基于深度强化学习的空天地一体化网络信息物理系统垂直切换策略

武艳[1]，潘广川[1]，姚明昨[1]，杨清海[1]，梁中明[2]
（1.西安电子科技大学空天地一体化综合业务网全国重点实验室，陕西 西安 710071；
2.深圳大学计算机与软件学院物联网研究中心，广东 深圳 518060）

摘　要：针对空天地一体化网络信息物理系统模型复杂、很难获得网络拓扑先验知识和模型化假设的特点，研究其基于深度强化学习的垂直切换策略。首先，综合考虑系统稳定性、切换开销和网络使用成本约束，将垂直切换策略问题建模为约束马尔可夫决策过程（CMDP），并给出保证可行解存在的充分条件；其次，提出约束–近端策略优化（CPPO）算法解决该问题，并在基站侧引入分布式强化学习机制加速训练收敛。相较于基准策略，仿真验证了所提垂直切换策略的优越性和有效性。

关键词：空天地一体化网络；信息物理系统；深度强化学习；垂直切换
中图分类号：TN929.5
文献标志码：A
DOI：10.11959/j.issn.1000–436x.2024140

Vertical handover policy for cyber-physical systems aided by SAGIN based on deep reinforcement learning

WU Yan[1], PAN Guangchuan[1], YAO Mingwu[1], YANG Qinghai[1], LEUNG Victor C.M.[2]
1. State Key Laboratory of Integrated Services Network, School of Telecommunications, Xidian University, Xi'an 710071, China
2. Internet of Things Research Center, School of Computer and Software, Shenzhen University, Shenzhen 518060, China

Abstract: The vertical handover policy of space-air-ground integrated cyber-physical systems based on deep reinforcement learning was studied, in which the challenges of complicated network model and difficulties in acquiring prior knowledge for network topology and model were addressed. By jointly taking the system stability, handover cost and network-using cost into account, the vertical handover policy problem was modeled as a constraint Markov decision process (CMDP), and a sufficient condition to ensure the existence of a feasible solution was derived. Furthermore, a constraint-proximal policy optimization (CPPO) algorithm was proposed to solve the CMDP, and also the distributed learning scheme at base station sides was introduced to accelerate the speed of converging. Simulation results verify the validation and superiority of the proposed vertical handover policy as compared with the baselines.

Keywords: space-air-ground integrated network, cyber-physical system, deep reinforcement learning, vertical handover

0　引言

信息物理系统（CPS, cyber-physical system）把制造过程中的资源、信息、物体以及人等生产要素紧密地联系在一起，将生产工厂变为一个智能环境，是实现工业4.0的重要基础[1]。CPS的实质是一个网络化控制系统，其主要部件包含一个动态受

收稿日期：2024–03–18；**修回日期：**2024–07–01
通信作者：姚明昨，mwyao@xidian.edu.cn
基金项目：国家重点研发计划基金资助项目（No.2020YFB1807700）；陕西省创新团队基金资助项目（No.2024RS-CXTD-01）
Foundation Items: The National Key Research and Development Program of China (No.2020YFB1807700), Innovation Capability Support Program of Shaanxi (No.2024RS-CXTD-01)

控对象、一个用于监测系统状态的传感器、一个能改变受控对象状态的执行器和一个生成控制决策的控制器，而通信网络负责发送受控对象状态信息到控制器，以及发送控制器的控制命令到执行器，所构成的“受控对象→传感器→控制器→执行器”控制闭环不断迭代促进了信息系统和物理系统的融合。近年来，随着制造业生产规模的不断扩大，面向新兴工业制造的CPS逐渐兴起，如智能工厂[2]、森林应急响应[3]、智能海上运输[4]等，CPS各部件在地理空间上的部署逐渐从传统陆地延伸至森林、海洋，甚至太空等地面网络难以覆盖的区域，因此需要引入一个综合移动通信网络实现广域范围设备随时随地接入，以保障CPS平稳运行。而空天地一体化网络（SAGIN, space-air-ground integrated network）将卫星网络、空基网络和地面网络无缝融合，可建立空天地多域多维度立体化的无缝通信覆盖，是承载新兴CPS业务的合适候选网络。

然而，将SAGIN作为连接CPS各部件的通信网络，实现空天地一体化网络信息物理系统的稳定运行面临诸多挑战，其中一个关键问题是空天地异构网络间垂直切换策略的选择。一方面，信息物理系统的本质属性要求SAGIN提供的服务必须稳定可靠，因此漫游在SAGIN内的CPS设备将对比空、天、地网络链路质量，选择最可靠的网络接入业务数据；另一方面，CPS设备及SAGIN中卫星和空中基站的移动会引起链路通信质量的高动态变化，由此将引起频繁切换，带来较大的切换开销（包含信令交互引起的时延和发送功率开销）。此外，垂直切换策略的选择还受限于CPS网络使用成本约束，如卫星使用价格通常高于地面网络价格，尽管卫星网络覆盖范围广，但为了降低费用，CPS用户可能会牺牲系统性能从而避免接入卫星网络。因此，在制定空天地一体化网络信息物理系统的垂直切换策略时，需要综合考虑CPS稳定性、切换开销和网络使用成本约束。

传统异构网络的垂直切换策略主要研究用户在无线局域网、Wi-Fi和蜂窝网络之间的连接转换。切换准则主要基于接收信号强度（RSS, received signal strength）、信干噪比（SINR, signal to interference plus noise ratio）、时延、抖动、丢包率、吞吐量等性能指标。文献[5]提出了一种保障端到端最小传输速率和最大容忍时延的分布式网络切片切换策略。文献[6]考虑包含宏基站、小基站和分簇用户的三层异构网络，基于RSS的切换策略提出了发生切换的理论边界和平均切换距离。文献[7]提出了一种异构的以用户为中心的分簇迁移方法，扩大了车联网络覆盖范围，减少了切换开销。然而，这些切换准则都不能直接用于空天地一体化网络信息物理系统的垂直切换，因为不同于传统异构网络，空天地一体化网络信息物理系统的首要目标是保障系统稳定运行，因此切换准则需要考虑控制系统的稳定而不是单纯表征网络服务能力的指标。

当前，针对空天地一体化网络信息物理系统切换策略的研究还很少，已有切换策略大多是针对如何提升空天地一体化网络的服务能力而提出。文献[8]提出了一种动态空天地一体化网络切换策略，利用软件定义传输控制解决控制器和切换节点之间信息传输瓶颈。文献[9]提出了一种空地协同计算的切换策略，其中空中基站与地面基站协同实现车辆分布式计算。此外，针对SAGIN网络模型复杂、很难获得网络拓扑先验知识和模型化假设的特点，近年来有学者提出利用深度学习方法制定网络服务策略[10]。文献[11]利用深度强化学习研究了最大化空天地一体化网络资源分配效率的多域资源优化策略。文献[12]针对空天存储容量受限问题，利用深度强化学习研究了一种任务调度策略，最大限度地减少了所有任务的卸载和计算时延。文献[13]研究了SAGIN中的多无人机合作任务卸载策略的优化目标是无人机选择合适的卸载目的地和任务数量，在无人机能量和卫星处理能力有限的情况下，使满足时延约束的任务数量最大化。然而，这些方法都不能直接用于空天地一体化网络信息物理系统的垂直切换，除了切换准则不同，当前利用深度强化学习方法解决SAGIN资源管理策略问题时，每种属性的权重都在训练前被确定，不能依据训练情况做自适应调整，因此无法满足约束需求，网络系统的可扩展性差。

综上所述，本文综合考虑CPS稳定性、切换开销和网络使用成本约束以及系统实现的可扩展性，研究基于深度强化学习的空天地一体化网络信息物理系统的跨域切换策略。具体而言，本文定义远程估计误差协方差的迹和切换开销作为空天地一体化网络信息物理系统的性能指标，以最小化该性能指标为目标、并将满足系统稳定性和网络使用成本约束的问题建模为约束马尔可夫决策过程（CMDP, constrained Markov decision process）问题。为解决

该问题，本文首先利用拉格朗日松弛技术将其转化为无约束问题，并提出了一种约束–近端策略优化（CPPO, constraint-proximal policy optimization）算法，该算法的惩罚系数可根据训练情况进行动态调整，使训练后的策略满足约束条件。更进一步，为加速训练收敛，本文提出了一种利用空、天、地基站实现分布式深度强化学习机制，该机制通过破坏样本间的相关性和降低样本方差，增强训练过程的稳定性，加快训练收敛速度。

