

$$\leq\frac{1}{b_m}\sum_{i=1}^{b_m}\mathbb{E}_{\mathcal{S}}\left\|\nabla\ell\left(\theta;z_m^{(t,i)}\right)-\nabla\ell\left(\theta;z_m^{(s,i)}\right)\right\|^2 \tag{170}$$

$$\leq\sigma_0^2. \tag{171}$$

因此,

$$\mathbb{E}_{\mathcal{S}}\left\|\nabla\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}(\theta)-\sum_{m=1}^{M}p_m\nabla\mathcal{L}_{\mathcal{M}_m^{(t)}}(\theta)\right\|^2\leq\sigma_0^2\,(M-M_{\text{hist}})\cdot\sum_{m=1}^{M}p_m^2 \tag{172}$$

结论 我们通过指出以下内容来完成证明：$\tilde{c}_0=(C_1+C_3)/\sqrt{T}+C_2/\sqrt{T^3}$,其中 C_1、C_2 和 C_3 是在定理 4.3 的证明中引入的常数。 □

(146) 中的第三项源于梯度随时间的变异性，该变异性由 (18) 中的 $\bar{\sigma}^2(\boldsymbol{\lambda})$ 表示。具体而言，它仅取决于新客户端的权重（因为历史客户端不存在梯度变异性）。(146) 中的第四项对应于 (18) 中目标分布 $\mathcal{P}^{(\boldsymbol{\alpha})}$ 与有效分布 $\mathcal{P}^{(\boldsymbol{p})}$ 之间的差异。正如预期的那样，当所有客户端具有相同的分布时，该项为零；对于给定的局部分布异质性，客户端的目标相对重要性与有效相对重要性越接近，该项就越小（即 $\boldsymbol{\alpha}$ 与 $\boldsymbol{p}$ 越接近）。最后，(146) 中的第五项对应于 (18) 中的 $\tilde{\mathcal{O}}\left(\sqrt{\text{VCdim}\,(\mathcal{H})/N_{\text{eff}}}\right)$ 项，因为 $N_{\text{eff}}=N/\left(\sum_{m=1}^{M}p_m^2/n_m\right)$ in this setting.

C.3 推论 5.1 的证明

推论 5.1。 考虑有 M_{hist} 个历史客户端和 $M-M_{hist}$ 个新客户端的情形。假设定理 *4.4* 的相同条件成立，且 $\boldsymbol{\alpha}=\boldsymbol{n}$，并且使用算法 *1*，采用客户端聚合权重 $\boldsymbol{p}=(p_m)_{m\in[M]}\in\Delta^{M-1}$，则

$$\epsilon_{true}\leq\psi(\boldsymbol{p};\boldsymbol{c})\triangleq c_0+c_1\cdot\sqrt{\sum_{m=M_{hist}+1}^{M}p_m^2}+c_2\cdot\sqrt{\sum_{m=1}^{M}\frac{p_m^2}{n_m}},$$

其中 $\boldsymbol{c}=(c_0,c_1,c_2)$ 是与 $\boldsymbol{p}$ 无关的非负常数，定义如下：

$$c_0=(C_1+C_3)+\frac{C_2}{T}$$

$$c_1=\sigma_0\sqrt{M-M_{hist}}\cdot\left(D+\frac{2}{\sqrt{T}}\right)$$

$$c_2 = 4 \cdot \sqrt{1 + \log\left(\frac{N}{\mathrm{VCdim}\left(\mathcal{H}\right)}\right)} \cdot \sqrt{\frac{\mathrm{VCdim}\left(\mathcal{H}\right)}{N}} + 2 \cdot \max_{m,m'} \mathrm{disc}\left(\mathcal{P}_m, \mathcal{P}_{m'}\right)$$

