# 基于强化学习的能量采集传感器传输功率控制方法

刘南君，贾日恒，许文韬，李明禄
（浙江师范大学计算机科学与技术学院，浙江 金华 321004）

**摘　要：**传感器可以通过能量收集技术从周围环境中采集能量，但自然环境中的能源供给通常具有不稳定性。为实现有效的功率控制，使传感器长期运行的同时提升数据吞吐量等性能指标，设计了基于强化学习的功率控制策略。考虑一个端到端通信系统，发送节点采集能量存储到电池中以用于数据传输，同时持续缓存待发送数据。实际应用中，通常无法完整地预知能量和数据到达的过程。该研究中发送节点仅能获取已收集能量、电池电量、已采集数据、数据缓存量、信道增益等当前状态信息，并基于此进行决策。采用了柔性演员–评论家（SAC, soft actor-critic）算法控制传输功率，并设计了合适的奖励函数和动作剪裁方法。仿真实验结果表明，该算法在性能上优于基线策略，并在部分场景中接近理论最优解。

**关键词：**柔性演员–评论家；无线传感器网络；能量采集；强化学习；功率控制
**中图分类号：**TP393
**文献标志码：**A
**doi：**10.11959/j.issn.2096−3750.2025.00445

# RL based transmission power control for energy harvesting sensor

LIU Nanjun, JIA Riheng, XU Wentao, LI Minglu
School of Computer Science and Technology, Zhejiang Normal University, Jinhua 321004, China

**Abstract:** Sensors can harvest energy from the surrounding environment, but the energy supply is always unstable. To achieve effective power control of sensors and enhance their performance metrics, such as data throughput, while ensuring long-term life, a reinforcement learning-based power control strategy was designed. Assume an end-to-end communication system, the sender harvests energy, stores it in a battery for data transmission, and continuously buffers data. In practical scenarios, the arrival of energy and data is random and unpredictable. In this study, the current state was only observed via the sender, which included harvested energy, battery level, collected data, data cache level, and channel gain. Decisions were made solely based on these limited observations. The soft actor-critic (SAC) algorithm was used to control transmission power, with an appropriate reward function and action clipping method. Experimental results demonstrate that the proposed algorithm outperformes baseline strategies and approaches the theoretical optimal in certain scenarios.

**Key words:** SAC, wireless sensor network, energy harvesting, reinforcement learning, power control

## 0　引言

在过去几十年里，低功耗微电子技术取得了显著进展，促进了无线设备和传感器的数量急剧增长，这些技术广泛应用于民用基础设施、家庭自动化设备、电子消费品、移动可穿戴设备、工业和农业监控等多个领域[1]。这些无线设备大多数依赖电池供电，但由于电池容量有限，大规模无线通信网络存在严重的性能瓶颈，特别是在某些特殊环境中，定期手动更换电池或进行有线充电的方

**收稿日期：**2024−09−19；**修回日期：**2024−11−29
**通信作者：**贾日恒，rihengjia@zjnu.edu.cn
**基金项目：**国家自然科学基金资助项目（No. 62272417）
**Foundation Item:** The National Natural Science Foundation of China (No. 62272417)

式成本高昂，甚至不可行，如对植入人体的传感器充电[2]。

能量收集（EH, energy harvesting）技术被视为实现无线通信设备可持续运作的关键技术之一。EH 技术具体定义为一种从周围环境中收集能量并转化为电能的技术。无线通信节点通过 EH 技术可以利用压电、射频能量来充电，也可利用自然能源，如太阳能、热能、风能等，以环保的方式为电池充电[3]。文献[4-5]探讨了基于太阳能供电的 EH 系统，这些研究利用太阳能电板和基于太阳能的直流到直流降压转换器来实现太阳能收集。另外，文献[6-7]研究了基于热能的 EH 系统，文献[8]研究了基于机械能的 EH 系统。EH 传感器有克服传感器节点能量限制的潜力，从而适应各种极端环境。然而，除太阳能外，多数情况下无线通信设备所能收集的自然能源都具有稀缺、不稳定、不可预测且分布不均匀的特性，这使优化 EH 系统的性能充满了挑战。

在基于 EH 的端到端无线通信系统中，功率控制方法可以分为离线和在线两大类。离线方法致力于在有限时间范围内设计最优功率控制策略，并且假设在整个时间范围内的所有相关信息都是已知的[9-10]。在这些假设下，功率控制问题可以归结为优化给定性能指标的静态优化问题，并可以通过传统优化技术来解决。然而，因为需要提前知道关于能量到达和信道状态的所有相关信息[11]，离线方法在实际应用中受到限制。

在线方法考虑 EH 过程以及信道增益等相关信息的随机性和不可预知性，更贴近实际运行节点所面临的情况。一些在线模型假设可以获取系统状态的统计信息[12-15]，但在实际场景中，这些统计信息可能难以获取。即使在特定时间内能够获得统计信息，其规律也容易受到外界因素的干扰。此外，系统环境本身也在不断变化，系统不可能将所有影响因素都纳入考量。以太阳能 EH 系统为例，能量的收集不仅依赖地理位置（如太阳辐射强度），还受天气条件（如晴天或阴天）和日出日落时间等因素的影响。基于强化学习（RL, reinforcement learning）的算法能够较好地适应这种未知且动态变化的环境。RL 算法通过不断与环境交互和学习，能够根据实时状态调整策略，从而实现复杂多变环境中的能量管理和功率控制的最优方案。这种方法具有良好的鲁棒性，能够有效地适应实际应用场景中的不确定性和变化。

现有的基于强化学习的算法通常假设数据流量负载是无限的，即只要具备足够的能量，系统便能够持续进行数据传输。然而在实际的无线传感器网络中，数据流量负载通常表现出随机性和动态性，即使在能量充足的情况下，也未必始终存在足够的数据需要传输。考虑一个能够从环境中采集能量的无线传感器节点，该节点不仅需要生成并传输自身的感知数据，还可能作为中继节点，协助传输其他节点的数据。节点的待发送数据是随机且动态的。因此，数据流量负载的动态性与可用能量的限制都对网络性能产生影响[16]。为此，在设计算法时需要同时考虑数据负载和可收集能量的随机动态特性，以有效地提升网络的整体性能和资源利用效率。

区别于现有研究，本文针对数据负载有限的情况，考虑了数据负载、能量采集和信道状态信息的随机性，研究了在此条件下端到端能量收集通信系统中的在线功率控制问题，目标是最大化系统的数据吞吐量。本文的主要贡献包括以下 3 个方面。

1) 同时考虑能量和数据分多次随机到来、能量和数据都有存储上限且可能溢出、信道增益随机变化的情况。系统性地研究了端到端能量收集通信系统的数据吞吐量优化问题。

2) 将问题建模为强化学习问题，并且基于柔性演员−评论家（SAC, soft actor-critic）算法设计了解决方案，用于优化端到端能量收集通信系统中的吞吐量。仿真实验结果表明，该方案能够在未来的数据、能量、信道状态等相关信息未知的情况下，有效地优化传输功率。本文通过与其他强化学习算法（如双延迟深度确定性策略梯度（TD3, twin delayed deep deterministic policy gradient）算法、近端策略优化（PPO, proximal policy optimization）算法）的对比仿真实验，进一步验证了 SAC 算法在解决该问题上的优越性。

3) 设计了一种专门针对端到端能量收集通信系统的强化学习环境，包含输入输出数据的归一化方法、动作剪裁与归一化策略，并对奖励机制的各个维度及其平衡进行了详细分析。该框架可为能量收集通信系统中基于强化学习的研究提供有价值的参考。

