

采用Q学习的能量收集多中继协作 网络 中继选择策略

孙宝玺 李巍 翁康怡 余文政 梁中华

(长安大学信息工程学院, 710018, 西安)

摘要: 针对采用太阳能的能量收集 (energy harvesting, EH) 多中继协作通信网络中断性能较差的问题, 基于强化学习理论提出了一种联合中继选择和功率分配策略的优化方法。首先采用基于太阳辐照度实测数据的随机EH模型用于描述能量收集动态状况; 然后建立了EH多中继协作网络的马尔可夫决策过程模型, 定义动作空间、状态空间、收益函数、传输策略和价值函数; 最后使用Q学习算法设计了网络传输策略的学习和优化过程, 明确了单步学习过程中的动作选择、状态转移和Q值更新方法。该优化方法能够根据太阳能EH状态、无线信道衰落状态和所有EH中继的电池电量, 动态选择单个中继并确定其发射功率转发信号, 以优化网络信息传输中断概率。此外, 进行了计算机仿真实验, 分析了网络参数对中断概率性能的影响, 结果表明: 网络中断概率在极高信噪比下存在饱和特性; 本文优化方法相对于DQN算法, 在网络中断概率为 10^{-2} ~ 10^{-3} 时可获得3~5 dB的信噪比增益; 相对于Actor-Critic和SARSA算法在网络中断概率为 10^0 ~ 10^{-3} 时可分别获得2~3 dB和3~7 dB的信噪比增益。

关键词: 能量收集; 协作通信; 中继选择; 中断概率; Q学习

中图分类号: TN929.5 **文献标志码:** AA

文章编号: 0253-987X(XXXX)XX-0001-19

Q-Learning Based Relay Selection Policy for Multi-Relay Cooperative Networks with Energy Harvesting

SUN Baoxi, LI Wei, WENG Kangyi, YU Wenzheng, LIANG Zhonghua

(School of Information Engineering, Chang'an University, Xi'an 710018, China)

Abstract: To tackle the issue of poor outage performance in multi-relay cooperative communication networks with solar energy harvesting (EH), an optimization method for joint relay selection and power allocation policy is proposed based on reinforcement learning theory. First, a stochastic EH model based on measured solar irradiance data is adopted to mimic the dynamic status of solar EH. Then, the Markov decision process model of the multi-relay cooperative networks with solar EH is established, to define the action space, state space, reward function, transmission policy, and value function. Finally, the learning and optimization procedure of the transmission policy is designed by exploiting the Q-learning algorithm, which specifies the action selection, system state transfer and Q-value updating in a single-step of the learning procedure. This optimization method enables the dynamic selection of a

收稿日期: 2025-08-30。

single relay and its corresponding transmission power for signal forwarding based on the solar EH status, the wireless channel fading and the battery power of all EH relays, so as to optimize the outage probability of information transmission. Moreover, the impact of the network parameters on the outage performance is analyzed in computer simulation experiments and a saturation structure for the outage probability is revealed. In addition, simulation results demonstrate that, compared with the DQN algorithm, the proposed method can achieve 3~5 dB of signal-to-noise ratio (SNR) gains within the outage probability ranges of 10^{-2} ~ 10^{-3} . Compared with the Actor-Critic and SARSA algorithms, the proposed method can achieve 2~3 dB and 3~7 dB of SNR gains within the outage probability ranges of 10^0 ~ 10^{-3} .

Keywords: energy harvesting; cooperative communication; relay selection; outage probability; Q-learning

随着物联网技术的迅速发展,大规模的无线传感器网络得到广泛应用。无线传感器网络的终端设备通常都是能量受限的低功耗设备^[1]。如果使用固定电源或电池为该类设备供电,不利于网络灵活部署,同时也会增加网络维护成本^[2]。针对这个问题,一种

绿色通信技术——能量收集(energy harvesting, EH)通信技术越来越受到重视^[3]。在EH通信网络中,无线通信节点能够使用自然环境中的可再生能源为其提供电能用于信息传输,能够降低网络对固定电源和传统电池的依赖,降低网络维护成本,延长通信节点工作寿命,甚至实现终端能源自给^[1-2]。

由于EH通信节点能量受限,它们之间常采用协作通信技术以提高网络传输性能和覆盖范围^[4-5]。但是,从自然环境中收集到的能量具有较强的不确定性和随机性,而且EH协作网络结构相对复杂,传输方案灵活多变。因此在EH协作网络中EH能量分配和网络传输策略会直接影响网络传输性能,具有较高

的研究价值^[6]。

目前,EH协作通信网络传输策略优化算法主要分为两大类:离线算法与在线算法^[5]。前者基于确知的EH模型,EH通信节点能够提前准确获知整个通信过程的EH状态;后者则基于随机EH模型,EH通信节点无法提前获知EH状态,只能获得随机能量到达的统计信息。离线算法一般采用传统优化理论解决问题;在线算法则通常采用动态规划(dynamic programming, DP)、Lyapunov、强化学习(reinforcement learning, RL)等方法优化网络传输性能^[7]。

文献[8-10]以EH中继协作网络为研究对象,基于确知EH模型使用离线优化方法求解网络传输策略。文献[8]研究了EH双跳多中继网络在物联网应用场景中的能效最大化问题,借助拉格朗日对偶分解与朗伯W函数得到最优功率分配策略。文献[9]将网络中继选择和功率分配分解为两个子问题,利用投影梯度法求解网络最优传输策略。文

献[10]将网络吞吐量作为优化目标,利用分步优化法将功率分配优化问题转化为凸优化问题,通过迭代算法求解最优功率分配。然而,从自然界中收集的能量通常具有较强的随机性和不确定性,在实际情况下很难获得确知的EH模型,因此离线算法具有较大的局限性。

文献[11-13]针对EH多中继协作网络,基于随机EH模型建立了马尔可夫决策过程(Markov decision process, MDP)模型,并使用DP方法求解了网络最优传输策略。文献[11]将网络中断概率作为优化目标,采用值迭代算法求解最优中继选择,并分析了网络配置参数对中断性能的影响。文献[12]针对EH双向多中继网络设计了一种基于空时网络编码的多中继传输协议,利用值迭代算法求解EH中继功率最优分配问题。文献[13]将网络吞吐量作为优化目标,使用值迭代算法设计了一种联合中继选择和功率分配优化方案,并证明了其传输性能接近离线优化方案。然而,在使用DP方法求解EH协作网络最优传输策略时,需要准确的系统状态转移概率,但是在实际应用中很难获取到复杂通信网络的状态转移概率模型,因此DP方法在应用中存在一定局限性。

除了DP方法,Lyapunov方法也被广泛应用于EH中继协作网络传输策略的在线优化。文献[14]针对EH双向中继网络,利用Lyapunov方法将长期优化问题转换为基于漂移加惩罚的分布式调度问题进行求解,从而获得最优传输

策略。文献[15]基于EH协作物联网,利用Lyapunov优化方法提出了一种在线随机资源分配算法,并证明了该算法在保证系统稳定性的同时能够提高网络吞吐量。虽然Lyapunov方法能够解决具有高维状态空间的动态系统优化问题,但对于部分复杂系统难以构造合适的Lyapunov函数,导致算法无法收敛。