C_1、 C_2 和 C_3 是定理 ***4.3*** 证明中定义的常数，而 σ_0 定义于注释 ***1***。

证明。我们提醒读者，推论 5.1′ 意味着

$$\begin{aligned}\epsilon_{\text{true}} \leq \frac{(C_1 + C_3)}{\sqrt{T}} + \frac{C_2}{\sqrt{T^3}} + \left(D + \frac{2}{\sqrt{T}}\right)\sigma_0\sqrt{M - M_{\text{hist}}}\sqrt{\sum_{m=M_{\text{hist}}+1}^{M} p_m^2} + 2 \cdot \max_{m,m'} \mathrm{disc}\left(\mathcal{P}_m, \mathcal{P}_{m'}\right) \cdot \left\|\boldsymbol{n} - \boldsymbol{p}\right\|_1 \\ + 4 \cdot \sqrt{1 + \log\left(\frac{N}{\mathrm{VCdim}\left(\mathcal{H}\right)}\right)} \cdot \sqrt{\frac{\mathrm{VCdim}\left(\mathcal{H}\right)}{N}} \cdot \sqrt{\sum_{m=1}^{M} \frac{p_m^2}{n_m}}. \end{aligned} \tag{173}$$

利用以下事实即可得出该结果，我们将在下文证明这一点。

$$\left\|\boldsymbol{p} - \boldsymbol{n}\right\|_1 = \sum_{m=1}^{M} \left|p_m - n_m\right| \tag{174}$$

$$= \sum_{m=1}^{M} \frac{\left|p_m - n_m\right|}{\sqrt{n_m}} \cdot \sqrt{n_m} \tag{175}$$

$$\leq \sqrt{\sum_{m=1}^{M} \frac{\left(p_m - n_m\right)^2}{n_m} \cdot \sum_{m=1}^{M} n_m} \tag{176}$$

$$= \sqrt{\sum_{m=1}^{M} \frac{\left(p_m - n_m\right)^2}{n_m}} \tag{177}$$

$$= \sqrt{\sum_{m=1}^{M} \frac{p_m^2}{n_m} - 2\sum_{m=1}^{M} \frac{p_m n_m}{n_m} + \sum_{m=1}^{M} \frac{n_m^2}{n_m}} \tag{178}$$

$$= \sqrt{\sum_{m=1}^{M} \frac{p_m^2}{n_m} - 1}, \tag{179}$$

其中，我们使用柯西-施瓦茨不等式对 $\sum_{m=1}^{M} \frac{\left|p_m - n_m\right|}{\sqrt{n_m}} \cdot \sqrt{n_m}$ 进行界定。 □

C.4 ψ 的凸性证明

我们提醒读者，对于 $\boldsymbol{p} \in \Delta^{M-1}$ 和 $\boldsymbol{c} \in \mathbb{R}_+^3$，有

$$\psi(\boldsymbol{p}; \boldsymbol{c}) = \frac{c_0}{\sqrt{T}} + c_1 \cdot \sqrt{\sum_{m=M_{\text{hist}}+1}^{M} p_m^2} + c_2 \cdot \sqrt{\sum_{m=1}^{M} \frac{p_m^2}{n_m}}. \tag{180}$$

为了证明 $\boldsymbol{p} \mapsto \sqrt{\sum_{m=1}^{M} \frac{p_m^2}{n_m}}$ 和 $\boldsymbol{p} \mapsto \sqrt{\sum_{m=M_{\text{hist}}}^{M} p_m^2}$ 的凸性，只需证明对于任意向量 $\boldsymbol{\beta} \in \mathbb{R}_+^M$，函数$\varphi_{\boldsymbol{\beta}}$: $\boldsymbol{p} \mapsto \sqrt{\sum_{m=1}^{M} \beta_m p_m^2}$ 是凸函数。令 $\boldsymbol{\beta} \in \mathbb{R}_+^M, \boldsymbol{p}, \tilde{\boldsymbol{p}} \in \Delta^M$ 和 $\gamma \in [0,1]$，则有

$$\varphi_{\boldsymbol{\beta}}^2\big(\gamma \cdot \boldsymbol{p} + (1-\gamma) \cdot \tilde{\boldsymbol{p}}\big) = \sum_{m=1}^{M} \beta_m \cdot \big(\gamma \cdot p_m + (1-\gamma) \cdot \tilde{p}_m\big)^2 \tag{181}$$

$$= \gamma^2 \cdot \sum_{m=1}^{M} \beta_m p_m^2 + (1-\gamma)^2 \cdot \sum_{m=1}^{M} \beta_m \tilde{p}_m^2 + 2\gamma(1-\gamma) \cdot \sum_{m=1}^{M} \beta_m p_m \tilde{p}_m \tag{182}$$

