基于 Decision Transformer 模型的 5G 基站节能控制策略优化新范式研究

夏鹏程[1]，朱银涛[1]，于霄[1]，何怡[2]，倪艺洋[3]
（1. 南京理工大学电子工程与光电技术学院，江苏 南京 210094；2. 南京理工大学网络空间安全学院，江苏 江阴 214443；
3. 江苏第二师范学院计算机工程学院，江苏 南京 211200）

摘 要：随着 5G 网络的持续深入应用，基站设备的增长势头迅猛，这不仅催生了通信行业对提升 5G 基站综合能效和实现节能减排的新需求，也对相关厂商提出了更高的要求。虽然传统的强化学习（RL, reinforcement learning）技术有望优化 5G 基站的节能策略，但此类方法需要大量的环境互动和模型训练时间。此外，面对动态变化的基站运行环境，状态和动作空间的变化也会导致 RL 难以学习到较好的策略，且传统 RL 模型的泛化能力也有限。为了解决这些问题，创新性地提出了基于 Decision Transformer（DT）模型的 5G 基站节能控制策略优化新框架，该框架将每个基站对应的状态和动作空间解耦，以适应不同的场景任务，并且基于轨迹先验改进了原始 DT 模型，通过轨迹数据的先验信息来优化模型的期望回报。仿真实验结果表明，所提方法与其他 RL 算法相比，可以在保障用户服务质量的前提下大幅降低系统功耗，且能够在不重新训练的情况下适应未知任务，所提方法在 5G 基站节能决策场景中展现出明显优势和应用潜力。

关键词：5G；基站节能；策略优化；强化学习；DT 模型
中图分类号：TN929.53
文献标志码：A
doi：10.11959/j.issn.2096−3750.2025.00481

Research on the new paradigm for optimizing 5G base station energy-saving control strategies utilizing the Decision Transformer model

XIA Pengcheng[1], ZHU Yintao[1], YU Xiao[1], HE Yi[2], NI Yiyang[3]
1. School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
2. School of Cyberspace and Security, Nanjing University of Science and Technology, Jiangyin 214443, China
3. School of Computer Engineering, Jiangsu Second Normal University, Nanjing 211200, China

Abstract: With the ongoing proliferation of 5G networks, there has been a surge in the deployment of base station equipment. This trend has not only given rise to new demands within the telecommunications industry for enhancing the overall energy efficiency of 5G base stations and achieving energy conservation and emission reduction but also imposed higher standards on related manufacturers. While traditional reinforcement learning (RL) techniques hold promise for optimizing energy-saving strategies for 5G base stations, they require extensive environmental interactions and model training time. Moreover, in the face of dynamically changing base station operating environments, the variability in state and action spaces can make it difficult for RL to learn effective strategies, and the generalization ability of traditional RL models is

收稿日期：2025−01−04；**修回日期：**2025−02−19
通信作者：夏鹏程，pengcheng.xia@njust.edu.cn
基金项目：国家自然科学基金资助项目（No. 62471204）；江苏省基础研究计划（自然科学基金）前沿引领技术基础研究专项（No. BK20212001）；江苏省高等学校基础科学（自然科学）研究重大项目（No. 24KJA510003）
Foundation Items: The National Natural Science Foundation of China (No. 62471204), The Natural Science Foundation on Frontier Leading Technology Basic Research Project of Jiangsu (No. BK20212001), Major Natural Science Foundation of the Higher Education Institutions of Jiangsu Province (No. 24KJA510003)

also limited. To address these challenges, a new framework for optimizing 5G base station energy-saving control strategies based on the Decision Transformer (DT) model was innovatively proposed. The framework decouples the state and action spaces corresponding to each base station for different scenario tasks and improves the original DT model based on the trajectory priors to optimize the expected return of the model through the prior information of the trajectory data. Simulation results demonstrate that compared to other RL algorithms, the proposed method can significantly reduce system power consumption while ensuring the quality of service for users, and it can adapt to unknown tasks without retraining, showcasing the distinct advantages and application potential of our approach in the context of 5G base station energy-saving decision-making.

Key words: 5G, energy saving of base station, strategy optimization, reinforcement learning, Decision Transformer model

0　引言

随着物联网、大数据、人工智能等技术的发展，5G网络作为新一代的移动通信技术，其部署与应用的需求也不断扩大。5G网络以其高速率、大连接、高可靠和低时延的特点，不仅为消费者带来了更加流畅的移动互联网体验，也为工业制造、智慧城市、远程医疗、智能交通等多个领域提供了强大的技术支撑。然而，5G网络的建设和运营也面临着能耗高和成本大的挑战，尤其是在基站功耗（5G基站主设备的功耗相当于4G基站的3~4倍）和电费支出方面[1-2]。为了积极响应国家“双碳”目标，提升5G网络能效并降低运营成本，运营商、通信设备制造商、科研机构以及高等院校等多方主体正积极投身于5G基站节能技术的创新与改进研究，共同探索可持续发展的新路径[3-6]。

近年来，通信运营商和设备商，如中国移动、中兴通讯等公司，提出了设备级、站点级和网络级3种类型的5G节能总体技术体系[3-4]，其中站点级技术因具有灵活性和实时性优势，已成为研究重点领域，其主要技术方案包括载波关停、基站休眠、通道关断和用户迁移等，其核心目标在于通过动态资源调整实现业务需求与能耗的最优匹配。针对此方式，很多研究者提出了具体的节能策略。文献[7]提出了一种功率感知用户关联和可再生能源配置方案，用于优化混合能源供电的异构蜂窝网络中的并网能源。文献[8]提出了一种不对称损耗函数的天线静音策略，通过选择性地激活天线子集来最小化基站功耗。文献[9]介绍了一种下行链路信号传输方案，该方案可以在频谱效率损失很小的情况下显著降低基站的天线传输功耗。另外，很多研究工作[10-13]针对基站的休眠策略进行优化，提出了多级休眠的方案，通过考虑合理切换基站的休眠方式来提高基站的能效。综合对比分析以上文献可知，载波关停方案在轻载时段的节能效果显著但存在业务中断风险，天线静默策略可实现较好的节能增益却受限于信道状态估计精度，而多级休眠机制通过分级功耗模式可明显提升能效，但面临状态切换时延与信令开销的挑战。相比之下，网络级节能则重点从多网协调角度，利用现网的配置和性能统计等基础数据，基于内置的策略算法，在保证业务质量的前提下对小区进行关断，以实现降低现网能耗的目标。然而，无论是站点级还是网络级技术，这些方法大多基于白盒规则和传统机器学习技术，虽然可以带来一定的节能效果，但普遍存在两个关键问题：一是基于简单模型的业务预测难以适配5G网络流量的时空特性；二是静态阈值控制机制会导致在突发业务场景下出现用户服务质量的下降，以及用户等待时延的增加。而深度学习技术不仅能够通过深度神经网络理解用户业务流量模式，还可构建高效的节能决策模型，通过联合优化站点级控制动作与网络级资源调度，在保证用户体验的前提下，实现较传统方法更大的系统能效提升。因此，有必要引入大数据分析和深度学习等技术，融合站点级和网络级节能方式，以寻求5G基站节能的最优策略。

强化学习（RL, reinforcement learning）作为一种通过智能体与环境交互来学习最优策略的强大机器学习范式，已经被越来越多的研究者应用于5G基站的能源调度和节能决策中。文献[14]提出了一种基于Q学习的基站能效优化方法，通过动态调整基站的睡眠和唤醒状态，显著提高了能效。文献[15]提出了一种在线流量感知的深度Q学习方法，以最大限度地实现长期节能，以可忽略的延迟用户数量为代价，实现了可观的节能。文献[16]提出了一种深度迁移RL框架，利用原基站数据训练好的RL模型迁移至新基站进行节能策略优化。文献[17]提出了一种

名为DeepBSC的动态基站控制框架，该框架利用深度RL方法来确定基站的活动/休眠模式，显著提高了能源效率和网络性能。文献[18]使用双向长短期记忆网络来预测用户流量，并使用RL训练的策略来停用空闲的基站。文献[19]介绍了一种用于基站激活的多智能体深度RL方法。然而，RL往往需要大量的现实世界交互，这不仅成本高昂，而且常常难以实现，其试错的学习方式也可能导致训练周期延长。面对环境的复杂性和动态变化，RL在处理高维状态空间和适应新环境等方面也面临着挑战[20]。

