电池寿命效用感知的低轨卫星分层联邦学习资源分配方法

摘　要：针对低轨卫星联邦学习中星地通信窗口受限、光照周期变化导致训练效率与电池寿命难以兼顾的问题，提出一种电池寿命效用感知的分层联邦学习资源分配方法。首先，利用同轨卫星相对拓扑稳定的特点，以轨道内卫星构成客户端集群，通过星间参数共享与轨道内局部聚合减少星地参数交互和训练等待时间。其次，综合考虑卫星光照状态、能量消耗、放电深度及电池循环特性，构建电池寿命效用指标，并建立训练时间与寿命效用协同优化模型；进一步设计基于规则掩模的分布式近端策略优化算法，以筛除不可行动作并缓解聚合等待。基于 Walker 星座、CIFAR-10 和 EuroSat 数据集的仿真结果表明，所提方法在保证模型精度的同时，可有效降低电池寿命损耗并改善训练效率。

关键词：卫星网络；联邦学习；电池寿命；资源分配；强化学习

doi：10.11959/j.issn.2096−3750.

Battery lifetime utility-aware resource allocation for hierarchical federated learning in LEO satellite networks

Abstract: To address the problem that training efficiency and battery lifetime could not be easily balanced in LEO(low-Earth-orbit) satellite federated learning due to limited satellite-ground communication windows and illumination cycle variations, a battery-lifetime-utility-aware hierarchical federated learning resource allocation method was proposed. First, the relatively stable topology of satellites within the same orbit was exploited, client clusters were formed by satellites in each orbit, and satellite-ground parameter interactions and training waiting time were reduced through inter-satellite parameter sharing and intra-orbit local aggregation. Second, a battery lifetime utility metric was constructed by comprehensively considering satellite illumination states, energy consumption, depth of discharge, and battery cycle characteristics, and a joint optimization model of training time and lifetime utility was established. Furthermore, a rule-mask-based distributed proximal policy optimization algorithm was designed to filter out infeasible actions and alleviate aggregation waiting. It was shown by simulation results based on the Walker constellation, the CIFAR-10 dataset, and the EuroSat dataset that model accuracy is ensured while battery lifetime loss is effectively reduced and training efficiency is improved by the proposed method.

Key words: satellite network, federated learning, battery life, resource allocation, reinforcement learning

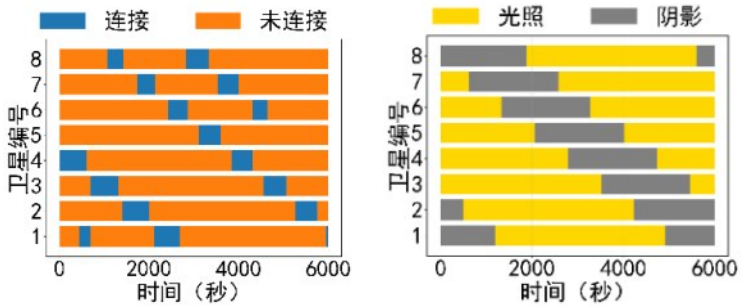
0　引言

低轨卫星在山火监测、土地分类等智能感知任务中产生了大量星上数据[1][2]。受星地链路容量和通信窗口限制，将大量原始数据持续传输至地面进行集中训练会产生较高的通信开销和传输时延。联邦学习通过在卫星本地完成模型训练、仅交换模型参数，可减少原始数据的星地传输需求，为低轨卫星协同智能训练提供新的实现方式[3][4]。然而，低轨卫星高速运动形成的间歇性连接以及周期性光照变化，使传统地面联邦学习机制难以直接适用于卫星网络。

首先，星地通信窗口受限与联邦学习的同步参数交换之间存在矛盾。低轨卫星处于高速运动状态，卫星与地面站之间的可见时间有限，且不同卫星与地面站的连接窗口并不同步。如图 1(a)所示，在 Walker 星座中，同一轨道内多颗卫星与地面站之间的通信窗口存在明显差异，导致客户端与参数服务器之间难以持续、稳定地进行模型参数交换。已有研究针对星地通信受限问题设计相应的联邦学习架构与聚合方法[5][6]，在一定程度上缓解了卫星联邦学习训练中的通信瓶颈。然而，现有方法对低轨卫星星座的轨道结构特征的利用仍不充分，星地参

收稿日期：XXXX−XX−XX；修回日期：XXXX−XX−XX

数交互频繁以及不同卫星训练进度不同所引起的聚合等待问题仍然存在。因此，如何减少星地参数交互和训练等待，是提升低轨卫星联邦学习效率需要解决的问题。



（a） 卫星的连接状态　　（b） 卫星的光照状态

图1　Walker星座卫星的连接和光照状态图

其次，卫星联邦学习中的资源使用还受到周期性能源供给和电池寿命的共同约束。低轨卫星的能源主要来自太阳能板和星载电池：光照期可利用太阳能板补充能量，阴影期则主要依赖电池供电[7]。由于卫星进入光照区和阴影区的时间不同步，如图1(b)所示，同一轨道内不同卫星在相同时刻可能处于不同光照状态。因此，在同一联邦学习轮次中，部分处于阴影期或低电量状态的卫星仍可能需要执行本地训练和参数传输，从而造成较深放电并加剧电池老化。由于星载电池难以在轨更换，其寿命直接影响卫星持续服务能力和星座运行周期[8]。已有研究主要通过降低星座总能耗[9-10]、减小放电深度或调整任务分配[7][8][11]等方式降低卫星电池寿命损耗，但能量消耗量难以反映不同放电深度对电池老化的非线性影响。以表1为例，方案2的放电量为方案1的2倍，但损耗仅为1.75倍；方案3的放电量为方案1的4倍，但损耗却超过10倍，表明不同放电深度下单位能量使用所对应的寿命代价存在显著差异。对卫星联邦学习而言，降低计算频率和传输功率虽然可以减少单轮能量消耗和电池损耗，却可能延长本地训练和参数传输时间；反之，提高资源投入又可能加剧电池放电。因此，如何根据卫星光照、电池状态和训练负载动态配置计算与通信资源，在联邦学习训练效率与电池寿命之间实现有效权衡，是本文需要解决的另一关键问题。

针对上述问题，本文充分结合卫星场景和联邦学习的特点，构建分层联邦学习架构，进一步研究电池寿命效用感知的资源分配问题。该问题受到卫星连接、光照、电池状态和训练负载等多种时变因素影响，且当前决策会通过电池状态演化影响后续资源配置。深度强化学习为此类动态优化提供了可行途径，但直接应用于大规模低轨卫星联邦学习仍面临以下挑战：一是卫星数量较多，集中式决策导致联合动作空间快速增长，增加策略学习难度[13]；二是各卫星资源决策共同影响轨道内局部聚合和全局聚合，完全独立决策难以兼顾整体优化目标[14]；三是资源配置受到放电深度、训练时间等约束，无约束探索容易产生大量低效或不可行动作，降低策略探索效率。

表1　电池不同放电深度与放电量和损耗对比[12]

方案	放电深度	放电量	归一化损耗
1	20%	总容量的20%	0.0735
2	40%	总容量的40%	0.1286
3	80%	总容量的80%	0.7979

因此，针对上述挑战，本文提出基于规则掩模的分布式近端策略优化算法DPPO-M（Distributed Proximal Policy Optimization with Masker），通过轨道级分布式策略执行与全局策略更新实现卫星间协同，并利用系统先验约束过滤不可行动作，以提高动态资源分配效率。

本文的主要贡献如下：

1) 设计了面向低轨卫星网络的分层联邦学习架构，充分利用同轨卫星相对拓扑稳定和星间链路可用于参数共享的特点，以轨道内卫星构成客户端集群，通过轨道内参数共享与局部聚合减少星地参数交互频次，降低通信窗口受限造成的聚合等待。

