Sumedh Dongare, Aleksandar Jovovic, Wanja de Sombre, Andrea Ortiz, and Anja Klein
”Minimizing the Age of Incorrect Information for Status Update Systems with Energy Harvesting”, in *IEEE International Conference on Communications (ICC), Denver, USA*, June 2024.

Minimizing the Age of Incorrect Information for Status Update Systems with Energy Harvesting

Sumedh Dongare, Aleksandar Jovovic, Wanja de Sombre, Andrea Ortiz, Anja Klein
Communications Engineering Lab, Technical University of Darmstadt, Germany.
{s.dongare, w.sombre, a.ortiz, a.klein}@nt.tu-darmstadt.de, aleksandar.jovovic@stud.tu-darmstadt.de

***Abstract*—Status Update Systems (SUSs) are central components in applications like environmental sensing or smart cities. They consist of a sender monitoring a remote process and sending the sensed information to a receiver. The sender aims to deliver fresh information about the monitored process's state to allow the receiver to timely respond to the process's changes. In SUSs, the sender is usually battery operated. Therefore, to increase the available energy we consider Energy Harvesting (EH). Moreover, as at the receiver the information transmitted by the sender is only relevant when the process's state changes, we measure the information's freshness using Age of Incorrect Information (AoII). Finding the optimal transmission strategy at the sender that minimizes the AoII requires perfect system knowledge, i.e., the behavior of the monitored process, the channel quality, and the available energy. However, in real applications this knowledge is usually not available. To overcome this challenge, we first establish the optimality of threshold-based policies for AoII minimization in SUSs with EH capabilities by proving that there exists an AoII value depending on the observed state of the monitored process, the battery level and the receiver's estimation of the monitored process's state beyond which transmitting is preferable over idling. Next, we exploit the threshold-based policies' structure and deploy a learning algorithm based on Finite-Difference Policy Gradient (FDPG). Our proposed approach finds the AoII thresholds without requiring perfect system knowledge. Simulations show that our approach outperforms reference algorithms by at least 20% and efficiently learns near-optimal policies for AoII minimization.**

I. Introduction

A Status Update System (SUS) consists of a sender and a receiver. The sender monitors a process and communicates the sensed information to the receiver. SUSs are useful in numerous applications in the Internet of Things (IoT), such as environmental monitoring [1], [2], industrial IoT [3], and in smart cities [4] including traffic monitoring [5]. Due to such wide variety of applications, SUSs have drawn significant interest from industry as well as from academia.

To enable fast reactions to changes in the monitored process, SUSs demand fresh information at the receiver. To measure the freshness of this information, the Age of Information (AoI) metric was proposed in [6]. The AoI indicates the time elapsed since the sensing of the most recent successfully received status update. The drawback of the AoI is that it does not consider the content of the available information at the receiver. In cases where the monitored process remains constant, or if the monitored process returns to a previous state, the AoI falls short as it triggers unnecessary transmissions of irrelevant information. This is particularly undesirable when the sender is battery limited. Although many works [7]–[9] consider Energy Harvesting (EH) capabilities in SUS to increase the available energy, this does not reduce the number of unnecessary transmissions. Furthermore, EH brings the additional challenge that the amount of energy harvested by the sender is not constant and varies over time.

A recent communication paradigm called semantic communication focuses on transmitting only the relevant information to improve resource utilization [10]–[12]. This idea can be transferred to SUSs by measuring the Age of Incorrect Information (AoII) instead of the AoI. The AoII as introduced in [10] measures the time elapsed since the receiver last had correct information about the monitored process. This allows the sender to transmit only in cases where the information at the receiver is incorrect, thus reducing the number of unnecessary transmissions as compared to the AoI. Note that the relevance of the information which the sender transmits highly depends on the state transitions of the monitored process, i.e., how the state of the process changes.

To make optimal decisions about transmissions, the sender needs perfect knowledge about the state transitions of the monitored process, the channel quality between the sender and the receiver, and the amounts of energy harvested by the sender. However, the availability of this knowledge at the sender is unrealistic to assume. In the literature, many works [7], [9], [13] consider different transmission strategies to ensure freshness of the information in SUS. In [7], the optimality of threshold-based policies to minimize the AoI is shown. In [13], assuming a constant power source, the authors prove that the optimal transmission policy to minimize AoII uses a threshold approach. This means that the sender attempts to transmit once a certain AoII value is reached. Their approach, however, relies on perfect causal knowledge. The authors of [9] propose a threshold-based solution to minimize AoII for an EH sender. However, the authors use a single AoII threshold irrespective of the battery state of the sender, which is in general suboptimal. Moreover, the works considering AoII [9], [13] make the simplifying assumption

This work has been funded by the German Research Foundation (DFG) as a part of the project C1 within the Collaborative Research Center (CRC) 1053 - MAKI (Nr. 210487104) and has been supported by the BMBF project Open6GHub under grant 16KISKO14, by DAAD with funds from the German Federal Ministry of Education and Research (BMBF) and by the LOEWE Center emergenCity.