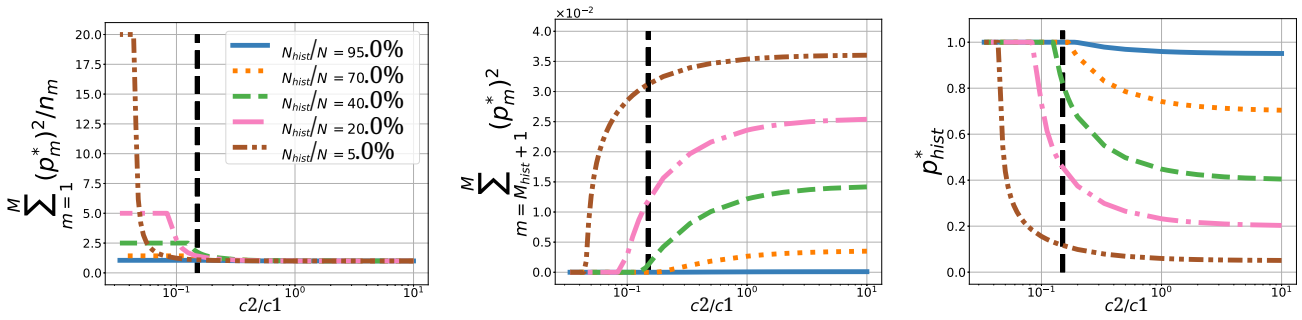


图 4：从左至右：c_2/c_1 对有效样本数、归一化梯度噪声以及历史客户端相对重要性的影响 p^*_{hist}，适用于 CIFAR-10 数据集（$N = 5 \times 10^5$）和不同的 N_{hist}/N取值，此时$M = 50$，且 $M_{\text{hist}} = 25$。虚线对应于我们对 CIFAR-10 实验中 c_2/c_1 的估计（$\hat{c}_2/\hat{c}_1 = 0.15$）。

$$\leq \gamma^2 \cdot \sum_{m=1}^{M} \beta_m p_m^2 + (1-\gamma)^2 \cdot \sum_{m=1}^{M} \beta_m \tilde{p}_m^2 + 2\gamma(1-\gamma) \cdot \sqrt{\sum_{m=1}^{M} \beta_m p_m^2} \cdot \sqrt{\sum_{m=1}^{M} \beta_m \tilde{p}_m^2} \tag{183}$$

$$= \left(\gamma \cdot \sqrt{\sum_{m=1}^{M} \beta_m p_m^2} + (1-\gamma) \cdot \sqrt{\sum_{m=1}^{M} \beta_m \tilde{p}_m^2} \right)^2 \tag{184}$$

$$= \left(\gamma \cdot \varphi_{\boldsymbol{\beta}}(\boldsymbol{p}) + (1-\gamma) \cdot \varphi_{\boldsymbol{\beta}}(\tilde{\boldsymbol{p}}) \right)^2, \tag{185}$$

其中我们使用柯西-施瓦茨不等式对 $\sum_{m=1}^{M} \beta_m p_m \tilde{p}_m$ 进行界定，如下所示

$$\sum_{m=1}^{M} \beta_m p_m \tilde{p}_m = \sum_{m=1}^{M} \left(p_m \sqrt{\beta_m} \right) \cdot \left(\tilde{p}_m \sqrt{\beta_m} \tilde{p}_m \right) \leq \sqrt{\sum_{m=1}^{M} \beta_m p_m^2} \cdot \sqrt{\sum_{m=1}^{M} \beta_m \tilde{p}_m^2}. \tag{186}$$

