# Wireless-Powered Multi-Channel Backscatter Communications Under Jamming: A Cooperative Reinforcement Learning Approach

Dara Ron
Wireless Cyber Center
George Mason University
Fairfax, Virginia, USA
dron@gmu.edu

Kai Zeng
Wireless Cyber Center
George Mason University
Fairfax, Virginia, USA
kzeng2@gmu.edu

## ABSTRACT

Wireless-Powered Backscatter Communication (WPBC) is emerging as a promising technology for battery-less solutions to many Internet-of-Things (IoT) applications. In this paper, we study the channel selection and operation mode control problem in a multi-channel WPBC system that turns jamming signal into energy harvesting opportunities. We propose a cooperative reinforcement learning (RL) approach that enables multiple agents, namely the access point (AP) and backscatter device (BD), to exploit unknown jamming pattern present. The learning of jamming channels and patterns not only enables the AP to communicate with the BD with interference-free but also empowers the BD to harvest energy from the jamming signals. Additionally, we address the limitation of BDs, which cannot decode information and harvest energy simultaneously, by incorporating this constraint into our design of a cooperative RL approach. This innovative strategy empowers the BD to make intelligent decisions regarding its operational mode—whether to prioritize communication or energy harvesting. Our aim is to optimize the trade-off between throughput and energy harvesting efficiency, ensuring that data reception requirements are met while adhering to constraint on the energy level stored in the battery. Unlike traditional approaches that only consider paired states and actions, our proposed cooperative RL algorithm incorporates channel state, jamming state, and action. These elements represent channel operation, jamming experience, and channel selection, respectively. The awareness of jamming experience derived from the jamming state enables the AP agent to select an action that can evade jamming. The proposed scheme experienced low computation and storage overhead as an inherent feature of the algorithm. Remarkably, the results show that the proposed method achieves optimal performance under static and round-robin jamming, and even in scenarios of random jamming.

## CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.



*WiseML '24, May 31, 2024, Seoul, Republic of Korea*

ACM ISBN 979-8-4007-0602-8/24/05.
https://doi.org/10.1145/3649403.3656489

## KEYWORDS

Reinforcement Learning, energy harvesting, Backscatter communication, jamming detection, and wireless power transfer.

**ACM Reference Format:**
Dara Ron and Kai Zeng. 2024. Wireless-Powered Multi-Channel Backscatter Communications Under Jamming: A Cooperative Reinforcement Learning Approach. In *Proceedings of the 2024 ACM Workshop on Wireless Security and Machine Learning (WiseML '24), May 31, 2024, Seoul, Republic of Korea.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3649403.3656489

## 1 INTRODUCTION

Wireless-Powered Backscatter Communication (WPBC) has emerged as a promising technology that offers a unique solution for "battery-less" communication. It empowers Internet-of-Things (IoT) backscatter devices (BDs) to transmit data by reflecting and modulating incident RF signals, thus eliminating the need for costly and power-hungry RF transmitters [8]. Furthermore, BDs can harvest energy from ambient radio frequency (RF) signals, including information signals and unwanted/jamming signals. With these key features, WPBC stands as a technology capable of supporting long-term, cost-effective, and simplified communications and networking for IoT devices [14]. Anti-jamming and jamming exploration for energy harvesting have been studied independently in [16] and [1], respectively. Another work considers both sides, i.e., BD and jammer, to be strategic and formulate the problem in game-theoretic frameworks [10]. However, that frameworks assume that the BD and jammer know each other's action spaces or beliefs. Such assumptions may not always hold in real-life application scenarios. Furthermore, existing methods usually incur significant computational overhead to find an optimal solution. Additionally, in the game-theoretic approaches, stationary or heuristic behavior of one side is assumed while countermeasures of the other side are investigated. This family of methods is heuristic or empirical, and the theoretical performance guarantees of their solutions are not readily available.

In this study, we develop a cooperative RL algorithm for two tasks: 1) channel selection at AP for anti-jamming WPBC in multi-channel networks, and 2) intelligent decision-making regarding operational modes at BDs, determining whether to prioritize communication or energy harvesting. Unlike existing solutions, our proposed scheme operates without the need for prior knowledge of the underlying environment or the statistical features of the channel states. Moreover, RL features low storage and computational overhead while dynamically selecting optimal actions to evade or

