arXiv:2301.03062v1 [cs.LG] 8 Jan 2023

AnycostFL：面向异构边缘设备的高效按需联邦学习

Peichun Li*, †、Guoliang Cheng*、Xumin Huang*, †、Jiawen Kang*、Rong Yu*、Yuan Wu†和 Miao Pan‡
*广东工业大学自动化学院，中国广州
†澳门大学智慧城市物联网国家重点实验室，中国澳门
‡休斯敦大学电气与计算机工程系，美国休斯敦 电子邮箱：peichun@mail2.gdut.edu.cn,
guoliang cheng@126.com，huangxu min@163.com,{kavinkang, yurong}@gdut.edu.cn,
yuanwu@um.edu.mo, mpan2@uh.edu

摘要——本文研究了在资源约束各异的异构边缘设备上按需进行联邦学习（FL）这一具有挑战性的问题。我们提出了一种名为 AnycostFL 的成本可调联邦学习框架，使各类边缘设备能够在广泛的效率约束下高效执行本地更新。为此，我们设计了模型收缩机制，以支持计算成本弹性的本地模型训练；同时设计了梯度压缩机制，以实现通信开销动态变化的参数传输。我们采用逐元素方式进行增强型参数聚合，以提升模型性能。针对 AnycostFL，我们进一步提出一种优化设计，在个性化时延和能耗约束下最小化全局训练损失。通过揭示收敛性分析的理论洞见，我们推导出适用于不同设备的个性化训练策略，使其与本地可用资源相匹配。实验结果表明，与最先进的高效联邦学习算法相比，为达到合理的全局测试准确率，我们的学习框架可将训练时延和能耗降低至原来的 1/1.9。此外，结果还表明，我们的方法显著提升了收敛后的全局准确率。

索引词——联邦学习、边缘智能、移动计算、资源管理。

I. 引言

联邦学习（FL）是一种新兴的分布式学习范式，使多个边缘设备能够在不共享各自数据的情况下训练一个共同的全局模型 [1]。这种面向海量设备、注重隐私保护的数据分析技术被视为实现泛在智能的有前景的解决方案[2]。然而，在许多现实应用领域中，移动设备通常配备不同的本地资源，这给按需开展本地训练带来了新的挑战[3]。鉴于本地资源状态各异（例如计算能力和通信信道状态），且个性化效率约束（例如时延和能耗）各不相同，为异构边缘设备定制训练策略至关重要。

我们深入分析了边缘设备执行本地模型更新所需的时间延迟和能耗。具体而言，我们评估并记录了三种不同 NVIDIA Jetson 系列平台（即 Nano、NX AGX 和 Xavier AGX）在不同信道状态（即良好、中等和较差）下的本地训练成本。一方面，我们观察到，学习效率会因不同的学习场景而显著不同

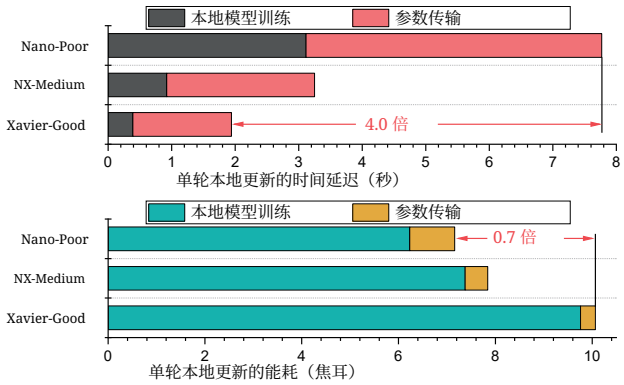


图1. 不同通信条件下，单轮本地更新在不同硬件平台上的时间延迟（上）和能耗（下）。

。如图1所示，在通信条件较差时，Nano 上单轮训练的训练时延约为通信条件良好时 Xavier AGX 的 4.0 倍，而其能耗则比后者低约 0.7 倍。另一方面，我们观察到，时延和能耗的瓶颈分别由参数传输和本地模型训练造成。

上述观察为合理设计按需联邦学习系统提供了启示。为应对资源异构性，建议降低本地设备的能耗和时延成本。更重要的是，应联合降低计算和通信开销，以实现高效的本地训练。在现有文献中，大多数研究要么采用资源分配和设备调度来降低系统成本 [4]–[10]，，要么设计梯度压缩方法以加快参数传输过程[11]–[17]。前一种方法沿袭了移动边缘系统的传统设计思路，没有考虑神经网络的优化；后一种方法则忽视了本地模型训练的计算成本。

本文提出了一种“任意成本”的联邦学习方法，称为 AnycostFL，旨在突破异构边缘设备按需分布式训练中的时延与能耗瓶颈。我们的目标是开发一种成本可调的联邦学习框架，使边缘设备能够在不同学习场景下执行本地更新。为此，我们首先设计了模型收缩和

梯度压缩，以支持具有不同计算与通信成本的自适应本地更新。同时，我们提出了一种增强型参数聚合方案，用于融合本地更新中的知识。随后，我们通过在个性化时延和能耗约束下调节本地模型结构、梯度压缩策略和计算频率，研究 AnycostFL 的按需学习。然而，针对不同学习场景定制训练策略并非易事，因为本地模型结构和压缩率如何影响全局准确率仍不明确。为解决这一问题，我们从理论上揭示了该框架的收敛规律，并进一步利用这些规律指导优化分析。最后，根据每台设备本地可用的资源，为其推导出最优训练策略。

我们的主要贡献总结如下。

- 我们提出了一种名为 AnycostFL 的新型联邦学习框架，使本地更新能够灵活调整计算成本和通信开销。
- 我们从理论上提出了 AnycostFL 的最优聚合方案，并分析了其收敛性。
- 我们研究了 AnycostFL 的按需训练问题，并设计了最优训练策略，以适应本地可用资源。
- 大量实验表明，在资源利用率和学习准确率方面，所提出的 AnycostFL 优于最先进的高效联邦学习方法。

本文其余部分安排如下。第二节介绍相关研究。第三节详细说明 AnycostFL 为实现单轮训练而执行的主要操作。第四节给出问题 ϕορм化、理论分析及相应的解决方案。第五节展示实验评估结果。最后，我们在第六节总结全文并讨论未来研究方向。

II. 相关工作

资源管理方法。资源管理方法旨在通过调配本地资源和系统资源，降低联邦学习系统成本。资源分配方法采用频率调度 [18], 发射功率控制[19], 和带宽分配 [20] 来平衡本地训练的成本。近期的设备选择方法会直接排除计算或通信能力较差的弱设备，以缩短收敛时间 [21]–[23]。此外，拓扑感知管理是改善网络吞吐量的另一种非常有效的方法 [18], [24], [25]。然而，这些方法沿用了传统移动系统的高效设计理念，忽视了神经网络的优化。

神经元感知技术。神经元感知技术致力于揭示神经网络的黑箱，以提升联邦学习系统的训练效率。早期的梯度压缩采用稀疏化 [11], [26], 和量化 [14],[27], [28] 来降低联邦学习系统的传输成本。此外，还可以通过特征图融合和知识蒸馏来提升信息聚合 [29], [30]。

此外，FedMask 提出为每台设备训练一个个性化掩码，以提升本地数据集上的测试准确率[31]。最近，模型结构剪枝使具有不同模型架构的多台设备能够训练共享全局模型 [32], [33]。这类方法可以降低本地训练的成本，但如何针对不同的学习场景定制最优训练策略（例如梯度压缩和模型剪枝策略）仍是未知的。

III. 使用 AnycostFL 进行训练

本节首先概述 AnycostFL 的整体设计，随后详细介绍我们框架中的关键技术，包括弹性模型收缩（EMS）、灵活梯度压缩（FGC）和一体化聚合（AIO）。

A. AnycostFL 概述

我们考虑一个由一组 I 边缘设备 $\mathcal{I}=\{1,2,\cdots,I\}$ 组成的通用联邦学习应用场景。我们用 $\mathcal{D}_i$ 表示设备 i 的本地训练数据，$\mathcal{D}=\cup_{i=1}^{I}\mathcal{D}_i$ 表示全局数据。令 $F_i(\boldsymbol{w})=\ell(\boldsymbol{w},\mathcal{D}_i)$ 表示设备 i 关于模型权重 $\boldsymbol{w}$ 的本地训练损失，其中 $\ell(\cdot,\cdot)$ 是预先设定的损失函数。联邦学习系统的目标是最小化以下全局损失函数

$$F(\boldsymbol{w}) \triangleq \sum_{i=1}^{I} \frac{|\mathcal{D}_i|}{|\mathcal{D}|} F_i(\boldsymbol{w}), \tag{1}$$

其中，$|\mathcal{D}_i|$ 是 $\mathcal{D}_i$ 的大小。对于给定的学习任务，单个样本的原始训练工作量 W 和未压缩梯度的数据大小 S 可以通过经验测得。

如图 2(a) 所示，为降低本地模型训练的计算复杂度和梯度更新传输的通信成本，我们提出了 AnycostFL，并采用两种设备端技术，即模型收缩和梯度压缩。在 AnycostFL 的第 t 次全局迭代中，设备 i 能够调整其训练工作量和梯度大小，分别表示为 $W_{t,i}=\alpha_{t,i}W$ 和 $S_{t,i}=\beta_{t,i}S$。其中，$\alpha_{t,i}\in(0,1]$ 和 $\beta_{t,i}\in(0,1]$ 分别定义为模型收缩因子和梯度压缩率。AnycostFL 的训练流程概述如下。

1）弹性本地训练：在第 t 个全局轮次中，设备 i 从参数服务器下载最新的全局模型 $\boldsymbol{w}_t$。利用预先计算的模型收缩因子 $\alpha_{t,i}$，可高效地通过收缩函数 $\text{shrink}(\boldsymbol{w}_t,\alpha_{t,i})$ 得到专用子模型 $\boldsymbol{w}_{t,i}^{\alpha}=$，其中函数 $\text{shrink}(\cdot,\cdot)$ 表示模型收缩操作。随后，使用子模型 $\boldsymbol{w}_{t,i}^{\alpha}$ 和本地数据 $\mathcal{D}_i$ 进行本地训练，得到更新后的本地子模型 $\boldsymbol{w}_{t+1,i}^{\alpha}$。此外，可获得本地梯度更新 $\boldsymbol{u}_{t,i}=\boldsymbol{w}_{t,i}^{\alpha}-\boldsymbol{w}_{t+1,i}^{\alpha}$。