RL是一种智能体面向未知环境的自适应学习理论,可以解决智能体的决策优化问题。由于DP和Lyapunov方法的局限性,RL方法在近年来被用于解决EH中继协作网络传输策略的优化问题。文献[16]针对EH中继协作网络,采用合作SARSA算法求解了EH源节点和中继的最优发射功率。文献[17]以EH协作网络吞吐量为优化目标,分别采用DQN和Actor-Critic算法求解了EH中继最优功率分配策略。文献[18]以EH多中继协作网络为研究对象,将网络吞吐量作为优化目标,提出了一种熵近端策略优化(EPPO)算法,以求解最优中继选择策略,并验证了所提策略具有高吞吐量和低中断概率的特点。文献[19]针对EH多中继协作网络,提出了基于Actor-Critic算法的最优中继选择策略,能够提高网络传输可靠性,同时兼顾传输时延和能量效率。然而,上述研究中所采用的随机EH模型与自然界中EH实测数据的匹配度有待探讨。如果EH模型与实测数据不符,会影响EH协作网络传输策略理论研究的应用价值。

针对上述问题,本文采用了一种基于太阳辐照度实测数据的随机EH模型来描述太阳能EH的随机特性^[20],并使用RL理论中的Q学习算法研究了EH多中继协作网络的传输策略优化问题。在该网络中,配置有一个源节点、一个目的节点和多个EH中继节点。源节点和目的节点之间没有直达的无线链路,它们需要中继节点转发信号。所有EH中继没有固定电源供电,它们均装备有一个容量有限的电池用于收集太阳能为信号转发提供能量。为了节约EH中继能耗,网络在一个传输周期内只选择一个中继转发信号。网络能够根据太阳能EH状态、信道衰落状态和EH中继电池电量动态选择EH中继并确定其发射功率,以优化信息传输中断概率。为解决该随机优化问题,首先建立了网络的MDP模型,定义了动作空间、状态空间,求解了收益函数,并明确了传输策略和价值函数;接下来使用Q学习算法设计了网络传输策略的学习和优化过程,详细说明了单步学习过程中的动作选择、状态转移和Q值更新方法。最后,在计算机仿真实验中分析了网络配置参数对传输性能的影响,并指出网络中断概率存在饱和特性:中断概率在极高信噪比时趋向于极限值,即所有EH中继电池电量耗尽概率。同时,通过对比网络在不同优化方法和传输策略下的中断概率性能,证实本文策略明显优于对比策略,仿真结果表明:本文优化方法相对于DQN算法^[17-18],在网络中断概

率为 $10^{-2} \sim 10^{-3}$ 时可获得3~5 dB的信噪比增益;相对于Actor-Critic^[17,19]和SARSA^[16]算法在中断概率为 $10^0 \sim 10^{-3}$ 时可分别获得2~3 dB和3~7 dB的信噪比增益。

1 EH多中继协作通信网络

1.1 网络模型

EH多中继协作通信网络如图1所示,受物理障碍遮挡或发射功率限制,源节点S和目的节点D之间没有直达的无线链路,S在 N 个中继节点 $R_i(i=1, \dots, N)$ 的协助下向D传输信息。网络中所有无线节点只配备单根天线,并且工作在半双工模式下。源节点使用固定电源供电,每个中继节点均配备太阳能电池面板和容量有限的电池,能够收集太阳能用于转发源节点的信号。针对协作通信网络部署在野外环境的应用场景,当信息传输速率较低时,可以假设网络中无线信道服从准静态且平坦的瑞利衰落模型,信道衰落系数之间相互独立且服从均值为0、方差为 γ_0 的复高斯随机分布。

在EH多中继协作网络中,所有EH中继节点均采用“收集-储存-使用”的能量管理模型^[20]。假设网络的能量管理周期(即信息传输周期)为 T_M ,在当前周期内所有EH中继节点收集太阳能,并将收集到的能量以基本能量单位 E_0 进行量化,然后储存在各自的电池中用于后续周期的信息转发。由于EH

中继节点的电池容量有限,当电池达到最大容量时,新收集的能量将无法储存。

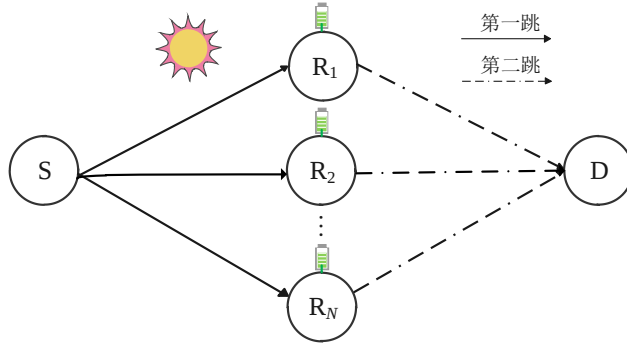


图 1 EH 多中继协作通信网络

Fig. 1 EH multi-relay cooperative communication networks

1.2 传输方案

在EH多中继协作网络中,为了避免多个EH中继节点同时转发源节点信息所带来的系统同步开销,降低无线传感器网络的系统复杂度,本文采用单中继传输方案,即在一个传输周期内根据网络状态仅选择一个EH中继节点转发源节点信息,其他EH中继节点则同时收集能量用于后续的信息转发。在每个传输周期内,当选择的单中继节点为R时, EH多中继协作网络就等效为传统的单中继协作网络。整个信息传输过程可以分为两个阶段:在第一阶段,源节点向中继R发送信息;在第二阶段, R采用放大转发协议将源节点信号转发至目的节点D。

在该传输过程中,中继R和目的节点D接收到的信号可以表示为:

$$\begin{cases} y_{S,R} = h_{S,R} \sqrt{P_S} x_S + n_R \\ y_{R,D} = g_{R,D} \beta y_{S,R} + n_D \end{cases} \quad (1)$$

式中: $y_{S,R}$ 和 $y_{R,D}$ 分别表示中继和目的节点接收到的信号; x_S 表示源节点S发送的符号; P_S 表示S的发射功率; $h_{S,R}$ 和 $g_{R,D}$ 分别表示无线链路 $S \rightarrow R$ 和 $R \rightarrow D$ 的信道衰落系数; n_R 和 n_D 分别表示中继和目的节点的加性高斯白噪声,服从均值为0、方差为 N_0 的复高斯分布; $\beta = \sqrt{P_R} / \sqrt{P_S |h_{S,R}|^2 + N_0}^{[21]}$ 表示中继R的放大转发系数; P_R 表示中继的发射功率。将 β 代入式(1)可以得到:

$$y_{R,D} = \frac{\sqrt{P_R} \sqrt{P_S}}{\sqrt{P_S |h_{S,R}|^2 + N_0}} h_{S,R} g_{R,D} x_S + n'_D \quad (2)$$

式中: n'_D 为目的节点接收到的等效噪声,服从均值为0的复高斯分布,其方差为:

$$N_0' = \left(\frac{P_R |g_{R,D}|^2}{P_S |h_{S,R}|^2 + N_0} + 1 \right) N_0 \quad (3)$$

根据式(2)和式(3), 目的节点D的接收信噪比为:

$$\rho_D = \frac{\gamma_1 \gamma_2 \eta_S \eta_R}{\gamma_1 \eta_S + \gamma_2 \eta_R + 1} \quad (4)$$

式中: 信道增益 $\gamma_1 = |h_{S,R}|^2$; 信道增益 $\gamma_2 = |g_{R,D}|^2$; $\eta_S = P_S/N_0$; $\eta_R = P_R/N_0$ 。网络信息传输中断事件定义为从源节点到目的节点的可达速率小于中断门限速率 R_{th} , 信息传输中断概率表示中断事件的发生概率, 如式(5)所示:

$$P_{out} = P_r \left\{ \frac{1}{2} \ln(1 + \rho_D) < R_{th} \right\} \quad (5)$$

式中: $P_r\{\cdot\}$ 表示概率函数。

2 马尔可夫决策过程建模

在EH多中继协作网络中, 以网络的长期平均中断概率为优化目标, 设计网络传输策略, 即根据当前传输周期的网络状态选择一个EH中继节点并确定其发射功率。网络传输策略的设计和优化受多个因素的影响, 例如: 太阳能辐照度的不确定性和随机性、EH中继节点电池电量、无线信道的衰落状态等。根据随机优化理论, 本文使用MDP对EH多中继协作网络进行建模, 网络的传输周期对应MDP模型的决策周期。MDP模型架构主要包括动作空间、状态空间、收益函数、传输策略和价值函数。

