Wanja de Sombre, Felipe Marques, Friedrich Pyttel, Andrea Ortiz and Anja Klein, ”A Unified Approach to Learn Transmission Strategies Using Age-Based Metrics in Point-to-Point Wireless Communication”, in *Proc. of the IEEE Global Communications Conference - (IEEE Globecom 2023)*, December 2023.

A Unified Approach to Learn Transmission Strategies Using Age-Based Metrics in Point-to-Point Wireless Communication

Wanja de Sombre, Felipe Marques, Friedrich Pyttel, Andrea Ortiz, Anja Klein
Communications Engineering Lab, Technical University of Darmstadt, Germany.
{w.sombre, a.ortiz, a.klein}@nt.tu-darmstadt.de, felipedrmarques@usp.br, friedrich.pyttel@stud.tu-darmstadt.de

***Abstract*—Based on the Age of Information as an optimization criterion, proposals for further age-based metrics have been made in recent years in the Internet of Things (IoT) domain. The research community's great interest in age-based metrics for point-to-point wireless communication has led to a multitude of different scenarios being investigated, including energy optimization, sensing, and risk-sensitivity. All these scenarios involve a sender-receiver pair and revolve around finding appropriate times for the sender to communicate status updates to the receiver. We propose a unified and modular framework that represents the aforementioned options in various combinations and enables transferring solutions developed for specific cases to a variety of scenarios. We generalize an existing optimization approach, which decides to transmit based on a threshold for the age-based metric, using this framework. We develop a unified and extended Q-learning-based algorithm with mechanisms to learn suitable solutions for all scenarios derived from our framework. These mechanisms accelerate the learning process and result in improved algorithmic performance compared to traditional Q-learning. Furthermore, we demonstrate the effectiveness of our solution in numerical simulations. Our unified solution outperforms several reference schemes in terms of age-based metrics, energy consumption, and risk. We present our findings as a starting point to investigate transmission strategies for more general settings with a more efficient approach.**

I. Introduction

Monitoring serves the purpose of collecting data, analyzing system performance, detecting anomalies, and providing feedback to enhance system operation [1]. As this ability to improve system performance, efficiency, and safety is necessary across a broad range of applications, like robotics, vehicular communication, and industrial plants, monitoring systems have become an indispensable tool.

The monitoring application scenarios share a common characteristic: An Internet of Things (IoT) device monitors an environment and sends the status updates to a receiver. The sender's main objective is to balance between the freshness of the status updates, the energy consumption required for transmission, and the potential risk associated with outdated information. To achieve this balance, the sender needs to find effective transmission strategies. However, the exact characteristics of the scenario under which the sender and receiver operate can vary significantly from application to application. These characteristics include the risk-sensitivity of the scenario to delayed updates [2]–[4], the use of ideal [5] or stochastic channel models [3], [6], and the consideration of continuously powered [3], [7], [8] or energy harvesting senders [9]. An additional characteristic is that the sender can sense the monitored environment continuously [6], at random [3], [4], or at specific time instances depending on its needs [7]. Furthermore, different age-based evaluation metrics, e.g., the Age of Information (AoI) as the most frequently used age-based metric [3], [4], [8], [9], the Query Age of Information (QAoI) [10], and the Age of Incorrect Information (AoII) [6] can be considered.

The currently dominant research approach is to optimize transmission strategies for only one specific scenario at a time. This approach suffers from limitations in terms of generalizing solutions to different scenarios. To overcome this, we present a unified framework and an algorithm that can find transmission strategies for a wide range of point-to-point scenarios. Using this framework enables researchers to directly address a large set of scenarios, including new and unexplored ones, without the need for individual treatment of each scenario. An example of a new and unexplored scenario addressed by our framework is the combination of risk with QAoI or with AoII, which is made possible by expanding the risk metric utilized in [4].

Our main contributions can be summarized as follows:

- We introduce a unified and modular framework for age-based metric minimization in point-to-point wireless communication scenarios. The model includes multiple options for sensing, power supply, channel quality, age-based evaluation metric, and risk-sensitivity. The challenge is to identify a transmission strategy that minimizes energy consumption and age-based metric, while also mitigating the risk of high values in the age-based metric. By using the new notion of a *configuration*, a wide spectrum of point-to-point scenarios can be derived from our proposed modular framework. The framework's implementation is publicly available on github: *https://github.com/wanjads/P2PFramework*
- To capture the dynamics of the different scenarios in our framework, we use a single Markov Decision Process

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 (Nr. 16KISK014) and the LOEWE Center EmergenCity.