1　系统模型

本文考虑的空天地一体化网络信息物理系统如图1所示。该系统包含需实时控制的动态受控对象、移动设备（MD, mobile device）（携带智能传感器和执行器）、负责生成控制决策信号的远程控制器，以及SAGIN连接MD和远程控制器与有线通信连接远程控制器和动态受控对象。具体来说，配备了智能传感器的MD可以测量并生成本地状态估计值，每个时隙它将本地状态估计值通过SAGIN发送到远程控制器；远程控制器负责生成远程状态估计值，并依据该估计值生成控制决策调节动态系统。其中，SAGIN垂直切换策略的选择会直接影响MD是否成功将本地状态估计值传给远程控制器，进而影响远程控制器对动态系统调节的精确度和系统稳定性。因此，为了提高受控系统性能，垂直切换策略倾向于选择更可靠的网络接入；然而，受限于切换开销和网络使用成本约束，垂直切换策略将尽量使用低成本网络接入并降低切换频次。为了阐明垂直切换策略与系统稳定性、切换开销和网络使用成本约束的关系，本节系统模型中将详细阐述动态受控对象模型、空天地一体化网络信息物理系统的稳定性、空天地一体化网络模型、切换开销和网络使用成本约束。

不失一般性，本文考虑使用旋翼无人机（UAV, unmanned aerial vehicles）作为空基网络的空中通信平台。由于UAV相较于地面基站具有更高的高度和更远的可视距离，因此可以认为其覆盖范围大于地面基站的覆盖范围。卫星网络的高度决定了其覆盖范围更广。此外，考虑MD同时具有卫星网络、空基网络和地面网络接口，它依靠网络切换策略来决定选择利用哪个网络进行数据传输。

1.1　动态受控对象模型

本文考虑一个按时隙工作的动态系统，其动态受控对象状态为$\boldsymbol{x}_t \in \mathbb{R}^n$，远程控制器输出的控制量为$\boldsymbol{u}_t \in \mathbb{R}^u$。将该动态系统建模为离散时间线性时不变系统，其状态演变如式(1)所示。

$$\boldsymbol{x}_{t+1} = \boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{B}\boldsymbol{u}_t + \boldsymbol{\omega}_t \tag{1}$$

其中，状态转移矩阵$\boldsymbol{A} \in \mathbb{R}^{n\times n}$，控制矩阵$\boldsymbol{B} \in \mathbb{R}^{n\times u}$为时不变参数，$\boldsymbol{\omega}_t \in \mathbb{R}^n$为每个时隙$t$中符合独立同分布的系统噪声，即$\boldsymbol{\omega}_t \sim \mathcal{N}\left(0,\boldsymbol{Q}_\omega\right)$。此外，由MD所携带的智能传感器采集到的动态受控对象测量值$\boldsymbol{y}_t \in \mathbb{R}^m$可表示为

$$\boldsymbol{y}_t = \boldsymbol{C}\boldsymbol{x}_t + \varsigma_t \tag{2}$$

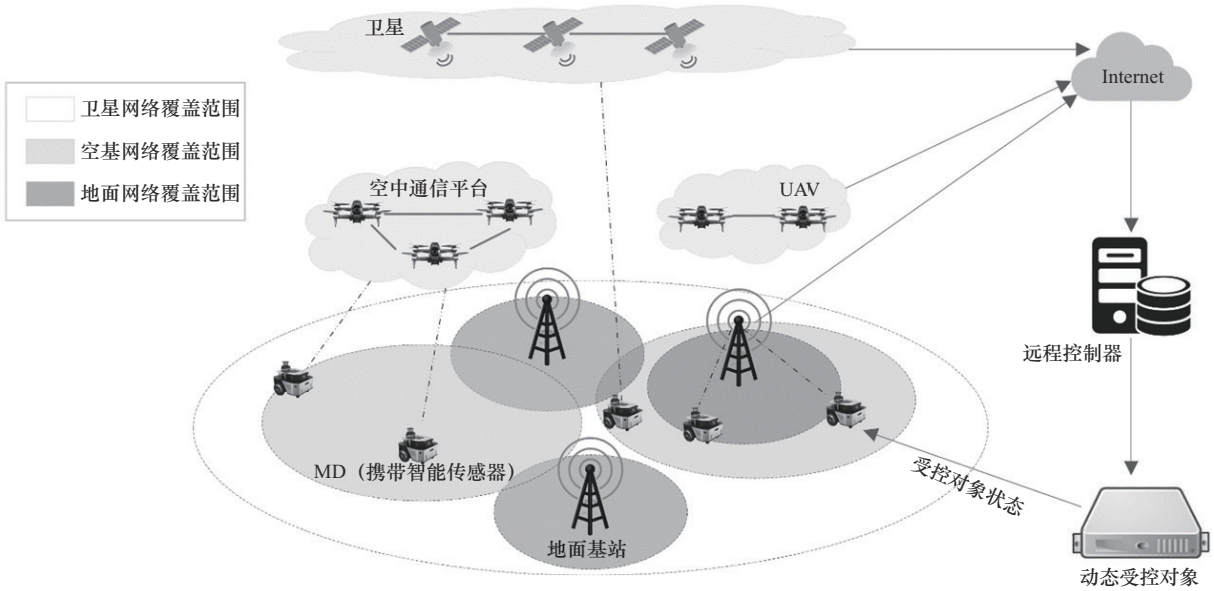


图1　空天地一体化网络信息物理系统

其中，测量矩阵 $\boldsymbol{C}\in\mathbb{R}^{m\times n}$ 为时不变参数，$\varsigma_t\in\mathbb{R}^m$ 为每个时隙 t 中具有独立同分布的测量噪声 $\varsigma_t\sim\mathcal{N}(0,\boldsymbol{Q}_\varsigma)$。$\boldsymbol{Y}_t=\{\boldsymbol{y}_1,\cdots,\boldsymbol{y}_t\}$ 为传感器从时隙 1 到时隙 t 采集到的所有测量数据。假设传感器能够运行本地卡尔曼滤波器[14]并估计系统的状态得到本地状态估计值 $\hat{\boldsymbol{x}}_{\text{local},t}$，表示为

$$\hat{\boldsymbol{x}}_{\text{local},t}=\mathbb{E}[\boldsymbol{x}_t|\boldsymbol{Y}_t] \tag{3}$$

相应的本地估计误差协方差为

$$\boldsymbol{V}_{\text{local},t}=\mathbb{E}[(\boldsymbol{x}_t-\hat{\boldsymbol{x}}_{\text{local},t})(\boldsymbol{x}_t-\hat{\boldsymbol{x}}_{\text{local},t})^{\mathrm{T}}|\boldsymbol{Y}_t] \tag{4}$$

假设 $(\boldsymbol{A},\boldsymbol{Q}_\omega)$ 和 $(\boldsymbol{A},\boldsymbol{B})$ 是可控的，$(\boldsymbol{A},\boldsymbol{C})$ 是可观察到的。令 $\mathcal{F}$ 和 $\mathcal{G}$ 分别是 Lyapunov 和 Riccati 算子，即 $\mathcal{F}(\boldsymbol{X})\triangleq\boldsymbol{A}\boldsymbol{X}\boldsymbol{A}^{\mathrm{T}}+\boldsymbol{Q}_\omega$，$\mathcal{G}(\boldsymbol{X})\triangleq\boldsymbol{X}-\boldsymbol{X}(\boldsymbol{Q}_\omega)^{\mathrm{T}}(\boldsymbol{Q}_\omega\boldsymbol{X}(\boldsymbol{Q}_\omega)^{\mathrm{T}}+\boldsymbol{Q}_\xi)^{-1}\boldsymbol{Q}_\omega\boldsymbol{X}$。$\mathcal{G}\circ\mathcal{F}(\bar{\boldsymbol{V}})=\bar{\boldsymbol{V}}$ 的唯一正半定解表示为 $\bar{\boldsymbol{V}}$，即稳态误差协方差。考虑到 $\boldsymbol{V}_{\text{local},t}$ 可以快速收敛到稳态[15]，本文假设本地卡尔曼滤波器在传感器侧已经进入稳态，即 $\boldsymbol{V}_{\text{local},t}=\bar{\boldsymbol{V}},t\geqslant 0$。

在远程控制器侧接收到的信息集表示为 $\boldsymbol{G}_t:=\{\gamma_0\hat{\boldsymbol{x}}_{\text{local},0},\cdots,\gamma_t\hat{\boldsymbol{x}}_{\text{local},,t}\}$，则远程状态估计值 $\tilde{\boldsymbol{x}}_t$ 表示为

$$\begin{aligned}\tilde{\boldsymbol{x}}_t&=\mathbb{E}[\boldsymbol{x}_t|\boldsymbol{G}_t]=\\&\gamma_t\hat{\boldsymbol{x}}_{\text{local},t}+(1-\gamma_t)(\boldsymbol{A}\tilde{\boldsymbol{x}}_{t-1}+\boldsymbol{B}\boldsymbol{u}_{t-1})\end{aligned} \tag{5}$$