## 1　相关工作

许多研究[17-18]指出，减少对传感器配置的手动干预是提高无线传感器网络性能的重要研究方向。文献[19]首先提出了基于能量预测模型的自适应算法，该算法能够根据预测的能量动态地调整节点的占空比，并引入了能量中性操作（ENO, energy neutral operation）的概念，将其定义为节点收集的能量大于或等于其消耗的能量。在此基础上，文献[20]进一步提出了能量中性距离（ENP, energy neutral performance）的概念，定义为当前操作状态与ENO状态之间的偏差。该研究将ENP引入奖励函数，并基于SARSA（state-action-reward-state-action）算法实现学习优化以提升系统的占空比。此外，该研究还特别关注天气条件对电池状态和自然能量收集的影响，并将天气预报纳入环境状态的建模，以提高预测精度和系统鲁棒性。文献[21]研究了在信道状态信息和能量到达信息的统计信息未知的条件下，使用SAC算法使系统的决策实现ENO，以最大化系统运行的占空比。

文献[22]使用Q学习算法调整传感器节点的睡眠时间，以降低能耗并减少传感器节点的失效率。该研究收集了光强度和超级电容器电压的测量值，并利用这些数据建立了模拟环境，最终在真实环境中进行了部署实验。文献[23]同样使用了Q学习方法，该方法利用一组二元函数拟合Q值以处理连续状态空间，但动作空间仍是离散的。为了实现更精细的控制，需要增加离散坐标的数量，这可能导致计算中的维数灾难问题。文献[24]使用线性函数拟合离散状态的价值函数，在处理连续状态空间的同时，降低了算法的能量开销。文献[25]提出了一种基于深度Q网络（DQN, deep Q-network）的策略来分配传输功率，并根据所获得的信息自适应地调整多元调制级别，以实现系统最大吞吐量。

文献[26]考虑了由单个基站和多个EH用户设备构成的简单上行链路系统，提出了基于DQN算法的功率分配策略，并重点考虑了实际电池的电量自损耗问题。该研究采用长短期记忆网络（LSTM, long short-term memory）对实际电量进行预测，并将预测结果作为智能体状态的输入，为智能体决策提供支持。文献[11]采用了改进演员–评论家架构的深度确定性策略梯度（DDPG, deep deterministic policy gradient）算法对包含能量中继节点的EH系统功率优化问题进行了研究。该研究以最大化系统的长期净比特率为目标，并基于真实的太阳光照数据进行了仿真验证。

文献[27]研究了如何在特定时间范围内最大化两跳半双工中继节点的数据吞吐量。为有效解决该问题，该文献采用了异步优势SAC（A3C, asynchronous advantage actor-critic）算法来解决这个问题。在该研究中，尽管中继节点具备数据缓存功能，但源节点的数据是无限的，并且由中继节点自行决定是否从源节点收集数据。因此，该问题的最优解不存在数据溢出和能量溢出的问题，其离线场景也可以转换为一个凸优化问题，并求出离线最优解作为对照。文献[28]研究了另一种中继节点系统性能优化问题，其中一个中继节点同时服务于多个源节点和多个接收节点。该文献将这个基于EH的中继通信系统链路选择问题转化为强化学习问题，优化目标是最大化系统长期平均效率。

文献[29]提出了一种基于平均场多智能体深度强化学习（MARL, multi agent reinforcement learning）的框架，用于优化大规模EH网络中的在线功率控制策略。由于该网络缺乏全局信息并且节点之间的状态交换开销过大，该研究通过分布式学习方法，使每个节点仅凭本地信息进行功率控制，并最大化系统的吞吐量。文献[30]也采用了基于多智能体深度强化学习的方法，旨在优化基于同时无线信息与能量传输（SWIPT, simultaneous wireless information and power transfer）技术的端到端通信蜂窝网络中的能量效率，其设备之间存在互相干扰，该方法通过多智能体之间的协作学习，有效地处理了干扰问题，并优化了资源分配，从而提升了网络的性能。

## 2　系统模型

受限于通信节点的能量和计算能力，训练任务通常无法在节点端完成。系统架构如图1所示，其中，传感器作为发送节点，基站作为接收节点，而智能体的训练和决策都在服务器端完成，并通过基站将决策指令回传给传感器。这种架构适用于通信时延小的有限规模网络。另一种架构方案是服务器定期将神经网络的参数广播到传感器节点，由传感器自主进行控制。这种方式的优势是能够实现更及

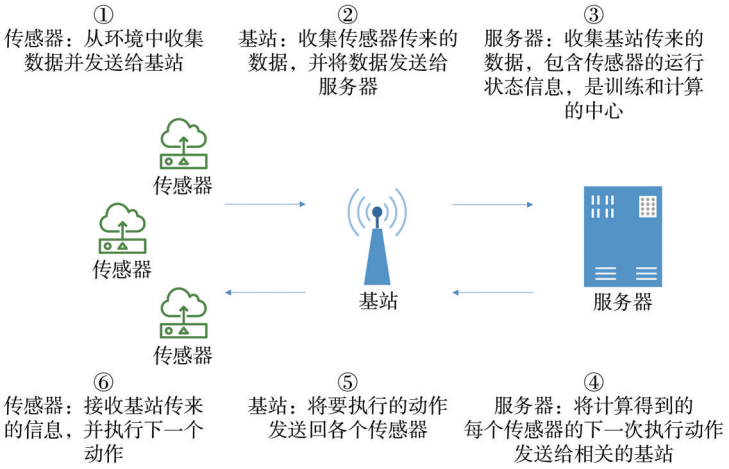


图1　系统架构

时的控制响应，并能够支持更大规模的网络部署。然而，这种方法会导致传感器的能量消耗增加，因此，采用何种方案需要在实际应用中进行权衡。

本文考虑由发送节点$S$和接收节点$R$组成的端到端通信场景（该场景可以扩展至多个发送节点和一个接收基站的配置，如图1所示）。端到端系统示意图如图2所示，在端到端通信系统中，发送节点$S$将收集到的数据和能量分别保存在有限的数据缓存和有限容量的电池中，因此，系统吞吐量受到能量和数据负载的共同约束。本文使用收集—存储—使用（HSU, harvest-store-use）模型[31]，即当前收集的能量和数据只能在后续的时间段中使用。时间被划分为等长的时隙，以$t_i$表示每个时隙的开始时刻。在每个时隙新收集的能量和数据分别表示为$E_i$和$D_i$，发送节点$S$的发送功率为$p_i$。因此，电池中的电量可以表示为

$$B_{i+1}=\min\{B_i-\tau p_i+E_i,B_{\max}\} \tag{1}$$

其中，$B_i$为$t_i$时刻电池的电量，$\tau$为一个时间间隔的长度，$p_i$为发送功率，$B_{\max}$为电池容量上限。

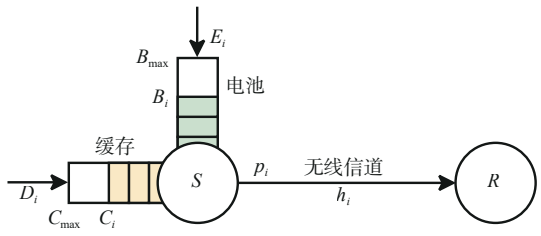


图2　端到端系统示意图

本文利用马尔可夫过程模拟数据产生和能量收集的过程。与其他数据生成方法相比，这种方法更能够模拟动态变化的环境，并减少可获得的统计信息。具体而言，本文使用随机游走模拟数据产生和能量收集的过程。$E_i\in\mathcal{E}_n\triangleq[0,E_{\max}]$是服从正态分布$f_{\mathcal{E}_n}$的连续随机变量，其均值和标准差分别为$E_{i-1}$和$\sigma_1$。$D_i\in\mathcal{D}_n\triangleq[0,D_{\max}]$是服从正态分布$f_{\mathcal{D}_n}$的连续随机变量，其均值和标准差分别为$D_{i-1}$和$\sigma_2$。信道增益$h_i$的统计信息对决策的影响比较独立，与其他因素无相互影响。为简化问题，本文采用固定均值的正态分布对信道增益进行模拟。信道增益的变化过程为$h_i\in\mathcal{H}_n\triangleq[0.1,1.0]$，$\mathcal{H}_n$是一个连续随机变量，服从正态分布$f_{\mathcal{H}_n}$，其均值为$h_c$，标准差为$\sigma_3$。

