B站 丸一口

FedAWARE [1,2]

Tackling Hybrid Heterogeneity on Federated Optimization via Gradient Diversity Maximization

作者：Dun Zeng 曾遁　　电子科技大学UESTC

[1]https://arxiv.org/abs/2310.02702

[2]https://github.com/dunzeng/FedAWARE

CONTENTS

PART 01

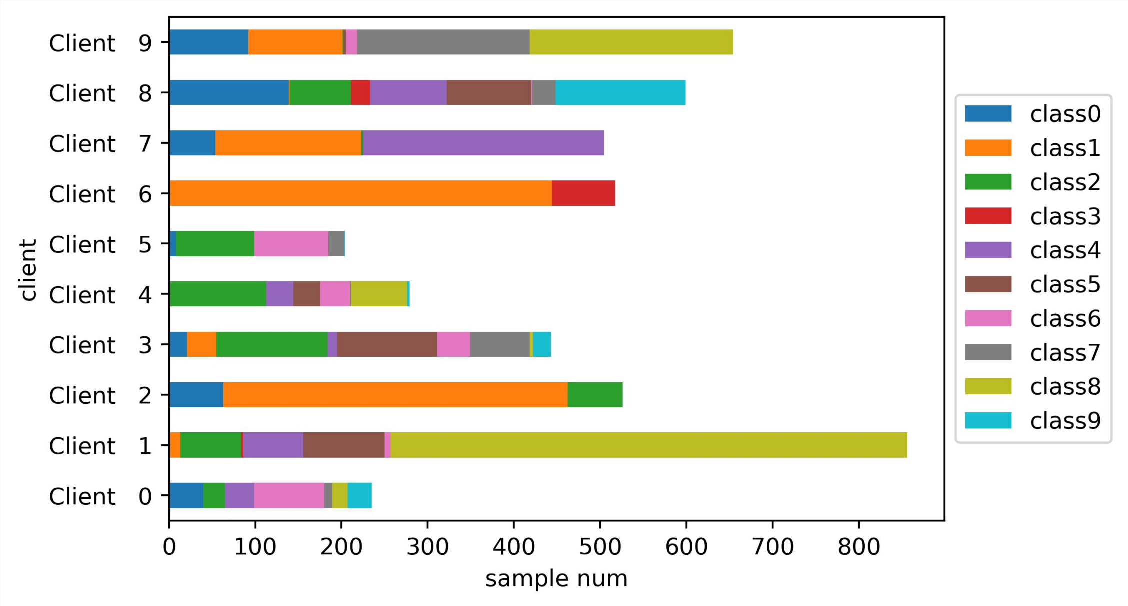
背 景 介 绍

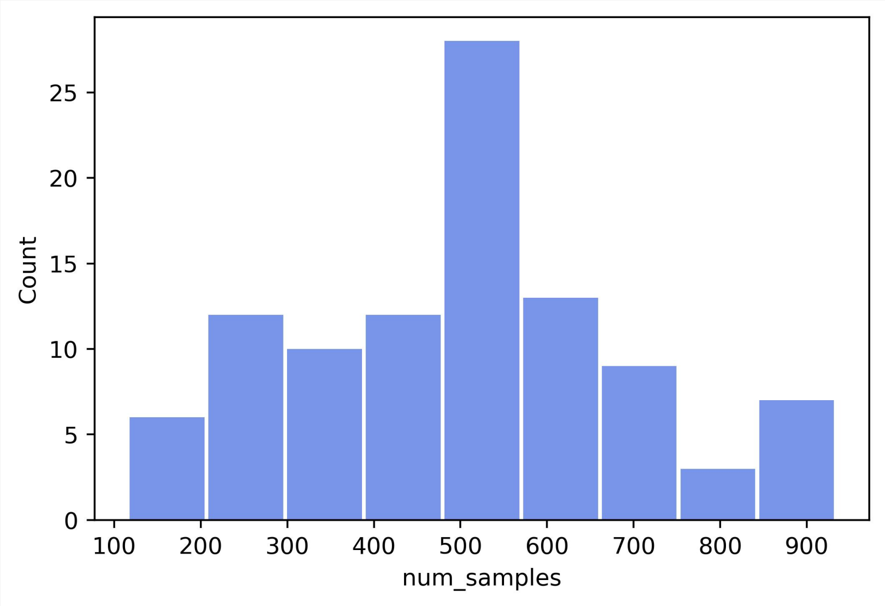
Introduction of Background

两种异构

Heterogeneities

Statistical Heterogeneity: Data Non-IID(数据非独立同分布)