2) 构建了电池寿命效用感知的资源分配模型，综合考虑卫星光照状态、能量消耗、放电深度和电

池循环特性，建立卫星星座寿命效用指标，并通过动态分配计算与通信资源，实现联邦学习训练时间与电池寿命效用的协同优化。

3) 提出了基于规则掩模的分布式近端策略优化算法DPPO-M，针对卫星数量多、资源决策相互耦合以及约束条件下无效探索等问题，采用轨道级局部策略执行与全局策略更新机制，并结合放电深度、训练时间等系统约束设计动作掩模，同时引入掉队者惩罚机制，缓解轨道内和轨道间的聚合等待。

4) 基于CIFAR-10[①]和EuroSat[②]数据集进行仿真实验，并与代表性卫星联邦学习方法进行对比。实验结果表明，本文方法能够在保证模型精度的同时降低电池寿命损耗并改善训练效率；消融实验进一步验证了电池寿命效用指标和DPPO-M算法各模块的有效性。

1 相关工作

和地面联邦学习相比，卫星联邦学习存在通信环境严峻与卫星寿命有限这两大问题。相关综述如下。

1.1 星地通信优化相关研究

对于卫星联邦学习中通信环境严峻问题的研究主要分为架构与流程设计、通信过程优化这两方面。对于前一个方面，文献[5]采用分布式聚合方法，避免中心化聚合时中心卫星的通信拥堵问题。文献[6]设计同步与异步相结合的聚合方法，缓解卫星与地面服务器连接窗口短带来的等待时间问题。文献[9]动态选择领导者卫星作为参数服务器，避免了冗长的星地通信时间。文献[15]设计星间多次局部聚合的方式，减少卫星客户端与地面服务器通信次数，提高了训练速度。

在卫星联邦学习中通信过程优化方面，文献[16]对星地通信进行渐进式量化压缩，以适应卫星与地面站间通信窗口短暂的问题。文献[17]根据信号传播延迟以及预测流量负载进行数据分配和服务器选择，减少卫星通信的时间消耗。文献[18]采用智能体对带宽分配进行优化，最小化卫星通信延迟。文献[19]提出基于分解和元深度强化学习的资源分配算法，通过调节设备运行频率改善联邦学习速度。文献[20]在联合优化设备状态时，把时间作为约束条件以保证通信效率。

然而，由于大型星座中的LEO（low-Earth-orbit，低地球轨道）卫星数量多，通信转发、路由变化和通信窗口问题限制了通信效率。针对现有研究未能充分结合卫星网络特征与联邦学习过程的问题，本文综合借鉴先前研究，在以卫星轨道为局部群组进行分层聚合的基础上，结合流程中的时间差实施动态资源分配，有效应对卫星联邦学习中的通信瓶颈，实现更高效的卫星联邦学习。

1.2 卫星寿命优化相关研究

对卫星联邦学习中卫星寿命优化的研究主要分为降低能量消耗以及调整放电深度两方面。在降低能量消耗方面，文献[9]使用强化学习方法，动态求解配对问题和聚合卫星的选择问题，以降低能量消耗。文献[10]通过控制星间通信链路来降低通信能耗。文献[21]把能耗作为约束条件，求解卫星联邦学习中数据分配与服务器联合选择这两层优化问题。在调整放电深度方面，文献[7]通过研究卫星电池放电深度和放电次数之间的联系，将优化方向设定为最小化放电深度，以追求循环次数最大化。文献[8]基于循环次数计算卫星电池的损耗，以均衡各颗卫星的放电深度。文献[11]引用了文献[8]的电池模型，对卫星的初始任务量进行分配，以均衡任务量的方式降低寿命损耗。

以上研究中，前一方面研究虽然通过节约能量消耗降低了卫星电池寿命损耗，但是忽略了电池的放电深度的影响。后一方面的研究虽然降低了电池损耗，但都是通过减少或均衡卫星的任务负载来实现的。这些做法在优化卫星寿命的同时往往会对联邦学习的训练效率或模型性能产生不利影响。文献[22]表明，相比于电池的充放电次数，电池的总工作时间更能反映其寿命。本文引入卫星星座寿命效用指标，综合考虑卫星工作时间与电池寿命损耗，表征星座任务持续能力及其相应的寿命代价。

2 系统模型

本文考虑一个面向低轨卫星星座的分层联邦学习系统，其中，低轨卫星作为联邦学习客户端，地面站作为参数服务器。低轨卫星分布在多条环绕地

①https://www.cs.toronto.edu/%7Ekriz/cifar.html

②https://github.com/phelber/eurosat

球运行的轨道内，设包含O条轨道，每条轨道o内包含N_o颗卫星，$\mathcal{N}=\{\mathcal{N}_1,\cdots,\mathcal{N}_o,\cdots,\mathcal{N}_O\}$。将联邦过程分为若干轮次，记为$\tau=\{1,\cdots,r,\cdots,R\}$。系统通信包括同轨星间通信和星地通信。同轨卫星通过激光链路交换模型参数，本文不考虑跨轨道星间链路[15]。卫星仅在地面站通信窗口内进行星地参数传输。各地面站通过高速光纤互联，其间通信时延忽略不计，因此在架构中被视为一个整体[16]。

2.1　模型训练时间

模型训练时间主要包括联邦学习一个训练轮次内的全局参数广播、卫星本地训练、卫星轨道内局部聚合、星地参数传输、全局聚合阶段的时间。

(1)全局参数广播阶段：第r轮中，地面站与轨道o内卫星j之间的星地传输速率，以及卫星j与其相邻卫星之间的星间传输速率分别为$c_{o,j}^{UL}(r)$和$c_{o,j}^{ISL}(r)$：

$$c_{o,j}^{UL}(r)=B^{GSL}log_2\left(1+\frac{p^{UL}}{B^{GSL}\left(d_{o,j}^{GSL}(r)\right)^2}\gamma^{GSL}\right) \tag{1}$$

$$c_{o,j}^{ISL}(r)=B^{ISL}log_2\left(1+\frac{p_{o,j}^{ISL}(r)}{B^{ISL}\left(d_{o,j}^{ISL}\right)^2}\gamma^{ISL}\right) \tag{2}$$

其中，B^{GSL}和B^{ISL}分别表示星地微波链路和星间激光链路的通信带宽；p^{UL}和$p_{o,j}^{ISL}(r)$分别表示地面站发射功率和卫星星间传输功率；$d_{o,j}^{GSL}(r)$和$d_{o,j}^{ISL}$分别表示卫星与地面站、相邻卫星之间的距离。考虑到收发天线参数、噪声谱密度等均为预设值，将其分别归并为等效信道参数γ^{GSL}与γ^{ISL}，用于描述星地链路与星间链路的信道状态。

第r轮中地面站向轨道o内的卫星j发送全局参数的时间$t_{o,j}^{UL}(r)$表示为：

$$t_{o,j}^{UL}(r)=\frac{w}{c_{o,j}^{UL}(r)}+\frac{d_{o,j}^{GSL}(r)}{v^{light}} \tag{3}$$

其中，$\frac{w}{c_{o,j}^{UL}(r)}$是地面站向卫星传输模型参数的时间，$\frac{d_{o,j}^{GSL}(r)}{v^{light}}$是地面站与卫星通信的传播时间。$w$是模型参数大小，$v^{light}$是无线信号的传播速度。

第r轮中轨道o内卫星j通过轨道内共享获取全局参数的时间$t_{o,j}^{share}(r)$，表示为：

$$t_{o,j}^{share}(r)=\sum_{m=1}^{m_{o,j}(r)}\left\{\frac{w}{c_{o,j,m}^{ISL}(r)}+\frac{d_{o,j,m}^{ISL}}{v^{light}}\right\},j\in\mathcal{N}_o^{off}(r) \tag{4}$$