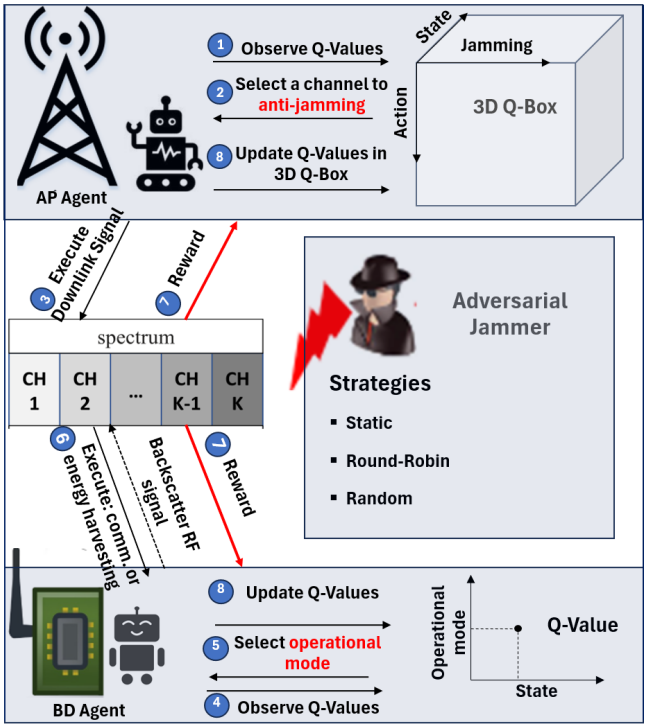


**Figure 1: A cooperative RL-based jamming approach.**

exploit jamming based on its operation mode without explicit assumptions about attacking policies. Thus, it proves highly suitable for resource-constrained BD. The jamming is generated based on three common strategies encountered in real-life application scenarios: static, round-robin, and random. To effectively address the jamming challenge, our proposed algorithm not only leverages a state-action pair to compute 2D Q-values but also introduces a new jamming state to obtain 3D Q-values. This enables the agent to consider previous experiences of channel jamming before making a decision to select a channel, thereby enhancing decision-making efficacy. Beyond merely learning to evade or exploit jamming, our approach empowers the BD to make decisions regarding its operational mode to optimize the trade-off between throughput and energy harvesting efficiency.

## 2 WPBC SYSTEM MODEL

We consider a scenario where both an AP and BD can operate across $K$ channels, while a jammer also operates on the same $K$ channels with the objective of disrupting communication between the AP and the BD, as illustrated in Figure 1. The AP has the ability to detect jamming signals, but the BD does not possess this capability. With this lacking channel detection ability, the BD must choose between harvesting energy or backscattering incident signals in-band. Additionally, communication between the AP and BD is prone to failure if the jammer operates on the same channel as the AP. The AP is equipped with self-interference cancellation capabilities, enabling it to mitigate interference that arises when simultaneously transmitting and receiving signals within the same frequency band.

Jamming is generated based on three distinct strategies: static, round-robin, and random. In the static strategy, the jamming signal remains fixed on a specific channel. Contrastingly, in the round-robin strategy, the jamming signal shifts sequentially across consecutive channels at each time step, looping back to the first channel once it reaches the last one. Finally, in the random strategy, the jamming signal randomly appears on a channel with uniform distribution.

At each time slot $t$, the AP endeavors to select a jamming-free channel with high-quality communication to achieve optimal throughput and ensure reliability, while the BD intelligently decides whether to operate in communication mode or energy harvesting mode, aiming to maximize the tradeoff between throughput and energy harvesting. Furthermore, if the BD's battery is running low, it will operate in the energy harvesting mode.

AP sends the following incident signal in channel $k$:

$$x_k(t) = \sqrt{2P_A}\cos\left(2\pi f_k t\right), \tag{1}$$

where $P_A$ is the incident signal power, and $f_k$ is the carrier frequency of the $k$-th channel expressed as $f_k = f_0 + k\Delta f$, where $\Delta f$ represents the channel spacing. Through the $k$-th channel, the received signal at the BD can be written as

$$\begin{aligned} y_{B,k}(t) &= \mathcal{R}\left\{h_k\sqrt{2P_A}\exp(j2\pi f_k t)\right\} + \eta_B(t) \\ &= |h_k|\sqrt{2P_A}\cos(2\pi f_k t + \theta_k) + \eta_B(t) \end{aligned} \tag{2}$$

where $h_k = |h_k|\exp(j\theta_k)$. Here, $h_k$ represents the channel coefficient between the AP and BD.

The attacker generates noise across the entire band in the $i$-th channel. With the presence of jamming, the received signal power is:

$$\begin{aligned} y_{B,k}(t) &= |h_k|\sqrt{2P_A}\cos(2\pi f_k t + \theta_k) \\ &\quad + \mathbf{1}_{(f_k == f_{J,i})}\left|h_{J,i}\right|U_{NAM}\cos\left(2\pi f_{J,i} t + \phi_J\right) + \eta_J(t), \end{aligned} \tag{3}$$

where $U_{NAM}$ represents the noise amplitude modulation with an amplitude power of $\frac{1}{2}|U_{NAM}| = P_J$, and $\mathbf{1}(fk = f_{J,i})$ is an indicator function that evaluates to 1 when the statement $(f_k = f_{J,i})$ is true [2, 11]. Assuming BD uses phase-shift keying (PSK) to modulate the incident signal, the signal reflected by the BD received at AP can be expressed as:

$$\begin{aligned} &y_{A,k}(t) \\ &= \sqrt{\alpha}\left|h_{B,k}h_k\right|\sqrt{2P_A}\cos(2\pi f_k t + 2\theta_k + \phi_k) + \mathbf{1}_{(f_k == f_{J,i})}\left|\hat{h}_{J,i}\right|U_{NAM} \\ &\cos\left(2\pi f_{J,i} t + \hat{\phi}_J\right) + \beta|h_{SI}|\sqrt{2P_A}\cos(2\pi f_k t) + \eta_A(t), \end{aligned} \tag{4}$$