为解决RL所面临的问题，Decision Transformer（DT）模型在快速收敛和出色的泛化性方面取得了重大进展，其应用范围已扩展至众多领域，包括智能通信、智慧交通、游戏策略等[20-25]。DT模型引入了一种新颖的序贯决策方法来克服传统RL在利用有限样本和适应新任务方面的低效性，并且通过学习离线样本，显著提升了对复杂环境的认知水平和策略优化能力。DT模型利用序列建模的优势，将历史状态、动作和奖励信息编码为序列输入，以预测未来的最佳决策。这种方法不仅可以处理高维状态空间，而且在随机和动态的环境中也表现得很好[20-25]。文献[24]讨论了相对于离线RL，DT模型在少量样本适应下对未见任务具有强大的泛化能力。此外，文献[25]将DT模型应用于交通流控制，并展示了其在实际应用中的潜力。所以，基于DT模型突出的优势，可以将其引入基站节能场景。面对严苛和动态的基站网络环境，DT模型能够适应相似任务（环境、状态或者动作存在部分差异）而无须重新训练，并能更灵活地利用休眠模式来优化基站的节能策略。

在上述背景下，本文利用DT模型的突出优势，将其融入基站节能决策场景中，设计了一个新颖的基于DT模型的5G基站节能控制策略优化框架，结合DT模型的Transformer解码器网络结构，将环境内每个基站的状态和动作空间解耦，从而解决基站数量动态变化带来的空间维度不一致性，并调整DT模型中的因果掩码机制来确保所提框架能够应对不同环境任务的变化，提升了框架的整体泛化能力。另外，本文还提出了一种轨迹先验辅助的DT模型架构（PDT, prior information-aided decision transformer），通过轨迹先验分布信息减小了期望回报与模型生成动作所产生实际回报的偏差，进一步解决了原始DT模型过度依赖数据集质量的缺陷，使得模型即使用非最佳数据集训练，也能够实现较好的性能。为了验证本文所提方案的有效性，设置了两种不同环境区域，对PDT、DT以及其他RL算法进行了仿真对比实验。实验结果表明，与其他算法相比，所提的PDT方法能够在保证用户服务质量的前提下最多降低120%的系统功耗，并且针对新环境具有良好的泛化迁移能力，在5G基站节能场景中呈现出了显著优势。

1　系统模型

1.1　基站蜂窝网络环境

5G基站蜂窝网络场景示意图如图1所示。本文考虑一个包含M个基站的蜂窝网络环境，环境中所有基站按照标准蜂窝结构排列，并在时隙$t(t=0,1,2,\cdots)$内为N个用户提供服务。5G基站包括激活和休眠两种工作模式，在激活模式下，基站正常工作并为用户提供服务，处于非休眠状态；而在休眠模式下，基站部分组件被关闭，从而导致用户服务的中断。此模式分为多个不同等级，每种等级具有其独特的功耗折扣因子，即从休眠状态恢复工作所需的唤醒时延。假设每个用户只有一根天线，在时隙t内只能接入一个基站$m(m=1,2,\cdots,M)$，并且除非基站服务队列已满或拒绝新用户接入，每个用户倾向接入信号最强的基站。

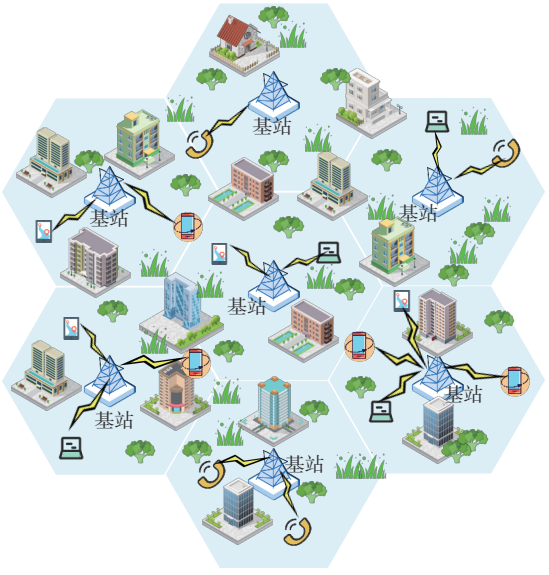


图1　5G基站蜂窝网络场景示意图

1.2　用户流量模型

为构建本文系统模型中的用户到达情况，如泊松分布、对数正态分布、超指数分布以及中断泊松过程等各类分布已被广泛应用。本系统考虑用户到达小区的过程服从泊松过程，这意味着在每个时隙t内，有$X(t)$个用户产生，即

$$X(t)=D(t)U\Delta t \tag{1}$$

其中，$D(t)$为用户密度，U为小区覆盖范围，Δt为时隙长度。

考虑每个用户$n(n=1,2,\cdots,N)$在网络中的服务时间和业务需求分别满足期望为λ_{T}和λ_{R}的指数分布[26]，即对每一个用户，其在网络中的服务时间预算φ_n^{T}和业务需求φ_n^{R}的概率密度分别为$\frac{1}{\lambda_{\mathrm{T}}}\exp\left(-\frac{1}{\lambda_{\mathrm{T}}}t\right)$和$\frac{1}{\lambda_{\mathrm{R}}}\exp\left(-\frac{1}{\lambda_{\mathrm{R}}}x\right)$，用户如果在时延预算$\varphi_n^{\mathrm{T}}$内完成所有业务需求（即传输完所有数据）或者用户网络时延超出预算范围，则会提前离开整个网络环境。

为了模拟网络中的用户到达情况和数据流量模型，假设在时隙t内网络中有$X(t)$个新用户到达，且网络中每个用户按照最小需求的数据传输速率接收服务，则在时隙t内系统共有数据流量大小$\lambda(t)$为

$$\lambda(t)=\sum_{n=1}^{N}r_n \tag{2}$$

$$E(\lambda(t))=E\left(\sum_{n=1}^{N}r_n\right)=E\sum_{t'=0}^{t}\left(\sum_{n=1}^{X(t')}\frac{\varphi_n^{\mathrm{R}}}{\varphi_n^{\mathrm{T}}}\int_{t-t'}^{+\infty}\frac{1}{\lambda_{\mathrm{T}}}\mathrm{e}^{-\frac{1}{\lambda_{\mathrm{T}}}t}\right)=$$

$$\sum_{t'=0}^{t}\left(\sum_{n=1}^{E(X(t'))}\frac{E(\varphi_n^{\mathrm{R}})}{E(\varphi_n^{\mathrm{T}})}\mathrm{e}^{-\frac{1}{\lambda_{\mathrm{T}}}(t-t')}\right)=\sum_{t'=0}^{t}\left(\frac{\lambda_{\mathrm{R}}}{\lambda_{\mathrm{T}}}\mathrm{e}^{-\frac{1}{\lambda_{\mathrm{T}}}(t-t')}E(X(t'))\right) \tag{3}$$

其中，r_n为用户n在整个蜂窝网络中的数据传输速率，t'为从初始时刻到当前时刻t的离散时间索引，$E(\lambda(t))$可以从真实的流量数据集获取，依据上述推导能够求出每一时隙t内网络中新到达的期望用户数量$E(X(t))$，进而模拟出本系统的用户到达状况以及数据流量模型。图2展示了模型的流量负载数据与真实数据对比[27]。

1.3　信道模型

在多基站的蜂窝网络中，为了获得每个用户的速率，根据香农公式，每个用户n从基站$m(m=1,2,\cdots,M)$可获得的数据传输速率r_{mn}为

$$r_{mn}=B\,\mathrm{lb}\,(1+\mathrm{SINR}_{mn}) \tag{4}$$

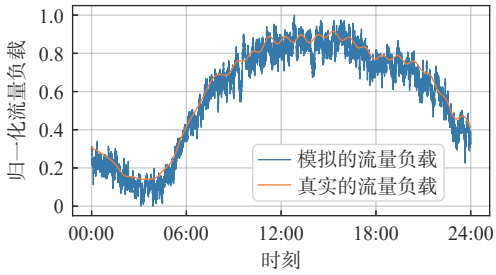


图2　模型的流量负载数据与真实数据对比

$$\mathrm{SINR}_{mn}=\frac{\beta_{mn}p_{mn}}{\sum_{j=1,j\neq m}^{M}\beta_{jn}p_{jn}+{\sigma_{\mathrm{noise}}}^2} \tag{5}$$

其中，SINR_{mn}为基站m到用户n的信噪比，β_{mn}为基站m到用户n的信道所产生的路径损耗，p_{mn}为基站m分配给用户n的发射功率，p_{jn}为其他基站的干扰功率，${\sigma_{\mathrm{noise}}}^2$为噪声功率，$B$为通信系统带宽。

