arXiv:2303.05206v1 [cs.LG] 9 Mar 2023

FedREP：一种拜占庭鲁棒、通信高效且隐私保护的联邦学习框架

Yi-Rui Yang YANGYR@SMAIL.NJU.EDU.CN
Kun Wang WANGK@SMAIL.NJU.EDU.CN
Wu-Jun Li* LIWUJUN@NJU.EDU.CN
新软件技术国家重点实验室，计算机科学与技术系，南京大学，中国

摘要

联邦学习（FL）近年来已成为热门研究课题，其中拜占庭鲁棒性、通信效率和隐私保护是三个重要方面。然而，这三个方面之间存在矛盾，因此难以兼顾。针对这一挑战，我们从理论上分析了通信压缩方法若要兼容现有的拜占庭鲁棒方法和隐私保护方法，必须满足哪些条件。受分析结果启发，我们提出了一种新型通信压缩方法，称为共识稀疏化（ConSpar）。据我们所知，ConSpar 是首个旨在同时兼容拜占庭鲁棒方法和隐私保护方法的通信压缩方法。在 ConSpar 的基础上，我们进一步提出了一种新型联邦学习框架 *FedREP*，兼具拜占庭鲁棒性、通信效率和隐私保护。我们从理论上证明了 FedREP 的拜占庭鲁棒性和收敛性。实证结果表明，FedREP 的表现显著优于通信高效的隐私保护基线方法。此外，与拜占庭鲁棒且通信高效的基线方法相比，FedREP 在实现准确率相当的同时，还具有隐私保护这一额外优势。

1. 引言

联邦学习（FL）让参与方（也称为客户端）在数据仍归其所有的同时协作训练学习模型，近来已成为热门研究课题（Konevcnˇy 等人，2016 年；McMahan 和 Ramage，2017 年）。与传统的基于数据中心的分布式学习（Haddadpour 等人，2019 年；Jaggi 等人，2014 年；Lee 等人，2017 年；Lian 等人，2017 年；Shamir 等人，2014 年；Sun 等人，2018 年；Yu 等人，2019a 年；Zhang 和 Kwok，2014 年；Zhao 等人，2017 年、2018 年；Zhou 等人，2018 年；Zinkevich 等人，2010 年）相比，在联邦学习应用中，服务提供商对客户端的控制较少，网络通常也不够稳定，带宽更低。此外，如果不采用隐私保护方法，参与方在联邦学习中也会面临隐私泄露的风险。因此，拜占庭鲁棒性、通信效率和隐私保护已成为联邦学习方法的三个重要方面（Kairouz 等，2021 年），近年来受到广泛关注。

*. 通讯作者。

Ta表 1：不同方法在联邦学习三个方面的比较

方法	拜占庭鲁棒	通信高效	隐私保护
RCGD（Ghosh 等，2021年）	✓	✓	-
联邦学习（Wang 等，2020）	✓	-	✓
SHARE（Velicheti 等，2021）	✓	-	✓
SparseSecAgg（Ergun 等，2021）	-	✓	✓
FedREP（我们的方法）	✓	✓	✓

拜占庭鲁棒性。在联邦学习应用中，客户端或网络传输发生故障后，可能无法及时发现和解决（Kairouz 等，2021 年）。此外，一些客户端可能受到对抗方攻击，故意发送错误甚至有害的信息。发生故障或遭受攻击的客户端也称为拜占庭客户端。要实现对拜占庭客户端的鲁棒性，主要有三种不同的方式，分别称为冗余计算、服务器验证和鲁棒聚合。冗余计算方法（Chen 等人，2018 年；Konstantinidis 和 Ramamoorthy，2021 年；Rajput 等人，2019 年）要求不同客户端针对相同的训练样本计算梯度。这些方法主要适用于传统的基于数据中心的分布式学习，但由于隐私原则，在联邦学习中无法使用。在服务器验证方法中（Xie 等人，2019b、2020b），服务器依据公共数据集验证客户端更新。然而，服务器验证方法的性能取决于训练样本的数量和质量。在许多场景下，很难获得大规模、高质量的公共数据集。第三种方式是用鲁棒聚合取代服务器上的均值聚合（Alistarh 等人，2018 年；Bernstein 等人，2019 年；Blanchard 等人，2017 年；Chen 等人，2017 年；Ghosh 等人，2020 年；Karimireddy 等人，2021 年；Li 等人，2019 年；Sohn 等人，2020 年；Yin 等人，2018 年、2019 年）。与冗余计算和服务器验证相比，鲁棒聚合通常适用范围更广。许多拜占庭鲁棒联邦学习方法（Wang 等，2020；Xie 等人，2019a）都采用了这种方式。

通信效率。在许多联邦学习应用中，服务器与客户端通过广域网（WAN）连接。与传统基于数据中心的分布式机器学习所使用的网络相比，广域网通常稳定性较差、带宽较低。因此，也应考虑通信成本。本地更新技术（Konevcn`y 等人，2016年；McMahan 等人，2017年；Yu 等人，2019b；Zhao 等人，2017年、2018年）被广泛应用于联邦学习方法中：客户端在全局聚合之前先在本地更新模型若干次迭代。通信成本还可以通过通信压缩技术降低，这些技术主要包括量化（Alistarh 等人，2017年；Faghri 等人，2020年；Gandikota 等人，2021年；Safaryan 和 Richt´arik，2021年；Seide 等人，2014年；Wen 等人，2017年）、稀疏化（Aji 和 Heafield，2017年；Chen 等人，2020年；Stich 等人，2018年；Wangni 等人，2018年）和草图化[1] （Rothchild 等人，2020年）。误差补偿（亦称误差反馈）技术（Gorbunov 等人，2020年；Wu 等人，2018年；Xie 等人，2020c）旨在缓解通信压缩方法导致的精度下降。此外，不同技术还可以结合使用，进一步降低通信成本（Basu 等人，2020年；Lin 等人，2018年）。

1. 草图化可通过多种方式降低通信成本或保护隐私。因此，草图化同时出现在通信高效方法和隐私保护方法中。

隐私保护。大多数现有的联邦学习方法在训练过程中会发送梯度或模型参数，同时依据隐私原则将数据保留在本地。然而，发送梯度或模型参数也可能导致隐私泄露问题（Kairouz 等，2021 年；Zhu 等人，2019年）。一些隐私保护技术采用随机噪声来隐藏真实输入值，例如差分隐私（差分隐私）（Abadi 等人，2016年；Jayaraman 等人，2018年；McMahan 等人，2018年）和草图化（Liu 等人，2019年；Zhang 和 Wang，2021年）。安全聚合（安全聚合）（Bonawitz 等人，2017年；Choi 等人，2020年）旨在确保计算隐私。安全聚合基于安全多方计算（安全多方计算）和 Shamir 的 t-out-of-n 秘密共享（Shamir，1979年），使服务器在不了解每个客户端的本地模型参数（或梯度）的情况下，仅能获取用于全局模型更新的平均值。由于在大多数现有联邦学习方法中，只需向随机梯度添加噪声即可实现输入隐私，因此本文主要关注如何将安全聚合与拜占庭鲁棒、通信高效的方法相结合。

此外，还有一些方法兼顾三个方面中的两个方面（拜占庭鲁棒性、通信效率和隐私保护），包括 RCGD（Ghosh 等，2021）、F^2ed-Learning（Wang 等，2020）、SHARE（Velicheti 等，2021）和 SparseSecAgg（Ergun 等，2021），我们在表1中对这些方法进行了总结。然而，这三个方面之间的张力使得同时兼顾所有三个方面变得困难。鉴于这一挑战，我们从理论上分析了拜占庭鲁棒性、通信效率和隐私保护之间的张力，并提出了一种名为 FedREP 的新型联邦学习框架。主要贡献如下：

- 我们从理论上分析了通信压缩方法要与拜占庭鲁棒方法和隐私保护方法兼容所需满足的条件。

- 受分析结果启发，我们提出了一种名为 共识稀疏化 （ConSpar）的新型通信压缩方法。据我们所知，Con- Spar 是首个旨在同时兼容拜占庭鲁棒方法和隐私保护方法的通信压缩方法。

- 基于 ConSpar，我们进一步提出了一种名为$FedREP$的新型联邦学习框架，该框架具有拜占庭鲁棒性、通信高效且保护隐私。 _ _

- 我们从理论上证明了 FedREP 的拜占庭鲁棒性和收敛性。

- 我们的实证结果表明，FedREP 的表现显著优于现有的通信高效、隐私保护基线方法。此外，与拜占庭鲁棒、通信高效的基线方法相比，FedREP 在实现相当准确率的同时，还具有隐私保护这一额外优势。

2. 预备知识

在本文中，我们主要关注由 m 个客户端和单个服务器组成的传统联邦学习设置（Kairouz 等，2021年），它们协作求解有限和

优化问题:

$$\min_{\mathbf{w}\in\mathbb{R}^d} F(\mathbf{w}) = \sum_{k=1}^{m} p_k F_k(\mathbf{w})$$
$$\text{s.t.} \quad F_k(\mathbf{w}) = \frac{1}{|\mathcal{D}_k|} \sum_{i\in\mathcal{D}_k} f_i(\mathbf{w}),\ k = 1, 2, \ldots, m, \tag{1}$$

其中，$\mathbf{w}$ 是模型参数，d 是参数的维度。 f_i**(w)**是参数$\mathbf{w}$ 在第 i个训练样本上的经验损失。 $\mathcal{D}_k$ 表示存储在第 k个客户端上的样本索引集， F_k**(w)**是第 k个客户端的局部损失函数。我们假设当 $k \neq k'$时， $\mathcal{D}_k \cap \mathcal{D}_{k'} = \emptyset$；并将不同客户端上取值相同的样本视为多个不同的样本。 p_k 是第 k个客户端的权重，满足 $p_k > 0$且 $\sum_{k=1}^{m} p_k = 1$。 p_k 的一种常见设置是 $p_k = |\mathcal{D}_k|/(\sum_{k=1}^{m} |\mathcal{D}_k|)$。为简单起见，我们假设对于所有 $k, k' \in [m]$， $|\mathcal{D}_k| = |\mathcal{D}_{k'}|$，因此 $p_k = 1/m$。本文中的分析可以类似地扩展到一般情形。

大多数用于求解问题 **(1)** 的联邦学习方法（Karimireddy 等人，2020年；McMahan 等人，2017年、2018年）都基于分布式随机梯度下降及其变体。在这些方法中，客户端根据各自的训练样本在本地更新模型参数，然后在每次迭代中与服务器通信以进行模型聚合。然而，许多广泛使用的模型（Devlin 等人，2018年；He 等，2016年）规模非常庞大，导致通信成本高昂。因此，联邦学习需要采用降低通信成本的技术。此外，在实际应用中，联邦学习方法还应能抵御潜在的拜占庭攻击和隐私攻击。

拜占庭攻击。 令 $[m] = \{1, 2, \ldots, m\}$ 表示客户端集合。 $\mathcal{G} \subseteq [m]$ 表示良性（非拜占庭）客户端的集合，它们会忠实地执行算法。其余客户端 $[m] \setminus \mathcal{G}$ 为拜占庭客户端，可能会恶意行事并发送任意值。服务器通常由服务提供商控制，也会忠实地执行算法。此拜占庭攻击模型与许多先前研究中的模型一致（Karimireddy 等人，2021年）。尽管有些研究（Burkhalter 等人，2021年）关注另一类称为后门攻击的攻击（Kairouz 等，2021年），但本文主要关注拜占庭攻击。拜占庭攻击旨在降低模型性能。防御拜占庭攻击的一种典型技术是鲁棒聚合（Kairouz 等，2021年），即使拜占庭客户端发送错误值，也能保证聚合误差有界。

隐私攻击。 在典型的联邦学习方法中，服务器负责使用客户端本地更新值的平均值来更新全局模型。然而，本地更新信息可能被用于还原客户端的训练样本（Zhu 等人，2019年），从而增加隐私泄露的风险。因此，出于隐私保护的要求，服务器不得直接接收单个客户端的更新信息（Kairouz 等，2021 年）。安全聚合（Bonawitz 等人，2017年）是一种典型的隐私保护方法，它仅允许服务器获取平均值以更新全局模型。

联邦学习设置主要有两种不同类型，分别称为跨机构联邦学习和跨设备联邦学习（Kairouz 等，2021 年）。本文主要关注跨机构联邦学习设置，其中客户端数量 m通常不会太多，且所有客户端都可以参与

每次训练迭代。同时，本文主要关注同步联邦学习方法。

3. 方法论

本节分析通信压缩方法要与拜占庭鲁棒方法和隐私保护方法兼容所需满足的条件。基于分析，我们提出了一种名为共识稀疏化 的新型通信压缩方法，以及一种名为 $FedREP$ 的新型联邦学习框架，该框架具有拜占庭鲁棒性、通信高效和隐私保护的特点。在 FedREP 中，我们采用鲁棒聚合技术来实现拜占庭鲁棒性，因为与冗余计算和服务器验证相比，该技术的适用范围更广。为保护隐私，我们主要关注安全聚合；这是一种在联邦学习中广泛使用的技术，可确保服务器只能访问客户端本地更新值的平均值。

3.1 动机

我们首先分析现有通信压缩方法与安全聚合（SecAgg）（Bonawitz 等人，2017年）的兼容性。SecAgg 通常与量化结合使用，因为它需要在有限域上运算才能保证隐私保护。将每个坐标表示为较低位数的传统量化方法，最多只能将浮点数（32 位）梯度压缩至原始大小的 $1/32$。即使采用量化，SecAgg 仍面临高昂的通信成本。因此，需要进行稀疏化以进一步降低通信成本（Ergun 等，2021）。然而，如果我们将传统稀疏化方法（例如，随机-K 和 top-K稀疏化）直接与 SecAgg 结合，SecAgg 中的随机掩码会破坏稀疏性。因此，非拜占庭客户端应就非零坐标达成一致，以保持 SecAgg 中的稀疏性。

接下来，我们分析稀疏化与鲁棒聚合的兼容性。正如先前的研究（Karimireddy 等人，2021 年）所示，要实现拜占庭鲁棒性，不同客户端的压缩更新之间的距离（即客户端之间的差异）必须较小。具体而言，我们在定义 1 中给出 $(\delta,\ c)$-鲁棒聚合器的定义。

定义 *1*（*(δ, c)*-鲁棒聚合器（*Karimireddy* 等人，*2021* 年，*2022* 年））。假设常数 $\delta \in [0, \frac{1}{2})$以及索引集 $\mathcal{G} \subseteq [m]$ 满足 $|\mathcal{G}| \geq (1-\delta)m$。假设给定 m个随机向量 $v_1, \ldots, v_m \in \mathbb{R}^d$，使得对于任意固定的 $k, k' \in \mathcal{G}$，$\mathbb{E}\|v_k - v_{k'}\|^2 \leq \rho^2$。如果 $k \in [m] \setminus \mathcal{G}$，则 v_k 可以是任意值。若聚合器 $Agg(\cdot)$ 的聚合误差 $e = Agg(\{v_k\}_{k=1}^m) - \frac{1}{|\mathcal{G}|}\sum_{k\in\mathcal{G}} v_k$ 满足

$$\mathbb{E}\|\mathbf{e}\|^2 \leq c\delta\rho^2. \tag{2}$$

如已有研究所示（Karimireddy 等人，2022），许多广泛使用的聚合器，例如 Krum（Blanchard 等人，2017 年）、geoMed（Chen 等人，2017 年）和逐坐标中位数（Yin 等人，2018 年），与缓冲区内平均（请参阅第 3.3 节）结合后，满足定义 1。此外，$O(\delta\rho^2)$是最紧的阶（Karimireddy 等人，2021 年）。因此，与鲁棒聚合兼容的压缩方法应满足

压缩后客户端更新之间的差异期望仍然较小这一条件。因此，我们从理论上分析了稀疏化后的差异期望。为节省篇幅，此处仅给出结果。证明细节见附录 B。

定理 *1.* 设 $\{\mathbf{v}_k\}_{k=1}^m$ 表示满足 $\mathbb{E}\|\mathbf{v}_k - \mathbf{v}_{k'}\|^2 = (\rho_{k,k'})^2$ 且 $\mathbb{E}\|\mathbf{v}_k\|^2 = (\mu_k)^2$ 的随机向量，其中任意固定的 $k, k' \in \mathcal{G}$ 均成立。更具体地说，$\mathbb{E}[(\mathbf{v}_k)_j - (\mathbf{v}_{k'})_j]^2 = \xi_{k,k',j}(\rho_{k,k'})^2$ 且 $\mathbb{E}[(\mathbf{v}_k)_j^2] = \zeta_{k,j}(\mu_k)^2$, , 其中 $\xi_{k,k',j} > 0$, $\mu_{k,j} > 0$, $\sum_{j\in[d]} \xi_{k,k',j} = 1$ 且 $\sum_{j\in[d]} \zeta_{k,j} = 1$ 对任意固定的 $k, k' \in \mathcal{G}$ 均成立。设 $\mathcal{C}(\cdot)$ 表示任意稀疏化算子，$\mathcal{N}_k$ 表示 $\mathcal{C}(\mathbf{v}_k)$ 中非零坐标的集合。对于任意固定的 $k, k' \in \mathcal{G}$，有：

$$\begin{aligned}\mathbb{E}\|\mathcal{C}(\mathbf{v}_k) - \mathcal{C}(\mathbf{v}_{k'})\|^2 = (\rho_{k,k'})^2 \cdot \sum_{j\in[d]} \Big(\xi_{k,k',j}\Pr[j \in \mathcal{N}_k \cap \mathcal{N}_{k'}]\Big) \\ + \ (\mu_k)^2 \cdot \sum_{j\in[d]} \Big(\zeta_{k,j}\Pr[j \in \mathcal{N}_k \setminus \mathcal{N}_{k'}]\Big) \\ + \ (\mu_{k'})^2 \cdot \sum_{j\in[d]} \Big(\zeta_{k',j}\Pr[j \in \mathcal{N}_{k'} \setminus \mathcal{N}_k]\Big). \qquad (3)\end{aligned}$$

请注意，当第 k 个客户端与第 k' 个客户端之间的差异不太大时，$(\mu_k)^2$ 和 $(\mu_{k'})^2$ 通常远大于 $(\rho_{k,\ k'})^2$。在公式 (3) 中，当且仅当 $\mathcal{N}_k \setminus \mathcal{N}_{k'} = \mathcal{N}_{k'} \setminus \mathcal{N}_k = \emptyset$ 以概率 1 成立时，项 $(\mu_k)^2$ 和 $(\mu_{k'})^2$ 才会消失，这等价于 $\mathcal{N}_k = \mathcal{N}_{k'}$ 以概率 1 成立。此外，为了降低任意一对非拜占庭客户端之间的差异，所有非拜占庭客户端都应在稀疏化向量的非零坐标上达成一致。基于这两方面的分析结果，我们提出了共识稀疏化。

3.2 共识稀疏化

本节介绍共识稀疏化（ConSpar）方法。为简单起见，我们假设超参数 K 是客户端数量 m 的倍数。我们用$\mathbf{u}_k^t$ 表示客户端 k 在第 t 次迭代时用于误差补偿的本地记忆（Stich 等人，2018年）。初始时，$\mathbf{u}_k^0 = \mathbf{0}$。

令$\mathbf{g}_k^t$ 表示在第 t 次迭代时从客户端 k 发送的更新向量。客户端 k首先根据 top-$\frac{K}{m}$ 稀疏化准则生成坐标集合 $\mathcal{T}_k^t$。具体而言，$\mathcal{T}_k^t$包含$\mathbf{g}_k^t$中绝对值最大的 $\frac{K}{m}$ 个坐标。随后，通过从 $\mathcal{T}_k^t$ 中随机选取 $(\frac{K}{m} - r_k^t)$ 个元素，生成集合 $\tilde{\mathcal{T}}_k^t$，其中随机变量 r_k^t 服从二项分布 B($\frac{K}{m}$, α)（$0 \le \alpha \le 1$）。因此，$|\tilde{\mathcal{T}}_k^t| = \frac{K}{m} - r_k^t$。然后，通过从 $[d] \setminus \tilde{\mathcal{T}}_k^t$ 中随机选取 r_k^t 个元素，生成集合 $\mathcal{R}_k^t$。最后，客户端 k 计算出 $\mathcal{I}_k^t = \tilde{\mathcal{T}}_k^t \cup \mathcal{R}_k^t$，并将 $\mathcal{I}_k^t$ 发送给服务器。上述操作的主要目的是对 top-$\frac{K}{m}$ 维度进行混淆，以实现隐私保护。较大的 α 可以更好地保护坐标隐私，但可能会降低模型准确率，正如我们将在第 5 节中通过实证展示的那样。当 $\alpha = 1$ 时，发送的坐标完全随机。

当服务器收到来自所有客户端的 $\{\mathcal{I}_k^t\}_{k=1}^m$ 后，它会确定稀疏化梯度 $\mathcal{I}^t = \cup_{k=1}^m \mathcal{I}_k^t$的坐标集合，并将 $\mathcal{I}^t$ 广播给所有客户端。最后，每个客户端收到 $\mathcal{I}^t$，并据此计算稀疏化后的 $\tilde{}\mathbf{g}_k^t$。[2] 更具体地说，$(\tilde{}\mathbf{g}_k^t)_j = (\mathbf{g}_k^t)_j$

2. 本文中，我们用波浪号表示稀疏向量，以便于区分。

如果 $j \in \mathcal{I}^t$ 且$(\tilde{\mathbf{g}}_k^t)_j = 0$，否则，其中$(\tilde{\mathbf{g}}_k^t)j$ 表示 $\tilde{\mathbf{g}}_k^t$的第 j个坐标的值。由于 $|\mathcal{I}_k^t| = \frac{K}{m}$，我们有 $|\mathcal{I}^t| = |\uplus_{k=1}^m \mathcal{I}_k^t| \leq \sum_{k=1}^m |\mathcal{I}_k^t| = K$。请注意，由于客户端和服务器已就稀疏化梯度的非零坐标 $\mathcal{I}^t$ 达成一致，因此客户端只需向服务器发送$(\tilde{\mathbf{g}}_k^t)_{\mathcal{I}^t}$。更具体地说，$(\tilde{\mathbf{g}}_k^t)_{\mathcal{I}^t} = ((\tilde{\mathbf{g}}_k^t)_{j_1}$, $\ldots, (\tilde{\mathbf{g}}_k^t)^{j_{|\mathcal{I}^t|}})$，其中 $j_s \in \mathcal{I}^t (s = 1, \ldots, |\mathcal{I}^t|)$ 按升序排列。当服务器收到 $\{(\tilde{\mathbf{g}}_k^t)_{\mathcal{I}^t}\}_{k=1}^m$ 后，会使用鲁棒聚合得到$(\tilde{\mathbf{G}}^t)_{\mathcal{I}^t} = \text{Agg}(\{(\tilde{\mathbf{g}}_k^t)_{\mathcal{I}^t}\}_{k=1}^m)$。请注意，$(\tilde{\mathbf{G}}^t)_{\mathcal{I}^t}$ 仍然是一个稀疏化向量。此外，由于 $\mathcal{I}^t$此前已发送给客户端，因此服务器只需广播$(\tilde{\mathbf{G}}^t)_{\mathcal{I}^t}$。因此，ConSpar 自然是一种双向稀疏化方法，无需采用 DoubleSqueeze 技术（Tang 等人，2019年）。接下来，我们分析共识稀疏化后客户端之间的差异。请注意，我们不对拜占庭客户端的行为作任何假设；它们可能发送任意的 $\mathcal{I}_k^t$。

命题 1. 令 $\{\tilde{\mathbf{g}}_k^t\}_{k=1}^m$ 表示向量 $\{\mathbf{g}_k^t\}_{k=1}^m$的共识稀疏化结果，则对于任意固定的 $k, k' \in \mathcal{G}$, 有 $\mathbb{E}\|\tilde{\mathbf{g}}_k^t - \tilde{\mathbf{g}}_{k'}^t\|^2 \leq \mathbb{E}\|\mathbf{g}_k^t - \mathbf{g}_{k'}^t\|^2$。

命题 1 表明，ConSpar 不会扩大客户端之间的差异，这与定理 1 一致。同时，第二轮通信可以使用安全聚合，并且只需在共识非零坐标上添加随机掩码。接下来，我们分析 ConSpar 中用于生成$\mathcal{I}_k^t$ 的机制的隐私保护。首先，我们在定义2中给出 ϵ-差分隐私（ϵ-差分隐私）的定义。

定义2. 设 $\epsilon > 0$ 为一个实数。如果对于任意两个相邻输入数据集 $\mathcal{T}_1$ 和 $\mathcal{T}_2$，以及可能的输出 $\mathcal{S}$的任意子集，随机机制 $\mathcal{M}$ 都能提供ϵ-差分隐私，则称其满足ϵ-差分隐私：

$$\Pr[\mathcal{M}(\mathcal{T}_1) \in \mathcal{S}] \leq \exp(\epsilon) \cdot \Pr[\mathcal{M}(\mathcal{T}_2) \in \mathcal{S}].$$