其中，$m_{o,j}(r)$表示卫星j获取参数所需的最大转发跳数，$\frac{w}{c_{o,j,m}^{ISL}(r)}$是传输模型参数的时间，$\frac{d_{o,j,m}^{ISL}}{v^{light}}$是传播时间，$c_{o,j,m}^{ISL}(r)$和$d_{o,j,m}^{ISL}$分别表示第$m$跳星间链路的数据传输速率和链路距离，$\mathcal{N}_o^{off}(r)$是需要等待转发的卫星的集合。

第r轮中轨道o内未连接地面站的卫星通过轨道内共享获取全局参数的时间$t_o^{share}(r)$，表示为：

$$t_o^{share}(r)=\max_{j\in\mathcal{N}_o^{off}(r)}\left\{t_{o,j}^{share}(r)\right\} \tag{5}$$

因此，第r轮中，地面站向轨道o内卫星j广播全局参数的时间$\mathrm{t}_{o,j}^{\mathrm{broc}}(r)$表示为：

$$t_{o,j}^{broc}(r)=t_o^{wait}(r)+t_{o,j}^{UL}(r)+t_o^{share}(r) \tag{6}$$

其中，$t_o^{wait}(r)$表示轨道o内无卫星处于地面站通信范围时产生的等待时间。

（2）卫星本地训练阶段：第r轮中，卫星j的计算时间$t_{o,j}^{comp}(r)$表示为：

$$t_{o,j}^{comp}(r)=\frac{\rho_{o,j}\chi_{o,j}(r)}{f_{o,j}^{comp}(r)} \tag{7}$$

其中，$\rho_{o,j}$表示处理单位比特数据所需的CPU周期数，$\chi_{o,j}(r)$表示卫星j在第r轮的本地训练数据量，$f_{o,j}^{comp}(r)$表示其计算频率，调整计算频率可调控本地计算所需的时间。

（3）卫星轨道内局部聚合阶段：第r轮中轨道o内局部聚合参数的时间$t_o^{RAR}(r)$表示为：

$$t_o^{RAR}(r)=\left(2N_o-2\right)\left(\frac{w}{N_oc_{o,j}^{ISL}(r)}+\frac{d_{o,j}^{ISL}}{v^{light}}\right)+t_{o,j}^{sum}(r)\left(N_o-1\right) \tag{8}$$

其中，公式（8）第一项表示通信时间，第二项表示计算时间。RingAllreduce将模型参数均分为N_o个数据块，共需$2\left(N_o-1\right)$次通信，每次通信时间包括数据传输时间$\frac{w}{N_oc_{o,j}^{ISL}(r)}$和链路传播时延$\frac{d_{o,j}^{ISL}}{v^{light}}$，单次计算时间是$t_{o,j}^{sum}$，共执行$\left(N_o-1\right)$次[15][23]。

（4）星地参数传输阶段：第r轮中轨道o内卫星j向地面站的发送速率$c_{o,j}^{DL}(r)$表示为：

$$c_{o,j}^{DL}(r)=B^{GSL}log_2\left(1+\frac{p_{o,j}^{DL}(r)}{B^{GSL}\left(d_{o,j}^{GSL}\right)^2}\gamma^{GSL}\right) \tag{9}$$

其中，$p_{o,j}^{DL}(r)$是卫星的传输功率，通过调节该传输功率，可以调控星地链路的数据传输时间。

第r轮中，轨道o内卫星向地面站并行发送局部聚合参数的时间表示为$t_o^{DL}(r)$，其主要由参数传输时延与链路传播时延之和最大的卫星决定：

$$t_o^{DL}(r)=\max_{j\in\mathcal{N}_o^{on}(r)}\left\{\frac{w}{c_{o,j}^{DL}(r)N_o^{on}(r)}+\frac{d_{o,j}^{GSL}(r)}{v^{light}}\right\} \tag{10}$$

其中，$\mathcal{N}_o^{on}(r)$表示处于地面站通信范围内的卫星集合，$N_o^{on}(r)$表示集合中的卫星数量。

（5）全局聚合阶段：第r轮中，地面站聚合全局参数的时间$t^{aggr}(r)$表示为：

$$t^{aggr}(r)=\frac{\mathrm{O}w\rho_{GS}}{f_{GS}^{comp}} \tag{11}$$

其中，f_{GS}^{comp}和ρ_{GS}分别表示地面站的计算频率和处理单位比特数据所需的CPU周期数。假设当地面站能源充足时，其始终以最大计算频率运行。

卫星联邦学习各阶段的时序关系如图2所示。综合考虑各阶段的计算与通信开销，整个卫星星座第r轮中的总训练时间$T(r)$表示为：

$$T(r)=\max_{o\in O}\left\{\max_{j\in\mathcal{N}_o}\left\{t_{o,j}^{broc}(r)+t_{o,j}^{comp}(r)\right\}+t_o^{RAR}(r)+t_o^{DL}\right\}+t^{aggr}(r) \tag{12}$$

由于采用同步分层聚合机制，轨道内局部聚合需等待该轨道所有卫星完成本地训练，全局聚合则需等待所有轨道完成局部聚合及星地参数传输。因此，整个星座的总训练时间由完成时间最长的轨道决定。

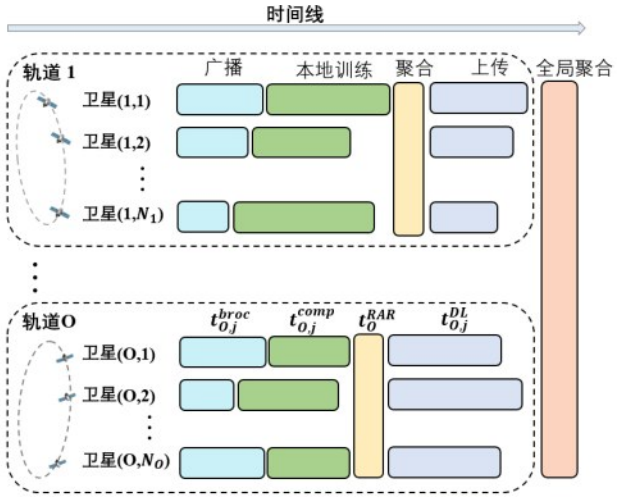


图2　联邦学习一个轮次中的时间步骤示意图

2.2　卫星电池寿命

(1)卫星的光照与能量补充：有光照时，假设阳光垂直照射时发电功率为p^{sun}，第r轮中光照方向与轨道o内卫星j轨道法线的夹角为$\partial_{o,j}(r)$，则卫星获取光照的功率$p_{o,j}^{supp}(\mathrm{r})$表示为：

$$p_{o,j}^{supp}(r)=p^{sun}cos\left(\partial_{o,j}(r)\right) \tag{13}$$

第r轮中轨道o内卫星j获取的能量$E_{o,j}^{supp}(\mathrm{r})$表示为：

$$E_{o,j}^{supp}(r)=\int_r^{r+1}p_{o,j}^{supp}(\mathrm{r})rdr \tag{14}$$

(2)卫星的能量消耗：第r轮中轨道o内卫星j的计算能耗$E_{o,j}^{comp}(r)$表示为：

$$E_{o,j}^{comp}(r)=\kappa\rho_{o,j}\chi_{o,j}(r)\left(f_{o,j}^{comp}(r)\right)^2 \tag{15}$$

其中，κ是能耗系数，$\chi_{o,j}(r)$，$\rho_{o,j}$，$f_{o,j}^{comp}(r)$与式（6）相同。

第r轮中轨道o内卫星j的星地通信能耗$E_{o,j}^{trans}(r)$表示为：

$$E_{o,j}^{trans}(r)=p_{o,j}^{trans}(r)t_{o,j}^{trans}(r) \tag{16}$$

其中，$p_{o,j}^{trans}(r)$是卫星的传输功率。

第r轮中轨道o内卫星j的星间通信能耗$E_{o,j}^{laser}(r)$表示为：

$$E_{o,j}^{laser}(r)=p_{o,j}^{laser}(r)t_{o,j}^{laser}(r) \tag{17}$$

其中，$p_{o,j}^{laser}(r)$和$t_{o,j}^{laser}(r)$分别是卫星间通信的功率和累计时间。

进而，第r轮中轨道o内卫星j消耗的总能量$E_{o,j}^{cons}(r)$表示为：

$$E_{o,j}^{cons}(r)=E_{o,j}^{comp}(r)+E_{o,j}^{trans}(r)+E_{o,j}^{laser}(r) \tag{18}$$