根据第三代合作伙伴计划（3GPP, 3rd Generation Partnership Project）标准[28]，本文假设通信场景为城市微小区（UMi, urban micro），该场景主要捕捉现实生活中真实场景，如车站广场、写字办公楼、城市街道等用户密集的开放区域。根据标准所述，并考虑用户与基站之间存在遮挡，信道模型为非视距（NLOS, non line of sight）传输模型，上述路径损耗β_{mn}可表示为

$$\beta_{mn}=35.3\lg(d_{mn})+22.4+21.3\lg(f_{\mathrm{c}})-0.3(h_n-1.5)+\chi_{\mathrm{SF}} \tag{6}$$

其中，χ_{SF}为阴影衰落，d_{mn}为基站m和用户n之间的空间距离，f_{c}为信息传输载波频率，h_n为用户位置高度。

1.4　基站工作模式与用户关联方式

本文考虑5G基站的工作模式分为激活和休眠两种[12, 29-30]，其中休眠模式是指基站的部分硬件停用从而进入睡眠状态（此模式下基站的天线全部处于关闭状态），该模式可包括多个不同的等级，低等级休眠方式功耗降低少，唤醒时延t_{SM}低，高等级休眠方式则可以减少更多功耗，但是相应地唤醒时延t_{SM}更高。基站在某一休眠状态下经过唤醒时延t_{SM}后转入下一休眠状态。

针对基站的用户关联方式，本文考虑以下几种：1)允许新用户接入；2)拒绝新用户接入；3)断开当前与该基站所有服务用户的连接同时新用户无法加入。原来的用户断开连接或者新用户被拒绝加入后，将会在下个时段寻找新的可接入基站。当基站处于激

活状态时，上述用户关联方式正常执行。但是当基站处于休眠状态时，通常会执行第3种方式，即转移所有服务用户，但是也可以执行前两种方式。针对第2种方式，当前用户可以保留在服务队列中（即等待基站重新激活后，正常提供通信服务）；针对第1种方式，新用户可以加入基站的等待队列中。

1.5 功耗模型

本文考虑每个5G基站的功率消耗分为激活功耗和休眠功耗两部分[12, 29-30]。基站在不同的工作模式下，具体功率消耗P_m可表示为

$$P_m=\begin{cases}\delta_{\mathrm{SM}}P_{\mathrm{const}}, & \text{休眠}\\ P_{\mathrm{const}}+L_mP_{\mathrm{PA}}, & \text{激活}\end{cases} \tag{7}$$

其中，P_{const}为基站m除去负载外的常态功率，L_m为基站m使用天线的数量，P_{PA}为天线功率放大器的消耗，δ_{SM}为基站休眠时与负载无关的常态功率消耗的折扣因子。

1.6 系统优化目标

在多基站通信网络中，当总业务需求或负载很小时，通过采取合适的方法对基站天线组的开关、工作模式的切换以及用户关联的选择等动作策略进行优化，能够很大程度地减小整个通信系统的能量开销。因此，本节的优化目标是最小化整个系统的能耗，具体如下

$$\begin{aligned}&\min_{\{a^{\mathrm{SM}}(t)\},\{a^{\mathrm{AS}}(t)\},\{a^{\mathrm{UA}}(t)\}}\sum_{t=0}^{T}\sum_{m=1}^{M}P_m(t)\Delta t\\ &\text{s.t.}\quad \text{C1:}\quad L_m\leqslant L^{\max},\qquad \forall m=1,2,\cdots,M\\ &\qquad\ \ \text{C2:}\quad r_n^{\mathrm{avg}}\geqslant r_n^{\min},\qquad \forall n=1,2,\cdots,N\end{aligned} \tag{8}$$

其中，优化变量a^{SM}为基站工作模式，a^{AS}为基站的天线开关，a^{UA}为基站的用户关联方式。上述优化问题中，在时隙t内约束每个基站开启的天线数量不超过最大可用数$L^{\max}$，并且约束每个用户进入网络后的平均传输速率r_n^{avg}（即每个用户当前已传输的总数据量与当前网络服务总时长的比值）能满足其最低传输速率要求$r_n^{\min}=\dfrac{\varphi_n^{\mathrm{R}}}{\varphi_n^{\mathrm{T}}}$。

2 基于DT的5G基站节能控制策略优化框架

2.1 马尔可夫决策过程建模

根据上文所述的系统模型及优化问题，本文考虑将该基站节能决策问题建模为马尔可夫决策过程（MDP, Markov decision process）。一个MDP通常包括状态空间S、动作空间A、奖励函数r等。在本文构建的网络系统场景中，将基站的状态定义为系统在任意时刻状态的描述，而动作则包括基站工作模式的切换、基站用户关联方式的选择和基站天线组的开关，奖励则定义为与所有基站总功耗密切相关的函数。具体定义如下。

1)状态空间

状态空间S是所有可能的环境状态的集合。在任一时隙t，每个基站m的状态$s_t^m\in S$包括基站当前休眠状态、下一休眠状态、切换到下一休眠状态所需时延、天线数量、用户关联模式、基站位置、总功率消耗、发射功率、服务的用户数量、用户服务时延、用户数据传输速率、用户最小需求的数据传输速率。

2)动作空间

动作空间A是所有基站可能采取的动作的集合。在任一时隙t，基站m在下一时刻所要采取的动作$a_t^m\in A$包括基站工作模式的切换、用户关联方式设置和基站天线组的开关，具体定义为

$$a_t^m=\left(a^{\mathrm{SM}}(t),a^{\mathrm{AS}}(t),a^{\mathrm{UA}}(t)\right) \tag{9}$$

其中，$a^{\mathrm{SM}}=\{a_{\mathrm{active}}^{\mathrm{SM}},a_{\mathrm{sleep1}}^{\mathrm{SM}},a_{\mathrm{sleep2}}^{\mathrm{SM}},\cdots,a_{\mathrm{sleepx}}^{\mathrm{SM}}\}$表示基站工作模式，包括基站激活模式或休眠模式中的任意一种等级，在基站采取动作a^{SM}后，经过一定唤醒时延后转入下一工作状态。$a^{\mathrm{AS}}=\{a_{\mathrm{open}}^{\mathrm{AS}},a_{\mathrm{shut}}^{\mathrm{AS}},a_{\mathrm{unchanged}}^{\mathrm{AS}}\}$表示天线组的开关，分别为打开一组天线，关闭一组天线和对基站天线不做变动。$a^{\mathrm{UA}}=\{a_{\mathrm{allow}}^{\mathrm{UA}},a_{\mathrm{deny}}^{\mathrm{UA}},a_{\mathrm{disconnect}}^{\mathrm{UA}}\}$表示基站的用户关联模式，分别为基站允许新用户接入，拒绝新用户接入，以及断开与该基站所有用户的连接，没有完成服务的用户会寻找新的基站。

3)奖励函数

奖励函数$r(s_t,a_t)$表示在状态$s_t=(s_t^1,s_t^2,\cdots,s_t^M)$下采取动作$a_t=(a_t^1,a_t^2,\cdots,a_t^M)$所获得的即时反馈。针对上述定义的优化目标，本文定义奖励函数为

$$r(s_t,a_t)=-\Phi(s_t,a_t)\pm\omega\Psi(s_t,a_t) \tag{10}$$

$$\Phi(s_t,a_t)=\sum_{m=1}^{M}P_m(t) \tag{11}$$

$$\Psi(s_t,a_t)=\frac{1}{N}\sum_{n=1}^{N}\begin{cases}\kappa\left(\dfrac{r_n^{\mathrm{avg}}}{r_n^{\min}}-1\right), & r_n^{\mathrm{avg}}\geqslant r_n^{\min}\\ \dfrac{r_n^{\mathrm{avg}}}{r_n^{\min}}-1, & r_n^{\mathrm{avg}}<r_n^{\min}\end{cases} \tag{12}$$

其中，$\Phi(s_t,a_t)$为在时隙t内基站的功率消耗，$\Psi(s_t,a_t)$为保障用户服务质量的惩罚函数，r_n^{avg}为用

户进入网络后的平均传输速率，如果平均传输速率 r_n^{avg} 小于最小需求速率 $r_n^{\min}$，则表示用户服务质量不满足要求，那么奖励上会加上一个负惩罚项；如果平均传输速率大于最小需求速率，尽管满足了服务质量，惩罚项会乘上一个缩放因子 $\kappa(0<\kappa<1)$，ω 为约束惩罚权重。

