基于元深度强化学习的蜂窝网链路自适应方法

叶小文，林恒羿，吴怡
（福建师范大学光电与信息工程学院，福建 福州 350000）

摘　要：针对蜂窝网络对可靠性与数据速率的高要求，提出一种高效且泛化能力强的链路自适应方法。首先，为保障无线通信传输可靠性，设计一种带约束的调制编码方案选择策略，以满足对误块率的限制。其次，针对传统算法在未知传输环境下泛化性较差的问题，将元学习与深度强化学习相结合，通过离线训练与在线微调，实现策略的快速收敛。仿真结果表明，在严格满足误块率要求的前提下，所提算法相较于传统链路自适应算法，具备更高的数据速率性能和更强的泛化能力。

关键词：链路自适应；深度强化学习；元学习；泛化性
中图分类号：TN929.5
文献标志码：A
DOI：10.11959/j.issn.1000–436x

Meta Deep Reinforcement Learning-Based Link Adaptation for Cellular Networks

YE Xiaowen, LIN Hengyi, WU Yi
College of Photonic and Electronic Engineering, Fujian Normal University, Fuzhou 350000, China

Abstract: To address the stringent requirements for reliability and data rate in cellular networks, an efficient and highly generalizable link adaptation method was proposed. First, to ensure reliable wireless communication transmission, a constrained modulation and coding scheme selection strategy was designed to meet the block error rate requirement. Second, to overcome the poor generalization capability of traditional algorithms in unknown transmission environments, a meta-learning mechanism was integrated with deep reinforcement learning. Through offline training followed by online fine-tuning, rapid policy convergence was achieved. Simulation results demonstrate that, while strictly satisfying the block error rate requirement, the proposed algorithm achieves higher data rate performance and stronger generalization capability compared with traditional link adaptation algorithms.

Keywords: link adaptation, deep reinforcement learning, meta-learning, generalization capability

0　引言

链路自适应是长期演进技术（long term evolution，LTE）和新空口（new radio, NR）系统的关键技术，其核心在于根据信道质量指示（channel quality indicator, CQI）动态地调整调制编码方案（modulation and coding scheme, MCS），使之与时变的信道状态相匹配，从而提升系统数据速率并降低误块率。

在蜂窝网络中，传统的下行链路自适应技术如外环链路自适应（outer-loop link adaptation, OLLA）主要基于用户上报的CQI来选择用于数据传输的MCS，并利用解码结果来优化选择策略。然而，

收稿日期：2026–01–28；修回日期：2026–04–02
通信作者：吴怡，wuyi@fjnu.edu.cn
基金项目：国家自然科学基金资助项目（No.62501157，No.U25A20398）；福建省青年科技人员育成项目（No.2025350410）
Foundation Items：The National Natural Science Foundation of China (No.62501157，No.U25A20398), The Foundation for Cultivated Young Talents of Fujian Province, China (No.2025350410)

以OLLA为代表的传统链路自适应技术，在CQI过期的情况下面临严重的性能下降。具体而言，由于系统的处理和传输延迟，用于MCS决策的CQI通常是滞后的。此外，为减少信令开销，实际系统中用户仅周期性上报CQI。这种过期的CQI与实际信道状态不匹配，将导致MCS决策出现偏差，从而严重损害系统性能。

为了克服传统链路自适应技术在CQI过期场景下的局限性，研究者提出了一系列更加智能的MCS选择方法。其中，作为人工智能的一项关键技术，强化学习被引入以实现高效的链路自适应决策[1-6]。在该框架中，智能体通过观察信道状态并执行相应动作与环境交互，随后获得反映决策质量的即时奖励。通过持续试错与策略更新，智能体逐步优化其行为以最大化累积奖励。基于强化学习算法，文献[1]针对OLLA参数配置难题，采用Q学习算法动态选择MCS和传输层数，以最大化系统数据速率。文献[2]同样基于Q学习算法选择MCS，以在维持低误块率的同时最大化频谱效率。文献[3]则运用Q学习算法对功率资源进行分配，旨在降低传输延迟并提升可靠性。文献[4]和文献[5]引入了基于上下文多臂赌博机算法，分别以用户速率、载波频率及解码对数似然比作为上下文信息，辅助MCS决策以提升传输可靠性。文献[6]进一步利用强化学习算法动态调整OLLA的步长，进而优化MCS选择，以提升网络数据速率。

尽管传统强化学习已能实现一定程度的智能链路自适应，但其在处理无线通信场景中高维状态空间时仍面临困难。为此，深度强化学习（deep reinforcement learning, DRL）被引入以应对大规模状态空间下的链路自适应问题。DRL将深度神经网络与强化学习结合，利用神经网络作为强大的函数逼近器，能够有效处理高维、连续的信道状态信息。在相关研究中，文献[7-13]利用深度Q网络及其改进算法，从能耗、信令开销、频谱资源等方面降低了链路自适应的系统开销。具体而言，文献[7]与[8]着眼于无线资源的优化利用，分别通过联合优化MCS与数据流数量或资源块分配，以最大化频谱效率和最小化资源块消耗。文献[9]在兼顾传输时延和可靠性要求的基础上，利用双重深度Q网络结合信道统计信息，优化传输模式与功率分配，从而降低系统能耗。文献[10]则采用深度Q网络结合分类经验重放和切换控制策略，在保证高数据速率与低误块率的同时，减少了MCS切换带来的额外开销。文献[11]面向下一代Wi-Fi网络场景，提出基于深度Q网络的动态功率调整方法，显著提升了系统能效。此外，针对DRL自身计算开销大的问题，文献[12]与文献[13]提出使用DRL对OLLA的偏移量或步长进行智能调整，而非直接决策MCS，在控制复杂度的同时仍能优化链路性能。不同于上述研究，文献[14]针对超可靠低时延通信场景中对误块率的严格要求与时变信道特性，提出采用DRL选择初始MCS，再通过OLLA进行微调的策略，在保证极高可靠性的同时最大化编码率。文献[15]面向综合数据与能量传输系统，设计了基于约束的参数化动作深度确定性策略梯度算法，用于联合优化调制方式与功率控制，在满足误块率等长期约束的前提下，最大化能量收集性能。文献[16]则聚焦Wi-Fi网络中帧长度与物理速率的权衡问题，利用双重深度Q网络对其进行联合优化，实现了系统数据速率与传输时延性能的综合提升。

上述文献提出的方案虽然在提升系统数据速率和资源利用率方面效果显著，但大多未对误块率施加显式约束，仅将其作为奖励函数中的惩罚项。例如，文献[15]通过设计包含约束项的奖励函数，引导智能体在最大化收益的同时学习满足约束。然而，这种基于奖励函数的约束方式对惩罚项的设置较为敏感，若惩罚值设置不当，可能导致智能体难以做出满足约束的决策。此外，尽管这些DRL方法在特定场景下表现优异，但在面对动态多变的无线通信环境时，仍存在泛化能力弱和冷启动问题。传统DRL算法在应对尚未涉及的无线传输环境（如CQI反馈延迟或上报周期发生改变）时，由于缺乏先验知识，通常需要从零开始重新训练。在收敛至最优策略之前，算法需进行大量的试错探索和消耗较长的训练时间。并且，在训练前期，智能体输出的动作常具有随机性或高风险，难以满足无线通信对性能的实际需求。为解决上述问题，本文融合元学习与深度强化学习，提出了一种具备良好泛化能力的链路自适应算法，称为泛化链路自适应（generalizable link adaptation, GLA）。与现有研究大多仅关注数据速率而忽略可靠性不同，本文将MCS选择过程建模为马尔可夫决策过程，旨在满足误块率约束的前提下最大化链路数据速率。首