2.1 动作空间

$\mathcal{A} = \mathcal{A}_s \times \mathcal{A}_p$ 表示MDP模型的2维动作空间, 其中 $\times$ 表示笛卡尔积; $\mathcal{A}_s = \{0, 1, 2, \dots, N\}$ 表示中继选择动作子空间; $\mathcal{A}_p = \{0, 1, 2, \dots, N_b - 1\}$ 表示中继发射功率动作子空间; N_b 表示电池容量。在该MDP模型中, 动作 $a_s \in \mathcal{A}_s$ 表示在当前传输周期内选择的EH中继节点序号, 动作 $a_p \in \mathcal{A}_p$ 表示在当前周期内该EH中继的发射功率。如果在当前传输周期中网络选择动作 $\mathbf{a} = (a_s, a_p) \in \mathcal{A}$, 那么EH中继 R_{a_s} 将以发射功率 $P_R = a_p P_U$ 转发源节点信号, 其中 P_U 表示基本发射功率, 它与基本能量单位 E_U 的关系可以表示为 $E_U = P_U T_M / 2$ 。此外, 如果中继选择动作 $a_s = 0$, 则表示在当前传输周期内无EH中继转发源节点信号。

2.2 状态空间

$$\mathcal{S} = \mathcal{S}_E \times \prod_{n=1}^N (\mathcal{S}_{B_n} \times \mathcal{S}_{H_n} \times \mathcal{S}_{G_n}) \quad \text{表示}$$

MDP模型的 $3N+1$ 维状态空间, 其中: $\mathcal{S}_E = \{0, 1, \dots, N_e - 1\}$ 代表太阳能EH状态子空间; $\mathcal{S}_{B_n} = \{0, 1, \dots, N_b - 1\}$ 代表EH中继 R_n 的电池电量状态子空间; $\mathcal{S}_{H_n}$ 和 $\mathcal{S}_{G_n}$ 分别代表无线链路 $S \rightarrow R_n$ 和 $R_n \rightarrow D$ 的信道衰落状态子空间, 它们均可用 $\mathcal{S}_c = \{0, 1, \dots, N_c - 1\}$ 表示。因此, 在一个传输周期内网络系统状态可以表示为 $\mathbf{s} = (e, b_1, h_1, g_1, b_2, h_2, g_2, \dots, b_N, h_N, g_N) \in \mathcal{S}$ 。

接下来, 分别详细说明太阳能EH状态、电池电量状态和信道衰落状态的定义。

本文采用具有 N_e 个状态的隐马尔可夫模型 (hidden Markov model, HMM) 描述太阳能 EH 随机状态, 模型所有参数是由太阳辐照度实测数据训练得到的^[20]。该实测数据来自于美国伊丽莎白城州立大学的太阳观测站点, 数据收集时段为 2008 至 2010 年, 数据采样周期为 300 秒。虽然文献[20]采用了北美地区的太阳辐照度实测数据, 但是其建模方法具有通用性。HMM 模型的每个状态对应一个正态分布, 不同状态所对应的正态分布参数不一样, 分别代表不同的太阳辐照强度。

在 MDP 模型中太阳能 EH 状态 $e \in \mathcal{S}_e$ 定义为 HMM 的状态序号, 代表在当前传输周期内太阳辐照度 x 服从正态分布 $N(\mu_e, \rho_e)$, 因此在当前周期内每个 EH 中继收集到的能量可以计算为 $E_H = x\Omega T_M v$, 其中 Ω 是太阳能面板的面积, v 是能量转换率。在单个传输周期内, 每个 EH 中继将收集到的能量 E_H 以基本能量单位 E_U 进行量化, 然后再将量化后的能量单元储存在其电池中用于后续周期的信息转发^[20]。

在电池能量管理中使用基本能量单位 E_U 量化每个 EH 中继的电池容量和可用电量。假设每个 EH 中继的电池容量有限, 且均为 $E_U(N_b - 1)$ 。在 MDP 模型中电池状态 $b_n \in \mathcal{S}_{b_n}$ 表示 EH 中继 R_n 电池中可用电量为 $E_U b_n$ 。在单个传输周

期中, 如果选择中继 R_n 转发源节点信号, 那么其转发信号所消耗的能量不能超过当前电池电量, 即: $a_p \leq b_n$ 。

本文采用具有 N_c 个状态的马尔可夫链描述无线信道衰落状态, 使用一组阈值 $\{0 = \Gamma_0, \Gamma_1, \dots, \Gamma_{N_c} = \infty\}$ 将无线信道增益的取值区间量化为 N_c 个子区间, 各个子区间对应不同的信道衰落状态^[22]。在当前周期中如果信道增益 γ 在子区间 $[\Gamma_i, \Gamma_{i+1})$ 内, 则定义该信道的衰落状态为 $i \in \mathcal{S}_{c_i}$ 。

2.3 收益函数

由于 MDP 的优化目标是 EH 多中继协作网络的长期平均中断概率, 因此 MDP 收益函数需要反映网络的信息传输中断概率。设 $P_{\text{out}}(\mathbf{s}, \mathbf{a})$ 表示网络在状态 $\mathbf{s} = (e, b_1, h_1, g_1, b_2, h_2, g_2, \dots, b_N, h_N, g_N)$ 下采取动作 $\mathbf{a} = (a_s, a_p)$ 时的条件中断概率, 它与网络在当前传输周期中的动作 $\mathbf{a}$ 和所选中继的信道状态直接相关, 因此 MDP 收益函数可以定义为:

$$R_a(\mathbf{s}) \triangleq 1 - P_{\text{out}}(\mathbf{s}, \mathbf{a}) = 1 - P_{\text{out}}(h_{a_s}, g_{a_s}, a_p) \quad (6)$$

EH 多中继协作网络在确定中继选择动作 a_s 后就等效为单中继网络, 在高信噪比下根据式(4)和式(5), 式(6)中的条件中断概率可以近似为:

$$\begin{aligned}
& P_{\text{out}}(h_{a_s}, g_{a_s}, a_p) \\
&= P_r \left\{ \frac{1}{2} \ln \left(1 + \frac{\gamma_1 \gamma_2 \eta_S \eta_R}{\gamma_1 \eta_S + \gamma_2 \eta_R} \right) < R_{\text{th}} \mid h_{S,R} = h_{a_s}, g_{R,D} = g_{a_s}, P_R = a_p P_U \right\} \\
&= P_r \left\{ \frac{\gamma_1 \gamma_2 \eta_S \eta_R}{\gamma_1 \eta_S + \gamma_2 \eta_R} < 2^{2R_{\text{th}}} - 1 \mid h_{S,R} = h_{a_s}, g_{R,D} = g_{a_s}, P_R = a_p P_U \right\}
\end{aligned} \quad (7)$$

根据不等式 $\frac{xy}{x+y} \leq \min(x, y)$ 可得:

$$\begin{aligned}
& P_{\text{out}}(h_{a_s}, g_{a_s}, a_p) \\
&\geq P_r \left\{ \min(\gamma_1 \eta_S, \gamma_2 \eta_R) < 2^{2R_{\text{th}}} - 1 \mid h_{a_s}, g_{a_s}, a_p \right\} \\
&\geq 1 - P_r \left\{ (\gamma_1 \eta_S \geq 2^{2R_{\text{th}}} - 1) \mid h_{a_s} \right\} \cdot P_r \left\{ (\gamma_2 \eta_R \geq 2^{2R_{\text{th}}} - 1) \mid g_{a_s}, a_p \right\} \\
&\geq 1 - P_r \left\{ \gamma_1 \geq \gamma_{\text{th1}} \mid \Gamma_h \leq \gamma_1 < \Gamma_{h+1} \right\} \cdot P_r \left\{ \gamma_2 \geq \gamma_{\text{th2}} \mid \Gamma_g \leq \gamma_2 < \Gamma_{g+1} \right\}
\end{aligned} \quad (8)$$