在我们用于在 ConSpar 中生成 $\mathcal{I}_k^t$ 的机制中， $\mathcal{T}_1$ 和 $\mathcal{T}_2$ 是排名前 $\frac{K}{m}$ 的坐标集合。定义 2 未明确相邻数据集的定义。在本研究中，满足 $\mathcal{T}_1$, $\mathcal{T}_2 \subseteq [d]$ 和 $|\mathcal{T}_1| = |\mathcal{T}_2| = \frac{K}{m}$ 的坐标集合 $\mathcal{T}_1$ 和 $\mathcal{T}_2$ 被定义为相邻，当且仅当 $\mathcal{T}_1$ 和 $\mathcal{T}_2$ 仅在一个元素上不同。Liu 等人（2020年）为仅选择一个坐标的稀疏化方法提供了差分隐私保证。我们的定义更加一般，并涵盖了 $|\mathcal{T}_1| = |\mathcal{T}_2| = 1$这一仅包含一个坐标的特殊情况。现在我们证明，坐标生成机制提供 ϵ-差分隐私。

定理 2. 对于任意 $\alpha \in (0, 1]$, ，共识稀疏化中以排名靠前的坐标集合 $\mathcal{T}_k^t$作为输入并输出 $\mathcal{I}_k^t$的机制提供 $\left(\ln\left(\binom{(1+\alpha)\cdot\frac{K}{m}}{2\alpha}\binom{d-\frac{K}{m}+1}{}\right)\right)$-差分隐私。

最后，我们分析 ConSpar 的通信复杂度。每次迭代中，客户端需要发送候选坐标集 $\mathcal{I}_k^t$, 接收 $\mathcal{I}^t$，以 $(\tilde{\mathbf{g}}_k^t)_{\mathcal{I}^t}$ 的形式发送局部梯度，然后接收 $(\tilde{\mathbf{G}}^t)_{\mathcal{I}^t}$。因此，每个客户端每次迭代最多需要通信 $(\frac{K}{m} + K)$ 个整数和 $2K$ 个浮点数。当每个整数或浮点数以 32 位（4 字节）表示时，每个客户端每次迭代的总通信负载不超过 $(96 + \frac{32}{m})K$ 位。其通信负载不高于普通的 top-K 稀疏化，后者会传输 $4\times 32K = 128K$ 位

每次迭代。与此同时，尽管 ConSpar 需要两个通信轮次，但额外的通信轮次是可以接受的。一方面，两轮之间的计算量很少，不会显著增加聚合过程中客户端断连的风险。另一方面，与安全聚合结合时，额外通信轮次的开销可以忽略不计，因为安全聚合本就需要多个通信轮次，并且能够处理离线客户端。

3.3 FedREP

如前所述，ConSpar 既兼容鲁棒聚合，也兼容安全聚合。然而，鲁棒聚合和安全聚合无法直接结合使用，因为安全聚合最初是为线性聚合（如求和和取平均）设计的，而鲁棒聚合通常是非线性的。这也被称为鲁棒性与隐私之间的张力（Kairouz 等，2021 年）。在拜占庭鲁棒机器学习领域，服务器端的缓冲区已得到广泛研究（Karimireddy 等人，2022；Velicheti 等人，2021；Wang 等，2020；Yang 和 Li，2021），可用于在鲁棒性和隐私之间进行权衡。我们也在 FedREP 中引入了缓冲区。FedREP 的细节见附录 A 中的算法1 和算法 2。

设整数 s 表示缓冲区大小。为简单起见，我们假设客户端数量 m 是缓冲区大小 s 的整数倍，因此服务器上有 $\frac{m}{s}$ 个缓冲区。在第 t次全局迭代开始时，每个客户端 k 使用优化算法$\mathcal{A}$ 和训练样本 $\mathcal{D}_k$，基于$\mathbf{w}^t$ 在本地训练模型，并获得模型参数$\mathbf{w}_k^{t+1} = \mathcal{A}(\mathbf{w}^t; \mathcal{D}_k)$。待发送的更新为$\mathbf{g}_k^t = \mathbf{u}_k^t + (\mathbf{w}^t - \mathbf{w}_k^{t+1})$，其中$\mathbf{u}_k^t$ 是用于误差补偿的本地记忆，且$\mathbf{u}_k^0 = \mathbf{0}$。随后，客户端 k 通过共识稀疏化（参见第 3.2 节）生成坐标集合 $\mathcal{I}_k^t$，并将 $\mathcal{I}_k^t$ 发送给服务器。

服务器收到所有客户端建议的坐标集合后，会向所有客户端广播 $\mathcal{I}^t = \cup_{k=1}^m \mathcal{I}_k^t$，即当前迭代要传输的坐标集合。此外，服务器还会为每个客户端随机分配一个缓冲区。具体来说，服务器从 $[m]$ 中随机选取一个排列 π，并将缓冲区 b_l 分配给客户端 $\{\pi(ls + k)\}_{k=1}^s (l = 0, 1, \ldots, \frac{m}{s} - 1)$。随后，通过安全聚合[3] 得到每个缓冲区对应的 $\mathrm{b}_l = \frac{1}{s}\sum_{k=1}^s (\tilde{\mathrm{g}}_{\pi(ls+k)}^t)_{\mathcal{I}^t}$，再通过对缓冲区进行鲁棒聚合得到全局更新 $(\tilde{\mathrm{G}}^t)_{\mathcal{I}^t} = \mathrm{Agg}(\{\mathrm{b}_l\}_{l=1}^{m/s})$。在此期间，客户端可以通过计算 $\mathrm{u}_k^{t+1} = \mathrm{g}_k^t - \tilde{\mathrm{g}}_k^t$ 来更新本地记忆，以补偿误差。最后，服务器将 $(\tilde{\mathrm{G}}^t)_{\mathcal{I}^t}$ 广播给所有客户端，以便通过 $\mathrm{w}^{t+1} = \mathrm{w}^t - \tilde{\mathrm{G}}^t$ 进行全局更新。

我们注意到，共识稀疏化与循环式本地 top-K 稀疏化（CLT-K）（Chen 等人，2020）相似，后者在每个通信轮次中由一个客户端决定所有客户端的非零坐标。然而，这两种稀疏化方法之间存在显著差异。CLT-K 旨在兼容 all-reduce，而共识稀疏化则旨在兼容联邦学习中的鲁棒聚合和安全聚合。此外，存在拜占庭客户端时，CLT-K 不满足 d'-收缩性质，因为拜占庭客户端可能故意发送错误的非零坐标。同时，一些研究（Karimireddy 等人，2022）表明，在鲁棒聚合之前进行分组平均（FedREP 即采用了这种做法），有助于增强异构数据集上学习方法的鲁棒性

3. 在安全聚合之前，可以直接采用随机量化，将数值转换为有限域上的值，以更好地保护隐私。不过，为简洁起见，此处不作介绍。

。我们将在未来的工作中进一步探讨这一方面，因为这超出了本文的研究范围。

最后，我们想进一步讨论隐私问题。理想情况下，服务器除了聚合结果外，不应了解任何其他信息。然而，要实现理想的隐私保护属性本身就很有挑战性，若还要同时保证拜占庭鲁棒性和通信效率，难度就更大了（Kairouz 等，2021 年）。在 FedREP 中，服务器可以访问部分聚合均值 $\mathbf{b}_l$ 和坐标集合 $\mathcal{I}_k^t$。不过，据我们所知，这两类信息带来的隐私泄露风险有限。服务器不知道每个客户端发送的动量，只能访问坐标集合 $\mathcal{I}_k^t$，而不知道对应的数值，甚至连其符号也不知道。尽管还需要进一步研究如何从这些坐标中获取信息，但据我们所知，几乎没有现有方法能够仅凭坐标还原训练数据。

4. 收敛

本节将从理论上证明 FedREP 的收敛性。证明细节见附录 B。首先，我们给出 d'-压缩算子的定义（Stich 等人，2018年）。

定义 3 (d'-收缩)。 $\mathcal{C}: \mathbb{R}^d \to \mathbb{R}^d$ 称为 d'-收缩算子 $(0 < d' \leq d)$，如果

$$\mathbb{E}\|\mathbf{x} - \mathcal{C}(\mathbf{x})\|^2 \leq \left(1 - d'/d\right)\|\mathbf{x}\|^2, \quad \forall \mathbf{x} \in \mathbb{R}^d. \tag{4}$$

The 共识稀疏化的 d'-收缩性质见命题 2。

命题 2。 如果拜占庭客户端的比例不超过 δ $(0 \leq \delta \leq \frac{1}{2})$，则共识稀疏化是一个 d'_{cons}-收缩算子，其中 $d'_{cons} = d(1 - e^{-\frac{\alpha K[(1-\delta)m-1]}{md}}) + me^{-\frac{\alpha K[(1-2\delta)m-1]}{md}}$。

因此，在不存在拜占庭攻击时，现有关于 d'-收缩算子的收敛结果（Stich 等人，2018年）可直接应用于共识稀疏化。接下来，我们从理论上分析 FedREP 的收敛性。为简化分析，我们将安全聚合与服务器上的鲁棒聚合视为一个单元安全鲁棒聚合器，记为 $\mathbf{SRAgg}(\cdot)$。因此，$\tilde{\mathbf{G}}^t = \mathbf{SRAgg}(\{\tilde{\mathbf{g}}_k^t\}_{k=1}^m)$。假设如下所列。

假设 1 (拜占庭设置)。 拜占庭客户端的比例不超过 $\delta(0 \leq \delta < \frac{1}{2})$，且安全鲁棒聚合器 $\mathbf{SRAgg}(\cdot)$ 是具有常数 $c \geq 0$ 的(δ, c)-鲁棒的。

假设 2 （下界）。 $F(\mathbf{w})$ 有下界： $\exists F^* \in \mathbb{R}$， $F(\mathbf{w}) \geq F^*$， $\forall \mathbf{w} \in \mathbb{R}^d$。

假设3 (L-光滑性)。 全局损失函数 $F(\mathbf{w})$ 可微且为 L-光滑：$||\nabla F(\mathbf{w}) - \nabla F(\mathbf{w}')|| \leq L||\mathbf{w} - \mathbf{w}'||$, $\forall \mathbf{w}, \mathbf{w}' \in \mathbb{R}^d$.

假设 4 (有界偏差)。 $\forall k \in \mathcal{G}$，我们有 $\mathbb{E}[\nabla f_{i_k^t}(\mathbf{w})] = \nabla F_k(\mathbf{w})$，并且存在$B \geq 0$ 使得 $\|\nabla F_k(\mathbf{w}) - \nabla F(\mathbf{w})\| \leq B$, $\forall \mathbf{w} \in \mathbb{R}^d$.

假设5 (有界梯度)。 $\forall k \in \mathcal{G}$，随机梯度 $\nabla f_{i_k^t}(\mathbf{w})$ 的期望有界： $\exists D \in \mathbb{R}_+$，存在 $\|\nabla F_k(\mathbf{w})\| \leq D, \forall \mathbf{w} \in \mathbb{R}^d$。

假设 1 是拜占庭鲁棒分布式机器学习中的常见假设，与先前的研究一致（Karimireddy 等人，2022）。其余假设是分布式随机优化中的常见假设。假设5广泛用于采用误差补偿的梯度压缩方法的分析。我们首先分析 FedREP 的一个特殊情形下的收敛性，其中训练算法 $\mathcal{A}$ 是采用学习率 η 和间隔 I的本地随机梯度下降。具体而言，$\mathbf{w}_k^{t+1}$ 通过以下过程计算：(i) $\mathbf{w}_k^{t+1\,0} = \mathbf{w}^t$; (ii)$\mathbf{w}_k^{t+1\,j+1} = \mathbf{w}_k^{t+1\,j} - \eta \cdot \nabla f_{i_k^{t,j}}(\mathbf{w}_k^{t+1\,j})$ $j = 0, 1, \ldots, I-1$; (iii) $\mathbf{w}_k^{t+1} = \mathbf{w}_k^{t+1\,I}$，其中 $i_k^{t,j}$ 从 $\mathcal{D}_k$中均匀采样。假设 6 是针对这一情形作出的。

假设6（有界方差）。随机梯度 $\nabla f_{i_k^t}(\mathbf{w})$ 是无偏的，且具有有界方差：$\mathbb{E}[\nabla f_{i_k^t}(\mathbf{w})] = F_k(\mathbf{w})$ 且 $\exists \sigma \in \mathbb{R}_+$，使得 $\mathbb{E}\|\nabla f_{i_k^t}(\mathbf{w}) - \nabla F_k(\mathbf{w})\|^2 \le \sigma^2, \forall \mathbf{w} \in \mathbb{R}^d, \forall k \in \mathcal{G}$。

根据假设5和假设6，随机梯度 $\nabla f_{i_k^t}(\mathbf{w})$ 的二阶矩由 $(D^2+\sigma^2)$ 界定。令 $\mathbf{u}^t = \frac{1}{|\mathcal{G}|}\sum_{k\in\mathcal{G}} \mathbf{u}_k^t$，并令 $\mathbf{e}^t = \mathbf{SRAgg}(\{\tilde{\mathbf{g}}_k^t\}_{k=1}^m) - \frac{1}{|\mathcal{G}|}\sum_{k\in\mathcal{G}} \tilde{\mathbf{g}}_k^t$表示聚合误差。我们证明 $\mathbb{E}\|\mathbf{u}_k^t\|^2$ 和 $\mathbb{E}\|\mathbf{e}^t\|^2$ 均有界。

引理1。在假设 *1*、*2*、*3*、*4*、*5* 和 *6* 下，设常数 $H = d/d'_{cons}$，并取学习率 $\eta_t = \eta > 0$，我们有

$$\mathbb{E}\|\mathbf{u}_k^t\|^2 \le 4H^2I^2(D^2+\sigma^2)\cdot\eta^2, \quad \forall k \in \mathcal{G}.$$

引理2。在与引理*1*相同的条件下，我们有

$$\mathbb{E}\|\mathbf{e}^t\|^2 \le 8c\delta I^2(4H^2+1)(D^2+\sigma^2)\cdot\eta^2.$$

根据引理1和引理2，我们得到以下定理。

定理 3。对于 *FedREP*，在与引理*1*和引理*2*相同的条件下，我们有：

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(\mathbf{w}^t)\|^2 \le \frac{2[F(\hat{\mathbf{w}}^0) - F^*]}{\eta IT} + \eta\gamma_1 + \eta^2\gamma_2 + \Delta,$$

其中 $\gamma_1 = 2IL\cdot[2(1-I^{-1})LD + 2HD\sqrt{D^2+\sigma^2} + (D^2+\sigma^2) + 8c\delta(4H^2+1)(D^2+\sigma^2)]$, $\gamma_2 = 8H^2I^2L^2(D^2+\sigma^2)$ 且 $\Delta = 2BD + 4\sqrt{2c\delta(4H^2+1)(D^2+\sigma^2)D}$。

取 $\eta = O(1/\sqrt{T})$ 时，定理 3 保证 FedREP 的收敛速度为 $O(1/\sqrt{T})$，并带有额外误差 Δ，该误差由两项组成。第一项 $2BD$ 源于随机梯度的偏差，反映了客户端之间的异质性程度。在满足 $B=0$ 的独立同分布情形下，该项会消失。第二项 $4\sqrt{2c\delta(4H^2+1)(D^2+\sigma^2)D}$源于聚合误差。当不存在拜占庭客户端（$\delta=0$）时，该项会消失。也就是说，在不存在拜占庭客户端的独立同分布情形下，额外误差 Δ 会消失。接下来，我们分析满足假设 7 的一般本地训练算法下 FedREP 的收敛性，这揭示了训练算法的两个重要性质。

假设 *7*。令 $w' = \mathcal{A}(w;\mathcal{D}_k)$。存在常数 $\eta_{\mathcal{A}} > 0$、$A_1 \ge 0$ 和 $A_2 > 0$，使得本地训练算法 $\mathcal{A}$ 满足 $\|\mathbb{E}[G_{\mathcal{A}}(w;\mathcal{D}_k)] - \nabla F_k(w)\| \le A_1$ 且$\mathbb{E}\|G_{\mathcal{A}}(w;\mathcal{D}_k)\|^2 \le (A_2)^2$，，其中 $G_{\mathcal{A}}(w;\mathcal{D}_k) = (w - w')/\eta_{\mathcal{A}}$，$\forall k \in [m]$。

在假设 7 中，$\mathbf{G}_{\mathcal{A}}(\mathbf{w};\mathcal{D}_k)$可视为算法 $\mathcal{A}$ 对梯度 $\nabla F_k(\mathbf{w})$ 的估计，且具有有界偏差 A_1 和有界二阶矩（A_2）^{2}。期望源于算法 $\mathcal{A}$中的随机性。许多广泛使用的算法都满足假设 7。对于普通随机梯度下降，令 $\eta_{\mathcal{A}}$ 为学习率，则 $\mathbf{G}_{\mathcal{A}}(\mathbf{w};\mathcal{D}_k)$恰好是随机梯度。因此，在假设 5 和 6 下，我们有 $A_1 = 0$ 和（A_2）$^2 = D^2 + \sigma^2$。此外，先前的研究（Allen-Zhu 等，2020；El-Mhamdi 等，2020；Karimireddy 等人，2021 年）表明，在拜占庭鲁棒机器学习中，利用动量等历史信息是必要的。我们在附录 B 的命题 3 中证明，本地动量随机梯度下降也满足假设 7。

定理 4. 设常数为 $H = d/d'_{cons}$。对于 *FedREP*，在假设 *1*、*2*、*3*、*4*、*5* 和 *7* 下，我们有：

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(\mathbf{w}^t)\|^2 \leq \frac{2[F(\hat{\mathbf{w}}^0) - F^*]}{\eta_{\mathcal{A}} T} + \eta_{\mathcal{A}}\gamma_{\mathcal{A},1} + (\eta_{\mathcal{A}})^2\gamma_{\mathcal{A},2} + \Delta_{\mathcal{A}},$$

其中 $\gamma_{\mathcal{A},\ 1} = 2(A_2)^2 L + 4HA_2 DL + 16c\delta(4H^2+1)(A_2)^2 L$，$\gamma_{\mathcal{A},\ 2} = 8H^2(A_2)^2 L^2$ 以及 $\Delta_{\mathcal{A}} = 2A_1 D + 2BD + 4\sqrt{2c\delta(4H^2+1)A_2 D}$。

与定理 3 中的误差 Δ 相比，$\Delta_{\mathcal{A}}$中多出了一项 $2A_1 D$，这是由算法 $\mathcal{A}$中的梯度估计偏差造成的。同时，我们想指出，定理 4 为一般算法提供了收敛保证。对于某些特定算法，采用特定的分析技术可能会得到更紧的上界。由于本文主要关注一般框架，我们将此留待未来工作。

5. 实验

在本节中，我们评估 FedREP 和基线方法在图像分类任务上的性能。每种方法均使用广泛应用的深度学习模型 ResNet-20（He 等，2016）在 CIFAR-10 数据集（Krizhevsky 等，2009）上进行评估。训练样本均匀且等量地分配给各客户端。本文的所有实验均通过 PyTorch 在带有 Docker 容器的分布式平台上进行。具体而言，我们将 32 个 Docker 容器设为客户端，其中 7 个为拜占庭客户端。另设一个 Docker 容器作为服务器。每个 Docker 容器均绑定一块 NVIDIA Tesla K80 GPU。除非另有说明，我们将 FedREP 的本地训练算法 $\mathcal{A}$设为本地动量随机梯度下降，动量超参数为 $\beta = 0.9$（见附录 B 中的公式（180））。每种方法均在相同环境下运行 120 个训练轮次。初始学习率从 $\{0.1, 0.2, 0.5, 1.0, 2.0, 5.0\}$ 中选取。按照 He 等（2016）的建议，在第 80 轮时，学习率将乘以 0。1。我们将各轮次中最佳的 Top-1 准确率作为最终指标。在比特翻转攻击、“少量即足够”（ALIE）攻击（Baruch 等，2019）和“帝国覆灭”（FoE）攻击（Xie 等，2020a）下测试每种方法。遭受比特翻转攻击的拜占庭客户端会发送方向相反的更新。ALIE 和 FoE 是两种全知攻击，假设攻击者知道所有客户端的更新，并利用这些更新实施攻击。我们将 FoE 攻击的攻击幅度超参数设为 0。5。对于 FedREP，我们测试鲁棒聚合器分别采用几何中位数（geoMed）（Chen 等人，2017 年）、逐坐标截尾均值（TMean）（Yin 等人，2018 年）和中心裁剪（CClip）（Karimireddy 等人，

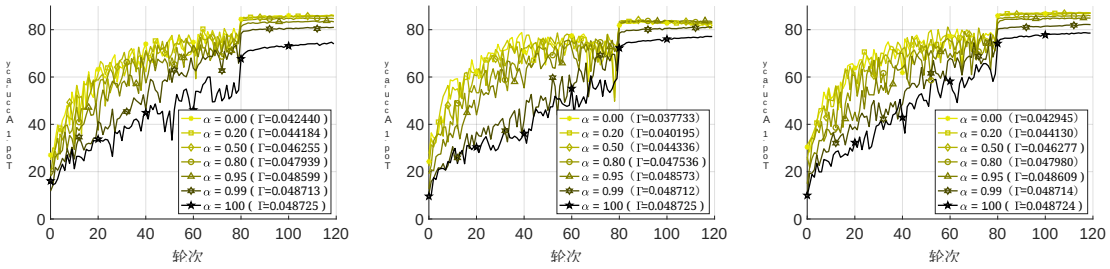


图1：在 7 个拜占庭客户端遭受比特翻转攻击（左）、ALIE 攻击（中）和 FoE 攻击（右）时，FedREP 搭配 CClip 的 Top-1 准确率随训练轮次的变化。

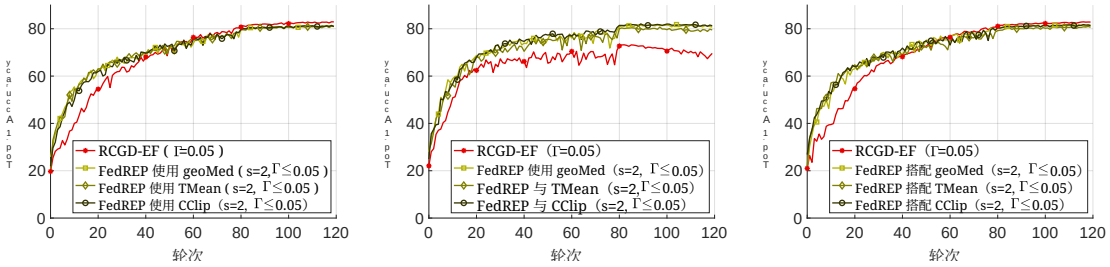


图2：在 7 个拜占庭客户端遭受比特翻转攻击（左）、ALIE 攻击（中）和 FoE 攻击（右）时，FedREP 和 RCGD-EF 的 Top-1 准确率随训练轮次的变化。

2021 年）时的性能。具体而言，我们采用 Weiszfeld 算法（Pillutla 等人，2019 年），将迭代次数设为 5，以计算 geoMed。TMean 的修剪比例设为 $7/16$。对于 CClip，我们将裁剪半径设为 0。5，迭代次数设为 5。批大小设为 25。

我们首先通过实证评估 α 对 FedREP 性能的影响。我们设置 $I=1$、$s=2$ 和 $K=0.05d$。我们比较 FedREP 在 $\alpha=0$、0。2、0。5、0。8、0。95、0。99 和 1 时的性能。如图1所示，当 α 在 0 到 0。95 之间变化时，搭配 CClip 的 FedREP 性能变化很小。当 α 继续增大时，最终准确率会迅速下降。一个可能的原因是，当 α 接近 1 时，本地误差补偿记忆中的某些坐标可能会变得非常大。附录 C.1 展示了 FedREP 搭配 geoMed 和 TMean 的更多结果。由于当 $0 \leq \alpha \leq 0.95$时，α 的影响很小，因此我们在以下实验中设置$\alpha=0$。此外，附录 C.2 中的实证结果表明，针对坐标的拜占庭攻击对 FedREP 的性能影响很小。因此，我们假设以下实验中坐标不受攻击。

随后，我们将 FedREP 与一种名为带误差反馈的鲁棒压缩梯度下降（RCGD-EF）（Ghosh 等，2021）的拜占庭鲁棒、通信高效基线方法进行比较。为确保公平，我们将 RCGD-EF 中的压缩算子 $\mathcal{Q}(\cdot)$设为 top-K