2.2　基于DT模型的基站节能决策框架

针对系统模型和MDP模型，本文考虑使用离线RL思想进行基站节能控制策略的优化。传统的离线RL方法需要提前收集具备高回报轨迹的专家数据集进行模型的预训练，但是在很多任务或场景中，可能无法收集一定数量的最优数据集，这些数据集往往可能是只包含部分最优轨迹的次优数据集，使用这些数据集训练模型作为决策智能体一定会导致非常低的样本学习效率或者模型根本无法收敛。在本文构建的5G基站节能网络环境中，高回报的轨迹序列和低回报的轨迹序列都可反映状态和执行动作中的内在联系。尽管离线RL不需要覆盖所有可能情况的数据集，但是数据集的多样性和广泛性对于训练出一个有效和具备泛化性的策略是非常重要的。所以，针对上述这种情况，为使模型也能从低回报的轨迹序列中学习与环境相关的知识，提高样本利用和学习效率，本文提出了基于DT模型的5G基站节能控制策略优化框架，如图3所示。

由图3可知，具体流程如下：首先，从先前多个不同环境中通过专家策略收集历史经验数据，构建MDP序列元组，包括状态、动作、奖励等信息（每个时隙 t 都会有一个元组 $<s_t, a_t, r_t>$）；然后，对收集到的数据进行预处理并构建成轨迹数据集，将数据集输入至DT模型中进行预训练；最后，将预训练好的DT模型部署到新环境下，在复杂多变的环境中做出科学合理的5G基站节能控制决策。

DT模型通过序列建模技术，重新定义了离线RL方法，与依赖估计值函数或计算策略梯度的传统RL方法不同，DT模型利用Transformer架构，根据过去的状态、动作和奖励序列来预测未来的动作。这种基于序列的方法使模型能够捕捉轨迹数据中的长期依赖关系和复杂模式，从而使学习过程更加灵活，并具有更强的泛化能力。因此，DT模型的这一特性在无线通信和物联网应用中尤为有用，该模型可以在场景应用中快速适应不断变化的环境并进行有效的策略优化。

DT模型结构如图4所示，其基本网络单元是Transformer模型的解码器。Transformer模型由多个残差连接堆叠的自注意力层组成，每一个自注意力层接收对应输入token的多个嵌入，并且向后一层输出同样大小的嵌入，接收的这些token将会通过线性矩阵将其分别映射到键矩阵 $\boldsymbol{K}$、查询矩阵 $\boldsymbol{Q}$ 和值矩阵 $\boldsymbol{V}$，然后用 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 的归一化点积对 $\boldsymbol{V}$ 进行加权。

在DT模型训练阶段，首先将MDP序列元组数

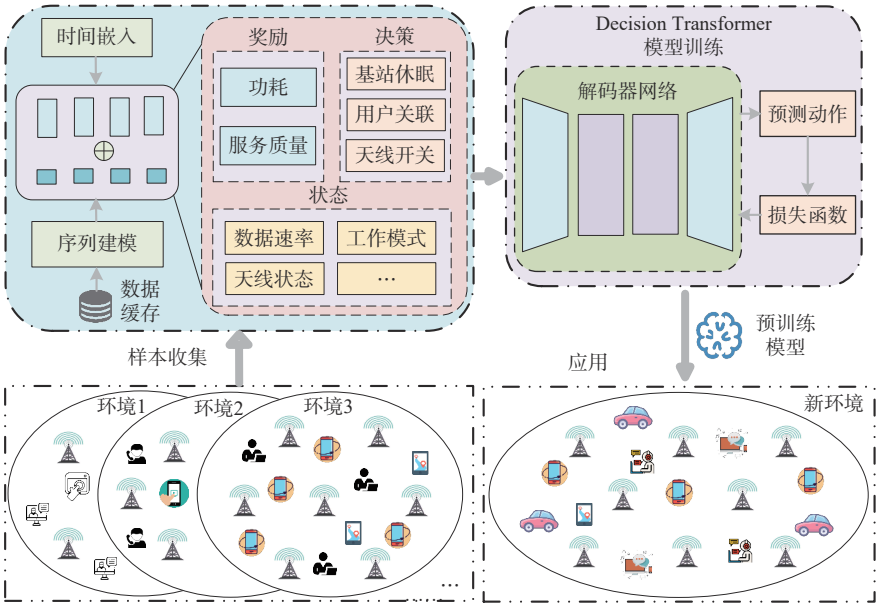


图3　基于DT模型的5G基站节能控制策略优化框架

据进一步处理成轨迹数据集

$$\tau = (\hat{R}_1, s_1, a_1, \hat{R}_2, s_2, a_2, \cdots, \hat{R}_T, s_T, a_T) \tag{13}$$

其中，T为一条完整轨迹的长度，$\hat{R}_t$为重新建模的t时刻累积的未来期望回报，而非通过过去的奖励来生成动作，因此计算如下

$$\hat{R}_t = \sum_{t'=t}^{T} r_{t'} \tag{14}$$

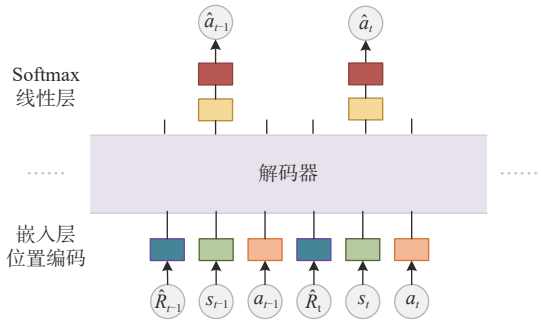


图4　DT模型结构

模型在执行完由当前期望回报$\hat{R}_t$和状态s_t生成的动作后，将目标回报减去所得奖励（$\hat{R}_{t+1} = \hat{R}_t - r_t$）即可得到下一时隙的目标回报，并以此重复直到结束。在模型构建中，目标回报$\hat{R}_t$也会被通过一个单独的线性层映射到嵌入向量，然后与状态和动作一起输入模型的每一层。

本文考虑DT模型能够在多种任务中更好地开展训练，如具有可变区域范围和可变数量的基站中，每个基站均有自身的状态与动作，并且状态和动作是一系列可观测变量的集合。倘若将每一个状态中的每一个变量单独视作一个token，那么输入、输出序列的长度将会极大地限制模型的训练及推理速度与能力；若将它们简单地拼接在一起被当作一个单独的token，那么缺失的基站状态和动作将会用padding方式进行填充，无法有效掩盖填充的内容，此时模型可能会将填充处的基站状态和动作认定为填充的内容而非缺失值，这些都可能对模型的训练和推理结果造成影响。故针对上述问题，本文将每一个基站m的状态和动作解耦出来作为单独的token输入，那么一个完整的轨迹片段τ_{DT}可建模为

$$\tau_{\mathrm{DT}} = \left(\hat{R}_{t-K+1}, s^1_{t-K+1}, s^2_{t-K+1}, \cdots, s^M_{t-K+1}, a^1_{t-K+1}, a^2_{t-K+1}, \cdots, a^M_{t-K+1}, \cdots\cdots, \hat{R}_t, s^1_t, s^2_t, \cdots, s^M_t, a^1_t, a^2_t, \cdots, a^M_t\right) \tag{15}$$

其中，M为蜂窝网络中的基站数量，K为Transformer的上下文长度，由于一条完整轨迹包括很多序列元组，并且期望通过过去一段而非所有时间内的状态变化和动作决策来产生当前时隙t的决策，所以设置合理的上下文长度K（即选取最后K个时间步的数据作为输入）能够有效地加速模型训练和推理，减轻计算和存储的压力。

鉴于基站节能决策环境中动作是离散的，模型最后的线性层无法很好地产生离散动作，为使模型动作得以生成并便于训练，在模型的最后输出中添加Softmax函数以生成动作的概率值，在模型参数θ反向更新时采用负对数似然损失作为模型更新的损失函数

$$L_{\text{loss - DT}} = -\frac{1}{K} q\left(a_k|\hat{R}, s, a\right) \log \pi_\theta\left(\hat{a}_k|\hat{R}, s, a\right) \tag{16}$$

其中，$q(a_k|\hat{R}, s, a)$是数据集中动作a_k的真实分布情况，$\log \pi_\theta\left(\hat{a}_k|\hat{R}, s, a\right)$是模型依据自身策略$\pi_\theta$生成的动作$\hat{a}_k$的分布，$\hat{R}$、$s$、$a$分别为轨迹片段中$a_k$之前所有重标记的轨迹期望回报集合、轨迹片段中$a_k$之前所有基站状态的集合、轨迹片段中$a_k$之前所有基站动作的集合。