式中: $\gamma_{\text{th1}} = (2^{2R_{\text{th}}} - 1)/\eta_S$; $\gamma_{\text{th2}} = (2^{2R_{\text{th}}} - 1)/\eta_R$;

$h = h_{a_s}$; $g = g_{a_s}$ 。接下来分情况计算条件中断概率。

● 当 $\gamma_{\text{th1}} \geq \Gamma_{h+1}$ 或 $\gamma_{\text{th2}} \geq \Gamma_{g+1}$ 时:

$$P_{\text{out}}(h_{a_s}, g_{a_s}, a_p) = 1 \quad (9)$$

● 当 $\gamma_{\text{th1}} \leq \Gamma_h$ 并且 $\gamma_{\text{th2}} \leq \Gamma_g$ 时:

$$P_{\text{out}}(h_{a_s}, g_{a_s}, a_p) = 0 \quad (10)$$

● 当其他情况时:

$$\begin{aligned}
& P_r \left\{ \gamma_1 \geq \gamma_{\text{th1}} \mid \Gamma_h \leq \gamma_1 < \Gamma_{h+1} \right\} \\
&= \frac{P_r \left\{ \max(\gamma_{\text{th1}}, \Gamma_h) \leq \gamma_1 < \Gamma_{h+1} \right\}}{P_r \left\{ \Gamma_h \leq \gamma_1 < \Gamma_{h+1} \right\}} = \frac{e^{-\max(\gamma_{\text{th1}}, \Gamma_h)/\gamma_0} - e^{-\Gamma_{h+1}/\gamma_0}}{e^{-\Gamma_h/\gamma_0} - e^{-\Gamma_{h+1}/\gamma_0}}
\end{aligned} \quad (11)$$

同理可得:

$$P_r \left\{ \gamma_2 \geq \gamma_{\text{th2}} \mid \Gamma_g \leq \gamma_2 < \Gamma_{g+1} \right\} = \frac{e^{-\max(\gamma_{\text{th2}}, \Gamma_g)/\gamma_0} - e^{-\Gamma_{g+1}/\gamma_0}}{e^{-\Gamma_g/\gamma_0} - e^{-\Gamma_{g+1}/\gamma_0}} \quad (12)$$

综合以上各种情况, 式(6)中的条件中断概率表示为:

$$P_{\text{out}}(h_{a_s}, g_{a_s}, a_p) = \begin{cases} = 1, \gamma_{\text{th1}} \geq \Gamma_{h+1} \text{ 或 } \gamma_{\text{th2}} \geq \Gamma_{g+1} \\ = 0, \gamma_{\text{th1}} \leq \Gamma_h \text{ 且 } \gamma_{\text{th2}} \leq \Gamma_g \\ \geq 1 - \frac{e^{-\max(\gamma_{\text{th1}}, \Gamma_h)/\gamma_0} - e^{-\Gamma_{h+1}/\gamma_0}}{e^{-\Gamma_h/\gamma_0} - e^{-\Gamma_{h+1}/\gamma_0}} \cdot \frac{e^{-\max(\gamma_{\text{th2}}, \Gamma_g)/\gamma_0} - e^{-\Gamma_{g+1}/\gamma_0}}{e^{-\Gamma_g/\gamma_0} - e^{-\Gamma_{g+1}/\gamma_0}}, \text{其他} \end{cases} \quad (13)$$

2.4 传输策略与价值函数

MDP 传输策略 $\pi: \mathcal{S} \rightarrow \mathcal{A}$ 表示 EH 多中继协作网络在给定状态 $\mathbf{s}$ 下的动作 $\mathbf{a}$, 即所选中继的序号和发射功率。MDP 的优化目标是网络的长期平均中断概

率,因此MDP动作价值函数 $q_\pi(\mathbf{s}, \mathbf{a})$ 定义为网络根据策略 π 从状态 $\mathbf{s}$ 开始执行动作 $\mathbf{a}$ 后,所获得的总收益的期望值^[23]:

$$q_\pi(\mathbf{s}, \mathbf{a}) = E_\pi \left[\sum_{k=0}^{\infty} \lambda^k R_{t+k} | S_t = \mathbf{s}, A_t = \mathbf{a} \right] \quad (14)$$

式中: $E_\pi[\cdot]$ 表示在给定策略 π 时的期望值; $\lambda \in [0, 1]$ 为折扣因子; t 表示起始周期; R_{t+k} 表示网络在决策周期 $t+k$ 时的收益。

在MDP中如果存在一个策略 π^* ,使得 $q_{\pi^*}(\mathbf{s}, \mathbf{a}) \geq q_\pi(\mathbf{s}, \mathbf{a}), \forall \pi, \forall \mathbf{s} \in \mathcal{S}, \forall \mathbf{a} \in \mathcal{A}$ 成立,则称 π^* 为最优传输策略, $q^*(\mathbf{s}, \mathbf{a})$ 为最优动作价值函数。最优动作价值函数满足贝尔曼最优方程,通过求解该方程即可获得网络的最优传输策略 π^* 。

3 基于Q学习的传输策略学习与优化

由于EH多中继协作网络包含的通信节点数量较多,其MDP模型的随机状态数目随着EH中继节点数目增多呈指数增长,属于超大规模MDP,网络在实际应用中很难获得准确的系统状态转移概率特性,因此无法直接采用传统的DP方法求解网络最优传输策略。

Q学习算法是一种基于动作价值函数的强化学习算法,无需状态转移概率便可求解贝尔曼最优方程,从而获得网络的最优传输策略 π^* ^[23]。在Q学习算法中,每一个“状态-动作”组合 $(\mathbf{s}, \mathbf{a})$ 对

应一个Q值 $Q(\mathbf{s}, \mathbf{a})$,它可以看作是动作价值函数 $q_\pi(\mathbf{s}, \mathbf{a})$ 的近似值,用于评估在特定状态下采取不同动作的总收益。该算法通过不断生成“状态-动作”组合 $(\mathbf{s}, \mathbf{a})$ 形成学习序列(episode),从而迭代更新Q值并优化传输策略。当算法收敛后Q值能以1的概率收敛到最优动作价值函数 q^* ,并可获得最优传输策略。

在基于Q学习算法的网络传输策略学习与优化过程中,首先需要初始化网络的所有Q值,然后随机选择一个系统状态启动一个学习序列。每一步(one step)学习包括根据状态生成动作、确定收益和下一步状态、更新Q值。不断重复单步学习直至该episode结束。如果Q值收敛,则停止更新Q值;反之,则启动新的episode直至Q值收敛。最后,基于收敛后的Q值为每个状态选择最优动作,从而得到网络的最优传输策略。该学习与优化方法描述如下:

输入 初始探索率 ε_i 、最终探索率 ε_f 、初始学习步长 α_i 、最终学习步长 α_f 、折扣因子 λ

输出 最优传输策略 π^*

1. Repeat

2. 初始化系统状态 $\mathbf{s}$

3. Repeat

4. 根据式(16)更新探索率 ε 和学习步长 α

5. 根据式(15),当前状态 $\mathbf{s}_t$ 采用 ε -greedy策略

选择动作 $\mathbf{a}$,根据式(6)和(13)计算

收益 R_t ,

6 根据式 (17) 更新电池状态, 根据文献 [20] 更

新太阳能 EH 状态和信道衰落状态, 得到下

一步状态 s_{t+1}

7. 根据式(18)更新 Q 值

8. 更新状态: $s_t \leftarrow s_{t+1}$

9. Until 当前 episode 结束

10. Until $\forall s, a, Q(s, a)$ 收敛

3.1 根据 MDP 状态选择动作

为了在学习过程中探索和遍历更多的“状态-动作”组合, 本文采用 ϵ -greedy 策略根据 MDP 当前状态 s_t 选择动作 a_t , 如式(15)所示:

$$a_t = \begin{cases} \arg \max_a Q(s_t, a), & P_t > \epsilon \\ \text{随机均匀选择}, & P_t \leq \epsilon \end{cases} \quad (15)$$

式中: $P_t \in [0, 1]$ 表示服从均匀分布的随机数; $\epsilon \in [0, 1]$ 表示探索率。为提高整体学习效果, 在学习过程前期可以设置较大的探索率 ϵ , 这有利于探索和遍历到更多的“状态-动作”组合, 避免陷入局部最优。随着学习步数的逐渐增加, 可以逐渐降低 ϵ , 有利于算法收敛。因此, 本文采用基于学习步数的线性衰减方法设置探索率, 使其在学习过程中稳定减小, 如式(16)所示^[16]:

$$\epsilon = \epsilon_i - \frac{(\epsilon_i - \epsilon_f) n_{\text{step_cur}}}{n_{\text{step_tot}}} \quad (16)$$

式中: $n_{\text{step_cur}}$ 表示当前学习步数; $n_{\text{step_tot}}$

表示总学习步数; ϵ_i 和 ϵ_f 分别是初始探索率、最终探索率。

3.2 确定收益与下一步状态

在生成当前“状态-动作”组合 (s_t, a_t) 后, 根据式(6)和(13)可以计算得到当前收益 R_t , 同时也能够根据 MDP 状态转移关系确定下一步状态 s_{t+1} 。在 EH 多中继协作网络中, 由于采用“收集-储存-使用”的能量管理模型, EH 中继节点在当前周期内收集的能量先存储在电池中, 然后再用于后续周期的信息转发。因此网络在当前周期采取动作 $a_t = (a_{s,t}, a_{p,t})$ 时, EH 中继节点的电池状态转移关系如式(17)所示:

$$b_{n,t+1} = \begin{cases} \min(b_{n,t} + q_t - a_{p,t}, N_b - 1), & n = a_{s,t} \\ \min(b_{n,t} + q_t, N_b - 1), & n \neq a_{s,t} \end{cases} \quad (17)$$

式中: $b_{n,t}$ 和 $b_{n,t+1}$ 分别表示 EH 中继 R_n 在当前周期和下一周期的电池状态; q_t 表示 EH 中继在当前周期内收集到的能量单元数目。除了 EH 中继的电池状态, 另外, 根据文献[20]可以更新太阳能 EH 状态和无线信道衰落状态。

3.3 更新 Q 值

在确定当前收益和下一步状态后, 可以使用时序差分方法更新当前“状态-动作”组合的 Q 值^[23]:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left[R_t + \lambda \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t) \right] \quad (18)$$

式中: $\alpha \in [0, 1]$ 为学习步长。 α 对算法性能有直接影响: 较大的 α 能够加快学习效率, 但同时也容易导致 Q 值不收敛; 而较小的 α 更有利于 Q 值收敛, 但是会降低收敛速度。因此, 本文采用与式(16)相同的基于学习步数的线性衰减方法更新 α , 使其在学习过程中稳定减小, 以兼顾收敛性和学习效率。

在 EH 多中继协作网络, MDP 系统状态数目 $n_s = (N_b N_c N_e)^N N_e$ 。因此随着中继节点数目 N 增加, 系统状态数目呈指数增长, 形成超大状态空间。本文所提优化方法使用了 Q 学习算法, 学习算法收敛所需的训练时序 step 数目与系统状态数目直接相关, 优化方法的时间复杂度取决于系统状态数目 n_s 。同时, 用于评价传输策略的 Q 函数表格所包含的元素个数 $n_o = (N_b N_c N_e)^N N_e N N_b$, 因此优化方法的空间复杂度表示为: $\mathcal{O}((N_b N_c N_e)^N N_e N N_b)$ 。

4 仿真结果与分析

本节描述了 EH 多中继协作网络传输性能的计算机仿真结果, 使用蒙特卡罗方法计算网络的长期平均中断概率。在相同 MDP 模型下通过与文献[16-19]中 SARSA、DQN、Actor-Critic 等动态优化方法进行对比来评估本文优化方法和传输策略性能。由于 EH 中继节点发射功率是根据传输策略动态生成的, 因此在仿真实验中使用基于 1mW 发射功率的归一化信噪比 (signal-to-noise-

ratio, SNR) 来表示通信链路质量。

仿真实验中 EH 多中继协作网络和 MDP 模型主要参数详见表 1。太阳能转换效率、EH 节点的太阳能面板和信号发射功率等关键参数均符合 EH 无线传感器在实际应用中的配置需求^[5]。为使得系统模型更加符合实际应用, 本文使用了文献[20]中基于太阳辐照度实测数据的随机 EH 模型来描述太阳能 EH 的随机特性, 也借鉴该文献将 MDP 决策周期设置为 300 秒。所有无线信道的衰落系数由 Jakes 信道模型生成^[20], 可以有效模拟真实环境中的信道衰落情况, 同时也分析了在实际应用中可能存在的信道估计误差对网络传输性能的影响。

表 1 仿真参数

Table 1 Simulation parameters

参数名称	参数值
MDP 决策周期 (T_M)	300 s
太阳能 EH 状态数目 (N_e)	4
能量转换效率 (ν)	20%
太阳能面板面积 (Ω)	4 cm ²
信道状态数目 (N_c)	4
信道量化门限	{0, 0.6, 2.0, 3.0, ∞ }
归一化多普勒频率 (f_D)	0.05
电池状态数目 (N_b)	4
基本发射功率 (P_U)	50 mW
源节点发射功率 (P_S)	P_U
网络目标速率 (R_{th})	2 bit/s/Hz
折扣因子 (λ)	0.99
平均信道增益 (γ_0)	1

图 2 展示了基于 Q 学习算法得到的 EH 双中继协作网络的最优动作 $\mathbf{a}^* = (a_s^*, a_p^*)$ 分别在低、中、高三种 SNR 条

件下的概率分布情况。其中,最优动作 a_s^* 和 a_p^* 的出现概率定义为该动作取值在系统状态空间中的出现比例。图 (a) 揭示了在不同 SNR 时 a_p^* 的概率分布。在低 SNR 条件下,动作 $a_p^*=2$ 和动作

$a_p^*=3$ 的出现概率明显高于在高 SNR 条件下的概率;随着 SNR 增加,动作 $a_p^*=1$ 的出现概率逐步上升。图 (b) 显示,中继选择动作 a_s^* 的分布在低、中、高三种 SNR 下均表现稳定。

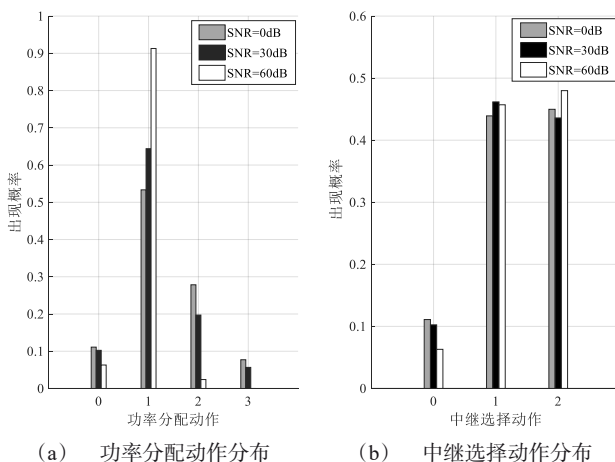


图2 最优功率分配与最优中继选择分布情况 ($P_U=50$ mW, $\Omega=4$ cm², $R_{th}=2$ bit/s/Hz)