先，针对传统DRL在新场景下收敛慢的问题，本文引入包含离线训练和在线微调两阶段的元学习机制。离线训练阶段，算法在多样化传输环境下进行多任务学习，提取跨任务的通用特征，得到一组优化的元参数。在线微调阶段，当面对尚未涉及的传输环境时，智能体以该元参数为起点，仅需少量实时交互样本即可快速微调，迅速收敛至适应新环境的最优策略。其次，传统DRL通常采用概率-贪婪策略进行动作选择，即以一定概率选择当前最优动作，一定概率进行随机探索，以平衡探索与利用。然而，该策略在探索过程中可能选择那些以牺牲误块率为代价换取高数据速率的高风险动作，从而损害传输可靠性。为满足误块率约束，本文提出一种带约束的动作选择算法。该算法首先根据各动作的历史传输结果统计其误块率，筛选出满足误块率要求的“安全动作”。在此基础上，若系统误块率超过预设阈值，则智能体仅从“安全动作”集合中选择MCS，以优先保障传输可靠性；若系统误块率满足要求，则智能体可在全局动作空间中搜索最优MCS，以最大化链路数据速率。

本文的主要贡献如下：

1）提出了带约束的动作选择算法：该算法的核心思想是通过动态调整“安全动作”空间，使智能体在满足误块率约束的前提下，最大化链路数据速率。

2）设计了基于元学习的泛化性增强机制。该机制通过构建元学习双层优化框架，使智能体能够从历史任务中提取跨任务的先验知识，训练得到泛化能力较强的元参数。在面对未知传输环境（例如未涉及过的CQI反馈延迟或上报周期）时，元参数可实现快速少样本适配，显著提升收敛速度。

3）通过仿真验证了所提方法的性能：实验结果表明，与其他链路自适应技术相比，所提GLA在满足目标误块率约束的前提下，显著提高了链路数据速率，并对不同环境参数表现出良好的鲁棒性。此外，所提出的元学习机制有效提升了GLA的收敛速度与泛化性能。

1　系统模型

本文考虑LTE/NR网络的下行链路传输场景，其中一个基站为单个用户提供服务。假设基站采用饱和缓冲模型，即始终有数据待发送。在每个时隙 $t \in \mathcal{T} = \{1,2,\cdots,T\}$，基站从给定集合 $\mathcal{M} = \{1,2,\cdots,M\}$ 中选择一个MCS索引 $m(t)$，其中 $m(t)=1$ 表示选用最低阶MCS，而索引 $m(t)=M$ 表示选用最高阶MCS。根据所选 $m(t)$，$D_m(t)$ 个信息比特被编码到可变长度的数据包中[17]，并通过时变信道传输给用户。MCS索引越高，所封装的信息比特也越多。定义 $x_m(t)$ 为时隙 t 上的MCS选择指示符

$$x_m(t)=\begin{cases}1, & \text{若索引}m(t)\text{被选中}\\ 0, & \text{否则}\end{cases} \tag{1}$$

由于每个时隙最多只能选择一种MCS，故有 $\sum_{m=1}^{M} x_m(t)=1, \forall t \in \mathcal{T}$。在时隙 t 上，与所选索引 m_t 对应的数据速率可表示为

$$\frac{1}{\Delta \mathrm{t}} \sum_{m=1}^{M} D_m(t)\, x_m(t) \tag{2}$$

其中，Δt为一个时隙的持续时间。用户接收数据包后，向基站提供两种反馈：确认应答（acknowledgment, ACK）/否定应答（negative acknowledgment, NACK），以及CQI。

ACK/NACK反馈：为实现可靠通信，用户通过ACK/NACK反馈将传输结果告知基站[18]。具体来说，由于无线信道的衰落特性，用户接收到的数据包可能无法被正确解码。为此，用户对解码后的数据包进行循环冗余校验，以判断接收是否成功，并向基站反馈ACK（解码成功）或NACK（解码失败）。令 $y_m(t)$ 表示解码结果指示符

$$y_m(t)=\begin{cases}1, & \text{解码成功}\\ 0, & \text{解码失败}\end{cases} \tag{3}$$

因此，用户端的有效数据速率可表示为

$$R(t)=\frac{1}{\Delta t} \sum_{m=1}^{M} D_m(t)\, x_m(t)\, y_m(t) \tag{4}$$

CQI反馈：为辅助基站进行MCS决策，用户周期性地向基站上报CQI，该值反映了当前信道的瞬时质量。具体而言，在每个时隙开始时，用户基于参考符号估计瞬时信噪比 γ，并通过量化函数 $f(\cdot)$ 将其映射为离散的CQI值 $k=f(\gamma)=\mathcal{K} \in \{0,1,\cdots,K\}$。然而，由于量化误差、传输延迟和无线信道的时变特性，基站实际可获得的CQI是不准确的。此外，为降低系统信令开销，用户仅周期性上报CQI，导致基站获得的CQI具有滞后性。

为区分用户在时隙 t 生成的CQI $k(t)$ 与基站实际可用的CQI，本文将后者记为 $k^{-}(t)$。

为了实现智能化的链路自适应，本文提出的系统模型及神经网络部署方案如图1所示。在该场景中，智能体部署于基站侧。基站根据接收到的CQI信息 $k^{-}(t)$ 及上一时刻的ACK/NACK反馈，利用深度神经网络对不同MCS进行评估，并通过带约束的动作选择策略选出最佳MCS。其中，智能体所涉及的状态空间、动作空间及奖励函数的具体定义将在本文第2章中详细阐述。

基于以上框架，系统的优化目标是在保证通信可靠性的前提下，最大化链路的数据速率。具体来说，若盲目选用高阶MCS，可能导致误块率急剧升高；而过于保守的MCS选择虽能满足误块率约束，却会显著降低数据速率。因此，系统需在提升链路数据速率的同时，确保系统误块率不超过预设阈值 $\eta_{\max}$。优化问题可表述为

$$\max \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}R(t)$$

$$\text{s.t.C}_1\text{:}1-\frac{1}{t}\sum_{t=1}^{t}\sum_{m=1}^{M}x_m(t)\,y_m(t)\le\eta_{\max}\qquad(5)$$

$$\text{C}_2\text{:}\sum_{m=1}^{M}x_m(t)=1,\forall t\in\mathcal{T}$$

$$\text{C}_3\text{:}x_m,y_m(t)\in\{0,1\},\forall m\in\mathcal{M}\&t\in\mathcal{T}$$

其中，$\eta_{\max}$ 为系统允许的最大误块率。约束 C_1 表示从时隙1到时隙 t 的系统误块率，约束 C_2 确保在每个时隙仅选择一种MCS索引，约束 C_3 定义了变量 x_m 和 y_m 的取值限制。

2　GLA算法

传统OLLA算法依赖实时、准确的CQI以选择MCS。然而，由于反馈延迟、周期性上报及量化误差等因素，基站实际可用的CQI往往是过期且不准确的，导致传统OLLA算法难以有效应对此类场景。为此，本文首先将MCS选择问题建模为马尔可夫决策过程，并引入DRL算法，通过与环境交互实现对MCS决策的动态优化。但DRL算法在面对未经训练的环境时需从零开始学习，存在收敛速度慢的问题。因此，本文提出一种融合元学习与深度强化学习的GLA算法，并结合带约束的动作选择机制，在严格满足系统误块率约束的条件下，显著提升链路数据速率。

2.1　马尔可夫决策过程

在本文建模中，用户与动态无线信道共同构成环境，而搭载GLA算法的基站被视为智能体。形式上，该马尔可夫决策过程包含四个要素：状态空间、动作空间、状态转移概率和奖励函数。