稀疏化，并为 FedREP 和 RCGD-EF 均设置 $\Gamma = 0.05$，即传输维度数与总维度数之比。每种方法的本地更新间隔 I 均设为 5。如图2所示，在比特翻转攻击和 FoE 攻击下，FedREP 的性能与 RCGD-EF 相当；但在 ALIE 攻击下，FedREP 的表现优于 RCGD-EF。同时，FedREP 天然采用双向稀疏化方法，而 RCGD-EF 并非如此。此外，与 RCGD-EF 相比，FedREP 还提供了额外的隐私保护。

更多实证结果见附录，下面简要概述：

- 附录C.3中的实证结果展示了动量超参数 β 对 FedREP 的影响。

- 附录C.4中的实证结果表明，FedREP 的表现显著优于通信高效的隐私保护基线方法 SparseSecAgg（Ergun 等，2021）。

- 附录C.5中的实证结果表明，与拜占庭鲁棒隐私保护基线 SHARE（Velicheti 等人，2021年）相比，FedREP 在收敛速度和准确率方面表现相当，而通信成本则低得多。

6. 结论

本文从理论上分析了联邦学习中拜占庭鲁棒性、通信效率与隐私保护之间的矛盾。基于分析结果，我们提出了一种新颖的拜占庭鲁棒、通信高效且保护隐私的联邦学习框架，称为 $FedREP$。本文给出了 FedREP 的拜占庭鲁棒性和收敛性的理论保证。实证结果表明，FedREP 的性能显著优于通信高效的隐私保护基线方法。此外，与拜占庭鲁棒的通信高效基线方法相比，FedREP 在实现相当准确率的同时，还具备隐私保护这一额外优势。

附录A：FedREP 详情

FedREP 在服务器端和客户端上的详细算法分别见算法1和算法2。

算法1 FedREP（服务器）

输入： 客户端数量 m，迭代次数 T，缓冲区大小 s，鲁棒聚合器 $\mathbf{Agg}(\cdot)$;

对于 $t=0$ **到** $T-1$，**执行**

接收 $\{\mathcal{I}_k^t\}_{k=1}^m$ 从所有客户端接收，并计算 $\mathcal{I}^t=\cup_{k=1}^m\mathcal{I}_k^t$;

向所有客户端广播 $\mathcal{I}^t$；

随机排列 π，取自 $[m]$;

将缓冲区 $\mathbf{b}_l$ 分配给客户端 $\{\pi(ls+k)\}_{k=1}^s$ 对于 $l=0,\ 1,\ \ldots,\frac{m}{s}-1$;

对于 $l=0$，**直到**$\frac{m}{s}-1$ **执行**

获取 $\mathbf{b}_l=\frac{1}{s}\sum_{k=1}^s(\tilde{\mathbf{g}}^t_{\pi(ls+k)})_{\mathcal{I}^t}$ 通过安全聚合协议;

结束循环

计算：$(\tilde{\mathbf{G}}^t)_{\mathcal{I}^t}=\mathbf{Agg}(\{\mathbf{b}_l\}_{l=1}^{m/s})$;

广播 $(\tilde{\mathbf{G}}^t)_{\mathcal{I}^t}$ 发送给所有客户端;

结束循环

算法 2 FedREP（客户端）

输入：客户端数量 m、迭代次数 T、本地优化算法 $\mathcal{A}$ 和稀疏化大小 K；初始化：模型参数 $\mathbf{w}^0$ 和本地记忆 $\mathbf{u}_k^0=\mathbf{0}$；对于 $t=0$ 到 $T-1$，执行 /* 本地训练学习模型并更新误差补偿 */ 本地训练学习模型并获得参数 $\mathbf{w}_k^{t+1}=\mathcal{A}(\mathbf{w}^t;\mathcal{D}_k)$; 计算 $\mathbf{g}_k^t=\mathbf{u}_k^t+(\mathbf{w}^t-\mathbf{w}_k^{t+1})$; /* 通过稀疏化进行两阶段聚合 */ 通过共识稀疏化生成坐标集合 $\mathcal{I}_k^t$（参见正文第 3.2 节）；将 $\mathcal{I}_k^t$ 发送给服务器；从服务器接收坐标集合 $\mathcal{I}^t$ 和分配的缓冲区编号 l；计算 $(\tilde{\mathbf{g}}_k^t)_{\mathcal{I}^t}$ 并通过安全聚合协议将其发送到指定缓冲区 $\mathbf{b}_l$；更新记忆：$\mathbf{u}_k^{t+1}=\mathbf{g}_k^t-\tilde{\mathbf{g}}_k^t$；从服务器接收 $(\tilde{\mathbf{G}}^t)_{\mathcal{I}^t}$，并根据 $\mathcal{I}^t$恢复 $\tilde{\mathbf{G}}^t$；/* 更新模型参数*/ 更新参数：$\mathbf{w}^{t+1}=\mathbf{w}^t-\tilde{\mathbf{G}}^t$；结束循环 输出模型参数 $\mathbf{w}^T$;

附录 B. 证明细节

In 本节将给出论文中理论结果的证明细节 er.

B.1 定理 1 的证明

证明。 对于任意固定的 $k, k' \in [m]$,

$$
\begin{aligned}
&\mathbb{E}\|\mathcal{C}(\mathbf{v}_k) - \mathcal{C}(\mathbf{v}_{k'})\|^2 \\
=&\mathbb{E}\left[\sum_{j\in\mathcal{N}_k\cap\mathcal{N}_{k'}}[(\mathbf{v}_k)_j - (\mathbf{v}_{k'})_j]^2\right] + \mathbb{E}\left[\sum_{j\in\mathcal{N}_k\setminus\mathcal{N}_{k'}}(\mathbf{v}_k)_j^2\right] + \mathbb{E}\left[\sum_{j\in\mathcal{N}_{k'}\setminus\mathcal{N}_k}(\mathbf{v}_{k'})_j^2\right] && (5)\\
=&\mathbb{E}\left[\sum_{j\in\mathcal{N}_k\cap\mathcal{N}_{k'}}\xi_{k,k',j}(\rho_{k,k'})^2\right] + \mathbb{E}\left[\sum_{j\in\mathcal{N}_k\setminus\mathcal{N}_{k'}}\zeta_{k,j}(\mu_k)^2\right] + \mathbb{E}\left[\sum_{j\in\mathcal{N}_{k'}\setminus\mathcal{N}_k}\zeta_{k',j}(\mu_{k'})^2\right] && (6)\\
=&\sum_{j\in[d]}\xi_{k,k',j}(\rho_{k,k'})^2\cdot\Pr[j\in\mathcal{N}_k\cap\mathcal{N}_{k'}] \\
&\quad+\sum_{j\in[d]}\zeta_{k,j}(\mu_k)^2\cdot\Pr[j\in\mathcal{N}_k\setminus\mathcal{N}_{k'}] + \sum_{j\in[d]}\zeta_{k',j}(\mu_{k'})^2\cdot\Pr[j\in\mathcal{N}_{k'}\setminus\mathcal{N}_k] && (7)\\
=&(\rho_{k,k'})^2\cdot\sum_{j\in[d]}\left(\xi_{k,k',j}\Pr[j\in\mathcal{N}_k\cap\mathcal{N}_{k'}]\right) \\
&\quad+(\mu_k)^2\cdot\sum_{j\in[d]}\left(\zeta_{k,j}\Pr[j\in\mathcal{N}_k\setminus\mathcal{N}_{k'}]\right) + (\mu_{k'})^2\cdot\sum_{j\in[d]}\left(\zeta_{k',j}\Pr[j\in\mathcal{N}_{k'}\setminus\mathcal{N}_k]\right). && (8)
\end{aligned}
$$

□

B.2 命题 1 的证明

证明。 令 $\mathcal{I}^t$ 表示共识稀疏化后的非零坐标集合。在一般情况下，对于任意固定的 k, $k' \in [m]$, 我们有:

$$
\begin{aligned}
\mathbb{E}\|\tilde{\mathbf{g}}_k^t - \tilde{\mathbf{g}}_{k'}^t\|^2 &= \sum_{j\in\mathcal{I}^t}\mathbb{E}[(\mathbf{g}_k^t)_j - (\mathbf{g}_{k'}^t)_j]^2 && (9)\\
&\le \sum_{j\in[d]}\mathbb{E}[(\mathbf{g}_k^t)_j - (\mathbf{g}_{k'}^t)_j]^2 && (10)\\
&= \mathbb{E}\|\mathbf{g}_k^t - \mathbf{g}_{k'}^t\|^2. && (11)
\end{aligned}
$$

□

B.3 定理 2 的证明

证明。 令 $\mathcal{M}$ 为共识稀疏化中的机制，它以排名靠前的坐标集合 $\mathcal{T}$ 作为输入，并输出一个随机坐标集合。 $\mathcal{T}_1$, $\mathcal{T}_2 \subseteq [d]$ 是满足 $|\mathcal{T}_1| = |\mathcal{T}_2| = \frac{K}{m}$ 的两个任意相邻输入坐标集合，且 $\mathcal{S}$ 是 $\mathcal{M}$的可能输出的任意子集。当 $\mathcal{S}$ 为空时，$\Pr[\mathcal{M}(\mathcal{T}_1) \in \mathcal{S}] = \Pr[\mathcal{M}(\mathcal{T}_2) \in \mathcal{S}] = 0$。因此，对于

任意 $\epsilon > 0$，我们有：

$$0 = \Pr[\mathcal{M}(\mathcal{T}_1) \in \mathcal{S}] \leq \exp(\epsilon) \cdot \Pr[\mathcal{M}(\mathcal{T}_2) \in \mathcal{S}] = 0. \tag{12}$$

不失一般性，假设 $\mathcal{S}$ 是非空的。对于任意 $\mathcal{I} \in \mathcal{S}$，令$|\mathcal{T}_1 \cap \mathcal{I}| = U_1$ 和 $|\mathcal{T}_2 \cap \mathcal{I}| = U_2$。由于 $\mathcal{T}_1$ 和 $\mathcal{T}_2$ 仅有一个元素不同，因此它们是~相邻的。因此，我们有 $U_2 = U_1 - 1$、U_1 或 $U_1 + 1$。集合 $\mathcal{T}_1$ 由从 $\mathcal{T}_1$中随机选取 $\binom{K}{m} - r_1$ 个元素生成，其中 r_1服从二项分布 $\mathbf{Pr}\left(\frac{K}{m},\ \alpha\right)$。因此，$\forall i = 0,\ 1,\ \ldots,\ K/m$,

$$\Pr[r_1 = i] = \binom{K/m}{i} \alpha^i (1-\alpha)^{K/m-i}. \tag{13}$$

要得到 $\mathcal{I}$作为最终输出，只能选取 $\mathcal{T}_1 \cap \mathcal{I}$中的元素。因此，$r_1$ 应等于或大于 $|\mathcal{T}_1 \setminus \mathcal{I}| = \frac{K}{m} - U_1$。此外，对于 $r_1 \geq \frac{K}{m} - U_1$，所有元素均从 $\mathcal{T}_1 \cap \mathcal{I}$ 中选取的概率为 $\binom{U_1}{K/m-r_1} / \binom{K/m}{K/m-r_1} = \binom{U_1}{r_1 - (K/m-U_1)} / \binom{K/m}{r_1}$。~最后，$\mathcal{I} \setminus \mathcal{T}_1$ 中的 r_1 个元素应从 $[d] \setminus \mathcal{T}_1$中选取，其概率为$1/\binom{d-K/m+r_1}{r_1}$，因为 $|[d] \setminus \tilde{\mathcal{T}_1}| = d - (K/m - r_1) = d - K/m + r_1$。因此，我们有：

$$\begin{aligned}
&\Pr[\mathcal{M}(\mathcal{T}_1) = \mathcal{I}] \\
&= \sum_{i=K/m-U_1}^{K/m} \left\{ \Pr[r_1 = i] \times \frac{\binom{U_1}{i-(K/m-U_1)}}{\binom{K/m}{i}} \times \frac{1}{\binom{d-K/m+i}{i}} \right\} && (14)\\
&= \sum_{i=K/m-U_1}^{K/m} \left\{ \left[\binom{K/m}{i} \alpha^i (1-\alpha)^{K/m-i} \right] \times \frac{\binom{U_1}{i-(K/m-U_1)}}{\binom{K/m}{i}} \times \frac{1}{\binom{d-K/m+i}{i}} \right\} && (15)\\
&= \sum_{i=K/m-U_1}^{K/m} \left\{ \alpha^i (1-\alpha)^{K/m-i} \binom{U_1}{i-(K/m-U_1)} \times \frac{1}{\binom{d-K/m+i}{i}} \right\} && (16)\\
&= \sum_{i=0}^{U_1} \left\{ \alpha^{K/m-i} (1-\alpha)^i \binom{U_1}{U_1-i} \times \frac{1}{\binom{d-i}{K/m-i}} \right\} && (17)\\
&= \alpha^{K/m} \cdot \sum_{i=0}^{U_1} \left\{ \left(\frac{1-\alpha}{\alpha}\right)^i \binom{U_1}{i} \times \frac{1}{\binom{d-i}{d-K/m}} \right\}. && (18)
\end{aligned}$$

因此，$\Pr[\mathcal{M}(\mathcal{T}_1) = \mathcal{I}]$ 相对于 U_1 单调递增。同理，

$$\Pr[\mathcal{M}(\mathcal{T}_2) = \mathcal{I}] = \alpha^{K/m} \cdot \sum_{i=0}^{U_2} \left\{ \left(\frac{1-\alpha}{\alpha}\right)^i \binom{U_2}{i} \times \frac{1}{\binom{d-i}{d-K/m}} \right\}, \tag{19}$$

其相对于 U_2 单调递增。因此，当 $U_1 = U_2 + 1$ 时，$\Pr[\mathcal{M}(\mathcal{T}_1)=\mathcal{I}] \Pr[\mathcal{M}(\mathcal{T}_2)=\mathcal{I}]$ 取最大值。因此，

$$\frac{\Pr[\mathcal{M}(\mathcal{T}_1) = \mathcal{I}]}{\Pr[\mathcal{M}(\mathcal{T}_2) = \mathcal{I}]} \leq \frac{\alpha^{K/m} \cdot \sum_{i=0}^{U_2+1} \left[\left(\frac{1-\alpha}{\alpha}\right)^i \binom{U_2+1}{i} \times \frac{1}{\binom{d-i}{d-K/m}} \right]}{\alpha^{K/m} \cdot \sum_{i=0}^{U_2} \left[\left(\frac{1-\alpha}{\alpha}\right)^i \binom{U_2}{i} \times \frac{1}{\binom{d-i}{d-K/m}} \right]} \tag{20}$$

$$= \frac{\sum_{i=0}^{U_2+1}\left[\left(\frac{1-\alpha}{\alpha}\right)^i \frac{1}{\binom{d-i}{d-K/m}} \times \binom{U_2+1}{i}\right]}{\sum_{i=0}^{U_2}\left[\left(\frac{1-\alpha}{\alpha}\right)^i \frac{1}{\binom{d-i}{d-K/m}} \times \binom{U_2}{i}\right]} \tag{21}$$

$$= \frac{\frac{1}{\binom{d}{d-K/m}} + \sum_{i=1}^{U_2+1}\left[\left(\frac{1-\alpha}{\alpha}\right)^i \frac{1}{\binom{d-i}{d-K/m}} \times \binom{U_2+1}{i}\right]}{\frac{1}{\binom{d}{d-K/m}} + \sum_{i=1}^{U_2}\left[\left(\frac{1-\alpha}{\alpha}\right)^i \frac{1}{\binom{d-i}{d-K/m}} \times \binom{U_2}{i}\right]} \tag{22}$$

$$= \frac{1 + \frac{1-\alpha}{\alpha} \cdot \sum_{i=1}^{U_2+1}\left[\left(\frac{1-\alpha}{\alpha}\right)^{i-1} \frac{\binom{d}{d-K/m}}{\binom{d-i}{d-K/m}} \times \binom{U_2+1}{i}\right]}{1 + \frac{1-\alpha}{\alpha} \cdot \sum_{i=1}^{U_2}\left[\left(\frac{1-\alpha}{\alpha}\right)^{i-1} \frac{\binom{d}{d-K/m}}{\binom{d-i}{d-K/m}} \times \binom{U_2}{i}\right]}. \tag{23}$$

Let

$$S_0(\alpha) = \sum_{i=1}^{U_2}\left[\left(\frac{1-\alpha}{\alpha}\right)^{i-1} \frac{\binom{d}{d-K/m}}{\binom{d-i}{d-K/m}} \times \binom{U_2}{i}\right] \tag{24}$$

and

$$S_1(\alpha) = \sum_{i=1}^{U_2+1}\left[\left(\frac{1-\alpha}{\alpha}\right)^{i-1} \frac{\binom{d}{d-K/m}}{\binom{d-i}{d-K/m}} \times \binom{U_2+1}{i}\right]. \tag{25}$$

我们有

$$\frac{\Pr[\mathcal{M}(\mathcal{T}_1) = \mathcal{I}]}{\Pr[\mathcal{M}(\mathcal{T}_2) = \mathcal{I}]} \leq \frac{1 + \frac{1-\alpha}{\alpha} \cdot S_1(\alpha)}{1 + \frac{1-\alpha}{\alpha} \cdot S_0(\alpha)} \tag{26}$$

$$= 1 + \frac{\frac{1-\alpha}{\alpha} \cdot (S_1(\alpha) - S_0(\alpha))}{1 + \frac{1-\alpha}{\alpha} \cdot S_0(\alpha)} \tag{27}$$

$$\leq 1 + \frac{\frac{1-\alpha}{\alpha} \cdot (S_1(\alpha) - S_0(\alpha))}{\frac{1-\alpha}{\alpha} \cdot S_0(\alpha)} \tag{28}$$

$$= \frac{S_1(\alpha)}{S_0(\alpha)}. \tag{29}$$

由于 $U_1 = U_2 + 1 \leq K/m$, 我们有 $U_2 \leq K/m - 1$。因此,

$$S_1(\alpha) = \sum_{i=1}^{U_2+1} \frac{\left(\frac{1-\alpha}{\alpha}\right)^{i-1} \binom{d}{d-K/m}\binom{U_2+1}{i}}{\binom{d-i}{d-K/m}} \tag{30}$$

$$\leq \sum_{i=1}^{U_2} \frac{\left(\frac{1-\alpha}{\alpha}\right)^{i-1} \binom{d}{d-K/m}\binom{U_2+1}{i}}{\binom{d-i}{d-K/m}} + \sum_{i=2}^{U_2+1} \frac{\left(\frac{1-\alpha}{\alpha}\right)^{i-1} \binom{d}{d-K/m}\binom{U_2+1}{i}}{\binom{d-i}{d-K/m}} \tag{31}$$

$$= \sum_{i=1}^{U_2} \frac{\left(\frac{1-\alpha}{\alpha}\right)^{i-1} \binom{d}{d-K/m}\binom{U_2+1}{i}}{\binom{d-i}{d-K/m}} + \sum_{i=1}^{U_2} \frac{\left(\frac{1-\alpha}{\alpha}\right)^{i} \binom{d}{d-K/m}\binom{U_2+1}{i+1}}{\binom{d-i-1}{d-K/m}} \tag{32}$$

$$= \sum_{i=1}^{U_2} \frac{\left(\frac{1-\alpha}{\alpha}\right)^{i-1} \binom{d}{d-K/m}\binom{U_2}{i}\frac{U_2+1}{U_2+1-i}}{\binom{d-i}{d-K/m}} + \sum_{i=1}^{U_2} \frac{\left(\frac{1-\alpha}{\alpha}\right)\left(\frac{1-\alpha}{\alpha}\right)^{i-1} \binom{d}{d-K/m}\binom{U_2}{i}\frac{U_2+1}{i+1}}{\binom{d-i}{d-K/m}\frac{K/m-i}{d-i}} \tag{33}$$

$$\leq \sum_{i=1}^{U_2} \frac{\left(\frac{1-\alpha}{\alpha}\right)^{i-1}\binom{d}{d-K/m}\binom{U_2}{i}\frac{U_2+1}{U_2+1-U_2}}{\binom{d-i}{d-K/m}} + \sum_{i=1}^{U_2} \frac{\left(\frac{1-\alpha}{\alpha}\right)\left(\frac{1-\alpha}{\alpha}\right)^{i-1}\binom{d}{d-K/m}\binom{U_2}{i}\frac{U_2+1}{1+1}}{\binom{d-i}{d-K/m}\frac{K/m-U_2}{d-U_2}} \tag{34}$$

$$= \left[(U_2+1) + \frac{\frac{1-\alpha}{\alpha}\cdot\frac{U_2+1}{2}}{\frac{K/m-U_2}{d-U_2}}\right] \cdot \sum_{i=1}^{U_2}\left[\left(\frac{1-\alpha}{\alpha}\right)^{i-1}\frac{\binom{d}{d-K/m}}{\binom{d-i}{d-K/m}} \times \binom{U_2}{i}\right] \tag{35}$$

$$\leq \left[(K/m-1+1) + \frac{\frac{1-\alpha}{\alpha}\cdot\frac{K/m-1+1}{2}}{\frac{K/m-K/m+1}{d-K/m+1}}\right] \cdot S_0(\alpha) \tag{36}$$

$$= \left(\frac{K}{m} + \frac{\frac{K}{m}(1-\alpha)(d-\frac{K}{m}+1)}{2\alpha}\right) \cdot S_0(\alpha) \tag{37}$$

$$\leq \left(\frac{K}{m}(d-\frac{K}{m}+1) + \frac{\frac{K}{m}(1-\alpha)(d-\frac{K}{m}+1)}{2\alpha}\right) \cdot S_0(\alpha) \tag{38}$$

$$\leq \frac{(1+\alpha)\cdot\frac{K}{m}(d-\frac{K}{m}+1)}{2\alpha} \cdot S_0(\alpha)。 \tag{39}$$

因此，

$$\frac{\Pr[\mathcal{M}(\mathcal{T}_1)=\mathcal{I}]}{\Pr[\mathcal{M}(\mathcal{T}_2)=\mathcal{I}]} \leq \frac{S_1(\alpha)}{S_0(\alpha)} \leq \frac{(1+\alpha)\cdot\frac{K}{m}(d-\frac{K}{m}+1)}{2\alpha}. \tag{40}$$

因此，

$$\Pr[\mathcal{M}(\mathcal{T}_1)=\mathcal{I}] \leq \exp\left(\ln\left(\frac{(1+\alpha)\cdot\frac{K}{m}(d-\frac{K}{m}+1)}{2\alpha}\right)\right)\cdot\Pr[\mathcal{M}(\mathcal{T}_2)=\mathcal{I}], \tag{41}$$

这表明 $\mathcal{M}$提供了 $\ln\left(\frac{(1+\alpha)\cdot\frac{K}{m}(d-\frac{K}{m}+1)}{2\alpha}\right)$-差分隐私。 □

接下来，我们针对 α 接近 1 的情况给出一个更强的结果。具体而言，当$\frac{1}{2} \leq \alpha \leq 1$时，

$$\frac{\Pr[\mathcal{M}(\mathcal{T}_1)=\mathcal{I}]}{\Pr[\mathcal{M}(\mathcal{T}_2)=\mathcal{I}]} \leq 1 + \frac{\frac{1-\alpha}{\alpha}\cdot(S_1(\alpha)-S_0(\alpha))}{1+\frac{1-\alpha}{\alpha}\cdot S_0(\alpha)} \leq 1 + \frac{1-\alpha}{\alpha}\cdot(S_1(\alpha)-S_0(\alpha)). \tag{42}$$