Fig. 2 Distribution of optimal power allocation and optimal relay selection ($P_U=50$ mW, $\Omega=4$ cm², $R_{th}=2$ bit/s/Hz)

图3展示了在不同太阳能面板面积 Ω 和基本发射功率 P_U 下, EH 双中继协作网络基于最优传输策略 π^* 的长期平均中断概率。从图中可以看到,网络中断概率存在饱和特性,在极高 SNR 条件下中断概率趋近于极限值,原因如下。文献[24]采用 DP 方法研究了太阳能 EH 点对点通信链路的中断概率性能,研究结果显示当链路 SNR 趋向于无穷大时 EH 通信链路的中断概率收敛到极限值,该极限值等于 EH 发射节点的电池电量耗尽概率。本文则采用 Q 学习算法研究了 EH 多中继协作网络的最

优传输策略和中断概率性能。因为在求解 MDP 时,分别采用 DP 方法和 Q 学习算法得到的最优策略 π^* 在理论上是等价的^[23],因此采用本文基于 Q 学习方法得到的 EH 多中继协作网络中断概率在极高 SNR 下也趋向于极限值,该极限值取决于所有 EH 节点电池电量耗尽概率。

图3结果表明,参数 Ω 对网络中断概率的影响主要表现在高 SNR 区域,此时 Ω 从 4 cm² 增加到 6 cm² 能够显著降低中断概率。这主要是因为: Ω 的大小直接影响 EH 中继在周期 T_M 内的能量收

集能力, 当增大 Ω 时可以收集更多能量, 降低电池电量耗尽概率, 从而提高网络在高SNR下的中断概率性能。同时, 在高SNR区域, 较低的 P_U 能够获得更好的中断性能。这是因为减小 P_U 可降低基本能量单位 E_U 的数值, 当收集的总能量不变时使得电池存储更多能量单元, 在极高SNR时可以降低电池电量耗尽概率从而提升中断性能的极限性能。另外, 在低SNR区域增加 P_U 可有效降低中断概率。

图4展示了在不同目标速率 R_{th} 和

源节点发射功率 P_s 下, EH双中继协作网络基于最优传输策略 π^* 的长期平均中断概率。图中结果表明, 目标速率 R_{th} 和源节点发射功率 P_s 对网络中断概率的影响主要体现在低SNR区域, 在该区域内降低 R_{th} 或增加 P_s 均可有效降低网络中断概率。目标速率 R_{th} 表示网络发生中断时的最低速率门限, 因此 R_{th} 的增大必然导致中断概率升高。源节点发射功率 P_s 则主要影响网络吞吐量, 增加 P_s 能够提高吞吐量从而降低网络中断概率。

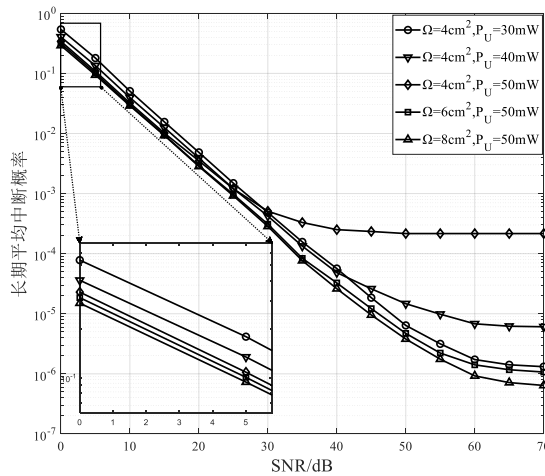


图3 不同 Ω 和 P_U 下的长期平均中断概率 ($R_{th}=2$ bit/s/Hz)

Fig. 3 Long-term outage probability for different Ω and P_U ($R_{th}=2$ bit/s/Hz)

图5对比了EH双中继协作网络在不同传输策略下的信息传输中断概率性能, 其中用于对比的传输策略包括: 分别采用强化学习SARSA算法^[16]、深度强化学习DQN^[17-18]和Actor-Critic^[17,19]算法获得的动态策略, 以及仅根据网络部分状态信息选择中继的四种短视策略。图5结果显示, 网络在本文Q学习方法

所得传输策略下的中断概率性能均优于其他对比策略。Q学习方法求解贝尔曼最优方程, 在理论上可以得到最优策略; 但是SARSA算法求解贝尔曼方程仅能获得较为保守的传输策略, 这会降低网络的传输性能^[23]。由于DQN和Actor-Critic算法使用神经网络近似Q函数, 神经网络在训练过程中会存在过度

估计和不稳定性等问题^[25], 因此这两种算法所对应的网络中断性能均逊色于本文优化方法。本文优化方法相对于DQN算法, 在中断概率为 $10^{-2} \sim 10^{-3}$ 时可获得3~5dB的信噪比增益; 相对于

Actor-Critic和SARSA算法在中断概率为 $10^0 \sim 10^{-3}$ 时可分别获得2~3dB和3~7dB的信噪比增益。在较高SNR区域, 本文优化方法对应的网络中断性能优势相较DQN和SARSA算法更加明显。

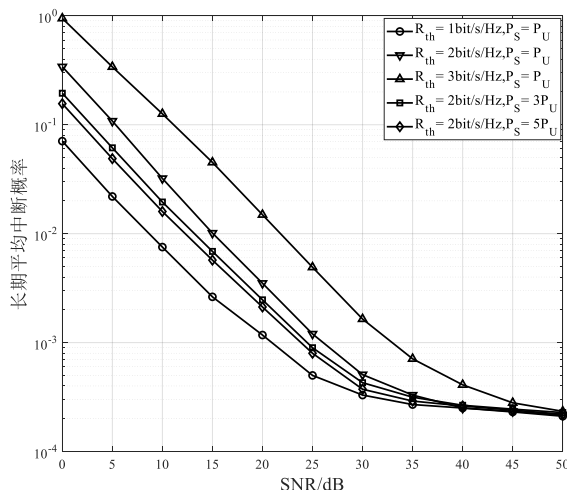


图4 不同 R_{th} 和 P_s 下的长期平均中断概率 ($P_U=50$ mW, $\Omega=4$ cm²)

Fig. 4 Long-term outage probability for different R_{th} and P_s ($P_U=50$ mW, $\Omega=4$ cm²)

另外, 在四种短视策略中电池状态策略1、2分别定义为: 选择电池电量最多的EH中继转发源节点信号, 但前者消耗电池所有电量转发信号, 后者仅消耗1个能量单元; 信道状态策略1、2分别定义为: 根据无线信道衰落状态动态选择EH中继, 但前者消耗电池所有电量, 后者仅消耗1个能量单元。由于这四种短视策略未能充分考虑网络的完整状态信息, 因此其对应的网络中断性能相对较差。如果短视策略消耗EH中继的全部电池电量, 则显著增加了电池电量耗尽概率, 会导致在高SNR下性能较差; 若短视策略消耗最少电量, 在高SNR区域网络中断性能较好, 但在

低SNR区域性能较差。

图6对比了EH双中继与三中继、四中继协作网络在最优传输策略 π^* 下的长期平均中断概率性能。从图中可以观察到, 对于相同的源节点发射功率 P_s , 网络中断概率随着中继数目 N 的增加而降低, 且随着SNR升高, 中断性能改善效果愈发明显。因为在极高SNR下, 网络的中断概率取决于所有EH中继电池电量耗尽概率。相较于双中继协作网络, 三中继和四中继协作网络的EH中继数目更多, 因此所有EH中继电池电量同时耗尽的概率更低, 从而在高SNR下表现出更优的中断性能。