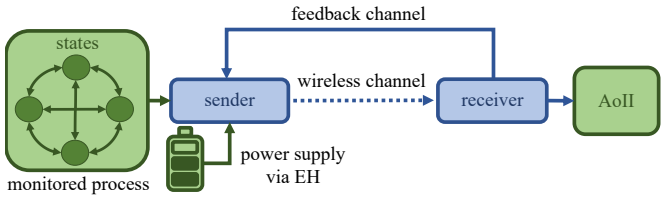

Fig. 1: System Model

that the monitored process has equiprobable transitions to any different state. This is unrealistic in real-world applications. Thus, the design of transmission strategies to minimize the AoII in EH SUS when perfect knowledge about the state transitions of the monitored process, channel quality, and the amounts of energy harvested by the sender, is not available is still an open research question.

In this work, we design transmission strategies for an EH sender in SUSs which uses the AoII to evaluate the freshness of the information. We prove that the optimal transmission policy which minimizes the AoII is threshold-based, i.e., for each battery level of the sender and each state of the monitored process, there exists an AoII value above which it is always optimal to transmit. However, the sender requires perfect knowledge about the SUS to evaluate the optimal transmission policy, which is an unrealistic assumption. In addition to this, the computational complexity of the optimal policy is directly dependent on the number of states in the monitored process, and the battery levels of the sender. To solve this, we propose a Reinforcement Learning (RL) solution based on Finite-Difference Policy Gradient (FDPG) algorithm which does not rely on perfect knowledge about the SUS and is computationally feasible. Moreover, FDPG exploits the threshold-based structure of the optimal transmission policy and learns different transmission policies depending on different battery levels and states of the monitored process. In contrast to other approaches in RL, FDPG optimizes the transmission policy directly based on the AoII thresholds. Through numerical simulations, we demonstrate that our approach performs close to the optimum determined through Value Iteration (VI). Additionally, we showcase the superior performance of our approach in minimizing AoII compared to baseline methods.

The rest of the paper is organized as follows: Section II presents the system model. Section III details the problem formulation aiming at minimizing the AoII. The proof of the threshold-based policy's optimality is provided in Section IV. Our novel approach employing RL is elaborated in Section V. Section VI provides extensive simulation results to validate the effectiveness of the proposed approach. Finally, conclusions are drawn in Section VII.

II. System Model

We consider a time-slotted SUS consisting of a sender and a receiver. Time is divided into discrete time steps with index $t \in \mathbb{N}_0$. Each time step has the same duration, which is assumed to be long enough to sense and transmit new information about the monitored process. The set of possible states of the monitored process is $\mathcal{X}$ with $|\mathcal{X}| = N$. In each time step t, the sender senses the current state $X_t \in \mathcal{X}$ of the monitored process. The monitored process is modeled as a Markov chain. This means that the transition to a new state only depends on the current state.

The sender decides, in each time step, whether to transmit or not. This decision is denoted by A_t, where $A_t = 0$ means that the sender idles and $A_t = 1$ means that the sender transmits. Status updates are transmitted via a wireless channel, which we model as a packet erasure channel. If the sender decides to transmit, the status update is correctly detected at the receiver with a probability of p_c. We use the Automatic Repeat Request (ARQ) protocol to inform the sender about whether a transmission was successful. This means that the receiver sends an error- and latency free feedback to the sender. If the receiver successfully decoded the transmitted information, $K_t = 1$ (ACK), otherwise $K_t = 0$ (NACK). The receiver updates its estimate $\hat{X}_t$ of the current state X_t:

$$\hat{X}_{t+1} := \begin{cases} X_{t+1}, & \text{if } K_t = 1, \\ \hat{X}_t, & \text{otherwise.} \end{cases} \tag{1}$$

Using the feedback K_t, the sender keeps track of the receiver's estimation $\hat{X}_t$ of X_t. The sender also keeps track of the AoII at the receiver at time $t+1$, termed Δ_{t+1}, which is then defined recursively by:

$$\Delta_{t+1} := \begin{cases} 0, & \text{if } X_{t+1} = \hat{X}_{t+1}, \\ \min(\Delta_t + 1, M), & \text{otherwise,} \end{cases} \tag{2}$$

where we set $\Delta_1 := 0$. We assume $M \in \mathbb{N}$ is the maximum allowable AoII. Above this limit, the estimate of the current state of the monitored process at the receiver is assumed to be too high for the receiver to make appropriate decisions. We further define the set of all allowable AoIIs at the receiver as $\mathcal{D} := \{0, 1, \ldots, M\}$.

