Hierarchical Learning Approach for Age-of-Information Minimization in Wireless Sensor Networks

Leiyang Cui*, Yusi Long*, Dinh Thai Hoang† and Shimin Gong*
*School of Intelligent Systems Engineering, Sun Yat-sen University, China
†School of Electrical and Data Engineering, University of Technology Sydney, Australia

***Abstract*—In this paper, we focus on a multi-user wireless network coordinated by a multi-antenna access point (AP). Each user can generate the sensing information randomly and report it to the AP. The freshness of information is measured by the age of information (AoI). We formulate the AoI minimization problem by jointly optimizing the users' scheduling and transmission control strategies. Moreover, we employ the intelligent reflecting surface (IRS) to enhance the channel conditions and thus reduce the transmission delay by controlling the AP's beamforming vector and the IRS's phase shifting matrices. The resulting AoI minimization becomes a mixed-integer program and difficult to solve due to uncertain information of the sensing data arrivals at individual users. By exploiting the problem structure, we devised a hierarchical deep reinforcement learning (DRL) framework to search for optimal solution in two iterative steps. Specifically, the users' scheduling strategy is firstly determined by the outer-loop DRL approach, and then the inner-loop optimization adapts either the uplink information transmission or downlink energy transfer to all users. Our numerical results verify that the proposed algorithm can outperform typical baselines in terms of the average AoI performance.**

***Index Terms*—AoI minimization, energy transfer, intelligent reflecting surface, deep reinforcement learning.**

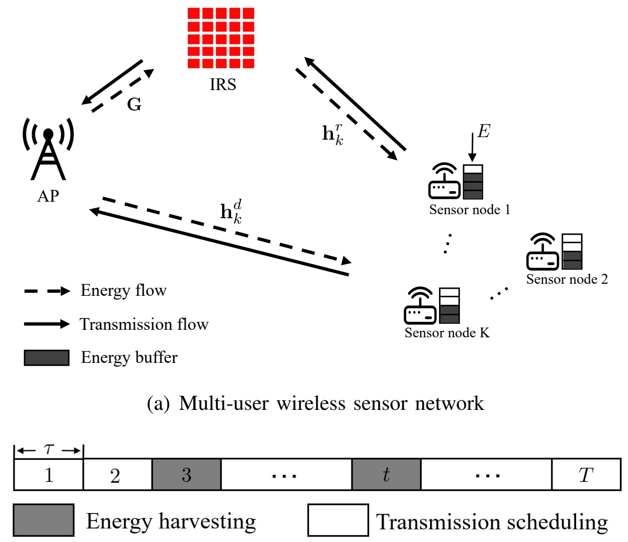
I. Introduction

With the development of the Internet of Things (IoT), a large portion of the emerging applications (e.g., autonomous driving, interactive gaming, and virtual reality) relies on timely transmission and processing of the IoT devices' sensing data of the physical world. For instance, the sensing data in autonomous driving has to be processed in real time to ensure safety. The user experience quality in interactive gaming would be degraded severely when the data transmission delay increases beyond some thresholds. Maintaining information freshness in such time-sensitive applications requires a new network performance metric, i.e., the age of information (AoI), which measures the freshness of information or overall time delay from its generation to the successful processing at the information sink, e.g., [1] and [2].

In wireless sensor network, the overall time delay of each sensor node's information update mainly includes the waiting delay or scheduling delay before information transmission and the in-the-air transmission delay. The waiting delay is usually determined by the multi-user scheduling strategy, while the transmission delay becomes worse off with limited transmit power or undesirable channel conditions, e.g., severe mutual interference or channel fading conditions. The AoI minimization problems in wireless networks were previously analyzed by the queueing theory, e.g., [3] and [4]. It was observed that the Last-Come-First-Served (LCFS) scheduling policy can achieve a smaller AoI than a few other scheduling policies [3]. The authors in [4] further proposed the Last-Generated-First-Served (LGFS) scheduling policy, which can achieve the minimum AoI in some queueing systems. The authors [5] extended the age-minimal transmission strategy from single-hop to the multi-hop scenario. Typically, the scheduling policy with exact channel information can achieve a better AoI performance comparing to that with unknown channel information [6].

However, for more complex wireless networks, e.g., with user mobility and limited resource constraints, it is still unclear how to design the multi-user scheduling policy that optimizes the overall AoI performance. Considering energy-constrained wireless sensor networks, the user scheduling becomes more challenging to balance the tradeoff between AoI and energy consumption. The authors in [7] and [8] studied wireless powered information updates by allowing the sensor nodes to harvest RF energy from the wireless environment. The RF energy harvested in each time slot is modelled as a stochastic energy arrival process. The joint optimization of multi-user scheduling and energy management is typically formulated as a high-dimensional dynamic program, due to the temporal- and spatial-interactions among different users. The AoI minimization problem is further complicated by the unknown dynamics of the channel conditions and the sensor nodes' data arrivals. Without complete system information, the scheduling strategy has to be adapted according to the users' historical AoI information. Instead of the conventional optimization approaches, the AoI minimization in a complex wireless system can be more flexibly solved by the model-free deep reinforcement learning (DRL) approaches, e.g., [9]. For example, the authors in [10] developed the DRL algorithm to jointly optimize the data offloading and scheduling polices to minimize the weighted sum of AoI and energy consumption. The authors in [11] studied AoI minimization in a device-to-device network with mutual interference. The DRL algorithm is employed to minimize the average AoI based on the users' transmission success probabilities in interfering channels.