状态空间 $\mathcal{S}$ 由一组能够充分表征系统环境的观测变量构成。在下行链路传输中，基站通常依赖用户反馈的CQI来评估信道条件。然而，由于量化误差、传输延迟和周期性上报等因素，基站实际可用

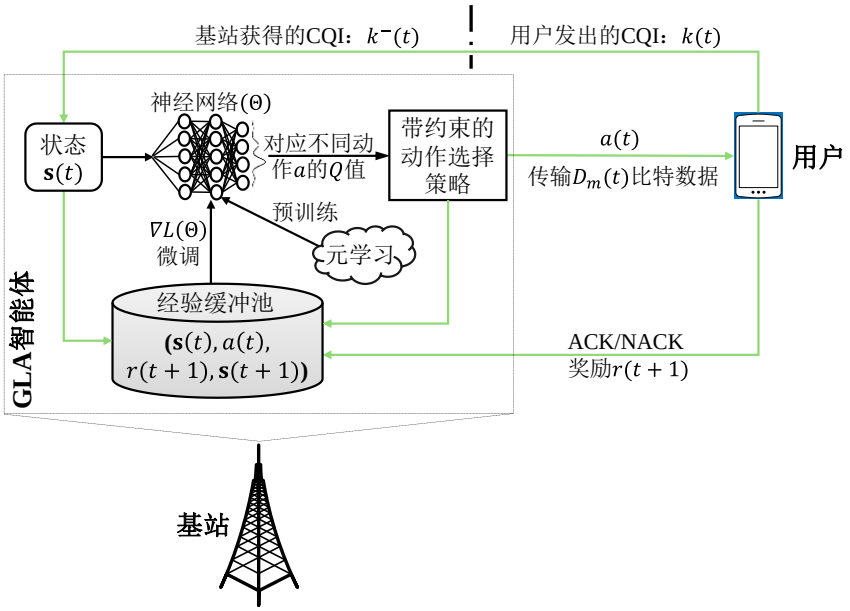

图1　系统模型及神经网络部署方案

的CQI往往无法准确反映当前瞬时信噪比。为此，本文将上一时隙动作对应的ACK/NACK反馈纳入状态空间。这种反馈虽然不如CQI精细，但其仍可为判断当前信道条件提供一定依据。综合来看，尽管这些观测数据是不理想的，但它们包含了关于信道时变特性的重要信息，对于指导链路自适应决策至关重要。考虑到系统的部分可观测特性，定义 $\mathbf{z}(t)=[m(t-1),y(t-1),k^{-}(t),\Delta k^{-}(t)]$ 作为基础观测变量。其中，$m(t-1)$ 为上一时隙选择的MCS索引，$y(t-1)$ 为上一时隙的解码结果反馈（ACK/NACK），$k^{-}(t)$ 表示基站当前可用的CQI，$\Delta k^{-}(t)$ 则为相邻CQI的差值，用以捕捉信道的时变趋势。为进一步弥补CQI信息的不完美性，本文引入历史观测序列来辅助决策。因此，时隙 t 的状态 $\mathbf{s}(t)$ 定义为

$$\boldsymbol{s}(t)=[\boldsymbol{z}(t),\boldsymbol{z}(t-1),\cdots,\boldsymbol{z}(t-L+1)] \tag{6}$$

其中，L 为状态历史长度。

动作空间 $\mathcal{A}$ 包含智能体所有可执行的动作。基于所建立的系统模型，动作空间为有限离散集合 $\mathcal{M}=\{1,2,\cdots,M\}$，其中 M 表示可用的MCS总数。每个动作 $a(t)$ 对应一个特定的MCS索引 $m(t)$：$a(t)=1$ 表示选择最低阶MCS，旨在恶劣信道条件下确保传输可靠性；$a(t)=M$ 表示选择最高阶MCS，用于在信道质量较好时最大化数据速率。因此，智能体在时隙 t 的动作可表示为

$$a(t)=m(t)\in\mathcal{M}=\{1,2,\cdots,M\} \tag{7}$$

状态转移概率 $p(\mathbf{s}(t+1)=\mathbf{s}'|\mathbf{s}(t)=\mathbf{s},a(t)=m)$ 表示系统在执行动作 $a(t)$ 后，从当前状态 $\mathbf{s}(t)$ 转移到下一个状态 $\mathbf{s}(t+1)$ 的概率[19]。在所考虑的下行链路场景中，信道状态的变化由时变瑞利衰落过程决定。由于智能体仅能获取环境的部分观测信息，无法获得完整的环境状态，因此显式的转移概率对于智能体而言是未知的，且决策 $a(t)$ 必须在不确定条件下进行。由此，该问题构成一个部分可观测马尔可夫决策过程问题。

奖励函数 $r(\mathbf{s}(t),a(t))$ 用于评估在状态 $\mathbf{s}(t)$ 下执行动作 $a(t)$ 所获得的即时收益。本文旨在学习一种高效的链路自适应策略，在保证传输可靠性的前提下最大化数据速率。本文设计了与实际数据速率相对应的标量奖励，并在第2.3节引入带约束的动作选择机制以满足误块率需求。具体地，在时隙 t 上，智能体的奖励 $r(t+1)$ 定义为

$$r(t+1)=\begin{cases}R(t), & \text{传输成功}\\ 0, & \text{传输失败}\end{cases} \tag{8}$$

在马尔可夫决策过程中，折扣因子 $\gamma\in[0,1]$ 用于调节即时奖励与未来长期奖励的相对重要性。智能体的目标是寻找一个最优策略 π，将观测状态映射到动作空间，以最大化累积折扣奖励

$$G(t)=\sum_{l=t}^{\infty}\gamma^{l-t}r(l+1) \tag{9}$$

此外，策略 π 和状态转移概率 p 共同决定了状态-动作对 $(\mathbf{s}(t),a(t))$ 的Q值，其定义为

$$Q_{\pi}(\mathbf{s}(t),a(t))=\mathbb{E}_{\pi}[G(t)|\mathbf{s}(t),a(t)] \tag{10}$$

至此，链路自适应问题已被建模为马尔可夫决策过程。若CQI实时且精确，则状态转移概率 $p(\mathbf{s}'|\mathbf{s},a)$ 易于获取，此时该马尔可夫决策过程可通过动态规划等方法有效求解。然而实际系统中，由于反馈延迟、周期性上报以及量化误差等因素，基站在每个时隙只能获得过期且不完美的CQI。因此，基站无法准确推断真实的信道状态转移模型，即 $p(\mathbf{s}'|\mathbf{s},a)$ 未知。基于模型的方法在此类场景下不再适用，这促使本文采用无模型的DRL技术来解决该问题。

2.2　神经网络架构

在所提出的GLA方案中，DRL智能体采用结构相同但参数独立的两组人工神经网络[20]：参数为 Θ 的主网络和参数为 Θ^{-} 的目标网络。主网络根据当前状态输入评估所有动作的Q值；目标网络则用于计算目标Q值，以维持训练过程的稳定性。该神经网络的架构包括输入层、门控循环单元层[21]、全连接层和输出层。

首先，为了从序列化的状态数据中有效提取潜在的时序相关性，状态向量 $\mathbf{s}(t)$ 被输入门控循环单元层。该层对时序特征进行聚合后，结果送入全连接层作进一步分析，最终输出所有MCS索引对应的Q值。

与主网络不同，目标网络根据下一状态 $\mathbf{s}(t+1)$ 计算目标Q值。具体而言，目标网络通过相同的网络结构处理下一状态，生成用于计算时序差分目标的Q值。为缓解训练过程中的不稳定性并促进算法收敛，目标网络采用周期性更新策略：目标网络