2) 灵活的梯度上传：为进一步减少上行流量，本地设备 i 会在参数传输前压缩梯度更新 $\boldsymbol{u}_{t,i}$。给定压缩率 $\beta_{t,i}$ 后，压缩后的梯度更新 $\tilde{\boldsymbol{u}}_{t,i}=\text{cmprs}(\boldsymbol{u}_{t,i},\beta_{t,i})$ 将上传至服务器，其中 $\text{cmprs}(\cdot,\cdot)$ 是梯度压缩函数。

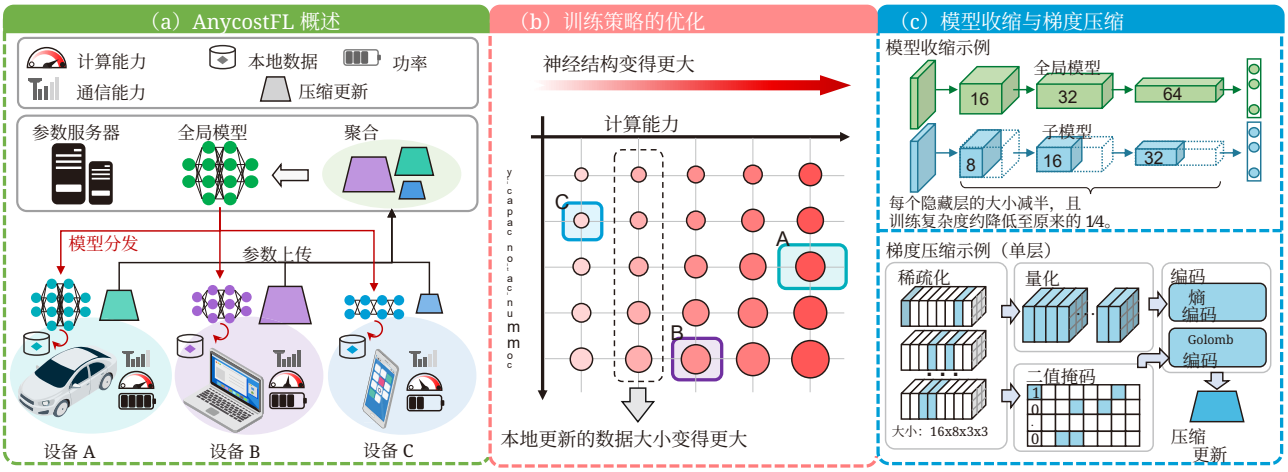


图 2。左：异构边缘设备上的 AnycostFL。中：根据设备本地可用资源，为不同设备定制神经结构和梯度压缩策略；颜色越深表示训练的计算复杂度越高，标记尺寸越大表示本地更新的数据大小越大。右：本地模型模型收缩和本地更新梯度压缩的示意图。

3) 参数聚合：服务器收集采用不同收缩因子$\{\alpha_{t,i}\}_{\forall i}$和压缩率 $\{\beta_{t,i}\}_{\forall i}$的压缩后本地更新 $\{\tilde{\boldsymbol{u}}_{t,i}\}_{\forall i}$。随后，通过 $\tilde{\boldsymbol{u}}_t = \text{aioagg}(\{\tilde{\boldsymbol{u}}_{t,i}\}_{\forall i})$ 计算全局更新，其中 aioagg(·) 是服务器端的一体化聚合。然后，计算得到更新后的全局模型 $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \tilde{\boldsymbol{u}}_t$。

经过上述三步迭代的 T 轮训练后，得到最终的全局模型 $\boldsymbol{w}_T$。在第四节介绍如何自定义 $\{\alpha_{t,i}\}_{\forall i}$ 和 $\{\beta_{t,i}\}_{\forall i}$ 的取值之前，本节剩余部分将详细说明模型收缩、梯度压缩和更新聚合。

B. 弹性模型收缩

我们的目标是通过缩小全局模型的宽度，从全局模型 $\boldsymbol{w}_t$ 中导出训练复杂度为 $\alpha_{t,i}W$ 的子模型$\boldsymbol{w}_{t,i}^{\alpha}$。收缩操作如下所示。

1）服务器端通道排序：为避免给边缘设备带来额外的内存开销，服务器会在模型分发前，先对最新的全局模型中的通道进行排序。给定全局模型某一层的权重，服务器按照 L2 范数值从大到小对当前层的输出通道进行排序，同时，下一层的输入通道也应按相同顺序相应排序，以保持整个模型 [34] 的置换不变性。

2）逐层均匀收缩：接下来，服务器按通道逐一广播全局模型各层的权重。设备无需下载完整的全局模型，只需接收全局模型中的重要参数，即可组成本地子模型。这里，我们对同一子模型中的每一层采用固定的收缩比例。根据经验，给定模型收缩因子 $\alpha_{t,i}$，我们可以将隐藏层大小缩减 $\sqrt{\alpha_{t,i}}$，以获得子模型。例如，如图 2(c) 所示，当收缩隐藏层大小为 $\{16, 32, 64\}$ 的全局模型，且采用 $\alpha_{t,i} = \frac{1}{4}$ 时，我们将每个隐藏层的大小大致缩减一半，即 $\{8, 16, 32\}$，从而构成子模型。

在第 t 个全局轮次开始时，所有设备通过从全局模型 $\boldsymbol{w}_t$ 中选择最重要的通道，来初始化其本地子模型 $\{\boldsymbol{w}_{t,i}^{\alpha}\}_{\forall i}$。这样既能显著降低训练复杂度，又能保持本地子模型的性能。随后，设备 k 使用子模型 $\boldsymbol{w}_{t,i}^{\alpha}$进行本地训练，并生成本地梯度 $\boldsymbol{u}_{t,i}$，其数据大小为 $\alpha_{t,i}S$。

C. 灵活梯度压缩

给定本地更新 $\boldsymbol{u}_{t,i}$ 以及目标压缩率 $\beta_{t,i}$，我们的目标是获得压缩更新 $\tilde{\boldsymbol{u}}_{t,i}$，其数据大小为 $\alpha_{t,i}\beta_{t,i}S$。设 $\rho_{t,i}$ 和 $L_{t,i}$ 分别表示稀疏率和量化级别数。梯度压缩方案的工作流程如下。

1) 逐卷积核稀疏化：为便于说明稀疏化过程，我们以卷积神经网络（CNN）为例。我们的目标是从 $\boldsymbol{u}_{t,i}$中获得稀疏更新 $\hat{\boldsymbol{u}}_{t,i}$。设 $\boldsymbol{u}_{t,i}[k]$ 表示 $\boldsymbol{u}_{t,i}$的第 k 个卷积核，且 $\boldsymbol{u}_{t,i} = \{\boldsymbol{u}_{t,i}[k]\}_{\forall k}$。我们衡量每个卷积核的重要性，并得到 $\mathcal{N} = \{\|\boldsymbol{u}_{t,i}[k]\|_2\}_{\forall k}$，其中 $\|\cdot\|_2$表示 L2 范数运算。接下来，选取 $\mathcal{N}$ 中的第$\lceil\rho_{t,i}K\rceil$大值作为阈值 Π，逐卷积核稀疏化可表示为

$$\hat{\boldsymbol{u}}_{t,i}[k] = \begin{cases} \boldsymbol{0} & \text{if } \|\boldsymbol{u}_{t,i}[k]\|_2 < \Pi, \\ \boldsymbol{u}_{t,i}[k] & \text{otherwise.} \end{cases} \tag{2}$$

同时，$\hat{\boldsymbol{u}}_{t,i}$ 的二值掩码记为 $\boldsymbol{m}_{t,i}$。

2) 概率量化：受 [35], [36]，中研究的启发，我们旨在利用给定的稀疏 $\hat{\boldsymbol{u}}_{t,i}$ 和量化级别 $L_{t,i}$，获得量化更新 $\tilde{\boldsymbol{u}}_{t,i}$。令 $u \in \hat{\boldsymbol{u}}_{t,i}$ 为一个标量值。首先，我们计算 $\hat{\boldsymbol{u}}_{t,i}$ 中非零元素的幅值范围，记为 $[u_{\min}, u_{\max}]$，其中 $u_{\min} = \text{分钟}\{|u|\}_{\forall u\neq 0}$，且 $u_{\max} = \max\{|u|\}_{\forall u\neq 0}$。接下来，令 $\mathcal{Q} = \{Q_l\}_{l=1}^{L_{t,i}}$ 表示量化点集合，其中 Q_l 的计算方式为

$$Q_l = \frac{l\,(u_{\max} - u_{\min})}{L_{t,i}} + u_{\min}. \tag{3}$$

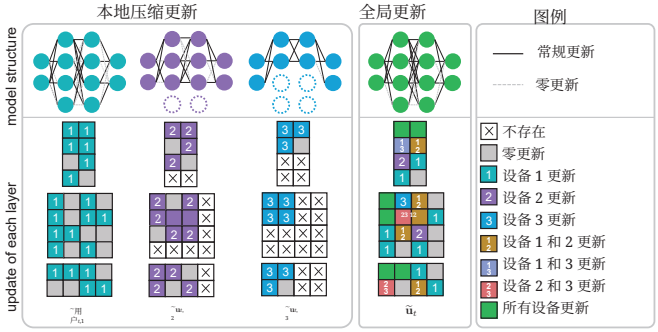


图3. 一体化聚合示意图。

对于任意 $u \in \hat{\boldsymbol{u}}_{t,i}$ 和 $u \neq 0$，我们总能找到一个量化区间 $[Q_l, Q_{l+1}]$，使得 $Q_l \leq |u| \leq Q_{l+1}$，其对应的量化值 $\tilde{u}$ 进一步按下式计算：

$$\tilde{u} = \begin{cases} \mathtt{sgn}(u) \cdot Q_l & \text{with probability } \frac{Q_{l+1}-|u|}{Q_{l+1}-Q_l}, \\ \mathtt{sgn}(u) \cdot Q_{l+1} & \text{otherwise}, \end{cases} \tag{4}$$

其中，sgn(·)用于计算给定标量的符号。此外，所有 $\tilde{u} \in \tilde{\boldsymbol{u}}_{t,i}$ 的量化索引集合记为 $\mathcal{L}_{t,i} = \{l, Q_l = \tilde{u}\}_{\forall \tilde{u} \neq 0}$。现在，$\tilde{\boldsymbol{u}}_{t,i}$ 可以表示为一个由 $\{u_{\min}, u_{\max}, L_{t,i}, \boldsymbol{m}_{t,i}, \mathcal{L}_{t,i}\}$组成的元组。