在本文中，仅考虑由数据传输过程引起的能量消耗，忽略其他形式的能量损耗，即假设所有收集到的能量均用于数据传输。本文假设数据传输过程中的噪声为独立同分布的加性高斯白噪声（AWGN, additive white Gaussian noise）。

在一个时间间隔$\tau$内，发送的数据量表示为

$$R_i(p_i,h_i)=\tau\,\mathrm{lb}\left(1+\frac{|h_i|^2p_i}{\sigma^2}\right) \tag{2}$$

其中，$h_i$为信道增益，$p_i$为发送功率，$\sigma^2$是噪声的方差。

因此，缓存中的数据量可以表示为

$$C_{i+1}=\min\{C_i-R_i+D_i,C_{\max}\} \tag{3}$$

其中，$C_{\max}$为缓存容量上限，$R_i$为发送的数据量，$D_i$为时隙内收集的数据量。

## 3　问题建模

在缺乏能量到来、数据到来和信道统计信息的前提下，本文通过在线学习节点的能源使用策略，实现从发送节点$S$到接收节点$R$的长期数据吞吐量

的最大化。在每个时隙中，数据吞吐量由发送功率$p_i$和信道增益$h_i$共同决定，而发送功率$p_i$又受到电池中剩余能量和数据缓存中剩余数据量的共同限制。根据式(2)，能够确定每个时隙内的数据吞吐量，于是问题可以定义为

$$\max \sum_{i=1}^{\infty} \tau \operatorname{lb}\left(1+\frac{|h_i|^2 p_i}{\sigma^2}\right) \tag{4}$$

$$\text{s.t. } \tau p_i \leqslant B_i \tag{4a}$$

$$R_i(p_i, h_i) \leqslant C_i \tag{4b}$$

$$B_{i+1} = \min\{B_i - \tau p_i + E_i, B_{\max}\} \tag{4c}$$

$$R_i(p_i, h_i) = \tau \operatorname{lb}\left(1+\frac{|h_i|^2 p_i}{\sigma^2}\right) \tag{4d}$$

$$C_{i+1} = \min\{C_i - R_i + D_i, C_{\max}\} \tag{4e}$$

式(4a)表明第$i$个时隙的数据发送功率$p_i$受到时隙开始时电池中的剩余电量$B_i$的限制，式(4b)表示$p_i$受到时隙开始时缓存中的数据量$C_i$的限制。在离线情况下，每个时刻开始时的电池电量$B_i$、缓存数据量$C_i$和每个时隙的信道增益$h_i$都是已知的。然而，在本文研究的在线场景中，未来的数据量$D_i$和采集能量$E_i$具有随机性。于是由式(4c)和式(4e)可知，下一时隙开始时电池电量$B_{i+1}$和缓存数据量$C_{i+1}$的变化也具有随机性。这意味着，在不同时隙中，可行解的范围受到当前系统状态的限制，而该状态又依赖于前一时隙的变化。因此，优化问题的约束条件在每个时隙中可能发生变化，从而影响全局最优解的求解。由此可见，该问题属于非凸优化问题。

特别地，当数据流量负载非常大，且数据缓存大小$C_{\max}$满足式(5)时，可以假设待发送的数据量近乎无限。

$$C_{\max} \geqslant \tau \operatorname{lb}\left(1+\frac{|h_{\max}|^2 \dfrac{E_{\max}}{\tau}}{\sigma^2}\right) \tag{5}$$

这时的子问题可定义为

$$\max \sum_{i=1}^{\infty} \tau \operatorname{lb}\left(1+\frac{|h_i|^2 p_i}{\sigma^2}\right) \tag{6}$$

$$\text{s.t.} \sum_{i=1}^{k} \tau p_i \leqslant \sum_{i=0}^{k-1} E_i, \forall k \in [1, \infty) \tag{6a}$$

$$\sum_{i=0}^{k-1} E_i - \sum_{i=1}^{k} L_i p_i \leqslant E_{\max}, \forall k \in [1, \infty) \tag{6b}$$

式(6)问题可视为式(4)问题在数据负载无限这一特定情境下的子问题。在未来的能量到达和信道增益可知的离线场景下，该问题呈现为凸优化问题，从而可以求得最优解。鉴于原问题离线场景求解最优解的困难性，本文设计了一组附加实验：将所提算法与这一特殊情形下的最优解进行对比分析，旨在验证算法的有效性。其中，式(6a)明确了每一步能耗不得超出当前剩余能量，这与先前问题的约束条件有所差异。在数据负载无限的前提下，最优策略不会导致能量过剩而发生溢出。式(6b)确保了能量无溢出，其中$E_0$代表电池初始电量。式(5)则表明，在任何时刻利用全部可用能量所能发送的最大数据量均不会超过缓存中的数据量。仅当式(5)成立时，才可将此问题视为数据流量无限制的情形。

本文研究的核心问题是有限数据负载下的在线传输功率控制问题，其中能量、数据到达和信道增益均具有随机性。算法需要根据已知的能量、数据到达和信道增益的相关信息合理地权衡能量分配。在此背景下，数据的发送速率与传输功率之间具有非线性关系，这意味着当前节省的能量在未来有机会发送更多的数据。此外，考虑信道增益的随机变化以及数据缓存和能量缓存的容量限制，该问题的复杂性进一步增加，从而使传统算法的设计和求解变得更加困难。

为应对上述复杂情况，本文采用强化学习方法，旨在使传感器能够通过持续学习，以最大化系统的长期平均吞吐量。尽管已有一些研究[11, 23, 27]采用强化学习来解决类似问题，但这些研究未充分考虑数据的随机性和数据缓存的有限性。除了基于能量水平决定是否传输数据，还需考虑数据与能量之间的相互依赖关系。在数据随机到达的情况下，即使能量水平较低，仍有可能选择传输数据，这不仅有助于提升数据吞吐量，还有助于增强数据的时效性。例如，在某些场景中，若在能量较低时选择不发送数据，可能会导致数据溢出；然而，若在随后的时段数据到达量减少且能量供应增加时，未及时传输的数据可能会因能量缓存溢出而被浪费。本文旨在优化传输决策策略，以避免这种资源浪费，从而提高整体数据传输效率。

## 4　强化学习解决方案

在本节中，将端到端EH通信场景建模为具有

连续状态空间和连续动作空间的马尔可夫决策过程（MDP, Markov decision process）。根据这一MDP模型，提出了一种基于SAC算法的能量分配方法，该方法根据发送节点的剩余电量、信道增益、随机到达能量、剩余数据量、随机到达数据量等状态变量，动态地选择最优发送功率，从而实现吞吐量最大化的目标。

### 4.1 MDP建模

在强化学习中，MDP通常用于描述环境信息，智能体通过与环境的持续交互实现自我学习和决策优化。在本文的研究场景中，发送节点被视为智能体。MDP由五元组（$\boldsymbol{S}, \boldsymbol{A}, \boldsymbol{P}, r, \gamma$）组成。

状态（$\boldsymbol{S}$）描述当前系统的状态，包括传感器的剩余电量、信道增益、随机到达的能量、剩余数据量、随机到达的数据量等。

动作（$\boldsymbol{A}$）表示在当前状态下智能体可以采取的动作。在本问题中，动作定义为每个时隙的发送功率$p_i$。在一个时隙内，发送功率保持恒定，因为发送功率与数据量之间存在非线性关系，保持恒定可以最大化能量利用效率。发送功率的大小受缓存剩余数据量和电池剩余能量的限制，超出限制时需要进行剪裁处理。

状态转移概率（$\boldsymbol{P}$）描述在某一状态下执行某个动作后，系统转移到下一个状态的概率分布。由于动作空间和状态空间是连续的，且采用无模型强化学习方法，因此不需要明确给出状态转移概率。

奖励（$r$）反映智能体在某个状态下执行某一动作后获得的即时回报。奖励包括数据吞吐量作为正向奖励，同时在智能体动作超出限制时进行剪裁，并给予相应的惩罚。与直接终止任务相比，这种方式有效地提高了智能体的学习效率。此外，本文还基于缓存数据的水平，对能量溢出情况进行惩罚，以促进智能体快速学习并优化决策策略。