的参数Θ在每个训练步进行梯度更新，而目标网络的参数Θ^-仅在固定的更新周期后通过直接复制主网络权重进行同步，即

$$\Theta^- \leftarrow \Theta \tag{11}$$

这一机制使目标网络的参数在多个时隙内保持不变，有利于稳定学习目标并提升算法的收敛性。

2.3 带约束的动作选择策略

本文的系统目标是在满足误块率约束的前提下，最大化数据速率。在2.1节中，本文已将最大化数据速率的目标嵌入奖励函数设计。然而，智能体的决策过程仍需应对传输可靠性的挑战。传统的DRL算法通常依赖概率-贪婪策略选择动作，但其无约束的探索机制往往导致智能体为追求高数据速率而频繁执行高风险动作，从而引发误块率急剧上升。为此，本节提出一种带约束的动作选择策略，以确保系统满足误块率需求。

与传统在全局动作空间$\mathcal{M}$中选择动作的策略不同，所提策略根据历史传输统计信息动态调整安全动作子集。具体来说，首先定义“安全动作”集合$\tilde{\mathcal{M}}(t)$，该集合从全局动作空间$\mathcal{M}$中筛选出满足可靠性约束的动作

$$\tilde{\mathcal{M}}(t)=\left\{a\in\mathcal{M}\,\middle|\,\tilde{\eta}(a,t)=\frac{N^{\text{failed}}(a,t)}{N^{\text{total}}(a,t)}\leqslant\eta_{max}\right\} \tag{12}$$

其中，$N^{\text{total}}(a,t)$表示截至时隙t动作a被选择的总次数，$N^{\text{failed}}(a,t)$为该动作导致传输失败的累计次数，$\tilde{\eta}(a,t)$为截至时隙t动作a的累计误块率，η_{max}为预设的误块率阈值。基于上述定义，本策略通过比较系统当前累计误块率η_t与阈值η_{max}，动态调整动作选择范围，动作选择逻辑如下

$$a(t)=\begin{cases}\underset{a\in\tilde{\mathcal{M}}(t)}{\text{argmax}}\,Q(\mathbf{s},a;\Theta), & \text{若}\eta(t)>\eta_{\max}(\text{概率}1-\epsilon)\\ \underset{a\in\tilde{\mathcal{M}}(t)}{\text{argmax}}\,Q(\mathbf{s},a;\Theta), & \text{若}\eta(t)\leqslant\eta_{\max}(\text{概率}1-\epsilon)\\ \text{随机选择}a\in\mathcal{M}, & \text{其他情况}(\text{概率}\epsilon)\end{cases} \tag{13}$$

该动作选择逻辑仍保留概率-贪婪策略的基本结构，通过概率ε平衡利用与探索。公式第一行对应于系统误块率$\eta(t)$超过阈值$\eta_{\max}$时，为降低后续时隙的误块率，智能体必须从安全动作集合$\tilde{\mathcal{M}}(t)$中选择最优动作。公式第二行表示当系统误块率$\eta(t)$满足需求时，智能体可在整个动作空间$\mathcal{M}$中进行选择；若所选动作原不在$\tilde{\mathcal{M}}(t)$中但传输成功，则后续可能被纳入安全集合；若传输失败导致$\eta(t)$升高并超过$\eta_{\max}$，则下一时隙将回退至第一行逻辑。公式第三行为随机探索环节，智能体以概率ε随机选择动作，以避免陷入既无法满足误块率约束又无法最大化数据速率的次优策略。

2.4 在线训练过程

为提高GLA方案的有效性，本文采用经验回放机制来训练神经网络[22]。智能体维护一个先进先出的经验缓冲池$\mathcal{E}$，用于存储在每个时隙获得的经验数据。具体而言，在时隙t，智能体的经验$e(t)$定义为

$$e(t)=(\mathbf{s}(t),a(t),r(t+1),\mathbf{s}(t+1)) \tag{14}$$

在每个训练回合中，智能体从经验缓冲池中随机采样N个经验构成一个批次$\mathcal{B}$。该随机采样机制有助于打破连续样本间的相关性，从而有效提升训练的稳定性。随后，通过最小化预测Q值与目标Q值之间的差异，更新主网络参数Θ。基于贝尔曼方程，对每一条采样经验$e_i\in\mathcal{B}$，本文利用目标网络计算其目标Q值

$$y_i=r_i+\gamma\max_{a'\in\mathcal{M}}Q\left(\mathbf{s}_i',a';\Theta^-\right) \tag{15}$$

其中，$\mathbf{s}_i'$表示经验e_i中的下一状态，a'表示下一个状态下可能的动作。为量化主网络的预测偏差并指导参数更新，定义损失函数$L(\Theta)$为批次样本上的均方误差

$$L(\Theta)=\frac{1}{N}\sum_{i=1}^{N}(y_i-Q(\mathbf{s}_i,a_i;\Theta))^2 \tag{16}$$

并采用均方根传播（root mean square propagation, RMSprop）优化器最小化该损失

$$\Theta\leftarrow\Theta-\alpha\cdot\text{RMSprop}(\nabla_\Theta L(\Theta)) \tag{17}$$

其中，$\alpha\in[0,1]$表示学习率，$\text{RMSprop}(\cdot)$表示对梯度进行自适应调整的优化操作，用于缓解无线信道随机衰弱带来的训练震荡。

2.5 元学习框架

尽管前述方案在特定传输环境下已能够实现高效的链路自适应，但其性能严重依赖于训练环境的统计特性。实际通信场景中，CQI反馈延迟和上报周期等环境参数往往高度动态变化。当智能体面对从未接触过的传输环境时，传统深度强化学习方法通常需从零开始进行大量重新训练，导致收敛速度慢、计算开销大。为应对这一挑战并提升GLA的

泛化能力，本文引入基于Reptile算法的元学习机制[23]。其核心思想并非仅针对单一任务优化策略，而是寻求一组通用的初始网络参数$\tilde{\Theta}$，使其能够捕捉不同传输环境间的共同特征。基于该参数，智能体在遇到新环境（不同的CQI反馈延迟和上报周期）时，仅需极少量样本进行微调，即可快速收敛至最优策略。该机制包含元训练与元适应两个阶段。

元训练阶段旨在通过跨任务学习获取泛化性强的元参数$\tilde{\Theta}$，使智能体能够根据不同的CQI反馈延迟与上报周期快速调整策略。该过程采用双层优化框架实现。第一层优化（任务级适配）：对于每个采样的传输环境（任务）$j=\mathcal{J}\in\{1,2,\cdots,J\}$，智能体首先将其主网络参数$\Theta_j$初始化为当前元参数$\tilde{\Theta}$。随后，智能体与该特定环境交互，并将生成的经验存入经验缓冲池。基于收集到的经验数据，智能体通过RMSprop优化器对参数Θ_j进行梯度更新，以获得针对该任务的最优参数

$$\Theta_j \leftarrow \Theta_j - \lambda \cdot \mathrm{RMSprop}\left(\nabla_{\Theta_j} L(\Theta_j)\right) \forall j = \mathcal{J}\in\{1,2,\cdots,J\} \tag{18}$$

其中，$\lambda\in[0,1]$为内循环更新步长。第二层优化（元参数更新）：完成所有采样任务的参数优化后，进入全局元参数更新阶段。尽管不同任务中的CQI反馈延迟与上报周期存在差异，但无线信道的物理特性具有一致性。本阶段的目标是通过评估各传输环境的信道共性，对元参数进行更新，使其逼近各任务最优策略的公共表征，优化问题可表述为

$$\min_{\Theta} \sum_{j=1}^{J} L\left(\Theta_j\right) \tag{19}$$

为高效求解此目标并避免高阶导数计算[24]，系统收集更新后的任务参数集$\{\Theta_1,\Theta_2,...,\Theta_J\}$，并采用低复杂度的Reptile算法融合多任务经验。该算法不显式计算元参数的梯度，而是提取各任务参数相对于元参数的更新方向，进而更新元参数

$$\tilde{\Theta} \leftarrow \tilde{\Theta} + \frac{\beta}{J}\sum_{j=1}^{J}(\Theta_j - \tilde{\Theta}) \tag{20}$$