(MDP). This mathematical model allows us to propose a unified solution to find transmissions strategies for all the included scenarios. An approach proposed in [4] uses a threshold for the age-based metric to decide whether the sender should decide to transmit a status update or not. Using this as a starting point, we provide an algorithm to find thresholds not only for one specific, but all the scenarios included in our framework. We then use this algorithm to determine suitable starting values for risk sensitive Q-learning as proposed in [4]. We additionally include a mechanism for incorporating the battery state into the learning process. These enhancements accelerate the traditional Q-learning algorithm and improve its final performance.

- We further demonstrate that our proposed solution outperforms the standard tabular Q-learning algorithm in optimizing energy consumption, age-based metrics, and mitigating high risk values across new and various wireless communication scenarios. To substantiate this, we evaluate the algorithm across an extensive range of configurations.

The rest of this paper is organized as follows. In Sec. II, the general system model for the framework is described. The MDP modelling the framework mathematically is given together with the optimization problem in Sec. III, followed by a description of the threshold-based approach and the modified Q-learning approach to solve the described optimization problem in Sec. IV. Sec. V contains numerical evaluations of this approach. Finally, we conclude in Section VI with a summary of our contributions.

II. System Model

Figure 1 shows the system model used to construct our framework. The blue components are part of all included scenarios, while the green components offer multiple options depending on the scenario characteristics. The specific configuration of the scenario determines which options are realized. For example, one configuration might use random sensing, energy harvesting, and the AoI, while another configuration might use actively controlled sensing, a continuously powered sender, and a risk-sensitive variant of the QAoI. Regardless of the specific configuration, the whole process is divided into discrete time steps $t \in \mathbb{N}$.

In the following subsections, we provide a detailed description of each component. The parameters for every component of the system model are given in Table A in Fig. 1. Next, we introduce the concept of a *configuration*.

A. System Core

The system core consists of a sender-receiver pair. The sender is equipped with a buffer to store the latest status update. Each time a new status update arrives at the sender, the currently stored status update is dropped and the new status update replaces the old one. In each time step t, the sender decides, whether it sends the currently stored status update to the receiver ($a_t = 1$), or whether it waits without transmitting ($a_t = 0$). If the sender transmits, it incurs discrete energy costs $\nu \in \mathbb{N}$, leading to a finite set of battery states. The set of battery states can be large to ensure fine-grained steps that capture the continuous reality.

To model the channel, we use Bernoulli distributed random variables with success probability $p \in (0, 1]$. If $p = 1$, the channel is assumed to be ideal in the sense that every transmission attempt is successful. After a transmission, the information whether a status update could be decoded successfully at the receiver is then communicated back to the sender through an error-free feedback channel.

B. Customizable Components

1) Monitored Environment: For the monitored environment, there are two options. The first option is selected whenever the age-based metric does not depend on the content of status updates (e.g., for AoI). In such cases, the dynamics of the specific environment can be modeled arbitrarily. Otherwise, the second option is chosen. In this case, we employ a Markov chain, as illustrated in Fig. 1, following the approach presented in [6]. At time step t, the current state of the environment is denoted by $X_t \in 1, ..., N$, where $N \in \mathbb{N}$ represents the number of states of the environment. The environment stays in the same state with a probability $p_r \in (0, 1]$ and transitions to any other state with a probability $p_c := \frac{1-p_r}{N-1}$.

When a transmitted update is successfully decoded, the estimated state at the receiver $\hat{X}_t$ is updated accordingly. This update is reflected in the definition of $\hat{X}_t$ as follows:

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

2) Sensing: The sender monitors the underlying environment through a sensing mechanism s, for which we consider three options, i.e., $s \in \{\text{active, random, perfect}\}$. The option $s = $ active allows the sender to decide when to generate a status update, incurring energy costs of $\mu \in \mathbb{N}$ if it chooses to sense and generate an update. The sender's sensing decision at time step t is denoted by $m_t \in \{0, 1\}$, where $m_t = 1$ indicates that the sender has decided to sense, and $m_t = 0$ indicates that it has decided not to sense. The option $s = $ random is based on randomly arriving status updates, where there is a probability $\lambda \in (0, 1]$ that a new status update arrives in each time step. This is modeled by independent Bernoulli random variables. The energy costs of random sampling can be included into the power supply dynamics, by adjusting the relevant parameters as explained in Sec. II-B3. The option $s = $ perfect assumes the sender has perfect knowledge about the underlying environment, i.e., $\lambda = 1$.