where $\alpha \in [0, 1]$ denotes the signal reflection coefficient and $\phi_k$ represents the phase modulation. $\beta$ $(0 \leq \beta \ll 1)$ represents the Self-Interference Cancellation (SIC) ability of the AP. A value of $\beta = 0$ indicates complete nullification of the self-interference signal. Furthermore, $h_{B,k}$ and $h_{SI}$ represent the reflected and self-interference channels, respectively [7]. The noise term $\eta(t) = \{\eta_B(t), \eta_J(t), \eta_A(t)\}$, following a complex Gaussian distribution $\mathcal{CN}(0, \sigma^2)$, represents Additive White Gaussian Noise (AWGN). For simplicity, we assume that the noise powers are equal across all channels since they are significantly lower than the power of the carrier signal.

# 3 OPTIMIZATION PROBLEM FORMULATION

The proposed cooperative learning algorithm empowers the AP to select an anti-channel that maximizes network throughput, while enabling the BD to choose an operational mode that optimizes the trade-off between communication and energy harvesting.

## 3.1 Problem Formulation for AP

The objective function of the AP is the network throughput, which is given by:

$$C_{A,k}(t) = \log\big(1 + \Gamma_{A,k}(t)\big), \tag{5}$$

where $\Gamma_A(t)$ represents the signal-to-jamming-plus-self-interference and noise ratio (SJSNR). From (4), it can be expressed as:

$$\Gamma_{A,k}(t) = \frac{2\alpha\left|h_{B,k}h_k\right|^2 P_A}{2\mathbf{1}_{(f_k==f_{J,i})}\left|\hat{h}_{J,i}\right|^2 P_J + 2\beta|h_{SI}|^2 P_A + \sigma^2}. \tag{6}$$

The optimization problem for AP can be formulated as

$$\begin{aligned}\textbf{P1:}\ \max_{\pi(A|J,S_A)} \lim_{t\to\infty} \frac{1}{N_t}\sum_{\tau=1}^{N_t}\log\big(1+\Gamma_{A,k}(\tau)\big)\\ \text{s.t. } \pi(A|J,S_A)\in[0,1],\end{aligned} \tag{7}$$

where $N_t$ is the number of time slots consumed at time $t$, and $\pi(A|J,S)$ is the learning policy that guides the AP to select a channel $A$ that maximizes the long-term average throughput.

## 3.2 Problem Formulation for BD

We assume that the AP generates a downlink frame for transmission at every time slot. Frame transmission will fail if the BD decides to operate in energy harvesting mode or if the transmission signal is jammed. Unsuccessful frame transmissions are stored in a queue and concatenated with the next frame for transmission in the following time slot. The frame transmission at the AP follows a first-come-first-serve (FCFS) policy, meaning that the upcoming frame will be transmitted only after the frame in the queue has been completely transmitted. With each transmission, information regarding the total number of frames in the queue and the next upcoming frame is included in the frame header and transmitted to the BD. Being aware of this information enables the BD to decide which mode to operate in that effectively nullifies the frames in the AP's queue and maximizes its energy harvesting potential. Let $W(t)$ represent the frame sizes (in bits) stored in the AP's queue at time $t$. Based on queueing theory in [4], the queue length model is given by:

$$W_q(t+1) = \max(W_q(t) + D_{next}(t) - \rho(t)D_{Tx}(t), 0), \tag{8}$$

where $D_{next}(t)$ is the upcoming frame generated at time $t$ for transmission in the next time slot, and $D_{Tx}(t)$ is the total data that has been completely transmitted during a time slot. $D_{Tx}(t)$ can be expressed as:

$$D_{Tx}(t) = T\Delta B\log\left(1 + \frac{2|h_k|^2 P_A}{2\mathbf{1}(fk = f_{J,i})|h_{J,i}|^2 P_J + \sigma^2}\right), \tag{9}$$

where $T$ denotes the time slot duration. One of BD's goals is to nullify the frames stored in the AP's queue. Thus, $W_q(t+1) = 0$ if it satisfies the following constraint:

$$W_q(t) + D_{next}(t) - \rho(t)D_{Tx}(t) \le 0. \tag{10}$$

The objective of the BD is not only to nullify the frames in the queue but also to maximize its battery lifetime through harvesting energy. This highlights the importance of formulating both the energy harvesting process and the associated energy constraints. Similar to (8), the battery model is described by:

$$\mathcal{B}(t+1) = \max(\mathcal{B}(t) + (1-\rho(t))E_{H,k}(t) - E_{Com}(t), B_{Low}), \tag{11}$$

where $B_{Low}$ is the low battery level, $E_{H,k}(t)$ denotes the energy harvesting, and $E_{Com}(t)$ represents the energy consumption. Let

$$X(t) = \min(\mathcal{B}(t) + (1-\rho(t))E_{H,k}(t) - E_{Com}(t), B_{Low}). \tag{12}$$

