Wanja de Sombre, Andrea Ortiz, Frank Aurzada, Anja Klein, “Risk-Sensitive Optimization and Learning for Minimizing Age of Information in Point-to-Point Wireless Communications,” in *IEEE International Conference on Communications (ICC)*, Rome, Italy, May 2023.

Risk-Sensitive Optimization and Learning for Minimizing Age of Information in Point-to-Point Wireless Communications

Wanja de Sombre*, Andrea Ortiz*, Frank Aurzada†, Anja Klein*
*Communications Engineering Lab, Technical University of Darmstadt, Germany.
†Probability and Statistics Group, Mathematics Departement, Technical University of Darmstadt, Germany.
{w.sombre, a.ortiz, a.klein}@nt.tu-darmstadt.de, aurzada@mathematik.tu-darmstadt.de

***Abstract*—When using Internet of Things (IoT) networks for monitoring, devices rely on fresh status updates about the monitored process. To measure the freshness of these status updates, the concept of Age of Information (AoI) is used. However, critical applications, e.g., those involving human safety, require not only fresh updates, but also a low risk of experiencing high AoI values. In this work, we introduce the notion of *risky states* for these high AoI events. We consider a point-to-point wireless communication scenario containing a sender transmitting randomly arriving status updates to a receiver through a wireless channel. The sender decides, when to send a status update and when to wait for a newer one. The sender's goal is to jointly minimize the AoI at the receiver, the required transmission energy and the frequency of visiting risky states. We present two solutions for this problem using optimization and learning, respectively For the optimization approach, we propose a family of threshold-based transmission strategies, which trigger a transmission whenever the difference between the AoI at the sender and at the receiver exceeds a certain threshold. Our proposed learning approach directly includes our notion of risky states into traditional Q-learning. As a result, it balances the minimization of AoI and the required transmission energy, with the frequency of visiting risky states. Through numerical results, we show that our proposed risk-aware approaches outperform relevant reference schemes. Moreover, and in contrast to value iteration, their computational complexity does not depend on the set of possible AoI values.**

I. Introduction

The resilience and robustness in applications, like robotics, vehicular communication, or even in critical infrastructure can be improved by exploiting IoT networks for monitoring [1]. Monitoring requires that IoT sensors transmit status updates regarding the monitored processes in a timely manner over unreliable, in general wireless communication channels. To measure the freshness of the received status updates, the concept of AoI was introduced in [2]. The AoI requirements for the status updates depend on the considered application. For example, applications concerning human safety usually have strict AoI requirements [3]. Moreover, these strict requirements may not only concern low average AoI, but also a low probability of experiencing large AoIs.

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.

Minimizing the average AoI in wireless communication systems and, in particular, in point-to-point wireless communication systems, has been the focus of recent research effort [4]–[12]. In [4], queuing theory is used to derive closed-form expressions for the average AoI at the receiver under different queue models. In [5] the authors consider a monitoring system. They propose a strategy at the sender to decide when to sample the monitored process and when to transmit a status update to the receiver. In a similar scenario, and assuming the sender monitors a dynamic Markov process with a fixed rate, the authors in [6] exploit differential encoding to increase the system's reliability against transmission errors. In [7], a capacity-constrained point-to-point scenario is considered. Assuming that the transmission of a status update requires multiple channel uses, the authors propose a transmission strategy to decide if an ongoing transmission should be aborted when a new status update arrives. In [8], the average AoI is minimized considering that the receiver is only interested in status updates at specific times.

All the previously mentioned works focus on the minimization of the average AoI, i.e., they optimize the AoI at the receiver over a long time horizon. However, average AoI minimization does not prevent the occurrence of events in which the AoI exceeds a predefined safety value. We term such events as *risky states* and use them to quantify and to minimize the risk of having large AoIs in the considered scenario. The name risky states comes from the fact that a large AoI at the receiver can compromise the system in critical applications like, for example, industrial IoT [12], [13]. For this reason, a new research direction has emerged which, in addition to the average AoI, focuses on the peak AoI. The goal of most of the current works considering peak AoI is to characterize the probability of reaching risky states under different assumptions, e.g., short status update packets [9], status update sources with and without retransmissions [10], and a customizable status update arrival rate at the sender [11].

However, the design of risk-aware transmission strategies at the sender has, so far, received little attention. The authors

in [12] take a step in this direction by proposing the use of value iteration to derive a risk-aware transmission strategy at the sender. Assuming that the probabilities for a status update arrival and for a successful transmission are a-priori known, the authors jointly minimize the average AoI at the receiver, the average energy required for the transmissions and a risk-measure. Although value iteration leads to the optimal transmission policy, it is computationally expensive and requires small sets of possible values of the AoI to derive the optimal policy in reasonable time.