综上，第r轮中轨道o内卫星j的电池剩余电量$E_{o,j}^{batt}(r)$表示为：

$$E_{o,j}^{batt}(r)=\begin{cases}E_{o,j}^{max},E_{o,j}^{batt}(r-1)+E_{o,j}^{supp}(r)-E_{o,j}^{cons}(r)>E_{o,j}^{max}\\0,E_{o,j}^{batt}(r-1)+E_{o,j}^{supp}(r)-E_{o,j}^{cons}(r)<0\\E_{o,j}^{batt}(r-1)-E_{o,j}^{cons}(r)+E_{o,j}^{supp}(r),otherwise\end{cases} \tag{19}$$

其中，$E_{o,j}^{max}$是此卫星电池所能存储的最大能量。$E_{o,j}^{batt}(r-1)$是此卫星上一轮次的电池剩余电量。$E_{o,j}^{cons}(r)$和$E_{o,j}^{supp}(r)$分别是此卫星当前轮次的能量消耗量与能量补充量。

(3)卫星星座的电池寿命效用：放电深度表示电池已释放能量占其额定容量的比例。第r轮中，轨道o内卫星j的放电深度$d_{o,j}(r)$表示为：

$$d_{oj}(r) = 1 - \frac{E_{oj}^{batt}(r)}{E_{oj}^{max}} \times 100\% \tag{20}$$

根据文献[7][12]，在其它电池属性不变的情况下，电池报废前的可充放电循环次数$Y\left(d_{oj}\right)$主要取决于放电深度，可表示为：$Y\left(d_{oj}\right) = k_1\left(d_{oj} + k_2\right)^{k_3}$，其中，$k_1$、$k_2$、$k_3$是电池寿命模型的常数参数。基于此，本文将放电深度为d_{oj}时的单次循环寿命损耗定义为：$U\left(d_{oj}\right) = 1/\left(Y\left(d_{oj}\right)\right)$[12][24]。为衡量单位电池寿命损耗所对应的放电效益，定义第r轮中卫星j的电池寿命效用$h_{oj}(r)$为：

$$h_{oj}(r) = \frac{d_{oj}(r+1) - d_{oj}(r)}{U\left(d_{oj}(r+1)\right) - U\left(d_{oj}(r)\right)} \tag{21}$$

其中，$d_{oj}(r) \in [0,1]$表示卫星j当前放电深度。当$d_{oj}(r+1) > d_{oj}(r)$时，计入本轮放电引起的电池寿命损耗。$h_{oj}(r)$越大，表示单位电池寿命损耗所对应的放电深度增量越大，即电池使用效益越高。

为综合衡量卫星星座的电池使用效益，将所有卫星的电池寿命效用取平均，定义第r轮中的星座寿命效用$H(r)$为：

$$H(\mathrm{r}) = \frac{\sum_{o=1}^{O}\sum_{j=1}^{N_o} h_{oj}(r)}{N} \tag{22}$$

2.3　问题定义

为了增加卫星星座的电池寿命效用，同时减少卫星联邦学习过程的总训练时间，本文对卫星的计算频率与传输功率进行调节。本问题可以形式化地表述为优化问题P1：

$$\mathrm{P1}: \min_{\boldsymbol{F}^{\mathrm{comp}}(\boldsymbol{r}),\boldsymbol{P}^{\mathrm{trans}}(\boldsymbol{r})} \sum_{r=1}^{R}\left(-\omega_1 H(r)' + \omega_2 T(r)'\right) \tag{23}$$

s.t.

$$C1: d_{oj}(r) < d^{max}$$

$$C2: T(r) < T^{max}$$

$$C3: f_{oj}^{comp}(r) \in \left\{0, f_1^{comp}, f_2^{comp}, \cdots, f_{max}^{comp}\right\}$$

$$C4: p_{oj}^{trans}(r) \in \left\{0, p_1^{trans}, p_2^{trans} \cdots p_{max}^{trans}\right\}$$

$$C5: \eta > \eta_{min}$$

其中，ω_1和ω_2是两个优化目标的重要性权重，$\omega_1 + \omega_2 = 1$。$H(r)$是第r轮中的卫星星座寿命效用，$T(r)$是第r轮中的训练时间，归一化为$H(r)'$和$T(r)'$。决策变量$F^{comp}(r)$和$P^{trans}(r)$分别是第r轮中所有卫星的计算频率和传输功率。d^{max}是为了维持卫星待机所允许的最大放电深度。T^{max}是单个训练轮次所允许的最大时间。f_{max}^{comp}和p_{max}^{trans}分别是卫星的最大计算频率和传输功率。η是联邦学习任务的准确率，$\eta_{\min}$是为预设的最低可接受准确率阈值。C1将放电深度控制在合理范围内，以维持卫星基本运转。C2限制联邦学习单轮训练的最大时间。C3和C4表明卫星计算与通信功率的可调节范围。C5保证基本训练精度要求。

2.4　基于MDP的问题构建

在低轨卫星分层联邦学习过程中，资源分配决策受到持续变化的卫星通信连接状态、电池放电深度等多种因素影响，固定资源分配策略难以适应。同时，当前轮次的资源分配不仅影响本轮的结果，也会通过电池放电深度和剩余能量的变化影响后续轮次的资源分配条件。因此，将该资源分配问题转化为马尔可夫决策过程$(\mathcal{S}, \mathcal{A}, \mathcal{R})$以方便求解，以刻画动态环境下资源分配策略与系统状态演化之间的关系。其中$\mathcal{S}$表示状态空间，$\mathcal{A}$表示动作空间，$\mathcal{R}$表示奖励。具体地，

状态空间□：第r轮中，将状态空间$s(r) \in \mathcal{S}$表示为：

$$s(r) = \left[\boldsymbol{D}(\boldsymbol{r}), \boldsymbol{X}(\boldsymbol{r}), \boldsymbol{U}(\boldsymbol{r})\right] \tag{24}$$

其中，$\boldsymbol{D}(\boldsymbol{r}) = \left\{d_{1,1}(r), \cdots d_{oj}(r), \cdots, d_{O,N_o}(r)\right\}$（$\mathrm{o} \in \mathcal{O}, \mathrm{j} \in \mathcal{N}_o$，下同）表示卫星当前轮次的放电深度，$\boldsymbol{X}(\boldsymbol{r})$表示卫星当前轮次本地待训练数据量的大小，$\boldsymbol{U}(\boldsymbol{r})$,表示卫星是否处于地面站的直接通信范围。

动作空间□：第r轮中，将动作空间$a(r) \in \mathcal{A}$表示为：

$$a(r) = \left[\boldsymbol{F}^{\mathrm{comp}}(\boldsymbol{r}), \boldsymbol{P}^{\mathrm{trans}}(\boldsymbol{r})\right] \tag{25}$$

其中，$\boldsymbol{F}^{\mathrm{comp}}(\boldsymbol{r})$是卫星的计算频率，$\boldsymbol{P}^{\mathrm{trans}}(\boldsymbol{r})$是卫星的传输功率。