3）无损编码：由于 $L_{t,i}$ 的分布特征是较小的索引可能出现得更频繁，因此我们采用熵编码来减小数据大小 [14]，[37]。此外，稀疏二进制矩阵 $\boldsymbol{m}_{t,i}$ 可以通过 Golomb 编码 [11], [38]进行压缩。

确定压缩方案后，我们可以改变 $\{\rho_{t,i}, L_{t,i}\}$ 的组合，并记录相应的压缩率。基于这些结果，我们可以构建一个分段线性函数，以根据给定的 $\beta_{t,i}$预测压缩策略$\{\rho_{t,i}, L_{t,i}\}$。值得注意的是，服务器可以通过少量公共训练数据（例如，16 个样本）以离线方式高效地拟合该函数。

D. 一体化聚合

所有设备上传其编码后的更新后，服务器接收、解码并重构压缩的本地更新 $\{\tilde{\boldsymbol{u}}_{t,i}\}_{\forall i}$。我们的目标是通过聚合 $\{\tilde{\boldsymbol{u}}_{t,i}\}_{\forall i}$来获得全局更新 $\tilde{\boldsymbol{u}}_t$。然而，我们框架中的本地更新由具有不同精度级别（即不同量化级别和稀疏度）的不同模型结构生成，因此无法采用传统的FedAvg [1], 对本地更新进行聚合。

为应对上述挑战，我们提出一种一体化聚合方案，以逐元素的方式融合本地更新。设 $\{1, 2, \cdots, J\}$ 表示全局更新 $\tilde{\boldsymbol{u}}_t$ 的索引元素，$\tilde{\boldsymbol{u}}_t^{[j]}$ 表示 $\tilde{\boldsymbol{u}}_t$ 的第 j 个元素。为对 $\tilde{\boldsymbol{u}}_t^{[j]}$ 执行聚合，我们首先确定设备子集 $\mathcal{I}_j \subseteq \mathcal{I}$，其中这些设备的本地模型结构也包含第 j 个元素。然后，我们有

$$\tilde{\boldsymbol{u}}_t^{[j]} = \begin{cases} 0 & \text{if } \sum\limits_{i \in \mathcal{I}_j} \boldsymbol{m}_{t,i}^{[j]} = 0, \\ \frac{1}{\sum\limits_{i \in \mathcal{I}_j} p_{t,i} \boldsymbol{m}_{t,i}^{[j]}} \sum\limits_{i \in \mathcal{I}_j} p_{t,i} \boldsymbol{m}_{t,i}^{[j]} \boldsymbol{u}_{t,i}^{[j]} & \text{otherwise}, \end{cases} \tag{5}$$

其中 $p_{t,i}$ 是第 t 个全局轮次中第 j 个设备的聚合系数。$\{p_{t,i}\}_{\forall i}$ 的最优取值将在第四节中进一步分析。图3给出了一个示例，以说明聚合的具体细节。具体而言，全局更新中的不同元素由不同的设备子集进行更新，而更重要的元素会从更多设备“吸收”知识。当设备 $\mathcal{I}_j$ 将第 j 个元素全部置零时，我们有 $\tilde{\boldsymbol{u}}_t^{[j]} = 0$。

IV. 理论分析与优化

本节将通过为不同设备定制训练策略，重点研究我们框架的优化。首先，我们构建 AnycostFL 的按需训练问题。然后，推导收敛率的上界，并揭示提升 AnycostFL 性能的关键要点。在此分析基础上，将优化问题转化为一种易于处理的形式，并推导出闭式解。

A. 无线网络中的 AnycostFL

本小节将为我们的框架建立计算模型和通信模型。随后，我们构建一个按需学习问题，在给定时延和能耗约束的情况下，最小化全局训练损失。

1) 计算模型：对于设备 i 在第 t 个全局轮次中，给定模型收缩因子 $\alpha_{t,i}$ 和计算频率 $f_{t,i}$，本地模型训练的时间消耗可表示为

$$T_{t,i}^{\text{cmp}} = \frac{\tau |\mathcal{D}_i| \alpha_{t,i} W}{f_{t,i}}, \tag{6}$$

其中，τ 表示本地轮数。同时，相应的能耗可表示为

$$E_{t,i}^{\text{cmp}} = \epsilon_i f_{t,i}^2 \tau |\mathcal{D}_i| \alpha_{t,i} W, \tag{7}$$

其中，ϵ_i 是设备 i 的硬件能耗系数。

2) 通信模型：我们采用频分多址（FDMA）方案传输本地梯度更新。对于设备 i 在第 t 个全局轮次中，可实现的传输速率可估算为

$$r_{t,i} = b_i \log_2 \Big(1 + \frac{|h_{t,i}| P_{t,i}^{\text{com}}}{N_0 b_i}\Big), \tag{8}$$

其中 $P_{t,i}^{\text{com}}$ 为发射功率；b_i 为可用带宽；$|h_{t,i}|$ 表示无线信道的路径损耗；N_0为加性高斯白噪声的功率谱密度。对于第 t 轮全局轮次中的设备 i，给定局部模型生成的更新 $\tilde{\boldsymbol{u}}_{t,i}$，其收缩因子为 $\alpha_{t,i}$，压缩率为 $\beta_{t,i}$，上行传输的所需时间 $T_{t,i}^{\text{com}}$，和能耗 $E_{t,i}^{\text{com}}$，可分别表示为

$$T_{t,i}^{\text{com}} = \frac{\alpha_{t,i} \beta_{t,i} S}{r_{t,i}}, \text{ and } E_{t,i}^{\text{com}} = T_{t,i}^{\text{com}} P_{t,i}^{\text{com}}. \tag{9}$$

基于上述计算和通信模型，接下来我们将重点讨论 AnycostFL 的优化问题。

3）问题表述：为了优化 AnycostFL，我们研究一个按需训练问题。具体而言，每轮的共享最大时延 $T^{\max}$ 由服务器确定。每轮的本地能耗预算 $E_{t\,i}^{\max}$，由设备自行设定。给定具有不同本地资源（例如计算、通信和数据）的多个设备，我们的目标是为每台设备定制训练策略，在满足个性化约束（例如时延和能耗）的前提下，使全局训练损失最小化。总之，在第 t 个全局轮次，我们旨在优化以下问题。

$$\text{(P1)} \quad \min F\big(\boldsymbol{w}_t;\{\alpha_{t,i}\}_{\forall i},\{\beta_{t,i}\}_{\forall i}\big) \tag{10}$$

$$\text{subject to:} \quad T_{t,i}^{\text{cmp}}+T_{t,i}^{\text{com}}\le T^{\max},\forall i, \tag{10a}$$

$$E_{t,i}^{\text{cmp}}+E_{t,i}^{\text{com}}\le E_{t,i}^{\max},\forall i, \tag{10b}$$

$$\alpha^{\min}\le\alpha_{t,i}\le 1,\forall i, \tag{10c}$$

$$0\le\beta_{t,i}\le\beta^{\max},\forall i, \tag{10d}$$

$$f_i^{\min}\le f_{t,i}\le f_i^{\max},\forall i, \tag{10e}$$

$$\text{变量:} \quad \{\alpha_{t,i},\beta_{t,i},f_{t,i}\}_{\forall i},$$

其中，$F\big(\boldsymbol{w}_t;\ \{\alpha_{t,i}\}_{\forall i},\{\beta_{t,i}\}_{\forall i}\big)$ 表示在训练策略 $\{\alpha_{t,i}\}_{\forall i}$ 和 $\{\beta_{t,i}\}_{\forall i}$下，给定全局模型权重 $\boldsymbol{w}_t$ 时第t轮的全局损失。在本节其余部分，我们将分析训练损失与训练策略之间的关系。随后，我们将基于理论见解进一步求解问题（P1）。

B. 假设与关键引理

与 [5], [39], 中的研究一致，我们对局部损失函数 $F_i,\forall i$ 作出如下假设。

假设1. F_i 是 λ-利普希茨连续的：$\|F_i(\boldsymbol{w})-F_i(\boldsymbol{w}')\|\le\lambda\|\boldsymbol{w}-\boldsymbol{w}'\|$，其中 $\lambda>0$。

假设2. F_i 是 ν-强凸的：$F_i(\boldsymbol{w})\ge F_i(\boldsymbol{w}')+(\boldsymbol{w}-\boldsymbol{w}')^\top\nabla F_i(\boldsymbol{w}')+\frac{\nu}{2}\|\boldsymbol{w}-\boldsymbol{w}'\|^2$。

假设3。 F_i 是二阶连续可微的。根据假设1和假设2，我们有 $\nu\mathrm{I}\preceq\nabla^2F_i(\boldsymbol{w})\preceq\lambda\mathrm{I}$。

假设4。 $\nabla F_i(\boldsymbol{w})$与$\nabla F(\boldsymbol{w})$的范数之比有界：$\|\nabla F_i(\boldsymbol{w})\|^2\le\varepsilon\|\nabla F(\boldsymbol{w})\|^2$，其中 $\varepsilon\ge$0为正数。

假设5。对于适中的收缩因子 $\alpha\ge\alpha^{\min}$，先收缩后训练可以近似为先训练后收缩：$\nabla F_i(\boldsymbol{w}^\alpha)=[\nabla F_i(\boldsymbol{w})]^\alpha$。这里，我们用 $[\nabla F_i(\boldsymbol{w})]^\alpha$ 表示对 $\nabla F_i(\boldsymbol{w})$执行的收缩操作。N

接下来，我们得 ive the followi到两个 definitions.

定义1（局部梯度散度）。局部梯度散度$\delta_{t,\ i}$ 定义为$\boldsymbol{u}_{t,\ i}$ 与$\tilde{\boldsymbol{u}}_{t,\ i}$之间的差异，表示为$\delta_{t,\ i}=\|\boldsymbol{u}_{t,\ i}-\tilde{\boldsymbol{u}}_{t,\ i}\|$。

定义 2（全局梯度散度）。全局梯度散度 Δ_t 定义为$\boldsymbol{u}_t$ 与 $\tilde{\boldsymbol{u}}_t$之间的差异，其度量方式为 $\Delta_t=\|\boldsymbol{u}_t-\tilde{\boldsymbol{u}}_t\|=\Big\|\sum_{i=1}^{I}p_{t,i}\boldsymbol{u}_{t,i}-\sum_{i=1}^{I}p_{t,i}\tilde{\boldsymbol{u}}_{t,i}\Big\|$.