The sender is powered through harvested energy. To store the harvested energy, the sender is equipped with a rechargeable battery. This battery has a capacity of $B_{\max} \in \mathbb{N}$, such that in each time step t, the current battery level b_t is from the set of possible battery levels $\mathcal{B} := \{0, 1, \ldots, B_{\max}\}$. At the end of each time step, the battery is recharged and a harvested amount $e_t \in \mathcal{E} = \{0, ..., E_{\max}\}$ of energy is added to the battery level. In real world scenarios, the amount of harvested energy is often time-correlated. For this reason, we assume that e_t depends on the previously harvested energy e_{t-1}. The probability that the harvested amount of energy in time step t is e_t is defined as:

$$p(e_t|e_{t-1}) := Pr(e_t|e_{t-1}) > 0. \tag{3}$$

For transmission ($A_t = 1$) the sender consumes an amount $E^{\text{tx}} \in \mathbb{N}$ of energy. We assume that E^{tx} remains constant over time. For simplicity, we omit the energy consumed by the sender when sensing and during idle mode. Constants modeling both could be easily added to the model. Consequently, the

battery level is not reduced if $A_t = 0$. With this, the battery in time step $t+1$ is calculated from the previous time step as

$$b_{t+1} = \min\left(b_t + e_t - E^{\text{tx}}\mathbb{1}[A_t = 1], B_{\max}\right). \quad (4)$$

Here, the $\mathbb{1}[A_t = 1]$ is an indicator function with a value of 1 if $A_t = 1$ and 0 otherwise. The battery level of the sender cannot be negative, i.e., the sender cannot choose an action $A_t = 1$ if the battery status b_t is lower than E^{tx}.

III. Problem Formulation

We formulate the AoII-minimization problem as a Markov Decision Process (MDP) $\mathcal{M} := (\mathcal{S}, \mathcal{A}, \mathcal{P}, c)$, where

- $\mathcal{S} := \mathcal{E} \times \mathcal{B} \times \mathcal{D} \times \mathcal{X} \times \mathcal{X}$ is the set of states at the sender,
- $\mathcal{A} := \{0, 1\}$ is the set of actions the sender can choose,
- $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the transition kernel and
- $c : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is the cost function.

The state $s_t = (e_t, b_t, \Delta_t, X_t, \hat{X}_t) \in \mathcal{S}$ in time step t consists of the amount of harvested energy e_t, the battery level b_t, the AoII Δ_t, the current state X_t of the monitored process at the sender, and the current estimate $\hat{X}_t$ of the monitored process at the receiver. For all $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$, the transition kernel $\mathcal{P}$ is defined by

$$\mathcal{P}(s, a, s') := Pr(s_{t+1} = s' | s_t = s, a_t = a). \quad (5)$$

Given all constants described in Sec. II, $Pr(s_{t+1} = s' | s_t = s, a_t = a)$ can be derived from Eq. (1), (2) and (4). The cost function c is defined by

$$c(s, a, s') := \Delta', \quad (6)$$

where Δ' is the AoII of state s'.

We aim to find a policy $\pi^* : \mathcal{S} \to \mathcal{A}$, which minimizes the average expected AoII.

The corresponding optimization problem related to the MDP $\mathcal{M}$ is given by:

$$J^* := \min_{\pi:\mathcal{S}\to\mathcal{A}} \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[c(s_t, \pi(s_t), s_{t+1})], \quad (7)$$

where the minimal average AoII J^* is achieved under the optimal policy π^*.

IV. Structure of the Optimal Policy

To solve the described problem, we first prove that the optimal transmission policy is threshold-based with respect to the AoII. Here, we roughly follow the structure of the proof of Theorem 1 in [7]. We first define the state action cost function Q. Intuitively, $Q(s, a)$ expresses the discounted expected future cost for action a in state s.

Definition 1. *The state action cost function Q is defined as*

$$Q(s, a) := \Delta + \mathbb{E}[h(s')|a, s], \quad (8)$$

where $s' = (e', b', \Delta', X', \hat{X}')$ is the successor of $s = (e, b, \Delta, X, \hat{X})$ and where the bias h satisfies

$$h(s) = \min_{a\in\mathcal{A}}(\Delta + \mathbb{E}[h(s')|s, a]) - J^*. \quad (9)$$

Theorem 2. *The policy π, minimizing the state action cost function Q is of threshold-type with respect to the AoII Δ.*

Proof. To show the threshold-structure of π, it is sufficient to show that the difference $D := Q(s, 0) - Q(s, 1)$ between the cost $Q(s, 0)$ of idling ($a = 0$) in a state $s = (e, b, \Delta, X, \hat{X})$ and the cost $Q(s, 1)$ of transmitting ($a = 1$) in the same state s is monotonically increasing with the AoII Δ. If $D < 0$, $a = 0$ is the optimal action, while if $D > 0$, $a = 1$ is the optimal action. If D is monotonically increasing in Δ, then there is a threshold AoII value from which it is always better to transmit, because from this threshold on, D remains positive.