In this paper, we aim to minimize the average AoI in a wireless powered sensor network. The intelligent reflecting surface (IRS) is also used to enhance the channel conditions and reduce the transmission delay. The IRS's passive beamforming can be jointly optimized with the sensors' scheduling policy to improve the overall AoI performance. Without complete information, we devise the DRL framework to adapt the scheduling the transmission control policies to minimize the long-term average AoI. To improve the learning efficiency, we propose a two-step hierarchical learning algorithm. The basic idea is to adapt the scheduling strategy by the outer-loop DRL algorithm, i.e., the proximal policy optimization (PPO) algorithm, and optimize the beamforming strategy by the inner-loop optimization module. Based on the AP's scheduling decision, the inner-loop routine either optimizes the AP's downlink energy transfer to all sensors or optimize individual sensor's uplink information transmission by efficient convex approximations. Our simulation results verify that the proposed learning framework can significantly reduce the average AoI compared with typical baseline strategies.



(a) Multi-user wireless sensor network

(b) Time-slotted structure for multi-user scheduling

Fig. 1. IRS-assisted and wireless-powered sensing information updates.

II. System Model

We consider an IRS-assisted wireless powered sensor network, consisting of an access point (AP) equipped with M antennas, an IRS with N reflection elements, and K single-antenna sensor nodes, as shown in Fig. 1(a). All sensor nodes are wireless powered by harvesting radio frequency (RF) energy from the AP's beamforming signals. Each RF-powered sensor node needs to report its sensing data to the AP for further processing, e.g., compression, classification, and prediction. Each sensor node can generate its sensing data irregularly depending on the status change of underlying physical process. We aim to collect all sensor nodes' data timely by scheduling their uplink data transmissions, based on their channel conditions, traffic demands, and energy status. To coordinate multiple sensor nodes' data transmissions, the transmission data frame can be divided into orthogonal time slots, as shown in Fig. 1(b). In each time slot, the uplink data transmission can be assisted by the IRS. Let $\theta_n(t) \in (0, 2\pi]$ denote the phase shift of the n-th reflection element of the IRS in the t-th time slot. Thus, we define the IRS's phase shifting vector as $\boldsymbol{\theta}(t) = [e^{j\theta_1(t)}, e^{j\theta_2(t)}, \ldots, e^{j\theta_N(t)}]$. The channel matrix from the AP to the IRS in t-th time slot is given by $\mathbf{G}(t) \in \mathbb{C}^{M\times N}$. The channel vectors from the IRS and the AP to the k-th sensor node are denoted by $\mathbf{h}_k^r(t) \in \mathbb{C}^{N\times 1}$ and $\mathbf{h}_k^d(t) \in \mathbb{C}^{M\times 1}$, respectively. All channels are assumed to be quasi-static flat-fading. The AP can estimate the channel information by a training period in the beginning of each time slot.

A. Mode Selection

As illustrated in Fig. 1(b), each data frame is equally divided into T time slots with unit length. In each time slot, we need to decide the AP's operation mode, i.e., either the downlink energy transfer or the uplink data transmission. We use the binary variable $\psi_0(t) \in \{0, 1\}$ to denote the mode selection strategy, i.e., $\psi_0(t) = 1$ indicates the AP's downlink energy transfer in time slot t and $\psi_0(t) = 0$ represents the uplink data transmission. We further use the binary variable $\psi_k(t)$ to denote the scheduling policy in the t-th time slot, i.e., $\psi_k(t) = 1$ represents that the k-th sensor node is allowed to access the channel for uploading its sensing information to the AP. To avoid interference, we require that at most one sensor node can access the channel in each time slot, which implies the following scheduling constraint:

$$\psi_0(t) + \sum_{k=1}^{K} \psi_k(t) \leq 1, \quad \forall\, t \in \mathcal{T} \text{ and } k \in \mathcal{K}. \tag{1}$$

where $\mathcal{K} = \{1, 2, \cdots, K\}$ is the set of sensor nodes. The AP can either choose to beamform RF power to all sensor nodes or receive the information update from one sensor node. We denote $\boldsymbol{\Psi}(t) = [\psi_0(t), \psi_1(t), \ldots, \psi_K(t)]$ as the scheduling policy. It is clear that the scheduling policy depends on the sensor nodes' traffic demands, channel conditions, and energy status.

B. Downlink Energy Transfer

When $\psi_0(t) = 1$, the downlink energy transfer to all sensor nodes ensures the sustainable operation of the system. With the IRS's assistance, the equivalent channel vector from the AP to the k-th sensor node can be expressed as follows:

$$\mathbf{f}_k(t) = \mathbf{h}_k^d(t) + \mathbf{G}_k(t)\boldsymbol{\Phi}(t)\mathbf{h}_k^r(t), \tag{2}$$

where $\boldsymbol{\Phi} \triangleq \text{Diag}(\boldsymbol{\theta})$ is a diagonal matrix with the diagonal element $\boldsymbol{\theta}$, denoting the IRS's phase shifting matrix. Let $\mathbf{w}(t) \in \mathbb{C}^{M\times 1}$ denote the AP's RF beamforming vector in the downlink energy transfer phase. Given the AP's transmit power p_s, the AP's beamforming signal is given by $\mathbf{x}(t) = \sqrt{p_s}\mathbf{w}(t)s_0(t)$, where $s_0(t) \in \mathbb{C}$ denotes a randomly generated complex symbol at the AP. Then, the received signal at the k-th sensor node is given as $y_k(t) = \mathbf{f}_k^H(t)\mathbf{x}(t) + n_k(t)$, where $n_k(t)$ denotes the Gaussian noise with zero mean and the normalized noise power. Considering a linear energy harvesting model, e.g., [12], [13], the energy harvested by the k-th sensor node is given by $E_k^h(t) = \eta p_s |\mathbf{f}_k^H(t)\mathbf{w}(t)|^2$, where η denotes the energy conversion efficiency. Due to the broadcast nature of wireless signals, the AP's wireless energy transfered to different sensor nodes is closely coupled via the AP's RF beamforming strategy $\mathbf{w}(t)$ and the IRS's passive beamforming strategy $\boldsymbol{\Phi}(t)$. The joint beamforming strategy $(\mathbf{w}(t), \boldsymbol{\Phi}(t))$ can be optimized to enhance the energy transfer to some sensor nodes with worse channel conditions or heavy traffic demands.