[3]

System Heterogeneity:

- hardware specifications
- operating systems
- software configurations

This can result in communication overheads, computational disparities, and bias

Hybrid Heterogeneity

[3]Dirichlet partition, images source from FedLab「https://github.com/SMILELab-FL/FedLab」

使用梯度差异性来衡量统计异质性状态的程度

Tackling Hybrid Heterogeneity on Federated Optimization via Gradient Diversity Maximization

Definition 3.1 (Gradient Diversity). *We define the gradient diversity as the following ratio:*

$$D(\boldsymbol{x}) := \sqrt{\frac{\sum_{i=1}^{N} \lambda_i \|\nabla f_i(\boldsymbol{x})\|^2}{\|\nabla f(\boldsymbol{x})\|^2}} \geq 1. \quad (3)$$

Assumption 3.1 Bounded global variance

$\longrightarrow$

Assumption 3.2 Bounded gradient dissimilarity

Corollary 3.1 (Bounded gradient diversity). *Let Assumption 3.1 hold, it induces that*

$$D(\boldsymbol{x}) \leq \sqrt{1 + \frac{\sigma_g^2}{\|\nabla f(\boldsymbol{x})\|^2}},$$

which is also connected to Assumption 3.2 with $G = 0$. In this case, the B is the upper bound of gradient diversity.

Bigger Gradient Diversity ↔ Smaller Norm of Global Gradient

Assumption 3.1 (Bounded global variance). *We assume the averaged global variance is bounded, i.e., $\sum_{i=1}^{N} \lambda_i \|\nabla f_i(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\|^2 \leq \sigma_g^2$ for all $x \in \mathcal{X}$.*

Assumption 3.2 (Bounded gradient dissimilarity). *There exist constants $B \geq 1, G \geq 0$ such that $\sum_{i=1}^{N} \lambda_i \|\nabla f_i(\boldsymbol{x})\|^2 \leq B^2 \|\nabla f(\boldsymbol{x})\|^2 + G^2$ for all $x \in \mathcal{X}$.*

混合异质性效应的上限

Tackling Hybrid Heterogeneity on Federated Optimization via Gradient Diversity Maximization

Lemma 3.2 (Upper bound of balanced local drift). *Let Assumptions 2.1 2.2 3.1 hold. For any client $i \in [N]$ with balanced local iteration steps $k \in [K]$ with local learning rate $\eta_l \leq \frac{1}{K}$, the average of local drift can be bounded by:*

$$\sum_{i=1}^{N} \lambda_i \mathbb{E} \left\| \boldsymbol{x}_i^{t,k} - \boldsymbol{x}^t \right\|^2 \leq 5\eta_l(\sigma_l^2 + 6K\sigma_g^2 + 6K\mathbb{E} \left\| \nabla f\left(\boldsymbol{x}^t\right) \right\|^2). \tag{4}$$

Corollary 3.2 (Loose upper bound of unbalanced local drift). *We denote the local drift of client i from the global gradient as $\zeta_i(\boldsymbol{x}) = \|\nabla f_i(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\|^2$. Let Assumption 2.1 2.2 3.1 hold. For all client $i \in [N]$ with arbitrary local iteration steps $k \in [K_i]$ with local learning rate $\eta_l \leq \frac{1}{K_i}$, the average of local drift can be bounded as follows:*

$$\sum_{i=1}^{N} \lambda_i \mathbb{E} \left\| \boldsymbol{x}_i^{t,k} - \boldsymbol{x}^t \right\|^2 \leq \Phi_{Hetero} + 5\eta_l(\sigma_l^2 + 6K_{min}^2\sigma_g^2 + 6\tilde{K}\mathbb{E} \left\| \nabla f\left(\boldsymbol{x}^t\right) \right\|^2), \tag{5}$$

Smaller Norm of Global Gradient ↔ Lower drift

Assumption 2.1 (Smothness). *Each objective $f_i(\boldsymbol{x})$ for all $i \in [N]$ is L-smooth, inducing that for all $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, it holds $\|\nabla f_i(\boldsymbol{x}) - \nabla f_i(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|$.*