折扣因子（$\gamma$）用于调整智能体在决策过程中对长期回报的重视程度。

### 4.2 动作剪裁方法

本文模型中，智能体的动作是选择数据传输功率$p_i$。当选择的动作超出限制时，通过动作剪裁机制进行调整，并对超限动作给予适当的惩罚，同时确保智能体能够继续执行任务，而非中断当前回合。智能体的动作空间受到缓存数据量$C_i$和电池中剩余能量$B_i$的限制。具体来说，发送功率$p_i \in [0, P_{\max,i}]$，其中，$P_{\max,i}$是第$i$个时隙开始时可选择的最大发送功率。$P_{\max,i}=\max\{P_{\max,i,\text{energy}}, P_{\max,i,\text{data}}\}$，其中，$P_{\max,i,\text{energy}}$和$P_{\max,i,\text{data}}$分别表示在当前能量限制和数据限制下可选择的最大功率，$P_{\max,i,\text{energy}}$和$P_{\max,i,\text{data}}$分别满足

$$\tau P_{\max,i,\text{energy}} \leqslant B_i \tag{7}$$

$$R(P_{\max,i,\text{data}}, h_i) \leqslant C_i \tag{8}$$

式(7)表明在一个时隙内，选择的发送功率所消耗的能量不超过电池剩余的能量。式(8)表明在一个时隙内，发送的数据量不超过缓存中当前剩余的数据量。对于不满足这些约束条件的动作，本文采用动作剪裁，令$p_{i,\text{clip}}=\max(\min(p_i, P_{\max,j}), 0)$，将$p_i$截断到合法范围内，从而确保智能体的行为符合能量和数据的双重约束，其中，$p_{i,\text{clip}}$为截断后的发送功率。

### 4.3 动作空间映射

动作归一化在强化学习算法中扮演着重要的角色，尤其是在处理连续动作空间的环境中。动作空间如果不归一化，可能严重影响智能体训练的稳定性，甚至导致训练无法收敛。动作归一化有助于确保算法在面对复杂且连续的动作空间时能够保持稳定，从而提高策略探索的效率，进而保证算法在实际应用中的稳健性与可靠性。在本文中，将动作空间映射到[−1,1]的范围内。具体的映射方法如式(9)所示。

$$p_{\text{norm},i} = \frac{p_{\text{original},i}}{E_{\max}/\tau} \cdot 2 - 1 \tag{9}$$

其中，$p_{\text{norm},i}$和$p_{\text{original},i}$分别代表归一化后的发送功率和归一化前的发送功率，$E_{\max}/\tau$为发送功率上限，且发送功率下限为0。

当环境接收到智能体的动作后，首先对该动作进行反归一化处理，然后执行相应的时间步，环境内的其他部分无须做任何修改。对于自定义环境，实际执行的动作值仍然位于原始动作空间的范围内。而对于智能体，经过归一化的动作空间有助于提高训练过程的稳定性和探索能力，从而加速智能体学习最优策略的过程。

### 4.4 奖励函数的详细设计

在本文中，奖励函数由4个部分组成，分别为数据传输量奖励$r_1$、动作能量超限惩罚$r_2$、动作数据超限惩罚$r_3$、能量溢出惩罚$r_4$，最后的单步即时奖励$r=\lambda_1 r_1+\lambda_2 r_2+\lambda_3 r_3+\lambda_4 r_4$，其中，$\lambda_1$、$\lambda_2$、$\lambda_3$

和$\lambda_4$为权重因子，用于平衡各项奖励的影响。由于每项奖励的性质和作用不同，需要进行合理的权衡。以下将分别阐述每个奖励项的设计思路。

数据传输量奖励$r_1$：为了最大化长期平均数据吞吐量，本文将单步数据传输量作为即时奖励的一部分。数据传输量的增加直接与任务的目标相关，奖励智能体能够鼓励其在执行任务时提高数据传输效率。

$$r_1 = \tau \mathrm{lb}\left(1 + \frac{p_{i,\mathrm{clip}}|h_i|^2}{\sigma^2}\right) \tag{10}$$

动作能量超限惩罚$r_2$：当智能体选择的发送功率所消耗的能量超出了自身所能提供的能量时，动作违反了能量约束。不同于传统方法通过终止训练并施加较大负奖励来处理此类超限问题的策略，本文采取了动作剪裁的策略，将超限的动作裁剪到合法范围内，并根据超限幅度给予相应的惩罚。本文选择不直接终止训练并给予较大惩罚的方式，原因是任务属于精细控制任务，智能体的动作受到电池能量和缓存数据量双重约束。由于剩余能量和数据量是动态变化的，且在某些时段内能量供应较为稳定，智能体的最优动作通常会接近这些边界。尤其是在早期探索阶段，智能体容易选择超出约束的动作。若此时频繁终止训练并给予较高惩罚，智能体的选择将趋于保守，从而限制了潜力。为了实现长期平均的最大数据吞吐量，本文对能量超限惩罚进行了适当的控制，以保证智能体能够顺利完成训练。

$$r_2 = \begin{cases} \tau \cdot \min\left\{p_i - P_{\max,i,\mathrm{energy}}, E_{\mathrm{upper}}\right\}, p_i > P_{\max,i,\mathrm{energy}} \\ 0, \qquad\qquad\qquad\qquad 其他 \end{cases} \tag{11}$$

为了避免惩罚值过大，本文设定了惩罚的上限$E_{\mathrm{upper}}$。由于数据传输量奖励$r_1$的计量单位是数据量，而能量超限惩罚$r_2$的计量单位是能量，能量消耗随发送数据量的变化呈现指数变化。这里没有将能量溢出直接取对数使它们在同一量纲，因为本文的问题主要是对能量进行管理，需要对能量消耗进行精细控制，而对数变换会使智能体对能量变化的敏感度降低，无法提供有效的反馈。因此，惩罚的上限$E_{\mathrm{upper}}$用于控制宏观惩罚上限，确保训练过程的稳定性。

动作数据超限惩罚$r_3$：当智能体选择的数据发送功率所能传输的数据量超出了缓存中剩余的数据量时，动作违反了数据约束。本文采用与动作能量超限惩罚相似的处理方法。由于发送功率与消耗的能量之间存在线性关系，而发送功率与传输数据量之间呈指数关系，并受到信道增益$h_i$的影响，因此，需要单独对数据超限进行处理，并通过奖励函数明确告知智能体这种复杂关系。为使数据超限惩罚$r_3$与能量超限惩罚$r_2$在同一量纲下，采用对数变换来平衡两者的影响。

$$r_3 = \begin{cases} \min\left\{2^{\left((\tau \cdot R(p_i,h_i)) - C_i\right)} - 1, E_{\mathrm{upper}}\right\}, p_i > P_{\max,i,\mathrm{data}} \\ 0, \qquad\qquad\qquad\qquad 其他 \end{cases} \tag{12}$$

能量溢出惩罚$r_4$：当能量溢出时，意味着能量未被充分利用，这对于智能体来说是应当避免的情况，因此本文对能量溢出给予惩罚。然而，在某些情况下，当没有足够的数据可以发送时，能量的溢出是无法避免的。为了解决这一问题，本文引入了一个与当前缓存数据量$C_i$和信道增益$h_i$相关的因子$L(C_i, h_i)$，用以调整惩罚的大小。具体来说，溢出的能量与$C_i$和$h_i$成正比，因此溢出的惩罚也随着这两个因子的变化而变化。

$$r_4 = \begin{cases} L(C_i,h_i) \cdot \min\left\{B_i - \tau p_i + E_i - E_{\max}, E_{\mathrm{upper}}\right\}, \\ \qquad\qquad\qquad\qquad B_i - \tau p_i + E_i > E_{\max} \\ 0, \qquad\qquad\qquad\qquad 其他 \end{cases} \tag{13}$$