由于 $\varphi_{\boldsymbol{\beta}}$ 是一个非负函数，我们有

$$\varphi_{\boldsymbol{\beta}}\left(\gamma \cdot \boldsymbol{p} + (1-\gamma) \cdot \boldsymbol{p}\right) \leq \gamma \cdot \varphi_{\boldsymbol{\beta}}(\boldsymbol{p}) + (1-\gamma) \cdot \varphi_{\boldsymbol{\beta}}(\tilde{\boldsymbol{p}}), \tag{187}$$

这证明了 $\varphi_{\boldsymbol{\beta}}$ 是凸的。

C.5 界限最小化的数值研究

图 4 展示了在特定条件 $M = 50$ 和 $M_{\text{hist}} = 25$下，解及重要系统量如何随比率 c_2/c_1和历史样本占比 N_{hist}/N变化。除具体数值外，还可以区分出两种极端情况。当 $c_2/c_1 \gg 1$时，最优解对应于最小化 $\sum_{m=1}^{M} p_m^2/n_m$，即最大化有效样本数，此时 $\sum_m (p_m^*)^2/n_m$。此时，最优聚合向量 $\boldsymbol{p}^*$ 为均匀策略：在整个训练过程中，每个样本都被赋予相同的重要性，而每个客户端的重要性则与其样本数成正比（$p_m^* = n_m$）。特别是，历史客户端的总相对重要性为 $p^*_{\text{hist}} = N_{\text{hist}}/N$。相反，当 $c_2/c_1 \ll 1$时，最优解对应于最小化 $\sum_{m>M_{\text{hist}}} p_m$，即梯度变异性。此时历史策略为最优：忽略新客户端，历史客户端获得与其本地数据集大小成正比的相对重要性（即，$p_m^* = N_m/N_{\text{hist}} = \frac{N}{N_{\text{hist}}} n_m$ 对应于 $m \in [M_{\text{hist}}]$ 和$p^*_{\text{hist}} = 1$）。图 4 验证了这些定性判断，同时也表明，两种状态之间的转换取决于 N_{hist}/N； N_{hist}/N越小，转换发生时的 c_2/c_1值也越小。

C.6 关于估计 c_2/c_1的细节

根据推论 5.1 中 c_1 和 c_2 的表达式，我们有

$$\frac{c_2}{c_1} = 2 \cdot \frac{\max_{m,m'} \operatorname{disc}\left(\mathcal{P}_m, \mathcal{P}_{m'}\right) + 2 \cdot \sqrt{1 + \log\left(\frac{N}{\operatorname{VCdim}(\mathcal{H})}\right)} \cdot \sqrt{\frac{\operatorname{VCdim}(\mathcal{H})}{N}}}{\sigma_0 \sqrt{M - M_{\text{hist}}} \cdot \left(D + \frac{2}{\sqrt{T}}\right)}. \tag{188}$$

我们使用以下近似

$$\sqrt{1 + \log\left(\frac{N}{\operatorname{VCdim}\left(\mathcal{H}\right)}\right)} \approx 1, \tag{189}$$

$$D + \frac{2}{\sqrt{T}} \approx D, \tag{190}$$

$$4\operatorname{VCdim}\left(\mathcal{H}\right) \approx d, \tag{191}$$

其中， d 是模型 $\theta \in \Theta \subset \mathbb{R}^d$ 的参数数量（见第 3 节）。我们回顾注释 1 中 σ_0 的定义：

$$\sigma_0 = \sqrt{\max_m \mathop{\mathbb{E}}_{\boldsymbol{z} \sim \mathcal{P}_m} \left[\sup_{\theta \in \Theta} \left\|\nabla \ell(\theta; \boldsymbol{z}) - \nabla \mathcal{L}_{\mathcal{P}_m}\left(\theta\right)\right\|^2\right]} \leq 2\sqrt{2} \cdot LB = 2G, \tag{192}$$

