面向无线联邦学习模型压缩的多维资源联合优化研究

朱光照，朱晓荣，徐鼎
（南京邮电大学通信与信息工程学院，江苏 南京 210003）

摘 要：针对边缘计算场景中，资源受限和网络动态的终端设备参与联邦学习产生的巨大时延和能耗问题，基于云–边–端三层联邦学习架构提出了一种高效训练和绿色节能的联邦学习算法。首先，将模型压缩技术引入三层联邦学习结构中，对三层联邦学习的模型收敛速率、训练时延和能耗进行理论分析。然后，根据理论分析结果进行问题建模，在一定的模型收敛速率下最小化全局模型训练时延和能耗，通过联合优化终端设备的发射功率、算力和模型压缩率，提高联邦学习的资源利用率。最后，将问题分解为 3 个优化子问题分别求解，设计了一种联合交替优化算法来获得原始问题的最优解。仿真结果表明，该算法可以适应大规模的边缘计算场景，在保证模型收敛速率的同时，与传统三层联邦学习算法相比，产生的时延和能耗分别减少了 71.54% 和 48.76%，有效地降低了全局模型训练产生的时延和能耗。

关键词：边缘计算；联邦学习；模型压缩；资源分配；收敛速率
中图分类号：TN929.5
文献标志码：A
doi：10.11959/j.issn.2096−3750.2025.00391

Joint multi-dimensional resource optimization for model compression in wireless federated learning

ZHU Guangzhao, ZHU Xiaorong, XU Ding
School of Communications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China

Abstract: In the edge computing scenarios, resource-constrained and participtionof the dynamically terminal devices of network in federated learning cause high latency and high energy consumption. An efficient and environmentally friendly federated learning algorithm based on a three-tier cloud-edge-terminal architecture was proposed. Firstly, by introducing model compression techniques into the three-tier federated learning structure, a theoretical analysis was conducted on the model convergence rate, training latency, and energy consumption. Subsequently, based on the theoretical analysis, a problem was formulated to minimize the global model training latency and energy consumption under a certain model convergence rate by jointly optimizing the terminal devices' transmission power, computing power, and model compression rate. Finally, by decomposing the problem into three sub-optimization problems and solving them alternately, a joint alternating optimization algorithm was designed to obtain the optimal solution for the original problem. Experimental results demonstrate that the proposed algorithm is adaptable to large-scale edge computing scenarios. It achieves reductions of 71.54% and 48.76%, respectively, in latency and energy consumption compared with traditional three-layer federated learning algorithms, while ensuring the convergence rate of the model, and effectively reduces the latency and energy consumption generated by global model training.

Key words: edge computing, federated learning, model compression, resource allocation, convergence rate

收稿日期：2024−01−26；**修回日期：**2024−04−14
通信作者：徐鼎，xuding@njupt.edu.cn
基金项目：国家自然科学基金资助项目（No. 92367102）；江苏省重点研发计划项目（No. BE2020084-3，No. BE2021013-3）
Foundation Items: The National Natural Science Foundation of China (No. 92367102), The Key Research and Development Plan of Jiangsu Province (No. BE2020084-3, No. BE2021013-3)

0　引言

随着对6G的探索，人们普遍认为6G将会建立在人工智能（AI, artificial intelligence）之上，需要推动AI与通信网络技术的进一步融合，实现6G网络的内生智能[1-3]。AI作为6G网络的主要应用场景和关键技术，主要体现在两个方面：AI驱动的新型无线通信系统（Al4NET, artificial intelligence for network）和无线网络驱动的AI服务（NET4AI, network for artificial intelligence）[4-6]。在云端集中式智能向深度边缘泛在智能演进的过程中，联邦学习作为一种主流的分布式学习范式被广泛应用到6G网络内生智能架构的设计中[7-9]。在联邦学习范式中，各个参与方交换与模型相关的训练参数，然后由中央服务器进行安全聚合并反馈给参与方，参与方负责根据反馈的模型信息进行己方模型的更新[10-11]。

在联邦学习的实际应用中，用户和服务器之间需要进行模型参数的多轮交互，这种学习机制训练出的泛化性能良好的全局模型能耗大，且会产生巨大的训练时延[12-13]。在边缘网络中，大量边缘设备参与联邦学习，然而每个边缘设备的存储、计算和通信能力因为硬件、网络连接和电源的变化会有所不同[14-15]，并且每个设备并不完全可靠，由于网络连接和能量的限制，还可能出现参与设备在一定迭代过程中退出的情况。在边缘设备的模型训练方面，联邦学习网络中各个边缘设备产生或收集到的数据各不相同，这些数据的统计特征服从非独立同分布（Non-IID, non independent identically distributed），这种数据分布严重影响全局模型的收敛速率。同时，边缘设备所处的通信环境复杂多变，边缘网络中的通信时间可能比本地模型训练时间低许多个数量级，这不仅会产生过高的汇聚时延影响全局模型的收敛速率，而且参与设备也会产生大量的能耗，造成能量资源的浪费[16-17]。

为解决联邦学习巨大的训练时延问题，文献[18-20]提出了模型参数传输的梯度压缩算法，通过减少传输模型的数据量，达到降低模型训练时延的预期，但是这些梯度压缩算法为所有终端设备设置相同的梯度压缩比，无法根据动态变化的通信环境和终端设备的模型训练情况自适应地调整梯度压缩比。在此基础上，文献[21]将终端设备选择和模型压缩率进行联合优化，通过自适应选择参与聚合的终端设备和梯度压缩率来降低模型训练时延。文献[22]采用异步联邦学习的机制，设计了一种边缘节点自适应选择局部模型更新频率和模型压缩比的控制算法，减少了模型聚合的等待时间，提高了训练效率。文献[23]提出了一种阈值自适应的梯度通信压缩机制，减少了终端设备与边缘服务器之间的冗余通信，有效地提升了模型训练的整体通信效率。文献[24]提出了一种自适应批量大小选择与梯度压缩算法，根据动态网络环境自适应地选择终端设备的训练批量大小和梯度压缩率，进而加快了模型的收敛速率。以上文献[18-24]提出的算法均采用模型梯度压缩的方法，从减少传输模型的数据量出发，降低联邦学习的训练时延，但是忽略了模型梯度压缩造成全局模型收敛速率下降的影响和参加模型训练的设备能耗问题。全局模型收敛速率的下降将额外增加模型交互的轮次，导致联邦学习训练时延和能耗的增加，同时，面对能量受限的终端设备，设备能耗是制约全局模型训练的重要因素。

在降低联邦学习训练能耗方面，文献[25]在非正交多址和时分多址两种传输协议下，通过联合通信与计算两个方面设计了提升系统能量效率的联邦学习算法。文献[26]通过联合建模联邦学习的训练时延、计算能耗和学习精度问题，表征终端设备计算和通信时延对终端设备能耗的影响，同时研究了模型训练时间和训练精度参数之间的权衡。文献[27]通过最小化学习性能约束下参数传输的能量消耗成本，提出了一种具有节能特性的联邦学习框架。文献[28]综合考虑了模型训练时延、能耗和模型精度之间的均衡，采用时机性选择集中式学习和联邦学习的方式，提出了一种联合机会式集中学习的联邦学习算法。以上文献[25-28]均从终端设备的通信和计算维度出发考虑模型训练时延与训练能耗之间的均衡问题，降低联邦学习的训练能耗，而忽视了模型压缩对联邦学习的训练时延、能耗和收敛速率的影响，没有通过模型压缩更进一步地减少联邦学习的训练能耗。

为了适应资源受限和网络动态的边缘计算场景，弥补模型压缩算法缺乏考虑终端设备能耗问题的不足，完善联邦学习模型压缩中训练时延和能耗的综合分析，本文在云–边–端三层联邦学习结构下应用模型压缩技术，综合考虑联邦学习中模型训练时延与能耗之间的均衡问题，理论分析了模型压缩

对于模型训练时延和能耗的影响，设计了一种具有高效训练和绿色节能特性的联邦学习算法。本文研究的主要贡献为以下 3 个方面。

1) 提出了针对云–边–端三层联邦学习结构的模型压缩策略，通过综合考虑终端层和边端层的模型参数特点，减少传输的模型参数信息。对联邦学习模型训练的性能进行理论分析，采用梯度平均 ℓ_2 范数的上界评估模型收敛速度，并从数学上推导出模型训练的时延和能耗。