Assumption 2.2 (Unbiasedness and Bounded Local Variance). *For each $i \in [N]$ and $\boldsymbol{x} \in \mathbb{R}^d$, we assume the access to an unbiased stochastic gradient $\nabla F_i(\boldsymbol{x}, \xi_i)$ of client's true gradient $\nabla f_i(\boldsymbol{x})$, i.e.,$\mathbb{E}_{\xi_i \sim \mathcal{D}_i}\left[\nabla F_i(\boldsymbol{x}, \xi_i)\right] = \nabla f_i(\boldsymbol{x})$. The function f_i have σ_l-bounded (local) variance i.e.,$\mathbb{E}_{\xi_i \sim \mathcal{D}_i}\left[\|\nabla F_i(\boldsymbol{x}, \xi_i) - \nabla f_i(\boldsymbol{x})\|^2\right] \leq \sigma_l^2$.*

Gradient Diversity

Tackling Hybrid Heterogeneity on Federated Optimization via Gradient Diversity Maximization

Pre-Experiments

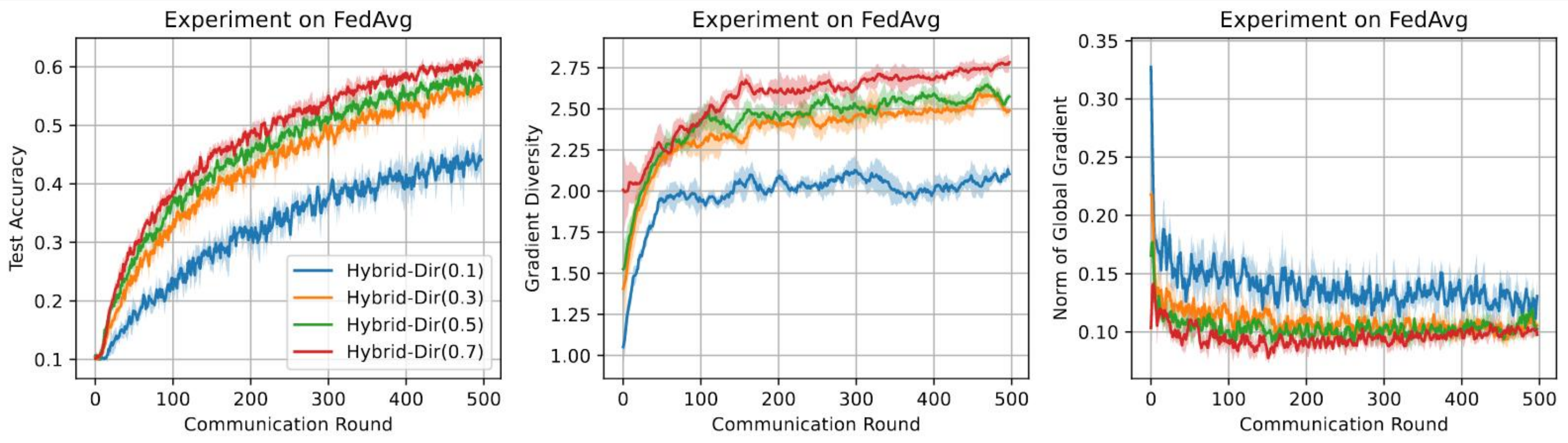


Figure 2. Observation: gradient diversity is related to convergence quality.

Bigger Acc ↔ Bigger Gradient Diversity ↔ Smaller Norm of Global Gradient

PART 02

算 法 介 绍

Introduction of Algorithm

都是梯度惹的祸

All heterogeneities can be reduce by decreasing Norm of Global Gradient

重构一组权重，通过自适应权重来减小 Norm of Global Gradient 以获取更好的ACC

formance. To this end, we use an adjustable $\tilde{\lambda}$ to study a *surrogate global objective*:

$$\tilde{f}(\boldsymbol{x}) = \sum_{i=1}^{N} \tilde{\lambda}_i f_i(\boldsymbol{x}).$$

$$\tilde{\lambda} = \min_{\lambda} \left\| \sum_{i=1}^{N} \lambda_i \nabla f_i(\boldsymbol{x}) \right\|^2, \qquad (8)$$

As the dimension of gradients can be millions, we use the **Frank-Wolfe algorithm** (Jaggi, 2013) to solve it.