N ot aby、 n ed on it 和 $\boldsymbol{u}_{t,i}$ 以及 $\tilde{\boldsymbol{u}}_{t,i}$ 的维度可能不同。在进行算术运算之前，我们将 $\tilde{\boldsymbol{u}}_{t,i}$ 中缺失的元素补零。接下来，我们关注训练策略 $\{\alpha_{t,i},\beta_{t,i}\}_{\forall i}$ 如何影响 $\{\delta_{t,i}\}_{\forall i}$ 和 Δ_t。我们推导出以下两个引理。

引理1。对于采用模型收缩因子$\alpha_{t,\ i}$ 和压缩率$\beta_{t,\ i}$的本地训练，局部梯度散度的平方有界于

$$\mathbb{E}\|\delta_{t,i}\|^2\le\big(1-\alpha_{t,i}(2-\alpha_{t,i})\sqrt{\beta_{t,i}}\big)^2\mathbb{E}\|\boldsymbol{u}_{t,i}\|^2. \tag{11}$$

证明。见附录A。 □

引理 2。对于本地更新 $\{\tilde{\boldsymbol{u}}_{t,i},\forall i\}$ 及其对应的训练策略 $\{\alpha_{t,i},\beta_{t,i}\}_{\forall i}$ 和聚合系数 $\{p_{t,i}\}_{\forall i}$，全局梯度散度的平方有界于

$$\mathbb{E}\|\Delta_t\|^2\le I\varepsilon\eta^2\sum_{i=1}^{I}p_{t,i}^2\big(1-\alpha_{t,i}(2-\alpha_{t,i})\sqrt{\beta_{t,i}}\big)^2\mathbb{E}\|\nabla F(\boldsymbol{w}_t)\|^2. \tag{12}$$

证明。见附录 B。 □

C. 最优聚合方案与收敛分析

直观而言，使用更大的$\boldsymbol{u}_{t,\ i}$ 生成的本地更新 $\{\alpha_{t,i},\beta_{t,i}\}$ 可能包含更准确的信息，因此在聚合时应赋予更大的$p_{t,\ i}$。基于引理 2，我们推导出以下定理。

定理 1（最优聚合方案）。给定本地更新 $\{\tilde{\boldsymbol{u}}_{t,i}\}_{\forall i}$ 及其对应的训练策略$\{\alpha_{t,i},\beta_{t,i}\}_{\forall i}$，最优聚合系数为

$$p_{t,i}^*=\frac{\frac{1}{\big(1-\alpha_{t,i}(2-\alpha_{t,i})\sqrt{\beta_{t,i}}\big)^2}}{\sum_i\frac{1}{\big(1-\alpha_{t,i}(2-\alpha_{t,i})\sqrt{\beta_{t,i}}\big)^2}},\forall i. \tag{13}$$

证明。基于引理 2，我们研究以下优化问题，以最小化全局梯度散度。

$$\text{(P2)} \quad \min_{\{p_{t,i}\}_{\forall i}}\sum_{i=1}^{I}p_{t,i}^2\big(1-\alpha_{t,i}(2-\alpha_{t,i})\sqrt{\beta_{t,i}}\big)^2 \tag{14}$$

$$\text{subject to:} \quad p_{t,i}\ge 0,\forall i, \tag{14a}$$

$$\sum_{i=1}^{I}p_{t,i}=1. \tag{14b}$$

可以验证，问题（P2）是一个凸优化问题。我们进一步利用 Karush–Kuhn–Tucker（KKT）条件求解该问题。令 $\{\varpi\}_{\forall i}$ 和 θ分别为约束（14a）和（14b）的拉格朗日乘子。于是，我们得到

$$\begin{aligned}&\varpi_i\ge 0,\ \ \varpi_i p_{t,i}=0,\ \ p_{t,i}\ge 0,\ \ \sum_{i=1}^{I}p_{t,i}=1,\\&2p_{t,i}\big(1-\alpha_{t,i}(2-\alpha_{t,i})\sqrt{\beta_{t,i}}\big)^2-\varpi_i+\theta=0,\ \forall i.\end{aligned} \tag{15}$$

与 [40], 中的研究一致，我们可以得到

$$p_{t,i}=-\frac{\theta}{2\big(1-\alpha_{t,i}(2-\alpha_{t,i})\sqrt{\beta_{t,i}}\big)^2}. \tag{16}$$

将式 (16) 代入式 (14b)，可得

$$\theta=-\frac{2}{\sum_k\frac{1}{\big(1-\alpha_{t,i}(2-\alpha_{t,i})\sqrt{\beta_{t,i}}\big)^2}}. \tag{17}$$

将公式（17）代入公式（16），证明即告完成。 □

通过最优聚合方案，我们研究 AnycostFL 的收敛率上界。

定义 3（局部和全局学习增益）。局部和全局学习增益分别定义为 $g_{t,\ i}=\alpha_{t,\ i}^4\beta_{t,\ i}$ 和$g_t=\sum_i g_{t,\ i}/I$。具体而言，局部和全局学习增益（即 $g_{t,\ i}\in[0,1]$ 和 $g_t\in[0,1]$）分别衡量局部和全局更新中所携带的有效信息量。

定理2（AnycostFL 的收敛率）。设 $g^{\min}$ =分钟$\{g_t\}_{\forall t}$ 为 T-轮次训练期间的最小全局学习增益。AnycostFL 收敛率的上界满足

$$\mathbb{E}\big(F(\boldsymbol{w}_T)-F(\boldsymbol{w}^*)\big)\le Z^{T-1}\mathbb{E}\big(F(\boldsymbol{w}_0)-F(\boldsymbol{w}^*)\big),\quad(18)$$

其中 $Z=1-\frac{\nu}{\lambda}\left(1-\varepsilon(1-g^{\min})\right)$。回顾一下，参数$\nu$、$\lambda$ 和 ϵ 在前文的假设1至4中定义。

证明。见附录 C。 □

基于定义 3 和定理2，我们推导出以下命题。

命题1。最小化 AnycostFL 训练损失的关键，是最大化每个全局轮次的学习增益 g_t。如果 $g_t=1\ \forall t$，AnycostFL 就会退化为不使用模型收缩和梯度压缩的传统联邦学习。

D. 问题（P1）的解

根据定理2和命题1，问题（P1）可转化为以下问题。

$$\text{(P3)}\qquad \max\ \frac{1}{I}\sum_{t=1}^{I}\alpha_{t,i}^4\beta_{t,i}\qquad(19)$$

subj约束: Constrains (10a) to (10e),

变量: $\{\alpha_{t,i},\beta_{t,i},f_{t,i}\}_{\forall i}$.

根据训练时延和能耗的约束（10a）和（10b），我们得到以下引理。

引理3. 在确定最优训练策略$\{\alpha_{t,i}^*,\beta_{t,i}^*,f_{t,i}^*\}_{\forall i}$时，约束（10a）和（10b）始终取等号，因此 $T_{t,i}^*=T\text{max}$ 和 $E_{t,\ i}=E_{t,i}^*\ \forall i$。证明。可通过反证法证明该引理。假设存在 i_0 使得 $T_{t,i_0}^*<T\text{max}$。我们可以为设备 i_0找到一个新解 $\{\alpha_{t,i_0}',\beta_{t,i_0}^*,f_{t,i_0}'\}$，并令 $\alpha_{t,i_0}'>\alpha_{t,i_0}^*$, $f_{t,i_0}'<f_{t,i_0}^*$, 使得 $T_{t,i_0}'=T\text{max}$ 且 $E_{t,t_0}'=E_{t\ i}^{\max}$,。由于全局学习增益随着 α_{t,i_0}的增加而增加，因此有 $g_t'>g_t^*$。同理，当 $E_{t,i_0}^*<E_{t\ i_0}^{\max}$, 时也会出现矛盾，由此完成证明。 □

根据引理3，我们为每个设备采用两个中间变量（即 $\phi_{t,\ i}$ 和 $\varphi_{t,\ i}$），对问题（P3）进行重新参数化。具体而言，$\phi_{t,\ i}\in[0,1]$ 和 $\varphi_{t,\ i}\in[0,1]$ 分别是时延和能耗的分配因子，使得

$$\begin{aligned}T_{t,i}^{\text{cmp}}&=\phi_{t,i}T^{\max},\ \ T_{t,i}^{\text{com}}=(1-\phi_{t,i})T^{\max},\\ E_{t,i}^{\text{cmp}}&=\varphi_{t,i}E_{t,i}^{\max},\ \ E_{t,i}^{\text{com}}=(1-\varphi_{t,i})E_{t,i}^{\max},\ \forall i.\end{aligned}\qquad(20)$$

结合公式（6）和（20），设备 i 在第 t 轮次的局部学习增益可改写为

$$g_{t,i}(\phi_{t,i})=\kappa_{t,i}\big(E_{t,i}^{\max}-(1-\phi_{t,i})T^{\max}P_{t,i}^{\text{com}}\big)(\phi_{t,i}^2-\phi_{t,i}^3),\quad(21)$$

其中 $\kappa_{t,\ i}=\frac{r_t}{S\epsilon_i}\left(\frac{T^{\max}}{\tau|\mathcal{D}_i|W}\right)^3$。

注意，由于每个设备的决策过程相互独立，问题（P3）可转化为 I 个子问题。根据式（21），第 i个子问题可表示为关于 $\phi_{t,\ i}$ 的单变量优化问题，如下所示。

$$\begin{aligned}\text{(P4)}\qquad &\max_{\phi_{t,i}}\ g_{t,i}\big(\phi_{t,i}\big)\qquad(22)\\ \text{约束条件:}\qquad &\phi_{t,i}^{\min}\le\phi_{t,i}\le\phi_{t,i}^{\max},\end{aligned}$$

其中，$\phi_{t,\ i}$ 的下限和上限可由下式得到

$$\begin{aligned}\phi_{t,i}^{\min}&=\max\Big\{\frac{\alpha^{\min}\tau|\mathcal{D}_i|W}{f_i^{\max}T^{\max}},1-\frac{\beta^{\max}S}{r_{t,i}T^{\max}}\Big\},\\ \phi_{t,i}^{\max}&=\min\Big\{\frac{\tau|\mathcal{D}_i|W}{f_i^{\min}T^{\max}},1-\frac{\alpha^{\min}\beta^{\min}S}{r_{t,i}T^{\max}}\Big\}.\end{aligned}\qquad(23)$$