3) Power Supply: We consider three common types of power supply: $e \in \{\text{unlimited, constrained, harvesting}\}$.

The first option, $e = $ unlimited, assumes an unlimited power supply that allows transmission in every time step [3], [4]. In this case, the energy cost is minimized together with the age-based metric chosen from the options in Sec. II-B4.

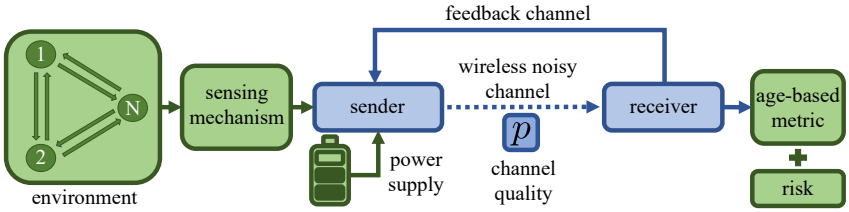

TABLE A: System Parameters

Component	Parameters
Environment	N, p_r
Sensing	s, λ
Power Supply	e, ν, μ, E, B, h_{max}
Channel	p
Metric	D, q
Risk	ρ

Fig. 1: System Model with Parameter Overview

The second option, $e =$ constrained, assumes the power supply is on average limited to a finite amount $E \in \mathbb{R}$. This assumption introduces the following constraint:

$$\lim_{n\to\infty}\left(\nu \cdot \frac{1}{n}\sum_{t=1}^{n} a_t + \mu \cdot \frac{1}{n}\sum_{t=1}^{n} m_t\right) \leq E. \tag{2}$$

Note that if the configuration uses randomly arriving status updates instead of active sampling, μ can be set to 0 and E can be adjusted accordingly.

The third option, $e =$ harvesting, considers that the power supply is a finite battery with capacity $B \in \mathbb{N}$ that can be recharged using energy harvesting. For simplicity, the harvested energy shares the same unit as the energy costs. At the beginning of each time step, an amount h_t of energy is harvested and stored in the battery. h_t is a realization of the random variable H, which follows a discrete and uniform distribution over the range $\{0, ..., h_{max}\}$ with $h_{max} \in \{1, ..., B\}$. Before performing the energy consuming actions $a_t = 1$ and $m_t = 1$, the battery is checked for sufficient charge $b_t \in \{0, ..., B\}$. The charge at the beginning of a time step b_{t+1} is then given by:

$$b_{t+1} := \min(b_t + h_t, B) - \mu \cdot m_t - \nu \cdot a_t, \tag{3}$$

where $\nu \in \mathbb{N}_0$ denotes the transmission energy costs and $\mu \in \mathbb{N}_0$ denotes the sensing energy costs.

4) Metric: We include three options for the age-based evaluation metric D, namely, $D \in \{\text{AoI, QAoI, AoII}\}$. To provide their definitions, we first define the AoI at the sender AoI_{Tx}. It is set to 0 each time a new status update arrives or is generated at the sender. If no new status update arrives in the next time step, AoI_{Tx} is increased by 1:

$$\text{AoI}_{\text{Tx},t+1} := \begin{cases} 0 & \text{if a new update arrives,} \\ \text{AoI}_{\text{Tx},t} + 1 & \text{otherwise.} \end{cases} \tag{4}$$

We set $\text{AoI}_{\text{Tx},1} = 0$.

The first option for the age-based metric, i.e., $D =$ AoI, considers the AoI at the receiver AoI_{Rx} which is defined as:

$$\text{AoI}_{\text{Rx},t+1} := \begin{cases} \text{AoI}_{\text{Tx},t} + 1 & \text{if a new update is decoded,} \\ \text{AoI}_{\text{Rx},t} + 1 & \text{otherwise.} \end{cases} \tag{5}$$

Furthermore, we set $\text{AoI}_{\text{Rx},1} = 1$.

The second option, $D =$ QAoI, is based on the definition of the AoI, but uses its value only in specific time steps called query time steps. In other time steps, the QAoI is set to 0. As in [10], query time steps are randomly selected by performing a Bernoulli trial for each time step with a probability $q \in (0, 1]$.