C. Sensing Information Updates

By channel reciprocity, we assume that the uplink channels are the same as the downlink channels in each time slot. Let $p_k(t)$ denote the transmit power of the k-th sensor node when it is scheduled in the t-th time slot, i.e., $\psi_k(t) = 1$. The signal received at the AP is given as follows:

$$\mathbf{y}_k = \sqrt{p_k(t)}(\mathbf{h}_k^{\rm d}(t) + \mathbf{G}_k(t)\boldsymbol{\Phi}(t)\mathbf{h}_k^{\rm r}(t))s_k + \mathbf{n}_k(t), \quad (3)$$

where $s_k(t)$ denotes the information symbol of the k-th sensor node and $\mathbf{n}_k(t)$ denotes the noise vector received by the AP. Without loss of generality, we can normalize the noise power to unit one. Hence, the received SNR at the AP can be characterized as follows:

$$\gamma_k = p_k(t)|\mathbf{w}(t)^H(\mathbf{h}_k^d(t) + \mathbf{G}_k(t)\boldsymbol{\Phi}(t)\mathbf{h}_k^r(t))|^2, \quad (4)$$

where $\mathbf{w}(t)$ represents the the AP's receiver beamforming vector. By maximum ratio combining (MRC), we can align $\mathbf{w}(t)$ to the phase of the IRS-enhanced channel. Hence, we can simplify the SNR as $\gamma_k = p_k(t)||\mathbf{h}_k^d(t) + \mathbf{G}_k(t)\boldsymbol{\Phi}(t)\mathbf{h}_k^r(t)||^2$, and thus the throughput of the k-th sensor node is given by $r_k(t) = \tau_k \log\big(1+\gamma_k\big)$, where $\tau_k \in (0,1)$ denotes the effective transmission time in one unit time slot. Given the data size d_k of the k-th sensor node, we require $r_k(t) \geq d_k$ to ensure the successful update of the sensing information.

In each time slot, the sensor node's energy consumption includes two parts, i.e., $E_k^c(t) = \tau_k(p_k(t) + p_c)$, where p_c denotes a constant operation power to maintain the node's activity. The choice of the RF transmit power $p_k(t)$ ensures the rate constraint $r_k(t) \geq d_k$, and thus it can vary with the channel conditions. We use B to denote the maximum capacity of the sensor nodes' batteries and denote $E_k(t)$ as the amount of remaining energy at the beginning of time slot t. Then, the available energy for the k-th sensor node evolves as follows:

$$E_k(t+1) = \min\Big\{\Big(E_k(t) + \psi_0(t)E_k^h - \psi_k(t)E_k^c(t)\Big)^+, B\Big\}, \quad (5)$$

Here we denote $(x)^+ \triangleq \max\{x, 0\}$ for simplicity. We require $E_k(t) \geq E_k^c(t)$ when the k-th sensor node is scheduled to update its information.

III. Age of Information Minimization

The sensing information can be regularly generated by the sensor devices, however the wireless transmission of the sensing information to the AP is limited by the channel capacity, and thus requires the coordination among different sensor nodes. It is clear that the overall information delay of each sensor node includes two parts, i.e., the caching or scheduling delay and the transmission delay over wireless channels. For each sensor node $k \in \mathcal{K}$, the caching delay depends on the AP's scheduling policy $\boldsymbol{\Psi}(t)$, while the transmission delay can be minimized by optimizing the uplink transmission control parameters, including the sensor node's transmit power $p_k(t)$, the joint active and passive beamforming strategy $(\mathbf{w}(t), \boldsymbol{\Phi}(t))$.

A. Hierarchical Learning for AoI Minimization

The AoI minimization requires the AP to adapt its scheduling strategy to collect the sensor's information timely once it is generated. Let $A_k(t)$ denote the AoI of the k-th sensor node. When the k-th sensor node is scheduled to access the channel, the AP can replace the obsolete information by the new information and thus update its AoI as $A_k(t+1) = 1$. Here we assume that the k-th sensor node can finish data transmission at the end of the time slot. If the k-th sensor node is not scheduled, its AoI will be updated as $A_k(t+1) = A_k(t) + 1$. Thus, we have the AoI dynamics as follows:

$$A_k(t+1) = \begin{cases} 1, & \text{if } o_k(t) = 1, \psi_k(t) = 1, \\ & r_k(t) \geq d_k, \text{ and } E_k(t) \geq E_k^c(t), \\ A_k(t) + 1, & \text{otherwise.} \end{cases} \quad (6)$$

Here $o_k(t) \in \{0, 1\}$ is a binary indicator, denoting the status of the caching space. When the cache is non-empty with $o_k(t) = 1$ and the sensor node is currently scheduled with $\psi_k(t) = 1$, the AP can update the sensing information from the k-th sensor node if the uplink transmission is successful, i.e., $r_k(t) \geq d_k$ and $E_k(t) \geq E_k^c(t)$ hold simultaneously.