由于 $U_1 = U_2+1 \leq K/m$，我们有 $U_2 \leq K/m-1$。因此，

$$S_1(\alpha)-S_0(\alpha) = \sum_{i=1}^{U_2+1}\frac{\left(\frac{1-\alpha}{\alpha}\right)^{i-1}\binom{d}{d-K/m}\binom{U_2+1}{i}}{\binom{d-i}{d-K/m}} - \sum_{i=1}^{U_2}\frac{\left(\frac{1-\alpha}{\alpha}\right)^{i-1}\binom{d}{d-K/m}\binom{U_2}{i}}{\binom{d-i}{d-K/m}} \tag{43}$$

$$= \sum_{i=1}^{U_2}\frac{\left(\frac{1-\alpha}{\alpha}\right)^{i-1}\binom{d}{d-K/m}\left[\binom{U_2+1}{i}-\binom{U_2}{i}\right]}{\binom{d-i}{d-K/m}} + \frac{\left(\frac{1-\alpha}{\alpha}\right)^{U_2}\binom{d}{d-K/m}}{\binom{d-U_2-1}{d-K/m}} \tag{44}$$

$$= \sum_{i=1}^{U_2}\frac{\left(\frac{1-\alpha}{\alpha}\right)^{i-1}\binom{d}{d-K/m}\binom{U_2}{i-1}}{\binom{d-i}{d-K/m}} + \frac{\left(\frac{1-\alpha}{\alpha}\right)^{U_2}\binom{d}{d-K/m}}{\binom{d-U_2-1}{d-K/m}} \tag{45}$$

$$\leq \frac{\binom{d}{d-K/m}}{\binom{d-U_2}{d-K/m}} \sum_{i=1}^{U_2} \left(\frac{1-\alpha}{\alpha}\right)^{i-1} \binom{U_2}{i-1} + \frac{\left(\frac{1-\alpha}{\alpha}\right)^{U_2} \binom{d}{d-K/m}}{\binom{d-U_2-1}{d-K/m}} \tag{46}$$

$$\leq \frac{\binom{d}{d-K/m}}{\binom{d-U_2}{d-K/m}} \sum_{i=0}^{U_2} \left(\frac{1-\alpha}{\alpha}\right)^{i} \binom{U_2}{i} + \frac{\left(\frac{1-\alpha}{\alpha}\right)^{U_2} \binom{d}{d-K/m}}{\binom{d-U_2-1}{d-K/m}} \tag{47}$$

$$= \frac{\binom{d}{d-K/m}}{\binom{d-U_2}{d-K/m}} \left(1 + \frac{1-\alpha}{\alpha}\right)^{U_2} + \frac{\left(\frac{1-\alpha}{\alpha}\right)^{U_2} \binom{d}{d-K/m}}{\binom{d-U_2-1}{d-K/m}} \tag{48}$$

$$\leq \frac{\binom{d}{d-K/m} \left(\frac{1}{\alpha}\right)^{U_2}}{\binom{d-U_2-1}{d-K/m}} + \frac{\left(\frac{1-\alpha}{\alpha}\right)^{U_2} \binom{d}{d-K/m}}{\binom{d-U_2-1}{d-K/m}} \tag{49}$$

$$\leq \frac{\binom{d}{d-K/m} 2^{U_2}}{\binom{d-U_2-1}{d-K/m}} + \frac{\left(\frac{1-\frac{1}{2}}{\frac{1}{2}}\right)^{U_2} \binom{d}{d-K/m}}{\binom{d-U_2-1}{d-K/m}} \tag{50}$$

$$= (2^{U_2}+1) \cdot \frac{\binom{d}{d-K/m}}{\binom{d-U_2-1}{d-K/m}} \tag{51}$$

$$\leq (2^{K/m}+1) \cdot \binom{d}{K/m} \tag{52}$$

$$\leq (2^{K/m}+1) \cdot d^{K/m}. \tag{53}$$

因此，

$$\Pr[\mathcal{M}(\mathcal{T}_1) = \mathcal{I}] \leq \exp\left(\ln\left(1 + \frac{1-\alpha}{\alpha} \cdot (2^{K/m}+1) \cdot d^{K/m}\right)\right) \cdot \Pr[\mathcal{M}(\mathcal{T}_2) = \mathcal{I}]. \tag{54}$$

这表明，当 $\frac{1}{2} \leq \alpha \leq 1$, $\mathcal{M}$ 提供 $\epsilon_{\mathcal{M}}$-差分隐私时，其中

$$\epsilon_{\mathcal{M}} = \ln\left(1 + \frac{1-\alpha}{\alpha} \cdot (2^{K/m}+1) \cdot d^{K/m}\right).$$

特别地，当 $\alpha = 1$, $\epsilon_{\mathcal{M}} = \ln(1+0) = 0$。这与当 $\alpha = 1$ 时坐标集合完全随机这一结论一致。

B.4 命题 2 的证明

证明。 $\forall k \in \mathcal{G}, \forall 0 \leq t < T$，我们有：

$$\mathcal{I}^t = \bigcup_{k' \in [m]} \mathcal{I}_{k'}^t \supseteq \bigcup_{k' \in \mathcal{G}} \mathcal{I}_{k'}^t = \left[\left(\bigcup_{k' \in \mathcal{G} \setminus \{k\}} \mathcal{I}_{k'}^t\right) \cup \mathcal{I}_k^t\right]. \tag{55}$$

因此，

$$\mathbb{E}[\|\tilde{\mathbf{g}}_k^t\|^2 | \mathcal{I}_k^t] = \mathbb{E}\left[\sum_{j \in \mathcal{I}^t} (\mathbf{g}_k^t)_j^2 \middle| \mathcal{I}_k^t\right] \tag{56}$$

$$= \mathbb{E}\left[\sum_{j\in\mathcal{I}_k^t}(\mathbf{g}_k^t)_j^2\middle|\mathcal{I}_k^t\right] + \mathbb{E}\left[\sum_{j\in(\mathcal{I}^t\setminus\mathcal{I}_k^t)}(\mathbf{g}_k^t)_j^2\middle|\mathcal{I}_k^t\right] \tag{57}$$

$$= \sum_{j\in\mathcal{I}_k^t}(\mathbf{g}_k^t)_j^2 + \mathbb{E}\left[\sum_{j\in(\mathcal{I}^t\setminus\mathcal{I}_k^t)}(\mathbf{g}_k^t)_j^2\middle|\mathcal{I}_k^t\right] \tag{58}$$

$$= \sum_{j\in\mathcal{I}_k^t}(\mathbf{g}_k^t)_j^2 + \sum_{j\notin\mathcal{I}_k^t}(\mathbf{g}_k^t)_j^2\cdot\Pr\left[j\in\mathcal{I}^t|\mathcal{I}_k^t\right]. \tag{59}$$

对于任意 $j\notin\mathcal{I}_k^t$,

$$\begin{aligned}
&\Pr\left[j\in\mathcal{I}^t|\mathcal{I}_k^t\right]\\
&\geq \Pr\left[j\in\left(\bigcup_{k'\in\mathcal{G}\setminus\{k\}}\mathcal{I}_{k'}^t\right)\middle|\mathcal{I}_k^t\right] &(60)\\
&= \Pr\left[j\in\left(\bigcup_{k'\in\mathcal{G}\setminus\{k\}}\left(\tilde{\mathcal{T}}_{k'}^t\cup\mathcal{R}_{k'}^t\right)\right)\middle|\mathcal{I}_k^t\right] &(61)\\
&= \Pr\left[j\in\left(\bigcup_{k'\in\mathcal{G}\setminus\{k\}}\tilde{\mathcal{T}}_{k'}^t\right)\cup\left(\bigcup_{k'\in\mathcal{G}\setminus\{k\}}\mathcal{R}_{k'}^t\right)\middle|\mathcal{I}_k^t\right] &(62)\\
&= \Pr\left[j\in\left(\bigcup_{k'\in\mathcal{G}\setminus\{k\}}\tilde{\mathcal{T}}_{k'}^t\right)\middle|\mathcal{I}_k^t\right] + \Pr\left[j\in\left(\bigcup_{k'\in\mathcal{G}\setminus\{k\}}\mathcal{R}_{k'}^t\right)\setminus\left(\bigcup_{k'\in\mathcal{G}\setminus\{k\}}\tilde{\mathcal{T}}_{k'}^t\right)\middle|\mathcal{I}_k^t\right]. &(63)
\end{aligned}$$

为简洁起见，令

$$\nu = \Pr\left[j\in\left(\bigcup_{k'\in\mathcal{G}\setminus\{k\}}\tilde{\mathcal{T}}_{k'}^t\right)\middle|\mathcal{I}_k^t\right]\in[0,1], \tag{64}$$

并且我们有:

$$\begin{aligned}
&\Pr\left[j\in\mathcal{I}^t|\mathcal{I}_k^t\right]\\
&= \nu + (1-\nu)\cdot\Pr\left[j\in\left(\bigcup_{k'\in\mathcal{G}\setminus\{k\}}\mathcal{R}_{k'}^t\right)\middle|\mathcal{I}_k^t, j\notin\left(\bigcup_{k'\in\mathcal{G}\setminus\{k\}}\tilde{\mathcal{T}}_{k'}^t\right)\right] &(65)\\
&= \nu + (1-\nu)\cdot\left\{1-\Pr\left[j\notin\left(\bigcup_{k'\in\mathcal{G}\setminus\{k\}}\mathcal{R}_{k'}^t\right)\middle|\mathcal{I}_k^t, j\notin\left(\bigcup_{k'\in\mathcal{G}\setminus\{k\}}\tilde{\mathcal{T}}_{k'}^t\right)\right]\right\} &(66)\\
&= \nu + (1-\nu)\cdot\left\{1-\prod_{k'\in\mathcal{G}\setminus\{k\}}\Pr\left[j\notin\mathcal{R}_{k'}^t\middle|\mathcal{I}_k^t, j\notin\left(\bigcup_{k'\in\mathcal{G}\setminus\{k\}}\tilde{\mathcal{T}}_{k'}^t\right)\right]\right\} &(67)\\
&\overset{\text{(i)}}{=} \nu + (1-\nu)\cdot\left\{1-\prod_{k'\in\mathcal{G}\setminus\{k\}}\left[\sum_{i=0}^{K/m}\Pr[r_{k'}^t=i]\cdot\left(1-\frac{i}{d-K/m+i}\right)\right]\right\} &(68)
\end{aligned}$$

$$\overset{\text{(ii)}}{\geq} \nu + (1-\nu)\cdot\left\{1 - \prod_{k'\in\mathcal{G}\setminus\{k\}}\left[\sum_{i=0}^{K/m}\Pr[r_{k'}^t = i]\cdot\left(1-\frac{i}{d}\right)\right]\right\} \tag{69}$$

$$= \nu + (1-\nu)\cdot\left\{1 - \prod_{k'\in\mathcal{G}\setminus\{k\}}\left[\sum_{i=0}^{K/m}\Pr[r_{k'}^t = i] - \sum_{i=0}^{K/m}\frac{i\cdot\Pr[r_{k'}^t = i]}{d}\right]\right\} \tag{70}$$

$$= \nu + (1-\nu)\cdot\left\{1 - \prod_{k'\in\mathcal{G}\setminus\{k\}}\left(1-\frac{\mathbb{E}[r_{k'}^t]}{d}\right)\right\} \tag{71}$$

$$\overset{\text{(iii)}}{=} \nu + (1-\nu)\cdot\left\{1 - \prod_{k'\in\mathcal{G}\setminus\{k\}}\left(1-\frac{\alpha K}{md}\right)\right\} \tag{72}$$

$$= \nu + (1-\nu)\cdot\left[1 - \left(1-\frac{\alpha K}{md}\right)^{|\mathcal{G}|-1}\right] \tag{73}$$

$$\geq \nu + (1-\nu)\cdot\left[1 - \left(1-\frac{\alpha K}{md}\right)^{(1-\delta)m-1}\right] \tag{74}$$

$$\overset{\text{(iv)}}{\geq} \nu\cdot\left[1 - \left(1-\frac{\alpha K}{md}\right)^{(1-\delta)m-1}\right] + (1-\nu)\cdot\left[1 - \left(1-\frac{\alpha K}{md}\right)^{(1-\delta)m-1}\right] \tag{75}$$

$$= 1 - \left(1-\frac{\alpha K}{md}\right)^{(1-\delta)m-1}, \tag{76}$$

其中，(i) 成立是因为当 $r_k^t = i$ 时，元素 j 是从 $[d]\setminus\tilde{\mathcal{T}}_k^t$ 中随机选取的 i 个元素之一的概率为 $\frac{i}{d-K/m+i}$，因为 $|[d]\setminus\tilde{\mathcal{T}}_k^t| = \cancel{d-K/m} + i$。不等式 (ii) 成立是因为 $i \leq K/m$。等式 (iii) 成立是因为 $r_{k'}^t$ 服从二项分布 $\mathrm{B}(\frac{K}{m},\alpha)$。不等式 (iv) 成立是因为 $1-\left(1-\frac{\mathbb{E}[r_k^t]}{d}\right)^{(1-\delta)m-1} \leq 1$。

由于 $0\leq\alpha\leq 1$ 和 $0<\frac{K}{m}<d$，我们有 $0\leq\frac{\alpha K}{md}<1$。因此，

$$\left(1-\frac{\alpha K}{md}\right)^{(1-\delta)m-1} = \left[\left(1-\frac{\alpha K}{md}\right)^{-\frac{md}{\alpha K}}\right]^{-\frac{\alpha K[(1-\delta)m-1]}{md}} \leq e^{-\frac{\alpha K[(1-\delta)m-1]}{md}}. \tag{77}$$

因此，

$$\Pr\left[j\in\mathcal{I}^t|\mathcal{I}_k^t\right] \geq 1 - e^{-\frac{\alpha K[(1-\delta)m-1]}{md}}. \tag{78}$$

将其代入式 (59)，可得

$$\mathbb{E}[\|\tilde{\mathbf{g}}_k^t\|^2|\mathcal{I}_k^t] \geq \sum_{j\in\mathcal{I}_k^t}(\mathbf{g}_k^t)_j^2 + \left(1-e^{-\frac{\alpha K[(1-\delta)m-1]}{md}}\right)\cdot\sum_{j\notin\mathcal{I}_k^t}(\mathbf{g}_k^t)_j^2 \tag{79}$$

$$= \sum_{j\in\mathcal{I}_k^t}(\mathbf{g}_k^t)_j^2 + \left(1-e^{-\frac{\alpha K[(1-\delta)m-1]}{md}}\right)\cdot\left(\|\mathbf{g}_k^t\|^2 - \sum_{j\in\mathcal{I}_k^t}(\mathbf{g}_k^t)_j^2\right) \tag{80}$$

$$= \left(1-e^{-\frac{\alpha K[(1-\delta)m-1]}{md}}\right)\cdot\|\mathbf{g}_k^t\|^2 + e^{-\frac{\alpha K[(1-\delta)m-1]}{md}}\sum_{j\in\mathcal{I}_k^t}(\mathbf{g}_k^t)_j^2. \tag{81}$$

取全期望可得:

$$\mathbb{E}\|\tilde{\mathbf{g}}_k^t\|^2 = \mathbb{E}[\mathbb{E}[\|\tilde{\mathbf{g}}_k^t\|^2|\mathcal{I}_k^t]] = \underline{\left(1 - e^{-\frac{\alpha K[(1-\delta)m-1]}{md}}\right)} \cdot \|\mathbf{g}_k^t\|^2 + \underline{e^{-\frac{\alpha K[(1-\delta)m-1]}{md}}} \cdot \mathbb{E}\left[\sum_{j\in\mathcal{I}_k^t}(\mathbf{g}_k^t)_j^2\right]. \tag{82}$$

此外,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{j\in\mathcal{I}_k^t}(\mathbf{g}_k^t)_j^2\Big|r_k^t\right] \\
=&\mathbb{E}\left[\sum_{j\in\tilde{\mathcal{T}}_k^t}(\mathbf{g}_k^t)_j^2\Big|r_k^t\right] + \mathbb{E}\left[\sum_{j\in\mathcal{R}_k^t}(\mathbf{g}_k^t)_j^2\Big|r_k^t\right] && (83)\\
=&\mathbb{E}\left[\sum_{j\in\tilde{\mathcal{T}}_k^t}(\mathbf{g}_k^t)_j^2\Big|r_k^t\right] + \frac{r_k^t}{d-K/m+r_k^t}\cdot\mathbb{E}\left[\sum_{j\notin\tilde{\mathcal{T}}_k^t}(\mathbf{g}_k^t)_j^2\Big|r_k^t\right] && (84)\\
=&\mathbb{E}\left[\sum_{j\in\tilde{\mathcal{T}}_k^t}(\mathbf{g}_k^t)_j^2\Big|r_k^t\right] + \frac{r_k^t}{d-K/m+r_k^t}\cdot\left(\|\mathbf{g}_k^t\|^2 - \mathbb{E}\left[\sum_{j\in\tilde{\mathcal{T}}_k^t}(\mathbf{g}_k^t)_j^2\Big|r_k^t\right]\right) && (85)\\
=&\frac{r_k^t}{d-K/m+r_k^t}\cdot\|\mathbf{g}_k^t\|^2 + \frac{d-K/m}{d-K/m+r_k^t}\cdot\mathbb{E}\left[\sum_{j\in\tilde{\mathcal{T}}_k^t}(\mathbf{g}_k^t)_j^2\Big|r_k^t\right] && (86)\\
=&\frac{r_k^t}{d-K/m+r_k^t}\cdot\|\mathbf{g}_k^t\|^2 + \frac{d-K/m}{d-K/m+r_k^t}\cdot\frac{K/m-r_k^t}{K/m}\cdot\mathbb{E}\left[\sum_{j\in\mathcal{T}_k^t}(\mathbf{g}_k^t)_j^2\Big|r_k^t\right] && (87)\\
\geq&\frac{r_k^t}{d-K/m+r_k^t}\cdot\|\mathbf{g}_k^t\|^2 + \frac{d-K/m}{d-K/m+r_k^t}\cdot\frac{K/m-r_k^t}{K/m}\cdot\frac{K/m}{d}\cdot\|\mathbf{g}_k^t\|^2 && (88)\\
=&\frac{r_k^t}{d-K/m+r_k^t}\cdot\|\mathbf{g}_k^t\|^2 + \frac{d-K/m}{d-K/m+r_k^t}\cdot\frac{K/m-r_k^t}{d}\cdot\|\mathbf{g}_k^t\|^2 && (89)\\
=&\frac{dr_k^t+(d-K/m)(K/m-r_k^t)}{d(d-K/m+r_k^t)}\cdot\|\mathbf{g}_k^t\|^2 && (90)\\
=&\frac{(K/m)\cdot(d-K/m+r_k^t)}{d(d-K/m+r_k^t)}\cdot\|\mathbf{g}_k^t\|^2 && (91)\\
=&\frac{K}{md}\|\mathbf{g}_k^t\|^2. && (92)
\end{aligned}$$

因此,

$$\mathbb{E}\left[\sum_{j\in\mathcal{I}_k^t}(\mathbf{g}_k^t)_j^2\right] = \mathbb{E}\left[\mathbb{E}\left[\sum_{j\in\mathcal{I}_k^t}(\mathbf{g}_k^t)_j^2\Big|r_k^t\right]\right] \geq \mathbb{E}\left[\frac{K}{md}\|\mathbf{g}_k^t\|^2\right] = \frac{K}{md}\|\mathbf{g}_k^t\|^2. \tag{93}$$

将（93）代入（82），可得：

$$\mathbb{E}\|\tilde{\mathbf{g}}_k^t\|^2 \geq \left(1 - e^{-\frac{\alpha K[(1-\delta)m-1]}{md}} + \frac{K}{md}e^{-\frac{\alpha K[(1-\delta)m-1]}{md}}\right) \cdot \|\mathbf{g}_k^t\|^2 \tag{94}$$

$$= \left(1 - \frac{(d - \frac{K}{m})e^{-\frac{\alpha K[(1-\delta)m-1]}{md}}}{d}\right) \cdot \|\mathbf{g}_k^t\|^2. \tag{95}$$

由于 $\tilde{\mathbf{g}}_k^t$ 是$\mathbf{g}_k^t$的共识稀疏化结果，因此有：

$$\mathbb{E}\|\tilde{\mathbf{g}}_k^t - \mathbf{g}_k^t\|^2 = \mathbb{E}\left[\sum_{j\in\mathcal{G}\setminus\mathcal{I}^t}(\mathbf{g}_k^t)_j^2\right] \tag{96}$$

$$= \mathbb{E}\left[\sum_{j\in\mathcal{G}}(\mathbf{g}_k^t)_j^2 - \sum_{j\in\mathcal{I}^t}(\mathbf{g}_k^t)_j^2\right] \tag{97}$$

$$= \|\mathbf{g}_k^t\|^2 - \mathbb{E}\|\tilde{\mathbf{g}}_k^t\|^2 \tag{98}$$

$$\leq \frac{(d - \frac{K}{m})e^{-\frac{\alpha K[(1-\delta)m-1]}{md}}}{d} \cdot \|\mathbf{g}_k^t\|^2 \tag{99}$$

$$= \left(1 - \frac{d(1 - e^{-\frac{\alpha K[(1-\delta)m-1]}{md}}) + \frac{K}{m}e^{-\frac{\alpha K[(1-\delta)m-1]}{md}}}{d}\right) \cdot \|\mathbf{g}_k^t\|^2. \tag{100}$$

根据定义，共识稀疏化是一个 d'_{cons}-压缩算子，其中

$$d'_{cons} = d\left(1 - e^{-\frac{\alpha K[(1-\delta)m-1]}{md}}\right) + \frac{K}{m}e^{-\frac{\alpha K[(1-\delta)m-1]}{md}}.$$

□

B.5 引理1的证明

证明。 当训练算法 $\mathcal{A}$ 是采用学习率 η进行 I 次迭代的本地随机梯度下降时，有$\mathbf{w}_k^{t+1\,0} = \mathbf{w}^t$, $\mathbf{w}_k^{t+1\,j+1} = \mathbf{w}_k^{t+1\,j} - \eta_t \cdot \nabla f_{i^{t,\,k}}(\mathbf{w}_k^{t+1\,j})$ $(j = 0,\ 1,\ \ldots\ldots,\ I-1)$ and $\mathbf{w}_k^{t+1} = \mathbf{w}_k^{t+1\,I}$，其中 $i^{t,\,k}$ 是从 $\mathcal{D}_k$中均匀采样得到的。因此，对所有$k \in \mathcal{G}$，都有以下不等式：

$$\mathbb{E}\|\mathbf{u}_k^{t+1}\|^2 = \mathbb{E}\|\mathbf{g}_k^t - \tilde{\mathbf{g}}_k^t\|^2 \tag{101}$$

$$\overset{\text{(i)}}{\leq} \left(1 - \frac{d'_{cons}}{d}\right)\mathbb{E}\|\mathbf{g}_k^t\|^2 \tag{102}$$

$$= \left(1 - \frac{d'_{cons}}{d}\right)\mathbb{E}\|\mathbf{u}_k^t + (\mathbf{w}^t - \mathbf{w}_k^{t+1})\|^2 \tag{103}$$

$$\overset{\text{(ii)}}{\leq} \left(1 - \frac{d'_{cons}}{d}\right)\left[(1 + \frac{d'_{cons}}{2d})\mathbb{E}\|\mathbf{u}_k^t\|^2 + (1 + \frac{2d}{d'_{cons}})\mathbb{E}\|\mathbf{w}_k^{t+1,0} - \mathbf{w}_k^{t+1,I}\|^2\right] \tag{104}$$

$$\overset{\text{(iii)}}{\leq} \left(1 - \frac{d'_{cons}}{2d}\right)\mathbb{E}\|\mathbf{u}_k^t\|^2 + \frac{2d}{d'_{cons}}\mathbb{E}\|\mathbf{w}_k^{t+1,0} - \mathbf{w}_k^{t+1,I}\|^2 \tag{105}$$

$$\leq \left(1-\frac{d'_{cons}}{2d}\right)\mathbb{E}\|\mathbf{u}_k^t\|^2+\frac{2Id}{d'_{cons}}\sum_{j=0}^{I-1}\mathbb{E}\|\mathbf{w}_k^{t+1,i}-\mathbf{w}_k^{t+1,i+1}\|^2 \tag{106}$$

$$=\left(1-\frac{d'_{cons}}{2d}\right)\mathbb{E}\|\mathbf{u}_k^t\|^2+\frac{2Id}{d'_{cons}}\sum_{j=0}^{I-1}\mathbb{E}\|\eta_t\cdot\nabla f_{i_k^{t,j}}(\mathbf{w}_k^{t+1,j})\|^2 \tag{107}$$

$$\overset{\text{(iv)}}{\leq}\left(1-\frac{d'_{cons}}{2d}\right)\mathbb{E}\|\mathbf{u}_k^t\|^2+\frac{2I^2d}{d'_{cons}}(\eta_t)^2(D^2+\sigma^2), \tag{108}$$

其中，(i) 根据命题 2 推导得出。(ii) 根据以下事实推导得出：$\|\mathbf{x}+\mathbf{y}\|^2\leq(1+\theta)\|\mathbf{x}\|^2+(1+\theta^{-1})\|\mathbf{y}\|^2$ 对于任意常数 $\theta>0$。(iii) 根据以下事实推导得出：$(1-\frac{d'_{cons}}{d})(1+\frac{d'_{cons}}{2d})<1-\frac{d'_{cons}}{2d}$以及 $(1-\frac{d'_{cons}}{d})(1+\frac{2d}{d'_{cons}})<\frac{2d}{d'_{cons}}$。(iv) 根据假设5和假设6推导得出。