其中，本文用 $\gamma_t\in\{0,1\}$ 作为每个时隙 t 的传输指示符。当本地状态估计值传输成功时，$\gamma_t=1$；否则，$\gamma_t=0$。式(5)表示，当本地状态估计值传输成功时，远程控制器根据 $\hat{\boldsymbol{x}}_{\text{local},t}$ 进行更新估计；否则，远程控制器根据式(1)进行预测。相应的远程估计误差协方差表示为

$$\begin{aligned}\boldsymbol{V}_t&=\mathbb{E}[(\boldsymbol{x}_t-\tilde{\boldsymbol{x}}_t)(\boldsymbol{x}_t-\tilde{\boldsymbol{x}}_t)^{\mathrm{T}}|\boldsymbol{G}_t]=\\&\gamma_t\boldsymbol{V}_{\text{local},t}+(1-\gamma_t)(\boldsymbol{A}\boldsymbol{V}_{t-1}\boldsymbol{A}^{\mathrm{T}}+\boldsymbol{Q}_\omega)=\\&\gamma_t\bar{\boldsymbol{V}}+(1-\gamma_t)\mathcal{F}(\boldsymbol{V}_{t-1})\end{aligned} \tag{6}$$

其中，$\gamma_t\bar{\boldsymbol{V}}+(1-\gamma_t)\mathcal{F}(\boldsymbol{V}_{t-1})$ 是由于 $\boldsymbol{V}_{\text{local},t}=\bar{\boldsymbol{V}},t\geqslant 0$。根据文献[16]中的引理3.1可知，1) 如果 $0\leqslant n_1\leqslant n_2$，则 $\mathcal{F}^{n_1}(\bar{\boldsymbol{V}})\leqslant\mathcal{F}^{n_2}(\bar{\boldsymbol{V}})$ 且 $\mathcal{F}(\bar{\boldsymbol{V}})\neq\mathcal{F}^0(\bar{\boldsymbol{V}})=\bar{\boldsymbol{V}}$；2) 对于任何 $n\geqslant 1$，$\mathrm{Tr}(\bar{\boldsymbol{V}})\leqslant\mathrm{Tr}(\mathcal{F}(\bar{\boldsymbol{V}}))\leqslant\mathrm{Tr}(\mathcal{F}^n(\bar{\boldsymbol{V}}))$，其中 $\mathrm{Tr}(\cdot)$ 是求矩阵迹。因此可以得出结论，远程估计误差协方差的迹可以当作衡量系统稳定性能的指标，如果迹越大，说明连续传输失败的次数越多，系统就越难保持稳定。此外，本文令 $\boldsymbol{V}_0=\bar{\boldsymbol{V}}$。

1.2　空天地一体化网络信息物理系统稳定性

对于空天地一体化网络信息物理系统，网络切换策略的主要目的是保证系统稳定。其系统稳定性标准遵循信息物理系统稳定性，定义如下。

定义 1[17]　如果动态系统的状态满足以下条件，则表明CPS是稳定的，即

$$\lim_{T\to\infty}\sup\frac{1}{T}\mathbb{E}\Big[\sum_{t=1}^{T}\|\boldsymbol{x}_t\|^2\Big]<\infty \tag{7}$$

定义1指出，CPS的系统稳定性取决于动态系统状态的极限平均能量。当 $\boldsymbol{x}_t$ 随时间收敛到稳定值时，它具有有限能量；当 $t\to\infty$ 且 $\boldsymbol{x}_t\to\infty$ 时，$\boldsymbol{x}_t$ 具有无限能量。换句话说，只要动态系统状态的极限平均能量有限，CPS系统就能保持稳定。

1.3　空天地一体化通信模型

根据香农容量定理，t 时隙天-地、空-地和地面通信模型分别为

$$\mathrm{Cap}^{\text{sat}}(t)\triangleq B^{\text{sat}}\mathrm{lb}\left(1+\frac{E^{\text{sat}}|h|^2}{P_N}\right) \tag{8}$$

$$\mathrm{Cap}^{\text{air}}(t)\triangleq B^{\text{air}}\mathrm{lb}\left(1+\frac{E^{\text{air}}\times 10^{-\frac{\mathrm{PL}^{\text{air}}(t)}{10}}}{P_N}\right) \tag{9}$$

$$\mathrm{Cap}^{\text{gr}}(t)\triangleq B^{\text{gr}}\mathrm{lb}\left(1+\frac{E^{\text{gr}}\times 10^{-\frac{\mathrm{PL}^{\text{gr}}(t)}{10}}}{P_N}\right) \tag{10}$$

其中，B^{sat}、B^{air} 和 B^{gr} 分别表示MD与卫星、UAV和地面基站之间链路的信道带宽，P_N 表示背景噪声的功率，E^{sat}、E^{air} 和 E^{gr} 分别表示MD对卫星、UAV和地面基站的发射功率，h 表示MD到卫星的信道增益。PL^{air} 和 PL^{gr} 分别表示空-地传输路径损耗和地面传输路径损耗。

不失一般性，本文假设MD与接入设备（包括地面基站、UAV和卫星）之间的信道容量大于或等于信道容量阈值 ϕ 时，其本地状态估计值能成功传输至远程控制器；否则，认为传输失败。记传输指标 $\gamma_t\in\{0,1\}$，表示为

$$\gamma_t=\begin{cases}1,\mathrm{Cap}^i(t)\geqslant\phi\\0,\mathrm{Cap}^i(t)<\phi\end{cases} \tag{11}$$

其中，$i=\{\text{sat},\text{air},\text{gr}\}$，本文考虑系统运行时间内大气状况相对稳定，卫星网络的通信容量固定不变且总是大于系统要求的信道容量阈值 ϕ，即如果MD选择连接卫星网络，则其传输一定成功。

1.4　切换开销

在SAGIN中，异构网络分布密集且相互重叠。MD的移动可能导致频繁切换或乒乓效应（反复来回切换），从而产生频繁的切换开销。这些切换开销主要体现在切换执行时的切换时延、信令开销和切换能耗上。本文假设在第 t 时隙产生的切换开销

为$j_{\mathrm{hc}}(t)$，表示为

$$j_{\mathrm{hc}}(t)=F_{\mathrm{hc}}\mathbb{I}(N_t\neq N_{t-1}) \tag{12}$$

其中，N_t表示系统在第t时隙使用的网络，$\mathbb{I}(\cdot)$表示指示算子，表示如果满足条件则等于1，否则等于0。F_{hc}表示切换的总体负面影响，包括切换时延、切换能耗和信令开销。

1.5 网络使用成本

网络使用成本是指网络运营商面向客户收取的网络使用费用，其具体定价是根据网络所需的技术成本、设备成本、维护成本等因素进行综合评估。通常情况下，卫星网络会因其自身高昂的制造、发射、维护和更新费用，导致其网络使用成本远远高于其他网络。而空基网络由于其需要先进的航空技术、自主飞行能力和能源系统，网络使用成本相对地面网络会较高。本文令$j_{\mathrm{mc}}(0)$，$j_{\mathrm{mc}}(1)$，$j_{\mathrm{mc}}(2)$和$j_{\mathrm{mc}}(3)$分别表示为空闲网络（未使用网络）、地面网络、空基网络和卫星网络的使用成本。假设$j_{\mathrm{mc}}(3)>j_{\mathrm{mc}}(2)>j_{\mathrm{mc}}(1)>j_{\mathrm{mc}}(0)=0$。其中$j_{\mathrm{mc}}(0)=0$是因为MD处于空闲状态，不与任何网络连接，所以不会产生网络使用成本。

2 空天地一体化网络信息物理系统垂直切换策略问题描述

本节同时考虑CPS稳定性、切换开销和网络使用成本约束，将空天地一体化网络信息物理系统垂直切换策略选择问题建模为CMDP问题，并给出保障系统稳定的充分条件。

2.1 CMDP问题建模

一般来说，一个马尔可夫决策过程（MDP, Markov decision process）可以用由4个对象组成的元组来描述，即状态空间$\mathcal{S}$、动作空间$\mathcal{A}$、转移概率函数Pr和奖励函数R。在同时考虑CPS系统稳定性、切换开销和网络使用成本约束的空天地一体化网络信息物理系统垂直切换策略问题中，4个对象描述如下。

1) 状态空间

状态空间s_t包含MD的通信状态和行为状态。

$$s_t=\{G_t,M_t\} \tag{13}$$

其中，$G_t=\{\mathrm{Cap}^{\mathrm{gr}}(t),\mathrm{Cap}^{\mathrm{air}}(t),\mathrm{Cap}^{\mathrm{sat}}(t)\}$表示MD在$t$时隙中观测到的通信状态，$M_t$表示$t$时隙MD的自身状态，包括速度$(v_x(t),v_y(t))$和位置$\big(X(t),Y(t)\big)$。