We aim to minimize the weighted average of all users' AoI:

$$\bar{A}(\mathbf{w}, \boldsymbol{\Phi}, \boldsymbol{\Psi}) = \lim_{T\to\infty} \frac{1}{TK}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k=1}^{K} \lambda_k A_k(t)\right], \quad (7)$$

where λ_k denotes a constant weight parameter for each sensor node, indicating the delay sensitivity of different sensing information. Till this point, we can formulate the AoI minimization problem as follows:

$$\begin{aligned} \mathcal{P}(A) \quad & \min_{\mathbf{w},\boldsymbol{\Phi},\boldsymbol{\Psi}} \quad \bar{A}(\mathbf{w}, \boldsymbol{\Phi}, \boldsymbol{\Psi}) & (8a) \\ & s.t. \quad (1) \text{ and } (5) - (6), & (8b) \\ & \qquad ||\mathbf{w}||^2 \leq 1, & (8c) \\ & \qquad \theta_n \in (0, 2\pi],\, n \in \mathcal{N}. & (8d) \end{aligned}$$

where $\mathcal{N} = \{1, 2, \cdots, N\}$ is the set of IRS's reflecting elements. Depending on the AP's mode selection $\psi_0(t)$ in each time slot, the joint beamforming optimization corresponds to either the downlink energy transfer or the uplink information transmission. The first difficulty of problem (8) lies in the unknown dynamics of the sensor nodes' data arrival processes. In the ideal case, if the AP knows exactly the time when the

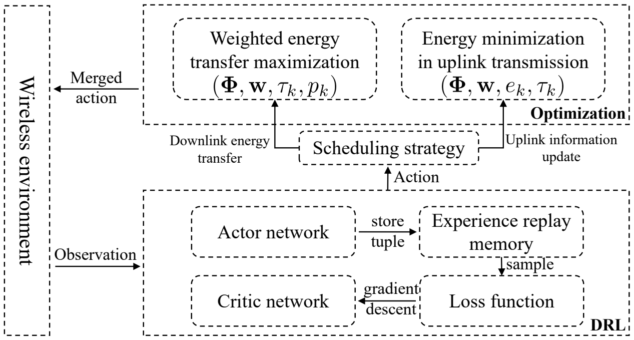


Fig. 2. The hierarchical DRL framework includes the outer-loop DRL and the inner-loop optimization methods.

sensing data arrives to the caching space, the AP can schedule each sensor node precisely to minimize the caching delay. However, without complete information, the AP has to adapt its scheduling policy based on the past observations of the sensor nodes' AoI dynamics. The second difficulty lies in the combinatorial nature of the AP's scheduling policy, resulting in a computation-demanding mix-integer dynamic programming.

To bypass these difficulties, we design a DRL agent at the AP and employ a model-free DRL method to learn the control variables $(\mathbf{w}, \mathbf{\Phi}, \mathbf{\Psi})$. However, this implies huge action space which may slow down the learning performance. Instead, we devise a hierarchical learning structure for problem (8) that decomposes the solution $(\mathbf{w}, \mathbf{\Phi}, \mathbf{\Psi})$ into two parts. The algorithm framework is shown in Fig. 2. The DRL agent firstly learns the scheduling policy $\mathbf{\Psi}(t)$ based on the past AoI performances of individual sensor nodes. Given the outer-loop scheduling decision $\mathbf{\Psi}(t)$, the inner-loop optimization of $(\mathbf{w}(t), \mathbf{\Phi}(t))$ becomes much easier by using the conventional alternating optimization (AO) and semi-definite relaxation (SDR) methods. Given the mode selection $\psi_0(t)$ in the outer loop, the AP will either maximize the downlink energy transfer to all sensor nodes or optimize the uplink information transmissions. By such a decomposition, the inner-loop optimization becomes computation-efficient, while the outer-loop learning also becomes time-efficient as it only adapts the combinatorial scheduling strategy with a smaller action space.

B. Inner-loop Optimization Problems

Given the AP's scheduling decision $\mathbf{\Psi}$, the AP either beamforms RF signals for downlink energy transfer or receives the sensing information from individual sensor node. In the sequel, we discuss the inner-loop optimization problems in two cases, respectively.

1) Weighted Energy Transfer Maximization: In downlink energy transfer with $\psi_0(t) = 1$, our intuition is to transfer more energy to those sensor nodes with worse AoI conditions. This allows them to increase their sensing frequencies and report their sensing data with a higher transmit power, and thus improve their AoI performances in the following sensing and reporting cycles. Therefore, we maximize the AoI-weighted energy transfer to all sensor nodes as follows:

$$\max_{\mathbf{w}, \mathbf{\Phi}, \tau_k, p_k} \quad \sum_{k=1}^{K} v_k(t) |\mathbf{f}_k^H(t)\mathbf{w}(t)|^2 \tag{9a}$$

$$s.t. \quad \tau_k \log(1 + p_k(t)|\mathbf{f}_k^H(t)\mathbf{w}(t)|^2) \geq d_k, \quad k \in \mathcal{K}, \tag{9b}$$

$$E_k^c(t) \leq E_k(t) + \eta |\mathbf{f}_k^H(t)\mathbf{w}(t)|^2, \quad k \in \mathcal{K}, \tag{9c}$$

$$||\mathbf{w}(t)||^2 \leq 1 \text{ and } \theta_n \in (0, 2\pi], \quad n \in \mathcal{N}, \tag{9d}$$

where $v_k(t) = A_k(t) + \alpha_k E_k^{-1}(t)$ is the weight parameter of the k-th sensor node proportional to the AoI value $A_k(t)$ while inversely proportional to the energy capacity $E_k(t)$. The constant parameter α_k represents the sensor node's sensitivity to energy shortage and AoI performance. The inequalities in (9b)-(9c) ensure that all sensor nodes have sufficient energy to upload their sensing data to the AP. It is clear that the weighted energy transfer scheme prioritizes the sensor nodes with worse energy and AoI conditions.