Consequently, the battery level can be simplified as

$$\mathcal{B}(t+1) + X(t) = \mathcal{B}(t) + (1-\rho(t))E_{H,k}(t) - E_{Com}(t) + B_{Low}. \tag{13}$$

When the battery is low, the BD will alert the sensor to cease any further tasks, such as computing or communication, focusing solely on energy harvesting. Consequently, $X(t) = \mathcal{B}(t) + E_{H,k}(t)$ if $\mathcal{B}(t) \le B_{Low}$ and $X(t) = B_{Low}$ otherwise. Let $\varepsilon_{Max}$ represent the maximum energy allowed to be consumed during a time slot. Hence, the energy constraint can be formulated as:

$$\begin{aligned}\varepsilon_{Max} \ge \mathbb{E}[\mathcal{B}(t) - \mathcal{B}(t+1)] = \mathbb{E}[E_{Com}(t)] - (1-\rho(t))\mathbb{E}[E_{H,k}(t)]\\ - B_{Low} + \mathbb{E}[X(t)].\end{aligned} \tag{14}$$

The expected value of $z(t)$ is determined by:

$$\mathbb{E}[z(t)] = \bar{z}_t = \left(1 - \frac{1}{N_t}\right)\bar{z}(t-1) + \frac{1}{N_t}z(t), \tag{15}$$

If the BD considers operating in energy harvesting mode, the energy consumption is given by $E_{Com}(t) = E_{Comput} + E_{Awake}$. Otherwise, $E_{Com}(t) = E_{Comput} + E_{Awake} + E_{Back}$, where both $E_{Comput}$ and $E_{Awake}$ represent the energy consumed for computing the learning algorithm and waking up the circuit to decode information or harvest energy, and $E_{Back}$ is the energy used for backscattering information signals. The energy consumed for computing the algorithm can be calculated using the equation $E_{Comput} = \xi(f_{clock})^2 D_{Comp}$, where $\xi$ represents the effective capacitance coefficient of the computing chipset, $f_{clock}$ denotes the computing speed of the CPU, and $D_{Comp}$ stands for the data required for computation. Similar to [3], the maximization of energy harvesting is equivalent to maximizing the DC current, which can be expressed as:

$$\begin{aligned}z_{DC}(t) &\approx A_2 R_{ant}\mathcal{E}\{y_{B,k}^2(t)\} + A_4 R_{ant}^2\mathcal{E}\{y_{B,k}^4(t)\}\\ &= 2A_2R_{ant}\left(|h_k|^2P_A + \mathbf{1}_{(f_k==f_{J,i})}\left|h_{J,i}\right|^2P_J\right)\\ &+ A_4R_{ant}^2\left(\frac{3}{8}|h_k|^4F_A + 4\mathbf{1}_{(f_k==f_{J,i})}\left(\left|h_{J,i}\right|^4P_J^2 + 6|h_k|^2\left|h_{J,i}\right|^2P_AP_J\right)\right),\end{aligned} \tag{16}$$

where $F_A = \frac{3P_A^2}{2}$ and $A_m$, $m = \{2,4\}$, denotes the reverse bias saturation current, and $R_{ant}$ represents the impedance. If the BD decides to operate in harvesting mode, the energy harvesting is given by $E_{H,k}(t) = \chi z_{DC}(t)T$, where $\chi$ represents the RF-to-DC

conversion efficiency. From (10) and (14), the optimization problem for DB can be formulated as

$$\begin{aligned}\textbf{P2:} \max_{\pi(\rho|S_B)} \lim_{t\to\infty} \frac{1}{N_t}\sum_{\tau=1}^{N_t} E_{H,k}(\tau)\\ \text{s.t. C1: } W_q(t) + D_{next}(t) - \rho(t)D_{Tx}(t) \le 0,\\ \text{C2: } \varepsilon_{Max} \ge \mathbb{E}[E_{Com}(t)] - (1-\rho(t))\mathbb{E}[E_{H,k}(t)] - B_{Low} + \mathbb{E}[X(t)].\end{aligned} \tag{17}$$

The proposed learning algorithm aims to optimize the policy $\pi(\cdot)$, which guides the BD in selecting an operational mode $\rho^*$ to maximize long-term average energy harvesting, while adhering to constraints C1 and C2.

# 4 COOPERATIVE RL-BASED ANTI-JAMMING IN MULTI-CHANNEL WPBC NETWORKS

The cooperative RL algorithm is one of federated learning approaches, which allows both agents, namely the AP and BD, to interact within WPBC networks to learn about the jammer's behavior and optimize the operational mode in a distributed manner.