根据一阶最优性条件$\partial g_{t,\ i}/\phi_{t,\ i}=0$，可得驻点如下

$$\phi_{t,i}^{\text{s1}}=\frac{\sqrt{\psi_{t,i}}-3E_{t,i}^{\max}}{8P_{t,i}^{\text{com}}T^{\max}}+\frac{3}{4},\ \phi_{t,i}^{\text{s2}}=-\frac{\sqrt{\psi_{t,i}}+3E_{t,i}^{\max}}{8P_{t,i}^{\text{com}}T^{\max}}-\frac{3}{4},\quad(24)$$

其中 $\psi_{t,\ i}=4(P_{t\ ,\ i}^{\text{com}}\ T_{\max})^2-4E_{t\ ,\ i}^{\max}\ P_{t\ ,\ i}^{\text{com}}\ T_{\max}+9(E_{t\ i}^{\max}\ ,\ )^2$。令 $\mathcal{S}_{t,\ i}=\{\phi_{t,i}^{\min},\phi_{t,i}^{\max},\phi_{t,i}^{\text{s1}},\phi_{t,i}^{\text{s2}}\}$ 表示问题（P4）的驻点与边界点的并集。则，$\mathcal{S}_{t,\ i}'=\{\phi_{t,i}|\phi_{t,i}\in[\phi_{t,i}^{\min},\phi_{t,i}^{\max}],\phi_{t,i}\in\mathcal{S}_{t,i}\}$ 是 $\mathcal{S}_{t,\ i}$的可行解集合。问题（P4）的最优解可通过下式求得

$$\phi_{t,i}^*=\arg\max_{\phi_{t,i}\in\mathcal{S}_{t,i}'}\ g_{t,i}(\phi_{t,i}).\qquad(25)$$

此外，将$\phi_{t,\ i}^*$ 代入以下方程，即可得到设备 i 在第t个全局轮次的最优解。

$$\begin{aligned}\varphi_{t,i}^*&=1-\frac{(1-\phi_{t,i}^*)T^{\max}P_{t,i}^{\text{com}}}{E_{t,i}^{\max}},\alpha_{t,i}^*=\sqrt[3]{\frac{(\phi_{t,i}^*T^{\max})^2\varphi_{t,i}^*E_{t,i}^{\max}}{\epsilon_i(\tau|\mathcal{D}_i|W)^3}},\\ \beta_{t,i}^*&=\frac{r_{t,i}(1-\phi_{t,i}^*)T^{\max}}{\alpha_{t,i}^*S},f_{t,i}^*=\frac{\alpha_{t,i}^*\tau|\mathcal{D}_i|W}{\phi_{t,i}^*T^{\max}}.\end{aligned}\qquad(26)$$

值得注意的是，每个设备的决策过程不涉及其他设备资源状态的辅助信息。在每个全局轮次开始时，每个设备都可以在本地确定其训练策略。

V. 实验评估

A. 实验设置

1) 联邦学习训练设置：我们考虑在 Fashion-MNIST 和 CIFAR-10 数据集上进行图像分类的联邦学习应用 [41], [42]。对于 Fashion-MNIST，我们使用一个小型卷积神经网络（CNN），模型更新的数据大小为 53.22Mb [1]。对于 CIFAR-10 数据集，我们采用 VGG-9，模型更新的数据大小为 111.7Mb [43]。对于独立同分布和非独立同分布数据设置，我们遵循 [34] 中的数据集划分策略。超参数方面，Fashion-MNIST 的学习率、批次大小和本地训练轮数分别设为 {0.01、32、 1}；CIFAR-10 数据集对应的参数分别设为 {0.08、64、 1}。最大时延设为 $T^{\max}=10$ 秒

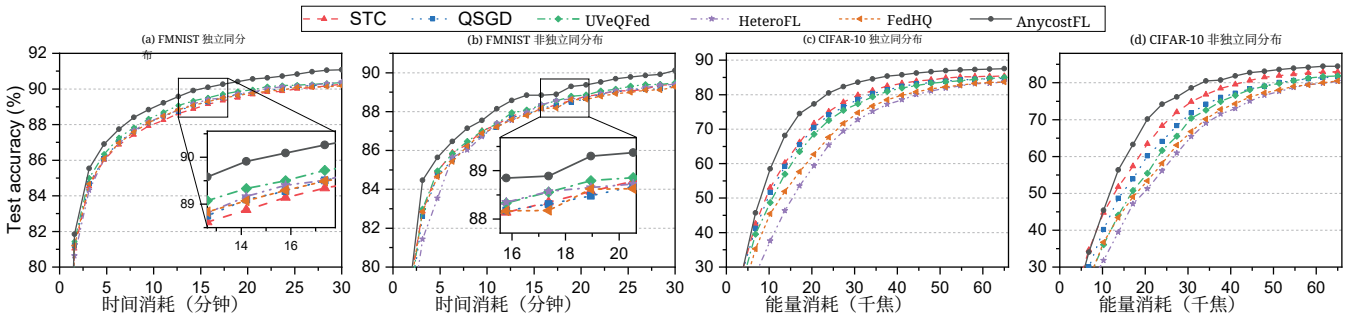


图4. 不同网络架构和数据集上的性能。（（a-b）：使用 Fashion MNIST 和 2 层卷积神经网络时，全局准确率与时间消耗的关系；（c-d）：使用 CIFAR-10 和 VGG-9 时，全局准确率与能耗的关系。）

表I ANYCOSTFL 与其他方法在 Fashion-MNIST 和 CIFAR-10 数据集上的性能比较。

数据集	方法	IID 轮数	IID 能耗 (KJ)	IID 时延 (分钟)	IID 计算 (万亿次浮点运算)	IID 通信量 (GB)	IID 最佳准确率 (%)	非独立同分布 轮数	非独立同分布 能耗 (KJ)	非独立同分布 时延 (分钟)	非独立同分布 计算 (万亿次浮点运算)	非独立同分布 通信量 (GB)	非独立同分布 最佳准确率 (%)
FMNIST {90%, 89%}*	STC	305 (1.7×)	10.94 (1.4×)	25.42 (1.7×)	152.71	0.71	90.28±0.18	283 (1.3×)	10.17 (1.1×)	23.56 (1.3×)	141.53	0.66	89.47±0.16
	QSGD	283 (1.6×)	11.40 (1.4×)	23.56 (1.6×)	141.53	0.80	90.39±0.04	279 (1.3×)	11.27 (1.2×)	23.28 (1.3×)	139.86	0.79	89.49±0.07
	UVeQFed	247 (1.4×)	11.36 (1.4×)	20.58 (1.4×)	123.67	0.72	90.44±0.10	266 (1.2×)	12.21 (1.3×)	22.14 (1.2×)	133.01	0.77	89.64±0.16
	HeteroFL	233 (1.3×)	12.03 (1.5×)	21.78 (1.5×)	92.21	0.57	90.43±0.13	242 (1.1×)	12.51 (1.3×)	22.62 (1.3×)	95.77	0.59	89.42±0.10
	FedHQ	288 (1.6×)	13.89 (1.7×)	24.03 (1.6×)	144.36	0.86	90.21±0.07	313 (1.5×)	14.96 (1.6×)	26.06 (1.5×)	156.55	0.93	89.27±0.19
	AnycostFL	179 (1.0×)	8.07(1.0×)	14.94 (1.0×)	**67.49**	**0.35**	91.**20**±0.09	214(1.0×)	9.63 (1.0×)	17.83 (1.0×)	**80.51**	**0.42**	90.**32**±0.14
CIFAR-10 {82%, 80%}*	STC	341 (1.2×)	35.39 (1.3×)	56.83 (1.2×)	4160.56	1.78	85.38±0.29	412（1.1×）	42.39（1.3×）	68.67（1.1×）	5026.84	2.15	83.09±0.53
	QSGD	337 (1.2×)	39.82 (1.5×)	56.17 (1.1×)	4111.76	2.14	84.83±0.54	430 (1.2×)	50.29 (1.5×)	71.61 (1.2×)	5242.39	2.73	81.94±0.13
	UVeQFed	296 (1.0×)	40.77 (1.5×)	49.28 (1.0×)	3607.45	2.12	85.09±0.16	377 (1.0×)	51.59 (1.5×)	62.89 (1.0×)	4603.87	2.71	82.30±0.28
	HeteroFL	332 (1.1×)	50.07 (1.9×)	69.14 (1.4×)	3222.26	1.65	83.75±0.55	413 (1.1×)	62.88 (1.9×)	85.78 (1.4×)	3990.49	2.05	80.68±0.45
	FedHQ	340 (1.2×)	48.95 (1.9×)	56.67 (1.2×)	4148.36	2.32	84.02±0.22	435 (1.2×)	61.99 (1.9×)	72.44 (1.2×)	5303.40	2.96	81.00±0.41
	AnycostFL	294 (1.0×)	26.43 (1.0×)	48.94 (1.0×)	**2459.92**	**1.56**	87.**72**±0.23	372 (1.0×)	33.51 (1.0×)	62.06 (1.0×)	**3118.60**	**1.98**	84.**91**±0.51

*$\{x, y\}$: x 和 y 分别表示独立同分布和非独立同分布数据设置下的目标全局模型精度。

CIFAR-10 数据集的能量预算设为 $E_t^{\max}{}_{,i} \sim U[3, 9]$ 焦耳，而 FMNIST 数据集对应的超参数默认减半。此外，我们设定 $\alpha^{\min} = 1/4$ 和 $\beta^{\max} = 1/15$。

2）移动系统设置：我们研究一个移动系统，其中 $I = 60$ 台设备位于半径为550米的圆形小区内，基站位于小区中心。为模拟移动性，每台设备的位置会在每轮开始时随机刷新 [44]。对于计算，能耗系数设为 $\epsilon_i \sim U[5 \times 10^{-27}, 1 \times 1^{-26}]$。对于通信，每台设备的带宽均设为1MHz，路径损耗指数为3.76。发射功率设为0.1W，且 N_0 设为 -114dBm/MHz。

B. 性能比较

我们使用三种不同的随机种子，将所提出的 AnycostFL与以下高效联邦学习算法进行比较。

- STC。采用稀疏三元压缩（STC）来降低上行参数传输的开销 [11]。
- QSGD。结合 TopK 稀疏化与概率量化来压缩本地梯度 [36]。
- UVeQFed。采用 TopK 稀疏化与通用向量量化来压缩本地梯度 [14]。
- HeteroFL。每台设备以不同宽度训练本地子模型，以匹配其计算容量 [32]。
- FedHQ。每台设备根据其信道状态 [40]，使用不同的量化级别压缩梯度。