The third option, $D =$ AoII, measures the number of time steps since the last time when the information at the receiver matched the state of the underlying environment. It is defined similar to the AoI, however, it is important to note that the AoII assumes perfect knowledge about the underlying environment. For configurations with AoII we hence assume that $s =$ perfect. This assumption results in AoI_{Tx} being constantly 0. The definition of the AoII is then given by

$$\text{AoII}_{t+1} := \begin{cases} 0 & \text{if } X_t = \hat{X}_t, \\ \text{AoII}_t + 1 & \text{otherwise.} \end{cases} \tag{6}$$

5) Risk: Real-world scenarios often exhibit risk-sensitivity, where high values of the age-based metric can lead to severe potential harm. To include this in our framework, we use our ideas presented in [4], specifically the concept of risky states. Risky states are states with a high value of the age-based metric and a risk-aware transmission strategy should aim at avoiding these states. The risk-sensitivity of a scenario is modelled using a parameter $\rho \geq 1$. Higher values of ρ indicate a higher sensitivity to risk. Here $\rho = 1$ means that the scenario is considered to be risk-neutral.

C. Configurations

We proceed with the definition of a configuration:

Definition 1. *A configuration C is a tuple*

$$C = (N, p_r, s, e, E, B, h_{max}, \nu, \mu, p, \lambda, D, q, \rho). \tag{7}$$

For a given configuration, the corresponding system model is a special case of the model depicted in Fig. 1 and consists of an underlying environment with $N \in \mathbb{N}$ states, a probability $p_r \in [0, 1]$ to remain in an environment state, a sensing mechanism $s \in \{$random, active, perfect$\}$, a type of power supply $e \in \{$unlimited, constrained, harvesting$\}$, a maximal average energy $E \in \mathbb{R}$, a battery capacity $B \in \mathbb{N}$, a maximal amount of harvested energy $h_{max} \in \{1, ..., B\}$, transmission energy costs $\nu \in \mathbb{N}_0$, sensing energy costs $\mu \in \mathbb{N}_0$, a channel quality $p \in (0, 1]$, a probability $\lambda \in (0, 1]$ of a new status update at the sender in each time step, an evaluation metric $D \in \{$AoI, QAoI, AoII$\}$, a probability $q \in [0, 1]$ of query time steps and the tolerated degree $\rho \geq 1$ of risk-sensitivity.

Remark 2. *It is evident that selecting certain parameters can render other parameters inconsequential for the resulting system model. For instance, if $D \neq$ AoII, the parameter N can be chosen arbitrarily without influencing the results of a simulation of the resulting system model. This does not compromise the effectiveness of the presented model.*

III. Problem Formulation

In this section, we mathematically define the MDP we use to model the system described in Sec. II.

Definition 3. *The MDP $\mathcal{M}_C = (\mathcal{S}, \mathcal{A}, P, c)$ which models the system in Fig. 1 is specified for configuration C by defining its individual components:*

$\mathcal{S} = \{0,1\} \times \{0, ..., B\} \times \mathbb{N}_0 \times \mathbb{N} \times \mathbb{N}_0$,

$\mathcal{A} = \{0,1\} \times \{0,1\}$,

$P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1]$ *with*

$P(S_t, (m_t, a_t), S_{t+1}) = P_{\text{proc},t} P_{\text{bat},t} P_{\text{sens},t} P_{\text{Tx},t} P_{\text{AoII},t}$,

$c : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1]$ *with*

$c_t := c(S_t, (a_t, m_t), S_{t+1}) := D_{t+1} + \mu \cdot m_t + \nu \cdot a_t$, *where*

- $\mathcal{S} = \{S_t = (I_t, b_t, AoI_{\text{Tx},t}, AoI_{\text{Rx},t}, AoII_t) | t = 1, 2, ...\}$ *is the set of states. I_t indicates, whether the information about the underlying environment corresponds to the currently stored state at the receiver, while b_t, $AoI_{\text{Tx},t}$, $AoI_{\text{Rx},t}$ and $AoII_t$ are defined in Sec. II.*
- $(m_t, a_t) \in \mathcal{A}$ *is the action at time step t, consisting of the sensing action m_t and the sending action a_t. If $s =$ random, m_t has no effect on state transitions.*
- D_t *is the value of the respective age-based metric indicated in the configuration at time step t,*
- $P_{\text{bat},t}$*, denoting the factor of the transition probability related to the battery state, depends on $S_t, (a_t, m_t), S_{t+1}$ and on e.*

  If $e = harvested$:

$$P_{\text{bat},t} := \begin{cases} \frac{1}{h_{max}+1} & \text{if } h \in \{0, ..., h_{max}\} \text{ and} \\ & b_{t+1} = \min(S_t + h, B) - \mu m_t - \nu a_t, \\ 0 & \text{otherwise}, \end{cases}$$