Solving Equation 8 typically **requires access to all local first-order gradients from clients**, which is often infeasible in FL systems. To overcome this limitation, FEDAWARE approximates utilizes the local updates **using the history momentum** of clients

Methodology

FedAWARE

Algorithm 1 FEDAWARE

Require: $\boldsymbol{x}^0, \boldsymbol{m}^0, \alpha$

for round $t \in [T]$ **do**

  Server sample clients S^t and broadcast model $\boldsymbol{x}^t$

  for client $i \in S^t$ in parallel **do**

    $\boldsymbol{x}_i^{t,0} = \boldsymbol{x}^t$

    for local update step $\boxed{k \in [K_i]}$ **do**

      $\boldsymbol{x}_i^{t,k} = \boldsymbol{x}_i^{t,k-1} - \eta_l \nabla F_i(\boldsymbol{x}_i^{t,k-1})$

    end for

    Client uploads local updates $\boldsymbol{g}_i^t = x^{t,0} - x^{t,K_i}$

  end for

  Server updates local momentum

$$\boldsymbol{m}_i^t = \begin{cases} \alpha \boldsymbol{m}_i^{t-1} + (1-\alpha)\boldsymbol{g}_i^t, & \text{if } i \in S^t \\ \boldsymbol{m}_i^{t-1}, & \text{if } i \notin S^t \end{cases}$$

  Server computes $\tilde{\lambda}^t$ by (8) with $\boldsymbol{m}^t$

  Server computes global estimates $\boldsymbol{d}^t = \sum_{i=i}^{N} \tilde{\lambda}_i^t \boldsymbol{m}_i^t$ and updates $\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta \boldsymbol{d}^t$

end for

As the dimension of gradients can be millions, we use the Frank-Wolfe algorithm (Jaggi, 2013) to solve it.

Solving Equation 8 typically **requires access to all local first-order gradients from clients**, which is often infeasible in FL systems. To overcome this limitation, FEDAWARE approximates utilizes the local updates **using the** history momentum of clients

$$\tilde{\lambda} = \min_{\lambda} \left\| \sum_{i=1}^{N} \lambda_i \nabla f_i(\boldsymbol{x}) \right\|^2, \tag{8}$$

PART 03

写 作 赏 析

Awesome Wrighting

Wrighting

Awesome

3. Analyses on Hybrid Heterogeneity

3.1. 测量统计异质性影响

3.2.混合异质性效应的上限

3.3.减轻混合异质性影响

Corollary 3.1 (Bounded gradient diversity). *Let Assumption 3.1 hold, it induces that*

$$D(\boldsymbol{x}) \leq \sqrt{1 + \frac{\sigma_g^2}{\|\nabla f(\boldsymbol{x})\|^2}},$$

which is also connected to Assumption 3.2 with $G = 0$. In this case, the B is the upper bound of gradient diversity.

Corollary 3.2 (Loose upper bound of unbalanced local drift). *We denote the local drift of client i from the global gradient as $\zeta_i(\boldsymbol{x}) = \|\nabla f_i(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\|^2$. Let Assumption 2.1 2.2 3.1 hold. For all client $i \in [N]$ with arbitrary local iteration steps $k \in [K_i]$ with local learning rate $\eta_l \leq \frac{1}{K_i}$, the average of local drift can be bounded as follows:*

$$\sum_{i=1}^{N} \lambda_i \mathbb{E} \left\| \boldsymbol{x}_i^{t,k} - \boldsymbol{x}^t \right\|^2 \leq \Phi_{Hetero} \quad (5)$$
$$+ 5\eta_l(\sigma_l^2 + 6K_{min}^2\sigma_g^2 + 6\tilde{K}\mathbb{E}\left\|\nabla f\left(\boldsymbol{x}^t\right)\right\|^2),$$