2) 动作空间

MD针对具体环境状态信息采取相应的动作$a_t\in\mathcal{A}$，动作空间$\mathcal{A}$表示为

$$\mathcal{A}=\{0,1,2,3\} \tag{14}$$

其中，a_t取0、1、2或3，分别表示MD在第t时隙切换至空闲状态、地面网络、空基网络或者卫星网络。当$a_t=a_{t-1}$时，意味着MD不做垂直切换动作，保持与原基站相连或做相应的水平切换。

3) 转移概率

MDP的转移概率由两部分组成，即通信状态转移和MD状态转移。由于两者不相关，因此在执行动作a_t后，从状态s_t到状态s_{t+1}的转移概率可以表示为

$$\begin{aligned}&\Pr(s_{t+1}|s_t,a_t)=\\&\Pr(G_{t+1}|G_t,a_t)\Pr(M_{t+1}|M_t,a_t)\end{aligned} \tag{15}$$

其中，G_t表示t时隙MD的通信状态，M_t表示t时隙MD的自身状态（位置和速度）。由于无线链路的高动态变化、MD移动位置、速度及通信环境的随机性，状态转移概率通常是未知的。

4) 奖励函数

本文定义MDP每一阶段的奖励函数为$R(s_t,a_t,s_{t+1})$，如式(16)所示。

$$\begin{aligned}&R(s_t,a_t,s_{t+1})=\\&\alpha_1\Gamma[\mathrm{Tr}(\boldsymbol{V}_t)]+\alpha_2\Gamma[j_{\mathrm{hc}}(t)]\end{aligned} \tag{16}$$

其中，奖励函数$R(s_t,a_t,s_{t+1})$是状态s_t、s_{t+1}和动作a_t的函数，为简化符号标识，可用R_t表示，$\mathrm{Tr}(\boldsymbol{V}_t)$是$t$时隙远程估计误差协方差的迹，即

$$\mathrm{Tr}(\boldsymbol{V}_t)=\gamma_t\bar{\boldsymbol{V}}+(1-\gamma_t)f(\boldsymbol{V}_{t-1})) \tag{17}$$

此外，$\Gamma[\cdot]$表示一个归一化函数，目的是统一远程估计误差协方差的迹和网络切换开销2种不同指标的量纲。这里出于对参数定义域区间的考虑，并为将“最小化开销”转化为“最大化奖励”，式(16)中的归一化函数分别将开销$\mathrm{Tr}(\boldsymbol{V}_t)$和$j_{\mathrm{hc}}(t)$从$[\mathrm{Tr}(\bar{\boldsymbol{V}}),V_{\max}]$和$[0,F_{\mathrm{hc}}]$映射到$[1,0]$，所采用的归一化方法分别基于sigmoid归一化和Min-Max归一化，具体定义为$\Gamma[\mathrm{Tr}(\boldsymbol{V}_t)]=\dfrac{2}{1+\mathrm{e}^{\mathrm{Tr}(\boldsymbol{V}_t)-\mathrm{Tr}(\bar{\boldsymbol{V}})}}$和$\Gamma[j_{\mathrm{hc}}(t)]=\dfrac{F-j_{\mathrm{hc}}(t)}{F}$。$j_{\mathrm{hc}}(t)$如式(12)所示表示切换开销。$\alpha_1$和$\alpha_2$是加权系数，满足$\alpha_1+\alpha_2=1$且$\alpha_1,\alpha_2>0$，其代表问题中2个指标的相对重要性。根据用户要求，可以设置不同的加权系数。

一方面，为保障空天地一体化网络信息物理系统稳定性，MD在移动过程中总希望切换到可靠性最高的网络，由此获取低的远程估计误差协方差进而提高系统稳定性；另一方面，频繁切换会带来额外的切换开销，而且是否切换到另一种网络还需要考虑网络使用成本约束。因此，本文同时考虑系统稳定性、切换开销和网络使用成本约束，通过最小化长远程估计误差协方差的迹和网络切换开销（与最大化奖励函数等效）保证系统稳定性和网络使用成本约束，将空天地一体化网络信息物理系统垂直切换问题建模为CMDP问题，具体如式(18)所示。

$$\begin{aligned}&\mathrm{P1}:\max_{\pi\in\Pi}\lim_{T\to\infty}\mathbb{E}^{\pi}\Big(\sum_{t=1}^{T}\lambda^{t}R_{t}\Big)\\&\text{s.t.}\lim_{T\to\infty}\frac{1}{T}\mathbb{E}^{\pi}\Big[\sum_{t=1}^{T}j_{\mathrm{mc}}(a_{t})\Big]\leqslant\sigma\\&\lim_{T\to\infty}\sup\frac{1}{T}\mathbb{E}\Big[\sum_{t=1}^{T}\left\|\boldsymbol{x}_{t}\right\|^{2}\Big]<\infty\end{aligned}\tag{18}$$

其中，$\sigma>0$定义为使用成本预算，表示平均网络使用成本的上限。$\lambda\in[0,1]$是折扣因子，表示当前时刻的决策对未来回报的重视程度。

考虑到问题P1可能在某些情况下无解。例如，当$\sigma\to 0$时，生成的切换策略将会一直保持空闲状态，导致系统不可能保持稳定。为了确保问题P1有解，接下来，本文将引入引理并分析预算和系统稳定的关系，给出保障系统稳定的充分条件。

引理1　[18]为了保证系统稳定，其成功传输的概率须满足

$$P(\gamma=1)>1-\frac{1}{(\max\{|\lambda_{i}|\})^{2}}\tag{19}$$

其中，λ_{i}表示$\boldsymbol{A}$的特征值。

基于引理1，本文得到了确保系统稳定的平均网络使用成本下限的定理1，该定理是保障空天地一体化网络信息物理系统稳定的充分条件。

定理1　设$\zeta=1-\frac{1}{(\max\{|\lambda_{i}|\})^{2}}$，如果平均网络使用成本满足

$$j_{\mathrm{mc}}(a_{t}=3)\zeta<\lim_{T\to\infty}\frac{1}{T}\mathbb{E}^{n}\Big[\sum_{t=1}^{T}j_{\mathrm{mc}}(a_{t})\Big]\tag{20}$$

此时，所有可行策略π能使系统保持稳定。

证明　令ξ表示网络接入场景，ξ^{*}表示其中一种特殊场景，MD没有可用的地面网络或空基网络，只有卫星网络覆盖。在ξ^{*}情况下，MD只能选择网络使用成本最高的卫星网络或者处于空闲状态的网络。另外，当网络使用成本为0时，MD将一直处于空闲状态，系统不能保持稳定；当不限制网络使用成本时，MD将一直选择卫星通信，系统能够保持稳定。因此，需要找到一个网络使用成本临界值μ，当平均网络使用成本大于μ时，能够确保系统稳定。

考虑到ξ^{*}场景下的临界值μ^{*}必然不小于μ。于是有

$$\begin{aligned}&\mu=\lim_{T\to\infty}\inf\frac{1}{T}\mathbb{E}^{\pi}\Big[\sum_{t=1}^{T}j_{\mathrm{mc}}(a_{t})\Big]\leqslant\mu^{*}=\\&\lim_{T\to\infty}\inf\frac{1}{T}\mathbb{E}^{\pi}\Big[\sum_{t=1}^{T}j_{\mathrm{mc}}(a_{t}=3)\mathbb{I}(a_{t}=3)\,\Big|\,\xi=\xi^{*}\Big]\end{aligned}\tag{21}$$

因此，在ξ^{*}场景下，为了保证系统稳定需满足

$$\begin{aligned}&P(\gamma=1|\xi=\xi^{*})=\\&\lim_{T\to\infty}\frac{1}{T}\mathbb{E}^{\pi}\Big[\sum_{t=1}^{T}P(\gamma=1|a_{t}=3,\xi=\xi^{*})\mathbb{I}(a_{t}=3)+\\&P(\gamma=1|a_{t}=2,\xi=\xi^{*})\mathbb{I}(a_{t}=2)+\\&P(\gamma=1|a_{t}=1,\xi=\xi^{*})\mathbb{I}(a_{t}=1)\Big]=\\&\lim_{T\to\infty}\frac{1}{T}\mathbb{E}^{\pi}\Big[\sum_{t=1}^{T}\mathbb{I}(a_{t}=3)|\xi=\xi^{*}\Big]>\\&\lim_{T\to\infty}\inf\frac{1}{T}\mathbb{E}^{\pi}\Big[\sum_{t=1}^{T}\mathbb{I}(a_{t}=3)\,\Big|\,\xi=\xi^{*}\Big]>\zeta\end{aligned}\tag{22}$$