  If $e = constrained$:

$$P_{\text{bat},t} := \begin{cases} 1 & \text{if } b_{t+1} = S_t + E - \mu m_t - \nu a_t, \\ 0 & \text{otherwise}, \end{cases}$$

  otherwise, if $e =$ unlimited, $P_{\text{bat},t} := 1$ and $b_t = \infty$ for all time steps t.
- *The definitions of $P_{\text{proc},t}$, $P_{\text{sens},t}$, $P_{\text{Tx},t}$ and $P_{\text{AoII},t}$, which represent the remaining factors of the transition probability associated with the environment, sensing, transmission, and AoII, respectively, are straightforward. For the sake of brevity, we omit their explanation here.*

We additionally define the set $\mathcal{R}_D$ of risky states, which depends on the age-based metric. We define:

$$\begin{aligned} \mathcal{R}_{\text{AoI}} &:= \{S_t : \text{AoI}_{\text{Rx},t} \geq \zeta_{\text{AoI}}\}, \\ \mathcal{R}_{\text{AoII}} &:= \{S_t : \text{AoII}_t \geq \zeta_{\text{AoII}}\}, \\ \mathcal{R}_{\text{QAoI}} &:= \{S_t : \text{AoI}_{\text{Rx},t} \geq \zeta_{\text{QAoI}} \wedge t \text{ query time step}\}, \end{aligned} \tag{8}$$

where ζ_{AoI}, ζ_{AoII} and ζ_{QAoI} are risk-thresholds.

Using the definitions above, our objective is to identify a strategy $\pi : \mathcal{S} \to \mathcal{A}$ for a given configuration C, with the aim of minimizing the average long-term costs expressed as:

$$costs(\pi) := \mathbb{E}\left[\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} c\left(S_t, \pi(S_t), S_{t+1}\right)\right]. \tag{9}$$

Algorithm 1: Threshold Finder

Data: sensing type s, power supply type e, energy bound E, age-based measure D, simulator for $\mathcal{M}_C$, action space $\mathcal{A}$, starting state S_1, no. of time steps per run T
Result: Best threshold $\mathcal{T}$

```
last_costs ← ∞
continue ← True
𝒯 ← 0
while continue do
    costs ← 0
    S_t ← S_1
    for i=1,...,T do
        (m_t, a_t) ← (0, 0)
        if D ∈ {AoI, QAoI} and AoI_Rx,t − AoI_Tx,t ≥ 𝒯 then
            (m_t, a_t) ← (1, 1)
        end
        if D = AoII and QAoI_t ≥ 𝒯 then
            (m_t, a_t) ← (1, 1)
        end
        sample S_{t+1}, c_t using M_C and a = (m_t, a_t)
        costs ← (1/i)(c_t + (i − 1) · costs)
        S_t ← S_{t+1}
    end
    if costs > last_costs then
        continue ← False
    else
        𝒯 ← 𝒯 + 1
        last_costs ← costs
    end
end
return 𝒯 − 1
```

If $e = constrained$, we add the constraint defined in Eq. (2) to the minimization problem. To evaluate the risk associated with a strategy, we utilize the frequency of risky states as defined in [4]:

$$f_{\mathcal{R}_D} := \mathbb{E}\left[\lim_{m\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}_{S_t \in \mathcal{R}_D}\right]. \tag{10}$$

IV. The Q-learning Based Solution

In this section, we present our Q-learning based solution for determining transmission strategies in the scenarios derived from our framework. Our solution is based on our threshold-based approach introduced in [4]. This threshold-based approach uses the idea of selecting time steps with high $\text{AoI}_{\text{Rx},t}$ and low $\text{AoI}_{\text{Tx},t}$ as the optimal transmission times, where the difference between the two serves as the decision criterion. We directly transfer this idea to QAoI and by replacing $\text{AoI}_{\text{Rx},t}$ with AoII_t, we generalize the concept to AoII. The central novelty of our algorithm is the combination of this threshold-based approach with risk-sensitive Q-learning. This is realized by initializing the Q-values according to the best threshold-based strategy.

The optimization approach presented in [4] for finding the optimal threshold for the age-based metric lacks generalizability. We instead propose Algorithm 1 to find the optimal threshold $\mathcal{T}$ for all the scenarios derived from out framework by simulations. The algorithm assumes $a_t = m_t$. This simplification allows to directly transfer the idea of a single threshold to all scenarios within the framework. The general case, including $a_t \neq m_t$, is considered by the Q-learning algorithm.