其中，$B_i - \tau p_i + E_i - E_{\max}$表示溢出的能量，$E_{\mathrm{upper}}$限制惩罚上限，保证训练稳定。

### 4.5　对于结束状态的处理

在强化学习任务中，存在终止状态（terminated）和截断状态（truncated）两种结束状态。终止状态表示任务已达到最终状态，后续没有进一步的动作或状态；而截断状态则表示任务因人为干预或达到预设的最大步数而中止，智能体实际上可以继续执行。两种结束状态在计算动作价值函数和状态价值函数的时序差分时，处理方式有所不同。

本文的任务目标是最大化长期回报，任务长期运行，不存在最终状态。但由于训练中的一个回合长度是有限的，本文设置一个单回合最大步数，因此，实际的结束状态都是截断状态。尽管任务在该步骤已中止，智能体实际上仍然可以继续执行，因此，在计算时序差分时，需将可能的下一个状态纳入考虑，以确保回报的计算更准确，进而优化智能体的决策。

### 4.6　基于SAC的能源管理算法

不同于传统强化学习算法以最大化智能体回报

为目标，SAC 算法在 DDPG 的框架下，引入最大熵的思想，同时最大化智能体的回报和动作的熵值。通过最大化信息熵，SAC 能够有效地平衡探索与利用。与其他强化学习算法（如 PPO 和 DDPG）相比，SAC 采用的最大熵策略增强了智能体在连续动作空间中的探索能力，同时在面对环境干扰时展现了更强的稳定性，能够更快速地进行策略调整。SAC 算法的网络架构如图 3 所示，由 1 个 Actor 网络、2 个 Critic 网络和 2 个 Critic 目标网络构成。Actor 网络接收输入状态、输出动作和对应的熵；Critic 网络对当前的状态—动作对的 Q 值进行评估；Critic 目标网络则通过滞后更新机制从主 Q 网络逐步获取更新，以提供更加稳定的 Q 值目标。

#### 4.6.1　最大熵强化学习

如果只有一个随机变量 $X$，且它的概率密度函数为 $p$，那么它的熵被定义为

$$H(X) = \mathbb{E}_{x\sim p}[-\lg p(x)] \tag{14}$$

使用 $H(\pi(\cdot|s))$ 表示策略 $\pi$ 在状态 $s$ 下的随机程度。最大熵强化学习的思想就是在最大化累积奖励的同时使策略更加随机。于是在强化学习的目标中加入了一项熵的正则项，从而最优策略可以定义为

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\pi}\left[\sum_t r(s_t,a_t) + \alpha H(\pi(\cdot|s_t))\right] \tag{15}$$

其中，$r$ 表示在状态 $s_t$ 下采取动作 $a_t$ 的即时奖励，$\pi^*$ 表示最优策略，$\pi(\cdot|s_t)$ 表示在状态 $s_t$ 下，根据策略 $\pi$ 选择动作的概率分布，$\alpha$ 是一个正则化的系数，用来控制熵的重要程度，并权衡探索与利用。

$\alpha$ 越大，智能体的探索能力就越强，这有助于找到更优的策略，并减少策略陷入较差的局部最优的可能性。但 $\alpha$ 过大会导致过度探索，使学习速度变慢，训练不稳定，在最优解周围徘徊。

#### 4.6.2　Critic 网络

Critic 网络（Q 网络）用于评估给定状态动作对的价值，SAC 使用两个 Q 网络（$Q_{\omega 1}$ 和 $Q_{\omega 2}$），以减小过估计，提高稳定性。在每次训练中，选择较小的 Q 值网络来进行更新，从而缓解 Q 值的过高估

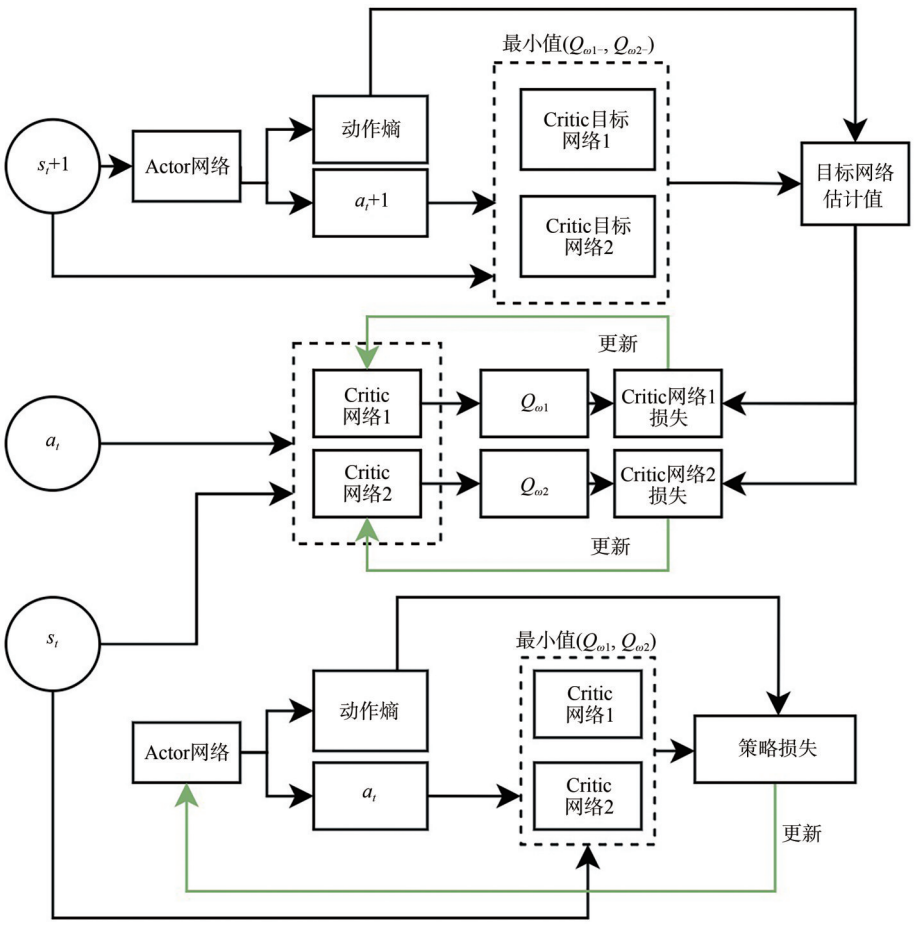


图 3　SAC 算法的网络架构

计。同时，SAC也使用目标网络机制，用滞后更新的方式将目标Q网络$Q_{\omega^-}$作为梯度更新的时序差分的目标，使算法训练过程更稳定。

状态价值函数$V(s_t)$定义为在状态$s_t$下，智能体遵循策略$\pi$时能够获得的期望累积奖励，表示为

$$V(s_t)=\mathbb{E}_{a_t\sim\pi}[Q(s_t,a_t)]+H(\pi(\cdot|s_t)) \tag{16}$$

其中，$Q(s_t,a_t)$表示Critic网络对于输入状态$(s_t,a_t)$计算出的$Q$值。

Q网络的损失函数为

$$L_Q(\omega)=\mathbb{E}_R\frac{1}{2}\left(Q_\omega(s_t,a_t)-(r_t+\gamma V_{\omega^-}(s_{t+1}))\right)=$$
$$\mathbb{E}_{\pi_\theta(\cdot|s_{t+1})}\left[\frac{1}{2}\left(Q_\omega(s_t,a_t)-(r_t+\gamma(\min_{j=1,2}Q_{\omega_j^-}(s_{t+1},a_{t+1})-\right.\right. \tag{17}$$
$$\left.\left.\alpha\lg\pi(a_{t+1}|s_{t+1})))\right)^2\right]$$