因此，根据最后一个不等式可得

$$\begin{aligned}&\lim_{T\to\infty}\inf\frac{1}{T}\mathbb{E}^{\pi}\Big[\sum_{t=1}^{T}j_{\mathrm{mc}}(a_{t}=3)\mathbb{I}(a_{t}=3)\,\Big|\,\xi=\xi^{*}\Big]=\\&\mu^{*}>j_{\mathrm{mc}}(a_{t}=3)\zeta\end{aligned}\tag{23}$$

式(23)表明，在ξ^{*}场景下，只要平均网络使用成本大于$j_{\mathrm{mc}}(a_{t}=3)\zeta$，就一定能保证系统稳定。又因$\mu\leqslant\mu^{*}$，那么在任何场景下，只要平均网络使用成本大于$j_{\mathrm{mc}}(a_{t}=3)\zeta$，均能保证系统稳定。证毕。

因此，系统设计时，为保证问题P1有可行解，平均使用成本须满足式(20)。引入该条件后，问题P1转化为问题P2。

$$\begin{aligned}&\mathrm{P2}:\max_{\pi\in\Pi}\lim_{T\to\infty}\mathbb{E}^{\pi}\Big(\sum_{t=1}^{T}\lambda^{t}R_{t}\Big)\\&\text{s.t.}\ j_{\mathrm{mc}}(a_{t}=3)\zeta<\lim_{T\to\infty}\frac{1}{T}\mathbb{E}^{\pi}\Big[\sum_{t=1}^{T}j_{\mathrm{mc}}(a_{t})\Big]\leqslant\sigma\end{aligned}\tag{24}$$

问题P2是一个典型的CMDP问题，为了解决该问题，传统的办法是动态规划[19]，但是此类方法需要先验知识，如完整的转移概率。综上所述，由于无线链路的高动态变化、MD移动位置、速度及通信环境的随机性，状态转移概率通常是未知

的。因此，本文采用不需要先验知识的无模型深度强化学习方法解决问题P2。

3 基于分布式约束-近端策略优化的空天地一体化网络垂直切换策略

本文将利用约束型近端策略优化（PPO, proximal policy optimization）算法[20]来解决问题P2，得到空天地一体化垂直切换策略。首先利用拉格朗日松弛技术将CMDP转化成无约束MDP，进而利用近端策略优化算法更新策略网络和价值网络，同时自适应调节惩罚系数，最终找到问题P2的可行解。更进一步，本文引入空、天、地基站分布式学习机制，该机制能够提高训练稳定性并加速训练收敛。

3.1 基于约束-近端策略优化算法解决问题P2

本节利用拉格朗日松弛技术将含有约束条件的问题P2转换成无约束的拉格朗日函数$\mathcal{L}(\pi,\psi)$，具体如式(25)所示。

$$\mathcal{L}(\pi,\psi)=\lim_{T\to\infty}\mathbb{E}^{\pi}\Big(\sum_{t=1}^{T}R_t\Big)+\psi(\sigma-J^{\pi}) \tag{25}$$

其中，$J^{\pi}=\lim\limits_{T\to\infty}\dfrac{1}{T}\mathbb{E}^{\pi}\Big[\sum\limits_{t=1}^{T}j_{\mathrm{mc}}(a_t)\Big]$，$\psi$表示惩罚系数（即拉格朗日乘子）。

在CPPO算法设计中，采用PPO算法的目标函数$L_A(\theta)$来满足最大化预期累积奖励值要求，其具体表示为

$$L_A(\theta)=\frac{1}{T}\sum_{t=1}^{T}\big(\min\{\mathrm{rate}(t,\theta)A(t),\ \max\{\min\{\mathrm{rate}(t,\theta),1+\omega\},1-\omega\}A(t)\big) \tag{26}$$

其中，θ表示垂直切换策略网络参数，$\mathrm{rate}(t,\theta)$表示旧策略$\pi_{\theta_{\mathrm{old}}}(a_t|s_t)$和新策略$\pi_{\theta}(a_t|s_t)$之间的概率比，$\omega$表示裁剪超参数，$A(t)$表示优势函数。优势是指在状态$s_t$下MD采取特定动作$a_t$相对于平均水平的额外累计奖励值。为了估计优势，CPPO算法采用文献[20]中提出的广义优势估计法，优势函数$A(t)$可表示为

$$A(t)=\sum_{k=0}^{T-t}(\lambda)^k\left(R_{t+k}+\lambda V_{\varphi}(s_{t+k+1})-V_{\varphi}(s_{t+k})\right) \tag{27}$$

其中，λ表示折扣因子，$V_{\varphi}(s_t)$表示状态值函数。在基于策略优化的强化学习算法中，通常使用价值网络近似估计状态值函数，φ为价值网络的网络参数。为了获得准确的状态值函数估计，采用时差法计算价值网络的目标函数$L_C(\varphi)$，如式(28)所示。

$$L_C(\varphi)=\frac{1}{T}\sum_{t=1}^{T}\left(R_t+\lambda V_{\varphi}(s_{t+1})-V_{\varphi}(s_t)\right)^2 \tag{28}$$

通过最小化目标函数$L_C(\varphi)$和优化价值网络，使其获得的状态值函数估计更加准确，则式(27)可改为

$$\begin{aligned}&\mathcal{L}(\theta,\psi)=L_A(\theta)+\psi(\sigma-J^{\pi})=\\&\frac{1}{T}\sum_{t=1}^{T}(\min\{\mathrm{rate}(t,\theta)A^c(t),\max\{\min\\&\{\mathrm{rate}(t,\theta),1+\omega\},1-\omega\}A^c(t))+\psi\sigma=\\&(-L_{A^c}(\theta))+\psi\sigma\end{aligned} \tag{29}$$

其中，$A^c(t)=A(t)-\psi j_{\mathrm{mc}}(a_t)$。

考虑到网络使用成本的影响，目标函数$L_A(\theta)$可改为

$$\begin{aligned}L_{A^c}(\theta)=-\frac{1}{T}\sum_{t=1}^{T}&(\min\{\mathrm{rate}(t,\theta)A^c(t),\\&\max\{\min\{\mathrm{rate}(t,\theta),1+\omega\},1-\omega\}A^c(t))\end{aligned} \tag{30}$$

根据式(31)可知，通过更新参数θ和ψ以最大化$\mathcal{L}(\theta,\psi)$，从而获得最优垂直切换策略π_{θ}。给定$\psi(k)\geqslant 0$、$\theta(k)$和$\varphi(k)$分别表示第k次迭代后的惩罚系数、垂直切换策略网络参数和价值网络参数，具体迭代表达式如式(31)所示。

$$\begin{aligned}\theta(k+1)&=[\theta(k)+\eta_{\theta}\nabla_{\theta}(-L_{A^c}(\theta))]^{+}\\\psi(k+1)&=\psi(k)-\eta_{\psi}(\sigma-J^{\pi})\\\varphi(k+1)&=\varphi(k)-\eta_{\varphi}\nabla_{\varphi}(L_C(\varphi))\end{aligned} \tag{31}$$

其中，η_{θ}、η_{ψ}和η_{φ}分别为参数θ、φ和ψ的学习率。$[\cdot]_{\theta}^{+}$为投影算子，它通过投影到一个紧凑的凸集合来保持迭代稳定[21]。基于CPPO算法求解空天地一体化网络信息物理系统垂直切换策略如算法1所示。

算法1 基于CPPO算法求解空天地一体化网络信息物理系统垂直切换策略

定义训练回合数N，优化步骤数K，初始化缓冲区，学习率η_{θ}、η_{ψ}和η_{φ}，折扣因子λ，垂直切换策略网络参数θ，价值网络参数φ和惩罚系数ψ

1) for n=1:1:N
2) 执行垂直切换策略$\pi(\cdot|\cdot,\theta_{\mathrm{old}})$
3) 获得经验$(s_t,a_t,s_{t+1},R_t,j_{\mathrm{mc}}(a_t))$，并存入缓冲区
4) 计算$A^c(t)$并存入缓冲区
5) for k=1:1:K
6) 根据式(31)更新参数θ和ψ
7) end for