π chooses the action minimizing $Q(s, a)$. If hence the difference $Q(s, 0) - Q(s, 1)$ increases with increasing Δ, π is of threshold type. To show this monotonicity, we compare the difference $Q(s, 0) - Q(s, 1)$ for the state s and a state $\check{s} = (e, b, \Delta + 1, X, \hat{X})$. $\check{s}$ differs from s only in the AoII-component, while all other components remain fixed.

Using $\check{s}$ and the definition of Q, we find the following:

$$\begin{aligned}
&Q(s, 0) - Q(s, 1) \le Q(\check{s}, 0) - Q(\check{s}, 1)\\
\Leftrightarrow&\mathbb{E}[h(s')|s, 0] - \mathbb{E}[h(s')|s, 1] \le \mathbb{E}[h(s')|\check{s}, 0] - \mathbb{E}[h(s')|\check{s}, 1]\\
\Leftrightarrow&\mathbb{E}[h(s')|s, 0] - \mathbb{E}[h(s')|\check{s}, 0] \le \mathbb{E}[h(s')|s, 1] - \mathbb{E}[h(s')|\check{s}, 1]\\
\Leftrightarrow&h((e', b', \Delta + 1, X', \hat{X}')) - h((e', b', \Delta + 2, X', \hat{X}'))\\
&\le (1 - p_c) \cdot (h((e', b', \Delta + 1, X', \hat{X}'))\\
&\qquad - h((e', b', \Delta + 2, X', \hat{X}')))\\
&\quad + p_c(h((e', b', 0, X', \hat{X}')) - h((e', b', 0, X', \hat{X}')))\\
\Leftrightarrow&h((e', b', \Delta + 1, X', \hat{X}')) - h((e', b', \Delta + 2, X', \hat{X}')) \le 0\\
\Leftrightarrow&h((e', b', \Delta + 1, X', \hat{X}')) \le h((e', b', \Delta + 2, X', \hat{X}'))\\
\Leftrightarrow&h \text{ is monotonically increasing in } \Delta.
\end{aligned}$$

Thus, if we show the monotonicity of h, we can deduce the monotonicity of the difference $Q(s, 0) - Q(s, 1)$ in Δ.

The monotonicity of h can be proven using induction. As a starting point, we use $s_{M-1} = (e, b, M - 1, X, \hat{X})$ and $s_M = (e, b, M, X, \hat{X})$. A priori, we do not know which actions a_{M-1} and a_M are optimizing the state action cost function Q for s_{M-1} and s_M. This results in four possible cases, namely

- $a_{M-1} = 0$ and $a_M = 0$:

$$\begin{aligned}
&h(s_M) - h(s_{M-1})\\
&= 1 + \mathbb{E}[h(s')|s_M, 0] - \mathbb{E}[h(s')|s_{M-1}, 0] \ge 0
\end{aligned}$$

- $a_{M-1} = 0$ and $a_M = 1$: Due to optimality of a_{M-1},

$$\begin{aligned}
&h(s_M) - h(s_{M-1})\\
&= 1 + \mathbb{E}[h(s')|s_M, 1] - \mathbb{E}[h(s')|s_{M-1}, 0]\\
&\ge 1 + \mathbb{E}[h(s')|s_M, 1] - \mathbb{E}[h(s')|s_{M-1}, 1] \ge 0
\end{aligned}$$

- $a_{M-1} = 1$ and $a_M = 1$:

$$\begin{aligned}
&h(s_M) - h(s_{M-1})\\
&= 1 + \mathbb{E}[h(s')|s_M, 1] - \mathbb{E}[h(s')|s_{M-1}, 1] \ge 0
\end{aligned}$$

- $a_{M-1} = 1$ and $a_M = 0$: Due to optimality of a_{M-1},

$$\begin{aligned} &h(s_M) - h(s_{M-1}) \\ &= 1 + \mathbb{E}[h(s')|s_M, 0] - \mathbb{E}[h(s')|s_{M-1}, 1] \\ &\geq 1 + \mathbb{E}[h(s')|s_M, 0] - \mathbb{E}[h(s')|s_{M-1}, 0] \geq 0 \end{aligned}$$

It remains to show the induction step. This time, we use $s = (e, b, \Delta, X, \hat{X})$ and $\check{s} = (e, b, \Delta + 1, X, \hat{X})$ and the respective optimal actions a and a_+. There are again four possible cases:

- $a = 0$ and $a' = 0$: Using the induction hypothesis,

$$\begin{aligned} h(s') - h(s) &= 1 + \mathbb{E}[h(s')|s', 0] - \mathbb{E}[h(s')|s, 0] \\ &= 1 + h((e', b', \Delta + 2, X', \hat{X}')) \\ &\quad - h((e', b', \Delta + 1, X', \hat{X}')) \geq 0. \end{aligned}$$