Finding the costwise best threshold, resulting from the trade-off between energy costs and age-based costs, is a convex

Algorithm 2: Q-learning Framework (QLF)

Data: transmission costs ν, sensing costs μ, sensing type s, power supply type e, energy bound E, age-based measure D, risk-factor ρ, simulator for $\mathcal{M}_C$, action space $\mathcal{A}$, starting state S_1, no. of time steps T, real learning rates $(\alpha_t)_{t\in\{1,\dots,T\}}$, discount factor γ, initial ε, decay factor δ, risk-threshold ζ_D, initial Q-values $Q^{(0)}$

Result: $Q^{(T)}$-values as estimates for Q-values

1 $\mathcal{R} \leftarrow \mathcal{R}_D$
2 $Q^{(1)} \leftarrow Q^{(0)}$
3 $S_t \leftarrow S_1$
4 **for** *t=1,...,T* **do**
5 sample a random action $a = (m_t, a_t)$ ε-greedy
6 update ε as $\varepsilon \leftarrow \delta \cdot \varepsilon$
7 sample next state S_{t+1} and cost c_t using the simulator for $\mathcal{M}$ and the action a
8 **if** $e \in$ {*constrained, harvested*} **then**
9 $c_t \leftarrow c_t + \beta(b_t) \cdot (\mu \cdot m_t + \nu \cdot a_t)$
10 **end**
11 **if** $S_{t+1} \in \mathcal{R}$ **then**
12 $c_t \leftarrow \rho \cdot c_t$
13 **end**
14 **for** $(S', a') \in \mathcal{S} \times \mathcal{A}$ **do**
15 **if** $S' = S_t$ & $a' = a$ **then**
16 $V(S_{t+1}) \leftarrow \max_{\hat{a}\in\mathcal{A}} Q^{(t-1)}(S_{t+1}, \hat{a})$ $Q^{(t)}(S', a') \leftarrow (1-\alpha_t)Q^{(t-1)}(S', a') + \alpha_t(c_t + \gamma V(S_{t+1}))$
17 **else**
18 $Q^{(t)}(S', a') \leftarrow Q^{(t-1)}(S', a')$
19 **end**
20 **end**
21 $S_t \leftarrow S_{t+1}$
22 **end**
23 **return** $Q^{(T)}$

problem. Algorithm 1 exploits this convexity property to numerically find the value of $\mathcal{T}$. The algorithm gradually increases $\mathcal{T}$ (lines 4-24) while evaluating the associated costs. At each threshold, the algorithm simulates the system using $\mathcal{M}_C$ (lines 5-18) and calculates the costs using incremental averaging (line 16). Initially, costs are high but they decrease as $\mathcal{T}$ increases. The algorithm stops increasing $\mathcal{T}$ as soon as the costs start to increase again (lines 19-21), indicating that the optimal threshold has been reached for the given configuration. Finally, the algorithm returns the optimal threshold (line 26).

We use the threshold found by Algorithm 1 to initialize the $Q^{(0)}$ values of the actions according to the resulting strategy. Specifically, we set the $Q^{(0)}$ values of the chosen actions to 0 and those of the remaining actions to a larger value $K_{\max}$. Based on these initial Q-values, we utilize ε-greedy tabular Q-learning. The pseudo-code for this algorithm is provided in Algorithm 2. Traditional Q-learning is implemented in lines 3-7 and 14-23. In each time step t, an action $a = (m_t, a_t)$ is chosen for the current state S_t (line 5), while balancing exploration and exploitation using the ε-greedy mechanism. Then ε is updated (line 6) and $\mathcal{M}_C$ is used to simulate a single step of the scenario (line 7). The resulting costs are later used to update the Q-value of the encountered state-action pair (lines 14-20). Before the next iteration starts, the current state is updated (line 21). To incorporate risk, we include risk-sensitivity as in [4], where the costs associated with transitioning to a risky state are scaled by a factor ρ. To this end, the set of risky states is defined in line 1 as per Eq. (8). By including energy costs in lines 8-10 for the cases where $e =$ constrained and $e =$ harvested, we accelerate the learning process. We introduce a dynamic weighting factor β of the energy costs, which is a strictly monotonically decreasing function of the battery charge: $\beta(b_t) := k_1 \cdot \exp\left(-k_2 \frac{b_t}{B}\right)$. Neglecting the energy costs associated with battery depletion would pose a significant challenge to the algorithm in comprehending the impact of an empty battery state. This restriction would force the algorithm to only learn about the impact of energy scarcity when the battery is empty, thus necessitating frequent visits to these states and increased sensitivity to future events. Incorporating energy costs linearly exhibited comparable issues. Our novel approach, which incorporates energy costs at each time step using the weighting factor β, overcomes this limitation, enabling the algorithm to converge more rapidly by eliminating the need to reach empty battery time steps.