8)　$\psi(n+1) \leftarrow [\psi(n) - \eta_{\psi}(\sigma - J^{\pi})]$

9) end for

3.2　基于DCPPO的空天地一体化网络信息物理系统垂直切换策略

为了加速训练收敛并提高系统稳定性，本文将空、天、地基站分布式学习机制引入CPPO算法，形成分布式约束−近端策略优化（DCPPO, distributed constraint-proximal policy optimization）算法加速空天地一体化网络信息物理系统垂直切换策略收敛。具体来讲，通过引入一个可在基站侧（包括地面基站、空中平台和卫星）训练数据的分布式强化学习框架，将计算本地梯度的任务从MD端转移到基站，减轻了MD的计算负担。此外，每个基站同时连接多个MD，将这些MD收集到的数据储存到同一存储空间，并从中随机抽取一定量的经验计算本地梯度。该过程通过随机选择经验、破坏数据间相关性、减少样本方差来提高样本利用率，从而可提高系统稳定性并加速训练收敛。

基于DCPPO算法求解空天地一体化网络信息物理系统垂直切换策略的实现流程为。① 同时部署总数为W的MD，其初始时刻具有相同垂直切换策略网络和价值网络；② 基站在预设的时间窗口内等待MD将所收集的数据上传，并利用本地学习器计算本地梯度（包括垂直切换策略网络和价值网络的梯度值），然后将其上传至中央学习器；③ 待所有本地梯度到达后，中央学习器计算它们的平均值，导出全局梯度并更新网络（包括垂直切换策略网络和价值网络）；④中央学习器下发更新后的网络参数给每个MD，不断循环以上过程直至训练结束。

值得注意的是，DCPPO算法将计算本地梯度的任务从MD端转移到基站计算梯度加速收敛，为MD带来了额外的通信消耗。考虑MD通信开销与算法收敛时间的关系需要厘清算法迭代过程中经历的各个步骤，并逐一建模。因篇幅所限，本文在算法学习过程中尚未深入分析由MD通信开销引起的算法收敛性能变化。此外，MD与基站之间的无线链路有断开的可能，例如，如果某基站覆盖范围内参与分布式学习的某些MD无法将数据传送给该基站，则该基站中实际参与分布式学习的MD数目将减少，这会直接影响基站收集到的数据集质量，进而影响全局计算效率。

3.3　算法复杂度分析

深度强化学习的算法复杂度与选取的状态空间、动作空间、训练回合数、优化步骤数和网络结构相关。将训练一个网络参数的算法复杂度记为L，则该值与选取的状态空间和动作空间有关。本文所提基于CPPO求解空天地一体化网络信息物理系统垂直切换策略算法的策略网络共有M^s网络参数，价值网络共有M^v网络参数，训练回合数为N，优化步骤数为K，因此其复杂度为$\mathcal{O}(N \times K \times M^v \times M^s \times L)$。DCPPO算法是CPPO算法的分布式演进，利用多MD同时训练提高系统稳定性和加速收敛速度。DCPPO算法本质是CPPO算法的并行计算，故其复杂度也为$\mathcal{O}(N \times K \times M^v \times M^s \times L)$。

4　仿真分析

本节首先对本文所提的求解空天地一体化网络信息物理系统垂直切换策略的DCPPO算法的有效性和优越性进行了仿真实验，对比了惩罚系数动态更新与惩罚系数固定时DCPPO算法的稳定性、收敛速度和对平均网络使用成本的约束效果，并比较了CPPO算法、基于本地学习的DCPPO算法与本文基站侧分布式学习的DCPPO算法性能。其次，从切换效率和切换频率2个性能指标入手，将本文策略（确保系统稳定和满足网络使用成本约束）与贪婪策略进行了比较（最低接入成本和最可靠网络接入）。

4.1　仿真环境

仿真环境为半径为2 700 m的圆形区域，该圆形区域内包含9个地面基站和6架旋翼无人机，该区域内卫星通信全覆盖。仿真环境中放置8辆MD并利用随机航点模型[22]生成MD运动轨迹。

假设地面基站安装在高度为20 m的信号塔顶，UAV的悬置高度为90 m。设置地面网络和空基网络的带宽B为20 MHz，信号发射功率E为200 mW，噪声功率P_N为−130 dBm/Hz，载波频率f为2000MHz，卫星网络固定带宽为150 Mbit/s，信道容量阈值ϕ = 120 Mbit/s。以上3种网络的使用成本分别为$j_{\mathrm{mc}}(a_t = 1) = 0.3$，$j_{\mathrm{mc}}(a_t = 2) = 0.9$ 和 $j_{\mathrm{mc}}(a_t = 3) = 2.7$，切换开销$j_{\mathrm{hc}}(a_t) = 1$。MD移动速度在5~12 m/s之间取随机值。

动态受控对象参数设置如下。状态转移矩阵$\boldsymbol{A} = [1.1,0\,;\,1,0.8]$，控制矩阵$\boldsymbol{B} = [2\,;\,1]$，测量矩阵

$\boldsymbol{C}=[0.9,0;0,0.9]$，系统噪声的协方差$\boldsymbol{Q}_{\omega}=[0.3;0.3]$以及测量噪声的协方差$\boldsymbol{Q}_{\varsigma}=[0.2;0.2]$。通过转移矩阵可以求得$\zeta\approx 0.174$，为保证可行解的存在，考虑到$\sigma > j_{\mathrm{mc}}(a_t=3)\zeta = 0.469\,8$，本文设置网络使用成本预算$\sigma = 1.20$。

深度强化学习参数设置如下。设置策略网络为全连接网络，包含输入层、3个全连接隐藏层和一个输出层。输入层维度与状态特征数量相匹配。每个隐藏层包括512个神经元，使用ReLU作为激活函数。输出层包含4个神经元，对应4个可能的动作，使用Softmax作为输出层激活函数。价值网络设置与策略网络类似，不同在于输出层为一维。实验使用Xavier初始化权重和偏置，Adam优化器进行训练。惩罚系数ψ的初始值设置为0.1，学习率η_{ψ}设置为0.001。具体强化学习算法的参数设置如表1所示。

表1　　强化学习算法的参数设置

参数	设置值
η_{φ}	0.000 1
η_{θ}	0.000 1
λ	0.99
优化器	Adam
缓存区容量	5 000
训练回合数	4 000
α_1	0.7
α_2	0.3

4.2　DCPPO算法性能

求解空天地一体化网络信息物理系统垂直切换策略问题P2时，除了利用本文所提DCPPO算法，还可以使用双重深度Q网络（DDQN, double deep Q-network）算法[23]和约束策略优化（CPO, constrained policy optimization）算法[24]。此外，所提DCPPO算法的惩罚系数ψ可动态更新。因此，本节将所提DCPPO算法与DDQN算法、CPO算法和固定惩罚系数ψ的强化学习训练算法进行了对比，其中固定惩罚系数ψ分别设置为0.5和0.6。

图2展示了不同算法累积奖励值随训练回合的变化情况。可以发现，本文所提的ψ可动态更新的DCPPO算法在训练第200回合时开始收敛，且训练较其他算法更为稳定，DDQN算法收敛特性与之相当，CPO算法收敛特性次之，固定惩罚系数ψ算法收敛特性最差，在4 000回合内尚未收敛。

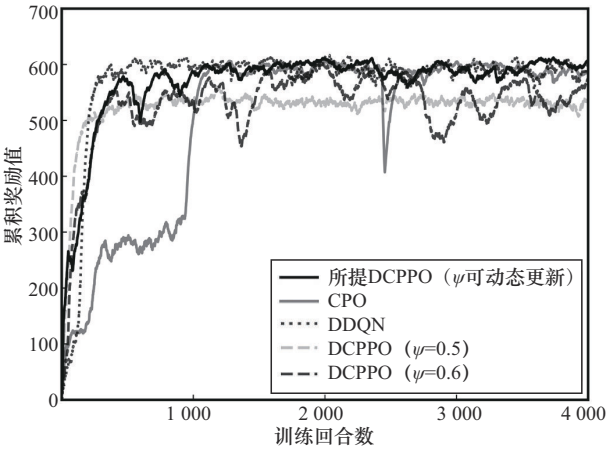


图2　不同算法累积奖励值随训练回合数的变化情况

图3展示了不同算法平均网络使用成本随训练回合数的变化情况，其中使用成本费用预算$\sigma = 1.20$（如4.1节所述该使用成本下系统稳定，且可行解存在）。可以看到，惩罚因子ψ可动态更新的DCPPO算法曲线能够紧贴使用成本费用预算曲线，而另外4条曲线的使用成本都远离使用成本费用预算曲线。ψ固定为0.5的DCPPO算法的平均使用成本稳定在2.7，DDQN算法平均使用成本约在1.5，均不满足约束。ψ为0.6的DCPPO算法和CPO算法虽然平均网络使用成本低于预算，但是联系图2可知，其收敛速度和稳定性较差。这是因为惩罚系数ψ需要合理选择，如固定为0.5的算法的惩罚系数设置太小，不能对训练起到有效约束，导致算法迅速收敛到只选卫星网络的策略，尽管收敛速度快、性能稳定，但不满足约束。设置为0.6，虽然费用预算成本有效约束，但是算法性能却变差。DDQN算法利用经验回放和固定Q目标技术虽然收敛特性表现良好，但因为本质上是固定的惩罚系数，导致在训练的过程中不能很好地控制约束力度；而CPO算法的约束控制效果虽然与DCPPO算法相当，但CPO算法的每一步都需要收集大量样本，以形成对状态值的准确估计，所以相对DCPPO算法更为耗时，且收敛速度慢。本文提出的ψ可动态更新的DCPPO算法会根据每回合计算的平均使用成本J^{π}微调惩罚系数，使得使用成本费用预算曲线更贴近基线，且训练过程更加稳定。