Wh其中常数 $b>0$ 和 $\lambda=\frac{4d}{d'_{cons}}$，右侧第二项 ——

$$\frac{2I^2d}{d'_{cons}}(\eta_t)^2(D^2+\sigma^2)=\frac{2I^2d}{d'_{cons}}(D^2+\sigma^2)\cdot\frac{b^2}{t+\lambda} \tag{109}$$

$$=\left(\frac{8I^2d^2b^2}{(d'_{cons})^2}(D^2+\sigma^2)\right)\cdot\frac{1}{t+\lambda}\cdot\frac{d'_{cons}}{4d} \tag{110}$$

$$=\left(\frac{8I^2d^2b^2}{(d'_{cons})^2}(D^2+\sigma^2)\right)\cdot\frac{1}{t+\lambda}\cdot(\frac{d'_{cons}}{2d}-\frac{d'_{cons}}{4d}) \tag{111}$$

$$=\left(\frac{8I^2d^2b^2}{(d'_{cons})^2}(D^2+\sigma^2)\right)\cdot\frac{1}{t+\lambda}\cdot(\frac{d'_{cons}}{2d}-\frac{1}{\lambda}) \tag{112}$$

$$\leq\left(\frac{8I^2d^2b^2}{(d'_{cons})^2}(D^2+\sigma^2)\right)\cdot\frac{1}{t+\lambda}\cdot(\frac{d'_{cons}}{2d}-\frac{1}{t+\lambda+1}) \tag{113}$$

$$=\left(\frac{8I^2d^2b^2}{(d'_{cons})^2}(D^2+\sigma^2)\right)\cdot\left(\frac{\frac{d'_{cons}}{2d}(t+\lambda+1)-1}{(t+\lambda)(t+\lambda+1)}\right) \tag{114}$$

$$=\left(\frac{8I^2d^2b^2}{(d'_{cons})^2}(D^2+\sigma^2)\right)\cdot\left(\frac{1}{t+\lambda+1}-\frac{(1-\frac{d'_{cons}}{2d})}{t+\lambda}\right). \tag{115}$$

结合式 (108) 和 (115)，我们得到

$$\mathbb{E}\|\mathbf{u}_k^{t+1}\|^2\leq\left(1-\frac{d'_{cons}}{2d}\right)\mathbb{E}\|\mathbf{u}_k^t\|^2+\left(\frac{8I^2d^2b^2}{(d'_{cons})^2}(D^2+\sigma^2)\right)\cdot\left(\frac{1}{t+\lambda+1}-\frac{(1-\frac{d'_{cons}}{2d})}{t+\lambda}\right). \tag{116}$$

因此，

$$\left(\mathbb{E}\|\mathbf{u}_k^{t+1}\|^2-\frac{8I^2d^2b^2(D^2+\sigma^2)}{(d'_{cons})^2(t+\lambda+1)}\right)\leq\left(1-\frac{d'_{cons}}{2d}\right)\left(\mathbb{E}\|\mathbf{u}_k^t\|^2-\frac{8I^2d^2b^2(D^2+\sigma^2)}{(d'_{cons})^2(t+\lambda)}\right). \tag{117}$$

递归地使用 (117)，可得

$$\left(\mathbb{E}\|\mathbf{u}_k^t\|^2-\frac{8I^2d^2b^2(D^2+\sigma^2)}{(d'_{cons})^2(t+\lambda)}\right)\leq\left(1-\frac{d'_{cons}}{2d}\right)^t\left(\mathbb{E}\|\mathbf{u}_k^0\|^2-\frac{8I^2d^2b^2(D^2+\sigma^2)}{(d'_{cons})^2\lambda}\right)<0. \tag{118}$$

因此，

$$\mathbb{E}\|\mathbf{u}_k^t\|^2 \leq \frac{8I^2d^2(D^2+\sigma^2)}{(d'_{cons})^2} \cdot \frac{b^2}{t+\lambda} = \frac{8I^2d^2(D^2+\sigma^2)}{(d'_{cons})^2} \cdot (\eta_t)^2. \tag{119}$$

最后，

$$\mathbb{E}\|\mathbf{u}^t\|^2 = \mathbb{E}\|\frac{1}{|\mathcal{G}|}\sum_{k\in\mathcal{G}}\mathbf{u}_k^t\|^2 \leq \frac{1}{|\mathcal{G}|}\sum_{k\in\mathcal{G}}\mathbb{E}\|\mathbf{u}_k^t\|^2 \leq \frac{8I^2d^2(D^2+\sigma^2)}{(d'_{cons})^2} \cdot (\eta_t)^2. \tag{120}$$

当 $\eta_t = \eta$ 时，由 (108) 可得

$$\mathbb{E}\|\mathbf{u}_k^{t+1}\|^2 \leq \left(1 - \frac{d'_{cons}}{2d}\right)\mathbb{E}\|\mathbf{u}_k^t\|^2 + \frac{2I^2d}{d'_{cons}}\eta^2(D^2+\sigma^2) \tag{121}$$

$$\left(\mathbb{E}\|\mathbf{u}_k^{t+1}\|^2 - \frac{4I^2d^2}{(d'_{cons})^2}\eta^2(D^2+\sigma^2)\right) \leq \left(1 - \frac{d'_{cons}}{2d}\right) \cdot \left(\mathbb{E}\|\mathbf{u}_k^t\|^2 - \frac{4I^2d^2}{(d'_{cons})^2}\eta^2(D^2+\sigma^2)\right). \tag{122}$$

递归地使用 (122)，可得

$$\mathbb{E}\|\mathbf{u}_k^t\|^2 - \frac{4I^2d^2}{(d'_{cons})^2}\eta^2(D^2+\sigma^2) \leq \left(1 - \frac{d'_{cons}}{2d}\right)^t \cdot \left(\mathbb{E}\|\mathbf{u}_k^0\|^2 - \frac{4I^2d^2}{(d'_{cons})^2}\eta^2(D^2+\sigma^2)\right) < 0. \tag{123}$$

因此，

$$\mathbb{E}\|\mathbf{u}_k^t\|^2 \leq \frac{4I^2d^2(D^2+\sigma^2)}{(d'_{cons})^2} \cdot \eta^2. \tag{124}$$

最后，

$$\mathbb{E}\|\mathbf{u}^t\|^2 = \mathbb{E}\|\frac{1}{|\mathcal{G}|}\sum_{k\in\mathcal{G}}\mathbf{u}_k^t\|^2 \leq \frac{1}{|\mathcal{G}|}\sum_{k\in\mathcal{G}}\mathbb{E}\|\mathbf{u}_k^t\|^2 \leq \frac{4I^2d^2(D^2+\sigma^2)}{(d'_{cons})^2} \cdot \eta^2. \tag{125}$$

□

B.6 引理2的证明

证明。 根据假设6和假设5，我们有 $\forall k \in \mathcal{G}$,

$$\begin{aligned}
\mathbb{E}\|\tilde{\mathbf{g}}_k^t\|^2 \leq \mathbb{E}\|\mathbf{g}_k^t\|^2 =& \mathbb{E}\|\mathbf{u}_k^t + (\mathbf{w}^t - \mathbf{w}_k^{t+1})\|^2 && (126)\\
\leq& 2\mathbb{E}\|\mathbf{u}_k^t\|^2 + 2\mathbb{E}\|\mathbf{w}^t - \mathbf{w}_k^{t+1}\|^2 && (127)\\
\leq& 2\mathbb{E}\|\mathbf{u}_k^t\|^2 + 2I^2(\eta_t)^2(D^2+\sigma^2). && (128)
\end{aligned}$$

根据引理1，如果 $\eta_t = \frac{b}{\sqrt{t}+\lambda}$ 其中常数 $b > 0$ 且 $\lambda = \frac{4d}{d'_{cons}}$，则有

$$\mathbb{E}\|\tilde{\mathbf{g}}_k^t\|^2 \leq 2I^2(8H^2+1)(D^2+\sigma^2) \cdot (\eta_t)^2, \quad \forall k \in \mathcal{G}. \tag{129}$$

$$\mathbb{E}_{k\neq k'}\left[\|\tilde{\mathbf{g}}_k^t - \tilde{\mathbf{g}}_{k'}^t\|^2\right] \leq 2\mathbb{E}\|\tilde{\mathbf{g}}_k^t\|^2 + 2\mathbb{E}\|\tilde{\mathbf{g}}_{k'}^t\|^2 \leq 8I^2(8H^2+1)(D^2+\sigma^2) \cdot (\eta_t)^2. \tag{130}$$

因此，根据定义 1 和式（130），

$$\mathbb{E}\|\mathbf{e}^t\|^2 = \mathbb{E}\left\|\mathbf{SRAgg}(\{\tilde{\mathbf{g}}_k^t\}_{k=1}^m) - \frac{1}{|\mathcal{G}|}\sum_{k\in\mathcal{G}}\tilde{\mathbf{g}}_k^t\right\|^2 \tag{131}$$

$$\leq c\delta\cdot\mathbb{E}_{k\neq k'}\left[\|\tilde{\mathbf{g}}_k^t - \tilde{\mathbf{g}}_{k'}^t\|^2\right] \tag{132}$$

$$\leq 8c\delta I^2(8H^2+1)(D^2+\sigma^2)\cdot(\eta_t)^2. \tag{133}$$

同样地，如果 $\eta_t = \eta > 0$，我们有

$$\mathbb{E}\|\tilde{\mathbf{g}}_k^t\|^2 \leq 2I^2(4H^2+1)(D^2+\sigma^2)\cdot\eta^2, \quad \forall k\in\mathcal{G}. \tag{134}$$

and

$$\mathbb{E}\|\mathbf{e}^t\|^2 \leq \ 8c\delta I^2(4H^2+1)(D^2+\sigma^2)\cdot\eta^2. \tag{135}$$

□

B.7 定理 3 的证明

证明。令 $\mathbf{u}^t = \frac{1}{|\mathcal{G}|}\sum_{k\in\mathcal{G}}\mathbf{u}_k^t$ 为非拜占庭客户端的平均记忆。$\hat{}\mathbf{w}$ 定义为 $\hat{}\mathbf{w}^t = \mathbf{w}^t - \mathbf{u}^t$。$\hat{}\mathbf{w}$ 的迭代规则推导如下：

$$\hat{\mathbf{w}}^{t+1} = \mathbf{w}^{t+1} - \mathbf{u}^{t+1} \tag{136}$$

$$= \left(\mathbf{w}^t - \mathbf{SRAgg}(\{\tilde{\mathbf{g}}_k^t\}_{k=1}^m)\right) - \frac{1}{|\mathcal{G}|}\sum_{k\in\mathcal{G}}\left(\mathbf{u}_k^t + (\mathbf{w}^t - \mathbf{w}_k^{t+1}) - \tilde{\mathbf{g}}_k^t\right) \tag{137}$$

$$= \mathbf{w}^t - \mathbf{SRAgg}(\{\tilde{\mathbf{g}}_k^t\}_{k=1}^m) - \left(\mathbf{u}^t + \frac{1}{|\mathcal{G}|}\sum_{k\in\mathcal{G}}(\mathbf{w}^t - \mathbf{w}_k^{t+1}) - \frac{1}{|\mathcal{G}|}\sum_{k\in\mathcal{G}}\tilde{\mathbf{g}}_k^t\right) \tag{138}$$

$$= (\mathbf{w}^t - \mathbf{u}^t) - \frac{1}{|\mathcal{G}|}\sum_{k\in\mathcal{G}}(\mathbf{w}^t - \mathbf{w}_k^{t+1}) - \left(\mathbf{SRAgg}(\{\tilde{\mathbf{g}}_k^t\}_{k=1}^m) - \frac{1}{|\mathcal{G}|}\sum_{k\in\mathcal{G}}\tilde{\mathbf{g}}_k^t\right) \tag{139}$$

$$= \hat{\mathbf{w}}^t - \left(\mathbf{w}^t - \frac{1}{|\mathcal{G}|}\sum_{k\in\mathcal{G}}\mathbf{w}_k^{t+1}\right) - \mathbf{e}^t, \tag{140}$$

其中，$\mathbf{e}^t = \mathbf{SRAgg}(\{\tilde{}\mathbf{g}_k^t\}_{k=1}^m) - \frac{1}{|\mathcal{G}|}\sum_{k\in\mathcal{G}}\tilde{}\mathbf{g}_k^t$ 是 $\frac{1}{|\mathcal{G}|}\sum_{k\in\mathcal{G}}\tilde{}\mathbf{g}_k^t$ 的估计误差。令 $\bar{}\mathbf{G}^t = (\mathbf{w}^t - \frac{1}{|\mathcal{G}|}\sum_{k\in\mathcal{G}}\mathbf{w}_k^{t+1})/(\eta I)$。于是我们有

$$\bar{\mathbf{G}}^t = \frac{1}{I|\mathcal{G}|}\sum_{k\in\mathcal{G}}\sum_{j=0}^{I-1}\nabla f_{i_k^{t,j}}(\mathbf{w}_k^{t+1,j}), \quad t = 0,1,\ldots,T-1, \tag{141}$$

and

$$\hat{\mathbf{w}}^{t+1} = \hat{\mathbf{w}}^t - \eta I\cdot\bar{\mathbf{G}}^t - \mathbf{e}^t, \quad t = 0,1,\ldots,T-1. \tag{142}$$

该等式可以解释为：$\hat{}\mathbf{w}^{t+1}$ 是通过对 $\hat{}\mathbf{w}$ 执行一次随机梯度下降步骤得到的 ^{t} 学习率为 ηI，梯度近似为 $\bar{}\mathbf{G}^t$，误差为 $\mathbf{e}^t$。根据假设3以及不等式 $\|\mathbf{x}+\mathbf{y}\|^2 \leq 2\|\mathbf{x}\|^2 + 2\|\mathbf{y}\|^2$,

$$F(\hat{\mathbf{w}}^{t+1}) = F(\hat{\mathbf{w}}^t - \eta I\cdot\bar{\mathbf{G}}^t - \mathbf{e}^t) \tag{143}$$

$$\leq F(\hat{\mathbf{w}}^t) - \nabla F(\hat{\mathbf{w}}^t)^T(\eta I \cdot \bar{\mathbf{G}}^t + \mathbf{e}^t) + \frac{L}{2}\|\eta I \cdot \bar{\mathbf{G}}^t + \mathbf{e}^t\|^2 \tag{144}$$

$$\leq F(\hat{\mathbf{w}}^t) - \eta I \cdot \nabla F(\hat{\mathbf{w}}^t)^T \bar{\mathbf{G}}^t - \nabla F(\hat{\mathbf{w}}^t)^T \mathbf{e}^t + \eta^2 I^2 L \|\bar{\mathbf{G}}^t\|^2 + L\|\mathbf{e}^t\|^2 \tag{145}$$

$$= F(\hat{\mathbf{w}}^t) - \eta I \cdot \Big(\|\nabla F(\hat{\mathbf{w}}^t)\|^2 + \nabla F(\hat{\mathbf{w}}^t)^T [\bar{\mathbf{G}}^t - \nabla F(\hat{\mathbf{w}}^t)] \Big)$$
$$- \nabla F(\hat{\mathbf{w}}^t)^T \mathbf{e}^t + \eta^2 I^2 L \|\bar{\mathbf{G}}^t\|^2 + L\|\mathbf{e}^t\|^2 \tag{146}$$

$$= F(\hat{\mathbf{w}}^t) - \eta I \cdot \|\nabla F(\hat{\mathbf{w}}^t)\|^2 - \eta I \cdot \nabla F(\hat{\mathbf{w}}^t)^T [\bar{\mathbf{G}}^t - \nabla F(\hat{\mathbf{w}}^t)]$$
$$- \nabla F(\hat{\mathbf{w}}^t)^T \mathbf{e}^t + \eta^2 I^2 L \|\bar{\mathbf{G}}^t\|^2 + L\|\mathbf{e}^t\|^2. \tag{147}$$

对两边取期望，得

$$\mathbb{E}[F(\hat{\mathbf{w}}^{t+1})|\mathbf{w}^t, \mathbf{u}^t] \leq F(\hat{\mathbf{w}}^t) - \eta I \cdot \|\nabla F(\hat{\mathbf{w}}^t)\|^2 - \eta I \cdot \mathbb{E}\Big[\nabla F(\hat{\mathbf{w}}^t)^T [\bar{\mathbf{G}}^t - \nabla F(\hat{\mathbf{w}}^t)] \Big| \mathbf{w}^t, \mathbf{u}^t\Big]$$
$$- \mathbb{E}[\nabla F(\hat{\mathbf{w}}^t)^T \mathbf{e}^t | \mathbf{w}^t, \mathbf{u}^t] + \eta^2 I^2 L \cdot \mathbb{E}[\|\bar{\mathbf{G}}^t\|^2 | \mathbf{w}^t, \mathbf{u}^t] + L \cdot \mathbb{E}[\|\mathbf{e}^t\|^2 | \mathbf{w}^t, \mathbf{u}^t]. \tag{148}$$

根据假设3以及 $-\|\mathbf{x}\|^2 \leq -\frac{1}{2}\|\mathbf{y}\|^2 + \|\mathbf{x} - \mathbf{y}\|^2$，我们有：

$$-\|\nabla F(\hat{\mathbf{w}}^t)\|^2 = -\|\nabla F(\mathbf{w}^t) + [\nabla F(\hat{\mathbf{w}}^t) - \nabla F(\mathbf{w}^t)]\|^2 \tag{149}$$

$$\leq -\frac{1}{2}\|\nabla F(\mathbf{w}^t)\|^2 + \|\nabla F(\hat{\mathbf{w}}^t) - \nabla F(\mathbf{w}^t)\|^2 \tag{150}$$

$$= -\frac{1}{2}\|\nabla F(\mathbf{w}^t)\|^2 + \|\nabla F(\mathbf{w}^t - \mathbf{u}^t) - \nabla F(\mathbf{w}^t)\|^2 \tag{151}$$

$$\leq -\frac{1}{2}\|\nabla F(\mathbf{w}^t)\|^2 + L^2\|\mathbf{u}^t\|^2. \tag{152}$$

此外，利用假设3、假设 4、假设5和公式 (141)，我们有：

$$-\mathbb{E}\Big[\nabla F(\hat{\mathbf{w}}^t)^T [\bar{\mathbf{G}}^t - \nabla F(\hat{\mathbf{w}}^t)] \Big| \mathbf{w}^t, \mathbf{u}^t\Big]$$
$$= -\nabla F(\hat{\mathbf{w}}^t)^T \cdot \mathbb{E}[\bar{\mathbf{G}}^t - \nabla F(\hat{\mathbf{w}}^t) | \mathbf{w}^t, \mathbf{u}^t]$$
$$\leq \|\nabla F(\hat{\mathbf{w}}^t)\| \cdot \Big\| \mathbb{E}[\bar{\mathbf{G}}^t - \nabla F(\hat{\mathbf{w}}^t) | \mathbf{w}^t, \mathbf{u}^t] \Big\| \tag{153}$$

$$\leq D \cdot \left\| \mathbb{E}\left[\nabla F(\hat{\mathbf{w}}^t) - \frac{1}{I|\mathcal{G}|} \sum_{k\in\mathcal{G}} \sum_{j=0}^{I-1} \nabla f_{i_k^{t,j}}(\mathbf{w}_k^{t+1,j}) \middle| \mathbf{w}^t, \mathbf{u}^t \right] \right\| \tag{154}$$

$$\leq D \cdot \left\| \frac{1}{I|\mathcal{G}|} \sum_{k\in\mathcal{G}} \sum_{j=0}^{I-1} \mathbb{E}\left[\nabla F(\mathbf{w}^t - \mathbf{u}^t) - \nabla F_k(\mathbf{w}_k^{t+1,j}) \Big| \mathbf{w}^t, \mathbf{u}^t \right] \right\| \tag{155}$$

$$\leq \frac{D}{I|\mathcal{G}|} \sum_{k\in\mathcal{G}} \sum_{j=0}^{I-1} \mathbb{E}\left[\Big\| \nabla F(\mathbf{w}_k^{t+1,0} - \mathbf{u}^t) - \nabla F(\mathbf{w}_k^{t+1,j}) \Big\| \Big| \mathbf{w}^t, \mathbf{u}^t \right]$$
$$+ \frac{D}{I|\mathcal{G}|} \sum_{k\in\mathcal{G}} \sum_{j=0}^{I-1} \mathbb{E}\left[\Big\| \nabla F(\mathbf{w}_k^{t+1,j}) - \nabla F_k(\mathbf{w}_k^{t+1,j}) \Big\| \Big| \mathbf{w}^t, \mathbf{u}^t \right] \tag{156}$$

$$\leq \frac{D}{I|\mathcal{G}|}\sum_{k\in\mathcal{G}}\sum_{j=0}^{I-1}\Big(L\cdot\mathbb{E}\Big[\Big\|\mathbf{w}_k^{t+1,0}-\mathbf{u}^t-\mathbf{w}_k^{t+1,j}\Big\|\Big|\mathbf{w}^t,\mathbf{u}^t\Big]+B\Big) \tag{157}$$

$$\leq \frac{DL}{I|\mathcal{G}|}\sum_{k\in\mathcal{G}}\sum_{j=0}^{I-1}\Big(\|\mathbf{u}^t\|+\mathbb{E}\Big[\Big\|\mathbf{w}_k^{t+1,0}-\mathbf{w}_k^{t+1,j}\Big\|\Big|\mathbf{w}^t,\mathbf{u}^t\Big]\Big)+BD \tag{158}$$

$$\leq \frac{DL}{I|\mathcal{G}|}\sum_{k\in\mathcal{G}}\sum_{j=0}^{I-1}\Big(\sum_{j'=0}^{j-1}\mathbb{E}\Big[\Big\|\mathbf{w}_k^{t+1,j'}-\mathbf{w}_k^{t+1,j'+1}\Big\|\Big|\mathbf{w}^t,\mathbf{u}^t\Big]\Big)+DL\cdot\|\mathbf{u}^t\|+BD. \tag{159}$$

根据假设5，我们有：

$$\mathbb{E}\Big[\Big\|\mathbf{w}_k^{t+1,j'}-\mathbf{w}_k^{t+1,j'+1}\Big\|\Big|\mathbf{w}^t,\mathbf{u}^t\Big]=\mathbb{E}\Big[\Big\|\eta\cdot\nabla f_{i_k^{t,j'}}(\mathbf{w}_k^{t+1,j'})\Big\|\Big|\mathbf{w}^t,\mathbf{u}^t\Big]\leq\eta D. \tag{160}$$

因此，

$$\begin{aligned}
&-\mathbb{E}\Big[\nabla F(\hat{\mathbf{w}}^t)^T[\bar{\mathbf{G}}^t-\nabla F(\hat{\mathbf{w}}^t)]\Big|\mathbf{w}^t,\mathbf{u}^t\Big]\\
&\leq \frac{DL}{I|\mathcal{G}|}\sum_{k\in\mathcal{G}}\sum_{j=0}^{I-1}\Big(\sum_{j'=0}^{j-1}\eta D\Big)+DL\cdot\|\mathbf{u}^t\|+BD &(161)\\
&=\frac{4L^2}{I|\mathcal{G}|}\sum_{k\in\mathcal{G}}\sum_{j=0}^{I-1}\Big(j\eta D\Big)+DL\cdot\|\mathbf{u}^t\|+BD &(162)\\
&=\frac{4L^2}{I|\mathcal{G}|}\cdot|\mathcal{G}|\frac{I(I-1)}{2}\eta D+DL\cdot\|\mathbf{u}^t\|+BD &(163)\\
&=2(I-1)\eta DL^2+DL\cdot\|\mathbf{u}^t\|+BD. &(164)
\end{aligned}$$

注意到 $\mathbb{E}[XY]\leq\sqrt{\mathbb{E}[X^2]\cdot\mathbb{E}[Y^2]}$。根据假设5和引理2，我们有：

$$\begin{aligned}
-\mathbb{E}[\nabla F(\hat{\mathbf{w}}^t)^T\mathbf{e}^t|\mathbf{w}^t,\mathbf{u}^t]&\leq\mathbb{E}[\|\nabla F(\hat{\mathbf{w}}^t)\|\cdot\|\mathbf{e}^t\||\mathbf{w}^t,\mathbf{u}^t] &(165)\\
&\leq\sqrt{\mathbb{E}[\|\nabla F(\hat{\mathbf{w}}^t)\|^2|\mathbf{w}^t,\mathbf{u}^t]\cdot\mathbb{E}[\|\mathbf{e}^t\|^2|\mathbf{w}^t,\mathbf{u}^t]} &(166)\\
&\leq\sqrt{8c\delta I^2(4H^2+1)D^2(D^2+\sigma^2)\cdot\eta^2} &(167)\\
&=\eta I\sqrt{8c\delta(4H^2+1)(D^2+\sigma^2)}D. &(168)
\end{aligned}$$