## 4.1 AP Agent

The fundamental concept of the proposed RL algorithm revolves around comprehending the behavior of the jammer across multiple channels and translating this comprehension into a learning policy. This policy guides the agent to select the optimal action based on its state and knowledge of the jammer's behavior that maximizes the reward function. With highlighting this concept, the learning parameters encompass the state, jamming knowledge, action, policy, and reward. Let $S_A \in \{CH_k|k = 1,\dots,K\}$ and $A \in \{CH_k|k = 1,\dots,K\}$ represent the state and action, respectively. These are defined as the channels selected at two consecutive time slots. $K$ is the total number of channels. For example, if the action selected at the current time slot is the $k$-th channel, the state transitions to this channel at the next time slot, and the action selection pertains to a new channel. Let $J$ represent the jamming knowledge, defined as:

$$J = \begin{cases} 1 & \text{If jammed} \\ 0 & \text{otherwise} \end{cases} \tag{18}$$

The AP selects an action $A$ based on the policy $\pi(A|J,S)$ to maximize the reward, which is determined by:

$$R_A = \begin{cases} C_A & \text{if a backscatter signal is received} \\ U^+ & \text{if not jammed but no backscatter signal is received} \\ U^- & \text{if jammed} \end{cases}, \tag{19}$$

where $C_A$ represents the objective function of problem **P1**. Based on the ergodic MDP property, the long-term average throughput is given by:

$$C_A = \lim_{t\to\infty}\sum_{\tau=1}^{N_t} \log\big(1+\Gamma_{A,k}(\tau)\big)/N_t = \mathbb{E}[\log\big(1+\Gamma_{A,k}(t)\big)] \tag{20}$$

Maximizing the reward is equivalent to maximizing the achievable rate at the AP, thereby solving problem **P1**. With full-duplex capability, the AP is able to simultaneously transmit and receive backscatter (or jamming) signals in-band. If the AP does not receive any signal (backscatter or jamming), it assumes that it can evade the jammer. Consequently, it sets its reward to a positive constant ($U^+$). However, once it receives the jamming signal, the agent imposes a penalty by setting the reward to a negative constant ($U^-$). By introducing new jamming knowledge, the Q-values are not confined to a 2D space but instead extend into a 3D space that incorporates information about channel transition and selection, as well as jamming knowledge. The AP assumes that channels with low Q-values are indicative of jamming channels, whereas those with high Q-values signify anti-jamming channels. Consequently, the AP prioritizes the selection of channels with higher Q-values. This distinction allows us to categorize the Q-values into two groups: the jamming group (or low Q group) and the anti-jamming group (or high Q group). Using TD error, the 3D Q-value can be expressed as

$$\begin{aligned}&Q_{3D}(S_A,J,A)\\ &= Q_{3D}(S_A,J,A) + \alpha\left(R_A + \gamma Q_{3D}(S'_A,J',A') - Q_{3D}(S_A,J,A)\right),\end{aligned} \tag{21}$$

where $S'_A$, $J'$, and $A'$ represent the next state, jamming knowledge, and action, respectively. The learning policy to select an action from action space is given by

$$\pi(A|J,S_A) = \begin{cases} 1+\epsilon-\frac{\epsilon}{K} & A^* = \arg\max_{A\in\mathcal{A}}(Q_{3D}(S_A,J,:)) \\ \frac{\epsilon}{K} & \text{otherwise} \end{cases}, \tag{22}$$

where $\mathcal{A} = \{CH_k|k = 1,\dots,K\}$ is the action space. The updating rule for $\epsilon$-greedy is defined as follows:

$$\epsilon = \epsilon_{\min} + (\epsilon_{\max} - \epsilon_{\min})\exp(-\lambda t), \tag{23}$$

where $\lambda$ represents the decay rate, and $\epsilon_{\min}$ and $\epsilon_{\max}$ are the minimum and maximum exploration rates, respectively. A low decay rate corresponds to more exploration, while a higher decay rate corresponds to more exploitation.

## 4.2 BD Agent

The BD lacks the capability to occupy two channels simultaneously, one for information decoding and another for jamming detection. It is aware of channel jamming only when it cannot decode information from the AP, thus switching to energy harvesting mode halfway through the time slot duration. This limitation prevents us from designing a learning algorithm for the BD to perform both tasks: information reception and jamming detection simultaneously. Nevertheless, it remains feasible to apply RL for the BD to determine when to harvest and when to communicate within the AP's channel based on the constraints outlined in problem **P2**. The BD agent tackles the optimization problem (**P2**) by transforming it into an Markov Decision Process (MDP) problem, which is defined by a tuple comprising states, actions, policies, and rewards. Here, the state and action correspond to the mode selection across two consecutive time slots. Let $S_B \in \{0,1\}$ and $\rho \in \{0,1\}$ be the state and action, respectively. For example, suppose the BD selects an action to operate in energy harvesting mode at time slot $t$, denoted as zero ($\rho = 0$). In the next time slot, it considers operating in communication mode, resulting in the next action being one ($\rho' = 1$), with the state transitioning to the previous selection, which is zero ($S_B = 0$). This selection is based on the learning policy $\pi(S_B,\rho)$. In the MDP formulation, the objective function and constraints of