奖励函数$\mathcal{R}$表示为：

$$R(r) = \omega_1 H(r)' - \omega_2 T(r)' - \omega_3 \varpi(r)' \tag{26}$$

其中，$\varpi(r)'$表示归一化后的聚合时间差异惩罚项。用于惩罚联邦学习中的掉队者（Straggler，即训练或传输滞后的卫星）引起的聚合等待，其由轨道内惩罚和轨道间惩罚构成，定义为：

$$\varpi(r) = \sigma^{intra} + \sigma^{inter} \tag{27}$$

其中，$\sigma^{intra} = sqrt(\frac{1}{N_o}\sum_{j=1}^{N_o}\left(\mathrm{t}_{oj}^{broc}(r) + t_{oj}^{comp}(r) -\right.$

$\frac{\sum_{j=1}^{N_o}[t_{o,j}^{broc}(r)+t_{o,j}^{comp}(r)]}{N_o}\Bigg)^2$)表示同一轨道内各卫星本地训练完成时间的标准差，$\sigma^{inter}=sqrt(\sum_{o=1}^{O}\left(\max_{j_o\in\mathcal{N}}\left\{t_{o,j}^{broc}(r)+t_{o,j}^{comp}(r)\right\}+t_o^{RAR}(r)+t_o^{DL}+t^{aggr}(r)-T(r)\right)^2)$表示各轨道完成局部训练、局部聚合及星地传输的时间与全局训练完成时间之间的差异。因此掉队者惩罚机制$\varpi(r)$能引导策略缩小轨道内外的训练时间差异，减少同步聚合等待，其中权重ω_3控制时间差异惩罚项的作用强度，与训练时间权重ω_2共同调节时间优化。

3　方法设计

3.1　方法概述

针对上述训练时间与电池寿命效用协同优化问题，本文设计一种电池寿命效用感知的计算与通信资源分配方法。如图3所示，在分层的卫星联邦学习架构中，各轨道需要持续感知电池状态与训练状态，动态配置卫星的计算频率和传输功率，使资源分配能够适应卫星运行状态变化，并在联邦学习训练效率与电池寿命之间实现动态权衡。由于卫星状态和资源需求持续变化，本文将上述资源分配问题转化为马尔可夫决策过程，并提出DPPO-M算法求解。具体的MDP建模与DPPO-M算法分别在2.4节和3.2节介绍。

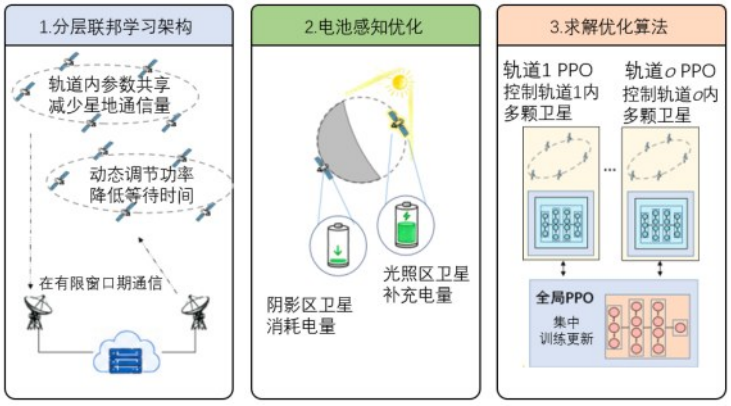


图3　本文主要方法示意图

3　2DPPO-M算法设计

PPO是一种基于策略梯度的强化学习算法，通过限制新旧策略之间的更新幅度，提高策略训练过程的稳定性[14]。相比之下，DDPG[25]和SAC[26]等算法主要采用离策略学习机制，在需要同时进行多个连续资源变量的动态控制场景下效果受限。考虑到本文资源分配问题受到卫星连接、电池状态和训练负载等动态因素影响，同时计算与通信资源决策受到放电深度、训练时间等约束，本文以PPO为基础，设计基于规则掩模的分布式近端策略优化算法DPPO-M。该算法采用轨道级分布式策略执行与全局策略更新方式，并利用系统先验约束筛除不可行动作，以提高策略探索效率[27][28]。DPPO-M算法框架如图4所示。

该框架由多个轨道级局部策略网络和1个全局PPO网络组成。每条轨道部署1个含Actor和Masker的局部策略网络：Actor根据当前状态生成计算与通信资源分配动作，Masker依据P1中的约束对候选动作进行可行性筛选。各局部策略网络执行动作后，将状态、动作、奖励和下一状态等交互轨迹存入轨迹缓存。全局PPO网络部署在地面站，由Actor、Critic和Masker组成，负责统一更新策略，并将更新后的Actor参数同步至各轨道局部策略网络。

每轮交互中，局部Actor根据状态$s(r)$输出候选动作分布，Masker根据放电深度和训练时间的约束，将不可行动作概率置为0，并对剩余可行动作重新归一化[27]：

$$M(s(r),a)=\begin{cases}1, d_{o,j}(r)<d^{\max}\&T(r)<T^{\max}\\ 0, else\end{cases}\tag{28}$$

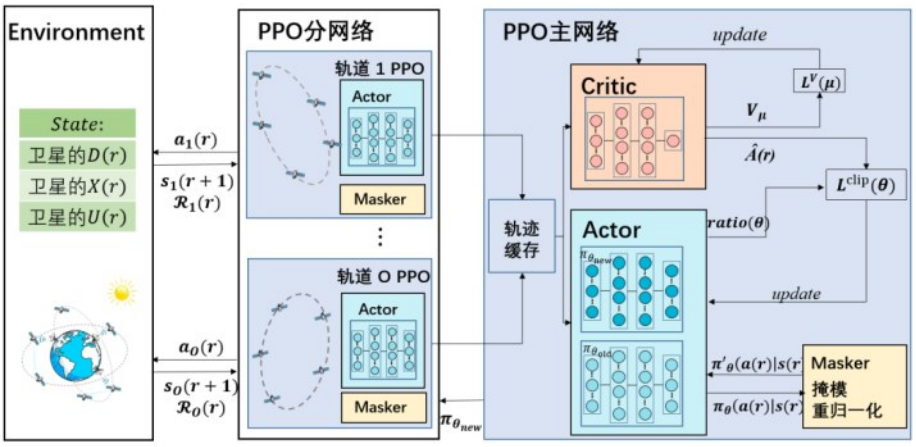


图4　DPPO-M 算法框架图

$$\tilde{\pi}_\theta(a|s(r)) = \frac{M(s(r),a)\pi_\theta(a|s(r))}{\sum_{a'\in\mathcal{A}} M(s(r),a')\pi_\theta(a'|s(r))} \quad (29)$$

其中$M(s(r),a)$为0代表动作不可行。局部Actor据此采样动作$a(r)$，配置所属轨道内卫星的计算频率和传输功率；环境返回奖励$\mathcal{R}$(r)和下一状态$s(r+1)$，轨迹存入缓存。

完成一批轨迹采集后，全局PPO网络读取各轨道交互信息，并按照PPO框架[14]进行参数更新。其中动作采样及策略概率比均采用式(29)的掩模后策略。Critic根据轨迹奖励和状态价值估计$V_\mu(s(r))$计算广义优势估计（GAE）$\hat{A}(r)$：

$$\delta(r) = R(r) + \gamma V_\mu(s(r+1)) - V_\mu(s(r)) \quad (30)$$

$$\hat{A}(r) = \sum_{l=0}^{R-r} (\gamma\lambda)^l \delta(r+l) \quad (31)$$

其中，$\delta(r)$为时序差分残差，γ为折扣因子，其中μ为Critic网络参数，λ为GAE参数，$R(r)$为式（24）定义的即时奖励值；终止状态时下一状态价值取0。构造目标价值$\hat{V}(r)$和价值损失函数$L^V(\mu)$：

$$\hat{V}(r) = \hat{A}(r) + V_{\mu_{old}}(s(r)) \quad (32)$$

$$L^V(\mu) = \hat{E}\left[\left[V_\mu(s(r)) - \hat{V}(r)\right]^2\right] \quad (33)$$