4.3　DCPPO分布式学习框架的优越性

为验证本文所提分布式学习框架带来的优越性，本节利用DCPPO算法来解决问题P2，并从稳定性、收敛速度和对平均使用成本的约束效果方面与用DPPO[25]、CPPO算法解决问题P2时进行对

比，如图4和图5所示。这里，CPPO算法未在策略训练中引入分布式学习。DPPO算法虽然使用了分布式学习框架，但其本地学习器部署于MD侧，与把本地学习器部署于基站侧的DCPPO相比，其所收集的数据相关性高，学习效率低。

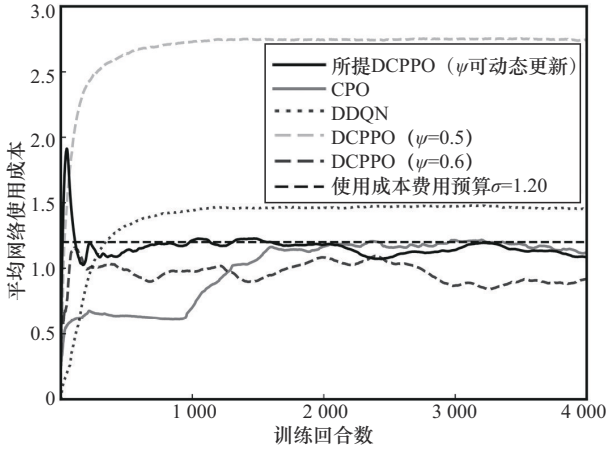


图3　不同算法平均网络使用成本随训练回合数的变化情况

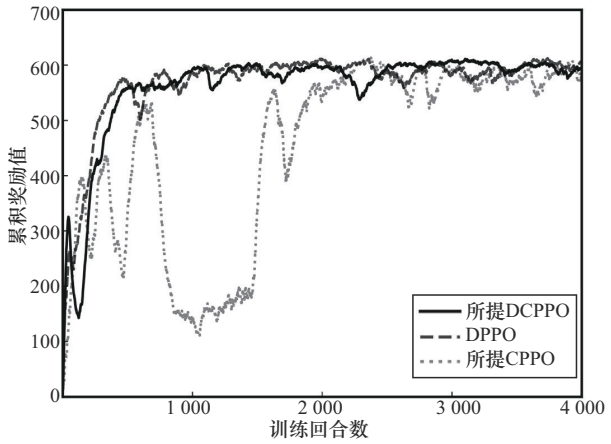


图4　累积奖励值随训练回合数的变化情况

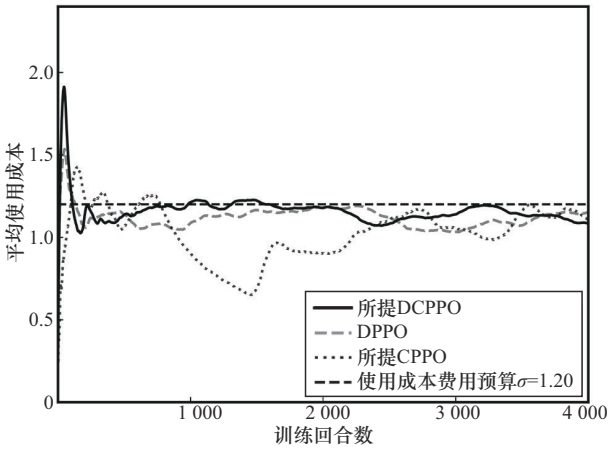


图5　平均网络使用成本随训练回合数的变化情况

图4展示了DCPPO、DPPO、CPPO算法解决问题P2，其累计奖励值随训练回合数的变化情况。可以看到，本文所提的分布式框架的算法DCPPO求解的垂直切换策略在大约第500轮训练回合时就能达到收敛，CPPO算法在大约第2 000轮训练回合时才达到收敛，DPPO算法在大约第800轮训练回合时收敛，可见本文所提框架对收敛速度的提升最大。图5展示了DCPPO、DPPO、CPPO算法求解问题P2时，其平均使用成本随训练回合数的变化情况。可以发现，虽然3种算法随着训练回合数的增加，最终都实现了对平均网络使用成本约束的紧密贴合，但本文所提DCPPO算法的收敛速度和稳定性均优于DPPO和CPPO算法。DCPPO算法所表现出的收敛速度快、收敛稳定和紧密贴合平均网络使用成本约束的优点是因为本文提出的分布式框架充分利用空天地一体化网络的特点，将本地学习器向上移动至基站。基站可以收集多个相互独立的MD的信息，这有利于数据经验去相关，提高学习效率。

4.4　基于DCPPO的空天地一体化网络信息物理系统的垂直切换策略性能验证

本节展示了本文所提空天地一体化网络信息物理系统垂直切换策略（即DCPPO算法且折扣因子$\lambda = 0.99$）与总选择当前时刻最低使用费用和最可靠网络接入的贪婪垂直切换策略，以及使用DCPPO算法折扣因子$\lambda = 0$和$\lambda = 0.9$时网络性能对比。性能对比如表2所示。这里，贪婪算法仅考虑当下短期回报，尽管从长期来看平均收益不高，但这类算法仅需根据当下的即时回报做出决策，无需对尚未发生的未来回报进行估计，计算轻便易于实施，因此，常被用作与CMDP算法对比的一个基准算法[26-27]。

首先定义2个度量指标，分别为切换频率和切换效率，据此衡量空天地一体化网络信息物理系统垂直切换策略的性能。

设MD采用特定策略在环境中共运行T时隙，则该策略的平均切换频率和平均切换效率分别为

$$F_h = \frac{N_h}{T} \tag{32}$$

$$\eta = \frac{\mathcal{T}}{\sigma} = \frac{1}{T\sigma}\sum_{t=1}^{T}(1-\beta_c)\mathrm{Cap}(t) \tag{33}$$

其中，N_h为切换发生次数，$\mathcal{T}$为有效吞吐量，σ为

使用成本，$\mathrm{Cap}(t)$为t时隙的信道容量，β_c为切换总时间。

表2展示了MD在移动速度区间下使用本文所提算法相较于对比算法的平均切换频率降低率（AHFRR, average vertical handover frequency reduction rate）和平均切换效率提升率（AHEIR, average vertical handover efficiency improvement rate）。其中，AHFRR定义为

$$\mathrm{AHFRR}=-\left(\frac{F_h^{\mathrm{DCPPO}(\lambda=0.99)}-F_h^{\mathrm{other}}}{F_h^{\mathrm{other}}}\right)\times 100\% \quad (34)$$

其中，$F_h^{\mathrm{DCPPO}(\lambda=0.99)}$和$F_h^{\mathrm{other}}$分别为本文所提算法和其他对比算法的平均切换频率。类似地，AHEIR定义为

$$\mathrm{AHEIR}=\frac{\eta^{\mathrm{DCPPO}(\lambda=0.99)}-\eta^{\mathrm{other}}}{\eta^{\mathrm{other}}}\times 100\% \quad (35)$$

其中，$\eta^{\mathrm{DCPPO}(\lambda=0.99)}$和$\eta^{\mathrm{other}}$分别为本文所提算法和其他对比算法的平均切换效率。这里，F_h^{other}和η^{other}为平均切换频率降低率和效率提升率的比较基准。

表2 垂直切换策略性能对比

垂直切换策略	AHFRR 5~8 m/s	AHFRR 9~12 m/s	AHEIR 5~8 m/s	AHEIR 9~12 m/s
DCPPO（$\lambda=0.9$）	+9%	+6%	+0.5%	+1.9%
DCPPO（$\lambda=0$）	+16%	+50%	+1.9%	+3.9%
贪婪算法（使用费用）	+16%	+18%	+1.2%	+3.2%
贪婪算法（可靠网络）	+9%	+20%	+7.1%	+9.6%