根据假设5和假设6，

$$\begin{aligned}
\mathbb{E}[\|\bar{\mathbf{G}}^t\|^2|\mathbf{w}^t,\mathbf{u}^t]&=\mathbb{E}\left[\left\|\frac{1}{I|\mathcal{G}|}\sum_{k\in\mathcal{G}}\sum_{j=0}^{I-1}\nabla f_{i_k^{t,j}}(\mathbf{w}_k^{t+1,j})\right\|^2\middle|\mathbf{w}^t,\mathbf{u}^t\right] &(169)\\
&\leq\frac{1}{I|\mathcal{G}|}\sum_{k\in\mathcal{G}}\sum_{j=0}^{I-1}\mathbb{E}\left[\left\|\nabla f_{i_k^{t,j}}(\mathbf{w}_k^{t+1,j})\right\|^2\middle|\mathbf{w}^t,\mathbf{u}^t\right] &(170)\\
&\leq\frac{1}{I|\mathcal{G}|}\sum_{k\in\mathcal{G}}\sum_{j=0}^{I-1}(D^2+\sigma^2) &(171)
\end{aligned}$$

$$= D^2 + \sigma^2. \tag{172}$$

将（135）、（152）、（164）、（168）和（172）代入（148），可得:

$$\begin{aligned}\mathbb{E}[F(\hat{\mathbf{w}}^{t+1})|\mathbf{w}^t, \mathbf{u}^t] \leq\ & F(\hat{\mathbf{w}}^t) - \eta I \cdot \|\nabla F(\hat{\mathbf{w}}^t)\|^2 - \eta I \cdot \mathbb{E}\Big[\nabla F(\hat{\mathbf{w}}^t)^T[\bar{\mathbf{G}}^t - \nabla F(\hat{\mathbf{w}}^t)]\Big|\mathbf{w}^t, \mathbf{u}^t\Big] \\ & - \mathbb{E}[\nabla F(\hat{\mathbf{w}}^t)^T \mathbf{e}^t|\mathbf{w}^t, \mathbf{u}^t] + \eta^2 I^2 L \cdot \mathbb{E}[\|\bar{\mathbf{G}}^t\|^2|\mathbf{w}^t, \mathbf{u}^t] + L \cdot \mathbb{E}[\|\mathbf{e}^t\|^2|\mathbf{w}^t, \mathbf{u}^t]\end{aligned} \tag{173}$$

$$\begin{aligned}& \leq F(\hat{\mathbf{w}}^t) - \frac{\eta I}{2}\|\nabla F(\mathbf{w}^t)\|^2 + \eta I L^2 \|\mathbf{u}^t\|^2 \\ & + \eta I\Big[2(I-1)\eta D L^2 + DL\|\mathbf{u}^t\| + BD\Big] \\ & + \eta I\sqrt{8c\delta(4H^2+1)(D^2+\sigma^2)}D \\ & + \eta^2 I^2 L(D^2+\sigma^2) + L \cdot [8c\delta I^2(4H^2+1)(D^2+\sigma^2)\cdot\eta^2].\end{aligned} \tag{174}$$

注意到 $\mathbb{E}\|\mathbf{u}^t\| = \sqrt{[\mathbb{E}\|\mathbf{u}^t\|]^2} \leq \sqrt{[\mathbb{E}\|\mathbf{u}^t\|^2]}$。对两边取全期望，并利用 $\mathbb{E}\|\mathbf{u}^t\|^2 \leq 4H^2I^2(D^2+\sigma^2)\cdot\eta^2$，可得:

$$\begin{aligned}\mathbb{E}[F(\hat{\mathbf{w}}^{t+1})] \leq\ & \mathbb{E}[F(\hat{\mathbf{w}}^t)] - \frac{\eta I}{2}\mathbb{E}\|\nabla F(\mathbf{w}^t)\|^2 + \eta I L^2[4H^2I^2(D^2+\sigma^2)\cdot\eta^2] \\ & + \eta I\Big[2(I-1)\eta D L^2 + DL\cdot\sqrt{4H^2I^2(D^2+\sigma^2)\cdot\eta^2} + BD\Big] \\ & + \eta I\sqrt{8c\delta(4H^2+1)(D^2+\sigma^2)}D \\ & + \eta^2 I^2 L(D^2+\sigma^2) + L\cdot[8c\delta I^2(4H^2+1)(D^2+\sigma^2)\cdot\eta^2].\end{aligned} \tag{175}$$

即,

$$\begin{aligned}\mathbb{E}[F(\hat{\mathbf{w}}^{t+1})] \leq\ & \mathbb{E}[F(\hat{\mathbf{w}}^t)] - \frac{\eta I}{2}\mathbb{E}\|\nabla F(\mathbf{w}^t)\|^2 \\ & + (\eta I)^2 L\Big[2(1-I^{-1})DL + 2HD\sqrt{D^2+\sigma^2} + (D^2+\sigma^2) + 8c\delta(4H^2+1)(D^2+\sigma^2)\Big] \\ & + (\eta I)^3\Big[4H^2L^2(D^2+\sigma^2)\Big] + (\eta I)\Big[BD + \sqrt{8c\delta(4H^2+1)(D^2+\sigma^2)}D\Big].\end{aligned} \tag{176}$$

对 $t=0$ 到 $T-1$ 求和，可得:

$$\begin{aligned}\mathbb{E}[F(\hat{\mathbf{w}}^T)] \leq\ & \mathbb{E}[F(\hat{\mathbf{w}}^0)] - \frac{\eta I}{2}\cdot\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(\mathbf{w}^t)\|^2 \\ & + T(\eta I)^2 L\Big[2(1-I^{-1})DL + 2HD\sqrt{D^2+\sigma^2} + (D^2+\sigma^2) + 8c\delta(4H^2+1)(D^2+\sigma^2)\Big] \\ & + T(\eta I)^3\Big[4H^2L^2(D^2+\sigma^2)\Big] + T(\eta I)\Big[BD + \sqrt{8c\delta(4H^2+1)(D^2+\sigma^2)}D\Big].\end{aligned} \tag{177}$$

注意，$\hat{\mathbf{w}}^0 = \mathbf{w}^0$和 $F(\hat{\mathbf{w}}^T) \geq F^*$。因此,

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(\mathbf{w}^t)\|^2 \leq \frac{2[F(\hat{\mathbf{w}}^0) - F^*]}{\eta I T}$$

$$+\eta\cdot 2IL\Big[2(1-I^{-1})DL+2HD\sqrt{D^2+\sigma^2}+(D^2+\sigma^2)+8c\delta(4H^2+1)(D^2+\sigma^2)\Big]$$
$$+\eta^2\cdot\Big[8H^2I^2L^2(D^2+\sigma^2)\Big]+2\Big[BD+\sqrt{8c\delta(4H^2+1)(D^2+\sigma^2)}D\Big]. \tag{178}$$

总之，

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(\mathbf{w}^t)\|^2\le\frac{2[F(\hat{\mathbf{w}}^0)-F^*]}{\eta IT}+\eta\gamma_1+\eta^2\gamma_2+\Delta, \tag{179}$$

其中 $\gamma_1=2IL\cdot[2(1-I^{-1})LD+2HD\sqrt{D^2+\sigma^2}+(D^2+\sigma^2)+8c\delta(4H^2+1)(D^2+\sigma^2)]$，$\gamma_2=8H^2I^2L^2(D^2+\sigma^2)$ 和 $\Delta=2BD+4\sqrt{2c\delta(4H^2+1)(D^2+\sigma^2)D}$。

□

B.8 本地动量随机梯度下降的分析

我们给出以下命题，说明当 $\mathcal{A}$ 设为本地动量随机梯度下降时，假设 7 成立。

命题 3. 在假设*3*、*5*和*6*下，本地动量随机梯度下降满足假设 *7*。此外，对于学习率为 $\eta>0$、更新间隔为 $I\in\mathbb{N}_+$、动量超参数为 $\beta\in[0,1)$ 的本地动量随机梯度下降，我们有 $\eta_{\mathcal{A}}=\eta I$, $A_1=\frac{\beta}{(1-\beta^I)I(1-\beta)}$ $D+\sqrt{D^2+\sigma^2}+\big(\frac{I-1}{2}+\frac{\beta^2}{(1-\beta^{I-1})I\overline{(1-\beta)^2}}-\frac{\beta}{(I-1)I(1-\beta)}\big)$ ~~$\cdot L\sqrt{D^2+\sigma^2}$~~ 并且 ~~$(A_2)^2$~~ $=D^2+\overline{\sigma^2}$。

证明。当 $\mathcal{A}$ 设置为学习率为 η、更新间隔为 I且动量超参数为 β 的本地动量随机梯度下降时，令 $\mathbf{m}_k^{0\,j}=\mathbf{0}$,为初始动量，且 $\mathbf{w}_k^{t+1}$ $(t=0,1,\ldots,T-1)$ 按照以下过程计算：

$$\begin{cases}\mathbf{m}_k^{t+1,0}=\mathbf{m}_k^{t,I};\\ \mathbf{w}_k^{t+1,0}=\mathbf{w}^t;\\ \mathbf{m}_k^{t+1,j+1}=\beta\cdot\mathbf{m}_k^{t+1,j}+(1-\beta)\cdot\nabla f_{i_k^{t,j}}(\mathbf{w}_k^{t+1,j}), & j=0,1,\ldots,I-1;\\ \mathbf{w}_k^{t+1,j+1}=\mathbf{w}_k^{t+1,j}-\eta\cdot\mathbf{m}_k^{t+1,j+1}, & j=0,1,\ldots,I-1;\\ \mathbf{w}_k^{t+1}=\mathbf{w}_k^{t+1,I}.\end{cases} \tag{180}$$

令 $\eta_{\mathcal{A}}=\eta I$，我们有

$$\mathbf{G}_{\mathcal{A}}(\mathbf{w}^t;\mathcal{D}_k)=(\mathbf{w}^t-\mathbf{w}_k^{t+1})/(\eta I)=\frac{1}{I}\sum_{j=0}^{I-1}(\mathbf{w}_k^{t+1,j}-\mathbf{w}_k^{t+1,j+1})=\frac{1}{I}\sum_{j=0}^{I-1}\mathbf{m}_k^{t+1,j+1}. \tag{181}$$

此外，

$$\mathbf{m}_k^{t+1,j+1}=\beta\cdot\mathbf{m}_k^{t+1,j}+(1-\beta)\cdot\nabla f_{i_k^{t,j}}(\mathbf{w}_k^{t+1,j}) \tag{182}$$
$$=\beta\cdot(\beta\cdot\mathbf{m}_k^{t+1,j-1}+(1-\beta)\cdot\nabla f_{i_k^{t,j-1}}(\mathbf{w}_k^{t+1,j-1}))+(1-\beta)\cdot\nabla f_{i_k^{t,j}}(\mathbf{w}_k^{t+1,j}) \tag{183}$$
$$=\beta^2\cdot\mathbf{m}_k^{t+1,j-1}+\beta(1-\beta)\cdot\nabla f_{i_k^{t,j-1}}(\mathbf{w}_k^{t+1,j-1})+(1-\beta)\cdot\nabla f_{i_k^{t,j}}(\mathbf{w}_k^{t+1,j}) \tag{184}$$

$$= \cdots\cdots\cdots\cdots$$

$$= \beta^{j+1}\mathbf{m}_k^{t+1,0} + (1-\beta)\sum_{j'=0}^{j}\beta^{j-j'}\nabla f_{i_k^{t,j'}}(\mathbf{w}_k^{t+1,j'}). \tag{185}$$

现在我们通过对 t 进行演绎，证明 $\mathbb{E}\|\mathbf{m}_k^{t\,j}\|^2 \le D^2+\sigma^2, (j=0,1,\ldots,I)$.

步骤 1。当 $t=0$ 时，我们有 $\mathbb{E}\|\mathbf{m}_k^{0\,j}\|^2 = 0 \le D^2+\sigma^2$。

步骤 2（推导）。假设 $\mathbb{E}\|\mathbf{m}_k^{t\,j}\|^2 \le D^2+\sigma^2$，我们有 $\mathbb{E}\|\mathbf{m}_k^{t+1\,0}\|^2 = \mathbb{E}\|\mathbf{m}_k^{t\,I}\|^2 \le D^2+\sigma^2$ 以及 $\mathbb{E}\|\nabla f_{i_k^{t,j'}}(\mathbf{w}_k^{t+1\,j'})\|^2 \le D^2+\sigma^2$。由于

$$\beta^{j+1} + (1-\beta)\sum_{j'=0}^{j}\beta^{j-j'} = \beta^{j+1} + (1-\beta)\frac{1-\beta^{j+1}}{1-\beta} = 1, \tag{186}$$

$\mathbf{m}_k^{t+1\,j+1}$，可以视为对$\mathbf{m}_k^{t+1\,0}$，和 $\{\nabla f_{i_k^{t,j'}}(\mathbf{w}_k^{t+1\,j'})\}_{j'=0}^{j}$进行加权取平均。因此，

$$\mathbb{E}\|\mathbf{m}_k^{t+1,j+1}\|^2 \le D^2+\sigma^2, \quad j=1,2,\ldots,I. \tag{187}$$

通过数学推导， $\forall t=0,1,\ldots,T$，我们得到

$$\mathbb{E}\|\mathbf{m}_k^{t,j}\|^2 \le D^2+\sigma^2, \quad j=0,1,2,\ldots,I. \tag{188}$$

因此，

$$\mathbb{E}\|\mathbf{G}_{\mathcal{A}}(\mathbf{w}^t;\mathcal{D}_k)\|^2 = \mathbb{E}\left\|\frac{1}{I}\sum_{j=0}^{I-1}\mathbf{m}_k^{t+1,j+1}\right\|^2 \le D^2+\sigma^2. \tag{189}$$

将 (185) 代入 (181)，可得:

$$\mathbf{G}_{\mathcal{A}}(\mathbf{w}^t;\mathcal{D}_k) = \frac{1}{I}\sum_{j=0}^{I-1}\left[\beta^{j+1}\mathbf{m}_k^{t+1,0} + (1-\beta)\sum_{j'=0}^{j}\beta^{j-j'}\nabla f_{i_k^{t,j'}}(\mathbf{w}_k^{t+1,j'})\right] \tag{190}$$

$$= \frac{\beta(1-\beta^I)}{I(1-\beta)}\mathbf{m}_k^{t+1,0} + \frac{1-\beta}{I}\sum_{j=0}^{I-1}\left[\sum_{j'=0}^{j}\beta^{j-j'}\nabla f_{i_k^{t,j'}}(\mathbf{w}_k^{t+1,j'})\right] \tag{191}$$

$$= \frac{\beta(1-\beta^I)}{I(1-\beta)}\mathbf{m}_k^{t+1,0} + \frac{1-\beta}{I}\sum_{j'=0}^{I-1}\left[\sum_{j=j'}^{I-1}\beta^{j-j'}\nabla f_{i_k^{t,j'}}(\mathbf{w}_k^{t+1,j'})\right] \tag{192}$$

$$= \frac{\beta(1-\beta^I)}{I(1-\beta)}\mathbf{m}_k^{t+1,0} + \frac{1-\beta}{I}\sum_{j'=0}^{I-1}\left[\frac{1-\beta^{I-j'}}{1-\beta}\nabla f_{i_k^{t,j'}}(\mathbf{w}_k^{t+1,j'})\right] \tag{193}$$

$$= \frac{1}{I}\left[\frac{\beta(1-\beta^I)}{1-\beta}\mathbf{m}_k^{t+1,0} + \sum_{j'=0}^{I-1}(1-\beta^{I-j'})\nabla f_{i_k^{t,j'}}(\mathbf{w}_k^{t+1,j'})\right]. \tag{194}$$

因此，

$$\mathbb{E}[\mathbf{G}_{\mathcal{A}}(\mathbf{w}^t;\mathcal{D}_k) - \nabla F_k(\mathbf{w}^t)] = \frac{1}{I}\left[\frac{\beta(1-\beta^I)}{1-\beta}\cdot\mathbb{E}[\mathbf{m}_k^{t+1,0} - \nabla F_k(\mathbf{w}^t)]\right.$$

$$+\sum_{j'=0}^{I-1}(1-\beta^{I-j'})\cdot\mathbb{E}[\nabla f_{i_k^{t,j'}}(\mathbf{w}_k^{t+1,j'})-\nabla F_k(\mathbf{w}^t)]\Bigg]. \quad (195)$$

由于 $\mathbb{E}\|\mathbf{m}_k^{t+1\,0}-\nabla F_{k,}(\mathbf{w}^t)\|\le\sqrt{D^2+\sigma^2}+D$ 并且

$$\begin{aligned}
&\mathbb{E}\|\nabla f_{i_k^{t,j'}}(\mathbf{w}_k^{t+1,j'})-\nabla F_k(\mathbf{w}^t)\| \\
&\le \mathbb{E}\|\nabla f_{i_k^{t,j'}}(\mathbf{w}_k^{t+1,j'})-\nabla F_k(\mathbf{w}_k^{t+1,j'})\|+\mathbb{E}\|\nabla F_k(\mathbf{w}_k^{t+1,j'})-\nabla F_k(\mathbf{w}^t)\| && (196)\\
&\le \sqrt{D^2+\sigma^2}+L\cdot\|\mathbf{w}_k^{t+1,j'}-\mathbf{w}^t\| && (197)\\
&\le \sqrt{D^2+\sigma^2}+L\sum_{j''=0}^{j'-1}\|\mathbf{m}_k^{t+1,j''+1}\| && (198)\\
&\le \sqrt{D^2+\sigma^2}+j'L\sqrt{D^2+\sigma^2}, && (199)
\end{aligned}$$

我们有

$$\begin{aligned}
&\|\mathbb{E}[\mathbf{G}_{\mathcal{A}}(\mathbf{w}^t;\mathcal{D}_k)]-\nabla F_k(\mathbf{w}^t)\| \\
&\le \frac{1}{I}\left[\frac{\beta(1-\beta^I)}{1-\beta}\cdot[\sqrt{D^2+\sigma^2}+D]+\sum_{j'=0}^{I-1}(1-\beta^{I-j'})\cdot[\sqrt{D^2+\sigma^2}+j'L\sqrt{D^2+\sigma^2}]\right] && (200)\\
&= \frac{1}{I}\left[\frac{\beta(1-\beta^I)}{1-\beta}D+I\sqrt{D^2+\sigma^2}+L\sqrt{D^2+\sigma^2}\cdot\sum_{j'=0}^{I-1}(j'-j'\beta^{I-j'})\right] && (201)\\
&= \frac{1}{I}\left[\frac{\beta(1-\beta^I)}{1-\beta}D+I\sqrt{D^2+\sigma^2}+L\sqrt{D^2+\sigma^2}\cdot\Big(\frac{I(I-1)}{2}+\frac{\beta^2(1-\beta^{I-1})}{(1-\beta)^2}-\frac{\beta(I-1)}{1-\beta}\Big)\right] && (202)\\
&= \frac{\beta(1-\beta^I)}{I(1-\beta)}D+\sqrt{D^2+\sigma^2}+\Big(\frac{I-1}{2}+\frac{\beta^2(1-\beta^{I-1})}{I(1-\beta)^2}-\frac{\beta(I-1)}{I(1-\beta)}\Big)\cdot L\sqrt{D^2+\sigma^2}. && (203)
\end{aligned}$$

□

B.9 定理 4 的证明

证明。 与引理1和引理2类似，对于满足假设 7 的一般训练算法 $\mathcal{A}$ ，我们有以下不等式，分别用于界定本地记忆和聚合误差:

$$\begin{aligned}
\mathbb{E}\|\mathbf{u}_k^{t+1}\|^2 &= \mathbb{E}\|\mathbf{g}_k^t-\tilde{\mathbf{g}}_k^t\|^2 && (204)\\
&\overset{\text{(i)}}{\le}\left(1-\frac{d'_{cons}}{d}\right)\mathbb{E}\|\mathbf{g}_k^t\|^2 && (205)\\
&=\left(1-\frac{d'_{cons}}{d}\right)\mathbb{E}\|\mathbf{u}_k^t+(\mathbf{w}^t-\mathbf{w}_k^{t+1})\|^2 && (206)\\
&\overset{\text{(ii)}}{\le}\left(1-\frac{d'_{cons}}{d}\right)\left[(1+\frac{d'_{cons}}{2d})\mathbb{E}\|\mathbf{u}_k^t\|^2+(1+\frac{2d}{d'_{cons}})\mathbb{E}\|\eta_{\mathcal{A}}\cdot\mathbf{G}_{\mathcal{A}}(\mathbf{w}^t;\mathcal{D}_k)\|^2\right] && (207)
\end{aligned}$$

$$\overset{\text{(iii)}}{\leq}\left(1-\frac{d'_{cons}}{2d}\right)\mathbb{E}\|\mathbf{u}_k^t\|^2+\frac{2d}{d'_{cons}}(\eta_{\mathcal{A}})^2\cdot\mathbb{E}\|\mathbf{G}_{\mathcal{A}}(\mathbf{w}^t;\mathcal{D}_k)\|^2 \tag{208}$$

$$\overset{\text{(iv)}}{\leq}\left(1-\frac{d'_{cons}}{2d}\right)\mathbb{E}\|\mathbf{u}_k^t\|^2+\frac{2d}{d'_{cons}}(\eta_{\mathcal{A}})^2(A_2)^2, \tag{209}$$

其中，(i) 根据命题 2 得出。(ii) 根据以下事实得出：$\|\mathbf{x}+\mathbf{y}\|^2\leq(1+\theta)\|\mathbf{x}\|^2+(1+\theta^{-1})\|\mathbf{y}\|^2$ 对任意常数 $\theta>0$ 均成立。(iii) 根据以下事实得出：$(1-\frac{d'_{cons}}{d})(1+\frac{d'_{cons}}{2d})<1-\frac{d'_{cons}}{2d}$ 且 $(1-\frac{d'_{cons}}{d})(1+\frac{2d}{d'_{cons}})<\frac{2d}{d'_{cons}}$。(iv) 根据假设 7 得出。因此，

$$\left(\mathbb{E}\|\mathbf{u}_k^{t+1}\|^2-\frac{4d^2}{(d'_{cons})^2}(\eta_{\mathcal{A}})^2(A_2)^2\right)\leq\left(1-\frac{d'_{cons}}{2d}\right)\cdot\left(\mathbb{E}\|\mathbf{u}_k^t\|^2-\frac{4d^2}{(d'_{cons})^2}(\eta_{\mathcal{A}})^2(A_2)^2\right). \tag{210}$$

递归地使用 (210)，可得

$$\left(\mathbb{E}\|\mathbf{u}_k^t\|^2-\frac{4d^2}{(d'_{cons})^2}(\eta_{\mathcal{A}})^2(A_2)^2\right)\leq\left(1-\frac{d'_{cons}}{2d}\right)^t\cdot\left(\mathbb{E}\|\mathbf{u}_k^0\|^2-\frac{4d^2}{(d'_{cons})^2}(\eta_{\mathcal{A}})^2(A_2)^2\right)<0. \tag{211}$$

因此，

$$\mathbb{E}\|\mathbf{u}_k^t\|^2\leq\frac{4d^2(A_2)^2}{(d'_{cons})^2}\cdot(\eta_{\mathcal{A}})^2. \tag{212}$$

令 $H=d'_{cons}/d$。最后，可得

$$\mathbb{E}\|\mathbf{u}^t\|^2=\mathbb{E}\|\frac{1}{|\mathcal{G}|}\sum_{k\in\mathcal{G}}\mathbf{u}_k^t\|^2\leq\frac{1}{|\mathcal{G}|}\sum_{k\in\mathcal{G}}\mathbb{E}\|\mathbf{u}_k^t\|^2\leq\frac{4d^2(A_2)^2}{(d'_{cons})^2}\cdot(\eta_{\mathcal{A}})^2=4H^2(A_2)^2(\eta_{\mathcal{A}})^2. \tag{213}$$

根据假设 7，可得 $\forall k\in\mathcal{G}$,

$$\begin{aligned}\mathbb{E}\|\tilde{\mathbf{g}}_k^t\|^2\leq\mathbb{E}\|\mathbf{g}_k^t\|^2=&\mathbb{E}\|\mathbf{u}_k^t+(\mathbf{w}^t-\mathbf{w}_k^{t+1})\|^2 &(214)\\ \leq&2\mathbb{E}\|\mathbf{u}_k^t\|^2+2\mathbb{E}\|\mathbf{w}^t-\mathbf{w}_k^{t+1}\|^2 &(215)\\ \leq&2\mathbb{E}\|\mathbf{u}_k^t\|^2+2(\eta_{\mathcal{A}})^2(A_2)^2 &(216)\\ \leq&2(4H^2+1)(A_2)^2\cdot(\eta_{\mathcal{A}})^2. &(217)\end{aligned}$$