The optimal solution to problem (9) also relates to the transmission parameters $(\tau_k, p_k)_{k\in\mathcal{K}}$, which depend on the rate and energy budget constraints in (9b) and (9c), respectively. We can easily observe that $(\tau_k, p_k)_{k\in\mathcal{K}}$ are coupled with the joint beamforming strategy $(\mathbf{w}, \mathbf{\Phi})$ in a non-convex form in the rate constraint (9b). This difficulty can be resolved by using the AO framework that optimizes $(\tau_k, p_k)_{k\in\mathcal{K}}$ and $(\mathbf{w}, \mathbf{\Phi})$ in two steps. With fixed $(\mathbf{w}, \mathbf{\Phi})$, the equivalent channel $\mathbf{f}_k^H(t)$ can be estimated by the AP and the channel gain $|\mathbf{f}_k^H(t)\mathbf{w}(t)|^2$ becomes known to the AP. As such, the inequalities in (9b)-(9c) define a convex set for (τ_k, E_k^c), and thus we can easily find a feasible solution to (9b)-(9c).

2) Energy Minimization in Uplink Transmission: When the k-th sensor node is allowed to update its sensing information in the t-th time slot, all other sensor nodes have to wait until the next scheduling time slot. In this case, our intuition is to minimize the energy consumption $E_k^c(t) = \tau_k(p_k(t) + p_c)$ of the k-th sensor node conditioned on the successful transmission of its sensing data, i.e., $r_k(t) \geq d_k$. This will preserve more energy for its future use. Thus, we have the following energy minimization problem:

$$\min_{\tau_k, p_k, \mathbf{\Phi}} \quad \tau_k(p_k(t) + p_c) \tag{10a}$$

$$s.t. \quad \tau_k \log(1 + p_k(t)||\mathbf{f}_k^H(t)||^2) \geq d_k, \tag{10b}$$

$$\tau_k \in (0, 1) \text{ and } \theta_n \in (0, 2\pi], \quad n \in \mathcal{N}. \tag{10c}$$

It is clear that the uplink transmission in each time slot only cares about one sensor node's rate constraint in (10b). The AP's receiver beamforming $\mathbf{w}(t)$ can be simply aligned with the channel vector $\mathbf{f}_k$. Hence, we focus on the optimization of $\mathbf{\Phi}$ to enhance the IRS-assisted channel gain $||\mathbf{f}_k||^2$. This is equivalent to the maximize the channel gain:

$$\max_{\mathbf{\Phi}} \; ||\mathbf{h}_k^d(t) + \mathbf{G}_k(t)\mathbf{\Phi}(t)\mathbf{h}_k^r(t)||^2,$$

which can be easily solved by the SDR method similar to that in [14]. Besides, the transmission control parameters (τ_k, p_k)

Algorithm 1 Energy-and-Age-Aware Scheduling and Transmission Optimization Algorithm for AoI Minimization

```
Initialize: DNN weight parameters $\theta$, policy network $\pi_{\theta_{\text{old}}}$
Initialize: $t \leftarrow 0$, $E_k(0) \leftarrow B$, $A_k(t) \leftarrow 0$
 1: for Episode $= \{1, 2, \dots, \text{MAX}\}$ do
 2:     while $t \neq T$ do
 3:         Observe the system state $(\mathbf{A}(t), \mathbf{E}(t))$
 4:         Choose the outer-loop action $\mathbf{\Psi}(t)$ for scheduling
 5:         if $\psi_0(t) = 1$ for downlink energy transfer then
 6:             Optimize $(\mathbf{w}(t), \mathbf{\Phi}(t))$ in (9)
 7:         else
 8:             Optimize $(\mathbf{w}(t), \mathbf{\Phi}(t))$ and $(\tau_k, p_k)$ in (10)
 9:         end if
10:         Execute the action $\mathbf{a}(t) \triangleq (\mathbf{\Psi}(t), \mathbf{w}(t), \mathbf{\Phi}(t))$
11:         Evaluate the reward $r(\mathbf{s}(t), \mathbf{a}(t))$
12:         Buffer the transition $(\mathbf{s}(t), \mathbf{a}(t), r(t), \mathbf{s}(t+1))$
13:         $t \leftarrow t + 1$
14:     end while
15:     Take samples from the experience replay buffer
16:     Update the DNN parameters by using PPO algorithm
17: end for
```

should be also optimized jointly to minimize the sensor node's energy consumption. Let $e_k = \tau_k p_k$ denote the sensor node's energy consumption in data transmission. Given fixed $\mathbf{\Phi}$, problem (10) can be simplified as follows:

$$\min_{\tau_k \in (0,1), e_k} \quad e_k + p_c \tau_k \tag{11a}$$

$$s.t. \quad \tau_k \log\left(1 + \frac{e_k}{\tau_k} ||\mathbf{f}_k^H(t)||^2\right) \geq d_k. \tag{11b}$$

It can be verified that problem (11) is convex in (τ_k, e_k). The optimal transmit power can be easily obtained as $p_k^* = e_k^* / \tau_k^*$, given the optimal solution (τ_k^*, e_k^*) to (11).