2) 提出了在受限的模型收敛速率下最小化全局模型训练时延和能耗的优化问题，通过联合优化终端设备发射功率、终端算力和终端模型压缩率，综合考虑终端层多维资源的利用情况，降低联邦学习的训练时延和能耗。

3) 设计了一种具有高效训练和绿色节能特性的联邦学习算法，通过理论分析将原始优化问题分解为 3 个子问题分别求解，利用 3 个子问题的交替优化获得终端设备的发射功率、终端算力和终端模型压缩率原始优化问题的最优解。

1　系统模型与问题建模

1.1　联邦学习模型

本文考虑一个三层无线联邦学习场景[29]，多个边端基站覆盖的终端设备联合完成全局模型训练任务，三层无线联邦学习场景主要由 3 个部分组成：终端层、边端层和云端层，三层无线联邦学习结构如图 1 所示。云端层由算力和存储资源不受限的云端服务器组成；边端层由多个独立且配备算力和存储资源不受限的基站服务器组成；终端层由算力、能量和通信资源受限的终端设备组成。在本场景中，边端层包含 K 个基站，集合表示为 $K_{\mathrm{BS}}=\{1,2,3,\cdots,K\}$，一个基站覆盖 N 个终端设备，集合为 $N_{\mathrm{UE}}=\{1,2,3,\cdots,N\}$。全局模型训练过程中，一次云端全局模型更新要依次执行终端模型训练、边端模型聚合和云端模型聚合 3 个主要步骤，最后广播更新的全局模型，具体执行内容如下。

1) 终端模型训练：终端设备利用本地数据集进行多轮次的本地训练，然后将本地训练的模型参数发送给边端服务器。在本场景中，采用随机梯度下降（SGD, stochastic gradient descent）算法进行模型训练。因此，第 k 个基站内的第 n 个终端模型的训练过程可以表示为

$$\boldsymbol{\omega}_{k,n}^{i+1}=\boldsymbol{\omega}_{k,n}^{i}-\eta\cdot\nabla\boldsymbol{F}_{k,n}\left(\boldsymbol{\omega}_{k,n}^{i}\right) \tag{1}$$

其中，η 表示模型训练的学习率，$\boldsymbol{\omega}_{k,n}^{i}$ 表示第 i 轮终端设备的模型，$\boldsymbol{\omega}_{k,n}^{i+1}$ 表示第 $i+1$ 轮终端设备的模型，$\nabla\boldsymbol{F}_{k,n}\left(\boldsymbol{\omega}_{k,n}^{i}\right)$ 表示模型训练的损失函数梯度。

2) 边端模型聚合：边端服务器接收来自终端设备的模型参数进行模型聚合，根据特定的边端聚合次数，将聚合的边端模型广播给终端设备或者发送给云端服务器。第 k 个基站的边端模型聚合过程可以表示为

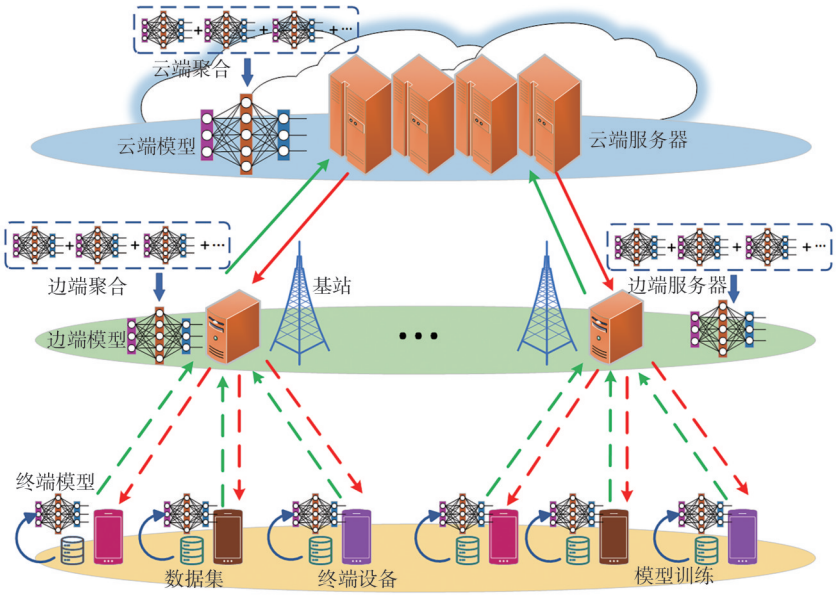

图 1　三层无线联邦学习结构

$$\Delta\boldsymbol{\omega}_k = \frac{1}{|D_k|}\sum_{n=1}^{N}|D_{k,n}|\Delta\boldsymbol{\omega}_{k,n} \tag{2}$$

其中，$|D_k|$和$\Delta\boldsymbol{\omega}_k$分别表示第$k$个基站所连接终端设备的所有数据集大小和基站服务器边端模型参数，$|D_{k,n}|$和$\Delta\boldsymbol{\omega}_{k,n}$分别表示第$k$个基站内的第$n$个终端设备的数据集大小和终端设备的终端模型参数。

3) 云端模型聚合：云端服务器接收来自边端服务器的模型参数进行模型聚合，然后将聚合的云端全局模型广播给边端服务器和终端设备。云端全局模型聚合过程可以表示为

$$\Delta\boldsymbol{\omega} = \frac{1}{|D|}\sum_{k=1}^{K}|D_k|\Delta\boldsymbol{\omega}_k \tag{3}$$

其中，$|D|$和$\Delta\boldsymbol{\omega}$分别表示全部基站连接终端设备的所有数据集大小和云端全局模型参数。

1.2　压缩模型

在终端层、边端层和云端层之间交互的模型信息通常具有稀疏性，即大部分的模型权重接近零，该部分信息几乎不促进全局模型的更新。因此，在保证模型收敛速度和模型精确度的前提下，本文采用gTop-K算法[19]，对模型参数进行压缩处理，减少模型交互的数据量。gTop-K算法的核心思想是选择模型参数中绝对数值最大的v个参数，可以表示为

$$\Delta\tilde{\boldsymbol{\omega}} = \text{Top-}K(\Delta\boldsymbol{\omega}) = \begin{cases}\Delta\boldsymbol{\omega}, |\Delta\boldsymbol{\omega}| > \Delta\boldsymbol{\omega}^{\text{th}} \\ 0, \text{其他}\end{cases} \tag{4}$$

其中，$\Delta\boldsymbol{\omega}^{\text{th}}$表示$|\Delta\boldsymbol{\omega}|$中的第$v$个大的参数，$\Delta\tilde{\boldsymbol{\omega}}$表示经过压缩之后的模型参数。于是，可以将模型压缩率定义为模型参数的总个数V与被选中传输模型参数个数v的比值，即

$$c = \frac{V}{v} \tag{5}$$

同时，gTop-K算法将产生的压缩误差进行累加，用于下一次算法的执行，即

$$\Delta\tilde{\boldsymbol{\omega}}^{t+1} = \text{Top-}K\left(\Delta\boldsymbol{\omega}^{t+1} + \boldsymbol{e}^{t}\right) \tag{6}$$

其中，$\boldsymbol{e}^{t} = \Delta\boldsymbol{\omega}^{t} - \Delta\tilde{\boldsymbol{\omega}}^{t}$表示累计的模型压缩误差。

原始gTop-K算法执行在两层联邦学习架构中，所有终端设备均采用相同固定的模型压缩率，终端设备之间需要相互通信进行模型的对比聚合，从而确定最终上传的模型参数。在云–边–端三层联邦学习架构中，为使gTop-K算法适应动态变化的通信环境并减少终端侧通信资源的浪费，本文将该算法做出改进：终端设备之间不进行相互通信，终端模型聚合由基站服务器完成，所有基站服务器采用相同的压缩率压缩模型参数，基站连接的终端设备根据自身的通信和算力资源自适应地选择模型压缩率，仅需保证终端模型的平均压缩率等于基站的压缩率，即

$$c = \frac{1}{N}\sum_{n=1}^{N}c_{k,n} \tag{7}$$

在三层无线联邦学习场景中，采用模型训练梯度的平均ℓ_2范数衡量全局模型训练的收敛速率，本文考虑一般的场景，其中每个终端设备可以根据计算能力和信道状态信息选择合适的模型压缩率。根据文献[30]对gTop-K算法模型训练收敛速率的分析，可以得到定理1。