其中，$\hat{E}$表示对当前批次轨迹样本取均值。优势估计和目标价值均在参数更新前计算，并在当前批次内保持固定。Critic通过最小化$L^V(\mu)$更新参数μ。Actor基于优势估计和掩模后的策略概率比，按照文献[14]中的PPO截断目标函数$L^{clip}(\theta)$更新Actor网络参数θ。完成若干轮更新后，全局PPO网络将Actor参数同步至各轨道局部策略网络，并清空当前批次轨迹，进入下一轮。具体过程如算法1所示。

算法1:本文提出的DPPO-M 算法

输入:环境,状态空间$\mathcal{S}$,动作空间$\mathcal{A}$
输出:训练后的actor参数θ, critic参数μ
Initialize network parameter θ, μ
　For episode=1 to K do
　　For orbit=1 to O do/*各轨道分布式执行*/
　　　For round=1 to R do
　　　　Observe current state $s(r)$
　　　　Get action probability distribution $\pi_\theta(\cdot|s(r))$
　　　　Mask and normalization to get $\tilde{\pi}_\theta(a|s(r))$
　　　　Execute $a(r)$, get reward $R(r)$ and next state $s(r+1)$
　　　　Store trajectory in memory
　　　End for
　　End for
　　For epoch = 1 to Z do/*集中训练*/
　　　Sample from trajectory memory
　　　Get GAE estimator $\hat{A}(r)$ with critic V_μ
　　　Update the actor network parameter θ by optimizing $L^{clip}(\theta)$, Update the critic network parameter μ by optimizing $L^V(\mu)$
　　End for
　　Synchronize actor network parameterθ for all orbits, Clear trajectory memory
　End for

4　实验评估

4.1　实验设置

(1)仿真设定：

在实验中，本文使用STK仿真了由6个轨道共72颗LEO卫星组成的Walker星座与8座地面站（地面站位置选用STK内置的Beijing、Berlin、

Cape Town、Rio De Janeiro、Sydney、Toronto、Ohio、Bengaluru），获取卫星在2026年1月1日至9日的光照状态和与地面站连接状态。本实验的联邦学习与强化学习部分使用pytorch完成，仿真实验中的主要参数参考相同研究场景的文献[5]、文献[9]和文献[11]设置，如表2所示。

表2　主要参数设置表

Parameters	Values	Parameters	Values
卫星高度	500千米	CPU频率	5 GHz
星地最小通信仰角	5°	卫星向地面站发送数据的功率	50 W
卫星轨道周期	90分钟	芯片能耗系数	5E-28
卫星之间的通信带宽	1 GHz	通信综合系数	2000 kW/Hz
卫星向地面站发送数据带宽	10 MHz	卫星同轨激光通信链路的功率	0.5 W
地面站发送数据的带宽	50 MHz	电池容量	10 Wh

(2)模型与数据：

本文实验采用Resnet18模型，使用图像数据集CIFAR-10[29]与真实遥感数据集EuroSat[30]进行联邦学习训练。CIFAR-10包含60000张图片，其中50000张为训练集；EuroSat包含27000张真实遥感图片，其中21600张为训练集。考虑到卫星和地球的运动，卫星本地数据有明显的非独立同分布特征，因此采用Dirichlet（参数为0.3）分布划分数据，以模拟数据在卫星上的分布状态[31]。

4.2　实验结果

(1)　DPPO-M算法的收敛性能分析

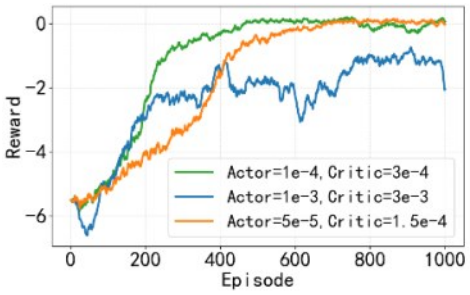


图5　不同学习率下DPPO-M的收敛曲线

图5给出了不同Actor和Critic学习率设置下DPPO-M的奖励收敛曲线。当Actor学习率为1e-4，Critic学习率为3e-4时，奖励值较快收敛，在250episode后逐渐稳定；在较大学习率下，DPPO-M的奖励值波动较为明显，且最终收敛水平较低；较小学习率则表现出较慢的收敛速度。

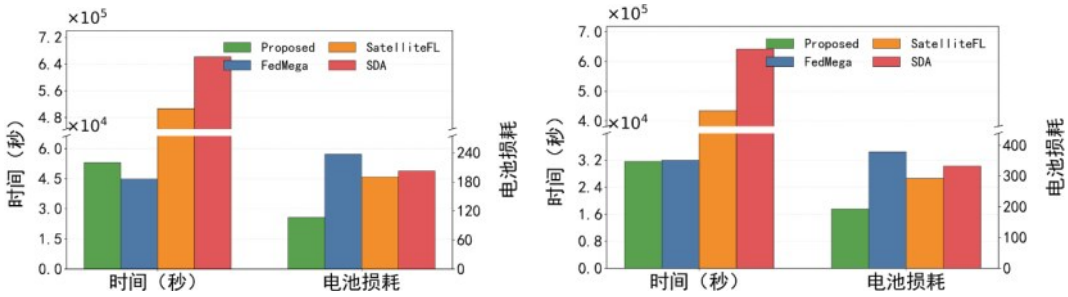
(2)　所提方法的综合性能评估

为评估所提方法的综合性能，选取以下3种方法作为对比：1)FedMega[15]：采用与本文相似的轨道内聚合架构，由地面站完成各轨道模型参数的全局聚合。为保证比较公平，实验中进一步根据卫星光照状态设置分档功率调整策略。2)SatelliteFL[16]：各卫星独立与地面站通信，在固定周期内选择可见卫星参与聚合，并根据通信窗口动态压缩模型参数以降低通信开销。3)SDA[32]：综合固定时间窗口与缓冲区容量设计动态聚合机制，由地面站根据连接状态确定聚合时机与参与卫星。

比较不同方法达到目标精度所需的训练时间与电池损耗，结果如图6所示。以CIFAR-10数据集为例，与FedMega相比，本文方法在训练时间增加18.3%的情况下，电池寿命损耗降低了54.98%；与SatelliteFL相比，训练时间和电池损耗分别降低了89.53%和43.84%。与SDA相比，本文方法的两项指标分别降低91.98%和47.31%。结果表明，本文方法能够在保持较高训练效率的同时显著降低电池寿命损耗，在训练时间与电池寿命之间取得较好的权衡。这是因为本文方法则根据当前综合状态动态优化资源分配，而FedMega缺乏动态优化，SatelliteFL和SDA的方法牺牲了部分精度，达到目标精度反而需要更大代价。

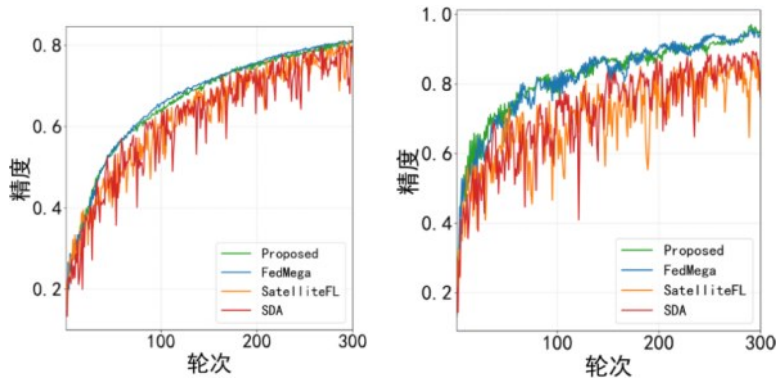
图7和图8进一步验证了以上分析，本文方法和FedMega的精度上升较稳定，而SatelliteFL和SDA存在明显波动。随着目标精度提高，各方法达到目标精度所需时间均增加，但本文方法和FedMega整体时间明显低于SatelliteFL和SDA。这是因为SatelliteFL和SDA每轮仅选择部分卫星参与聚合，增加了模型更新差异，从而使模型收敛产生较大波动并增加了耗时。