2.3　基于轨迹先验改进的DT模型

尽管DT模型在一定程度上克服了离线RL的缺陷，提高了模型在非最优数据集上样本利用效率和模型的训练效果，但该模型在部分场景下，面对以回报条件为指导的任务时，其期望回报可能与模型生成动作所产生的实际回报产生很大的偏差，即动作期望收益与目标收益不一致[31]。这就可能导致即使模型使用了包含高回报轨迹序列但未知分布的次优数据集进行训练，也会在推理阶段无法生成与期望目标一致的动作（即便在给定高期望回报目标的情况下），即达不到既定期望目标。这种偏差性对于5G基站节能决策的现实场景是非常不利的。

因此，针对上述问题，本文在5G基站节能决策的场景中，提出了使用轨迹数据的先验信息来辅助目标回报条件进行动作生成，即PDT模型，PDT模型结构如图5所示。

本文所提的PDT模型在生成动作时不仅依据初始给定的目标回报条件，也依据从预训练的数据集中获取到的期望回报分布Θ^{dis}。PDT中的轨迹先验即轨迹期望回报与期望回报分布的偏差程度，可以利用标准分数（Z分数）进行计算

$$\mathcal{Z}_t = Z\left(\hat{R}_t, \Theta^{\mathrm{dis}}_t\right) = \frac{\hat{R}_t - \mu_t}{\sigma_t} \tag{17}$$

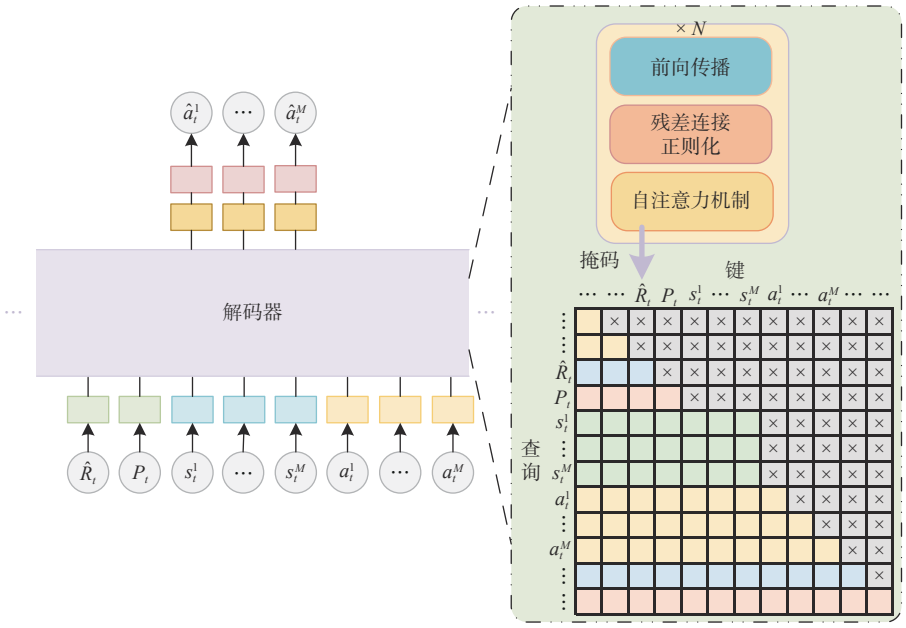


图5　PDT模型结构

其中，μ_t为t时刻期望回报分布的均值，σ_t为t时刻期望回报分布的标准差，Θ_t^{dis}为描述t时刻期望回报分布的参数。在每一个时隙，轨迹元组不仅依据奖励r_t来更新期望回报$\hat{R}_t$，还会根据期望回报分布Θ^{dis}来更新轨迹先验。在推理时，给定一个高期望回报，那么PDT模型除了依据该回报目标生成动作外，还会根据该期望回报的得分做一个辅助决策，如果相对原数据集中期望回报分布Θ^{dis}得分较高，那么模型会认为需要较好的表现才会实现这个期望回报；如果得分较低，则模型会认为自己要生成表现差的动作。关于期望回报分布Θ^{dis}，需要求得所有轨迹序列对应轨迹回报的均值和标准差，所以本文采取Bootstrap方法。该方法核心理念即是不断从现存数据里随机抽取样本，准许重复抽取相同数据来对新的样本集进行模拟，从而估量所需求的统计数据（如平均值、中位数或者标准差等）。

尽管上述轨迹先验使用了Z分数进行标准化，但过高或者过低的得分可能会对模型训练和推理产生影响，这些异常值即便对轨迹序列来说是正常的，也可能会影响模型对所生成动作效果的判断。因此，为了更好地统一高低分之间的差异与范围，本文采用Sigmoid函数将Z分数归一化到0、1之间，那么PDT模型中的轨迹先验P_t可表示为

$$P_t = \frac{1}{1+\exp(-\mathcal{Z}_t)} \tag{18}$$

其中，$\mathcal{Z}_t$为t时刻期望回报$\hat{R}_t$与该时刻期望回报分布Θ_t^{dis}通过标准分数计算得到的标准化值。

假设$x \in X$，$y \in Y$，且存在一个相应法则f，使得每个x都有唯一确定的y与之对应，如果条件概率函数$p(z|x) \geqslant p(z|y)$，那么应该存在$p(z|x,y) \geqslant p(z|x)$。据此，模型中的轨迹先验$P_t$是轨迹期望回报$\hat{R}_t$的一一映射，即$P_t = h(\hat{R}_t)$，相反，轨迹期望回报$\hat{R}_t$也是轨迹先验$P_t$的一一映射$\hat{R}_t = h^{-1}(P_t)$，那么给定序列包括期望回报$\hat{R}_t$和轨迹先验$P_t$，模型学习到的策略$\pi_\theta$生成最优动作$\hat{a}_k$的概率应满足如下关系

$$\pi_\theta(\hat{a}_k|\hat{R},P,s,a) \geqslant \max\left(\pi_\theta(\hat{a}_k|\hat{R},s,a),\pi_\theta(\hat{a}_k|P,s,a)\right) \tag{19}$$

其中，P是轨迹片段中a_k之前所有轨迹先验的集合。

上述推论可以表明：加入先验信息的轨迹序列后，PDT模型能更好地进行策略的学习，减小轨迹期望回报与模型生成动作所产生实际回报的偏差。故在PDT模型训练阶段，轨迹先验P_t也将通过一个单独的线性层映射到嵌入向量，然后与期望回报、状态和动作一起输入进模型的每一层，则一个完整的轨迹片段τ_{PDT}可表示为

$$\begin{aligned}\tau_{\text{PDT}} = \big(&\hat{R}_{t-K+1},P_{t-K+1},s_{t-K+1}^1{:}s_{t-K+1}^M,a_{t-K+1}^1{:}\\&a_{t-K+1}^M,\cdots,\hat{R}_t,P_t,s_t^1{:}s_t^M,a_t^1{:}a_t^M\big)\end{aligned} \tag{20}$$

其中，P_{t-K+1}、P_t分别为轨迹片段（片段长度为K）中的首个和最后一个轨迹先验值。轨迹τ_{PDT}输入

PDT模型后，按照解码器的网络结构，每一层的注意力机制可表示为

$$\text{Attention}\left(\boldsymbol{Q}_{\tau_{\text{PDT}}},\boldsymbol{K}_{\tau_{\text{PDT}}},\boldsymbol{V}_{\tau_{\text{PDT}}}\right)=\text{Softmax}\left(\frac{\boldsymbol{Q}_{\tau_{\text{PDT}}}\boldsymbol{K}_{\tau_{\text{PDT}}}^{\text{T}}}{\sqrt{d_{\boldsymbol{K}}}}\right)\boldsymbol{V}_{\tau_{\text{PDT}}} \tag{21}$$

其中，$\boldsymbol{Q}_{\tau_{\text{PDT}}}$、$\boldsymbol{K}_{\tau_{\text{PDT}}}$和$\boldsymbol{V}_{\tau_{\text{PDT}}}$分别为通过轨迹片段$\tau_{\text{PDT}}$映射的查询、键和值，$d_{\boldsymbol{K}}$为键矩阵维度。

考虑每个注意力机制中的掩码都为因果掩码，则对于上述重构的轨迹序列τ_{PDT}而言，t时刻模型的输出是基于所观测基站m的状态和之前时刻的序列数据来生成对应基站的动作，即在t时刻很显然处于m位置的基站只会关注到它之前位置的基站状态信息$\left(s_t^1,s_t^2,\cdots,s_t^{m-1}\right)$，而不能关注到$\left(s_t^{m+1},s_t^{m+2},\cdots,s_t^M\right)$，这对所有基站来说是不公平的。所以，针对上述情况需要修正注意力机制中相对应的掩码，使得在t时刻基站m可以关注到其他所有基站的状态$\left(s_t^1,s_t^2,\cdots,s_t^{m-1},s_t^{m+1},\cdots,s_t^M\right)$，则每一层的输出可表示为