定理1　当终端设备以模型压缩率c进行联邦学习时，训练S轮后模型梯度平均ℓ_2范数的上界可以表示为

$$\vartheta = \frac{\tau}{\sqrt{bS}} + \frac{\mu\left(c^3 - c\right)}{S} \tag{8}$$

其中，τ和μ是与训练模型有关的参数，c表示模型参数的压缩率，S表示模型训练的轮次，b表示终端设备模型训练的批量大小。

在定理1中，ϑ表示模型梯度平均ℓ_2范数的上界，ϑ越小，模型的收敛速率越大。从定理1可知，当不采用模型参数压缩算法，模型训练具有最大的收敛速率，但是传输大量的模型参数会产生巨大的时延和能耗。因此，在模型信息交互时，要采用合适的模型压缩率，用于平衡模型训练的收敛速率与通信成本之间的相互影响。

1.3　通信模型

在三层联邦学习系统中，为了提高信道频谱效率，终端层与边端层之间采用正交频分多址（OFDMA, orthogonal frequency division multiple access）技术共享信道资源，不考虑终端设备之间的干扰。由于实际通信环境中，信道状态信息的时间尺度远小于模型训练的时间尺度，本文利用训练时间内平均信道状态信息来代替瞬时信道状态信息。因此，第k个基站内的第n个终端设备传输模型参数的上行速率可以表示为

$$R_{k,n} = B\,\text{lb}\left(1 + \frac{p_{k,n}h_{k,n}}{N_0}\right) \tag{9}$$

其中，B表示分配给每个终端设备的带宽，$p_{k,n}$表示终端设备的发射功率，$h_{k,n}$表示信道的功率增益，

N_0表示信道存在的噪声功率。在模型参数下行传输阶段，基站可调度所有下行信道资源用于模型参数的广播，该部分的时延可以忽略不计。边端层与云端层之间采用有线连接的方式进行通信，每个基站的上行数据传输速率可以表示为R。同样地，云端层广播模型的时延也可以忽略不计。

1) 训练时延分析

在云–边–端三层联邦学习场景中，全局模型训练时延主要由终端模型训练时延、终端模型传输时延和边端模型传输时延3部分组成，模型聚合时延和模型更新时延非常小，可以忽略不计。

终端模型训练时延：第k个基站内的第n个终端设备模型训练时延可以表示为

$$T_{k,n}^{\mathrm{C}}=\frac{m_1\left|D_{k,n}\right|f_0}{f_{k,n}} \tag{10}$$

其中，m_1表示终端设备本地模型的训练次数，f_0表示训练1 bit数据所需要的CPU周期数，$f_{k,n}$表示终端设备训练时的调配算力。

终端模型传输时延：第k个基站内的第n个终端设备上行传输模型的时延可以表示为

$$T_{k,n}^{\mathrm{U}}=\frac{qV}{c_{k,n}R_{k,n}} \tag{11}$$

其中，q表示每个模型参数的量化位数，V表示终端设备全部模型参数的个数，$c_{k,n}$表示终端设备的模型压缩率。

边端模型传输时延：第k个基站上行传输模型的时延可以表示为

$$T_k=\frac{qV}{cR} \tag{12}$$

本文场景采用同步模型聚合机制，因此一次边端模型聚合的时延可以表示为

$$t_k=\max_{k,n}\left\{T_{k,n}^{\mathrm{C}}+T_{k,n}^{\mathrm{U}}\right\} \tag{13}$$

边端层的服务器需要进行m_2轮次的边端模型聚合才进行一次云端全局模型聚合，为了方便表示每轮边端模型聚合的时延，可以将边端层和云端层交互的时延平均划分为m_2个部分[31]，那么一轮的边端模型汇聚的时延可以表示为

$$T=t_k+\frac{T_k}{m_2} \tag{14}$$

2) 训练能耗分析

本文终端设备处于资源受限的状态，仅考虑终端层中全局模型训练的能耗情况，因此全局模型训练能耗由终端模型训练能耗和终端模型传输能耗两个部分组成。

终端模型训练能耗：第k个基站内的第n个终端设备模型训练能耗可以表示为

$$E_{k,n}^{\mathrm{C}}=m_1\left|D_{k,n}\right|f_0\xi f_{k,n}^2 \tag{15}$$

其中，ξ表示与芯片结构有关的有效转换容量[32]。

终端模型传输能耗：第k个基站内的第n个终端设备上行传输模型能耗可以表示为

$$E_{k,n}^{\mathrm{U}}=\frac{qVp_{k,n}}{c_{k,n}R_{k,n}} \tag{16}$$

因此，在一次边端模型汇聚过程中，全部基站所包含终端设备的能耗可以表示为

$$E=\sum_{k=1}^{K}\sum_{n=1}^{N}\left(E_{k,n}^{\mathrm{C}}+E_{k,n}^{\mathrm{U}}\right) \tag{17}$$

1.4　问题建模

根据模型压缩收敛速率、模型训练时延和能耗的分析，本文考虑如何在保证模型收敛速率的前提下，减少模型训练的时延和能耗。一方面，采用较大的模型压缩率，联邦学习的时延和能耗会大大降低，但也会导致模型收敛速率明显下降。另一方面，采用较小的模型压缩率，会获得较高的模型收敛速率，但会产生巨大的训练时延和能耗。因此，为了确保模型收敛速率、降低模型训练的时延和能耗，需要为每个终端设备自适应地选择终端发射功率、终端算力和终端模型压缩率，其中，终端模型压缩率衡量终端AI模型参数的数据式资源利用率。通过联合优化终端发射功率资源、终端算力资源和终端AI模型参数的数据式资源，进行多维资源联合优化算法的研究设计。综上所述，对于终端层资源受限的三层无线联邦学习系统，优化目标可以表述为：在给定模型收敛速率下，最小化模型训练的时延和能耗。此优化问题的数学表述为

$$\mathrm{P1}:\min_{c_{k,n},p_{k,n},f_{k,n}}P_1=\gamma_{\mathrm{T}}S_0T+\gamma_{\mathrm{E}}S_0E \tag{18}$$

$$\text{s.t.}\ \frac{\tau}{\sqrt{bS}}+\frac{\mu\left(c^3-c\right)}{S}\leqslant\vartheta_0 \tag{18a}$$

$$c=\frac{1}{N}\sum_{n=1}^{N}c_{k,n},\forall k\in K_{\mathrm{BS}},\forall n\in N_{\mathrm{UE}} \tag{18b}$$

$$1\leqslant c_{k,n}\leqslant c_{\max},\forall k\in K_{\mathrm{BS}},\forall n\in N_{\mathrm{UE}} \tag{18c}$$

$$0<f_{k,n}\leqslant f_{\max},\forall k\in K_{\mathrm{BS}},\forall n\in N_{\mathrm{UE}} \tag{18d}$$

$$0<p_{k,n}\leqslant p_{\max},\forall k\in K_{\mathrm{BS}},\forall n\in N_{\mathrm{UE}} \tag{18e}$$

其中，γ_{T}和γ_{E}分别表示模型训练时延和能耗的归一化权重系数，$\gamma_{\mathrm{T}}+\gamma_{\mathrm{E}}=1$，$S_0$表示边端模型聚合

的次数，ϑ_0表示模型压缩指定的最大模型训练梯度平均ℓ_2范数，$c_{\max}$表示终端设备模型压缩的最大压缩率，$f_{\max}$表示终端设备最大的配备算力，$p_{\max}$表示终端设备最大的发射功率。具体来说，式(18a)约束了三层联邦学习中模型训练的最小模型收敛速率；式(18b)保证了每个边端模型相同的压缩率和每个基站覆盖范围平均模型压缩率的相同；式(18c)确保了终端模型压缩率的合理性，杜绝过高模型压缩率的出现；式(18d)和式(18e)分别约束了终端层的终端发射功率和终端算力的受限取值范围。