In this work, we focus on the development of transmission strategies considering the average AoI and the average energy required for transmissions. However, and in contrast to the previous works, our focus lies on the design of *scalable* transmission strategies that *jointly* reduce the average AoI, the average required transmission energy and the occurrence of risky states. To this end, we consider the cost of the strategy and the risk associated to it. The cost is defined as the average over the costs in each time step, i.e, the average over the weighted sum of the AoI at the receiver and the transmit energy at the sender. The risk of the strategy is defined as the frequency with which the strategy's execution leads to *risky states*. To minimize cost and risk, we follow two approaches: offline optimization and reinforcement learning.
The solutions we contribute can be summarized as follows:

- Concerning offline optimization, we propose a family of threshold-based transmission strategies. Each strategy in this family has a unique threshold and triggers the transmission of status updates whenever the difference between the AoI at the sender and at the receiver exceeds this threshold. Depending on this threshold, each of the strategies in our family of threshold-based strategies has different properties. For example, strategies with lower thresholds cause lower risk, but may also cause higher costs. We provide the costwise optimal threshold-based strategy (TB-Opt) out of the proposed family of threshold-based strategies. Additionally, we provide methods to balance cost and risk for these strategies. We further derive a closed-form expression for the strategies' average cost. Likewise, we provide an expression for the frequency with which risky states will be visited under a threshold-based strategy. In contrast to value iteration, our proposed optimization approach is able to handle large sets of possible AoIs.
- Concerning reinforcement learning, we propose a novel risk-sensitive variation of Q-learning. We directly include our proposed notion of risky states into a risk-aware learning algorithm, which we call *Q-learning using risky states* (Q+RS). Q+RS is able to balance cost and risk using a tunable risk-parameter. At the same time, Q+RS does not depend on a-priori knowledge of the probabilities of a new status update arrival and of a successful transmission. By means of numerical simulations, we show that, compared to strategies based on traditional learning approaches, the risk-aware strategy derived from Q+RS not only reduces the occurrence of risky states, but also the cost in the system.

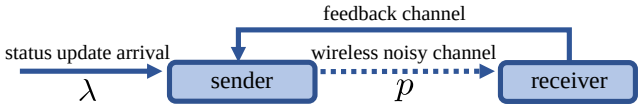

Fig. 1: System Model

The rest of the paper is organized as follows. The considered system model and the formulation of the optimization problem are presented in Sec. II and Sec. III, respectively. In Sec. IV, we introduce threshold-based strategies including TB-Opt . Our new risk-aware learning strategy Q+RS is introduced in Sec. V. The numerical evaluation of the proposed strategies is presented in Sec. VI, Sec. VII concludes the paper.

II. System Model

As considered in [12] and as shown in Fig. 1, the system consists of a sender, a receiver and a wireless packet erasure channel connecting both. The sender can be an IoT-device receiving status updates (e.g. distances or temperatures). The receiver relies on fresh information from this IoT-device.

The system uses discrete and equidistant time steps, indexed by $t \in \mathbb{N}$. The status update arrival process is modeled as a Bernoulli process, such that at the beginning of each time step t, an update arrives at the sender with probability λ. The sender has a buffer, able to store only the freshest status update. This means that, as soon as a new status update arrives, the currently stored update is replaced by the new one. Assume that status updates arrive at random time steps $t = t_i$, where $i \in \mathbb{N}$. Then the AoI at the sender evaluated at time step $t \in [t_i, t_{i+1} - 1]$, is denoted as $\mathrm{AoI}_{\mathrm{Tx},t} \in \mathbb{N}_0$ and it is defined as

$$\mathrm{AoI}_{\mathrm{Tx},t} := t - t_i, \quad \text{for } t \in [t_i, t_{i+1} - 1]. \tag{1}$$

Hence, the minimal $\mathrm{AoI}_{\mathrm{Tx},t}$ is 0.

In each time step t, the sender has then to decide, whether it wants to send the currently stored status update to the receiver or not. If the sender decides to transmit, the status update can correctly be detected at the receiver with a probability of p. This probability p models the quality of the wireless noisy channel. The transmission power is constant, i.e. each sending attempt needs the same amount of energy ν. We assume that the sender receives information of whether the packet could correctly be detected at the receiver or not via a perfect feedback channel. This allows the sender to keep track of the AoI at the receiver. The AoI at the receiver is denoted as $\mathrm{AoI}_{\mathrm{Rx},t} \in \mathbb{N}$ and is defined as

$$\mathrm{AoI}_{\mathrm{Rx},t+1} := \begin{cases} \mathrm{AoI}_{\mathrm{Tx},t} + 1, & \text{if a transmission} \\ & \text{attempt succeeds at } t, \\ \mathrm{AoI}_{\mathrm{Rx},t} + 1, & \text{otherwise.} \end{cases} \tag{2}$$

Note that the lowest possible value of $\mathrm{AoI}_{\mathrm{Rx}}$ is 1, while the lowest possible value of $\mathrm{AoI}_{\mathrm{Tx}}$ is 0. This modeling decision was adopted from [12].