Output =

$$F\left(\text{Attention}\left(\begin{matrix}\boldsymbol{Q}_{s_t^m},\boldsymbol{K}_{\hat{R}_{t-K+1},P_{t-K+1},s_{t-K+1}^1,a_{t-K+1}^1:\hat{R}_t,P_t,s_t^M,a_{t-1}^M},\\ \boldsymbol{V}_{\hat{R}_{t-K+1},P_{t-K+1},s_{t-K+1}^1,a_{t-K+1}^1:\hat{R}_t,P_t,s_t^M,a_{t-1}^M}\end{matrix}\right)\right) \tag{22}$$

其中，$\boldsymbol{Q}_{s_t^m}$为基站m在t时刻的状态经过嵌入后生成的查询向量，$\boldsymbol{K}_{\hat{R}_{t-K+1},P_{t-K+1},s_{t-K+1}^1,a_{t-K+1}^1:\hat{R}_t,P_t,s_t^M,a_{t-1}^M}$是指轨迹片段$\tau$中从$\hat{R}_{t-K+1}$到$a_{t-1}^M$之间的所有序列元素映射的键，同样$\boldsymbol{V}_{\hat{R}_{t-K+1},P_{t-K+1},s_{t-K+1}^1,a_{t-K+1}^1:\hat{R}_t,P_t,s_t^M,a_{t-1}^M}$是指轨迹片段$\tau$中从$\hat{R}_{t-K+1}$到$a_{t-1}^M$之间的所有序列元素映射的值。

相应地，PDT模型的损失函数可表示为

$$L_{\text{loss - PDT}}=-\frac{1}{K}\sum_{k=1}^{K}q\left(a_k|\hat{R},P,s,a\right)\log\pi_\theta\left(\hat{a}_k|\hat{R},P,s,a\right) \tag{23}$$

综上，PDT模型训练阶段的伪代码如下。

算法1　PDT模型训练算法

输入：离线数据集D，迭代次数N，模型参数θ，批次大小$\mathcal{B}$，上下文本长度K，损失函数L_{loss}，学习率α，期望回报分布Θ^{dis}

初始化模型参数θ

for episode = 1 to N **do**

从D中采样$\mathcal{B}$个上下文长度为K的轨迹片段

for 每一个轨迹片段τ' **do**

重标记轨迹期望回报$\hat{R}$

计算轨迹先验$P\leftarrow\text{Sigmoid}\left(Z\left(\hat{R},\Theta^{\text{dis}}\right)\right)$

重构模型输入$\tau^{\text{input}}\leftarrow(\hat{R},P,s,a)$

end for

将$\mathcal{B}$个样本输入模型，模型生成动作概率分布$\pi_\theta(\hat{a}|\hat{R},P,s,a)$

计算损失函数L_{loss}

更新模型参数θ

end for

输出：最优模型策略π_θ^*

首先，利用Bootstrap算法获取轨迹分布Θ^{dis}。随后，在每次训练迭代期间，从D中随机采样上下文长度为K的$\mathcal{B}$个轨迹片段。对于每个选定的轨迹片段τ'，获得重新标记的期望回报$\hat{R}$，计算轨迹先验P，并将$\hat{R}$和P都添加到轨迹序列输入中。最后，使用损失函数$L_{\text{loss - PDT}}$和梯度下降法更新PDT模型参数θ。

3　实验与分析

3.1　实验环境设置

本文的仿真实验基于一台高性能计算服务器，运行64位的Windows 11操作系统，配备64 GB内存、1 TB硬盘、13代英特尔酷睿i9-13900KF处理器，并配有两块NVIDIA RTX 4090GPU显卡。实验环境基于Python 3.8，并使用PyTorch 2.1.1作为主要深度学习框架。此外，仿真还使用了多个常用的Python库来进行数据处理和结果展示，包括Pandas、Numpy、Matplotlib和Scikit-learn等。

为验证本文所提5G基站节能决策方法的有效性，本文考虑一个1 000 m×1 000 m的矩形区域（区域1），在这个区域内，以蜂窝网络结构部署7个基站，相邻基站间隔为400 m，5G基站仿真参数见表1，基站的4种工作模式对应参数见表2[12, 30]。另外，为了评估PDT和DT在5G基站节能决策场景

表1　　**5G基站仿真参数**

载波频率 f_c/GHz	系统带宽 B/MHz	基站高度（用于计算基站和用户之间的空间距离d/m	用户高度 h_n/m	噪声功率谱密度 $\sigma_{\text{noise}}^2/B$/(dBm·Hz^{-1})	基站常态功率 P_{const}/W	基站天线最大可用数量$L^{\max}$	天线平均输出功率P_{PA}/W
6	100	10	1.5	−174	90	64	0.8

中的泛化性，本文还考虑了一个1 800 m×900 m的矩形区域（区域2），该区域部署了13个基站，并且基站的用户和流量密度产生了一些变动。

表2　基站的4种工作模式对应参数

工作模式	功耗折扣因子δ_{SM}	唤醒时延t_{SM}/ms
Active	1	0
SM1	0.69	1
SM2	0.50	10
SM3	0.29	100

根据本文第2.1节和表2的描述，实验中基站的动作空间设置为$a^{SM}=\{a^{SM}_{active}=0,a^{SM}_{sleep1}=1,a^{SM}_{sleep2}=2,a^{SM}_{sleep3}=3\}$，$a^{AS}=\{a^{AS}_{open}=1,a^{AS}_{shut}=2,a^{AS}_{unchanged}=0\}$，$a^{UA}=\{a^{UA}_{allow}=0,a^{UA}_{deny}=1,a^{UA}_{disconnect}=2\}$。整个实验的基站运行时间周期设置为1周，步长设定为1 ms。本文利用基于近端策略优化（PPO, proximal policy optimization）算法[16]训练好的专家策略与环境进行交互来采集相关样本，从中先选取成功和次优的策略数据作为历史经验数据，形成用于DT和PDT训练的离线数据集，DT和PDT模型训练的超参数见表3。

表3　DT和PDT模型训练的超参数

模型层数	注意力头数	嵌入层维度	激活函数	批大小$\mathcal{B}$	上下文长度K	学习率α
6、8（环境迁移泛化实验）	1	384	ReLU函数	64	20	5×10^{-3}

实验中，除了本文所提的PDT和DT模型，还使用PPO[16]、优势演员—评论家（A2C, advantage actor-critic）[32]和信任区域策略优化（TRPO, trust region policy optimization）[33]3种算法作为对比方案。

3.2　DT模型和PDT模型训练效果分析

为了验证PDT模型相较于DT模型的优化效果，图6、图7分别展示了第1种、第2种数据集分布情况下PDT模型和DT模型条件行为表现对比。从图6、图7中可以看到，在面对不同数据集质量或轨迹分布情况时，原始DT模型的条件行为（即根据期望回报生成动作）表现较差，很难生成与期望回报一致的动作，其实际回报和期望回报产生很大的偏差，即使在训练数据集中包括了最佳的轨迹样本，但受数据集质量影响，模型也可能无法找到最佳的行动策略。相较于原始DT模型，加入轨迹先验信息的PDT模型在期望回报和实际回报的误差上有更佳的表现，有效地克服了原始DT模型的缺陷。