综合图6、图7和图8及以上分析可以得出，本文方法在电池损耗远低于其它三种方法的同时，准确率上升速度、联邦学习稳定性都达到了四种方法中的最优水平。



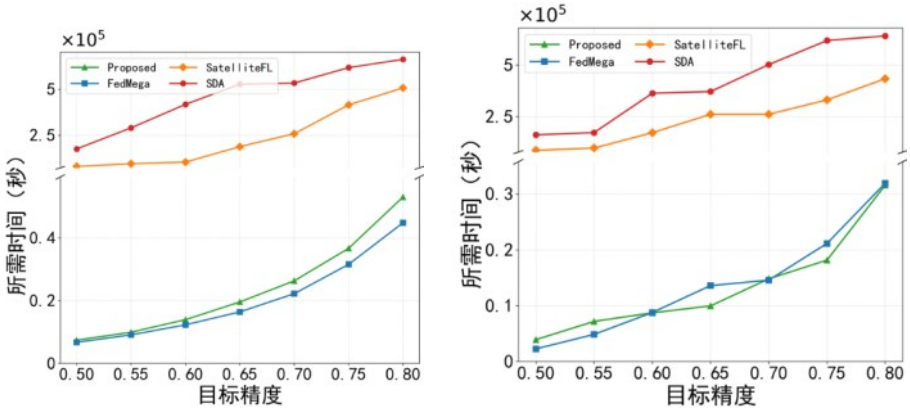
（a） CIFAR-10　　（b） EuroSat

图6　不同方法造成的时间和电池损耗



（a） CIFAR-10　　（b） EuroSat

图7　不同方法的模型精度随训练时间的变化



（a） CIFAR-10　　（b） EuroSat

图8　不同方法达到目标精度所需的训练时间

(3) 电池寿命效用指标的有效性分析

为验证本文电池寿命效用优化方案的效果，将本文寿命效用优化方案与最小化电池损耗(Degradation)、最小化放电深度(DoD)、最小化能量消耗(Energy)三种优化方案进行比较，结果如图9所示。

本文构建的寿命效用指标综合考虑不同放电深度下的任务执行收益与电池寿命代价，能够反映电池损耗随放电深度变化的非线性特征。与Degradation方案相比，本文方法的电池寿命损耗和训练时间分别降低0.06%和6.37%；与DoD方案相比，电池寿命损耗和训练时间分别降低13.33%和4.78%；与Energy方案相比，两项指标分别降低24.06%和3.06%。结果表明，相较于单独优化电池损耗、放电深度或能量消耗，所构建寿命效用指标能够在训练时间与电池寿命损耗之间取得更好的权衡。

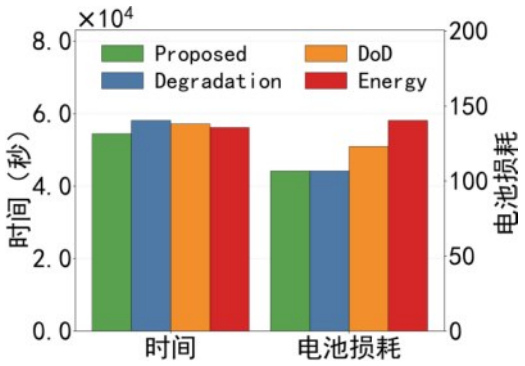
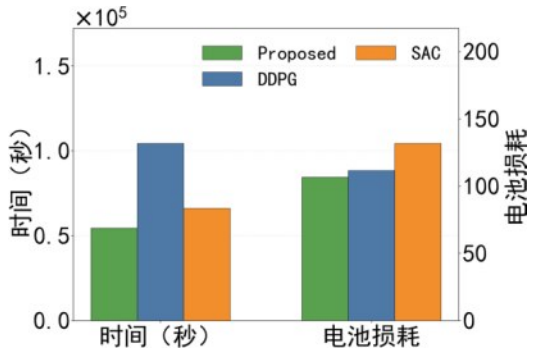


图9　电池寿命优化方案对时间和电池损耗的影响

(4) *求解算法的效果分析*

为验证本文求解算法的效果，设置了两组实验。

第一组将本文求解方法DPPO-M与DDPG、SAC对比，结果如图10所示。与SAC相比，本文方法的电池寿命损耗降低19.05%，训练时间降低17.69%；与DDPG相比，本文方法的电池寿命损耗降低4.43%，训练时间降低47.84%。这是因为DPPO-M拥有较低的决策维度、降低无效探索的掩模以及PPO相对稳定的策略更新。



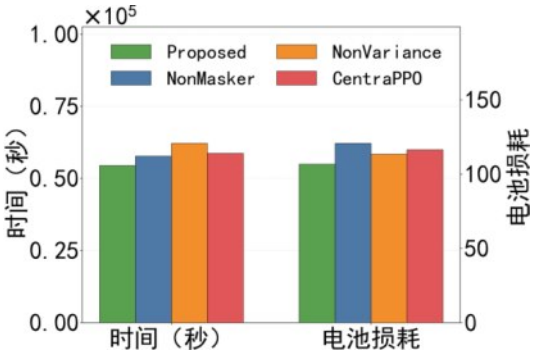


图10　对比传统强化学习方法的效果　图11不同算法模块的消融效果

另一组是对本文DPPO-M算法的各个模块的消融对比，包括去除规则掩模模块的NonMasker、移除聚合时间差异惩罚项（掉队者惩罚）模块的NonVariance、将轨道级网络替换为单一集中式PPO网络的CentraPPO，结果如图11所示。与NonMasker相比，本文方法的电池寿命损耗降低11.62%，训练时间降低5.67%；与NonVariance相比，本文方法的电池寿命损耗降低5.94%，训练时间降低12.43%；与CentraPPO相比，本文方法的电池寿命损耗降低8.36%，训练时间降低7.25%。这说明，规则掩模模块主要作用是优化了深度放电带来的电池损耗，聚合时间差异惩罚项主要作用是精准降低了训练时间，轨道级网络也对改善效果具有积极作用。

(5) *优化目标权重参数变化的影响分析*

为验证本文P1中卫星星座寿命效用的权重ω_1和训练时间的权重ω_2的变化影响，本文调整权重参数ω_1进行实验，结果如图12所示：

结果表明，以ω_1=0.5为基准，当ω_1降低到0.3时，训练时间降低10.94%，但电池寿命损耗增加16.69%。当ω_1增长到0.7时，训练时间增加11.18%，而电池寿命损耗降低2.33%。这是因为ω_1减小时，训练时间的相对权重增大，策略倾向于加快训练，但会增加电池寿命损耗；反之，ω_1增大时，策略更加注重电池寿命效用，通过减少深度放电降低寿命损耗，但需要付出训练时间增加的代价。

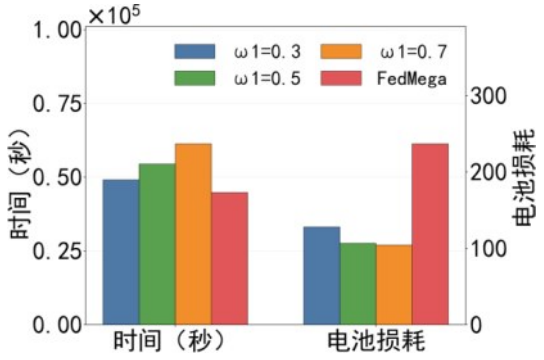


图12　优化目标权重变化的影响

5　结论

本文针对低轨卫星联邦学习中训练效率与电池寿命难以兼顾的问题，提出了一种电池寿命效用感知的分层联邦学习资源分配方法。利用低轨卫星相

对拓扑稳定的特点，构建了分层联邦学习架构，并综合考虑卫星光照与电池状态变化构建卫星寿命效用指标。将计算与通信资源分配问题建模为训练时间与寿命效用的协同优化问题，并采用基于规则掩模的分布式近端策略优化算法进行求解。实验表明，所提方法在保证模型精度的同时，有效降低电池寿命损耗，并在训练效率与电池寿命之间取得较好的权衡。

本文侧重于研究电池寿命效用指标在卫星联邦学习中的作用，对不同架构与任务场景的探讨不足。此外，本文研究中的电池模型缺乏实际数据的实验。未来可能会考虑更全面的卫星架构与场景，并引入NASA等机构公布的卫星电池数据集对模型进行验证，提升此方法在大规模低轨卫星网络中的适用性。

参考文献：

[1] YOU L, LI K, WANG J, et al. Massive MIMO Transmission for LEO satellite communications[J]. IEEE Journal on Selected Areas in Communications, 2020, 38(8):1851-1865.