2　联合优化算法设计

在优化问题P1中，3个优化变量$c_{k,n}$、$f_{k,n}$和$p_{k,n}$相互耦合，导致了优化问题P1的非凸性，无法使用现有的数学方法直接求出优化问题的最优解。本文通过交替优化方法[33]分别求解终端发射功率$p_{k,n}$、终端算力$f_{k,n}$和终端模型压缩率$c_{k,n}$。首先，给定$f_{k,n}$和$c_{k,n}$求解最优的$p_{k,n}$；然后，给定$p_{k,n}$和$c_{k,n}$求解最优的$f_{k,n}$；最后，将$p_{k,n}$和$f_{k,n}$代入优化问题P1求解$c_{k,n}$，不断交替求解，直到目标函数收敛。即将优化问题P1解耦成3个独立的优化子问题，分别为终端发射功率优化子问题、终端算力优化子问题和终端模型压缩优化子问题。针对这3个优化子问题分别求解出各自的最优解，根据3个子问题的最优解设计出一种联合交替优化算法，最后获得优化问题P1的最优解。

2.1　终端发射功率优化子问题

优化问题P1中的优化变量$p_{k,n}$决定终端设备的传输时延和传输能耗，仅接受式(18e)的条件约束。本场景中，不同基站的终端设备之间相互独立，优化问题P1可以分解为K个基站内部的终端发射功率优化子问题。因此，每个基站的终端发射功率优化子问题可以表示为

$$\text{P2:}\min_{p_{k,n}} P_2^k=\gamma_{\mathrm{T}}S_0\max_n\left\{T_{k,n}^{\mathrm{C}}+T_{k,n}^{\mathrm{U}}\right\}+\gamma_{\mathrm{E}}S_0\sum_{n=1}^{N}E_{k,n}^{\mathrm{U}} \tag{19}$$

$$\text{s.t. } 0<p_{k,n}\leqslant p_{\max},\forall n\in N_{\mathrm{UE}} \tag{19a}$$

对于优化问题P2，令$x_k=\max_n\left\{x_{k,n}\right\}=\max_n\left\{T_{k,n}^{\mathrm{C}}+T_{k,n}^{\mathrm{U}}\right\}$，优化问题P2可以等价地转化为$N$个独立的优化问题，表示为

$$\text{P3:}\min_{x_{k,n}} P_3^k=\gamma_{\mathrm{T}}S_0x_k+\gamma_{\mathrm{E}}S_0\sum_{n=1}^{N}\left(x_{k,n}-T_{k,n}^{\mathrm{C}}\right)\frac{N_0}{h_{k,n}}\left(2^{\frac{qV}{c_{k,n}B\left(x_{k,n}-T_{k,n}^{\mathrm{C}}\right)}}-1\right) \tag{20}$$

$$\text{s.t. } x_{k,n}\geqslant T_{k,n}^{\mathrm{C}}+\frac{qV}{c_{k,n}B\,\mathrm{lb}\left(1+\dfrac{p_{\max}h_{k,n}}{N_0}\right)},\forall n\in N_{\mathrm{UE}} \tag{20a}$$

$$x_k\geqslant x_{k,n},\forall n\in N_{\mathrm{UE}} \tag{20b}$$

优化问题P3的目标函数P_3^k是关于$x_{k,1},x_{k,2},\cdots,x_{k,N}$的凸函数，同时，式(20a)和式(20b)满足线性规范性条件[28]，采用KKT（Karush-Kuhn-Tucker）条件进行求解，则优化问题P3的拉格朗日函数表示为

$$\begin{aligned}L_1^k=&\gamma_{\mathrm{T}}S_0x_k+\sum_{n=1}^{N}\beta_{k,n}^p\left(x_{k,n}-x_k\right)+\\&\gamma_{\mathrm{E}}S_0\sum_{n=1}^{N}\left(x_{k,n}-T_{k,n}^{\mathrm{C}}\right)\frac{N_0}{h_{k,n}}\left(2^{\frac{qV}{c_{k,n}B\left(x_{k,n}-T_{k,n}^{\mathrm{C}}\right)}}-1\right)+\\&\sum_{n=1}^{N}\alpha_{k,n}^p\left(T_{k,n}^{\mathrm{C}}+\frac{qV}{c_{k,n}B\,\mathrm{lb}\left(1+\dfrac{p_{\max}h_{k,n}}{N_0}\right)}-x_{k,n}\right)\end{aligned} \tag{21}$$

其中，$\alpha_{k,n}^p$和$\beta_{k,n}^p$是优化问题P3中约束条件对应的拉格朗日乘子。

KKT条件为

$$\begin{aligned}&\frac{\partial L_1^k}{\partial x_{k,n}}=0,\frac{\partial L_1^k}{\partial x_k}=0,\\&\alpha_{k,n}^p\left(T_{k,n}^{\mathrm{C}}+\frac{qV}{c_{k,n}B\,\mathrm{lb}\left(1+\dfrac{p_{\max}h_{k,n}}{N_0}\right)}-x_{k,n}\right)=0,\\&\beta_{k,n}^p\left(x_{k,n}-x_k\right)=0,\alpha_{k,n}^p\geqslant 0,\beta_{k,n}^p\geqslant 0,\forall n\in N_{\mathrm{UE}}\end{aligned} \tag{22}$$

将$\partial L_V^k/\partial x_{k,n}=0$展开，可得

$$\frac{\partial L_1^k}{\partial x_{k,n}}=\gamma_{\mathrm{E}}S_0\frac{N_0}{h_{k,n}}\left(u_{k,n}-1-u_{k,n}\ln u_{k,n}\right)-\alpha_{k,n}^p+\beta_{k,n}^p \tag{23}$$

其中，$u_{k,n}=2^{\frac{qV}{c_{k,n}B\left(x_{k,n}-T_{k,n}^{\mathrm{C}}\right)}}$，由于式(23)中$\left(u_{k,n}-1-u_{k,n}\ln u_{k,n}\right)<0$，根据KKT条件可以得到$x_{k,n}=x_k$，$\forall n\in N_{\mathrm{UE}}$。

因此，优化问题P3可以转化为关于x_k的单变量优化问题，表示为

$$\text{P4:}\min_{x_k} P_4^k=\gamma_{\mathrm{T}}S_0x_k+\gamma_{\mathrm{E}}S_0\sum_{n=1}^{N}\left(x_k-T_{k,n}^{\mathrm{C}}\right)\frac{N_0}{h_{k,n}}\left(2^{\frac{qV}{c_{k,n}B\left(x_k-T_{k,n}^{\mathrm{C}}\right)}}-1\right) \tag{24}$$

$$\text{s.t. } x_k \geqslant \max_n \left\{ T_{k,n}^{\mathrm{C}} + \frac{qV}{c_{k,n} B \operatorname{lb}\left(1 + \frac{p_{\max} h_{k,n}}{N_0}\right)} \right\}, \forall n \in N_{\mathrm{UE}} \tag{24a}$$

由于优化问题P4的目标函数P_4^k满足$\mathrm{d}^2 P_4^k / \mathrm{d}x_k^{\,2} \geqslant 0$，因此优化问题P4是关于$x_k$的凸问题，最优解可以表示为

$$x_k^{\mathrm{opt}} = \begin{cases} x_k^*, x_k^* \geqslant x_{\min}^k \\ x_{\min}^k, x_k^* < x_{\min}^k \end{cases} \tag{25}$$

其中，$x_{\min}^k = \max_n \left\{ T_{k,n}^{\mathrm{C}} + \frac{qV}{c_{k,n} B \operatorname{lb}\left(1 + \frac{p_{\max} h_{k,n}}{N_0}\right)} \right\}$，$x_k^*$为$\mathrm{d}P_1^k / \mathrm{d}x_k = 0$的解。

因此，终端发射功率优化子问题的最优解可以表示为

$$p_{k,n}^{\mathrm{opt}} = \frac{N_0}{h_{k,n}} \left(2^{\frac{qV}{c_{k,n} B \left(x_k^{\mathrm{opt}} - T_{k,n}^{\mathrm{C}} \right)}} - 1 \right), \forall n \in N_{\mathrm{UE}} \tag{26}$$

2.2　终端算力优化子问题

优化问题P1中的优化变量$f_{k,n}$决定终端设备的计算时延和计算能耗，仅接受式(18d)条件的约束。与终端发射功率优化子问题类似，优化问题P1可以分解为K个基站内部的终端算力优化子问题。因此，每个基站的终端算力优化子问题可以表示为

$$\text{P5: } \min_{f_{k,n}} P_5^k = \gamma_{\mathrm{T}} S_0 \max_n \left\{ T_{k,n}^{\mathrm{C}} + T_{k,n}^{\mathrm{U}} \right\} + \gamma_{\mathrm{E}} S_0 \sum_{n=1}^{N} E_{k,n}^{\mathrm{C}} \tag{27}$$

$$\text{s.t. } 0 < f_{k,n} \leqslant f_{\max}, \forall n \in N_{\mathrm{UE}} \tag{27a}$$