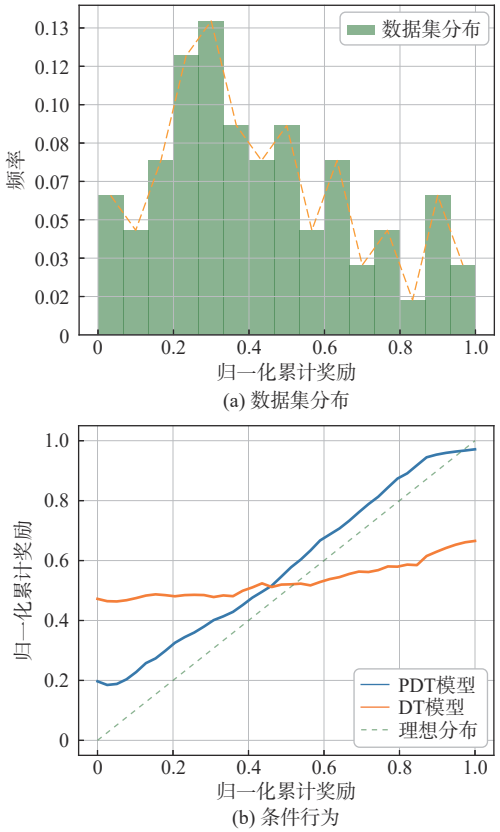


图6　第1种数据集分布情况下PDT模型和DT模型条件行为表现对比

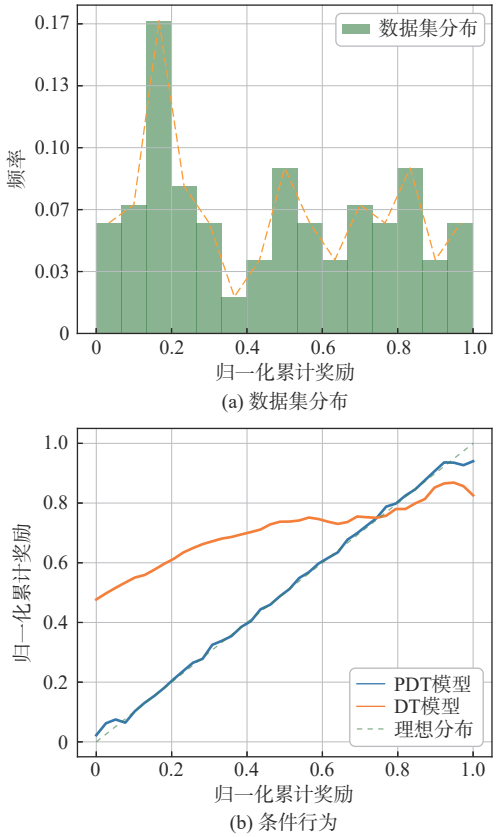


图7　第2种数据集分布情况下PDT模型和DT模型条件行为表现对比

另外，图 8 展示了 DT 模型和 PDT 模型的离线训练收敛曲线。从图 8 中可知，随着训练轮次的增加，两种 DT 模型都能够收敛至稳定，并且由于 PDT 模型具备轨迹先验信息的优势，其收敛速度和训练性能都优于 DT 模型。

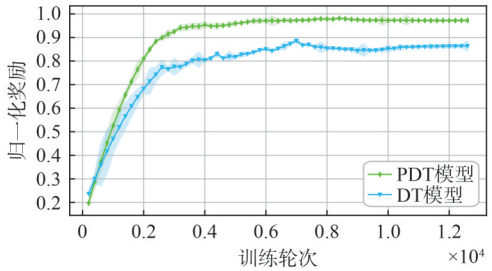


图 8　DT 模型和 PDT 模型的离线训练收敛曲线

3.3　模型推理性能分析

首先，本文针对区域 1 的环境，分别用 PDT、DT、PPO、A2C 和 TRPO 算法进行了 5G 基站节能决策的仿真实验，针对系统的功耗和用户掉线率性能进行了对比，其中用户掉线率表示在给定时间内因未完成服务而提前退出网络的用户数量与离开用户总数的比率。不同算法在线与环境（区域 1，7 个基站）交互下的一周内系统功耗性能对比、用户掉线率性能对比分别如图 9、图 10 所示。从图 9 可以看出，A2C 算法明显未能有效地从环境中学习到一个较好的策略，DT 算法在面对不同数据集质量或轨迹分布情况时，很难产生与期望回报一致的动作，实际回报和期望回报产生很大的偏差，数据集的次优策略轨迹也可能阻碍其学习最优策略，故该方法也未能产生较好的效果。PDT 算法解决了 DT 算法的局限性，减轻了数据集分布的影响，并且与 PPO、A2C 和 TRPO 等算法相比，功耗减少了 5%~70%。从图 10 中可以看出，相比于其他 4 种算法，在用户服务质量方面，PDT 算法也在大部分时段实现了最佳性能，这表明其能够在有效保证用户质量的同时最大限度地降低基站功耗，展示了 PDT 模型强大的轨迹学习能力。

随后，本文使用新环境（区域 2）来评估所提 PDT 算法的泛化性能。DT 算法和 PDT 算法都在先前的区域 1 中进行过预训练（从区域 1 中同时收集具有 5、7 和 10 个基站的节能控制策略数据集），能够灵活地应对这种变化，而不需要再重新训练。然而，由于环境和基站数量的变化，MDP 的状态和动作空间均发生了变化，PPO、A2C 和 TRPO 等算法则需要

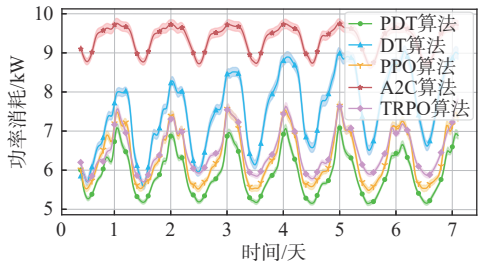


图 9　不同算法在线与环境（区域 1，7 个基站）交互下的一周内系统功耗性能对比

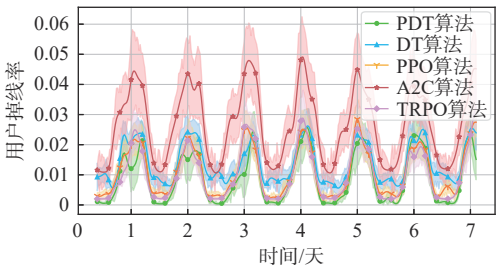


图 10　不同算法在线与环境（区域 1，7 个基站）交互下的一周内系统用户掉线率性能对比

重新训练以适应这一变化。另外，考虑新环境的轨迹回报的分布是未知的，故 PDT 算法中轨迹先验的值只需设置为训练任务上计算的最大轨迹先验值，该先验信息可以指导模型朝着最优的策略行动，并且降低原始 DT 算法对新环境上的奖励及回报分布的依赖性。不同算法在线与环境（区域 2，13 个基站）下的一周内系统的功耗性能对比、用户掉线率性能对比分别如图 11、图 12 所示。从图 11、图 12 可以看出，A2C 和 TRPO 算法都没有学习到较好的节能策略，尽管 TRPO 算法的用户掉线率整体呈现较低的趋势，但是 TRPO 算法没有很好地考虑系统功耗。PDT 算法可以从历史经验中吸取知识并有效应用于新环境，且明显优于原始 DT 算法。与其他 RL 算法相比，PDT 算法能够在保证用户服务质量的同时保证更低的系统功耗（降低了 2%~120%），且具备良好的泛化能力，在 5G 基站节能场景中呈现了显著的优势。

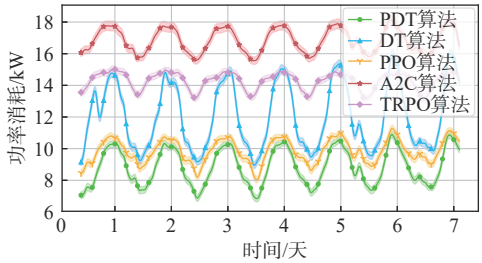


图 11　不同算法在线与环境（区域 2，13 个基站）交互下的一周内系统功耗性能对比

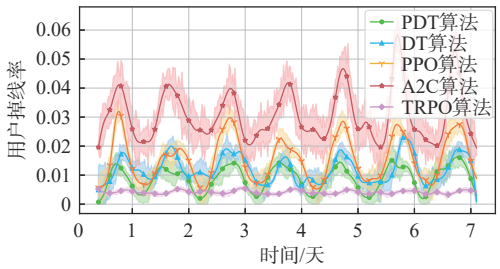


图12　不同算法在线与环境（区域2，13个基站）交互下的一周内系统用户掉线率性能对比

4　结束语

本文研究了DT模型在5G基站节能决策场景中应用的新范式，针对传统RL遇到的问题，提出了优化无线网络系统能耗的PDT框架，该方法基于Decision Transformer模型，同时引入了辅助优化轨迹期望回报的先验信息，使模型在基站节能场景中产生理想的高回报策略。由于每个基站的状态和动作输入是解耦的，本文所提的PDT框架可以有效地适应动态环境中的任务迁移，具备良好的泛化能力。仿真结果表明，与其他RL算法相比，PDT算法既能满足用户服务质量的约束，又能降低能耗，还能在不重新训练的情况下适应未知任务，展现出该方法在基站节能决策场景中的强大应用潜力。笔者未来将针对更复杂的异构无线网络环境、绿色能源辅助的无线基站能耗环境，以及严格约束的基站节能决策环境，研究改进后的DT新框架在这些环境中的有效性和泛化性。

参考文献：

[1] PIOVESAN N, LÓPEZ-PÉREZ D, DE DOMENICO A, et al. Machine learning and analytical power consumption models for 5G base stations[J]. IEEE Communications Magazine, 2022, 60(10): 56-62.