- $a = 1$ and $a' = 1$: Using the induction hypothesis,

$$\begin{aligned} h(s') - h(s) &= 1 + \mathbb{E}[h(s')|s', 1] - \mathbb{E}[h(s')|s, 1] \\ &= 1 + (1 - p_c)(h((e', b', \Delta + 2, X', \hat{X}')) \\ &\quad - h((e', b', \Delta + 1, X', \hat{X}'))) \\ &\quad + p_c(h((e', b', 0, X', \hat{X}')) \\ &\quad - h((e', b', 0, X', \hat{X}'))) \\ &= 1 + (1 - p_c)(h((e', b', \Delta + 2, X', \hat{X}')) \\ &\quad - h((e', b', \Delta + 1, X', \hat{X}'))) \geq 0. \end{aligned}$$

- $a = 0$ and $a' = 1$: Due to optimality of a,

$$\begin{aligned} h(s') - h(s) &= 1 + \mathbb{E}[h(s')|s', 1] - \mathbb{E}[h(s')|s, 0] \\ &\geq 1 + \mathbb{E}[h(s')|s', 1] - \mathbb{E}[h(s')|s, 1] \geq 0. \end{aligned}$$

- $a = 1$ and $a' = 0$: Due to optimality of a,

$$\begin{aligned} h(s') - h(s) &= 1 + \mathbb{E}[h(s')|s', 0] - \mathbb{E}[h(s')|s, 1] \\ &\geq 1 + \mathbb{E}[h(s')|s', 0] - \mathbb{E}[h(s')|s, 0] \geq 0. \end{aligned}$$

□

V. Proposed solution

In this section, we present our proposed solution based on the RL algorithm FDPG and the proof of the optimal policy's structure in Sec. IV. By using RL, we do not rely on perfect knowledge of the state transitions of the monitored process, the channel quality, and the amounts of harvested energy. FDPG exploits the threshold structure of the optimal policy by directly optimizing the AoII thresholds.

FDPG is a policy based RL algorithm. This means that the learned policy π_θ is directly parameterized by parameters θ, which are optimized during learning.

For our problem, and based on Theorem 2, we use the AoII threshold values as parameters θ for the policy. The policy uses a separate AoII threshold $\theta(e, b, X, \hat{X})$ for each tuple $e \in \mathcal{E}, b \in \mathcal{B}, X, \hat{X} \in \mathcal{X}$. This threshold $\theta(e, b, X, \hat{X})$ determines, whether the sender should remain idle or transmit. However, as it is generally more efficient for FDPG to use stochastic policies during learning [14], the policy's decision is not deterministic. Instead, we use a commonly used parameterized sigmoid function and define the probability to transmit

$$Pr(\pi_\theta(s) = 1) := \frac{1}{1 + e^{-\frac{\Delta - \theta(e,b,X,\hat{X})}{\tau}}}, \tag{10}$$

with τ as a so called temperature parameter. By reducing τ to 0, we ensure that after a sufficient number of steps, the stochastic policy converges to a deterministic policy.

In the remaining part of this section, we describe, how the parameters θ are learned using Alg. 1. We first initialize the temperature parameter as τ_0 and a corresponding decay factor ς. To initialize $\bar{\theta}_0$, we set all thresholds to 0, such that initially, the sender will transmit for all battery levels and states of the monitored process (line 1). Considering the energy constraint, we then set the thresholds for states where $b < E^{\text{tx}}$ to a value larger than the maximum allowable AoII M. This initial policy assumes that the sender always attempts to transmit information if the battery status is sufficiently high. Starting from the initial parameters $\bar{\theta}_0$, Alg. 1 explores the parameter space by repeatedly generating a random perturbation vector $\mathbf{d}_n$ for iterations numbered $n \in \{1, ..., n_{max}\}$. $\mathbf{d}_n$ contains 0 or 1 with equal probability in each position (line 3). The current parameters $\bar{\theta}_n$ are perturbed in both directions by adding $\pm\beta_n \cdot \mathbf{d}_n$, where β_n is a decaying step size parameter (line 4). This results in parameters $\theta_n^\pm$ from which we derive the strategies $\pi_{\theta_n^\pm}$ according to Eq. (10). Using the strategies $\pi_{\theta_n^\pm}$, we estimate,

$$J^\pm := \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[c(s_t, \pi(s_t), s_{t+1})], \tag{11}$$

by running a simulation for $T \in \mathbb{N}$ time steps (lines 5-9). The resulting estimates $\hat{J}_n^\pm$ are then used to compute the gradient in line 10:

$$\frac{\partial \hat{J}}{\partial \bar{\theta}_n} := (\mathbf{d}_n^\intercal \mathbf{d}_n)^{-1} \mathbf{d}_n^\intercal \frac{\hat{J}^+ - \hat{J}^-}{2\beta_n}. \tag{12}$$

Using this gradient, we update the parameters $\bar{\theta}_{n+1}$ in line 11 according to

$$\bar{\theta}_{n+1} := \bar{\theta}_n - \gamma_n \frac{\partial \hat{J}}{\partial \bar{\theta}}, \tag{13}$$

where γ_n represents a learning rate that converges to zero. Finally, the temperature τ is updated (line 12). This process (lines 2-13) is repeated until iteration $n = n_{max}$ is reached.