其中，$R$代表经验回放池，$\omega$代表网络的参数，$Q_\omega$代表任意一个Q网络。

目标网络的更新采取的是滞后更新的方式，使目标网络$Q_{\omega^-}$缓慢更新，逐渐接近网络$Q_\omega$，如式(18)所示。

$$\omega^-\leftarrow\tau\omega+(1-\tau)\omega^- \tag{18}$$

**4.6.3**　Actor网络

在连续动作空间中，Actor网络输出动作的高斯分布，由于从高斯分布中采样是不可导的，因此采用重参数化技巧进行处理。首先，从标准正态分布中采样噪声$\epsilon_t$~N(0,1)，然后，通过将噪声与标准差$\sigma$相乘并加上均值$\mu$，生成最终的动作$a=\mu(s;\theta)+\sigma\cdot\epsilon_t$，其中，$\mu(s;\theta)$和$\sigma(s;\theta)$分别是Actor网络根据当前状态$s_t$输出的均值和标准差。通过这种方式，使采样过程可导，并能通过反向传播进行优化。策略的损失函数为

$$L_\pi(\theta)=\mathbb{E}_{s_t\sim R,\epsilon_t\sim\mathrm{N}}\left[\alpha\lg(\pi_\theta(a_t|s_t))-\min_{j=1,2}Q_{\omega_j}(s_t,a_t)\right] \tag{19}$$

其中，$s_t$~$R$表示状态$s_t$是从经验回放池$R$中采样的。$\pi_\theta$中$\theta$表示策略网络的参数，$\pi_\theta(a_t|s_t)$表示策略网络$\pi$在$s_t$状态下输出$a_t$函数的概率。

**4.6.4**　熵自适应调整

SAC算法基于最大熵强化学习，因此选择熵的系数$\alpha$非常重要。为了自动调整熵正则项，SAC将强化学习的目标转化为一个带约束的优化问题

$$\max_\pi\mathbb{E}_\pi\left[\sum_t r(s_t,a_t)\right]$$
$$\text{s.t.}\quad\mathbb{E}_{(s_t,a_t)\sim\rho_\pi}[-\lg(\pi_t(a_t|s_t))]\geqslant\mathcal{H}_0 \tag{20}$$

式(20)的目标是最大化期望回报，同时约束熵的均值大于阈值$\mathcal{H}_0$，表示在策略$\pi$下得到的累计回报的均值。$\mathbb{E}_{(s_t,a_t)\sim\rho_\pi}$中$(s_t,a_t)$~$\rho_\pi$表示状态动作对取自$\rho_\pi$。

为了处理这个约束，引入一个拉格朗日乘子，并构造拉格朗日函数为

$$L(\pi,\alpha)=\mathbb{E}_\pi\left[\sum_t r(s_t,a_t)\right]+$$
$$\alpha\left(\mathbb{E}_{(s_t,a_t)\sim\rho_\pi}[-\lg(\pi(a_t|s_t))]-\mathcal{H}_0\right) \tag{21}$$

化简后得到$\alpha$的损失函数为

$$L(\alpha)=\mathbb{E}_{s_t\sim R,a_t\sim\pi(s_t)}[-\alpha\lg\pi(a_t|s_t)-\alpha\mathcal{H}_0] \tag{22}$$

**算法1**　基于SAC的传输功率控制算法

**初始化**　用随机的网络参数$\omega_1$，$\omega_2$和$\theta$分别初始化Critic网络$Q_{\omega_1}(s,a)$，$Q_{\omega_2}(s,a)$和Actor网络$\pi_\theta(s)$，复制与Critic网络相同的参数，分别初始化目标网络和$Q_{\omega_2^-}$；初始化经验回放池大小，采样批大小$N$，折扣因子$\gamma$；初始化发送节点$S$电池电量$B_0$，缓存中数据量$C_0$，信道增益$h_0$，收集能量$E_0$

**for** 迭代次数$i$=1→$I$ **do**
　初始化状态$s_0$；
　**for** 时间步$t$=1→$T$ **do**
　　将$s_t$输入策略网络，得到动作$a_t$（功率）；
　　发送节点$S$执行动作$a_t$，获得奖励$r_t$，环境状态变为$s_{t+1}$；
　　将（$s_t$, $a_t$, $r_t$, $s_{t+1}$）元组存入经验回放池；
　　**if** 经验回放池中数据数量 > $N$ **then**
　　**for** 训练轮数$k$=1→$K$ **do**
　　从经验回放池中采样$N$个元组；
　　对于每个元组，用目标网络计算其更新目标值$y_i$，然后根据式(17)计算损失函数并更新两个Critic网络；
　　重参数化，采样$\tilde{a}_t$和动作熵，根据式(19)中的损失函数更新当前Actor网络；
　　根据式(22)中的损失函数更新熵的系数$\alpha$以权衡智能体的探索与奖励提升；
　　根据式(18)软更新Critic目标网络参数，使目标网络的参数缓慢追上Ctiric网络，保证训练稳定性；
　　**end for**
　　**end if**
　**end for**
**end for**

算法 1 的大致执行流程可以描述如下：首先，初始化 Actor 网络和 Critic 网络的参数，并将 Critic 网络的参数复制至目标网络。然后，初始化经验回放池、批采样大小、折扣因子等环境相关参数。在每一轮迭代中，发送节点将当前的环境状态（电池电量、缓存数据量、信道状态等）作为输入，Actor 网络输出一个动作，即数据发送的功率大小。传感器节点执行该动作，通过发送数据来获得奖励，并更新状态（电池电量、缓存数据量等）。此时，传感器节点将当前的经验（状态、动作、奖励和下一状态）存入经验回放池中。当经验回放池的样本量达到设定值时，从中随机采样一批经验，利用目标网络计算目标值，并通过 Critic 网络的损失函数来更新 Critic 网络的参数。接着，通过重参数化技巧和动作熵更新 Actor 网络的参数，以更好地选择合适的发送功率。此外，熵系数也会被更新，以确保策略在环境中能够有效地进行探索。最后，Critic 目标网络参数通过软更新的方式进行调整，使目标网络参数慢慢接近 Critic 网络，保证训练过程的稳定性。这一过程持续进行，直至算法收敛，从而实现传感器在不同环境条件下选择最佳的发送功率，以达到高效的数据传输和能量管理。

## 5　仿真实验

本节中，在仿真端到端通信场景中对所提算法进行了详细的评估。假设每个回合持续 512 s，时间间隔 $\tau$ 为 1 s，噪声功率 $\sigma^2$=1。经过参数网格化测试，确定了较优的参数设置。折扣因子 $\gamma$=0.995，学习率 $\beta$=0.000 4。针对有限数据情形，模拟环境的参数设置如下，所用电池的最大容量 $B_{\max}$ 为 100 J。数据缓存的容量 $C_{\max}$ 由式(2)计算得出，即 $C_{\max}=2R(B_{\max},1)$，以确保数据缓存上限和电池容量上限处于相同量级。每个时隙的到达能量 $E_i \in \mathcal{E}_n \triangleq [0, E_{\max}]$ 服从正态分布 $f_{\mathcal{E}_n}$ 的连续随机变量，$f_{\mathcal{E}_n}$ 的标准差为 $E_{\max}/5$，均值等于 $E_{i-1}$，其中，$E_{\max}=B_{\max}/4$。缓存数据增量 $D_i \in \mathcal{D}_n \triangleq [0, D_{\max}]$ 是服从正态分布 $f_{\mathcal{D}_n}$ 的连续随机变量，$f_{\mathcal{D}_n}$ 的标准差为 $D_{\max}/5$，均值等于 $D_{i-1}$，其中，$D_{\max}$ 由式(2)获得，即 $D_{\max}=R(E_{\max}, 1)$。$E_i$ 和 $D_i$ 的生成过程可以视为一个随机游走过程。这种设置的好处在于，它使数据缓存和能量缓存的容量量级相近，从而能够真实地模拟能量溢出和数据溢出的场景。信道增益的变化过程是一个正态分布 $h_i \in \mathcal{H}_n \triangleq [0.1, 1.0]$，$\mathcal{H}_n$ 是一个连续随机变量，服从正态分布 $f_{\mathcal{H}_n}$，标准差为 0.25，均值等于 0.5。