其中，$\beta\in[0,1]$为元更新步长。

元适应阶段旨在利用元训练获得的知识快速适配新环境。当面对一个尚未涉及的传输环境时，智能体将元训练阶段获取的元参数$\tilde{\Theta}$作为初始参数赋予主网络

$$\Theta_{\mathrm{new}} \leftarrow \tilde{\Theta} \tag{21}$$

由于该元参数已具备较强的泛化能力，智能体无需重新学习信道特性，仅需与新环境进行少量交互以收集样本，并基于这些样本对网络进行微调，从而适应当前传输环境中的CQI反馈延迟与周期。具体而言，通过RMSprop优化器对参数进行更新

$$\Theta_{\mathrm{new}} \leftarrow \Theta_{\mathrm{new}} - \lambda \cdot \mathrm{RMSprop}\left(\nabla_{\Theta_{\mathrm{new}}} L(\Theta_{\mathrm{new}})\right) \tag{22}$$

通过上述机制，智能体能够以极低的时间成本快速收敛至适应新环境的最优策略。

3　仿真分析

为了评估本文所提GLA算法性能，采用Python 3.12.7仿真平台进行实验，根据NR物理层标准[25]设置仿真参数。具体来说，MCS索引数量为29，CQI值的数量为16，时隙的持续时间设置为1ms，平均信噪比为15dB，信道类型为采用归一化多普勒频率为0.01的瑞利衰弱信道，误块率阈值$\eta_{\max}$设置为0.1。参照文献[25]对5GNR系统移动场景的划分，归一化多普勒频移在市区车载环境下通常处于0.001至0.01范围，而在高速铁路场景下则分布在0.01至0.1之间。为了兼顾算法在不同移动速率下的鲁棒性，本文选取0.01作为归一化多普勒频移的仿真值。算法参数如表1所示。

表1　　算法参数

仿真参数	参数值
经验缓冲池	10000
批次大小$\mathcal{B}$	128
学习率α	0.001
折扣因子γ	0.9
初始探索概率ε	0.1
探索衰减因子	0.995
最小探索概率	0.001
目标网络更新步长	100
状态历史长度L	80

首先，为验证本文算法在不完美CQI下的可行性，仿真中将CQI的反馈延迟与上报周期均设为10个时隙，以模拟实际通信中信息滞后的典型场景。同时，为评估所提GLA算法的有效性，将其与OLLA算法[26]、贝叶斯算法[17]、深度Q网络算法[10]以及带约束的深度Q网络算法进行比较。图2

展示了各算法下链路数据速率与误块率随时隙变化情况。（1）GLA算法与贝叶斯算法对比：如图2(a)和图2(b)所示，贝叶斯算法因未施加误块率约束，尽管数据速率较高，但其误块率远超0.1的目标阈值，严重违反系统可靠性要求。相比之下，GLA算法通过引入带约束的动作选择策略，能够在严格满足误块率要求的前提下，实现相比贝叶斯算法7.48%的数据速率增益。（2）GLA算法与深度Q网络算法对比：深度Q网络凭借神经网络强大的拟合与推理能力，其数据速率与误块率性能均优于贝叶斯算法。然而，由于缺乏显式的误块率约束，其误块率仍超过目标阈值。GLA算法在严格满足误块率约束的同时优化数据速率，有效弥补了深度Q网络在可靠性方面的不足。（3）GLA算法与OLLA算法对比：OLLA作为传统链路自适应算法，虽能满足误块率约束，但其策略较为保守，且对过期CQI敏感，数据速率处于对比方案中的最低水平。相比之下，GLA算法具备对复杂时序关系的学习能力，即使在过期CQI的影响下仍能保持优越的性能。因此，在同样严格满足误块率约束的条件下，GLA算法的数据速率显著优于OLLA，对应增益约为26.48%。（4）GLA算法与带约束深度Q网络算法对比：两种算法均引入了带约束的动作选择策略，且均能严格满足误码率需求。然而，GLA算法引入了元学习机制，不仅在训练初期表现出更快的收敛速度，在收敛后的数据速率性能上也优于带约束深度Q网络算法，对应增益约为3.69%。这一结果充分验证了元参数具备跨任务的可迁移泛化能力。（5）GLA算法与GLA-z算法对比：为验证在线微调的必要性，本文将元训练后不进行在线微调的算法记为GLA-z。如图2(a)所示，GLA-z算法得益于元训练获得的初始元参数，其收敛速度接近GLA算法。但由于缺乏在线微调，难以精确适配当前传输环境，数据速率低于GLA算法，对应增益约为22.45%。尽管如此，GLA-z算法相较OLLA算法仍有3.84%的数据速率增益，表明元学习中的元训练机制能够显著提升算法的泛化能力。综上所述，本文所提出的GLA算法在确保通信可靠性的前提下，相较对比算法具备更优的数据速率性能与更强的泛化能力。

其次，为验证所提GLA算法对不同误块率需求的鲁棒性，仿真在保持其他参数不变的情况下，将误块率需求$\eta_{\max}$分别设置为0.04、0.06、0.08和0.1。不同误块率约束下各算法的数据速率与误块率性能如图3所示。由图3(a)和图3(b)可知，深度Q网络算法和贝叶斯算法在设定上未施加误块率限制，因此其性能不受$\eta_{\max}$变化的影响；二者虽维持了较高的数据速率，但实现的误块率远超目标阈值。相比之下，随着$\eta_{\max}$降低，带误块率约束的各算法所获得的数据速率均呈下降趋势。这主要源于数据速率和误块率间的权衡关系：为满足更严格的误块率需求，必然需牺牲一定的数据速率性能。在所有测试的$\eta_{\max}$条件下，GLA算法、GLA-z算法、带约束的深度Q网络算法以及OLLA算法能够始终满足误块率约束，并且GLA算法的数据速率性能始终优于其他三种算法。具体而言，在各种场景下，GLA算法相较OLLA算法实现至少10.10%的数据速率提升，相比带约束的深度Q网络算法也具