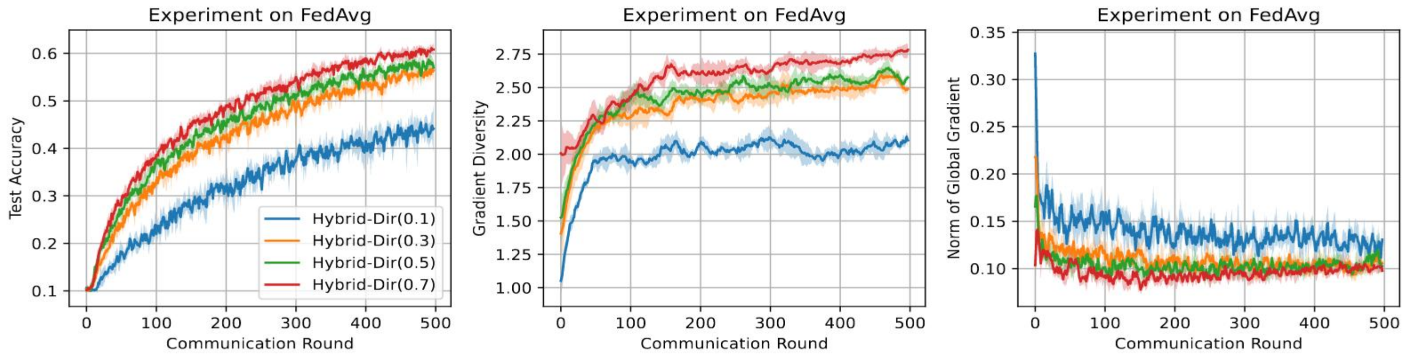


Figure 2. Observation: gradient diversity is related to convergence quality.

两节理论 + 一节实验得到优化对象

Bigger Acc ↔ Bigger Gradient Diversity ↔ Smaller $\boxed{\text{Norm of Global Gradient}}$

推论 3.2（非平衡本地漂移的松散上界）。我们标记客户端 i 与全局梯度的本地漂移为 $\zeta_i(\boldsymbol{x}) = \|\nabla f_i(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\|^2$。假设 2.1、2 .2、3 .1 成立。对于所有客户端 $i \in [N]$，其任意本地迭代步数 $k \in [K_i]$，且本地学习率 $\eta_l \le \frac{1}{K_i}$，本地漂移的平均值可以被限制如下：

$$\sum_{i=1}^{N} \lambda_i \mathbb{E}\left\|\boldsymbol{x}_i^{t,k} - \boldsymbol{x}^t\right\|^2 \le \Phi_{\text{Hetero}} + 5\eta_l\left(\sigma_l^2 + 6K_{\min}^2\sigma_g^2 + 6\tilde{K}\mathbb{E}\left\|\nabla f\left(\boldsymbol{x}^t\right)\right\|^2\right), \tag{5}$$

其中 $\tilde{K} = \sum_{i=1}^{N} \lambda_i K_i, K_{\min} = \min(K_1, \ldots, K_N)$，以及 $\Phi_{\text{Hetero}} = \sum_{i=1}^{N} 30\eta_l (K_i - K_{\min})\zeta_i\left(\boldsymbol{x}^t\right)$。

评注 3.1（推论 3.2 的解释）。本文没有假设 Φ_{Hetero} 中本地差异性 $\zeta_i\left(x^t\right)$ 的界限。因此，当本地更新步数变得不平衡时，上界 (4) 被 (5) 替代，并引入了混合异质性项 Φ_{Hetero}。这是因为至少会有一个客户端 i 使得 $K_i - K_{\min} = 0$，使得假设 3.1 不适用。此外，根据系统异质性（非平衡本地步骤），由于额外的本地步骤，Φ_{Hetero} 项被放大，使得 (5) 成为一个非常松散的界限。因此，这可能会对联邦优化的性能产生负面影响。

对混合异质性的洞察。混合异质性影响是由局部差异和不平衡局部步骤协同引起的。联邦 优化的先前工作隐含地最小化上限 (5) 以提高优化性能。例如，fedprox (li et al., 2020c) 利用惩罚项来减少推论 3.2 中的局部漂移 $\mathbb{E}\|x_i^{t,K_i} - x^t\|^2$。类似地，SCAFFLOD(karimireddy et al., 2020)、fedavgm (hsu et al., 2019) 和 feddyn (acar et al., 2020) 使用方差正则化项校正本地更新以缩小方差 σ_l 和 σ_g。fednova (wang et al.， 2020)基于局部步骤对本地更新进行剪辑，以减少$K_i, \forall i$ 的尺度效应。<u>总之，以前的工作主要是通过操纵本地更新来减轻异质性</u>。