C. Outer-loop Learning for Scheduling

The outer-loop DRL approach aims to update the AP's scheduling policy $\mathbf{\Psi}$ by continuously interacting with the uncertain network environment. To proceed, we can reformulate the scheduling optimization problem into the Markov decision process (MDP), which can be characterized by a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{R})$. The state space $\mathcal{S}$ denotes the set of all observations of the system. In each decision epoch, the AP's observation $\mathbf{s}(t) \in \mathcal{S}$ includes all sensor nodes' AoI values, denoted as a vector $\mathbf{A}(t) = [A_1(t), A_2(t), \dots, A_K(t)]$, and the energy status denoted as $\mathbf{E}(t) = [E_1(t), E_2(t), \dots, E_K(t)]$. Hence, we can define the system state as $\mathbf{s}(t) = (\mathbf{A}(t), \mathbf{E}(t))$. The action space $\mathcal{A}$ denotes the set of all feasible scheduling decisions $\mathbf{a} \in \{0, 1\}^{K+1}$ that satisfies the inequality in (1). The reward $\mathcal{R}$ defines an immediate reward value $r(\mathbf{s}(t), \mathbf{a}(t))$, which characterizes the quality of the action $\mathbf{a}(t)$ taking on the current state $\mathbf{s}(t)$. It also drives the DRL agent to improve its policy to maximize the long-term reward.

Many DRL approaches have been proposed in the literature and demonstrated to be effective for network optimization problems, e.g., the deep Q-learning and policy gradient (PG) algorithms. However, Q-learning may have some difficulties in convergence and have more applications in discrete control problems [15], [16]. The PG algorithms generally have unsatisfactory sampling efficiency, e.g., one gradient update per data sample. Comparing to traditional PG algorithms, the proximal policy optimization (PPO) has better sample complexity as it enables multiple epochs of mini-batch updates per data sample [16]. The PPO algorithm uses the off-policy learning strategy, including two copies of the deep neutral networks (DNNs) to approximate the policy network. The old policy network is used to interact with the environment and store the transition samples in the experience replay buffer. The mini-batch randomly sampled from the experience replay buffer is then used to update the weight parameters the policy network. Besides, the PPO algorithm uses importance sampling to calculate the probability ratio of two policies, and then restricts the probability ratio within $[1-\epsilon, 1+\epsilon]$ during the DNN training episodes. The parameter ϵ ensures that the data distribution in two policy networks is similar to each other.

Considering the preferable learning efficiency and robustness, we employ the PPO algorithm to adapt the user scheduling policy in the outer-loop of the hierarchical learning framework, as listed in Algorithm 1. At the initialization stage, we randomly initialize the DNN weight parameters θ for the policy network. In each learning episode, The AP collects observations $(\mathbf{A}(t), \mathbf{E}(t))$ of the system in each time slot t, and then executes an action $a(t)$ from the DRL agent according to the old policy network $\pi_{\theta_{\text{old}}}$, as shown in line 4 of Algorithm 1. Given the outer-loop scheduling decision $\mathbf{\Psi}(\mathbf{t})$, the AP needs to optimize the joint beamforming strategy $(\mathbf{w}(t), \mathbf{\Phi}(t))$ for either downlink energy transfer or uplink information tranmission, as shown in problems (9) and (10), respectively. This corresponds to lines $5-9$ of Algorithm 1. When we determine both the outer-loop and inner-loop decision variables, we will execute the decision variables $(\mathbf{\Psi}(t), \mathbf{w}(t), \mathbf{\Phi}(t))$ in the wireless system and evaluate the reward, which relates to the AoI and energy status of all sensor nodes, as shown in lines $10-12$ of Algorithm 1. The DNN training of the PPO algorithm is based on the mini-batch from the experience replay buffer, as shown in lines $15-16$ of Algorithm 1.

IV. Simulation Results

In this section, we present simulation results to verify the performance gain of Algorithm 1 for the IRS-assisted and wireless-powered wireless sensor network. The (x, y, z)-coordinates of the AP and the IRS in meters are given by $(100, 100, 0)$ and $(0, 0, 0)$, respectively. The sensor nodes are distributed randomly in a rectangular area $[5, 35] \times [-35, 35]$ in the (x, y)-plane with $z = -20$. We assume that the direct channel from the AP to each sensor node-k follows the Rayleigh fading distribution, i.e., $\mathbf{h}_k^{\text{d}}(t) = \beta_{0,k} \tilde{\mathbf{h}}_k^{\text{d}}(t)$, where $\tilde{\mathbf{h}}_k^{\text{d}}(t) \sim \mathcal{CN}(0, I)$ and $\beta_{0,k}$ denotes the path-loss of the direct channel modeled as $\beta_{0,k} = 32.6 + 36.7 \log(d_k^{\text{HS}})$, where d_k^{HS} is the distance from the AP to the sensor node-k. A similar channel model is employed in [17]. The IRS-sensor

TABLE I
PARAMETER SETTINGS

Parameters	Values
Number of AP's antennas	4
AP's transmit power p_s	30dBm
Energy conversion efficiency η	0.9
Noise power σ^2	-75dBm
Sensor nodes' data size D	5Kbits
Actor's learning rate	0.0003
Critic's learning rate	0.001
Reward discount	0.99
Number of DNN hidden layers	3
Number of neurons	64
Activation function	Tanh and Softmax
Optimizer	Adam