因此，

$$\mathbb{E}_{k\neq k'}\left[\|\tilde{\mathbf{g}}_k^t-\tilde{\mathbf{g}}_{k'}^t\|^2\right]\leq2\mathbb{E}\|\tilde{\mathbf{g}}_k^t\|^2+2\mathbb{E}\|\tilde{\mathbf{g}}_{k'}^t\|^2\leq8(4H^2+1)(A_2)^2\cdot(\eta_{\mathcal{A}})^2. \tag{218}$$

因此，根据定义 1 和（218），

$$\begin{aligned}\mathbb{E}\|\mathbf{e}^t\|^2=&\mathbb{E}\left\|\mathbf{SRAgg}(\{\tilde{\mathbf{g}}_k^t\}_{k=1}^m)-\frac{1}{|\mathcal{G}|}\sum_{k\in\mathcal{G}}\tilde{\mathbf{g}}_k^t\right\|^2 &(219)\\ \leq&c\delta\cdot\mathbb{E}_{k\neq k'}\left[\|\tilde{\mathbf{g}}_k^t-\tilde{\mathbf{g}}_{k'}^t\|^2\right] &(220)\\ \leq&8c\delta(4H^2+1)(A_2)^2\cdot(\eta_{\mathcal{A}})^2. &(221)\end{aligned}$$

令 $\bar{\mathbf{w}}^{t+1} = \frac{1}{|\mathcal{G}|}\sum_{k\in\mathcal{G}}\mathbf{w}_k^{t+1}$。结合方程（140），可得：

$$\hat{\mathbf{w}}^{t+1} = \hat{\mathbf{w}}^t - (\mathbf{w}^t - \bar{\mathbf{w}}^{t+1}) - \mathbf{e}^t. \tag{222}$$

该等式可以解释为：在 $\hat{\mathbf{w}}^{t+1}$ 上加上一个小项 $-(\mathbf{w}^t - \bar{\mathbf{w}}^{t+1})$，并带有误差 $\mathbf{e}^t$，即可得到 $\hat{\mathbf{w}}^t$。因此，

$$F(\hat{\mathbf{w}}^{t+1}) = F(\hat{\mathbf{w}}^t - (\mathbf{w}^t - \bar{\mathbf{w}}^{t+1}) - \mathbf{e}^t) \tag{223}$$

$$\leq F(\hat{\mathbf{w}}^t) - \nabla F(\hat{\mathbf{w}}^t)^T(\mathbf{w}^t - \bar{\mathbf{w}}^{t+1} + \mathbf{e}^t) + \frac{L}{2}\|\mathbf{w}^t - \bar{\mathbf{w}}^{t+1} + \mathbf{e}^t\|^2 \tag{224}$$

$$\leq F(\hat{\mathbf{w}}^t) - \nabla F(\hat{\mathbf{w}}^t)^T(\mathbf{w}^t - \bar{\mathbf{w}}^{t+1}) - \nabla F(\hat{\mathbf{w}}^t)^T\mathbf{e}^t + \eta^2 I^2 L\|\mathbf{w}^t - \bar{\mathbf{w}}^{t+1}\|^2 + L\|\mathbf{e}^t\|^2 \tag{225}$$

$$= F(\hat{\mathbf{w}}^t) - \frac{1}{|\mathcal{G}|}\sum_{k\in\mathcal{G}}\nabla F(\hat{\mathbf{w}}^t)^T(\mathbf{w}^t - \mathbf{w}_k^{t+1}) - \nabla F(\hat{\mathbf{w}}^t)^T\mathbf{e}^t + L\left\|\mathbf{w}^t - \frac{1}{|\mathcal{G}|}\sum_{k\in\mathcal{G}}\mathbf{w}_k^{t+1}\right\|^2 + L\|\mathbf{e}^t\|^2 \tag{226}$$

$$\leq F(\hat{\mathbf{w}}^t) - \eta_{\mathcal{A}}\|\nabla F(\hat{\mathbf{w}}^t)\|^2 - \frac{1}{|\mathcal{G}|}\sum_{k\in\mathcal{G}}\nabla F(\hat{\mathbf{w}}^t)^T[\mathbf{w}^t - \mathbf{w}_k^{t+1} - \eta_{\mathcal{A}}\cdot\nabla F(\hat{\mathbf{w}}^t)] - \nabla F(\hat{\mathbf{w}}^t)^T\mathbf{e}^t + \frac{L}{|\mathcal{G}|}\sum_{k\in\mathcal{G}}\left\|\mathbf{w}^t - \mathbf{w}_k^{t+1}\right\|^2 + L\|\mathbf{e}^t\|^2 \tag{227}$$

$$= F(\hat{\mathbf{w}}^t) - \eta_{\mathcal{A}}\|\nabla F(\hat{\mathbf{w}}^t)\|^2 - \frac{1}{|\mathcal{G}|}\sum_{k\in\mathcal{G}}\nabla F(\hat{\mathbf{w}}^t)^T[\eta_{\mathcal{A}}\cdot\mathbf{G}_{\mathcal{A}}(\mathbf{w}^t;\mathcal{D}_k) - \eta_{\mathcal{A}}\cdot\nabla F(\hat{\mathbf{w}}^t)] - \nabla F(\hat{\mathbf{w}}^t)^T\mathbf{e}^t + \frac{L}{|\mathcal{G}|}\sum_{k\in\mathcal{G}}\left\|\eta_{\mathcal{A}}\cdot\mathbf{G}_{\mathcal{A}}(\mathbf{w}^t;\mathcal{D}_k)\right\|^2 + L\|\mathbf{e}^t\|^2. \tag{228}$$

对两边取期望，得

$$\begin{aligned}&\mathbb{E}[F(\hat{\mathbf{w}}^{t+1})|\mathbf{w}^t,\mathbf{u}^t]\\ \leq\ & F(\hat{\mathbf{w}}^t) - \eta_{\mathcal{A}}\|\nabla F(\hat{\mathbf{w}}^t)\|^2 - \frac{\eta_{\mathcal{A}}}{|\mathcal{G}|}\sum_{k\in\mathcal{G}}\mathbb{E}\Big[\nabla F(\hat{\mathbf{w}}^t)^T[\mathbf{G}_{\mathcal{A}}(\mathbf{w}^t;\mathcal{D}_k) - \nabla F(\hat{\mathbf{w}}^t)]\Big|\mathbf{w}^t,\mathbf{u}^t\Big]\\ -\ & \mathbb{E}[\nabla F(\hat{\mathbf{w}}^t)^T\mathbf{e}^t|\mathbf{w}^t,\mathbf{u}^t] + \frac{(\eta_{\mathcal{A}})^2 L}{|\mathcal{G}|}\sum_{k\in\mathcal{G}}\mathbb{E}\Big[\left\|\mathbf{G}_{\mathcal{A}}(\mathbf{w}^t;\mathcal{D}_k)\right\|^2\Big|\mathbf{w}^t,\mathbf{u}^t\Big] + L\cdot\mathbb{E}[\|\mathbf{e}^t\|^2|\mathbf{w}^t,\mathbf{u}^t].\end{aligned} \tag{229}$$

利用假设3、假设 4、假设 7，我们得到：

$$\begin{aligned}&-\mathbb{E}\Big[\nabla F(\hat{\mathbf{w}}^t)^T[\mathbf{G}_{\mathcal{A}}(\mathbf{w}^t;\mathcal{D}_k) - \nabla F(\hat{\mathbf{w}}^t)]\Big|\mathbf{w}^t,\mathbf{u}^t\Big]\\ =\ & -\nabla F(\hat{\mathbf{w}}^t)^T\big[\mathbb{E}[\mathbf{G}_{\mathcal{A}}(\mathbf{w}^t;\mathcal{D}_k)|\mathbf{w}^t,\mathbf{u}^t] - \nabla F(\hat{\mathbf{w}}^t)\big] \qquad (230)\\ \leq\ & \|\nabla F(\hat{\mathbf{w}}^t)\|\cdot\big\|\mathbb{E}[\mathbf{G}_{\mathcal{A}}(\mathbf{w}^t;\mathcal{D}_k)|\mathbf{w}^t,\mathbf{u}^t] - \nabla F(\hat{\mathbf{w}}^t)\big\| \qquad (231)\\ \leq\ & \|\nabla F(\hat{\mathbf{w}}^t)\|\cdot\Big\{\big\|\mathbb{E}[\mathbf{G}_{\mathcal{A}}(\mathbf{w}^t;\mathcal{D}_k)|\mathbf{w}^t,\mathbf{u}^t] - \nabla F_k(\mathbf{w}^t)\big\|\end{aligned}$$

$$+\|\nabla F_k(\mathbf{w}^t)-\nabla F(\mathbf{w}^t)\|+\|\nabla F(\mathbf{w}^t)-\nabla F(\hat{\mathbf{w}}^t)\|\Big\} \tag{232}$$

$$\leq D\cdot(A_1+B+L\|\mathbf{w}^t-\hat{\mathbf{w}}^t\|) \tag{233}$$

$$=A_1D+BD+DL\|\mathbf{u}^t\|. \tag{234}$$

注意，$\mathbb{E}[XY]\leq\sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}$。根据假设5、假设 7 和（221），我们得到：

$$-\mathbb{E}[\nabla F(\hat{\mathbf{w}}^t)^T\mathbf{e}^t|\mathbf{w}^t,\mathbf{u}^t]\leq\mathbb{E}[\|\nabla F(\hat{\mathbf{w}}^t)\|\cdot\|\mathbf{e}^t\||\mathbf{w}^t,\mathbf{u}^t] \tag{235}$$

$$\leq\sqrt{\mathbb{E}[\|\nabla F(\hat{\mathbf{w}}^t)\|^2|\mathbf{w}^t,\mathbf{u}^t]\cdot\mathbb{E}[\|\mathbf{e}^t\|^2|\mathbf{w}^t,\mathbf{u}^t]} \tag{236}$$

$$\leq\sqrt{D^2\cdot 8c\delta(4H^2+1)(A_2)^2(\eta_{\mathcal{A}})^2} \tag{237}$$

$$=\eta_{\mathcal{A}}\cdot\sqrt{8c\delta(4H^2+1)}A_2D. \tag{238}$$

根据假设 7,

$$\mathbb{E}[\|\mathbf{G}_{\mathcal{A}}(\mathbf{w}^t;\mathcal{D}_k)\|^2|\mathbf{w}^t,\mathbf{u}^t]\leq(A_2)^2. \tag{239}$$

将 (152)、(221)、(234)、(238) 和 (239) 代入 (229)，可得：

$$\begin{aligned}\mathbb{E}[F(\hat{\mathbf{w}}^{t+1})|\mathbf{w}^t,\mathbf{u}^t]\leq{}&F(\hat{\mathbf{w}}^t)-\frac{\eta_{\mathcal{A}}}{2}\|\nabla F(\mathbf{w}^t)\|^2+\eta_{\mathcal{A}}L^2\|\mathbf{u}^t\|^2\\&+\eta_{\mathcal{A}}\Big[A_1D+BD+DL\|\mathbf{u}^t\|\Big]+\eta_{\mathcal{A}}\cdot\sqrt{8c\delta(4H^2+1)}A_2D\\&+(\eta_{\mathcal{A}})^2L(A_2)^2+L\cdot[8c\delta(4H^2+1)(A_2)^2\cdot(\eta_{\mathcal{A}})^2].\end{aligned} \tag{240}$$

注意，$\mathbb{E}\|\mathbf{u}^t\|=\sqrt{[\mathbb{E}\|\mathbf{u}^t\|]^2}\leq\sqrt{[\mathbb{E}\|\mathbf{u}^t\|^2]}$，并且 $\mathbb{E}\|\mathbf{u}^t\|^2=4H^2(A_2)^2(\eta_{\mathcal{A}})^2$。对两边取全期望，可得：

$$\begin{aligned}\mathbb{E}[F(\hat{\mathbf{w}}^{t+1})]\leq{}&\mathbb{E}[F(\hat{\mathbf{w}}^t)]-\frac{\eta_{\mathcal{A}}}{2}\mathbb{E}\|\nabla F(\mathbf{w}^t)\|^2+\eta_{\mathcal{A}}L^2[4H^2(A_2)^2(\eta_{\mathcal{A}})^2]\\&+\eta_{\mathcal{A}}\Big[(A_1D+BD+2HA_2DL\eta_{\mathcal{A}})+\sqrt{8c\delta(4H^2+1)}A_2D\Big]\\&+(\eta_{\mathcal{A}})^2L(A_2)^2+L\cdot[8c\delta(4H^2+1)(A_2)^2\cdot(\eta_{\mathcal{A}})^2].\end{aligned} \tag{241}$$

即,

$$\begin{aligned}\mathbb{E}[F(\hat{\mathbf{w}}^{t+1})]\leq{}&\mathbb{E}[F(\hat{\mathbf{w}}^t)]-\frac{\eta_{\mathcal{A}}}{2}\mathbb{E}\|\nabla F(\mathbf{w}^t)\|^2+\eta_{\mathcal{A}}[A_1D+BD+\sqrt{8c\delta(4H^2+1)}A_2D]\\&+(\eta_{\mathcal{A}})^3[4H^2(A_2)^2L^2]\ +(\eta_{\mathcal{A}})^2[(A_2)^2L+2HA_2DL+8c\delta(4H^2+1)(A_2)^2L].\end{aligned} \tag{242}$$

对 $t=0$ 到 $T-1$ 求和，我们得到：

$$\begin{aligned}\mathbb{E}[F(\hat{\mathbf{w}}^T)]\leq{}&\mathbb{E}[F(\hat{\mathbf{w}}^0)]-\frac{\eta_{\mathcal{A}}}{2}\cdot\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(\mathbf{w}^t)\|^2+T\eta_{\mathcal{A}}[A_1D+BD+\sqrt{8c\delta(4H^2+1)}A_2D]\\&+T(\eta_{\mathcal{A}})^3[4H^2(A_2)^2L^2]+T(\eta_{\mathcal{A}})^2[(A_2)^2L+2HA_2DL+8c\delta(4H^2+1)(A_2)^2L].\end{aligned} \tag{243}$$

注意，$\hat{\mathbf{w}}^0 = \mathbf{w}^0$和 $F(\hat{\mathbf{w}}^T) \geq F^*$。因此，

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(\mathbf{w}^t)\|^2 \leq \frac{2[F(\hat{\mathbf{w}}^0) - F^*]}{\eta_{\mathcal{A}} T} + [2A_1 D + 2BD + 4\sqrt{2c\delta(4H^2+1)}A_2 D] \\ + (\eta_{\mathcal{A}})^2 \cdot [8H^2 (A_2)^2 L^2] + \eta_{\mathcal{A}} \cdot [2(A_2)^2 L + 4HA_2 DL + 16c\delta(4H^2+1)(A_2)^2 L]. \tag{244}$$

总而言之，

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(\mathbf{w}^t)\|^2 \leq \frac{2[F(\hat{\mathbf{w}}^0) - F^*]}{\eta_{\mathcal{A}} T} + \eta_{\mathcal{A}}\gamma_{\mathcal{A},1} + (\eta_{\mathcal{A}})^2 \gamma_{\mathcal{A},2} + \Delta_{\mathcal{A}}, \tag{245}$$

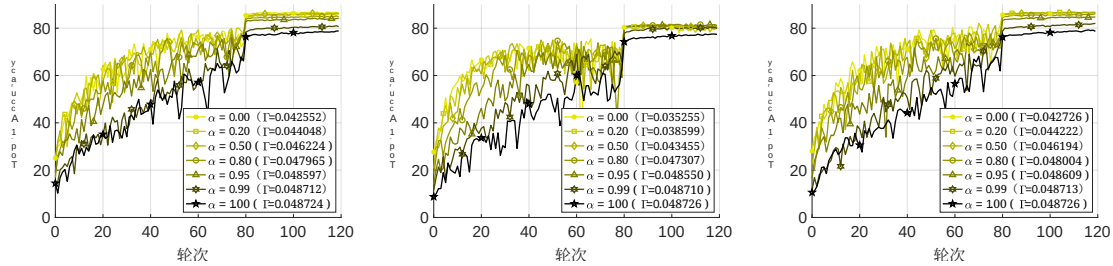
其中 $\gamma_{\mathcal{A},\ 1} = 2(A_2)^2 L + 4HA_2 DL + 16c\delta(4H^2+1)(A_2)^2 L$，$\gamma_{\mathcal{A},\ 2} = 8H^2 (A_2)^2 L^2$ 和 $\Delta_{\mathcal{A}} = 2A_1 D + 2BD + 4\sqrt{2c\delta(4H^2+1)}A_2 D$。 □

附录 C. 更多实验结果

本节将展示更多实证结果，这些结果与本文正文中的结果一致，并进一步支持我们的结论。

C.1 关于 Alpha 影响的更多实验

本节展示了 FedREP 使用 geoMed 和 TMean 聚合器时的更多实证结果。实验设置与正文中的相同。如图 3 所示，实证结果与正文中的结果一致。此外，我们还注意到，FedREP 使用 TMean 时在 ALIE 攻击下的性能不够稳定。一个可能的原因是，聚合器 TMean 对 ALIE 攻击的鲁棒性不足，因为 FedREP 使用另外两种聚合器（geoMed 和 CClip）时，实证结果都相对稳定。我们将在未来的工作中进一步研究这一现象。



(a) geoMed，比特翻转攻击　(b) geoMed，ALIE 攻击　(c) geoMed，FoE 攻击

(d) TMean，比特翻转攻击　(e) TMean，ALIE 攻击　(f) TMean，FoE 攻击

图 3：比特翻转攻击（左列）、ALIE 攻击（中列）和 FoE 攻击（右列）下，采用 geoMed（上排）和 TMean（下排）的 FedREP 随训练轮次变化的 Top-1 准确率。

C.2 坐标上的拜占庭攻击实验

在 FedREP 的每次迭代中，客户端都会向服务器发送坐标集合 $\mathcal{I}_k^t$ 。然而，拜占庭客户端可能会发送任意坐标。尽管

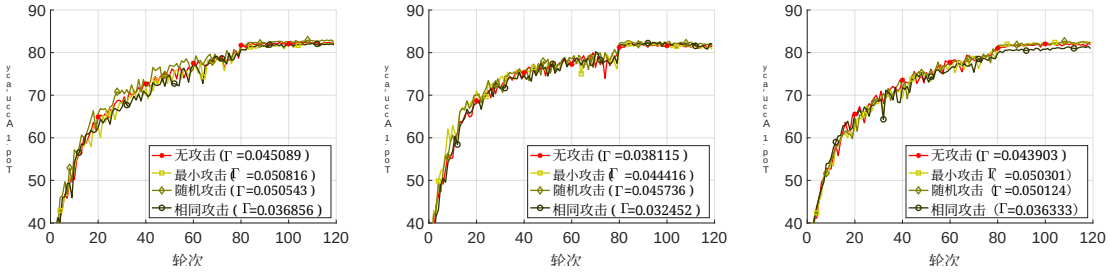


图 4：当有 7 个拜占庭客户端分别进行位翻转攻击（左）、ALIE 攻击（中）和 FoE 攻击（右）时，FedREP 在发送坐标方面采用不同拜占庭行为的 Top-1 准确率随训练轮次的变化。

正文已将这种情况纳入理论分析，我们还提供了关于发送坐标时拜占庭行为的实证结果。我们为非拜占庭客户端设置 $\alpha = 0$，并考虑四种不同的拜占庭设置：拜占庭客户端分别发送正确的坐标（无攻击）、发送 $\frac{K}{m}$ 个绝对值最小的坐标（最小攻击）、发送随机坐标（随机攻击），以及发送与某个非拜占庭客户端相同的坐标（相同攻击）。我们设置 $K = 0.065d$，其他设置与正文第 5 节中的设置相同。如图 4 所示，尽管针对坐标的拜占庭攻击会略微改变通信成本，但对收敛速度和最终 Top-1 准确率影响甚微。主要原因是，在 FedREP 中，每个非拜占庭客户端的 Top-$\frac{K}{m}$ 个坐标始终可以发送给服务器，无论拜占庭客户端发送什么内容。

C.3 局部动量实验

先前的研究（Karimireddy 等人，2021 年）表明，使用动量有助于降低随机梯度的方差，并增强拜占庭鲁棒性。本节也提供了有关动量的实证结果。实验设置与正文第 5 节相同。如图 5 所示，使用局部动量可以提高 FedREP 对 ALIE 拜占庭攻击的鲁棒性，这与先前的研究（Karimireddy 等人，2021 年）一致。

C.4 与 SparseSecAgg 的比较

我们首先在不存在攻击的情况下，通过实证比较 FedREP 与通信高效的隐私保护联邦学习基线 SparseSecAgg（Ergun 等，2021）的性能。我们分别测试缓冲区大小为 $s = 4$、8 和 16 时 FedREP 的性能。在两组实验中，我们分别为 SparseSecAgg 设置$\Gamma = 0.05$ 和 0。1。相应地，我们为 FedREP 在两组实验中分别设置 $K = 0.05d$ 和 $K = 0.1d$，因为 FedREP 的传输维度数不确定，但不会大于 K。因此，FedREP 的 $\Gamma \leq K/d$。图 6 展示了 Top-1 准确率随训练轮次的变化。结果表明，在不存在拜占庭攻击的情况下，FedREP 的性能显著优于现有的通信高效隐私保护基线 SparseSecAgg。

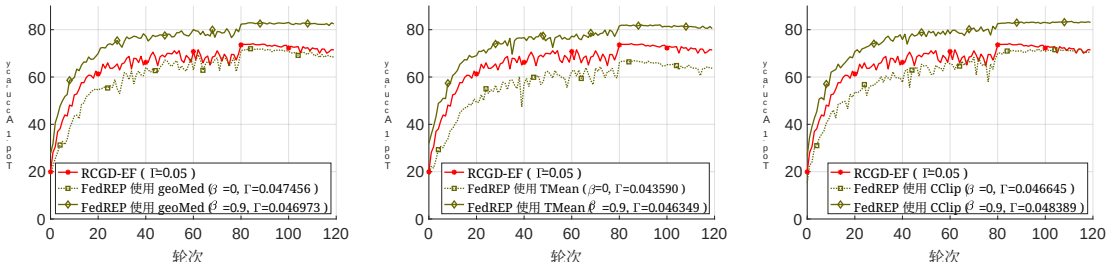


图 5：存在 7 个采用 ALIE 攻击的拜占庭客户端时，Top-1 准确率随训练轮次的变化。 β 是局部动量的超参数。当 $\beta = 0$时，不使用局部动量。FedREP 中的鲁棒聚合器分别设为 geoMed（左）、TMean（中）和 CClip（右）。

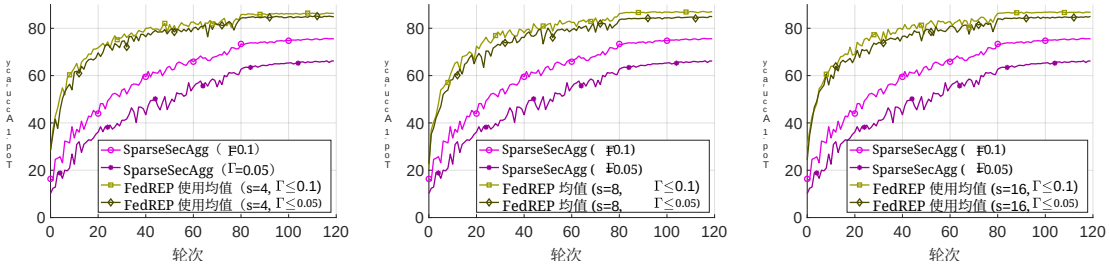


图 6：无攻击时 FedREP 和 SparseSecAgg 的 Top-1 准确率随训练轮次的变化。FedREP 的缓冲区大小 s 分别设为 4（左）、8（中）和 16（右）。

此外，我们尝试了为 SparseSecAgg 设置不同的学习率，发现学习率为 5 时其性能最佳。如图 7 所示，在通信成本相近时，FedREP 的 Top-1 准确率显著优于 SparseSecAgg。当 Top-1 准确率相当时，SparseSecAgg 的通信成本远高于 FedREP。一方面，FedREP 基于 Top-K 稀疏化，而 SparseSecAgg 基于随机-K 稀疏化。另一方面，FedREP 采用了误差补偿技术，而 SparseSecAgg 没有采用。

C.5 与 SHARE 的比较

本节通过实证比较 FedREP 和 SHARE（Velicheti 等，2021）的性能。FedREP 和 SHARE 都是联邦学习框架，可与客户端上的各种本地训练算法配合使用。与 SHARE 相比，FedREP 的主要优势在于共识稀疏化。此外，当共识稀疏化超参数 $K = md$ 时，FedREP 会退化为 SHARE。因此，我们比较了