V. Numerical Evaluation

A. Reference schemes

This section presents a numerical comparison of four strategies. The first strategy (rand) randomly chooses the action $(m_t, a_t) = (1, 1)$ with a probability of p_{random}, depending on the considered scenario. Otherwise, the random strategy chooses $(m_t, a_t) = (0, 0)$. The second strategy (TQL) uses traditional Q-learning. The third strategy (TB) utilizes thresholds generated by Algorithm 1, while our proposed strategy (QLF) employs the Q-values learned by Algorithm 2.

B. Simulation setup

To show the variety of possible scenarios in the framework and the performance of our solution, we consider nine configurations C_1 up to C_9. The specific parameters of each configuration are listed in Table I.

In the following, 100 realizations are considered. For the learning algorithms, we use $T = 10^5$ training time steps in each realization. Afterward, we test all four strategies in each realization over 10^4 time steps. Both Q-learning algorithms use $\varepsilon_0 = 0.9$, a factor $\delta = (0.0001/0.9)^{1/T}$ to decrease ε, $\gamma = 0.7$, and a constant learning rate of $\alpha_t = 0.007$. The risk thresholds are set to $\zeta_{\text{AoI}} = 5$ and $\zeta_{\text{AoII}} = 3$. For the case of $e =$ constrained, we set $p_{\text{random}} = E(\mu + \nu)^{-1}$, and for the case of $e =$ harvested, we set $p_{\text{random}} = B(2(\mu+\nu))^{-1}$, with the objective of maximizing energy utilization. We set $k_1 := 2$ and $k_2 := 5$ and $K_{\max}$ to $K_{\max} := 10$.

C. Numerical results

We present our numerical results in Fig. 2, Fig. 3, Fig. 4, and Fig. 5, demonstrating the effectiveness of our framework in comparing a vast number of scenarios without the need for completely re-implementing individual scenarios. Note that the ordinate scales in Fig. 2 and Fig. 3 are logarithmic to ensure all bars are visible. Each figure includes a sketch of the considered scenario, with the highlighted component varied between configurations. Notably, our proposed solution exhibits broad applicability, proving to be effective across all tested configurations, including AoI, AoII, risky states, and QAoI metrics. Only rand consistently performs poorly, with the exception of C_2, where it achieves a low frequency of risky states due to the choice of $p_{\text{random}} = 1$. Our results highlight

config	N	p_r	s	e	E	B	h_{max}	ν	μ	p	λ	D	q	ρ
$C_1/C_2/C_3$	10	0.5	perfect	constr./unl./harv.	$0.5/\infty/\infty$	0/0/10	0/0/1	1	2	0.9	1	AoII	1	2
$C_4/C_5/C_6$	10	0.5	act./rand./perf.	unlimited	∞	0	0	4	1	0.9	0/0.5/1	AoI	1	2
$C_7/C_8/C_9$	10	0.5	random	harvested	∞	5	1	0	1	0.9	0.8	QAoI	1/0.75/0.5	1

TABLE I: Parameters for the configurations C_1 to C_9

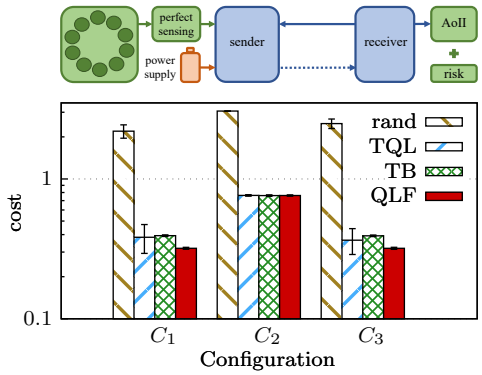

Fig. 2: Average cost after 10^5 learning time steps for configurations C_1 to C_3

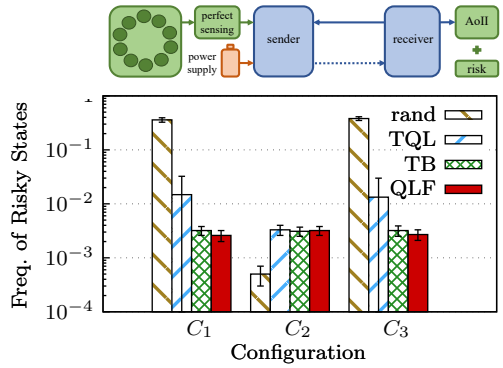

Fig. 3: Freq. of risky states after 10^5 learning time steps for configurations C_1 to C_3