其中， G 在（60）中定义。我们使用近似 $\sigma_0 \approx 2G$。最后，我们指出 $\max_{m,\ m'} \operatorname{disc}\left(\mathcal{P}_m, \mathcal{P}_{m'}\right) \leq B$，因此我们将 c_2/c_1 近似为

$$\frac{\hat{c}_2}{\hat{c}_1} \approx \frac{B + \sqrt{d/N}}{GD\sqrt{M - M_{\text{hist}}}}. \tag{193}$$

在我们的实验中，客户端使用各自历史样本的一部分，协作估计 $\hat{c}_2/\hat{c}_1$。其特别之处在于， D 被估计为 $\hat{D} = \max_{m=1}^{M} \left\|\hat{\theta}_m^* - \theta^{(1)}\right\|$，其中 $\hat{\theta}_m^*$ 是使用客户端 $m \in [M]$ 的部分历史数据，经过少量随机梯度下降迭代后得到的模型。

表 3： 在 $N_{\text{hist}}/N = 50\%$的设置下，不同数据集上各客户端的平均测试准确率。

数据集	D	G	B	d
合成	1.9	0.4	0.7	21
CIFAR-10	1.0	5.5	2.3	$3,353,034$
CIFAR-100	1.0	4.7	4.6	$3,537,444$
FEMNIST	5.9	12.9	3.5	$867,390$
莎士比亚	2.6	1.4	6.1	$226,180$

D 实验设置详情

D.1 数据集和模型

本节将详细介绍我们实验中使用的数据集和模型。我们考虑了五个用于不同机器学习任务的联邦基准数据集：图像分类（CIFAR-10 和 CIFAR-100（Krizhevsky，2009年））、手写字符识别（FEMNIST（Caldas 等人，2018年））和语言建模（莎士比亚（Caldas 等人，2018年；McMahan 等人，2017年）），以及附录 D.1.1 中描述的一个合成数据集。对于莎士比亚和 FEMNIST 数据集，可以通过客户端以自然的方式划分数据（分别按角色和书写者划分）。对于 CIFAR-10 和 CIFAR-100，我们采用了文献中的常见方法来抽取异构本地数据集。下文将详细介绍这些数据集，以及针对每个数据集所考虑的模型/任务。

D.1.1 合成数据集

我们的合成数据集生成过程如下：

1. 从均值为零、方差为单位值的 d维多元正态分布中采样 $\theta_0 \in \mathbb{R}^d \sim \mathcal{N}(0,\ I_d)$。2. 从以 θ_0 为中心、方差等于 ε^2的 d维多元正态分布中采样 $\theta_m \in \mathbb{R}^d \sim \mathcal{N}(\theta_0,\ \varepsilon^2 I_d)\ m \in [M]$。3. 对于 $m \in [M]$ 和 $i \in [N_m]$，从 $[-1,1]^d$上的均匀分布中采样 $\boldsymbol{x}_m^{(i)} \sim \mathcal{U}\left([-1,1]^d\right)$。4. 对于 $m \in [M]$ 和 $i \in [N_m]$，采样 $y_m^{(i)} \sim \mathcal{B}\left(\text{sigmoid}\ \left(\langle \boldsymbol{x}_t^{(i)}, \theta_m \rangle\right)\right)$，其中 $\mathcal{B}$ 是标准伯努利分布

D.1.2 CIFAR-10/ CIFAR-100

我们按照 (Wang 等人，2020年) 的做法，根据参数为 0.4 的对称狄利克雷分布，将标签相同的样本分配给各个客户端，从而创建了联邦版 CIFAR-10。对于 CIFAR-100，我们利用“粗粒度”和“细粒度”标签，采用两阶段 Pachinko 分配方法 (Li 和 McCallum，2006) 将样本分配给各个客户端，做法与 (Reddi 等人，2021年) 相同。我们使用浅层卷积神经网络训练 CIFAR-10/100 数据集。

D.1.3 FEMNIST

FEMNIST（联邦扩展 MNIST）是一个包含 62 个类别的图像分类数据集，它将扩展 MNIST 的数据按照数字/字符的书写者进行划分而构建。我们使用两层全连接神经网络训练 FEMNIST 数据集