**Table 1: Network Parameters**

| Parameter Settings | Values |
|---|---|
| Clock rate $f_{clock}$ | 22MHz [13] |
| Capacitance coefficient $\xi$ | $10^{-26}$ [15] |
| Wake-up power $W_{Awake}$ | $2.8\mu W$ [12] |
| Backscatter power $W_{Back}$ | $28\mu W$ [12] |
| RF-to-DC harvesting efficiency $\chi$ | 0.46 [9] |
| SIC ability $\beta$ | $10^{-9}$ [7] |
| Frame size $D_{next}$ | 100 kbps [8] |
| Lithium battery $B_{max}$ | 3v and 120 mAh [5] |
| Expected lifetime $T_{Tol}$ | 10-year [5] |
| Modulation | QPSK |

problem (**P2**) are transformed into a reward function, defined as follows:

$$R_B = C_1 C_2 O, \tag{24}$$

where $O$ denotes the objective function and $C1$ and $C2$ represent the constraints C1 and C2 in problem **P2**. According to the ergodic property, the objective function is defined as $O = \lim_{t\to\infty} \sum_{\tau=1}^{N_t} E_{H,k}(\tau)/N_t = \mathbb{E}[E_{H,k}(t)]$. Both constraints are defined as follows:

$$C_1 = \exp\left(-ReLu(Y_1)\right) \text{ and } C_2 = \exp\left(-ReLu(Y_2)\right), \tag{25}$$

where $Y_1 = D_{Tx}(t) - W_q(t) - D_{next}(t)$ and $Y_2 = \varepsilon_{Max} + (1-\rho)\mathbb{E}[E_{H,k}(t)] + \mathbb{E}[e_{M,k}(t)] + B_{Low} - \mathbb{E}[E_{Com}(t)] - \mathbb{E}[X(t)]$. With this design, $C_1 = 1$ and $C_2 = 1$ if both constraints are met; otherwise, $C_1 < 1$ and $C_2 < 1$. Thus, maximizing the reward function is equivalent to maximizing energy harvesting while adhering to the communication and energy constraints. The expected long-term reward, also known as the Q-value, can be updated based on the TD error as follows:

$$Q_{2D}(S_B, \rho) = Q_{2D}(S_B, \rho) + \alpha\left(R_B + \gamma Q_{2D}(S'_B, \rho') - Q_{2D}(S_B, \rho)\right), \tag{26}$$

where $\alpha$ and $\gamma$ are the learning rate and discount factor, respectively. Lastly, the BD updates its learning policy $\pi(S_B|\rho)$ using the $\epsilon$-greedy strategy, as outlined in Equation (22).

## 5 PERFORMANCE EVALUATION

From [6], the communication, jamming, and backscatter channels are modeled as $H = \frac{1}{2}P_{\text{Los}}\hat{h}$, where $H = \{h_k, h_{J,k}, \hat{h}_{J,k}, h_{B,k}\}$, $\hat{h} \sim \mathcal{CN}(0,1)$ represents the quasi-static Rayleigh fading, and $P_{\text{Los}}$ denotes the free-space path loss. The reference frequency $f_0$ is set to 900 MHz, and the distances from the BD to the AP, and from the jammer to the BD and AP are $d_{BA} = 3m$, $d_{JB} = d_{JA} = 3.1m$, respectively [15]. With perfect channel reciprocity, the correlation between the communication and backscatter channels is one. According to [7], the self-interference channel is represented as $h_{SI} = \sqrt{\frac{1}{K+1}}g_{SI}^{LoS} + \sqrt{\frac{K}{K+1}}g_{SI}^{NLoS}$, where $K = 0$ dB, $\left\|g_{SI}^{LoS}\right\|^2 = 1$, and $g_{SI}^{NLoS} \sim \mathcal{CN}(0,1)$. Other simulation parameters are summarized in Table 1.

The performance of the proposed learning algorithm is evaluated under three jamming strategies: static, round-robin, and random. Subsequently, it is compared to the baseline, which is randomly selecting actions. Perfectly evading jamming is achieved when the jammer strategically generates jamming signals using static and round-robin methods. With static jamming, the AP can easily learn the jamming behavior because the jamming signal is present only in one channel. The proposed algorithm can also handle dynamic jamming, such as round-robin, perfectly. This is thanks to the periodic channel jamming, which enables the algorithm to learn about its behavior and understand when the channel will be jammed again. It also helps determine if the channel is jammed at the current time slot, and predicts which channel will be jammed in the next slot. Thus, it results in achieving the same perfect performance as with static jamming. Network throughput is maximized, and communication and energy constraints are adhered to, as depicted in **Figure 2**. Surprisingly, the proposed learning algorithm not only handles static and round-robin jamming but also has the capability to evade jamming under uniform random jamming. This indicates the power of introducing a jamming state that can classify the Q-values into two groups during the exploration phase: the low Q group and the high Q group. This classification is based on the jamming knowledge and the achievable reward when transitioning from one channel to another at two consecutive time slots. During the exploitation phase, the learning policy of the algorithm guides it to focus solely on selecting channels within the high Q group, thus resulting in superior performance compared to random channel selection, as illustrated in **Figures 2, 3**, and **4**. The decay rate for exploration and exploitation is set to $\lambda = 0.001$. Lastly, the reason behind the degradation of network throughput of the AP in **Figure 4** compared to that of the BD in **Figure 3** is the reflection coefficient dropping below 1.