VI. Numerical Results

In this section, we present and discuss the numerical simulations used to evaluate the performance of our proposed approach in comparison with the reference schemes. We first establish the simulation setup. The results are generated and averaged over 600 independent runs of the respective simulation. In each of these realizations, we consider a time horizon of 2×10^6 time steps. We set a limit $M = 40$ for the maximum allowable AoII. The probability of successful transmission through the channel is $p_c = 0.9$.

The states of the monitored process are modelled as integers $\mathcal{X} = \{1, ..., N\}$. As described in Section I, we consider a

Algorithm 1: FDPG

Initialize: $\tau_0, \varsigma, \bar{\theta}_0$
for $n = \{1, 2, ..., n_{max}\}$ **do**
  Generate a random perturbation vector $\mathbf{d}_n$
  Perturb the parameters $\bar{\theta}_n$
    $\bar{\theta}_n^+ = \bar{\theta}_n + \beta_n \cdot \mathbf{d}_n,\ \bar{\theta}_n^- = \bar{\theta}_n - \beta_n \cdot \mathbf{d}_n$
  Estimate $\hat{J}_n^{\pm}$ from simulations of the MDP using policies $\pi_{\theta\pm}$:
  for $t \in \{1, 2, ..., T\}$ **do**
    Observe current state s_t and USE policy $\pi_{\theta\pm}$
  end for
  Estimate $\hat{J}_n^{\pm}$ as:
    $\hat{J}_n^{\pm} = \frac{1}{T} \sum_{t=1}^{T} \Delta_t$
  Compute the estimate of the gradient $\frac{\partial \hat{J}}{\partial \bar{\theta}_n}$
    $\frac{\partial \hat{J}}{\partial \bar{\theta}_n} \leftarrow (\mathbf{d}_n^\intercal \mathbf{d}_n)^{-1} \mathbf{d}_n^\intercal \frac{\hat{J}^+ - \hat{J}^-}{2\beta_n}$
  Update
    $\bar{\theta}_{n+1} = \bar{\theta}_n - \gamma_n \frac{\partial \hat{J}}{\partial \bar{\theta}}$
  $\tau_{n+1} \leftarrow \varsigma \tau_n$
end for

realistic scenario in which the current state of the monitored process transitions to a different state with a certain probability. This state transition probability is modelled as,

$$Pr(X, Y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(Y-X)^2}{2\sigma^2}} / \sum_{Z \in \mathcal{X}} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(Z-X)^2}{2\sigma^2}}, \quad (14)$$

where $X, Y \in \mathcal{X}$ are the states of the monitored process and σ is a deviation factor. Using these probabilities, we model that a small change in the monitored process is more likely than a larger change. In Eq. 14 a higher difference between X and Y results in a smaller argument of the exponential in the numerator and therefore in a smaller transition probability. For example, when monitoring temperature, if the current temperature state is $X_t = 10°\text{C}$, then temperatures around $10°\text{C}$ are more probable in the next time step $t + 1$.

We set the sender's battery capacity to $B_{\max} = 5$ and $E_{max} = 1$. This means that the possible amounts of harvested energy are $\mathcal{E} = \{0, 1\}$. We consider a correlated energy harvesting process. The conditional probabilities to harvest energy in the current time step t based on the harvested energy in the previous time step $t - 1$ are given by $p(1|1) = p(0|0) = 0.7$ and $p(1|0) = p(0|1) = 0.3$.

We consider the following algorithms to compare the performance of our proposed approach:

Value Iteration (VI): This approach determines the optimal transmission strategy based on perfect knowledge about the channel quality, the amounts of harvested energy and the behaviour of the monitored process.

Greedy policy: This is a simple approach in which the sender always transmits ($A = 1$) the information, if it has sufficient energy in the battery, otherwise it stays idle ($A = 0$). Apart from the battery level, this policy does not take the current state into account.

Q-learning: This algorithm uses Q-learning to learn when to transmit according to the current state at the transmitter. To this end, this method learns a *state-action cost value* for each state-action combination in the MDP.