D.1.4 莎士比亚

莎士比亚数据集是一个语言建模数据集，由威廉·莎士比亚的全集构建而成。在这个数据集中，每个客户端对应一个至少有两句台词的说话角色。任务是预测下一个字符。我们使用一个循环神经网络，先将一系列字符作为输入，再将每个字符嵌入一个通过学习得到的 8维空间中。随后，嵌入后的字符会经过 2 个循环神经网络层，每层包含 256 个节点，最后连接一个全连接 Softmax 输出层。我们将每个说话角色的台词划分为长度为 80 个字符的序列，必要时进行填充。

D.2 训练详情。

在所有实验中，我们使用验证集在网格 $\{10^{-3.5}, 10^{-3}, 10^{-2.5}, 10^{-2}, 10^{-1.5}, 10^{-1}\}$ 上通过网格搜索调整学习率。选定学习率后，我们在训练集和验证集的合并数据上重新训练模型。每项实验使用随机数生成器的三个不同种子重复进行；我们报告均值和 95% 置信界。

D.3 到达过程

对于 CIFAR-10/100 数据集，我们考虑这样一种到达过程： $M_{\text{hist}} = 25$ 个客户端拥有"历史"数据集，这些数据集在训练期间不会发生变化；另有 $M - M_{\text{hist}} = 25$ 个客户端以恒定速率收集"新鲜"样本$\{b_m > 0, m \in [\![M_{\text{hist}} + 1, M]\!]\}$，并且由于内存限制，仅存储最近的 b_m 个样本（即 $C_m = b_m$）。对于给定的 N_{hist}/N值，我们将原始 CIFAR-10/100 的训练部分划分为历史组和新鲜组，分别包含 N_{hist} 和 $N - N_{\text{hist}}$ 个样本。随后，我们将历史组（相应地，新鲜组）中的样本分配到M_{hist} 个历史客户端（相应地， $M - M_{\text{hist}}$ 个新鲜客户端）上。对于 CIFAR-10，采用对称狄利克雷分布；对于 CIFAR-100，则采用 Pachinko 分配方法。

莎士比亚数据集和 FEMNIST 数据集天然地按字符和书写者划分为不同客户端。在我们的实验中，我们将 FEMNIST 和莎士比亚数据集的自然客户端分为历史组和新鲜组，分别包含 M_{hist} and$M - M_{\text{hist}}$ 个客户端。历史客户端参与每一轮通信，而在 FEMNIST 的情况下，每个新鲜客户端只在单个通信轮次中可用；对于莎士比亚数据集，每个新鲜客户端最多可在连续两个通信轮次中可用。

D.4 $\hat{c}_2/\hat{c}_1$ 的数值

表 3 给出了用于估计比值$\hat{c}_2/\hat{c}_1$ 的 D、 G、 B 和 d and 的取值。

表4： 在 $N_{\text{hist}}/N = 5\%$的设置下，不同数据集的客户端平均测试准确率。

数据集	$\hat{c}_2/\hat{c}_1$	$p^*_{\text{历史}}$	测试准确率				
			新鲜	历史	均匀策略	OURS	最优
合成	0.094	0.06	82.4 ± 1.89	68.1 ± 2.39	$\mathbf{82.7} \pm 1.94$	$\mathbf{82.7} \pm 1.90$	82.9 ± 2.17
CIFAR-10	0.150	0.12	59.5 ± 0.77	48.2 ± 0.21	60.7 ± 0.58	$\mathbf{61.0} \pm 0.42$	63.7 ± 0.57
CIFAR-100	0.284	0.08	23.5 ± 0.65	13.5 ± 0.41	24.4 ± 0.54	$\mathbf{25.2} \pm 0.66$	27.8 ± 0.39
FEMNIST	0.001	1.00	55.2 ± 1.79	$\mathbf{65.7} \pm 0.09$	58.4 ± 1.80	$\mathbf{65.7} \pm 0.09$	65.7 ± 0.09
莎士比亚	0.064	1.00	40.2 ± 0.34	$\mathbf{49.0} \pm 0.06$	41.0 ± 1.33	$\mathbf{49.0} \pm 0.06$	49.0 ± 0.06