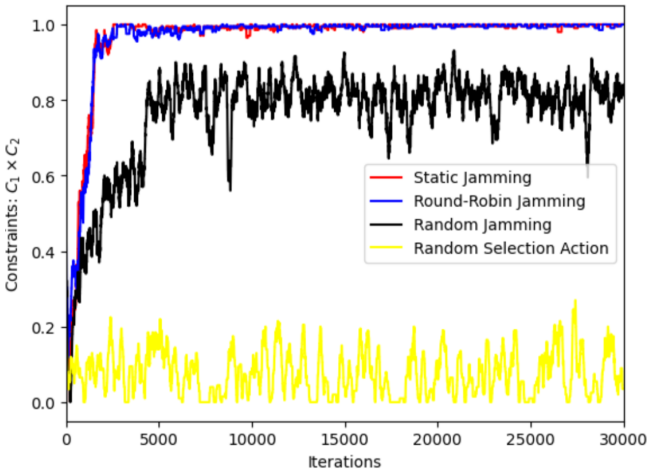


**Figure 2: Communication and energy constraints.**

## 6 CONCLUSION

Inspired by the potential of backscatter technology as a battery-less solution, our study introduces a cooperative RL approach, empowered both the AP and BD to effectively navigate and exploit jamming in multi-channel environments. Anti-jamming capabilities equip the AP to strategically select channels for communications with BD without interference, thus maximizing throughput. Conversely, jamming exploitation enables the BD to identify channels occupied

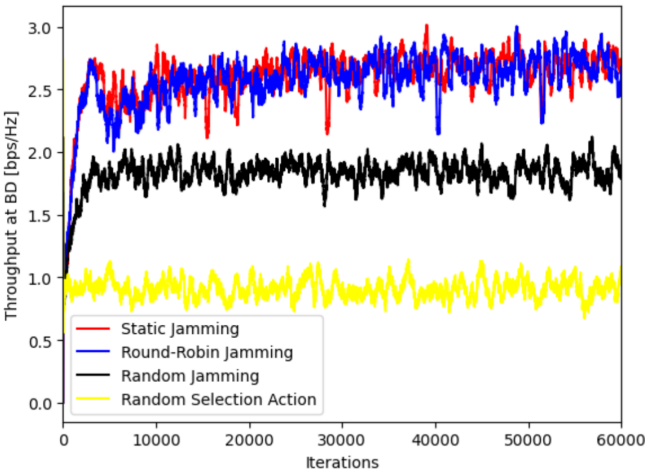


**Figure 3: Achievable throughput at BD**

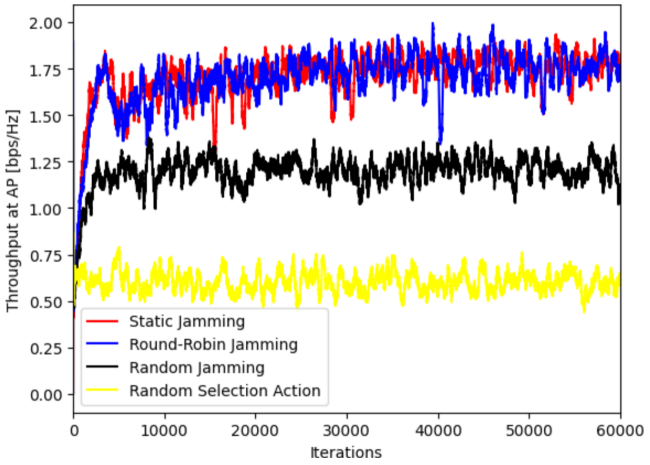


**Figure 4: Achievable throughput at AP**

by jamming signals for its energy harvesting operation. This approach further empowers the BD agent to make decisions regarding its operational mode—whether to prioritize communication or energy harvesting—to optimize the trade-off between throughput and energy harvesting efficiency. Additionally, rather than attempting to extract information from a jammed RF signal, the BD capitalizes on this opportunity for energy harvesting, thereby transforming jamming signal into a valuable energy harvesting source.

## ACKNOWLEDGMENTS

This work was supported in part the U.S. National Science Foundation (NSF) through the Networking Technology and Systems (NeTS) Program under Grant No. 2131507 and the Secure and Trustworthy Cyberspace (SaTC) Program under Grant No. 2318796, and Army Research Office (ARO) under Grant No. W911NF-21-1-0187.

## REFERENCES

[1] Hayder Al-Hraishawi, Osamah Abdullah, Symeon Chatzinotas, and Björn Ottersten. 2023. Energy Harvesting From Jamming Attacks in Multi-User Massive MIMO Networks. *IEEE Transactions on Green Communications and Networking* 7, 3 (2023), 1181–1191. https://doi.org/10.1109/TGCN.2023.3280036

[2] Kuiyu Chen, Jingyi Zhang, Si Chen, Shuning Zhang, and Huichang Zhao. 2023. Active Jamming Mitigation for Short-Range Detection System. *IEEE Transactions on Vehicular Technology* 72, 9 (2023), 11446–11457. https://doi.org/10.1109/TVT.2023.3266380

[3] Bruno Clerckx and Ekaterina Bayguzina. 2016. Waveform Design for Wireless Power Transfer. *IEEE Transactions on Signal Processing* 64, 23 (2016), 6313–6328. https://doi.org/10.1109/TSP.2016.2601284

[4] James M. Thompson Carl M. Harris Donald Gross, John F. Shortle. 2013. *Fundamentals of Queueing Theory, 4th Edition* (4th. ed.). Wiley.