从表2可以看到，相较于对比策略，本文所提DCPPO（$\lambda=0.99$）具有较低的AHFRR和较高的AHEIR，DCPPO（$\lambda=0.9$）次之，DCPPO（$\lambda=0$）和另外2种贪婪算法性能较差。这主要是因为这两项指标都依赖于长期回报，较小的折扣因子和贪婪算法对未来奖励的重视程度较低，导致算法更倾向于关注即时奖励，而忽视了长期累积奖励的潜在价值。特别地，所提算法与DCPPO（$\lambda=0$）相比，在AHFRR和AHEIR上的提升非常显著。这是因为当$\lambda=0$时，算法不再考虑长期回报，但惩罚系数对其算法的约束效果并未改变。这样MD会在满足使用成本费用预算的条件下选择奖励最大的动作，导致其在局部区域会通过反复切换网络来实现约束使用成本费用预算的效果。因此，它的平均切换频率较高，移动速度越快性能越差。与DCPPO（$\lambda=0$）类似，另外2种贪婪算法仅在每一时刻寻找当下能用且费用最低或者最可靠的网络，当MD移动速度加快时，其迅速从一个网域移动到另一个网域，造成频繁切换，因此AHFRR性能较差。此外，这2种贪婪算法并未考虑使用费用约束，也就是说其垂直切换策略选择不计成本，因此AHEIR性能也差。

5 结束语

为保障空天地一体化网络信息物理系统的稳定运行，本文综合考虑系统稳定性、切换开销、网络使用成本约束以及空天地一体化网络结构的复杂性，研究了基于深度强化学习的空天地一体化网络信息物理系统垂直切换策略。首先将最小化远程估计误差协方差的迹和切换开销作为目标、并将满足系统稳定性和网络使用成本约束的问题建模为CMDP问题，提出一种DCPPO算法解决了该问题。该算法通过动态调整惩罚系数，达到了约束策略训练过程的目的，并利用空、天、地基站分布式强化学习机制加速训练收敛。仿真结果表明，本文策略具有很强的稳定性和收敛性，且与基准策略相比，本文所提策略具有更低的切换频率和更高的平均吞吐量。

参考文献：

[1] 陈明, 梁乃民. 智能制造之路: 数字化工厂[M]. 北京: 机械工业出版社, 2022.
CHEN M, LIANG N M. The road to intelligent manufacturing-digital factory[M]. Beijing: Machinery Industry Press, 2022.

[2] ZHANG K W, SHI Y, KARNOUSKOS S, et al. Advancements in industrial cyber-physical systems: an overview and perspectives[J]. IEEE Transactions on Industrial Informatics, 2023, 19(1): 716-729.

[3] BHATIA V, KUMAWAT S, JAGLAN V. Overview of the role of the Internet of things and cyber-physical systems in various applications[M]. Boca Raton: Apple Academic Press, 2022.

[4] HU J, KAUR K, LIN H, et al. Intelligent anomaly detection of trajectories for IoT empowered maritime transportation systems[J]. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(2): 2382-2391.

[5] WU W J, YANG F, GAO Y, et al. Distributed handoff problem in heterogeneous networks with end-to-end network slicing: decentralized Markov decision process-based modeling and solution[J]. IEEE Transactions on Wireless Communications, 2022, 21(12): 11222-11236.

[6] ZHOU H, ZHOU H B, LI J G, et al. Heterogeneous ultradense networks with traffic hotspots: a unified handover analysis[J]. IEEE Internet of Things Journal, 2023, 10(10): 8825-8838.

[7] LIN Y, ZHANG Z M, HUANG Y M, et al. Heterogeneous user-centric cluster migration improves the connectivity-handover trade-off in vehicular networks[J]. IEEE Transactions on Vehicular Technology, 2020,

69(12): 16027-16043.

[8] GUO C, GONG C, XU H T, et al. A dynamic handover software-defined transmission control scheme in space-air-ground integrated networks[J]. IEEE Transactions on Wireless Communications, 2022, 21(8): 6110-6124.

[9] REN Q Q, ABBASI O, KURT G K, et al. Handoff-aware distributed computing in high altitude platform station (HAPS) – assisted vehicular networks[J]. IEEE Transactions on Wireless Communications, 2023, 22(12): 8814-8827.

[10] LI H, OTA K, DONG M X. AI in SAGIN: building deep learning service-oriented space-air-ground integrated networks[J]. IEEE Network, 2023, 37(2): 154-159.

[11] ZHANG P Y, WANG C, KUMAR N, et al. Space-air-ground integrated multi-domain network resource orchestration based on virtual network architecture: a DRL method[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(3): 2798-2808.

[12] ZHOU C H, WU W, HE H L, et al. Deep reinforcement learning for delay-oriented IoT task scheduling in SAGIN[J]. IEEE Transactions on Wireless Communications, 2021, 20(2): 911-925.

[13] ZHANG S B, LIU A J, HAN C, et al. Multiagent reinforcement learning-based orbital edge offloading in SAGIN supporting Internet of remote things[J]. IEEE Internet of Things Journal, 2023, 10(23): 20472-20483.

[14] CAI S F, LAU V K N. Modulation-free M2M communications for mission-critical applications[J]. IEEE Transactions on Signal and Information Processing Over Networks, 2018, 4(2): 248-263.

[15] CAO Y L, LIU Q H, ZUO Y, et al. Receiver-assisted cellular/WiFi handover management for efficient multipath multimedia delivery in heterogeneous wireless networks[J]. EURASIP Journal on Wireless Communications and Networking, 2016, 2016: 1-13.

[16] ANDERSON B D O, MOORE J B. Optimal filtering[M]. Englewood: Prentice-Hall, 1979.

[17] SHI L, ZHANG H S. Scheduling two Gauss-Markov systems: an optimal solution for remote state estimation under bandwidth constraint[J]. IEEE Transactions on Signal Processing, 2012, 60(4): 2038-2042.

[18] XU Y G, HESPANHA J P. Estimation under uncontrolled and controlled communications in networked control systems[C]//Proceedings of the 44th IEEE Conference on Decision and Control. Piscataway: IEEE Press, 2005: 842-847.

[19] WANG C, LEI S B, JU P, et al. MDP-based distribution network reconfiguration with renewable distributed generation: approximate dynamic programming approach[J]. IEEE Transactions on Smart Grid, 2020, 11(4): 3620-3631.

[20] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms[J]. arXiv Preprint, arXiv: 1707.06347, 2017.

[21] TESSLER C, MANKOWITZ D J, MANNOR S. Reward constrained policy optimization[J]. arXiv Preprint, arXiv: 1805.11074, 2018.

[22] BETTSTETTER C, HARTENSTEIN H, PÉREZ-COSTA X. Stochastic properties of the random waypoint mobility model[J]. Wireless Networks, 2004, 10(5): 555-567.

[23] HASSELT H V, GUEZ A, SILVER D. Deep reinforcement learning with double Q-learning[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2016, 30(1).

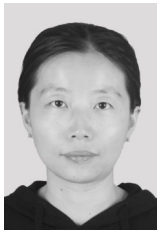
[24] ACHIAM J, HELD D, TAMAR A, et al. Constrained policy optimization[J]. arXiv Preprint, arXiv: 1705.10528, 2017.

[25] HEESS N, TB D, SRIRAM S, et al. Emergence of locomotion behaviours in rich environments[J]. arXiv Preprint, arXiv: 1707.02286, 2017.

[26] WU Y, YANG Q H, LIU X F, et al. Delay-constrained optimal transmission with proactive spectrum handoff in cognitive radio networks[J]. IEEE Transactions on Communications, 2016, 64(7): 2767-2779.

[27] WU Y Q, HU F, KUMAR S, et al. A learning-based QoE-driven spectrum handoff scheme for multimedia transmissions over cognitive radio networks[J]. IEEE Journal on Selected Areas in Communications, 2014, 32(11): 2134-2148.

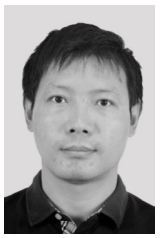
[作者简介]



武艳（1986– ），女，内蒙古鄂尔多斯人，博士，西安电子科技大学讲师、硕士生导师，主要研究方向为移动信息物理系统、工业物联网、人工智能等。

潘广川（2000– ），男，重庆人，西安电子科技大学硕士生，主要研究方向为人工智能、移动信息物理系统。

姚明昉（1975– ），男，陕西西安人，博士，西安电子科技大学副教授、博士生导师，主要研究方向为时间敏感网络、确定性网络的测试与探知、高动态无人机自组织网络中的组网与业务质量保证技术等。



杨清海（1976– ），男，山东高密人，博士，西安电子科技大学教授、博士生导师，主要研究方向为人工智能与现代通信、自主通信、信息融合、网络融合、数据分析。

梁中明（1955– ），男，加拿大人，博士，加拿大皇家学会科学院院士、加拿大工程院院士、深圳大学特聘教授，主要研究方向为无线网络、移动系统。