具体而言，本文考虑了数据负载有限和数据负载无限两种情况。

### 5.1　数据负载无限场景

首先，本文关注数据量无限的场景，即式(6)对应的子问题。此类场景在需要持续发送数据的环境中具有实际意义。该问题的离线场景，即能量到达和信道状态信息在初始时刻已完全已知，可以转化为一个凸优化问题。本文利用分裂圆锥求解器（SCS, splitting conic solver）求解此离线场景的数值最优解，并将其与在线算法的执行结果进行对比。在实验中，随机生成并固定 100 个环境种子（这些种子不参与智能体的训练过程）。在训练过程中，每次迭代智能体经历 5 个回合（即 512×5 步），并不断更新策略。每次迭代后，利用这 500 个种子生成 500 个测试回合。通过比较当前智能体策略与最优解之间的差异，并绘制迭代次数与策略差异的曲线，可以评估算法的效果。

随机生成并固定 100 个环境种子（这些种子不会参与智能体的训练）。在训练智能体的过程中，每次迭代（智能体经历 5 个回合，即 512×5 步，并不断更新），本文利用这 500 个种子生成 500 个测试回合。通过比较当前智能体策略与最优解之间的差异，并根据迭代次数绘制曲线，能够评估算法的效果。此外，本文还与贪心策略、PPO 和 TD3 算法进行了对比。模拟该场景的另一个目的是通过离线最优解验证算法的可行性和有效性。由于在数据负载有限的情况下，问题的离线场景为非凸优化问题，难以获得离线最优解，而在数据负载无限的情况下，该问题是凸优化问题，可以求得数值最优解。

平均累计吞吐量随迭代次数的变化（数据负载无限）如图 4 所示，平均累计奖励随迭代次数的变化（数据负载无限）如图 5 所示。结果表明，SAC 和 TD3 算法的最终效果相近，并能够接近最优解，而 PPO 算法的性能逊色于前两者。在算法稳定性方面，SAC 算法表现较强的稳定性，并且收敛速度最快。在训练初期，所有算法都出现了较明显的震荡，这是本文在环境中对违法动作和能量溢出进行了惩罚，导致智能体在早期为了最大化吞吐量而可能选择违法动作。例如，发送功率超出剩余能量，或者数据不足但能量充足时选择超过剩余数据量的

发送功率。这些违法动作会引起较大的惩罚，因此训练曲线在初期呈现较大震荡。

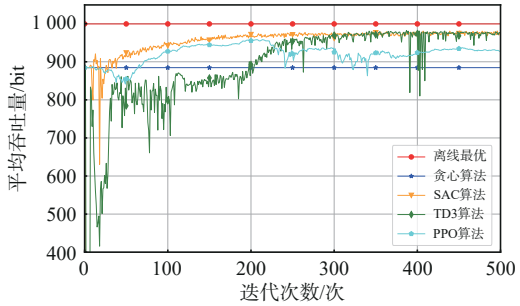


图4　平均累计吞吐量随迭代次数的变化（数据负载无限）

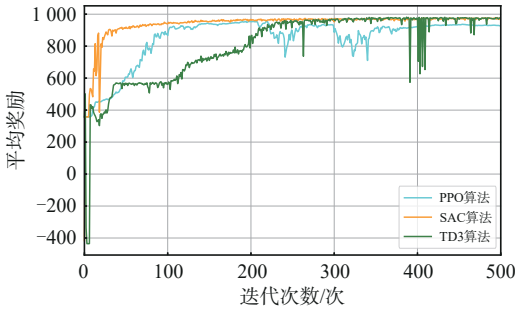


图5　平均累计奖励随迭代次数的变化（数据负载无限）

## 5.2　数据负载有限场景

接着，本文考虑更为真实的应用场景，即式(4)对应的问题，数据负载有限的场景。在此场景中，仍然使用贪心策略、TD3和PPO作为对比，验证所提算法在有限数据负载情况下的有效性。考虑数据的随机到来和可能发生的数据缓存溢出，环境变得更复杂，为此，本文设计了更复杂的奖励函数和动作剪裁方法，如第4.2节与第4.4节所述。实验设计的思路与数据负载无限场景类似，平均累计吞吐量随迭代次数的变化如图6所示，平均累计奖励随迭代次数的变化如图7所示。结果表明，在该复杂环境下，PPO算法未能有效地训练出优于贪婪算法的策略，而SAC算法和TD3算法可以有效地训练。这表明SAC和TD3算法在复杂环境中的连续动作和状态空间控制任务中表现更出色，且对于该任务SAC算法的性能和稳定性优于TD3。

## 5.3　排除信道增益测试

信道增益的变化对数据吞吐量有显著影响，在信道增益较高时，相同能量消耗能够实现更大的数据吞吐量。为了验证所提算法可以学习到信道增益的影响，本文设计了另一组实验。该组实验保持信道增益不变，排除了信道增益这一变量对算法表现

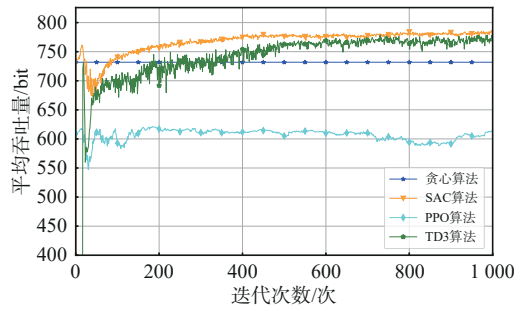


图6　平均累计吞吐量随迭代次数的变化

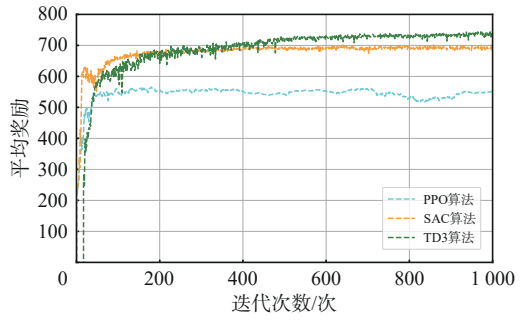


图7　平均累计奖励随迭代次数的变化

的影响。平均累计吞吐量随迭代次数的变化（信道增益为常数，数据负载无限）如图8所示，平均累计吞吐量随迭代次数的变化（信道增益为常数）如图9所示，结果表明，所提算法不仅能够有效地学习信道增益的变化，还能够理解并应对数据和能量之间

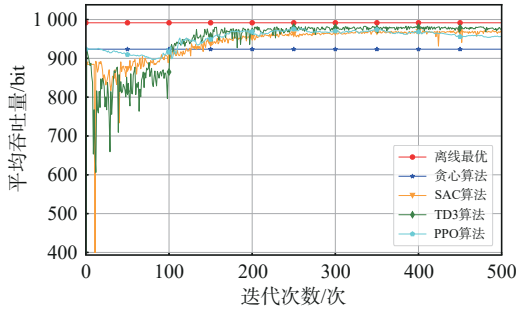


图8　平均累计吞吐量随迭代次数的变化（信道增益为常数，数据负载无限）

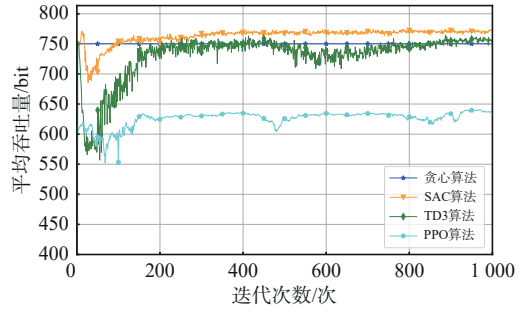


图9　平均累计吞吐量随迭代次数的变化（信道增益为常数）

的复杂关系。这进一步验证了所提算法在面对复杂的环境条件时，具备较强的适应能力和鲁棒性。

## 6 结束语

本文针对无线可充电网络中的端到端通信问题，在能量和数据随机到达的条件下，提出了一种基于深度强化学习的在线控制决策方法，旨在最大化系统的长期数据吞吐量。该方法不需要预先了解能量和数据到达的具体分布，而是基于SAC算法实现在线控制。本文与其他强化学习算法及贪婪算法进行了仿真实验对比。仿真结果表明，所提方法在性能上优于基线算法，证明了其有效性。此外，本文研究了在某些特定场景中问题的转化。在数据负载无限的情况下，问题可通过凸优化方法求解，获得其离线最优解。然而，本文仅讨论了单个传感器与基站进行直接通信的情形，未考虑并发通信中可能存在的干扰问题。未来的工作中，可以进一步探讨在多对一或多对多通信场景中，如何有效地避免干扰，以提高算法在更复杂通信环境中的适用性和鲁棒性。

## 参考文献：


[1] CLERCKX B, ZHANG R, SCHOBER R, et al. Fundamentals of wireless information and power transfer: from RF energy harvester models to signal and system designs[J]. IEEE Journal on Selected Areas in Communications, 2019, 37(1): 4-33.