[5] Marco Gonzalez, Pengcheng Xu, Rémi Dekimpe, Maxime Schramme, Ivan Stupia, Thibault Pirson, and David Bol. 2023. Technical and Ecological Limits of 2.45-GHz Wireless Power Transfer for Battery-Less Sensors. *IEEE Internet of Things Journal* 10, 17 (2023), 15431–15442. https://doi.org/10.1109/JIOT.2023.3263976

[6] Joshua D. Griffin and Gregory D. Durgin. 2009. Complete Link Budgets for Backscatter-Radio and RFID Systems. *IEEE Antennas and Propagation Magazine* 51, 2 (2009), 11–25. https://doi.org/10.1109/MAP.2009.5162013

[7] Azar Hakimi, Shayan Zargari, Chintha Tellambura, and Sanjeewa Herath. 2023. Sum Rate Maximization of MIMO Monostatic Backscatter Networks by Suppressing Residual Self-Interference. *IEEE Transactions on Communications* 71, 1 (2023), 512–526. https://doi.org/10.1109/TCOMM.2022.3223716

[8] Tao Jiang, Yu Zhang, Wenyuan Ma, Miaoran Peng, Yuxiang Peng, Mingjie Feng, and Guanghua Liu. 2023. Backscatter Communication Meets Practical Battery-Free Internet of Things: A Survey and Outlook. *IEEE Communications Surveys Tutorials* 25, 3 (2023), 2021–2051. https://doi.org/10.1109/COMST.2023.3278239

[9] Cong Ding Zhen Sun Bu-Yun Yu Lu Ju Xin-Hua Liang Zhao-Min Chen Hao Chen Yong-Hao Jia Zhen-Guo Liu Tie-Jun Cui Jun-Lin Zhan, Wei-Bing Lu. 2024. Flexible and wearable battery-free backscatter wireless communication system for colour imaging. *npj Flex Electron* 8, 19 (2024). https://doi.org/10.1038/s41528-024-00304-4

[10] Yasin Khan, Aaqib Afzal, Ankit Dubey, and Alok Saxena. 2024. Secrecy Performance of Energy-Harvesting Backscatter Communication Network Under Different Tag Selection Schemes. *IEEE Journal of Radio Frequency Identification* 8 (2024), 43–48. https://doi.org/10.1109/JRFID.2024.3371877

[11] Mingqian Liu, Zhenju Zhang, Yunfei Chen, Jianhua Ge, and Nan Zhao. 2024. Adversarial Attack and Defense on Deep Learning for Air Transportation Communication Jamming. *IEEE Transactions on Intelligent Transportation Systems* 25, 1 (2024), 973–986. https://doi.org/10.1109/TITS.2023.3262347

[12] Po-Han Peter Wang, Chi Zhang, Hongsen Yang, Manideep Dunna, Dinesh Bharadia, and Patrick P. Mercier. 2020. A Low-Power Backscatter Modulation System Communicating Across Tens of Meters With Standards-Compliant Wi-Fi Transceivers. *IEEE Journal of Solid-State Circuits* 55, 11 (2020), 2959–2969. https://doi.org/10.1109/JSSC.2020.3023956

[13] TANG Xiao-qing, CUI Yong-qiang, SHE Ya-jun, XIE Gui-hui, LIU Xin, and ZHANG Shuai. 2019. Battery-free Wi-Fi: Making Wi-Fi transmission simpler and practical. In *2019 IEEE 28th International Symposium on Industrial Electronics (ISIE)*. 1575–1582. https://doi.org/10.1109/ISIE.2019.8781331

[14] Fang Xu, Touseef Hussain, Manzoor Ahmed, Khurshed Ali, Muhammad Ayzed Mirza, Wali Ullah Khan, Asim Ihsan, and Zhu Han. 2023. The State of AI-Empowered Backscatter Communications: A Comprehensive Survey. *IEEE Internet of Things Journal* 10, 24 (2023), 21763–21786. https://doi.org/10.1109/JIOT.2023.3299210

[15] Jia Yan, Suzhi Bi, and Ying Jun Angela Zhang. 2020. Offloading and Resource Allocation With General Task Graph in Mobile Edge Computing: A Deep Reinforcement Learning Approach. *IEEE Transactions on Wireless Communications* 19, 8 (2020), 5404–5419. https://doi.org/10.1109/TWC.2020.2993071

[16] Long Zhang, Zekun Wang, Hongliang Zhang, Minghui Min, Chao Wang, Dusit Niyato, and Zhu Han. 2024. Anti-Jamming Colonel Blotto Game for Underwater Acoustic Backscatter Communication. *IEEE Transactions on Vehicular Technology* (2024), 1–15. https://doi.org/10.1109/TVT.2024.3367935