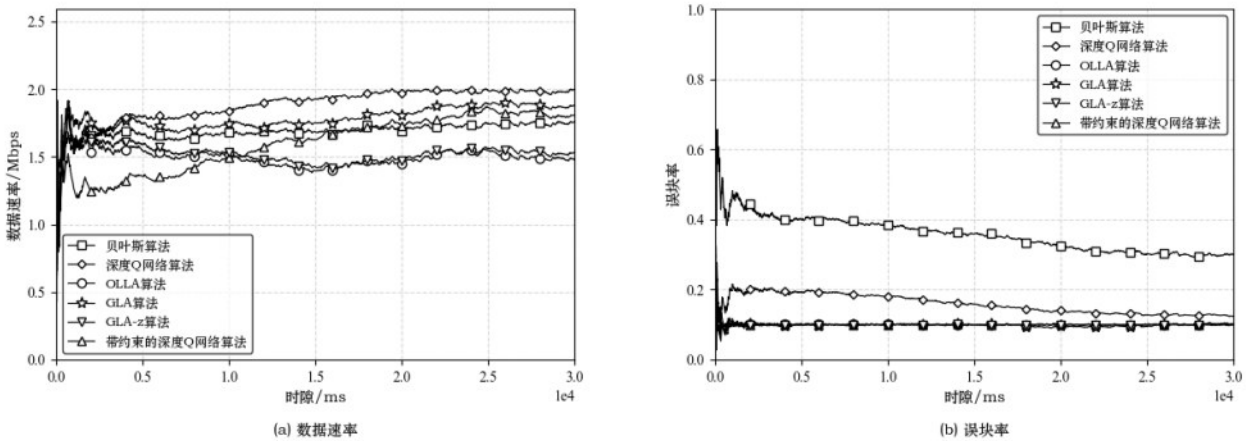

图2　各算法下链路数据速率与误块率随时隙变化

备至少 2.28% 的性能优势。值得注意的是，在 $\eta_{\max}=0.04$ 的严格约束下带约束的深度Q网络算法数据速率性能低于OLLA算法的原因是在探索阶段会因频繁触发误块率约束机制而选择过于保守的MCS，以保证误块率不超标。对于GLA-z算法，虽然其利用元参数实现了快速启动，但由于缺乏在线微调难以完全适配当前传输环境，因此其数据速率性能会有所降低。特别是在 $\eta_{\max}=0.04$ 的严格约束下，GLA-z算法无法通过持续的反馈迭代以达到较高的数据速率。而GLA算法得益于元学习所提供的先验知识，能够快速探索到既满足误块率约束又具有更高数据速率的MCS，因而相比带约束的深度Q网络算法及GLA-z算法展现出更佳的数据速率性能。综上所述，与其他算法相比，所提GLA算法对不同 $\eta_{\max}$ 具备更好的鲁棒性。

为了研究所提GLA算法对不同信道信噪比的鲁棒性，仿真在保持其他参数不变的情况下，将平均信噪比以2为步长从13dB逐步增加到19dB，不同平均信噪比下各算法的数据速率与误块率性能如图4所示，由图4(a)可知，随着信噪比的增加，所有算法的数据速率均呈现上升趋势。这是因为信道质量的改善，促使各算法倾向选择更高阶的MCS以提升数据速率。其中，贝叶斯算法与深度Q网络算法由于没有误块率的约束，以误块率性能为代价选择了更为激进的MCS，因此数据速率上高于其余算法。而在施加误块率约束的四种算法中，GLA算法相比与OLLA算法，GLA-z算法，带约束的深度Q网络算法分别实现了超过24.71%，21.75%，2.79%的数据速率增益。由图4(b)和图4(c)可以看出，没有误块率约束的贝叶斯算法与深度Q网络算法的误块率随着信噪比的增加而降低，但始终超过0.1的目标误块率。相比之下，由于误块率约束，OLLA算法的误块率始终恒定在0.1。GLA算法，GLA-z算法以及带约束的深度Q网络算法由于使用了带约束的动作选择策略，其误块率严格控制在0.1以下，且随着平均信噪比的增加，误块率进一

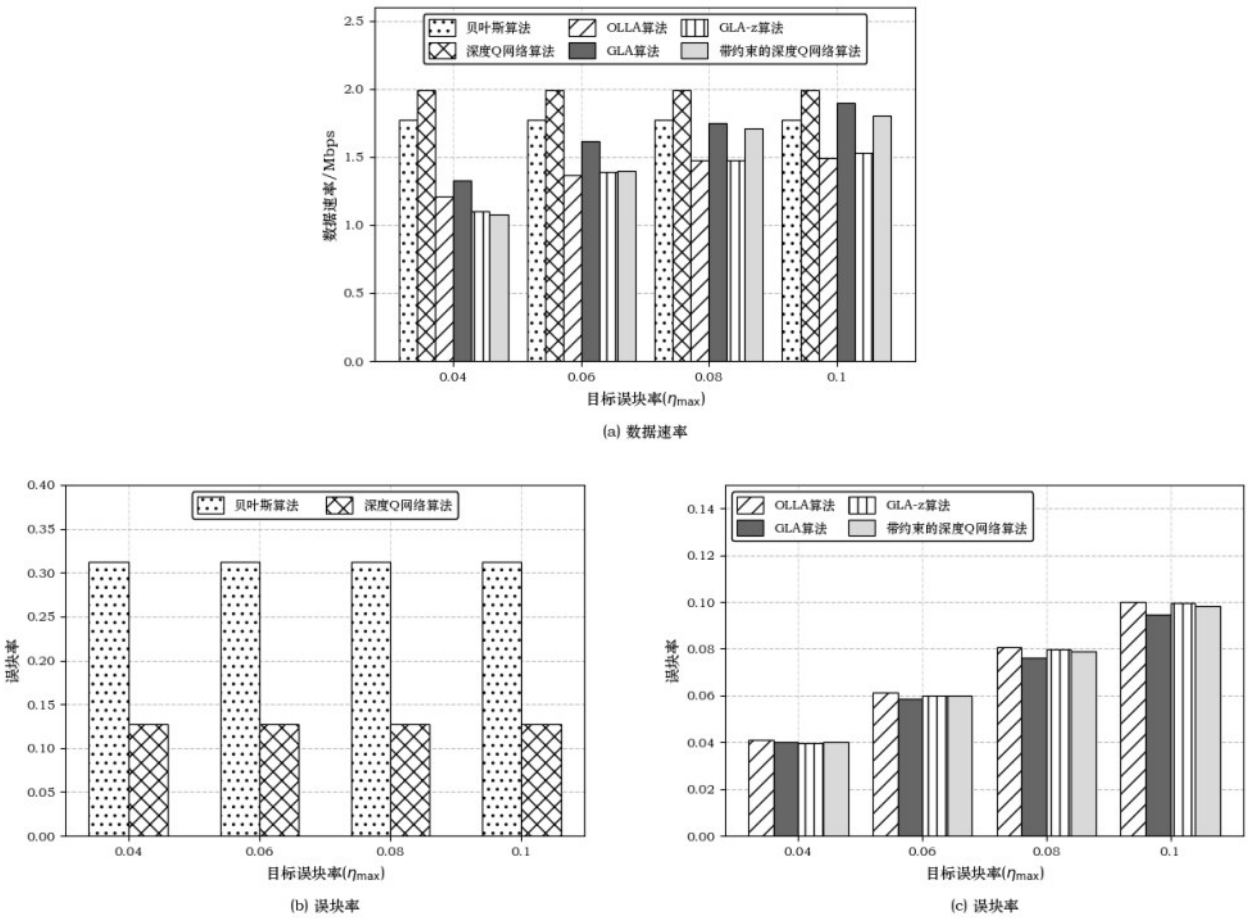

图3　不同误块率约束下各算法的数据速率与误块率性能

步降低。综上所述，GLA算法不仅在特定信噪比下有着性能优势，更能够根据不同平均信噪比自适应调整MCS选择策略，在保障可靠性的同时最大化数据速率，表现出优异的鲁棒性与泛化能力。

进一步，本文研究了任务数量J对GLA算法性能的影响。仿真在保持其他参数不变的情况下，通过改变任务数量J进行实验，不同任务数量对GLA算法传输性能与训练时长的影响如图5所示。由图5(a)可知，随着任务数量的增加，算法的数据速率呈明显上升趋势，且误块率始终保持在约束阈值之下。具体而言，当任务数量较少时（例如$J=1$），训练样本的有限性使得GLA算法难以从少数预训练样本/预传输环境中提取出泛化能力强的元参数，因此数据速率相对较低。随着任务数量增加到12，数据速率显著提升，增益约为7.34%；任务数量进一步增加到24时，数据速率的增长幅度趋于平缓，增益仅为1.74%。此外，从图5(b)可见，随着任务数量的增加，元训练所需的仿真运行时间呈线性增长。上述结果表明，当任务数量达到一定规模后，所得元参数已能有效捕捉不同传输环境的通用特征。此时继续增加任务数量对性能的提升有限，反而会增加元训练阶段的计算开销和时间成本。因此，综合权衡GLA算法的性能与计算效率，在实际部署中选择$J=12$或$J=18$可保证数据速率和误块率性能的同时维持较低的训练开销。