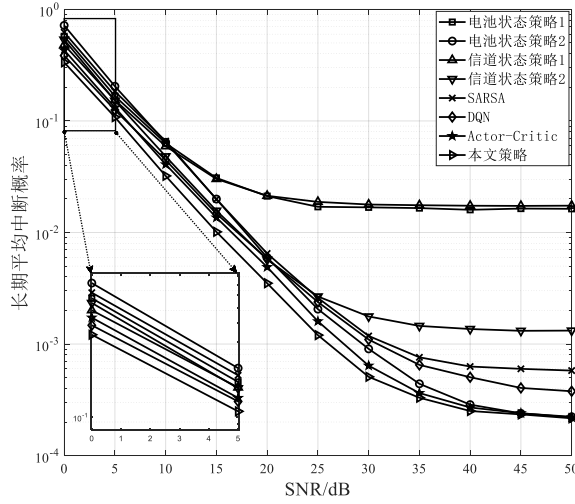


图 5 不同传输策略下的长期平均中断概率 ($\Omega=4 \text{ cm}^2$, $P_U=50 \text{ mW}$, $R_{th}=2 \text{ bit/s/Hz}$)

Fig. 5 Long-term outage probability for different policies ($\Omega=4 \text{ cm}^2$, $P_U=50 \text{ mW}$, $R_{th}=2 \text{ bit/s/Hz}$)

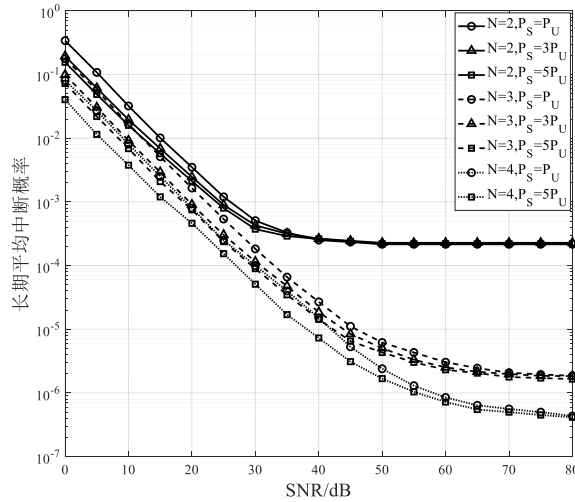


图 6 不同 P_s 和 N 下的长期平均中断概率 ($\Omega=4 \text{ cm}^2$, $P_U=50 \text{ mW}$, $R_{th}=2 \text{ bit/s/Hz}$)

Fig. 6 Long-term outage probability for different P_s and N ($\Omega=4 \text{ cm}^2$, $P_U=50 \text{ mW}$, $R_{th}=2 \text{ bit/s/Hz}$)

为了分析在实际应用中信道估计误差对本文所提优化方法性能的影响, 根据文献[26], 假设无线信道衰落系数的

估计值相对于真实值存在加性误差, 并且信道估计误差服从均值为0、方差为 σ_e^2 的复高斯分布。当方差 σ_e^2 为0时表示

不存在信道估计误差。图 7 展示了在不同信道估计误差的概率方差 σ_e^2 和 MDP 决策周期 T_M 下, EH 双中继协作网络基于最优传输策略 π^* 的长期平均中断概率。可以看到, 信道估计误差对网络中断性能的影响程度在不同 SNR 下存在明显差异。根据文献[24], 在极高 SNR 下 EH 通信节点的最优传输策略仅与电池电量状态相关, 与信道状态无关。因此在极高 SNR 下信道估计误差不会影响网络中断概率性能。同时, 在中低 SNR 区域, 网络最优传输策略与信道衰落状态和电池电量状态均相关, 因此信道估计误差会影响网络的中继选择和功率分配, 从而影响网络中断概率性能。当 σ_e^2 越大时, 网络中断性能下降

越明显。当 σ_e^2 小于 0.1 时, 信道估计误差对网络中断性能没有显著影响。

同时从图 7 可以看到, 网络中断概率性能随 MDP 决策周期 T_M 缩短而提高, 在高 SNR 下提高效果尤为显著。这主要是因为, 当 T_M 缩短后电池能够存储的能量单元数目增多, 在极高 SNR 下 EH 中继的电池电量耗尽概率会减小, 从而有效提高了网络在高 SNR 下的中断性能。同时缩短 T_M 也可使网络在中低 SNR 下的中断性能也得到改善。但是当 T_M 缩短后, 电池状态子空间的状态数目 N_b 增大, 导致优化方法的空间复杂度和时间复杂度相应变高, 降低了学习和优化效率。

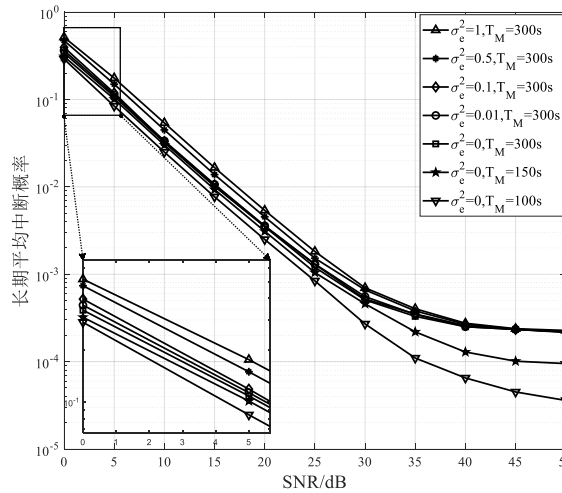


图 7 不同 σ_e^2 和 T_M 下的长期平均中断概率 ($\Omega=4 \text{ cm}^2$, $P_U=50 \text{ mW}$, $R_{th}=2 \text{ bit/s/Hz}$)

Fig. 7 Long-term outage probability for different σ_e^2 and T_M ($\Omega=4 \text{ cm}^2$, $P_U=50 \text{ mW}$, $R_{th}=2 \text{ bit/s/Hz}$)

5 结论

本文采用基于太阳能实测数据的随机EH模型,使用强化学习理论研究了EH多中继协作网络中继选择策略的优化问题。首先建立了网络的MDP模型,然后使用Q学习算法设计了网络传输策略的学习与优化过程。所设计的优化方法能够根据太阳能EH状态、信道衰落状态和EH中继电池电量,动态选择单个EH中继并确定其发射功率转发信号,以优化网络的信息传输中断概率。最后在计算机仿真实验中分析了网络配置参数对中断性能的影响,指出了网络中断概率存在饱和特性。同时将本文优化方法与SARSA、DQN和Actor-Critic等强化学习算法进行了对比,仿真结果显示网络在本文策略下的中断概率性能均明显优于对比策略。接下来会针对大规模EH多中继协作网络,进一步改进传输策略优化方法,尝试采用深度强化学习和多智能体强化学习等算法,降低优化方法的时间复杂度和空间复杂度,提升学习和优化效率。同时进一步改进网络部署方案和配置参数,采用混合供电等更加实际的EH解决方案,不断提高理论研究工作的实际应用价值。