对于优化问题P5，令$y_k = \max_n \left\{ y_{k,n} \right\} = \max_n \left\{ T_{k,n}^{\mathrm{C}} + T_{k,n}^{\mathrm{U}} \right\}$，优化问题P5同样转化为$N$个独立的优化问题，表示为

$$\text{P6: } \min_{y_{k,n}} P_6^k = \gamma_{\mathrm{T}} S_0 y_k + \gamma_{\mathrm{E}} S_0 \sum_{n=1}^{N} \frac{\xi \left(m_1 \left| D_{k,n} \right| f_0 \right)^3}{\left(y_{k,n} - T_{k,n}^{\mathrm{U}} \right)^2} \tag{28}$$

$$\text{s.t. } y_{k,n} \geqslant \frac{m_1 \left| D_{k,n} \right| f_0}{f_{\max}} + T_{k,n}^{\mathrm{U}}, \forall n \in N_{\mathrm{UE}} \tag{28a}$$

$$y_k \geqslant y_{k,n}, \forall n \in N_{\mathrm{UE}} \tag{28b}$$

优化问题P6的目标函数P_6^k是关于$y_{k,1}, y_{k,2}, \cdots, y_{k,N}$的凸函数，同时，式(28a)和式(28b)满足线性规范性条件[28]，采用KKT条件进行求解，则优化问题P6的拉格朗日函数表示为

$$\begin{aligned} L_2^k = {} & \gamma_{\mathrm{T}} S_0 y_k + \gamma_{\mathrm{E}} S_0 \sum_{n=1}^{N} \frac{\xi \left(m_1 \left| D_{k,n} \right| f_0 \right)^3}{\left(y_{k,n} - T_{k,n}^{\mathrm{U}} \right)^2} + \\ & \sum_{n=1}^{N} \alpha_{k,n}^f \left(\frac{m_1 \left| D_{k,n} \right| f_0}{f_{\max}} + T_{k,n}^{\mathrm{U}} - y_{k,n} \right) + \\ & \sum_{n=1}^{N} \beta_{k,n}^f \left(y_{k,n} - y_k \right) \end{aligned} \tag{29}$$

其中，$\alpha_{k,n}^f$和$\beta_{k,n}^f$是优化问题P6中约束条件对应的拉格朗日乘子。

KKT条件为

$$\begin{aligned} & \frac{\partial L_2^k}{\partial y_{k,n}} = 0, \frac{\partial L_2^k}{\partial y_k} = 0, \\ & \alpha_{k,n}^f \left(\frac{m_1 \left| D_{k,n} \right| f_0}{f_{\max}} + T_{k,n}^{\mathrm{U}} - y_{k,n} \right) = 0, \\ & \beta_{k,n}^f \left(y_{k,n} - y_k \right) = 0, \alpha_{k,n}^f \geqslant 0, \beta_{k,n}^f \geqslant 0, \forall n \in N_{\mathrm{UE}} \end{aligned} \tag{30}$$

将$\partial L_2^k / \partial y_{k,n} = 0$展开，可得

$$\frac{\partial L_2^k}{\partial y_{k,n}} = -2\gamma_{\mathrm{E}} S_0 \frac{\xi \left(m_1 \left| D_{k,n} \right| f_0 \right)^3}{\left(y_{k,n} - T_{k,n}^{\mathrm{U}} \right)^3} - \alpha_{k,n}^f + \beta_{k,n}^f \tag{31}$$

根据KKT条件可以得到$y_{k,n} = y_k, \forall n \in N_{\mathrm{UE}}$。

因此，优化问题P6可以转化为关于y_k的单变量优化问题，表示为

$$\text{P7: } \min_{y_k} P_7^k = \gamma_{\mathrm{T}} S_0 y_k + \gamma_{\mathrm{E}} S_0 \sum_{n=1}^{N} \frac{\xi \left(m_1 \left| D_{k,n} \right| f_0 \right)^3}{\left(y_k - T_{k,n}^{\mathrm{U}} \right)^2} \tag{32}$$

$$\text{s.t. } y_k \geqslant \max_n \left\{ \frac{m_1 \left| D_{k,n} \right| f_0}{f_{\max}} + T_{k,n}^{\mathrm{U}} \right\}, \forall n \in N_{\mathrm{UE}} \tag{32a}$$

同样地，优化问题P7的目标函数P_7^k满足$\mathrm{d}^2 P_7^k / \mathrm{d}y_k^{\,2} \geqslant 0$，因此优化问题P7是关于$y_k$的凸问题，最优解可以表示为

$$y_k^{\mathrm{opt}} = \begin{cases} y_k^*, y_k^* \geqslant y_{\min}^k \\ y_{\min}^k, y_k^* < y_{\min}^k \end{cases} \tag{33}$$

其中，$y_{\min}^k = \max_n \left\{ \frac{m_1 \left| D_{k,n} \right| f_0}{f_{\max}} + T_{k,n}^{\mathrm{U}} \right\}$，$\forall n \in N_{\mathrm{UE}}$，$y_k^*$为$\mathrm{d}P_7^k / \mathrm{d}y_k = 0$的解。

因此，终端算力优化子问题的最优解可以表示为

$$f_{k,n}^{\mathrm{opt}} = \frac{m_1 \left| D_{k,n} \right| f_0}{y_k^{\mathrm{opt}} - T_{k,n}^{\mathrm{U}}}, \forall n \in N_{\mathrm{UE}} \tag{34}$$

2.3　终端模型压缩优化子问题

优化问题P1中的优化变量$c_{k,n}$决定终端设备的

传输时延和传输能耗，接受式(18a)至式(18c)的条件约束。观察发现式(18a)条件中，ϑ是关于$c\in[1,+\infty)$的单调递增函数，则在给定$\vartheta\leqslant\vartheta_0$的情况下，对应$c_0$的值唯一确定。因此，可以将式(18a)和式(18b)条件重写为

$$\frac{1}{N}\sum_{n=1}^{N}c_{k,n}\leqslant c_0,\forall k\in K_{\mathrm{BS}},\forall n\in N_{\mathrm{UE}} \tag{35}$$

与前两个优化子问题类似，优化问题P1可以分解为K个基站内部的终端模型压缩优化子问题，同时将目标函数中$\max\limits_{n}\left\{T_{k,n}^{\mathrm{C}}+T_{k,n}^{\mathrm{U}}\right\}$近似为$\left(T_{k,n}^{\mathrm{C}}+T_{k,n}^{\mathrm{U}}\right)$。因此，每个基站的终端模型压缩优化子问题可以表示为

$$\mathrm{P8}:\min_{c_{k,n}}P_8^k=\gamma_{\mathrm{T}}S_0\left(T_{k,n}^{\mathrm{C}}+T_{k,n}^{\mathrm{U}}\right)+\gamma_{\mathrm{E}}S_0\sum_{n=1}^{N}E_{k,n}^{\mathrm{U}} \tag{36}$$

$$\text{s.t.}\ \frac{1}{N}\sum_{n=1}^{N}c_{k,n}\leqslant c_0,\forall n\in N_{\mathrm{UE}} \tag{36a}$$

$$1\leqslant c_{k,n}\leqslant c_{\max},\forall n\in N_{\mathrm{UE}} \tag{36b}$$

优化问题P8是一个凸优化问题[8]，采用KKT条件求解，则优化问题P8的拉格朗日函数表示为

$$\begin{aligned}L_3^k=&\gamma_{\mathrm{T}}S_0\left[T_{k,n}^{\mathrm{C}}+\frac{qV}{c_{k,n}B\,\mathrm{lb}\left(1+\dfrac{p_{k,n}h_{k,n}}{N_0}\right)}\right]+\\&\gamma_{\mathrm{E}}S_0\sum_{n=1}^{N}\frac{qVp_{k,n}}{c_{k,n}B\,\mathrm{lb}\left(1+\dfrac{p_{k,n}h_{k,n}}{N_0}\right)}+\\&\alpha_k^c\left(\frac{1}{N}\sum_{n=1}^{N}c_{k,n}-c_0\right)\end{aligned} \tag{37}$$