表 5： 在 $N_{\text{hist}}/N = 50\%$的设置下，不同数据集的客户端平均测试准确率。

数据集	$\hat{c}_2/\hat{c}_1$	p_{HIST}	测试准确率				
			新鲜	历史	均匀策略	OURS	最优
合成	0.085	0.50	84.2 ± 1.27	84.8 ± 1.58	$\mathbf{86.5} \pm 1.20$	$\mathbf{86.5} \pm 1.20$	86.5 ± 1.20
CIFAR-10	0.150	0.95	52.1 ± 2.98	64.1 ± 5.60	65.1 ± 0.66	$\mathbf{68.7} \pm 0.37$	69.4 ± 0.25
CIFAR-100	0.284	0.69	17.5 ± 0.57	29.4 ± 1.40	29.7 ± 0.55	$\mathbf{34.4} \pm 0.31$	34.4 ± 0.31
FEMNIST	0.001	1.00	48.3 ± 2.98	$\mathbf{66.2} \pm 0.23$	57.8 ± 1.93	$\mathbf{66.2} \pm 0.23$	66.2 ± 0.23
莎士比亚	0.095	1.00	30.9 ± 0.51	$\mathbf{44.1} \pm 0.27$	41.1 ± 0.56	$\mathbf{44.1} \pm 0.27$	44.1 ± 0.27

E 补充实验结果

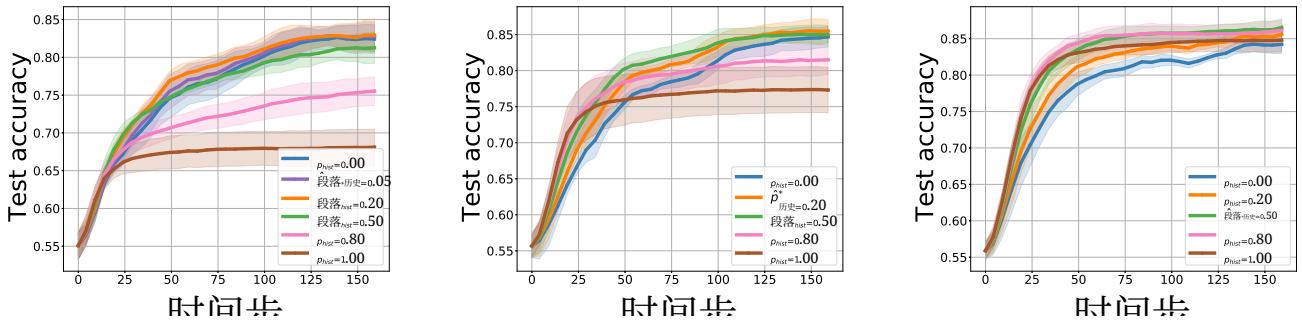


图 5：合成数据集在使用不同 p_{hist} 值时测试准确率的变化，其中 $N_{\text{hist}}/N = 5\%$（左）、20%（中）和 50%（右）。