参考文献

- [1] 李向阳, 尚飞, 闫宇博, 等. 智能物联网低功耗感知综述 [J]. 计算机研究与发展, 2024, 61(11): 2754-2775.
LI Xiangyang, SHANG Fei, YAN Yubo, et al. Survey on Low Power Sensing of AIoT [J]. Journal of Computer Research and Development, 2024, 61(11): 2754-2775.
- [2] 王新奕, 费泽松, 周一青, 等. 面向物联网的通感算智融合: 关键技术与未来展望 [J]. 电子与信息学报, 2025, 47(4): 888-908.
WANG Xinyi, FEI Zesong, ZHOU Yiqing, et al. Integrated Sensing, Communication, Computation, and Intelligence Towards IoT: Key Technologies and Future Directions [J]. Journal of Electronics & Information Technology, 2025, 47 (4): 888-908.
- [3] WILLIAMS A J, TORQUATO M F, CAMERON I M, et al. Survey of Energy Harvesting Technologies for Wireless Sensor Networks [J]. IEEE Access, 2021, 9: 77493-77510.
- [4] 郭潇瑶, 刘超, 包建荣, 等. 低复杂度放大转发多中继选择及功率分配 [J]. 电子学报, 2022, 50(11): 2602-2609.
GUO Xiaoyao, LIU Chao, BAO Jianrong, et al. Multi-Relay Selection and Power Allocation of Low Complexity Amplify-and-Forward Cooperation [J]. Acta Electronica Sinica, 2022, 50(11): 2602-2609.
- [5] MA D, LAN G, HASSAN M, et al. Sensing, computing, and communications for energy harvesting IoTs: A survey [J]. IEEE Communications Surveys and Tutorials, 2020, 22(2): 1222-1250.
- [6] 周欣彤, 肖琨. 综合上下行链路的无线能量收集协作网络资源分配 [J]. 浙江大学学报(工学版), 2023, 57(12): 2544-2552.
ZHOU Xintong, XIAO Kun. Resource allocation algorithm for wireless energy harvesting cooperative network integrating uplink and downlink [J]. Journal of Zhejiang University (Engineering Science),

- 2023, 57(12): 2544-2552.
- [7] MAO Bomin, TANG Fengxiao, KAWAMOTO Y, et al. AI Models for Green Communications Towards 6G [J]. IEEE Communications Surveys and Tutorials, 2022, 24(1): 210-247.
- [8] 陈珍珍, 冀保峰, 陈苏丹. 基于能量收集的中继协同无线物联传输方案 [J]. 计算机应用与软件, 2022, 39(11): 99-104.
- CHEN Zhenzhen, JI Baofeng, CHEN Sudan. Transmission scheme in wireless internet of things with relay cooperation based on energy harvesting [J]. Computer Applications and Software, 2022, 39(11): 99-104.
- [9] WU Yuan, QIAN Liping, HUANG Liang, et al. Optimal relay selection and power control for energy-harvesting wireless relay networks [J]. IEEE Transactions on Green Communications and Networking, 2018, 2(2): 471-481.
- [10] 张士兵, 韩刘可, 张美娟. 基于能量收集的全双工认知中继网络功率分配算法 [J]. 通信学报, 2020, 41(9): 139-146.
- ZHANG Shibing, HAN Liuke, ZHANG Meijuan. Power allocation algorithm of full duplex cognitive relay network based on energy harvesting [J]. Journal on Communications, 2020, 41(9): 139-146.
- [11] RAO S. Relay Selection for Energy Harvesting Cooperative NOMA [C]//2019 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS). Goa, India: IEEE, 2019: 1-6.
- [12] LI Wei, KU M L, CHEN Yan, et al. ZHU Shihua. Performance Analysis for Two-Way Network-Coded Dual-Relay Networks With Stochastic Energy Harvesting [J]. IEEE Transactions on Wireless Communications, 2017, 16(9): 5747-5761.
- [13] AHMED I, AHMED I, HOSSAIN J. Optimal stochastic power allocation and relay selection for energy harvesting systems [J]. IEEE Wireless Communications Letters, 2017, 6(4): 546-549.
- [14] HU Yang, QIU Chengrun, CHEN Yan. Lyapunov-Optimized Two-Way Relay Networks With Stochastic Energy Harvesting [J]. IEEE Transactions on Wireless Communications, 2018, 17(9): 6280-6292.
- [15] CHEN Xuehan, LIU Yong, CAI L X, et al. Resource Allocation for Wireless Cooperative IoT Network With Energy Harvesting [J]. IEEE Transactions on Wireless Communications, 2020, 19 (7): 4879-4893.
- [16] ORTIZ A, WEBER T, KLEIN A. Multi-Agent Reinforcement Learning for Energy Harvesting Two-Hop Communications With a Partially Observable System State [J]. IEEE Transactions on Green Communications and Networking, 2022, 5 (1): 442-456.
- [17] ZHAO Bin, ZHAO Xiaohui. Deep Reinforcement Learning Resource Allocation in Wireless Sensor Networks With Energy Harvesting and Relay [J]. IEEE Internet of Things Journal, 2022, 9(3): 2330-2345.
- [18] WANG Jindong, LIU Zhibin, LIU Xiaowu. Relay Selection for Energy-Harvesting Collaborative Communication Systems: A Deep Reinforcement Learning Approach [J]. IEEE Transactions on Industrial Informatics, 2025, 21(1): 771-780.
- [19] WANG Tong, WU Shiting, WANG Zhenyong, et al. A Multi-Featured Actor-Critic Relay Selection Scheme for Large-Scale Energy Harvesting WSNs [J]. IEEE Wire-

- less Communications Letters, 2022, 10(1): 180-184.
- [20] KU M L, CHEN Yan, LIU K J R. Data-driven stochastic transmission policies for energy harvesting sensor communications [J]. IEEE Journal on Selected Areas in Communications, 2015, 33(8): 1505-1520.
- [21] 邵彦彰, 王磊, 薛奕蕾. 采用分布式 Alamouti 编码的虚拟全双工中继通信方案 [J]. 西安交通大学学报, 2020, 54(12): 45-53.
- SHAO Yanzhang, WANG Lei, XUE Yilei. A Virtual Full-Duplex Relay Communication Scheme Based on Distributed Alamouti Protocol [J]. Journal of Xi'an Jiaotong University, 2020, 54(12): 45-53.
- [22] LOPEZ-SALAMANCA J J, SEMAN L O, BEREJUCK M D, et al. Finite-State Markov Chains Channel Model for CubeSats Communication Uplink [J]. IEEE Transactions on Aerospace and Electronic Systems, 2020, 56(1): 142-154.
- [23] 温广辉, 杨涛, 周佳玲, 等. 强化学习与自适应动态规划:从基础理论到多智能体系统中的应用进展综述 [J]. 控制与决策, 2023, 38(5): 1200-1230.
- WEN Guanghui, YANG Tao, ZHOU Jialing, et al. Reinforcement learning and adaptive/approximate dynamic programming: A survey from theory to applications in multi-agent systems [J]. Control and Decision, 2023, 38(5): 1200-1230.
- [24] LI Wei, KU M L, CHEN Yan, et al. On Outage Probability for Stochastic Energy Harvesting Communications in Fading Channels [J]. IEEE Signal Processing Letters, 2015, 22 (11): 1893-1897.
- [25] 丁世飞, 杜威, 张健, 等. 多智能体深度强化学习研究进展 [J]. 计算机学报, 2024, 47(7): 1547-1567.
- DING Shifei, DU Wei, ZHANG Jian, et al. Research progress of multi-agent deep reinforcement learning [J]. CHINESE JOURNAL OF COMPUTERS, 2024, 47 (7): 1547-1567.
- [26] 丁长文, 杨霖, 李高祥. 能量收集双向中继网络的高能效联合中继选择和功率分配算法 [J]. 电子学报, 2017, 45(5): 1124-1129.
- DING Changwen, YANG Lin, LI Gaoxiang. Energy-Efficient Relay Selection and Power Allocation for Energy Harvesting Two-Way Relay Network [J]. ACTA ELECTRONICA SINICA, 2017, 45(5): 1124-1129.