其中，α_k^c是优化问题P8中约束条件对应的拉格朗日乘子。

KKT条件为

$$\begin{aligned}&\frac{\partial L_3^k}{\partial c_{k,n}}=\frac{\alpha_k^c}{N}-\frac{S_0qV\left(\gamma_{\mathrm{T}}+\gamma_{\mathrm{E}}p_{k,n}\right)}{{c_{k,n}}^2B\,\mathrm{lb}\left(1+\dfrac{p_{k,n}h_{k,n}}{N_0}\right)}=0,\\&\alpha_k^c\left(\frac{1}{N}\sum_{n=1}^{N}c_{k,n}-c_0\right)=0,\alpha_k^c\geqslant 0,\forall n\in N_{\mathrm{UE}}\end{aligned} \tag{38}$$

根据式(38)的KKT条件，可得

$$\frac{1}{N}\sum_{n=1}^{N}c_{k,n}-c_0=0,\forall n\in N_{\mathrm{UE}} \tag{39}$$

根据式(38)和式(39)，可以推导出

$$c_{k,n}^{*}=\sqrt{\frac{S_0qVN\left(\gamma_{\mathrm{T}}+\gamma_{\mathrm{E}}p_{k,n}\right)}{\alpha_k^cB\,\mathrm{lb}\left(1+\dfrac{p_{k,n}h_{k,n}}{N_0}\right)}},\forall n\in N_{\mathrm{UE}} \tag{40}$$

因此，终端模型压缩优化子问题的最优解可以表示为

$$c_{k,n}^{\mathrm{opt}}=\sqrt{\frac{S_0qVN\left(\gamma_{\mathrm{T}}+\gamma_{\mathrm{E}}p_{k,n}\right)}{\alpha_k^cB\,\mathrm{lb}\left(1+\dfrac{p_{k,n}h_{k,n}}{N_0}\right)}},\forall n\in N_{\mathrm{UE}} \tag{41}$$

2.4　联合交替优化算法

根据终端发射功率、终端算力和终端模型压缩3个子问题的最优解，本文设计出了一种联合交替优化算法，通过3个子问题的交替求解，获得所提的优化问题P1的近似最优解。由于优化问题P1是一个非线性、非凸问题，通过交替优化算法可能求解出局部最优解，无法通过一次迭代求解出最优解。因此，本文算法通过随机的设置多个不同的初始点来获得多个局部最优解，基于这些局部最优解获得全局近似最优解。综上，终端自适应发射功率、算力与模型压缩算法见算法1。

算法1　终端自适应发射功率、算力与模型压缩算法

输入　模型的收敛速率ϑ_0，终端最大的发射功率、算力和模型压缩率$\left\{p_{\max},f_{\max},c_{\max}\right\}$

初始化　基站和终端设备数量K和N，外循环和内循环算法迭代次数I和$t_{\max}$以及算法精度ϵ_0

for $i=1{:}I$ **do**

　初始化$\left\{p_{k,n},f_{k,n},c_{k,n}\right\}$和迭代步数$t=1$

　while $\left|P_1(t)-P_1(t-1)\right|\geqslant\epsilon_0$ **and** $t\leqslant t_{\max}$ **do**

　　基于$f_{k,n}$和$c_{k,n}$，根据式(26)更新$p_{k,n}$；

　　基于$p_{k,n}$和$c_{k,n}$，根据式(34)更新$f_{k,n}$；

　　基于$p_{k,n}$和$f_{k,n}$，根据式(41)更新$c_{k,n}$；

　　基于$p_{k,n}$、$f_{k,n}$和$c_{k,n}$，更新$P_1(t+1)$；

　　$t=t+1$；

　end while

end for

输出　全局近似最优的终端发射功率、终端算力和终端模型压缩率$\left\{p_{k,n}^{\mathrm{opt}},f_{k,n}^{\mathrm{opt}},c_{k,n}^{\mathrm{opt}}\right\}$

算法1由内循环和外循环两个部分组成。内循环是初始化优化变量的情况下求解子问题的最优解，在内循环算法的第t次迭代过程中，通过依次

求解终端发射功率、终端算力和终端模型压缩优化子问题，在保证模型收敛速率的情况下，逐步减少联邦学习的时延和能耗，经过多轮的迭代使$\left|P_1(t)-P_1(t-1)\right|$逐渐地变小，使内循环算法收敛到稳定的解。外循环是通过多次求解局部最优解来获取全局解。假设内循环达到收敛所需要的步数为J，外循环所需要的次数为I。算法1的复杂度主要由内循环中求解优化子问题的复杂度决定，3个子问题的计算复杂度都与基站覆盖范围内终端设备的数量N有关，因此，3个子问题的复杂度可以分别表示为$O(Nt_1)$、$O(Nt_2)$和$O(Nt_3)$，其中，t_1、t_2和t_3表示计算时间是固定的常数。因此，算法1的计算复杂度可以表示为$O\left(IJ\left(Nt_1+Nt_2+Nt_3\right)\right)$。

3　仿真分析

3.1　仿真设置

本文的仿真基于Pytorch1.13版本的深度学习框架，设置了1个云端服务器连接4个边端基站服务器，每个边端基站服务器随机连接不同数量终端设备的三层联邦学习结构。在终端层，每个终端设备的最大发射功率$p_{\max}=28$ dBm，最大的终端设备算力$f_{\max}=2.4$ GHz，最大的终端模型压缩率$c_{\max}=50$，终端设备处理数据的CPU周期数$f_0=800$ cycles/bit，终端设备与芯片结构有关的有效转换容量$\xi=1.0\times10^{-27}$。终端设备与基站之间的无线信道符合的路径损耗模型为$128.1+37.6\lg d$，单位为dB，其中，d为终端设备与基站之间的距离，单位为km。每个边端基站的覆盖范围为300 m，每个终端设备被分配的无线信道带宽$B=1$ MHz，无线信道的噪声功率密度为−145 dBm/Hz，边端服务器的模型压缩率$c=10$，即模型收敛速率对应的模型压缩率。

在联邦学习过程中，边端聚合时终端设备模型训练的次数$m_1=2$，云端聚合时边端服务器聚合的次数$m_2=5$，模型信息交互所传输模型的参数数量$V=1.0\times10^6$，每个模型参数的量化位数$q=8$ bit。联邦学习的数据集使用服装数据（FMNIST, fashion modified national institute of standards and technology MNIST），该数据集包含T恤、牛仔裤、套衫、裙子、外套、凉鞋、衬衫、运动鞋、包和短靴10种数据类别，每种类别分别包含6 000张训练图像数据和1 000张测试图像数据，数据集总共包含70 000张图像数据，使用狄利克雷分布来模拟终端层终端设备数据的非独立同分布情况。联邦学习的训练模型采用经典的卷积神经网络，该神经网络的隐藏层包含两个卷积层、两个最大池化层和两个全连接层。

为了验证本文所提算法的性能，选择文献[24, 28-29]提出的联邦学习算法作为对比方案。其中，文献[29]提出了三层联邦学习算法，不进行任何优化方案，所有终端设备使用相同的发射功率和终端算力，并且不采用模型参数压缩策略，以下简称基准算法[29]；文献[24]不考虑终端设备发射功率和终端算力的影响，根据每个终端设备所处的动态网络环境自适应地选择模型参数压缩率，以下简称基准算法[24]；文献[28]不考虑模型参数压缩策略，综合考虑模型的训练精度和训练成本，自适应地选择终端设备最优的发射功率和终端算力，以下简称基准算法[28]。

3.2　仿真结果与分析

本节比较了本文算法与基准算法[29]、基准算法[24]和基准算法[28]在联邦学习模型训练的时延、能耗以及训练效果方面的表现，并且针对仿真结果进行了详细分析。其中，基准算法[24]仅针对联邦学习的训练时延进行优化；基准算法[28]针对联邦学习的训练时延和能耗同时进行优化。

当终端层存在100个终端设备时，一次云端聚合时延和能耗的对比如图2所示。从图2可知，在一次云端全局模型聚合过程中，本文算法产生最低的时延和能耗，与基准算法[29]传统的三层联邦学习算法相比较减少时延71.54%，并节约能耗48.76%。同时，基准算法[24]采用的模型压缩策略和基准算法[28]采用的终端侧资源优化策略均可以达到减少