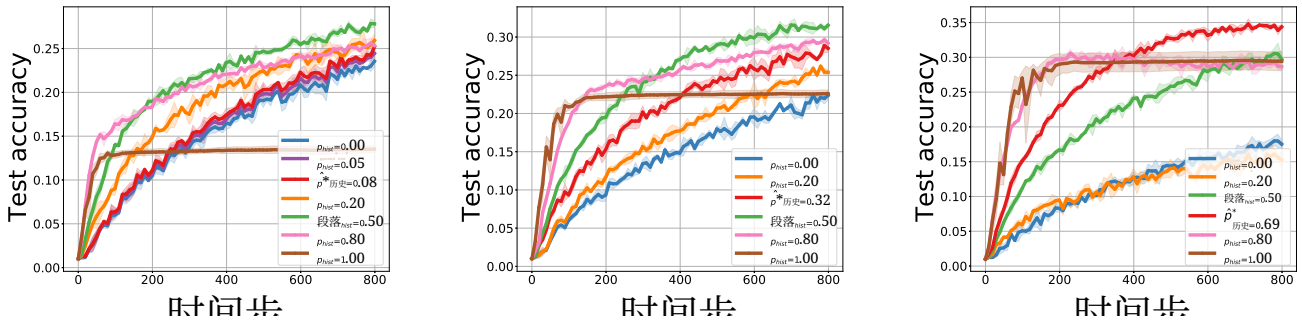


图 6：CIFAR-100 数据集在使用不同 p_{hist} 值时测试准确率的变化，其中 $N_{\text{hist}}/N = 5\%$（左）、20%（中）和 50%（右）。

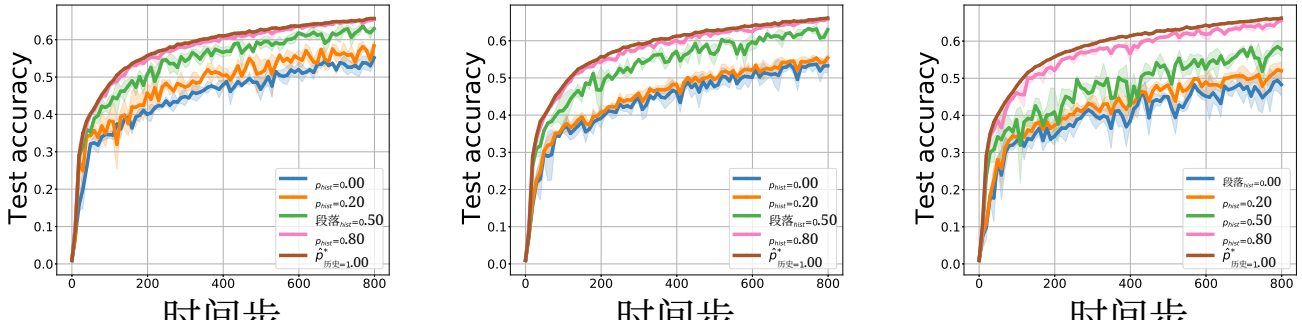


图 7：FEMNIST 数据集在使用不同 p_{hist} 值时测试准确率的变化，其中 $M_{\text{hist}}/M = 5\%$（左）、20%（中）和 50%（右）。

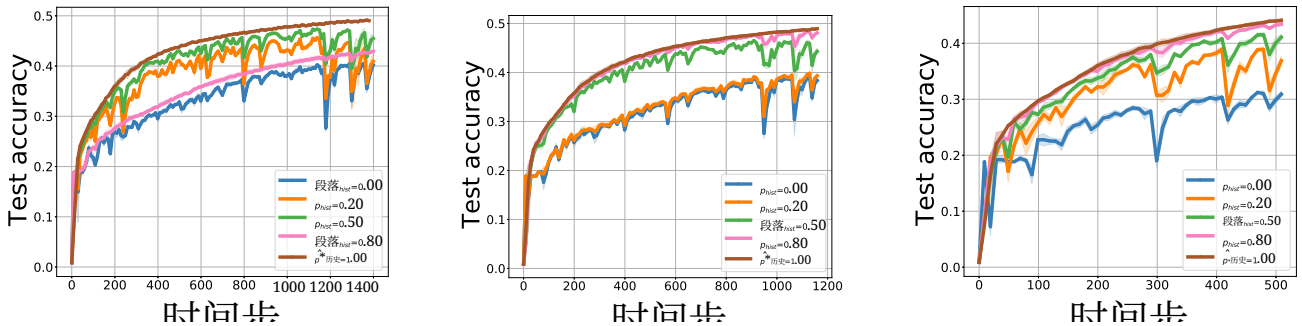


图 8：莎士比亚数据集在使用不同 p_{hist} 值时测试准确率的变化，其中 $M_{\text{hist}}/M = 5\%$（左）、20%（中）和 50%（右）。