[2] WEI Z Q, YU X H, NG D W K, et al. Resource allocation for simultaneous wireless information and power transfer systems: a tutorial overview[J]. Proceedings of the IEEE, 2022, 110(1): 127-149.

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

[4] CHEN Y, ZHAO F J, LU Y G, et al. Dynamic task offloading for mobile edge computing with hybrid energy supply[J]. Tsinghua Science and Technology, 2023, 28(3): 421-432.

[5] XIAO H, QI N J, YIN Y J, et al. Investigation of self-powered IoT sensor nodes for harvesting hybrid indoor ambient light and heat energy[J]. Sensors, 2023, 23(8): 3796.

[6] KIM W G, KIM D, LEE H M, et al. Wearable fabric-based hybrid energy harvester from body motion and body heat[J]. Nano Energy, 2022, 100: 107485.

[7] BAKYTBEKOV A, NGUYEN T Q, ZHANG G, et al. Synergistic multi-source ambient RF and thermal energy harvester for green IoT applications[J]. Energy Reports, 2023, 9: 1875-1885.

[8] BAI S M, CUI J, ZHENG Y Q, et al. Electromagnetic-triboelectric energy harvester based on vibration-to-rotation conversion for human motion energy exploitation[J]. Applied Energy, 2023, 329: 120292.

[9] MASADEH A, WANG Z D, KAMAL A E. Reinforcement learning exploration algorithms for energy harvesting communications systems[C]//Proceedings of the 2018 IEEE International Conference on Communications (ICC). Piscataway: IEEE Press, 2018: 1-6.

[10] BLASCO P, GUNDUZ D, DOHLER M. A learning theoretic approach to energy harvesting communication system optimization[J]. IEEE Transactions on Wireless Communications, 2013, 12(4): 1872-1882.

[11] QIU C R, HU Y, CHEN Y, et al. Deep deterministic policy gradient (DDPG)-based energy harvesting wireless communications[J]. IEEE Internet of Things Journal, 2019, 6(5): 8577-8588.

[12] DHILLON S, MADHU C, KAUR D, et al. A solar energy forecast model using neural networks: application for prediction of power for wireless sensor networks in precision agriculture[J]. Wireless Personal Communications, 2020, 112(4): 2741-2760.

[13] WANG Z, AGGARWAL V, WANG X D. Power allocation for energy harvesting transmitter with causal information[J]. IEEE Transactions on Communications, 2014, 62(11): 4080-4093.

[14] LIANG H, ZHAO X H, ZHANG W. Optimal power allocations for multichannel energy harvesting cognitive radio[C]//Proceedings of the 2017 IEEE 18th International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM). Piscataway: IEEE Press, 2017: 1-6.

[15] HAKAMI V, DEHGHAN M. Distributed power control for delay optimization in energy harvesting cooperative relay networks[J]. IEEE Transactions on Vehicular Technology, 2017, 66(6): 4742-4755.

[16] JIA R H, ZHANG J B, LIU X Y, et al. Optimal rate control for energy-harvesting systems with random data and energy arrivals[J]. ACM Transactions on Sensor Networks, 2019, 15(1): 1-30.

[17] DIAS G M, NURCHIS M, BELLALTA B. Adapting sampling interval of sensor networks using on-line reinforcement learning[C]//Proceedings of the 2016 IEEE 3rd World Forum on Internet of Things (WF-IoT). Piscataway: IEEE Press, 2016: 460-465.

[18] FRATERNALI F, BALAJI B, AGARWAL Y, et al. Pible: battery-free mote for perpetual indoor BLE applications[C]//Proceedings of the 5th Conference on Systems for Built Environments. New York: ACM Press, 2018: 184-185.

[19] KANSAL A, HSU J, ZAHEDI S, et al. Power management in energy harvesting sensor networks[J]. ACM Transactions on Embedded Computing Systems, 2007, 6(4): 32-es.

[20] SHRESTHAMALI S, KONDO M, NAKAMURA H. Adaptive power management in solar energy harvesting sensor node using reinforcement learning[J]. ACM Transactions on Embedded Computing Systems, 2017, 16(5s): 1-21.

[21] SAWAGUCHI S, CHRISTMANN J F, MOLNOS A, et al. Multi-

agent actor-critic method for joint duty-cycle and transmission power control[C]//Proceedings of the 2020 Design, Automation & Test in Europe Conference & Exhibition (DATE). Piscataway: IEEE Press, 2020: 1015-1018.

[22] FRATERNALI F, BALAJI B, AGARWAL Y, et al. ACES: automatic configuration of energy harvesting sensors with reinforcement learning[J]. ACM Transactions on Sensor Networks, 2020, 16(4): 1-31.

[23] ORTIZ A, AL-SHATRI H, LI X, et al. Reinforcement learning for energy harvesting point-to-point communications[C]//Proceedings of the 2016 IEEE International Conference on Communications (ICC). Piscataway: IEEE Press, 2016: 1-6.

[24] AIT AOUDIA F, GAUTIER M, BERDER O. RLMan: an energy manager based on reinforcement learning for energy harvesting wireless sensor networks[J]. IEEE Transactions on Green Communications and Networking, 2018, 2(2): 408-417.

[25] LI M Y, ZHAO X H, LIANG H, et al. Deep reinforcement learning optimal transmission policy for communication systems with energy harvesting and adaptive MQAM[J]. IEEE Transactions on Vehicular Technology, 2019, 68(6): 5782-5793.

[26] CHU M, LIAO X W, LI H, et al. Power control in energy harvesting multiple access system with reinforcement learning[J]. IEEE Internet of Things Journal, 2019, 6(5): 9175-9186.

[27] QIAN L P, FENG A Q, FENG X, et al. Deep RL-based time scheduling and power allocation in EH relay communication networks[C]//Proceedings of the ICC 2019-2019 IEEE International Conference on Communications (ICC). Piscataway: IEEE Press, 2019: 1-7.

[28] ZHANG H D, ZHAN D, ZHANG C J, et al. Deep reinforcement learning-based access control for buffer-aided relaying systems with energy harvesting[J]. IEEE Access, 2020, 8: 145006-145017.

[29] SHARMA M K, ZAPPONE A, DEBBAH M, et al. Multi-agent deep reinforcement learning based power control for large energy harvesting networks[C]//Proceedings of the 2019 International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOPT). Piscataway: IEEE Press, 2019: 1-7.

[30] MUY S, RON D, LEE J R. Energy efficiency optimization for SWIPT-based D2D-underlaid cellular networks using multiagent deep reinforcement learning[J]. IEEE Systems Journal, 2022, 16(2): 3130-3138.

[31] GREGORI M, GÓMEZ-VILARDEBÒ J. Online learning algorithms for wireless energy harvesting nodes[C]//Proceedings of the 2016 IEEE International Conference on Communications (ICC). Piscataway: IEEE Press, 2016: 1-6.

[作者简介]

**刘南君**（2000– ），男，浙江师范大学计算机科学与技术学院硕士生，主要研究方向为无线可充电网络、智能物联网等。

**贾日恒**（1989– ），男，博士，浙江师范大学计算机科学与技术学院副教授，主要研究方向为物联网、无线可充电传感器网络、无人机网络、强化学习等。

**许文韬**（1998– ），男，浙江师范大学计算机科学与技术学院硕士生，主要研究方向为无线可充电网络、智能物联网等。

**李明禄**（1965– ），男，博士，浙江师范大学计算机科学与技术学院教授，主要研究方向为物联网、无线传感器网络、并行计算等。