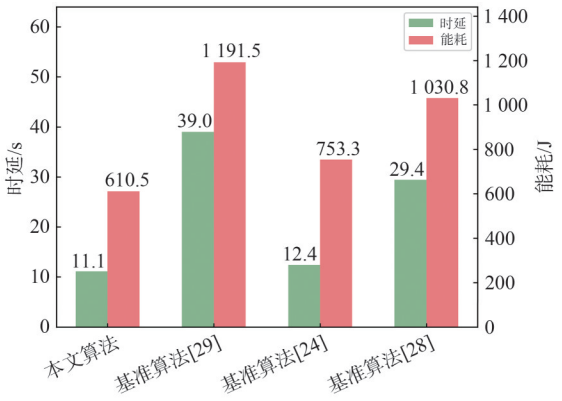

图2　一次云端聚合时延和能耗的对比

联邦学习模型训练时延和能耗的目的，其中模型压缩策略更优于终端侧的资源优化策略。本文算法与基准算法[24]和基准算法[28]相比，同时采用模型自适应压缩和终端资源优化两种策略，均衡优化模型训练产生的时延和能耗，在保证模型收敛速率的前提下，大大降低了联邦学习的时延和能耗。图2显示出在联邦学习高效训练和绿色节能方面，本文算法相较于其他对比算法具有较好的优势。

当终端层存在100个终端设备时，全局模型测试精确度随时延的变化如图3所示，全局模型测试精确度随能耗的变化如图4所示。从图3和图4可知，随着联邦学习模型训练时间和能耗的增加，4种算法的全局模型测试精确度逐渐增高，其中，本文算法的全局模型训练速度相对较快，可以在产生最少的训练时延和能耗下得到高精确度的全局模型。在图3中，尽管本文算法和基准算法[24]采用的模型压缩策略会造成全局模型收敛速率的下降，但是通过模型压缩缩减的训练时延充分弥补了模型压缩造成的模型训练收敛速率下降，与采用终端侧资源优化策略的基准算法[28]和传统三层联邦学习的基准算法[29]相比，可以在更短的时间内得到泛

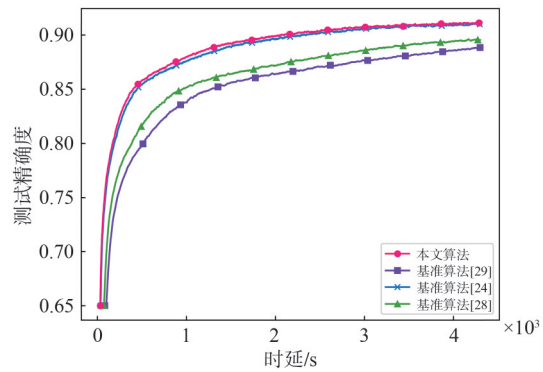

图3　全局模型测试精确度随时延的变化

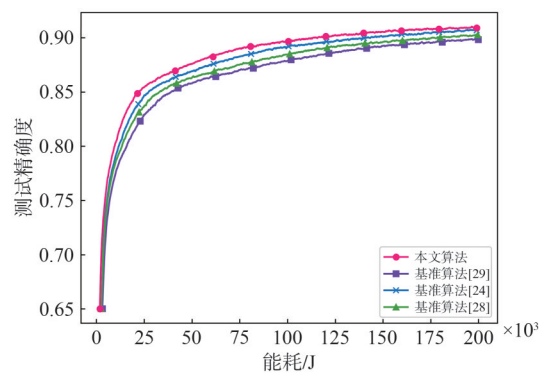

图4　全局模型测试精确度随能耗的变化

化性能良好的全局模型。在图4中，本文算法同样优于其他3种对比算法，这是由于本文算法通过终端侧资源优化策略对模型压缩收敛速率下降的缺陷进行了补偿，在联邦学习相同能耗情况下，本文算法相较于其他基准算法可以得到更高精确度的全局模型。

图3和图4同样展示了在联邦学习模型训练时，模型收敛速率的4种算法对比，云端全局模型测试精确度变化曲线的斜率可以反映出模型训练的收敛速率，即曲线斜率越大，模型训练收敛速率越高。从图3和图4中可以看出，4种算法的模型收敛速率随着训练时延和能耗的增加而逐渐降低，与其他算法相比，本文算法在模型训练阶段保持更高的模型收敛速率，使本文算法训练的全局模型可以得到更高的测试精确度，加快云端全局模型的训练进程。图3和图4体现了本文算法在模型训练方面的优势，可以在产生较低时延和能耗的情况下，训练出高测试精确度的云端全局模型。

当终端层存在不同数量的终端设备时，一次云端聚合的时延随终端设备数量的变化如图5所示，一次云端聚合的能耗随终端设备数量的变化如图6所示，可以观察到随着终端设备数量的增加，4种算法模型训练的时延和能耗均呈现出不同程度的上升趋势，其中，本文算法在不同的终端设备数量下都产生最低的时延和能耗。在图5中，终端设备的数据集大小服从非独立同分布，随着终端设备数量的增加，更容易出现差异性较大的终端数据集，这种较大的差异性导致了模型训练时延随着终端设备数量的增加而缓慢增加。在图6中，参与模型训练终端设备数量的增加直接导致联邦学习能耗的增加，导致了模型训练能耗随着终端设备数量的增加呈现快速增加的趋势。图5和图6表明了本文算法

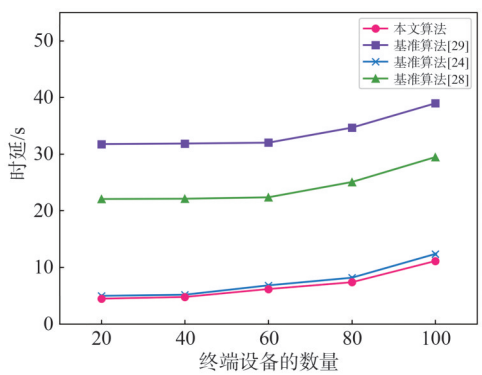

图5　一次云端聚合的时延随终端设备数量的变化

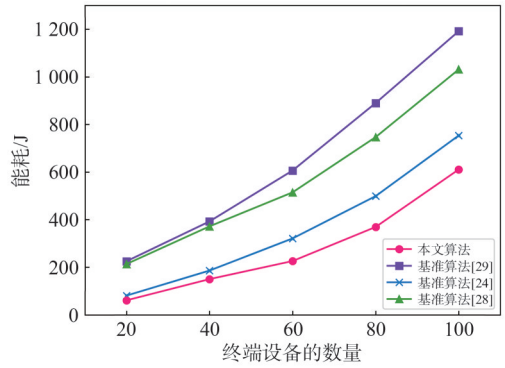

图6　一次云端聚合的能耗随终端设备数量的变化

的性能不受终端设备数量的影响，对于大规模不定数量终端设备的三层联邦学习场景具有良好的适用性。

4　结束语

为了应对资源受限和网络动态的边缘计算场景，降低联邦学习训练全局模型所产生的时延和能耗，本文基于云–边–端三层联邦学习架构提出了一种具有高效训练和绿色节能特性的联邦学习算法。本文综合分析了模型压缩策略对三层联邦学习模型训练时延和能耗的影响，建立了一定模型收敛速率下最小化联邦学习训练时延和能耗的优化问题，本文对终端设备的发射功率、终端算力和模型压缩率进行联合优化。接着，将优化问题分解为终端发射功率优化子问题、终端算力优化子问题和终端模型压缩优化子问题并分别求解，根据3个优化子问题的最优解，设计出了一种联合交替优化算法来获得原始问题的最优解。仿真结果表明，本文所提算法可以在保证全局模型收敛速率的情况下，显著地降低联邦学习模型训练的时延和能耗，且不降低全局模型的精确度。下一步研究工作将从终端层的终端设备选择方面出发，优化每轮次联邦聚合时终端模型的选择，进一步深入研究边缘计算中联邦学习的时延和能耗问题，提高联邦学习的模型训练效率。

参考文献：

[1] SHAHRAKI A, ABBASI M, PIRAN M J, et al. A comprehensive survey on 6G networks: applications, core services, enabling technologies, and future challenges[J]. arXiv preprint, 2021, arXiv: 2101.12475.