[2] ISRAR A, YANG Q, ISRAR A. Renewable energy provision and energy-efficient operational management for sustainable 5G infrastructures[J]. IEEE Transactions on Network and Service Management, 2023, 20(3): 2698-2710.

[3] 中兴通讯股份有限公司. PowerPilot 4G/5G网络节能降耗技术白皮书[R]. 2020.
ZTE Corporation. PowerPilot: white paper on energy saving and consumption reducing technologies for 4G/5G networks[R]. 2020.

[4] 中国移动研究院. 5G基站节能技术白皮书[R]. 2020.
China Mobile Research Institute. White paper on energy-saving technologies for 5G base stations[R]. 2020.

[5] Huawei Technologies Co., Ltd. 5G power whitepaper[R]. 2019.

[6] 刘俊勇, 潘力, 何迈. 能源物联网及其关键技术[J]. 物联网学报, 2020, 4(4): 9-16.
LIU J Y, PAN L, HE M. Internet of energy things and its key technologies[J]. Chinese Journal on Internet of Things, 2020, 4(4): 9-16.

[7] QIU B, MAO S X, XIAO H L, et al. Power-aware user association and renewable energy configuration for the optimized on-grid energy in hybrid-energy heterogeneous cellular networks[J]. IEEE Systems Journal, 2024, 18(1): 162-173.

[8] RAJAPAKSHA N, MOHAMMADI J, WESEMANN S, et al. Minimizing energy consumption in MU-MIMO via antenna muting by neural networks with asymmetric loss[J]. IEEE Transactions on Vehicular Technology, 2024, 73(5): 6600-6613.

[9] LEE B M, YANG H. Energy-efficient massive MIMO in massive industrial Internet of things networks[J]. IEEE Internet of Things Journal, 2022, 9(5): 3657-3671.

[10] ANDERSSON G, VÄSTBERG A, DEVLIC A, et al. Energy efficient heterogeneous network deployment with cell DTX[C]//Proceedings of the 2016 IEEE International Conference on Communications (ICC). Piscataway: IEEE Press, 2016: 1-6.

[11] FRENGER P, HELMERSSON K W. Massive MIMO muting using dual-polarized and array-size invariant beamforming[C]//Proceedings of the 2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring). Piscataway: IEEE Press, 2021: 1-6.

[12] PEESAPATI S K G, OLSSON M, MASOUDI M, et al. An analytical energy performance evaluation methodology for 5G base stations[C]//Proceedings of the 2021 17th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob). Piscataway: IEEE Press, 2021: 169-174.

[13] PERVAIZ H, ONIRETI O, MOHAMED A, et al. Energy-efficient and load-proportional eNodeB for 5G user-centric networks: a multilevel sleep strategy mechanism[J]. IEEE Vehicular Technology Magazine, 2018, 13(4): 51-59.

[14] AGGARWAL A, SELVAMUTHU D. Energy optimization of a base station using Q-learning algorithm[C]//Proceedings of the 2023 17th International Conference on Telecommunications (ConTEL). Piscataway: IEEE Press, 2023: 1-5.

[15] MASOUDI M, SOROUSH E, ZANDER J, et al. Digital twin assisted risk-aware sleep mode management using deep Q-networks[J]. IEEE Transactions on Vehicular Technology, 2023, 72(1): 1224-1239.

[16] WU D, XU Y T, JENKIN M, et al. Energy saving in cellular wireless networks via transfer deep reinforcement learning[C]//Proceedings of the GLOBECOM 2023-2023 IEEE Global Communications Conference. Piscataway: IEEE Press, 2023: 7019-7024.

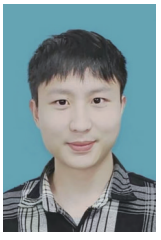
[17] WU Q, CHEN X, ZHOU Z, et al. Deep reinforcement learning with spatio-temporal traffic forecasting for data-driven base station sleep control[J]. IEEE/ACM Transactions on Networking, 2021, 29(2): 935-948.

[18] WANG X Y, LYU B C, GUO C, et al. A base station sleeping strategy in heterogeneous cellular networks based on user traffic prediction[J]. IEEE Transactions on Green Communications and Net-

working, 2024, 8(1): 134-149.
[19] LOTFOLAHI A, FERNG H W. A multi-agent proximal policy optimized joint mechanism in mmWave HetNets with CoMP toward energy efficiency maximization[J]. IEEE Transactions on Green Communications and Networking, 2024, 8(1): 265-278.
[20] ZHANG J, LI J, WANG Z, et al. Decision transformer for wireless communications: a new paradigm of resource management[J]. IEEE Wireless Communications,2025,32 (2):180-186.
[21] CHEN L L, LU K, RAJESWARAN A, et al. Decision transformer: reinforcement learning via sequence modeling[C]//Advances in Neural Information Processing Systems (NeurIPS). Cambridge: MIT Press, 2021: 15084-15097.
[22] LEE K H, NACHUM O, YANG M J, et al. Multigame decision transformers[C]//Advances in Neural Information Processing Systems (NeurIPS). Cambridge: MIT Press, 2022: 27921-27936.
[23] XIA P C, ZANG M F, ZHAO J, et al. Decision transformer-based efficient data offloading in LEO-IoT[J]. Entropy, 2024, 26(10): 846.
[24] XU M D, SHEN Y K, ZHANG S, et al. Prompting decision transformer for few-shot policy generalization[EB]. arXiv preprint, 2022, arXiv: 2206.13499.
[25] NAWAR M, SAAD A, ELHALAWANY B M, et al. Decision transformer-powered traffic control system[C]//Proceedings of the 2023 11th International Japan-Africa Conference on Electronics, Communications, and Computations (JAC-ECC). Piscataway: IEEE Press, 2023: 138-141.
[26] 曾德泽, 李跃鹏, 赵宇阳, 等. 基于强化学习的高能效基站动态调度方法[J]. 计算机科学, 2021, 48(11): 363-371.
ZENG D Z, LI Y P, ZHAO Y Y, et al. Reinforcement learning based dynamic basestation orchestration for high energy efficiency[J]. Computer Science, 2021, 48(11): 363-371.
[27] XU F L, LI Y, WANG H D, et al. Understanding mobile traffic patterns of large scale cellular towers in urban environment[J]. IEEE/ACM Transactions on Networking, 2017, 25(2): 1147-1161.
[28] 3GPP. Radio access network: study on channel model for frequencies from 0.5 to 100 GHz v15.0.0：TR 38.901[S]. 2018.
[29] 温醒醒, 贾亦真, 黄康. 基于异构蜂窝业务量预测的基站休眠策略[J]. 重庆邮电大学学报(自然科学版), 2021, 33(6): 921-929.
WEN X X, JIA Y Z, HUANG K. Base station sleeping strategy based on heterogeneous cellular traffic prediction[J]. Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition), 2021, 33(6): 921-929.
[30] TOMBAZ S, FRENGER P, OLSSON M, et al. Energy performance of 5G-NX radio access at country level[C]//Proceedings of the 2016 IEEE 12th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob). Piscataway: IEEE Press, 2016: 1-6.
[31] WANG Y F, YANG C, WEN Y, et al. Critic-guided decision transformer for offline reinforcement learning[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(14): 15706-15714.
[32] CHEN Z Y, HU J, MIN G Y. Learning-based resource allocation in cloud data center using advantage actor-critic[C]//Proceedings of the ICC 2019 - 2019 IEEE International Conference on Communications (ICC). Piscataway: IEEE Press, 2019: 1-6.
[33] MENG W J, ZHENG Q, SHI Y, et al. An off-policy trust region policy optimization method with monotonic improvement guarantee for deep reinforcement learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(5): 2223-2235.

[作者简介]

夏鹏程（1993– ），男，南京理工大学电子工程与光电技术学院博士生，主要研究方向为无线通信系统节能、工业互联网、工业区块链。



朱银涛（1999– ），男，南京理工大学电子工程与光电技术学院硕士生，主要研究方向为无线决策大模型、强化学习、大语言模型、信息与通信系统。

于霄（2000– ），男，南京理工大学电子工程与光电技术学院硕士生，主要研究方向为深度强化学习、无线通信系统、5G基站节能。

何怡（1998– ），男，南京理工大学网络空间安全学院硕士生，主要研究方向为强化学习、5G基站节能决策。

倪艺洋（1986– ），女，博士，江苏第二师范学院计算机工程学院教授，主要研究方向为无线通信、物联网等。