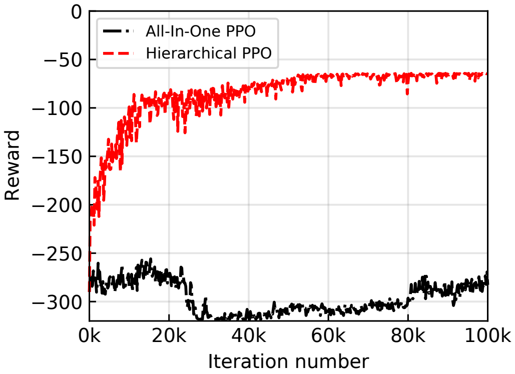


Fig. 3. Performance comparison between the hierarchical PPO and the conventional PPO algorithms.

channel $\mathbf{h}_k^{\mathrm{r}}(t)$ and the AP-IRS channel $\mathbf{G}(t)$ are modelled similarly. The other parameters and the PPO hyperparameters are summarized in Table I.

In Fig. 3, we compare the reward performance of the hierarchical learning Algorithm 1 with the conventional PPO, in which all decision variables $(\mathbf{\Psi}(t), \mathbf{w}(t), \mathbf{\Phi}(t))$ are adapted simultaneously in the PPO framework. Hence, we denote the conventional PPO as the All-in-One PPO algorithm in Fig. 3. The dashed line in red represents the dynamics of the AoI performance in the proposed hierarchical PPO algorithm, while the dash-dotted line in black denotes the conventional All-in-One PPO algorithm. It is clear that the All-in-One PPO may not converge effectively due to a huge action space in the mixed discrete and continuous domain. However, our hierarchical PPO in Algorithm 1, assisted by the inner-loop optimization modules, can reduce the action space for outer-loop reinforcement learning and thus achieve a significant performance gain as shown in Fig. 3. Considering the time complexity in inner-loop optimization, we only need to run the optimization module in the early stage of learning process, and cache the optimization results for similar network scenarios. Our intuition is that there will be numerous similar optimization scenarios between different episodes. The caching scheme can avoid unnecessary time consumption in inner-loop optimization. Our measurement results indicate that the training process in Algorithm 1 costs 4 hours 17 minutes on GPU GeForce RTX 2080TI, when the number of IRS's reflecting elements is set to 100.

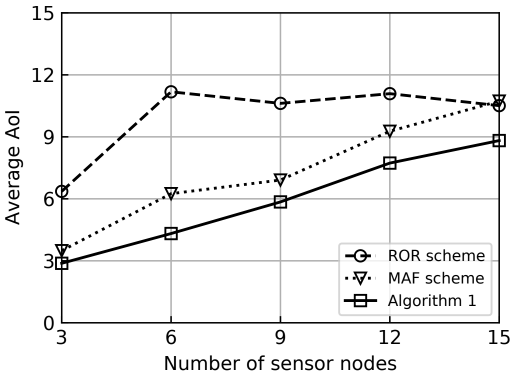


Fig. 4. Algorithm 1 achives the minimum AoI.

We also develop two baselines policies to verify the performance gain of our proposed Algorithm 1. The first baseline is the round-robin (ROR) scheduling policy which periodically selects one sensor node to upload its status-update information. In each scheduling period, we jointly optimize the active and passive beamforming strategy to enhance the information transmission. The second baseline is the Max-Age-First (MAF) scheduling policy, i.e., the AP selects the sensor node with the highest AoI value to upload its sensing information. Both baseline strategies rely on the same energy harvesting policy, i.e., the AP starts downlink energy transfer only when the scheduled sensor node has insufficient energy capacity, e.g., below some threshold value. In Fig. 4, we show the AoI performances in different algorithms with the increase in the number of sensor nodes. For different algorithms, we set the same coordinates for the AP and the IRS. The frame duration T is take as $T = 30$. Generally, different scheduling policies have a small AoI value when the number of sensor nodes is small. Besides, the MAF policy performs better than the ROR policy, as it gives higher priority to the sensor nodes with unsatisfactory AoI performance. As the number of sensor nodes increases, the proposed Algorithm 1 always outperforms the other baselines by adapting the scheduling strategy according to the stochastic data arrivals at individual sensor nodes.

In Fig. 5, we compare the fairness of different scheduling policies by showing the average AoI of different sensor nodes. For example, the average AoI of the k-th sensor node is evaluated by $\bar{A}_k = \frac{1}{T}\sum_{t=1}^{T} A_k(t)$. For fair comparison, we set the same weight parameters for all sensor nodes in (7). Obviously, it can be seen from Fig. 5 that the proposed Algorithm 1 achieves a smaller average AoI value for each sensor node. Moreover, different sensor nodes can achieve very similar AoI values, which implies a better fairness in the scheduling policy of Algorithm 1. An interesting observation is that the MAF scheduling policy also has a small deviation of AoI values among different sensor nodes, compared to that

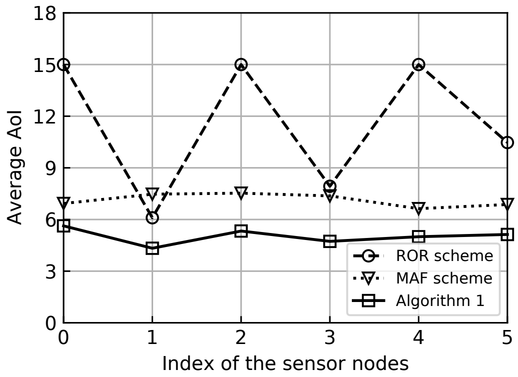


Fig. 5. Algorithm 1 enhances the scheduling fairness among sensor nodes.