这段写得是真好啊，把其他人的工作全部分析清楚了

理论分析出前人工作的优化方式

Wrighting

Awesome

重构一组权重，通过自适应权重来减小 Norm of Global Gradient 以获取更好的ACC

提出异于前人工作的优化方式：自适应权重

formance. To this end, we use an adjustable $\tilde{\lambda}$ to study a *surrogate global objective*:

$$\tilde{f}(\boldsymbol{x}) = \sum_{i=1}^{N} \tilde{\lambda}_i f_i(\boldsymbol{x}).$$

$$\tilde{\lambda} = \min_{\lambda} \left\| \sum_{i=1}^{N} \lambda_i \nabla f_i(\boldsymbol{x}) \right\|^2, \qquad (8)$$

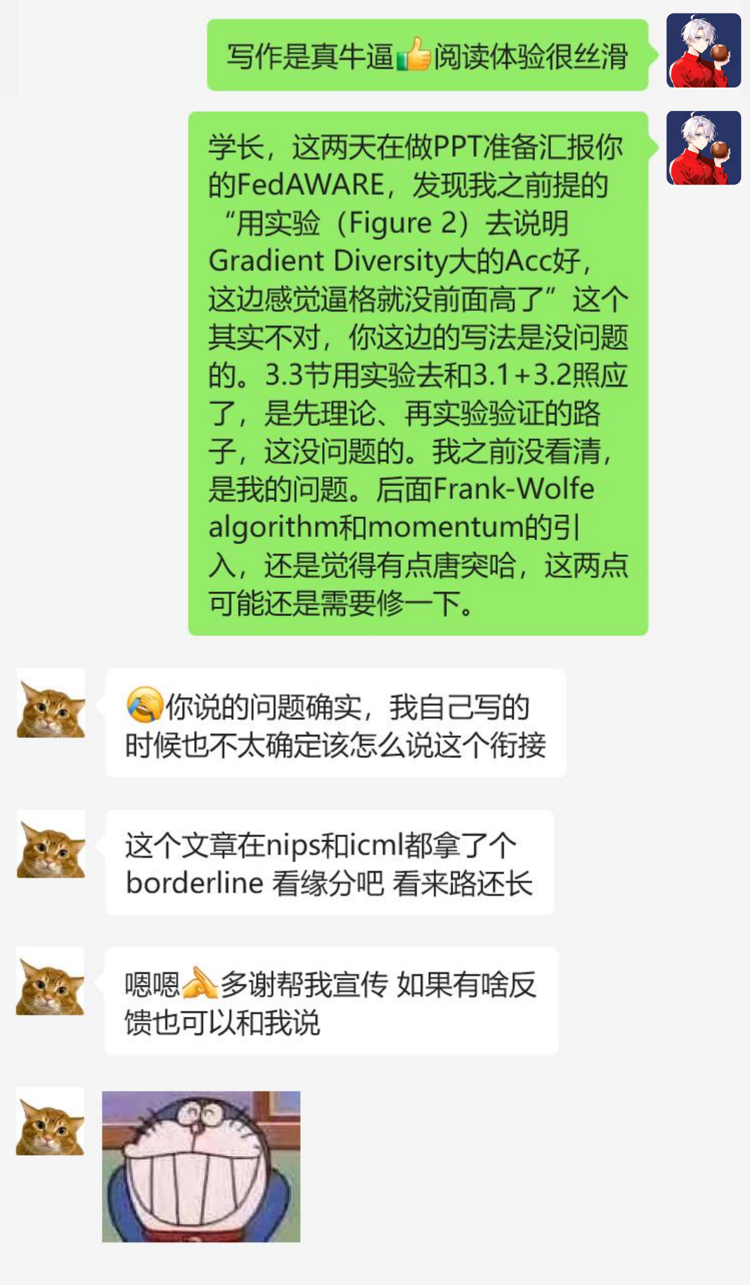
As the dimension of gradients can be millions, we use the **Frank-Wolfe algorithm** (Jaggi, 2013) to solve it.

Solving Equation 8 typically **requires access to all local first-order gradients from clients**, which is often infeasible in FL systems. To overcome this limitation, FEDAWARE approximates utilizes the local updates **using the history momentum** of clients

PART 04

交 流 评 价

Outline, Contribution, Argue Comments



胡乱拼接的聊天记录

交流评价

Communications

① The connection between theory and methodology needs to be improved.

② 写得很好，中稿只是时间问题。

B站 丸一口

感谢您的三连指导

2024/04/22