图4展示了在独立同分布和非独立同分布数据设置下，全局模型的测试准确率随时间消耗和能耗的变化情况。在训练效率（即时间和能耗）相同的情况下，所提出的 AnycostFL 始终优于基线方案，能够提升全局模型的测试准确率。同时，表I列出了达到指定测试准确率所需的系统成本及可达到的最高准确率。具体而言，与 HeterFL 和 FedHQ 相比，在独立同分布设置下，AnycostFL 在 CIFAR-10 数据集上达到 82% 的测试准确率时，能耗最多可降低至原来的 1/1.9。在独立同分布设置下，与 STC 相比，AnycostFL 在 FMNIST 数据集上达到 90% 的测试准确率时，时间消耗最多可降低至原来的 1/1.7。此外，在 CIFAR-10 数据集上，我们的框架在独立同分布和非独立同分布设置下，分别将全局模型的最高准确率提升了 2.33% 和 1.82%。

C. 关键机制与超参数的影响

图 5(a) 验证了 AnycostFL 主要技术的优势。我们逐步移除弹性模型收缩（w/o EMS）、灵活梯度压缩（w/o FGC）和一体化聚合（w/o AIO），并记录在独立同分布设置下使用 CIFAR-10 数据集达到 80% 测试准确率所需的系统成本。我们发现，所提出的 EMS 和 FGC 分别能够显著降低能耗和训练时间。此外，AIO 有助于同时节省能耗和时间。

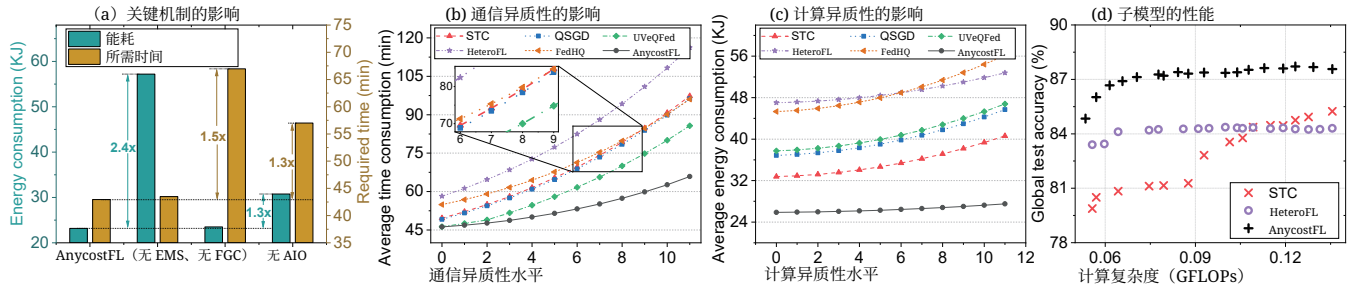


图 5. AnycostF 的主要优势　　L.（(a)：关键机制的影响；(b-c)：系统异质性的影响；(d)：子模型的性能。　　）

接下来，我们在图 5(b-c) 中评估资源异构性对训练效率的影响。我们将平均能耗系数 ϵ_i 设为 7.5×10^{-27}，并将基站与边缘设备之间的平均距离设为 400 米，然后改变它们的方差，分别模拟计算和通信异构性。方差越大，表示系统异质性越高。正如预期，面对高水平的系统异质性，所提出的 AnycostFL 比其他基线方法表现出更强的适应能力。

我们还在图 5(d) 中评估了不同宽度子模型的性能。具体而言，我们将 AnycostFL 与 HeteroFL（即采用不同宽度进行本地训练）和 STC（即性能最佳的纯压缩方法）进行了比较。这些子模型是从训练良好的全局模型中导出的，无需进一步重新训练。令人惊讶的是，由 AnycostFL 训练的全局模型的子模型仍能保持令人满意的测试准确率，从而在训练时间结束后为各种边缘设备提供动态推理。

VI. 结论

本文提出了 AnycostFL，这是一种面向联邦学习的计算与通信高效联合框架，使拥有不同资源的边缘设备能够训练共享全局模型。我们的目标是在给定的个性化时延和能耗约束下，最小化全局训练损失。通过利用 AnycostFL 的理论洞见，我们将优化问题分解为多个子问题。随后，根据每台设备本地可用的资源，为其推导出最优训练策略。实验表明，与最先进的方法相比，我们的框架在提升系统效率和模型性能方面具有优势。

致谢

Rong Yu 和 Yuan Wu 为通讯作者。本研究部分得到国家重点研发计划项目（资助编号：2020YFB1807802）、国家自然科学基金项目（资助编号：61971148、62102099、U22A2054 和 62001125）、澳门特别行政区科学技术发展基金项目（资助编号：0162/2019/A3）、FDCT-MOST联合项目（资助编号：0066/2019/AMJ）、广东省基础与应用基础研究基金（2022A1515011287）以及美国国家科学基金会项目（资助编号：CNS-2107057）的资助。

附录A 引理1的证明

证明。对于给定的收缩因子为 $\alpha_{t,i}$ 、梯度压缩率为 $\beta_{t,i}$ 的本地梯度 $\tilde{\boldsymbol{u}}_{t,i}$ ，我们旨在捕捉 $\tilde{\boldsymbol{u}}_{t,i}$ 与 $\boldsymbol{u}_{t,i}$ 之间的差异。假设 $\boldsymbol{u}_{t,i}$ 中元素的绝对值服从均匀分布 $|u| \sim U(0, u_{\max})$，且 $u_{\max} = \max\{|u|\}_{\forall u \in \boldsymbol{u}_{t,i}}$。

为清晰起见，我们按升序排列 $\boldsymbol{u}_{t,i}$ 的逐元素绝对值。$[\boldsymbol{u}_{t,i}^{[1]}, \ldots, \boldsymbol{u}_{t,i}^{[j]}, \ldots, \boldsymbol{u}_{t,i}^{[J]}]^\top$ $|\boldsymbol{u}_{t,i}^{[j]}| \le |\boldsymbol{u}_{t,i}^{[j+1]}|$ 于是，得到 $\mathbb{E}\|\boldsymbol{u}_{t,i}\|^2 = \mathbb{E}\sum_{j=1}^{J}|\boldsymbol{u}_{t,i}^{[j]}|^2 = J\mathbb{E}|\boldsymbol{u}_{t,i}^{[j]}|^2 = \frac{Ju_{\max}^2}{3}$。

$$\mathbb{E}\|\boldsymbol{u}_{t,i}\|^2 = \mathbb{E}\sum_{j=1}^{J}|\boldsymbol{u}_{t,i}^{[j]}|^2 = J\mathbb{E}|\boldsymbol{u}_{t,i}^{[j]}|^2 = \frac{Ju_{\max}^2}{3}. \tag{27}$$

根据假设5，由本地训练生成的更新 $\boldsymbol{w}_{t,i}^{\alpha}$ 等于 $\mathrm{shrink}(\boldsymbol{u}_{t,i}, \alpha_{t,i})$。对 $\boldsymbol{u}_{t,i}$ 进行模型收缩，且 $\alpha_{t,i}$ ，可视为从 $\boldsymbol{u}_{t,i}$ 中移除值最小的 $(1-\alpha_{t,i})J$ 个元素。然后，我们得到 $\mathrm{shrink}(\boldsymbol{u}_{t,i}, \alpha_{t,i}) = [0, \ldots, 0, \boldsymbol{u}_{t,i}^{[(1-\alpha_{t,i})J+1]}, \ldots, \boldsymbol{u}_{t,i}^{[J]}]^\top$。因此，我们有

$$\begin{aligned}\mathbb{E}\|\boldsymbol{u}_{t,i} - \mathrm{shrink}(\boldsymbol{u}_{t,i}, \alpha_{t,i})\|^2 &= \mathbb{E}\sum_{j=1}^{(1-\alpha_{t,i})J}|\boldsymbol{u}_{t,i}^{[j]}|^2 \\ &= J(1-\alpha_{t,i})^3 u_{\max}^2/3 = (1-\alpha_{t,i})^3\mathbb{E}\|\boldsymbol{u}_{t,i}\|^2.\end{aligned} \tag{28}$$

接下来，我们关注梯度压缩。对 $\boldsymbol{u}_{t,i}$ 进行梯度稀疏化，稀疏度为 $\rho_{t,i}$ ，可视为从 $\boldsymbol{u}_{t,i}$ 中移除值最小的 $\rho_{t,i}J$ 个元素。然后，对 $\hat{\boldsymbol{u}}_{t,i} = [0, \ldots, 0, \tilde{\boldsymbol{u}}_{t,i}^{[\rho_{t,i}J+1]}, \ldots, \tilde{\boldsymbol{u}}_{t,i}^{[J]}]^\top$ 中的非零元素进行量化，得到 $\mathrm{cmprs}(\boldsymbol{u}_{t,i}, \beta_{t,i})$。此外，我们有

$$\begin{aligned}&\mathbb{E}\|\boldsymbol{u}_{t,i} - \mathrm{cmprs}(\boldsymbol{u}_{t,i}, \beta_{t,i})\|^2 \\ &= \underbrace{\mathbb{E}\sum_{j=1}^{\rho_{t,i}J}|\boldsymbol{u}_{t,i}^{[j]}|^2}_{\text{(A)}} + \underbrace{\mathbb{E}\sum_{j=\rho_{t,i}J+1}^{J}|\boldsymbol{u}_{t,i}^{[j]} - \tilde{\boldsymbol{u}}_{t,i}^{[j]}|^2}_{\text{(B)}}.\end{aligned} \tag{29}$$

同公式（28）类似，我们有 (A)$= \rho_{t,i}^3\mathbb{E}\|\boldsymbol{u}_{t,i}\|^2$。根据公式（4）以及 $\boldsymbol{u}_t$ 的统计特性，可得 (B) $=(1-\rho_{t,i})^3\mathbb{E}\|\boldsymbol{u}_{t,i}\|^2/(2L_{t,i}^2)$。

给定采用 32 位浮点数表示的普通更新 $\boldsymbol{u}_{t,i}$ 以及期望的压缩率 $\beta_{t,i}$，我们可以在分析中设定 $\rho_{t,i} = 1-\sqrt{\beta_{t,i}}$ 和 $L_{t,i} = 2^{32}\sqrt{\beta_{t,i}}$。这样，相关操作

中的稀疏化和量化对梯度压缩的贡献相同。此外，我们有

$$\mathbb{E}\|\boldsymbol{u}_{t,i}-\mathtt{cmprs}(\boldsymbol{u}_{t,i},\beta_{t,i})\|^2\le(1-\beta_{t,i})^2\mathbb{E}\|\boldsymbol{u}_{t,i}\|^2. \tag{30}$$