[2] 栾宁, 熊轲, 张煜, 等. 6G: 典型应用、关键技术与面临挑战[J]. 物联网学报, 2022, 6(1): 29-43.
LUAN N, XIONG K, ZHANG Y, et al. 6G: typical applications, key technologies and challenges[J]. Chinese Journal on Internet of Things, 2022, 6(1): 29-43.

[3] 李文璟, 喻鹏, 张平. 6G智能内生网络架构及关键技术分析[J]. 中兴通讯技术, 2023, 29(5): 2-8.
LI W J, YU P, ZHANG P. Analysis of the architecture and key technologies of 6G intelligent[J]. ZTE Technology Journal, 2023, 29(5): 2-8.

[4] 耿光磊, 高博, 熊轲, 等. 联邦学习赋能6G网络综述[J]. 物联网学报, 2023, 7(2): 50-66.
GENG G L, GAO B, XIONG K, et al. A survey of federated learning for 6G networks[J]. Chinese Journal on Internet of Things, 2023, 7(2): 50-66.

[5] AL-QURAAN M, MOHJAZI L, BARIAH L, et al. Edge-native intelligence for 6G communications driven by federated learning: a survey of trends and challenges[J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2023, 7(3): 957-979.

[6] XIAO Y, SHI G M, KRUNZ M. Towards ubiquitous AI in 6G with federated learning[J]. arXiv preprint, 2020, arXiv: 2004.13563.

[7] ZHU G X, LYU Z H, JIAO X, et al. Pushing AI to wireless network edge: an overview on integrated sensing, communication, and computation towards 6G[J]. Science China Information Sciences, 2023, 66(3): 130301.

[8] 刘胜利. 无线分布式学习系统的模型优化与资源管理[D]. 杭州: 浙江大学, 2022.
LIU S L. Model optimization and resource management of wireless distributed learning system[D]. Hangzhou: Zhejiang University, 2022.

[9] BOUZINIS P S, DIAMANTOULAKIS P D, KARAGIANNIDIS G K. Wireless federated learning (WFL) for 6G Networks⁴Part I: research challenges and future trends[J]. IEEE Communications Letters, 2022, 26(1): 3-7.

[10] YANG Z H, CHEN M Z, WONG K K, et al. Federated learning for 6G: applications, challenges, and opportunities[J]. Engineering, 2022, 8: 33-41.

[11] MCMAHAN H B, MOORE E, RAMAGE D, et al. Communication-efficient learning of deep networks from decentralized data[J]. arXiv preprint, 2016, arXiv: 1602.05629.

[12] LI T, SAHU A K, TALWALKAR A, et al. Federated learning: challenges, methods, and future directions[J]. IEEE Signal Processing Magazine, 2020, 37(3): 50-60.

[13] LIU Y, YUAN X L, XIONG Z H, et al. Federated learning for 6G communications: challenges, methods, and future directions[J]. China Communications, 2020, 17(9): 105-118.

[14] QIN Z J, LI G Y, YE H. Federated learning and wireless communications[J]. IEEE Wireless Communications, 2021, 28(5): 134-140.

[15] SHI W Q, ZHOU S, NIU Z S. Device scheduling with fast convergence for wireless federated learning[C]//Proceedings of the ICC

2020 - 2020 IEEE International Conference on Communications (ICC). Piscataway: IEEE Press, 2020: 1-6.

[16] LI Q B, DIAO Y Q, CHEN Q, et al. Federated learning on non-IID data silos: an experimental study[C]//Proceedings of the 2022 IEEE 38th International Conference on Data Engineering (ICDE). Piscataway: IEEE Press, 2022: 965-978.

[17] YAMASAKI Y, TAKASE H. F2MKD: fog-enabled federated learning with mutual knowledge distillation[C]//Proceedings of the 2023 IEEE 20th Consumer Communications & Networking Conference (CCNC). Piscataway: IEEE Press, 2023: 682-683.

[18] ALISTARH D, HOEFLER T, JOHANSSON M, et al. The convergence of sparsified gradient methods[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. New York: Curran Associates Inc, 2018: 5977-5987.

[19] SHI S H, WANG Q, ZHAO K Y, et al. A distributed synchronous SGD algorithm with global Top-*K* sparsification for low bandwidth networks[C]//Proceedings of the 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS). Piscataway: IEEE Press, 2019: 2238-2247.

[20] SATTLER F, WIEDEMANN S, MULLER K R, et al. Robust and communication-efficient federated learning from non-i.i.d. data[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 31(9): 3400-3413.

[21] JIANG Z D, XU Y, XU H L, et al. Adaptive control of client selection and gradient compression for efficient federated learning[J]. arXiv preprint, 2022, arXiv: 2212.09483.

[22] XU Y, LIAO Y M, XU H L, et al. Adaptive control of local updating and model compression for efficient federated learning[J]. IEEE Transactions on Mobile Computing, 2023, 22(10): 5675-5689.

[23] 唐伦，汪智平，蒲昊，等. 基于自适应梯度压缩的高效联邦学习通信机制研究[J]. 电子与信息学报，2023，45(1): 227-234.
TANG L, WANG Z P, PU H, et al. Research on efficient federated learning communication mechanism based on adaptive gradient compression[J]. Journal of Electronics & Information Technology, 2023, 45(1): 227-234.

[24] LIU S L, YU G D, YIN R, et al. Communication and computation efficient federated learning for Internet of vehicles with a constrained latency[J]. IEEE Transactions on Vehicular Technology, 2024, 73(1): 1038-1052.

[25] MO X P, XU J. Energy-efficient federated edge learning with joint communication and computation design[J]. Journal of Communications and Information Networks, 2021, 6(2): 110-124.

[26] TRAN N H, BAO W, ZOMAYA A, et al. Federated learning over wireless networks: optimization model design and analysis[C]//Proceedings of the IEEE INFOCOM 2019 - IEEE Conference on Computer Communications. Piscataway: IEEE Press, 2019: 1387-1395.

[27] LI P C, CHENG G L, HUANG X M, et al. Snowball: energy efficient and accurate federated learning with coarse-to-fine compression over heterogeneous wireless edge devices[J]. IEEE Transactions on Wireless Communications, 2023, 22(10): 6778-6792.

[28] HONG W, LUO X T, ZHAO Z Y, et al. Optimal design of hybrid federated and centralized learning in the mobile edge computing systems[C]//Proceedings of the 2021 IEEE International Conference on Communications Workshops (ICC Workshops). Piscataway: IEEE Press, 2021: 1-6.

[29] LIU L M, ZHANG J, SONG S H, et al. Client-edge-cloud hierarchical federated learning[C]//Proceedings of the ICC 2020 - 2020 IEEE International Conference on Communications (ICC). Piscataway: IEEE Press, 2020: 1-6.

[30] SHI S H, ZHAO K Y, WANG Q, et al. A convergence analysis of distributed SGD with communication-efficient gradient sparsification[C]//Proceedings of the 28th International Joint Conference on Artificial Intelligence. Menlo Park: AAAI Press, 2019: 3411-3417.

[31] LIU S L, YU G D, CHEN X F, et al. Joint user association and resource allocation for wireless hierarchical federated learning with Non-IID data[C]//Proceedings of ICC 2022 - IEEE International Conference on Communications. Piscataway: IEEE Press, 2022: 74-79.

[32] PAN Y J, CHEN M, YANG Z H, et al. Energy-efficient NOMA-based mobile edge computing offloading[J]. IEEE Communications Letters, 2019, 23(2): 310-313.

[33] TUN Y K, PARK Y M, TRAN N H, et al. Energy-efficient resource management in UAV-assisted mobile edge computing[J]. IEEE Communications Letters, 2021, 25(1): 249-253.

[作者简介]

朱光照（2000– ），男，南京邮电大学通信与信息工程学院硕士生，主要研究方向为边缘计算、联邦学习。

朱晓荣（1977– ），女，博士，南京邮电大学通信与信息工程学院教授、博士生导师，主要研究方向为5G/6G网络、智能物联网、网络大数据、区块链、群体智能等。

徐鼎（1983– ），男，博士，南京邮电大学通信与信息工程学院副教授、硕士生导师，主要研究方向为无线网络边缘计算、无线通信非正交多址接入、无线通信物理层安全、天地一体化网络、物联网等。