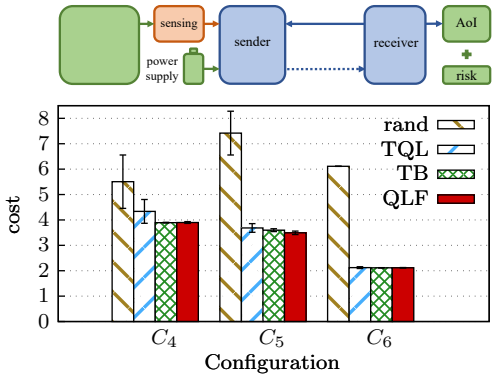

Fig. 4: Average cost after 10^5 learning time steps for configurations C_4 to C_6

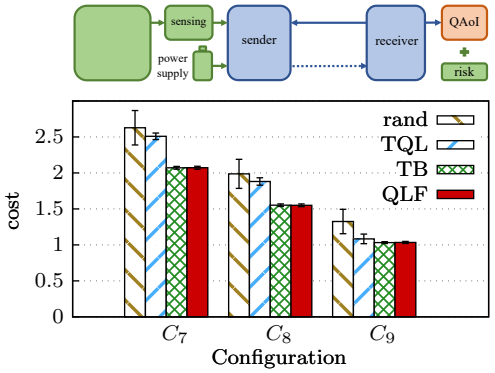

Fig. 5: Average cost after 10^5 learning time steps for configurations C_7 to C_9

QLF's clear advantage over TQL, while QLF and TB produce comparable results in most cases, with QLF outperforming TB in certain configurations. The clear advantage of QLF over TQL can be attributed to TB's ability to provide a strong starting point for QLF's learning process. Furthermore, in cases where e = harvested and e = constrained, learning is accelerated by using the weighting factor β.

VI. Conclusions

In this paper, we introduce a unified framework that can model a wide range of point-to-point communication scenarios, including various options for sensing, power supply, age-based metrics, and risk-sensitivity. Our framework allows for generalizing solutions developed for specific cases to a large set of related applications. By utilizing an MDP, we present a mathematical model of this framework and provide a risk-sensitive Q-learning algorithm to find solutions for all the scenarios represented by the framework. For that, we use an approach based on a threshold for the age-based metric. Furthermore, we demonstrate the versatility of our approach for new and various configurations. Our results clearly demonstrate the effectiveness of our solution in several scenarios and show that it provides a solid foundation for future research.

References

[1] L. Russell *et al.*, "Agile IoT for critical infrastructure resilience: Cross-modal sensing as part of a situational awareness approach," *IEEE Internet Things J.*, vol. 5, no. 6, pp. 4454–4465, 2018.

[2] Abd-Elmagid *et al.*, "On the role of age of information in the internet of things," *IEEE Commun. Mag.*, vol. 57, no. 12, pp. 72–77, 2019.

[3] B. Zhou *et al.*, "Risk-aware optimization of age of information in the internet of things," in *IEEE Int. Conf. Commun. (ICC)*, 2020, pp. 1–6.

[4] W. de Sombre *et al.*, "Risk-sensitive optimization and learning for minimizing age of information in point-to-point wireless communications," *IEEE Int. Conf. Commun. (ICC)*, pp. 1–6, 2023.

[5] O. Ayan *et al.*, "Age-of-information vs. value-of-information scheduling for cellular networked control systems," *CoRR*, vol. abs/1903.05356, 2019. [Online]. Available: http://arxiv.org/abs/1903.05356

[6] A. Maatouk *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.

[7] B. Zhou *et al.*, "Joint status sampling and updating for minimizing age of information in the internet of things," *IEEE Trans. Commun.*, vol. 67, no. 11, pp. 7468–7482, 2019.

[8] E. T. Ceran *et al.*, "A reinforcement learning approach to age of information in multi-user networks," in *2018 IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 2018, pp. 1967–1971.

[9] B. T. Bacinoglu *et al.*, "Age of information under energy replenishment constraints," in *2015 Information Theory and Applications Workshop (ITA)*, 2015, pp. 25–31.

[10] F. Chiariotti *et al.*, "Query age of information: Freshness in pull-based communication," *IEEE Trans. Commun.*, vol. 70, no. 3, pp. 1606–1622, 2022.