接下来，我们关注相对于$\delta_{t,\ i}$ $\alpha_{t,\ i}$ 和$\beta_{t,\ i}$的局部散度。根据定义1，我们有

$$\begin{aligned}&\mathbb{E}\|\delta_{t,i}\|^2=\mathbb{E}\|\boldsymbol{u}_{t,i}-\mathtt{cmprs}([\boldsymbol{u}_{t,i}]^\alpha,\beta_{t,i})\|^2\\&=\mathbb{E}\|\boldsymbol{u}_{t,i}-[\boldsymbol{u}_{t,i}]^\alpha\|^2+\mathbb{E}\|[\boldsymbol{u}_{t,i}]^\alpha-\mathtt{cmprs}([\boldsymbol{u}_{t,i}]^\alpha,\beta_{t,i})\|^2\\&\quad+2\underbrace{<\boldsymbol{u}_{t,i}-[\boldsymbol{u}_{t,i}]^\alpha,[\boldsymbol{u}_{t,i}]^\alpha-\mathtt{cmprs}([\boldsymbol{u}_{t,i}]^\alpha,\beta_{t,i})>}_{\text{(C)}}.\end{aligned} \tag{31}$$

可以验证，项（C）中的两个向量正交，因此得到（C）$=0$。根据公式（28）和（30），我们进一步得到

$$\begin{aligned}&\mathbb{E}\|\delta_{t,i}\|^2\le(1-\alpha_{t,i})^3\mathbb{E}\|\boldsymbol{u}_{t,i}\|^2+(1-\sqrt{\beta_{t,i}})^2\mathbb{E}\|[\boldsymbol{u}_{t,i}]^\alpha\|^2\\&\overset{\text{(a)}}{\le}(1-\alpha_{t,i})^3\mathbb{E}\|\boldsymbol{u}_{t,i}\|^2\\&\quad+(1-\sqrt{\beta_{t,i}})^2\alpha_{t,i}(\alpha_{t,i}^2-3\alpha_{t,i}+3)\mathbb{E}\|\boldsymbol{u}_{t,i}\|^2\\&\overset{\text{(b)}}{\le}\big(1-\alpha_{t,i}(2-\alpha_{t,i})\sqrt{\beta_{t,i}}\big)^2\mathbb{E}\|\boldsymbol{u}_{t,i}\|^2.\end{aligned} \tag{32}$$

与公式（28）类似，不等式（a）源于以下事实：$\mathbb{E}\|[\boldsymbol{u}_{t,i}]^\alpha\|^2=\alpha_{t,\ i}(\alpha_{t,\ i}^2-3\alpha_{t,\ i}+3)\mathbb{E}\|\boldsymbol{u}_{t,\ i}\|^2$。此外，不等式（b）对所有$\alpha_{t,\ i}\in[\alpha^{\min},1]$和$\beta_{t,\ i}\in[0,\beta^{\max}]$均成立。因此，我们完成了证明。

□

APPENDIX B
PROOF OF LEMMA 2

证明。根据定义2和引理1，我们有

$$\begin{aligned}\mathbb{E}\|\Delta_t\|^2&=\mathbb{E}\Big\|\sum_{i=1}^{I}p_{t,i}\boldsymbol{u}_{t,i}-\sum_{i=1}^{I}p_{t,i}\tilde{\boldsymbol{u}}_{t,i}\Big\|^2\\&\le\mathbb{E}\Big(\sum_{i=1}^{I}p_{t,i}\big(1-\alpha_{t,i}(2-\alpha_{t,i})\sqrt{\beta_{t,i}}\big)\|\boldsymbol{u}_{t,i}\|\Big)^2.\end{aligned} \tag{33}$$

我们用 η 表示学习率，并且$\boldsymbol{u}_{t,\ i}=\eta\nabla F_i(\boldsymbol{w}_t)$。根据假设4，我们得到

$$\mathbb{E}\|\Delta_t\|^2\le\varepsilon\eta^2\Big(\sum_{i=1}^{I}p_{t,i}\big(1-\alpha_{t,i}(2-\alpha_{t,i})\sqrt{\beta_{t,i}}\big)\Big)^2\mathbb{E}\|\nabla F(\boldsymbol{w}_t)\|^2. \tag{34}$$

根据柯西–施瓦茨不等式，我们得到

$$\mathbb{E}\|\Delta_t\|^2\le I\varepsilon\eta^2\sum_{i=1}^{I}p_{t,i}^2\big(1-\alpha_{t,i}(2-\alpha_{t,i})\sqrt{\beta_{t,i}}\big)^2\mathbb{E}\|\nabla F(\boldsymbol{w}_t)\|^2. \tag{35}$$

至此，证明完毕。 □

附录 C：AnycostFL 的收敛性

证明。受 [5], [39], 中研究的启发，我们推导出 AnycostFL 的收敛分析。根据泰勒展开和假设 3，可得

$$\begin{aligned}F(\boldsymbol{w}_{t+1})&\le F(\boldsymbol{w}_t)+(\boldsymbol{w}_{t+1}-\boldsymbol{w}_t)^\top\nabla F(\boldsymbol{w}_t)+\frac{\lambda}{2}\|\boldsymbol{w}_{t+1}-\boldsymbol{w}_t\|^2\\&=F(\boldsymbol{w}_t)-\tilde{\boldsymbol{u}}_t^\top\nabla F(\boldsymbol{w}_t)+\frac{\lambda}{2}\|\tilde{\boldsymbol{u}}_t\|^2.\end{aligned} \tag{36}$$

使用学习率 $\eta=\frac{1}{\lambda}$，可得

$$\begin{aligned}\mathbb{E}\big(F(\boldsymbol{w}_{t+1})\big)&\le\mathbb{E}\Big(F(\boldsymbol{w}_t)-\lambda(\boldsymbol{u}_t-\Delta_t)^\top\boldsymbol{u}_t+\frac{\lambda}{2}\|\boldsymbol{u}_t-\Delta_t\|^2\Big)\\&=\mathbb{E}\Big(F(\boldsymbol{w}_t)-\frac{1}{2\lambda}\|\nabla F(\boldsymbol{w}_t)\|^2+\frac{\lambda}{2}\|\Delta_t\|^2\Big).\end{aligned} \tag{37}$$

现在我们关注 $\|\Delta_t\|^2$ 的上界。根据詹森不等式和公式 (34)，我们得到

$$\mathbb{E}\|\Delta_t\|^2\le\varepsilon\eta^2\underbrace{\sum_{i=1}^{I}p_{t,i}\big(1-\alpha_{t,i}(2-\alpha_{t,i})\sqrt{\beta_{t,i}}\big)^2}_{\text{(D)}}\mathbb{E}\|\nabla F(\boldsymbol{w}_t)\|^2. \tag{38}$$

将公式 (13) 代入 (A)，可得

$$\mathbb{E}\|\mathrm{D}\|\le\mathbb{E}\left\|\frac{I}{\sum_{i=1}^{I}\frac{1}{\left(1-\alpha_{t,i}(2-\alpha_{t,i})\sqrt{\beta_{t,i}}\right)^2}}\right\|\overset{\text{(c)}}{\le}\mathbb{E}\left\|\frac{I}{\sum_{i=1}^{I}\frac{1}{1-\alpha_{t,i}^4\beta_{t,i}}}\right\|, \tag{39}$$

其中，对于$\alpha_{t,\ i}\in[0,1]$ 和$\beta_{t,\ i}\in[0,1]$，(c) 始终成立。根据定义 3，我们有 g_t $_i$ $g_t=$，$_i=\alpha_{t,\ i}^4\beta_{t,\ i}$ 以及$\frac{\sum_i g_{t\,i}/I}{,}$ $1/(\sum_i\frac{1}{1-g_{t,i}})$。由于 . 是关于 $g_{t,\ i}$的凹函数，根据詹森不等式，我们得到

$$\mathbb{E}\|\mathrm{A}\|\le\frac{I}{\sum_i\frac{1}{1-\mathbb{E}(\alpha_{t,i}^4\beta_{t,i})}}=1-g_t. \tag{40}$$

由于各设备的训练策略与全局数据梯度的范数 $\|\nabla F(\boldsymbol{w}_t)\|$ 相互独立，将公式（40）代回公式（38），可得

$$\mathbb{E}\|\Delta_t\|^2\le\mathbb{E}\Big(\varepsilon\eta^2\big(1-g_t\big)\|\nabla F(\boldsymbol{w}_t)\|^2\Big). \tag{41}$$

接下来，将公式（41）代回公式（37），可得

$$\mathbb{E}\big(F(\boldsymbol{w}_{t+1})\big)\le\mathbb{E}\Big(F(\boldsymbol{w}_t)-\frac{1+\varepsilon\big(g_t-1\big)}{2\lambda}\|\nabla F(\boldsymbol{w}_t)\|^2\Big). \tag{42}$$

在公式（42）两边同时减去 $F(\boldsymbol{w}^*)$，可得

$$\begin{aligned}&\mathbb{E}\big(F(\boldsymbol{w}_{t+1}-F(\boldsymbol{w}^*)\big)\\&\le\mathbb{E}\Big(F(\boldsymbol{w}_t)-\frac{1+\varepsilon(g_t-1)}{2\lambda}\|\nabla F(\boldsymbol{w}_t)\|^2-F(\boldsymbol{w}^*)\Big).\end{aligned} \tag{43}$$

根据假设 2 和 3，我们有 [5], [45]

$$\|\nabla F(\boldsymbol{w}_t)\|^2\ge2\nu\big(F(\boldsymbol{w}_t)-F(\boldsymbol{w}^*)\big). \tag{44}$$

将公式（44）代入公式（43），可得

$$\mathbb{E}\big(F(\boldsymbol{w}_{t+1})-F(\boldsymbol{w}^*)\big)\le Z_t\mathbb{E}\big(F(\boldsymbol{w}_t)-F(\boldsymbol{w}^*)\big), \tag{45}$$

其中 $Z_t=1-\frac{\nu}{\lambda}(1-\varepsilon(1-g_t))$.