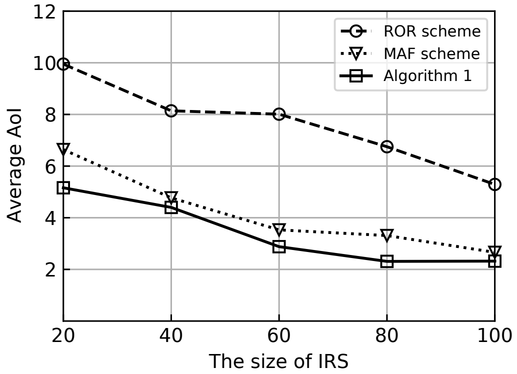


Fig. 6. Larger-size IRS improves the AoI performance.

of the ROR scheduling policy. This implies that the energy-aware MAF policy will be more efficient than the ROR policy, given the knowledge of the sensor nodes' energy status.

In Fig. 6, we show the dynamics of the average AoI by gradually increasing the size of IRS from 20 to 100. The increase in the number of the IRS's reflecting elements makes it more flexible to reshape the wireless channels and therefore improves the uplink transmission success probability. This can implicitly minimize the sensor nodes' AoI values in a long run. However, the AoI values will not keep decreasing as the size of IRS increases, as shown in Fig. 6. This implies that the AoI performance will not be affected by the transmission success probability or transmission delay when the size of IRS becomes larger enough. Instead, the AoI performance with a large-size IRS will be closely related to the AP's scheduling delay.

V. Conclusions

In this paper, we have focused on an IRS-assisted and wireless-powered sensor network and aimed to minimize the overall age-of-information (AoI) for sensing information updates. We have formulated the AoI minimization problem using a mixed-integer program and devised a novel hierarchical learning framework, which includes the outer-loop model-free learning algorithm and the inner-loop optimization methods. The interactions between the outer- and inner-loops can greatly improve the stability and learning efficiency. Simulation results demonstrate that our algorithm can significantly reduce the AoI and achieve fairness among different sensor nodes.

References

[1] S. K. Kaul, M. Gruteser, V. Rai, and J. B. Kenney, "Minimizing age of information in vehicular networks," in *Proc. SECON*, Jun. 2011, pp. 350–358.

[2] S. K. Kaul, R. D. Yates, and M. Gruteser, "Real-time status: How often should one update?" in *Proc. IEEE INFOCOM*, Mar. 2012, pp. 2731–2735.

[3] N. Pappas, J. Gunnarsson, L. Kratz, M. Kountouris, and V. Angelakis, "Age of information of multiple sources with queue management," in *Proc. IEEE ICC*, Jun. 2015, pp. 5935–5940.

[4] A. M. Bedewy, Y. Sun, and N. B. Shroff, "Optimizing data freshness, throughput, and delay in multi-server information-update systems," in *Proc. IEEE ISIT*, Jul. 2016, pp. 2569–2573.

[5] A. Arafa and S. Ulukus, "Age-minimal transmission in energy harvesting two-hop networks," in *Proc. IEEE GLOBECOM*, Dec. 2017, pp. 1–6.

[6] R. Talak, S. Karaman, and E. Modiano, "Optimizing age of information in wireless networks with perfect channel state information," in *Proc. IEEE WiOpt*, May 2018, pp. 1–8.

[7] I. Krikidis, "Average age of information in wireless powered sensor networks," *IEEE Wireless Commun. Lett.*, vol. 8, no. 2, pp. 628–631, Apr. 2019.

[8] O. M. Sleem, S. Leng, and A. Yener, "Age of information minimization in wireless powered stochastic energy harvesting networks," in *2020 54th Annual Conference on Information Sciences and Systems (CISS)*, Mar. 2020, pp. 1–6.

[9] S. Leng and A. Yener, "An actor-critic reinforcement learning approach to minimum age of information scheduling in energy harvesting networks," in *Proc. IEEE ICASSP*, Jun. 2021, pp. 8128–8132.

[10] X. Xie, H. Wang, and M. Weng, "A reinforcement learning approach for optimizing the age of computing enabled IoT," *IEEE Internet Things J.*, pp. 1–1, Jun. 2021.

[11] Z. Liu, Z. Chen, L. Luo, M. Hua, W. Li, and B. Xia, "Age of information-based scheduling for wireless device-to-device communications using deep learning," in *Proc. IEEE WCNC*, Jun. 2021, pp. 1–6.

[12] I. Krikidis, "Average age of information in wireless powered sensor networks," *IEEE Wireless Commun. Lett.*, vol. 8, no. 2, pp. 628–631, Apr. 2019.

[13] J. Yao and N. Ansari, "Wireless power and energy harvesting control in IoD by deep reinforcement learning," *IEEE Trans. Green Commun. Netw.*, vol. 5, no. 2, pp. 980–989, Jun. 2021.

[14] B. Lyu, D. T. Hoang, S. Gong, D. Niyato, and D. I. Kim, "IRS-based wireless jamming attacks: When jammers can attack without power," *IEEE Wireless Commun. Lett.*, vol. 9, no. 10, pp. 1663–1667, Oct. 2020.

[15] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "Openai gym," 2016. [Online]. Available: http://arxiv.org/abs/1606.01540

[16] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017. [Online]. Available: http://arxiv.org/abs/1707.06347

[17] T. Jiang, H. V. Cheng, and W. Yu, "Learning to reflect and to beamform for intelligent reflecting surface with implicit channel estimation," *IEEE J. Sel. Area. Commun.*, vol. 39, no. 7, pp. 1931–1945, Jul. 2021.