最后，为研究GLA算法对外界环境动态变化的自适应能力，本文将实验的前30000个时隙的误块率约束设为$\eta_{\max}=0.1$，并在第30000个时隙将约束调整为$\eta_{\max}=0.06$。GLA算法的数据速率和误块率性能随时隙变化如图6所示。从图6(a)和图6(b)可以看出，在前30000个时隙内，算法快速收敛，且误块率始终保持在0.1以下。当目标误块率变为0.06后，算法迅速调整传输策略，将链路误块率降低至0.06以下。由于误块率要求更为严格，智能体倾向于选择更为保守的MCS策略，导致数据速率出现下降并逐渐趋于稳定。该实验结果充分表明

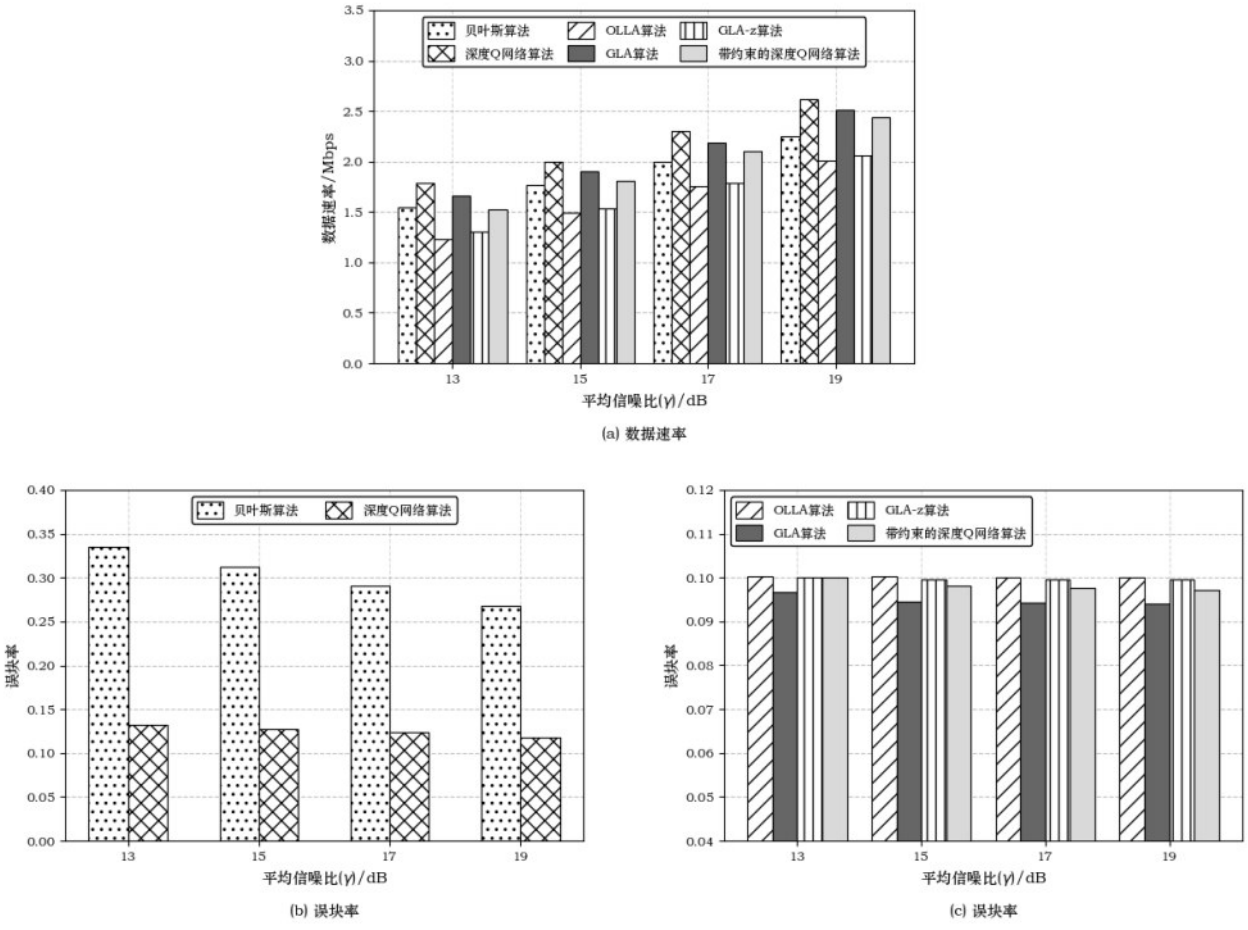

图4　不同平均信噪比下各算法的数据速率与误块率性能

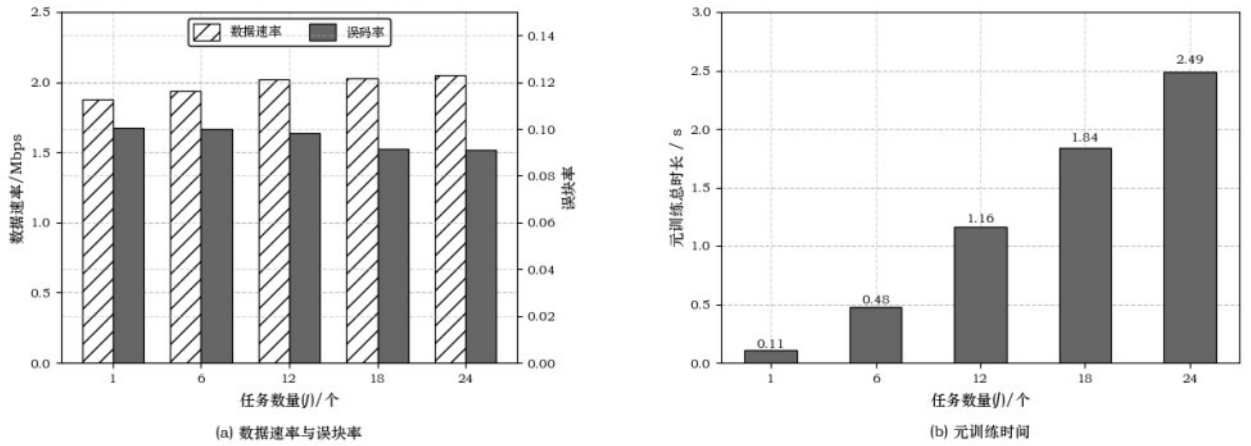

图5　不同任务数量对GLA算法传输性能与训练时长的影响

GLA 算法具备较强的在线学习能力，仅需通过少量样本的在线微调，即可快速适应约束条件的动态变化，表现出优异的自适应能力。

4　结束语

本文面向蜂窝网络下行链路传输场景，研究了非完美 CQI 下的链路自适应问题。首先，将 MCS 选择建模为马尔可夫决策过程，提出了融合元学习与深度强化学习的 GLA 算法，以解决传统算法在不同传输环境中有效性低与泛化性差的问题。其次，针对无线通信系统对可靠性的要求，设计了带约束的 MCS 选择策略。仿真结果表明，在各种传输环境下，所提 GLA 算法均能在满足误块率需求的同时，显著提升收敛速度和数据速率性能，并表现出较良好的鲁棒性和泛化能力。此外，所提算法在环境动态变化时也展现出较强的适应能力，从而验证了其在链路自适应优化中的有效性。