We visualize the threshold-based structure of the optimal transmission policy found by Value Iteration in Fig. 2. Value

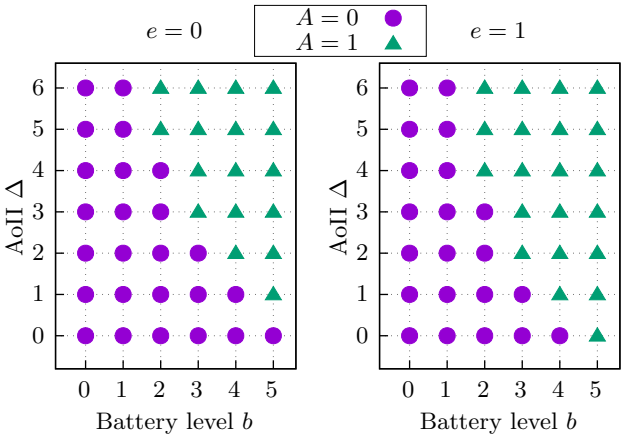

Fig. 2: Behaviour of the Value Iteration policy for $X = 2$, $\hat{X} = 1$, $e \in \{0, 1\}$ and different battery states and AoII values

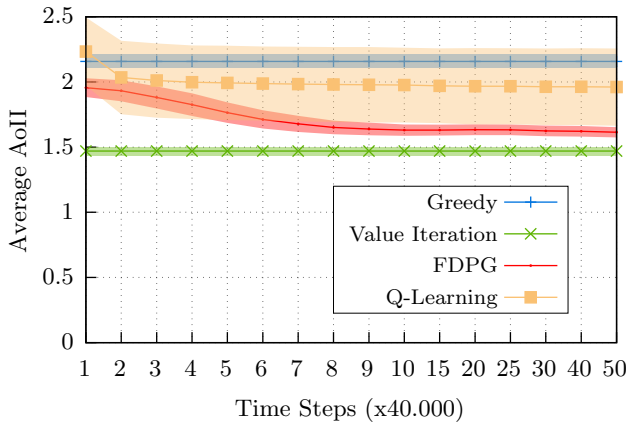

Fig. 3: Comparison of Average AoII over time for reference algorithms including standard deviation areas

Iteration exploits its perfect knowledge about the considered scenario. We show in which states $s = (e, b, \Delta, X, \hat{X})$ it is optimal to transmit (green triangles) and in which states it is optimal to idle (purple circles). We set $N = 5$, keep $X = 2$ and $\hat{X} = 1$ fixed and show the optimal actions for the battery levels $b = 0, ..., 5$, the AoII values $\Delta = 0, ..., 6$, and harvested energy $e = 0$ on the left and $e = 1$ on the right. Fig. 2 shows that for low battery levels ($b = 0$ and $b = 1$) it is optimal to idle. For $b = 2$ and $e = 0$, the sender should transmit as soon as the AoII exceeds 4. This threshold gets lower for $e = 1$ and for higher battery levels. The illustration confirms the threshold-based structure of the optimal transmission policy proven in Section IV. As soon as a specific AoII value depending on b, e, X and $\hat{X}$ is reached, transmitting is preferred over idling. Our proposed approach exploits this threshold-based structure of the optimal transmission policy to minimize the AoII.

In Fig. 3, we study the evolution of average AoII over time by evaluating different transmission strategies. We compare the performance of our proposed approach with the reference algorithms mentioned above. We set the number of states

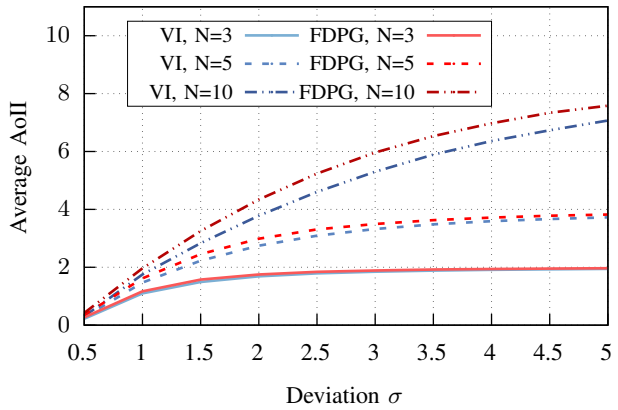

Fig. 4: Effect of number N of states and their transitions in the monitored process on the average AoII

in the monitored process to $N = 5$ and choose $\sigma = 1$ for the state transition probability distribution. Value Iteration performs best with an average AoII value of 1.5. Our proposed approach reaches an average AoII value of 1.6. It outperforms the greedy policy which reaches an average AoII of 2.2 by 28% and the Q-learning based policy which reaches an average AoII of 2.0 by 20%. The performance of the Q-learning based policy improves over time but stays above an average AoII of 2. This results from the fact that the Q-learning based policy needs to learn the cost, i.e., the AoII for every state-action pair, which corresponds to $|\mathcal{E}| \cdot B_{\max} \cdot (M+1) \cdot N \cdot N \cdot |\mathcal{A}|$ possible combinations of all states and all actions. Our proposed approach exploits the threshold based structure of the optimal transmission policy. It therefore only needs to learn $|\mathcal{E}| \cdot B_{max} \cdot N \cdot N$ thresholds and thus learns significantly faster than the Q-learning based policy.