令 $g^{\min}=$ 分钟$\{g_t\}_{\forall t}$ 为 T 个全局轮次中的最小全局学习增益。通过从第 0 次迭代轮次递归应用上述不等式，直到 T，我们可以得到

$$\mathbb{E}\big(F(\boldsymbol{w}_T)-F(\boldsymbol{w}^*)\big)\le Z^{T-1}\mathbb{E}\big(F(\boldsymbol{w}_0)-F(\boldsymbol{w}^*)\big), \tag{46}$$

其中 $Z=1-\frac{\nu}{\lambda}\big(1-\varepsilon(1-g^{\min})\big)$。至此，证明完成。

□

参考文献

[1] H. B. McMahan、E. Moore、D. Ramage、S. Hampson 和 B. A. y. Arcas，“从去中心化数据中高效通信地学习深度网络”，发表于 AISTATS 2017 会议论文集，2017 年 4 月 20–22 日，第 1273–1282 页。[2] Y. Qu、C. Dong、J. Zheng、H. Dai、F. Wu、S. Guo 和 A. Anpala- gan，“通过空地一体化联邦学习赋能边缘智能”，IEEE Network，第 35 卷，第 5 期，第 34–41 页，2021 年。[3] R. Yu 和 P. Li，“迈向移动边缘计算中的资源高效联邦学习”，IEEE Network，第 35 卷，第 1 期，第 148–155 页，2021 年。[4] T. Nishio 和 R. Yonetani，“移动边缘中异构资源下的客户端选择”，发表于 ICC。IEEE，2019 年，第 1–7 页。[5] M. Chen、Z. Yang、W. Saad、C. Yin、H. V. Poor 和 S. Cui，“面向无线网络联邦学习的联合学习与通信框架”，IEEE Transactions on Wireless Communications，第 20 卷，第 1 期，第 269–283 页，2020 年。[6] Z. Yang、M. Chen、W. Saad、C. S. Hong 和 M. Shikh-Bahaei，“无线通信网络中的节能联邦学习”，IEEE Transactions on Wireless Communications，第 20 卷，第 3 期，第 1935–1949 页，2020 年。[7] C. T. Dinh 、 N. H. Tran 、 M. N. Nguyen 、C. S. Hong、W. Bao、A. Y. Zomaya 和 V. Gramoli，“无线网络中的联邦学习：收敛分析与资源分配”，IEEE/ACM Transactions on Networking，第 29 卷，第 1 期，第 398–409 页，2020 年。[8] J. Yao 和 N. Ansari，“通过 CPU频率和无线功率控制增强雾辅助物联网中的联邦学习”，IEEE Internet of Things Journal，第 8 卷，第 5 期，第 3438–3445 页，2020 年。[9] N. H. Tran、W. Bao、A. Zomaya、M. N. Nguyen 和 C. S. Hong，“无线网络中的联邦学习：优化模型设计与分析”，发表于 INFOCOM。IEEE，2019 年，第 1387–1395 页。[10] Y. Wu、Y. Song、T. Wang、L. Qian 和 T. Q. Quek，“无线能量传输辅助的非正交多址接入联邦学习：一种成本高效的方法”，IEEE Transactions on Communications，第 70 卷，第 4 期，第 2853–2869 页，2022 年。[11] F. Sattler、S. Wiedemann、K.-R. M¨uller 和 W. Samek，“基于非独立同分布数据的稳健且通信高效的联邦学习”，IEEE Transactions on Neural Networks and Learning Systems，第 31 卷，第 9 期，第 3400–3413 页，2019 年。[12] L. Li、D. Shi、R. Hou、H. Li、M. Pan 和 Z. Han，“交流还是工作：面向异构移动边缘设备节能联邦学习的灵活通信压缩”，发表于 INFOCOM。IEEE，2021 年，第 1–10 页。[13] P. Li、X. Huang、M. Pan 和 R. Yu，“Fedgreen：采用细粒度梯度压缩的绿色移动边缘计算联邦学习”，发表于 GLOBECOM，2021 年，第 1–6 页。[14] N. Shlezinger、M. Chen、Y. C. Eldar、H. V. Poor 和 S. Cui，“UVeQFed：用于联邦学习的通用向量量化”，IEEE Transactions on Signal Processing，第 69 卷，第 500–514 页，2021 年。[15] L. Cui、X. Su、Y. Zhou 和 J. Liu，“压缩通信联邦学习中的最优速率自适应”，发表于 INFOCOM。IEEE，2022 年，第 1459–1468 页。[16] J. Kang、X. Li、J. Nie、Y. Liu、M. Xu、Z. Xiong、D. Niyato 和 Q. Yan，“面向物联网人工智能的通信高效、跨链赋能联邦学习”，IEEE Transactions on Network Science and Engineering，2022 年。[17] P. Prakash、J. Ding、M. Shu、J. Wang、W. Xu 和 M. Pan，“Squafl：受草图量化启发的通信高效联邦学习”，发表于 2021 IEEE/ACM 边缘计算研讨会（SEC）。IEEE，2021 年，第 350–354 页。[18] P. Li、Y. Zhong、C. Zhang、Y. Wu 和 R. Yu，“Fedrelay：面向 6G 移动边缘智能的联邦中继学习”，IEEE Transactions on Vehicular Technology，第 1–14 页，2022 年。[19] D. Liu 和 O. Simeone，“免费的隐私：通过自适应功率控制和无编码传输实现无线联邦学习”，IEEE Journal on Selected Areas in Communications，第 39 卷，第 1 期，第 170–185 页，2020 年。[20] J. Xu 、 H. Wang 和 L. Chen，“ 无线边缘网络中多种联邦学习服务的带宽分配”，IEEE Transactions on Wireless Communications，第 21 卷，第 4 期，第 2534–2546 页，2021 年。[21] S. Wang、M. Lee、S. Hosseinalipour、R. Morabito、M. Chiang 和 C. G. Brinton，“异构联邦学习的设备采样：理论、算法与实现”，发表于 INFOCOM。IEEE，2021 年，第 1–10 页。[22] W. Shi、S. Zhou、Z. Niu、M. Jiang 和 L. Geng，“面向时延受限无线联邦学习的联合设备调度与资源分配ing”，IEEE Transactions on Wireless Communications，第20卷，第1期，453–467页，2020年。[23] Y. Deng、F. Lyu、J. Ren、H. Wu、Y. Zhou、Y. Zhang 和 X. Shen，“Auction：一种面向高效联邦学习的自动化质量感知客户端选择框架”，IEEE Transactions on Parallel and Distributed Systems，第33卷，第8期，1996–2009页，2021年。[24] L. Liu、J. Zhang、S. Song 和 K. B. Letaief，“客户端-边缘-云分层联邦学习”，载于 ICC。IEEE，2020年，第1–6页。[25] Z. Wang、H. Xu、J. Liu、H. Huang、C. Qiao 和 Y. Zhao，“边缘计算中采用分层聚合的资源高效联邦学习”，载于 INFOCOM。IEEE，2021年，第1–10页。[26] P. Han、S. Wang 和 K. K. Leung，“面向高效联邦学习的自适应梯度稀疏化：一种在线学习方法”，载于 ICDCS。IEEE，2020年，第300–310页。[27] A. Reisizadeh、A. Mokhtari、H. Hassani、A. Jadbabaie 和 R. Pedarsani，“Fedpaq：一种采用周期性平均和量化的通信高效联邦学习方法”，载于人工智能与统计国际会议。PMLR，2020年，第2021–2031页。[28] N. Shlezinger、M. Chen、Y. C. Eldar、H. V. Poor 和 S. Cui，“具有量化约束的联邦学习”，载于 ICASSP。IEEE，2020年，第8851–8855页。[29] T. Lin、L. Kong、S. U. Stich 和 M. Jaggi，“用于联邦学习中鲁棒模型融合的集成蒸馏”，载于 NeurIPS，第33卷，2020年，第2351–2363页。[30] Z. Zhu、J. Hong 和 J. Zhou，“面向异构联邦学习的无数据知识蒸馏”，载于 ICML。PMLR，2021年，第12 878–12 889页。[31] A. Li、J. Sun、X. Zeng、M. Zhang、H. Li 和 Y. Chen，“Fedmask：通过异构掩蔽实现计算与通信高效的个性化联邦学习”，载于第19届 ACM 嵌入式网络传感器系统会议论文集，2021年，第42–55页。[32] E. Diao、J. Ding 和 V. Tarokh，“Heterofl：面向异构客户端的计算与通信高效联邦学习”，载于 ICLR，2021年。[33] H. Baek、W. J. Yun、Y. Kwak、S. Jung、M. Ji、M. Bennis、J. Park 和 J. Kim，“面向多宽度神经网络的联邦学习联合叠加编码与训练”，载于 INFOCOM。IEEE，2022年，第1729–1738页。[34] H. Wang、M. Yurochkin、Y. Sun、D. Papailiopoulos 和 Y. Khazaeni，“采用匹配平均的联邦学习”，载于 ICLR，2020年。[35] J. Kone˘cn`y、H. B. McMahan、F. X. Yu、P. Richt´arik、A. T. Suresh 和 D. Bacon，“联邦学习：提升通信效率的策略”，arXiv 预印本 arXiv:1610.05492，2016年。[36] D. Alistarh、D. Grubic、J. Li、R. Tomioka 和 M. Vojnovic，“Qsgd：通过梯度量化与编码实现通信高效的随机梯度下降”，载于 NeurIPS，2017年。[37] S. Han、H. Mao 和 W. J. Dally，“深度压缩：通过剪枝、训练后量化和霍夫曼编码压缩深度神经网络”，载于 ICLR，2016年。[38] S. Golomb，“游程编码（通信）”，IEEE 信息论汇刊，第12卷，第3期，399–401页，1966年。[39] M. Chen、H. V. Poor、W. Saad 和 S. Cui，“无线网络中联邦学习的收敛时间优化”，IEEE Transactions on Wireless Communications，第20卷，第4期，2457–2471页，2020年。[40] S. Chen、C. Shen、L. Zhang 和 Y. Tang，“联邦学习中异构量化的动态聚合”，IEEE Transactions on Wireless Communications，第20卷，第10期，6804–6819页，2021年。[41] H. Xiao、K. Rasul 和 R. Vollgraf，“Fashion-MNIST：用于机器学习算法基准测试的新型图像数据集”，arXiv 预印本 arXiv:1708.07747，2017年。[42] A. Krizhevsky、G. Hinton 等，“从微小图像中学习多层特征”，2009年。[43] K. Simonyan 和 A. Zisserman，“用于大规模图像识别的超深卷积网络”，载于 ICLR，2015年。[44] W. Shi、S. Zhou、Z. Niu、M. Jiang 和 L. Geng，“时延受限无线联邦学习的联合设备调度与资源分配”，IEEE Transactions on Wireless Communications，第20卷，第1期，453–467页，2021年。[45] S. Boyd、S. P. Boyd 和 L. Vandenberghe，《凸优化》。剑桥大学出版社，2004年。