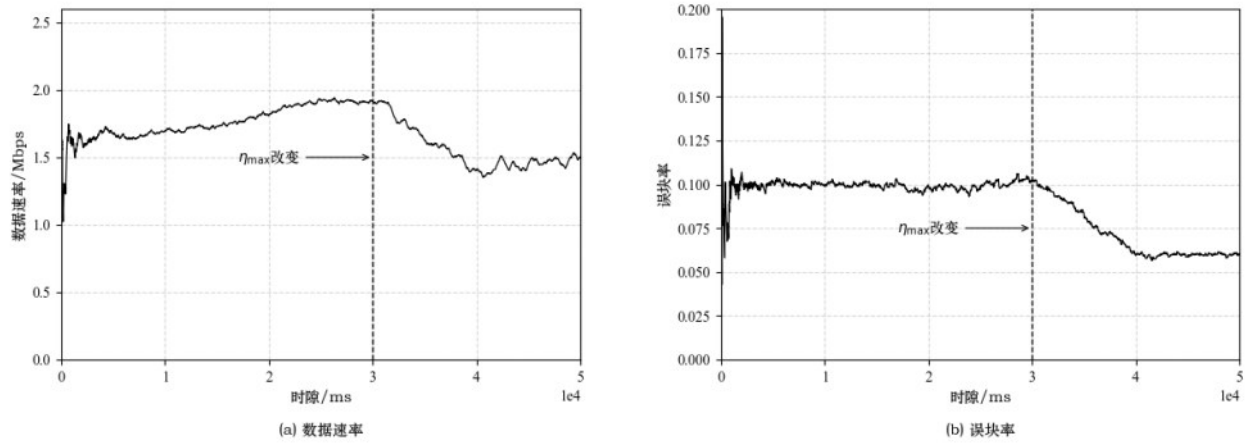

图6　GLA算法的数据速率和误块率性能随时隙变化

参考文献：

[1] WU S, TSOUKANERI G, MOUHOUCHE B. Q-learning based link adaptation in 5G[C]//Proceedings of the 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications. Piscataway: IEEE, 2020: 1-6.

[2] MOTA M P, ARAUJO D C, NETO F H C, et al. Adaptive modulation and coding based on reinforcement learning for 5G networks[C]//Proceedings of the 2019 IEEE Globecom Workshops (GC Wkshps). Piscataway: IEEE, 2019: 1-6.

[3] ELSAYED M, EROL-KANTARCI M. Reinforcement learning-based joint power and resource allocation for URLLC in 5G[C]//Proceedings of the 2019 IEEE Global Communications Conference (GLOBECOM). Piscataway: IEEE, 2019: 1-6.

[4] PRAVEEN S, KHAN J, JACOB L. Reinforcement learning based link adaptation in 5G URLLC[C]//Proceedings of the 2021 8th International Conference on Smart Computing and Communications (ICSCC). Piscataway: IEEE, 2021: 159-163.

[5] KU S, LEE C. Contextual multi-armed bandit-based link adaptation for URLLC[J]. IEEE Transactions on Vehicular Technology, 2024, 73(11): 17305-17315.

[6] CHEN J, MA J, HE Y, et al. Deployment-friendly link adaptation in wireless local-area network based on on-line reinforcement learning[J]. IEEE Communications Letters, 2023, 27(12): 3424-3428.

[7] LIAO Y, YANG Z, YIN Z, et al. DQN-based adaptive MCS and SDM for 5G massive MIMO-OFDM downlink[J]. IEEE Communications Letters, 2023, 27(1): 185-189.

[8] YE X, FU L. Joint MCS adaptation and RB allocation in cellular networks based on deep reinforcement learning with stable matching[J]. IEEE Transactions on Mobile Computing, 2024, 23(1): 549-565.

[9] ZHAO D, QIN H, SONG B, et al. A reinforcement learning method for joint mode selection and power adaptation in the V2V communication network in 5G[J]. IEEE Transactions on Cognitive Communications and Networking, 2020, 6(2): 452-463.

[10] YE X, YU Y, FU L. Deep reinforcement learning based link adaptation technique for LTE/NR systems[J]. IEEE Transactions on Vehicular Technology, 2023, 72(6): 7364-7379.

[11] EL JAMOUS Z, DAVASLIOGLU K, SAGDUYU Y E. Deep reinforcement learning for power control in next-generation WiFi network systems[C]//Proceedings of the MILCOM 2022 - 2022 IEEE Military Communications Conference (MILCOM). Piscataway: IEEE, 2022: 547-552.

[12] KELA P, HÖHNE T, VEIJALAINEN T, et al. Reinforcement learning for delay sensitive uplink outer-loop link adaptation[C]//Proceedings of the 2022 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit). Piscataway: IEEE, 2022: 59-64.

[13] PANG G, YOU L, FU L. An efficient DRL-based link adaptation for cellular networks with low overhead[C]//Proceedings of the 2024 IEEE Wireless Communications and Networking Conference (WCNC). Piscataway: IEEE, 2024: 1-6.

[14] GAO W, ZHENG P, HU Y, et al. A novel link adaptation approach for URLLC: a DRL-based method with OLLA[C]//Proceedings of the 2024 IEEE Wireless Communications and Networking Conference (WCNC). Piscataway: IEEE, 2024: 1-6.

[15] LIANG G, HU J, ZHAO Y, et al. Intelligent link adaptation for integrated data and energy transfer: an enhanced DRL approach for long-term constraints[J]. IEEE Transactions on Communications, 2024, 72 (11): 6956-6972.

[16] ZHOU L, FANG X, HE R, et al. Deep reinforcement learning-based joint frame length and rate adaption for WLAN network[C]//Proceedings of the 2023 IEEE 98th Vehicular Technology Conference (VTC2023-Fall). Piscataway: IEEE, 2023: 1-5.

[17] SAXENA V, TULLBERG H, JALDÉN J. Reinforcement learning for efficient and tuning-free link adaptation[J]. IEEE Transactions on Wireless Communications, 2022, 21(2): 768-780.

[18] 3GPP. Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures: 3GPP TS 36.213 V16.5.0[S]. 2021.

[19] SUTTON R S, BARTO A G. Reinforcement learning: an introduction [M]. 2nd ed. Cambridge, MA: MIT Press, 2018.

[20] GOODFELLOW I, BENGIO Y, COURVILLE A. Deep learning[M]. Cambridge, MA: MIT Press, 2016.

[21] CHO K, VAN MERRIËNBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha: Association for Computational Linguistics, 2014: 1724-1734.

[22] LIN L J. Self-improving reactive agents based on reinforcement learning, planning and teaching[J]. Machine Learning, 1992, 8(3/4): 293-321.

[23] NICHOL A, ACHIAM J, SCHULMAN J. On first-order meta-learning algorithms[EB/OL]. arXiv preprint arXiv:1803.02999, 2018.]

[24] LUONG N C, HOANG D T, GONG S, et al. Applications of deep reinforcement learning in communications and networking: A survey[J]. IEEE Communications Surveys & Tutorials, 2019, 21(4): 3133-3174.

[25] 3GPP. Physical layer procedures for data: 3GPP TS 38.214 V16.2.0[S]. 2020.

[26] KUMAR P S, HOLTZMAN J M. On setting reverse link target SIR in a CDMA system[C]//Proceedings of the 1997 IEEE 47th Vehicular Technology Conference. Piscataway: IEEE, 1997: 929-933.

叶小文（1996– ），男，福建南平人，博士，福建师范大学副教授，主要研究方向为智能无线通信、水声通信组网、通感一体化等。

林恒羿（2002- ），男，福建福州人，福建师范大学硕士生，主要研究方向为无线通信、强化学习等。

吴怡（1970- ），女，福建福州人，博士，福建师范大学教授、博士生导师，主要研究方向为海上/水下通信、信道接入、车联网、智能通信等。