[2] WANG S, LI Q. Satellite computing: vision and challenges[J]. IEEE Internet of Things Journal,2023,10(24):22514-22529.

[3] FARAJZADEH A, YADAV A, YANIKOMEROGLU H. Federated learning in NTN: design, architecture, and challenges [J]. IEEE Communications Magazine, 2025, 63(6): 26-33.

[4] MCMAHAN H B, MOORE E, RAMAGE D, et al. Communication-efficient learning of deep networks from decentralized data [C] // Proceedings of the 20th International Conference on Artificial Intelligence and Statistics(AISTATS). New York: PMLR, 2017: 1273-1282

[5] ZHAO L, GENG S, TANG X, et al. ALANINE: A novel decentralized personalized federated learning for heterogeneous LEO satellite constellation [J]. IEEE Transactions on Mobile Computing, 2025, 24(8): 6945-6960.

[6] LIN Z, CHEN Z, FANG Z, et al. FedSN: A federated learning framework over heterogeneous LEO satellite networks [J]. IEEE Transactions on Mobile Computing, 2025, 24(3): 1293-1307.

[7] LIU W, LAI Z, WU Q, et al. In-orbit processing or not? sunlight-aware task scheduling for energy-efficient space edge computing networks [C] // Proceedings of IEEE International Conference on Computer Communications (INFOCOM) 2024. Piscataway: IEEE, 2024: 881-890.

[8] YANG Y, XU M W, WANG D, et al. Towards energy-efficient routing in satellite networks [J]. IEEE Journal on Selected Areas in Communications, 2016, 34(12): 3869-3886.

[9] ZHANG H Y, ZHAO H B, LIU R K, et al. Leader federated learning optimization using deep reinforcement learning for distributed satellite edge intelligence [J]. IEEE Transactions on Services Computing, 2024, 17(5): 2544-2557.

[10] WANG G H, YANG F, SONG J, et al. Dynamic laser inter-satellite link scheduling based on federated reinforcement learning: An asynchronous hierarchical architecture [J]. IEEE Transactions on Wireless Communications, 2024, 23(10): 14273-14288.

[11] LEI C J, WU S H, YANG Y, et al. Joint partitioning, allocation, and transmission optimization for federated learning in satellite constellations via multi-task MARL [J]. IEEE Transactions on Mobile Computing, 2025, 24(10): 10345-10362.

[12] HAN X B, LU L G, ZHENG Y J, et al. A review on the key issues of the lithium ion battery degradation among the whole life cycle [J]. eTransportation, 2019, 1: 100005.

[13] LOWE R, WU Y, TAMAR A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments [C] // In Proceedings of the 31th Neural Information Processing Systems (NeurIPS). Massachusetts: MIT Press. 2017, 30: 6379-6390.

[14] CHAO Y, AKASH VELU, EUGENE VINITSKY , et al. The surprising effectiveness of PPO in cooperative multi-agent games. In Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS). Massachusetts: MIT Press. 2022: 24611 – 24624.

[15] SHI Y M, ZENG L, ZHU J Y, et al. Satellite federated edge learning: architecture design and convergence analysis [J]. IEEE Transactions on Wireless Communications, 2024, 23(10): 15212-15229.

[16] YANG C, YUAN J L, WU Y Z, et al. Communication-efficient satellite-ground federated learning through progressive weight quantization [J]. IEEE Transactions on Mobile Computing, 2024, 23(9): 8999-9011.

[17] HUANG C, CHEN G J, XIAO P, et al. Fair resource allocation for hierarchical federated edge learning in space-air-ground integrated networks via deep reinforcement learning with hybrid control [J]. IEEE Journal on Selected Areas in Communications, 2024, 42 (12): 3618-3631.

[18] HAN D J, FANG W, Hosseinalipour S, et al. Orchestrating federated learning in space-air-ground integrated networks: adaptive data offloading and seamless handover [J]. IEEE Journal on Selected Areas in Communications, 2024, 42(12): 3505-3520.

[19] ZHOU Y J, LEI L, ZHAO X, et al. Decomposition and meta-DRL based multi-objective optimization for asynchronous federated learning in 6G-satellite systems [J]. IEEE Journal on Selected Areas in Communications, 2024, 42(5): 1115-1129.

[20] WANG W D, LI P, LI S Q, et al. Optimizing proximity strategy for federated learning node selection in the space-air-ground information network for smart cities [J]. IEEE Internet of Things Journal, 2025, 12(6): 6418-6430.

[21] QIN P Y, XU D X, LIU L X, et al. Joint data allocation and LSTM-based server selection with parallelized federated learning in LEO satellite IoT networks [J]. IEEE Transactions on Network Science

and Engineering, 2024, 11(6): 6259-6271.

[22] OHRELIUS M, BERG M, WRELAND LINDSTROM R, et al. Lifetime limitations in multi-service battery energy storage systems [J]. Energies, 2023, 16(7): 3003.

[23] FANG Q, ZHAI Z, YU S, et al. Olive branch learning: A topology-aware federated learning framework for space-air-ground integrated network [J]. IEEE Transactions on Wireless Communications, 2023, 22(7): 4534-4551.

[24] FERREIRA M A M, PEREIRINHA P G, TROVAO J P F. Assessing lithium-ion battery aging in urban electric buses through rainflow-based cycle counting [J]. World Electric Vehicle Journal, 2026, 17(5): 245.

[25] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning [C] //International Conference on Learning Representations(ICLR).Open Publishing: OpenReview.net, 2016: 1-14.

[26] HAARNOJA T, ZHOU A, ABBEEL P, et al. Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor [C] //Proceedings of the 35th International Conference on Machine Learning(ICML). New York: PMLR, 2018, 80: 1861-1870.

[27] STOLZ R, KRASOWSKI H, THUMM J, et al. Excluding the irrelevant: focusing reinforcement learning through continuous action masking [C] // In Proceedings of the 38th Neural Information Processing Systems (NeurIPS). Massachusetts: MIT Press. 2024, 37.

[28] ACHIAM J, HELD D, TAMAR A, et al. Constrained policy optimization [C] //Proceedings of the 34th International Conference on Machine Learning. (ICML). New York: PMLR, 2017, 70: 22-31.

[29] KRIZHEVSKY A, HINTON G. Learning multiple layers of features from tiny images [R]. Toronto, Ontario: University of Toronto, 2009.

[30] HELBER P, BISCHKE B, DENGEL A, et al. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification [J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2019, 12(7): 2217-2226.

[31] KARIMIREDDY S P, KALE S, MOHRI M, et al. SCAFFOLD: Stochastic controlled averaging for federated learning [C] //Proceedings of the 37th International Conference on Machine Learning(ICML). New York: PMLR, 2020, 119: 5132-5143.

[32] LIN J, XU J, LI Y, et al. Federated learning with dynamic aggregation based on connection density at satellites and ground stations [C]. Proceedings of 2022 IEEE International Conference on Satellite Computing (Satellite), Piscataway: IEEE, 2022: 31-36,