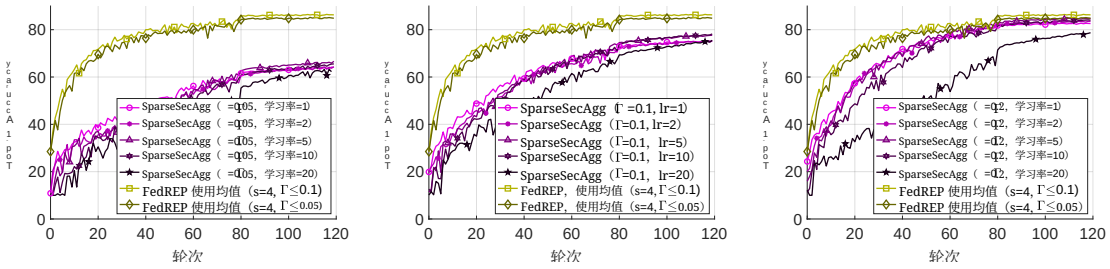


图 7：无攻击时 FedREP 和 SparseSecAgg 随训练轮次变化的 Top-1 准确率。SparseSecAgg 的 Γ 分别设为 0.05（左）、0.1（中）和 0.2（右）。FedREP 的学习率（lr）设为 0.5。

其他条件保持不变时不同的 Γ。具体而言，我们将每种方法的缓冲区大小（即 SHARE 中的集群大小）设为 4，并采用标准随机梯度下降作为本地训练算法。其他设置与本文正文第 5 节中的设置相同。由于 FedREP 每个通信轮次实际传输的维度数并不确定，我们统计所有通信轮次的平均传输维度数，并以此衡量通信成本。图 8 展示了不存在拜占庭客户端时 FedREP 和 SHARE 的实验结果。图 9、图 10 和图 11 分别展示了存在 3 个拜占庭客户端，且遭受比特翻转攻击、ALIE 攻击和 FoE 攻击时 FedREP 和 SHARE 的实验结果。从实证结果可以看出，与 SHARE 相比，当 FedREP 的 Γ 约为 0.079 时，其收敛速度和最终准确率几乎没有损失。此外，当 Γ 低至约 0.017 时，最终准确率也仅有轻微损失。

有趣的是，在遭受 ALIE 攻击时，实证结果显示，FedREP 的最终准确率甚至高于 SHARE。我们额外进行了一项实验，比较存在 7 个遭受 ALIE 攻击的拜占庭客户端时 FedREP 和 SHARE 的性能。在该实验中，我们将本地训练算法设为动量随机梯度下降，$\beta = 0.9$，缓冲区大小为 $s = 2$。其他设置均相同。如图 12 所示，在这种设置下，FedREP 仍然优于 SHARE。一个可能的原因是，FedREP 中的共识稀疏化能够降低不同客户端更新之间的差异，从而减小聚合误差（更多详情请参见正文中的定义 1）。不过，要进一步探究这一点还需要更多工作，我们将其留待未来研究。总之，FedREP 在收敛速度和最终准确率方面与 SHARE 表现相当，但通信成本远低于 SHARE。

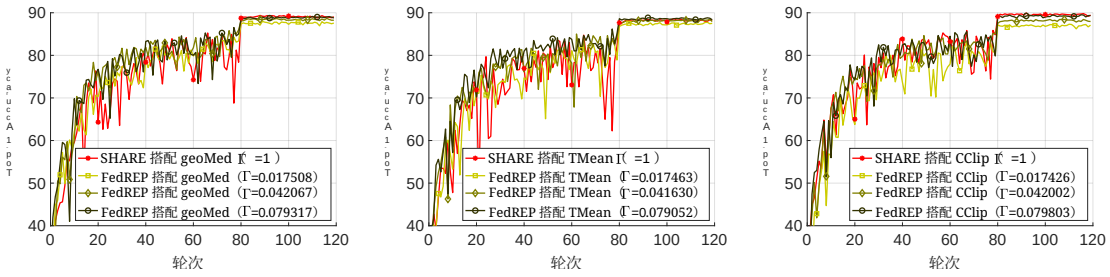


图 8：无拜占庭客户端时，FedREP 和 SHARE 的 Top-1 准确率随训练轮次的变化。

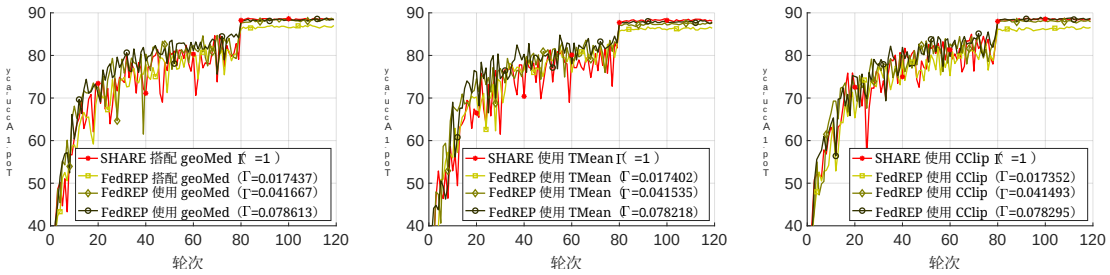


图 9：有 3 个客户端发起比特翻转攻击时，Top-1 准确率随训练轮次的变化。

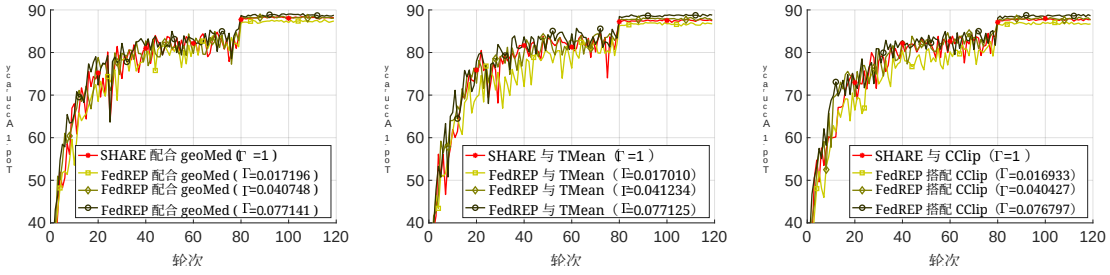


图 10：有 3 个客户端发起 ALIE 攻击时，Top-1 准确率随训练轮次的变化。

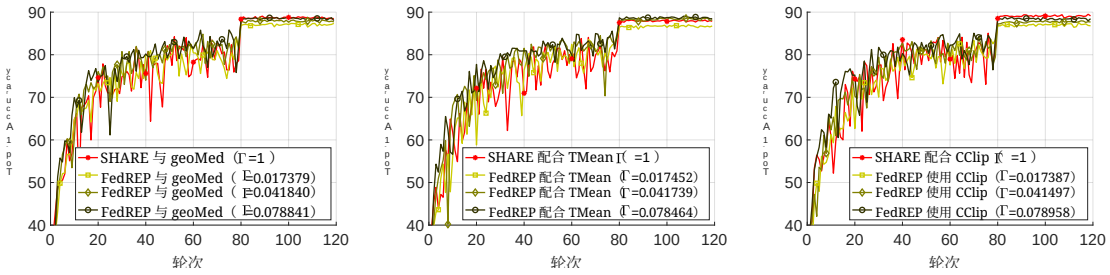


图 11：存在 3 个采用 FoE 攻击的拜占庭客户端时，Top-1 准确率随训练轮次的变化。

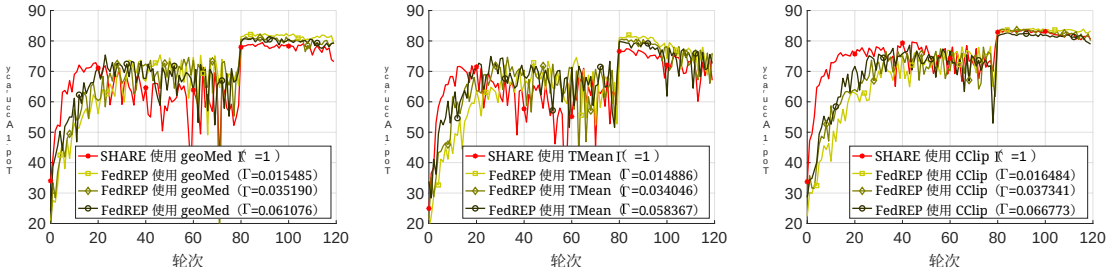


图 12：存在 7 个采用 ALIE 攻击的拜占庭客户端时，Top-1 准确率随训练轮次的变化。

参考文献

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

Alham Fikri Aji 和 Kenneth Heafield。用于分布式梯度下降的稀疏通信。载于自然语言处理实证方法会议论文集，第440–445页，2017年。

Dan Alistarh、Demjan Grubic、Jerry Li、Ryota Tomioka 和 Milan Vojnovic。QSGD：通过梯度量化和编码实现通信高效随机梯度下降。载于神经信息处理系统进展，第1709–1720页，2017年。

Dan Alistarh、Zeyuan Allen-Zhu 和 Jerry Li。拜占庭随机梯度下降。载于神经信息处理系统进展，第4613–4623页，2018年。

Zeyuan Allen-Zhu、Faeze Ebrahimian、Jerry Li 和 Dan Alistarh。拜占庭容错非凸随机梯度下降。*arXiv*预印本 *arXiv:2012.14368*，2020年。

Gilad Baruch、Moran Baruch 和 Yoav Goldberg。少即是多：规避分布式学习的防御措施。载于神经信息处理系统进展，第8635–8645页，2019年。

Debraj Basu、Deepesh Data、Can Karakus 和 Suhas N Diggavi。Qsparse-local-SGD：采用量化、稀疏化和本地计算的分布式随机梯度下降。*IEEE*信息论精选领域期刊，1(1)：217–226，2020年。

Jeremy Bernstein、Jiawei Zhao、Kamyar Azizzadenesheli 和 Anima Anandkumar。采用多数投票的signSGD具有通信效率和容错能力。载于国际学习表征会议论文集，2019年。

Peva Blanchard、Rachid Guerraoui、Julien Stainer 等。面向对抗者的机器学习：拜占庭容错梯度下降。载于神经信息处理系统进展，第119–129页，2017年。

Keith Bonawitz、Vladimir Ivanov、Ben Kreuter、Antonio Marcedone、H Brendan McMahan、Sarvar Patel、Daniel Ramage、Aaron Segal 和Karn Seth。用于隐私保护机器学习的实用安全聚合。载于*2017*年*ACM SIGSAC*计算机与通信安全会议论文集，第1175–1191页，2017年。

Lukas Burkhalter、Hidde Lycklama、AlexanderViand、Nicolas K¨uchler 和 Anwar Hith-nawi。RoFL：面向安全联邦学习的可证明鲁棒性。*arXiv*预印本 *arXiv:2107.03311*，2021年。

Chia-Yu Chen、Jiamin Ni、Songtao Lu、Xiaodong Cui、Pin-Yu Chen、Xiao Sun、Naigang Wang、Swagath Venkataramani、VijayalakshmiViji Srinivasan、Wei Zhang 等。Scalecom：面向通信高效分布式训练的可扩展稀疏化梯度压缩。载于神经信息处理系统进展，第 13551–13563 页，2020年。

Lingjiao Chen、Hongyi Wang、Zachary Charles 和 Dimitris Papailiopoulos。Draco：通过冗余梯度实现拜占庭容错分布式训练。载于第*35*届国际机器学习大会论文集，第 903–912 页，2018年。Yudong Chen、Lili Su 和 Jiaming Xu。对抗环境下的分布式统计机器学习：拜占庭梯度下降。*ACM*计算系统测量与分析汇刊，1(2):1–25，2017年。Beongjun Choi、Jy-yong Sohn、Dong-Jun Han 和 Jaekyun Moon。面向联邦学习的通信与计算高效安全聚合。*arXiv*预印本 *arXiv:2012.05433*，2020年。Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。BERT：用于语言理解的深度双向Transformer预训练。*arXiv*预印本 *arXiv:1810.04805*，2018年。El-Mahdi El-Mhamdi、Rachid Guerraoui 和 S´ebastien Rouault。面向拜占庭容错学习的分布式动量。*arXiv*预印本 *arXiv:2003.00010*，2020年。Irem Ergun、Hasin Us Sami 和 Basak Guler。面向隐私保护联邦学习的稀疏化安全聚合。*arXiv*预印本 *arXiv:2112.12872*，2021年。Fartash Faghri、Iman Tabrizian、Ilia Markov、Dan Alistarh、Daniel M Roy 和 Ali Ramezani-Kebrya。面向数据并行随机梯度下降的自适应梯度量化。载于神经信息处理系统进展，第 3174–3185 页，2020年。Venkata Gandikota、Daniel Kane、Raj Kumar Maity 和 Arya Mazumdar。VQSGD：向量量化随机梯度下降。载于国际人工智能与统计学会议，第 2197–2205 页，2021年。Avishek Ghosh、Raj Kumar Maity 和 Arya Mazumdar。分布式牛顿法能够减少通信并抵御拜占庭工作节点。载于神经信息处理系统进展，第 18028–18038 页，2020年。Avishek Ghosh、Raj Kumar Maity、Swanand Kadhe、Arya Mazumdar 和 Kannan Ramchandran。结合误差反馈的通信高效且拜占庭鲁棒的分布式学习。*IEEE*信息论精选领域期刊，2(3):942–953，2021年。Eduard Gorbunov、Dmitry Kovalev、Dmitry Makarenko 和 Peter Richt´arik。线性收敛误差补偿随机梯度下降。载于神经信息处理系统进展，第 20889–20900 页，2020年。Farzin Haddadpour、Mohammad Mahdi Kamani、Mehrdad Mahdavi 和 Viveck Cadambe。以冗余换取通信：加速面向非凸优化的分布式随机梯度下降。载于国际机器学习大会论文集，第 2545–2554 页，2019年。Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。用于图像识别的深度残差学习。载于*IEEE*计算机视觉与模式识别会议论文集，第 770–778 页，2016年。

Martin Jaggi、Virginia Smith、Martin Tak´ac、Jonathan Terhorst、Sanjay Krishnan、Thomas Hofmann 和 Michael I Jordan。通信高效的分布式对偶坐标上升。载于神经信息处理系统进展，第3068–3076页，2014年。Bargav Jayaraman、Lingxiao Wang、David Evans 和 Quanquan Gu。无忧分布式学习：隐私保护经验风险最小化。载于神经信息处理系统进展，第6346–6357页，2018年。Peter Kairouz、H Brendan McMahan、Brendan Avent、Aur´elien Bellet、Mehdi Bennis、Arjun Nitin Bhagoji、Kallista Bonawitz、Zachary Charles、Graham Cormode、Rachel Cummings 等。联邦学习的进展与开放问题。基础与趋势® 机器学习，14(1–2)：1–210，2021年。Sai Praneeth Karimireddy、Satyen Kale、Mehryar Mohri、Sashank Reddi、Sebastian Stich 和 Ananda Theertha Suresh。Scaffold：用于联邦学习的随机控制平均。载于国际机器学习大会，第5132–5143页，2020年。Sai Praneeth Karimireddy、Lie He 和 Martin Jaggi。通过历史信息学习以实现拜占庭鲁棒优化。载于国际机器学习大会论文集，第5311–5319页，2021年。Sai Praneeth Karimireddy、Lie He 和 Martin Jaggi。通过分桶在异构数据集上进行拜占庭鲁棒学习。载于国际学习表征会议论文集，2022年。Jakub Konevcn`y、H Brendan McMahan、Felix X Yu、Peter Richt´arik、Ananda Theertha Suresh 和 Dave Bacon。联邦学习：提高通信效率的策略。*arXiv:1610.05492*，2016年。Konstantinos Konstantinidis 和 Aditya Ramamoorthy。Byzshield：一种用于分布式训练的高效鲁棒系统。机器学习与系统会议论文集，3：812–828，2021年。Alex Krizhevsky、Geoffrey Hinton 等。从微型图像中学习多层特征。技术报告，2009年。Jason D Lee、Qihang Lin、Tengyu Ma 和 Tianbao Yang。通过有放回地采样额外数据实现分布式随机方差缩减梯度方法。机器学习研究期刊，18(1)：4404–4446，2017年。Liping Li、Wei Xu、Tianyi Chen、Georgios B Giannakis 和 Qing Ling。RSA：用于从异构数据集中进行分布式学习的拜占庭鲁棒随机聚合方法。载于*AAAI*人工智能大会论文集，第33卷，第1544–1551页，2019年。Xiangru Lian、Ce Zhang、Huan Zhang、Cho-Jui Hsieh、Wei Zhang 和 Ji Liu。去中心化算法能否胜过集中式算法？以去中心化并行随机梯度下降为例。载于神经信息处理系统进展，第5330–5340页，2017年。

Yujun Lin、Song Han、Huizi Mao、Yu Wang 和 Bill Dally。深度梯度压缩：降低分布式训练的通信带宽。载于国际学习表征会议论文集，2018年。Ruixuan Liu、Yang Cao、Masatoshi Yoshikawa 和 Hong Chen。Fedsel：在本地差分隐私下采用top-k维度选择的联邦随机梯度下降。载于国际高级应用数据库系统会议，第485–501页，2020年。Zaoxing Liu、Tian Li、Virginia Smith 和 Vyas Sekar。通过草图化增强联邦学习的隐私保护。*arXiv*预印本 *arXiv:1911.01812*，2019年。Brendan McMahan 和 Daniel Ramage。联邦学习：无需集中式训练数据的协作式机器学习。*Google Research Blog*，3，2017年。Brendan McMahan、Eider Moore、Daniel Ramage、Seth Hampson 和 Blaise Aguera y Arcas。利用去中心化数据进行深度网络的通信高效学习。载于人工智能与统计学，第1273–1282页，2017年。H. Brendan McMahan、Daniel Ramage、Kunal Talwar 和 Li Zhang。学习差分隐私循环语言模型。载于国际学习表征会议论文集，2018年。Krishna Pillutla、Sham M Kakade 和 Zaid Harchaoui。联邦学习的鲁棒聚合。*arXiv*预印本 *arXiv:1912.13445*，2019年。Shashank Rajput、Hongyi Wang、Zachary Charles 和 Dimitris Papailiopoulos。Detox：一种基于冗余的框架，用于实现更快速、更鲁棒的梯度聚合。载于神经信息处理系统进展，第10320–10330页，2019年。Daniel Rothchild、Ashwinee Panda、Enayat Ullah、Nikita Ivkin、Ion Stoica、Vladimir Braver-man、Joseph Gonzalez 和 Raman Arora。Fetchsgd：通过草图化实现通信高效的联邦学习。载于国际机器学习大会论文集，第8253–8265页，2020年。Mher Safaryan 和 Peter Richt´arik。随机符号下降方法：新算法与更优理论。载于国际机器学习大会论文集，第9224–9234页，2021年。Frank Seide、Hao Fu、Jasha Droppo、Gang Li 和 Dong Yu。1位随机梯度下降及其在语音深度神经网络数据并行分布式训练中的应用。载于国际语音通信协会年会，2014年。Adi Shamir。如何共享秘密。《*ACM*通讯》，22(11)：612–613，1979年。Ohad Shamir、Nati Srebro 和 Tong Zhang。使用近似牛顿型方法实现通信高效的分布式优化。载于国际机器学习大会论文集，第1000–1008页，2014年。Jy-yong Sohn、Dong-Jun Han、Beongjun Choi 和 Jaekyun Moon。用于分布式学习的选举编码：保护signSGD免受拜占庭攻击。载于神经信息处理系统进展，第14615–14625页，2020年。

Sebastian U Stich、Jean-Baptiste Cordonnier 和 Martin Jaggi。带记忆的稀疏化随机梯度下降。载于神经信息处理系统进展，第4447–4458页，2018年。

Shizhao Sun、Wei Chen、Jiang Bian、Xiaoguang Liu 和 Tie-Yan Liu。Slim-DP：一种用于通信高效的分布式深度学习的多智能体系统。载于第*17*届自主智能体与多智能体系统国际会议论文集，第721–729页，2018年。

Hanlin Tang、Chen Yu、Xiangru Lian、Tong Zhang 和 Ji Liu。DoubleSqueeze：采用双遍误差补偿压缩的并行随机梯度下降。载于国际机器学习大会论文集，第6155–6165页，2019年。

Raj Kiriti Velicheti、Derek Xia 和 Oluwasanmi Koyejo。通过聚类实现安全的拜占庭鲁棒分布式学习。*arXiv*预印本 *arXiv:2110.02940*，2021年。

Lun Wang、Qi Pang、Shuai Wang 和 Dawn Song。迈向联邦学习中的双向保护。*arXiv* 预印本 *arXiv:2010.01175*，2020年。

Jianqiao Wangni、Jialei Wang、Ji Liu 和 Tong Zhang。面向通信高效分布式优化的梯度稀疏化。载于神经信息处理系统进展，第 1299–1309 页，2018年。

Wei Wen、Cong Xu、Feng Yan、Chunpeng Wu、Yandan Wang、Yiran Chen 和 Hai Li。Terngrad：通过三元梯度减少分布式深度学习中的通信开销。载于神经信息处理系统进展，第 1509–1519 页，2017年。

Jiaxiang Wu、Weidong Huang、Junzhou Huang 和 Tong Zhang。误差补偿量化随机梯度下降及其在大规模分布式优化中的应用。载于国际机器学习大会论文集，第 5325–5333 页，2018年。

Cong Xie、Oluwasanmi Koyejo 和 Indranil Gupta。SLSGD：安全且高效的分布式端侧机器学习。欧洲机器学习与数据库知识发现联合会议，第 213–228 页，2019a。

Cong Xie、Sanmi Koyejo 和 Indranil Gupta。Zeno：具备基于怀疑的容错能力的分布式随机梯度下降。载于 国际机器学习大会论文集，第 6893–6901 页，2019b。

Cong Xie、Oluwasanmi Koyejo 和 Indranil Gupta。帝国的陨落：通过内积操纵攻破拜占庭容错随机梯度下降。载于人工智能不确定性会议论文集，第 261–270 页，2020a。

Cong Xie、Sanmi Koyejo 和 Indranil Gupta。Zeno++：鲁棒的全异步随机梯度下降。载于国际机器学习大会论文集，第 10495–10503 页，2020b。

Cong Xie、Shuai Zheng、Oluwasanmi O Koyejo、Indranil Gupta、Mu Li 和 Haibin Lin。CSER：带误差重置的通信高效随机梯度下降。发表于神经信息处理系统进展，第 12593–12603 页，2020c。

Yi-Rui Yang 和 Wu-Jun Li。BASGD：用于拜占庭学习的缓冲异步随机梯度下降。发表于国际机器学习大会论文集，第 11751–11761 页，2021年。Dong Yin、Yudong Chen、Ramchandran Kannan 和 Peter Bartlett。拜占庭鲁棒分布式学习：迈向最优统计速率。发表于国际机器学习大会论文集，第 5650–5659 页，2018年。Dong Yin、Yudong Chen、Ramchandran Kannan 和 Peter Bartlett。防御拜占庭鲁棒分布式学习中的鞍点攻击。发表于国际机器学习大会论文集，第 7074–7084 页，2019年。Hao Yu、Rong Jin 和 Sen Yang。关于用于分布式非凸优化的通信高效动量随机梯度下降的线性加速分析。发表于国际机器学习大会论文集，第 7184–7193 页，2019a。Hao Yu、Sen Yang 和 Shenghuo Zhu。并行重启随机梯度下降：更快收敛、更少通信——揭示模型平均为何适用于深度学习。发表于*AAAI*人工智能大会论文集，第 5693–5700 页，2019b。Mengjiao Zhang 和 Shusen Wang。用于安全协作机器学习的矩阵草图。发表于国际机器学习大会，第 12589–12599 页，2021年。Ruiliang Zhang 和 James Kwok。用于共识优化的异步分布式交替方向乘子法。发表于国际机器学习大会论文集，第 1701–1709 页，2014年。Shen-Yi Zhao、Ru Xiang、Ying-Hao Shi、Peng Gao 和 Wu-Jun Li。SCOPE：用于 Spark 学习的可扩展复合优化。发表于*AAAI*人工智能大会论文集，第 2928–2934 页，2017年。Shen-Yi Zhao、Gong-Duo Zhang、Ming-Wei Li 和 Wu-Jun Li。用于分布式稀疏学习的近端SCOPE。发表于神经信息处理系统进展，第 6551–6560 页，2018年。Yi Zhou、Yingbin Liang、Yaoliang Yu、Wei Dai 和 Eric P Xing。用于部分异步计算机集群的分布式近端梯度算法。机器学习研究期刊，19(1)：733–764，2018年。Ligeng Zhu、Zhijian Liu 和 Song Han。梯度深度泄露。发表于神经信息处理系统进展，第 14747–14756 页，2019年。Martin Zinkevich、Markus Weimer、Lihong Li 和 Alex J Smola。并行化随机梯度下降。发表于神经信息处理系统进展，第 2595–2603 页，2010年。