To show that the results of our approach are consistently close to those of Value Iteration, we additionally provide a comparison for different monitored processes. On the horizontal axis in Fig. 4, the value of the deviation factor σ is depicted. On the vertical axis, we measure the average AoII of our proposed policy (red) and the optimal policy obtained via Value Iteration (green). We further investigate the parameter N, representing the total number of states in the monitored process. For higher deviation σ, the average AoII increases. This is the result of a lower probability to remain in the same state of the monitored process in processes with high σ. If N increases, the average AoII also increases. This is again the result of a lower probability to remain in the same state of the monitored process and a lower probability to return to a certain state in the monitored process. In all the considered cases, on average, our proposed solution achieves an AoII only 9.2% above the Value Iteration policy. For $N = 3$ states in the monitored process, this difference is only 4.9% on average.

VII. Conclusions

In this work, we designed a transmission policy for a sender with EH capabilities in a SUS. We first proved that the threshold-based policy, which requires perfect system knowledge at the sender, i.e., knowledge about the state transitions of the monitored process, the channel quality of the link between sender and receiver, and the amount of harvested energy, minimizes the AoII and this, ensures the freshness of the sensed information at the receiver. As finding the optimal thresholds is computationally expensive, we proposed a strategy based on (FDPG) to minimizes the AoII. Our strategy exploits the threshold-based structure of the optimal threshold policy without requiring perfect system knowledge and while being computationally feasible. Our proposed approach shows near optimal performance and outperforms the greedy transmission policy by 28%, the Q-learning based policy by 20%.

References

[1] P. Corke, T. Wark *et al.*, "Environmental wireless sensor networks," *Proceedings of the IEEE*, vol. 98, no. 11, pp. 1903–1917, 2010.

[2] J. K. Hart and K. Martinez, "Environmental sensor networks: A revolution in the earth system science?" *Earth-Science Reviews*, vol. 78, no. 3, pp. 177–191, 2006.

[3] J. Zhao, Y. Wang *et al.*, "Timely device status updates in industrial wireless monitoring systems under resource constraints," *IEEE Internet of Things Journal*, vol. 9, no. 19, pp. 18 791–18 805, 2022.

[4] R. Grodi, D. B. Rawat *et al.*, "Smart parking: Parking occupancy monitoring and visualization system for smart cities," in *SoutheastCon 2016*, 2016, pp. 1–5.

[5] R. Jabbar, M. Shinoy *et al.*, "Urban traffic monitoring and modeling system: An IoT solution for enhancing road safety," in *2019 International Conference on Internet of Things, Embedded Systems and Communications*, 2019, pp. 13–18.

[6] S. Kaul, R. Yates *et al.*, "Real-time status: How often should one update?" in *Proceedings IEEE International Conference on Computer Communications*, 2012, pp. 2731–2735.

[7] E. T. Ceran, D. Gündüz *et al.*, "Learning to minimize age of information over an unreliable channel with energy harvesting," *Computing Research Repository [preprint]*, 2021.

[8] E. T. Ceran, D. Gündüz *et al.*, "A reinforcement learning approach to age of information in multi-user networks with HARQ," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 5, pp. 1412–1426, 2021.

[9] W. de Sombre, F. Marques *et al.*, "A unified approach to learn transmission strategies using age-based metrics in point-to-point wireless communication," in *IEEE Global Communications Conference*, 2023.

[10] A. Maatouk, S. Kriouile *et al.*, "The age of incorrect information: A new performance metric for status updates," *IEEE/ACM Transactions on Networking*, vol. 28, no. 5, pp. 2215–2228, 2020.

[11] A. Maatouk, M. Assaad *et al.*, "The age of incorrect information: An enabler of semantics-empowered communication," *IEEE Transactions on Wireless Communications*, vol. 22, no. 4, pp. 2621–2635, 2023.

[12] ——, "Semantics-empowered communications through the age of incorrect information," in *IEEE International Conference on Communications*, 2022, pp. 3995–4000.

[13] Y. Chen and A. Ephremides, "Minimizing age of incorrect information for unreliable channel with power constraint," in *IEEE Global Communications Conference*, 2021, pp. 1–6.

[14] J. C. Spall, *Introduction to stochastic search and optimization - estimation, simulation, and control*, ser. Wiley-Interscience series in discrete mathematics and optimization. Wiley, 2003.