The decision of the sender in time step t results in a cost C_t *associated with the single time step* t. C_t is defined as the

weighted sum of the age of information AoI_{Rx} at the receiver and the transmission energy ν. This means that a persistently high AoI_{Rx} will result in high costs. A successful transmission will result in a lower AoI_{Rx} and hence in lower costs. Formally, C_t is defined as

$$C_t = \begin{cases} \alpha \text{AoI}_{\text{Rx},t} + \beta\nu & \text{if the sender sends,} \\ \alpha \text{AoI}_{\text{Rx},t} & \text{otherwise,} \end{cases} \tag{3}$$

where α and β are weights on the AoI and the energy cost.

The cost *of a strategy* π is defined to be the long-term average of the costs of all single time steps:

$$cost(\pi) := \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}[C_t|\pi], \tag{4}$$

where $\mathbb{E}[C_t|\pi]$ denotes the expected costs in time step t under strategy π.

We additionally introduce the concept of *risky states* as events in which the $\text{AoI}_{\text{Rx}} \geq \zeta$, where $\zeta \in \mathbb{N}$ is a predefined safety value, which is measured in time steps. In the context of the given application, this safety value ζ quantifies the idea that the information at the receiver might be too old, which possibly results in undesirable system behaviour.

The details of how cost and risk are considered in the proposed approaches are explained in Sec. IV and Sec. V.

III. Problem Formulation

In this section, we formulate the problem as an average-cost Markov Decision Process (MDP) $\mathcal{M}$. For given parameters $p, \lambda \in (0,1)$ and $\nu \geq 0$, the MDP $\mathcal{M}$ modeling the described system consists of a set $\mathcal{S} := \mathbb{N}_0 \times \mathbb{N}$ of states, a set $\mathcal{A} := \{0,1\}$ of actions, a cost function c and state transition probabilities given by a function P. Each state $s \in \mathcal{S}$ is a pair of natural numbers modeling the AoI at the sender and at the receiver, i.e., $s = (\text{AoI}_{\text{Tx}}, \text{AoI}_{\text{Rx}})$. The action space $\mathcal{A}$ contains two actions. 0 means that the sender waits, 1 corresponds to a sending attempt. The cost function c returns the cost of a state-transition (s_t, a, s_{t+1}), i.e., the cost arising from transitioning from state $s_t = (\text{AoI}_{\text{Tx},t}, \text{AoI}_{\text{Rx},t})$, at time step t, to state $s_{t+1} = (\text{AoI}_{\text{Tx},t+1}, \text{AoI}_{\text{Rx},t+1}) \in \mathcal{S}$ at time step $t+1$ after taking action $a \in \mathcal{A}$. We define the function $c : \mathcal{S}\times\mathcal{A}\times\mathcal{S} \to \mathbb{R}$ as $c(s_t, a, s_{t+1}) = C_{t+1}$ using C_{t+1} defined in (3). According to the previously described system, the transition probability function $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1]$ is defined, such that the probability for a new packet ($\text{AoI}_{\text{Tx,t+1}} = 0$) is λ. Otherwise, $\text{AoI}_{\text{Tx,t+1}} = \text{AoI}_{\text{Tx,t}} + 1$. Independently, if $a_t = 1$, the probability for a successful transmission ($\text{AoI}_{\text{Rx,t+1}} = \text{AoI}_{\text{Tx,t}} + 1$) is p. Otherwise, or if $a_t = 0$, $\text{AoI}_{\text{Rx,t+1}} = \text{AoI}_{\text{Rx,t}} + 1$. Strategies π for the solution of this MDP are maps from $\mathcal{S}$ to $\mathcal{A}$. Expressing the average cost as defined in (4) for the MDP, we write

$$cost(\pi) := \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}[c(s_t, \pi(s_t), s_{t+1})], \tag{5}$$

where the occurrence of the state s_t at time step t depends on the transition probabilities of $\mathcal{M}$. $\mathbb{E}[c(s_t, \pi(s_t), s_{t+1})]$ denotes the expected cost of the transition $(s_t, \pi(s_t), s_{t+1})$.

For a safety value ζ, the set $\mathcal{R}$ of risky states is given as

$$\mathcal{R} := \{s = (\text{AoI}_{\text{Tx}}, \text{AoI}_{\text{Rx}}) \in \mathcal{S} | \text{AoI}_{\text{Rx}} \geq \zeta\}. \tag{6}$$

IV. The Threshold-Based Approach

In this section, we present our proposed threshold-based transmission strategy based on optimization. This strategy $\pi_{TB}(n)$ is characterized by a threshold n. According to $\pi_{TB}(n)$, the transmitter sends, if and only if the difference of the AoI at the receiver and that at the sender is equal to or larger than n. We define $\pi_{TB}(n)$ as the following strategy:

$$\pi_{TB}(n)((\text{AoI}_{\text{Tx}}, \text{AoI}_{\text{Rx}})) = \begin{cases} 0 & \text{for } \text{AoI}_{\text{Rx}} - \text{AoI}_{\text{Tx}} < n, \\ 1 & \text{for } \text{AoI}_{\text{Rx}} - \text{AoI}_{\text{Tx}} \geq n. \end{cases}$$

The intuitive idea behind the threshold-based strategy is that the AoI_{Rx} is reduced by the difference of both AoIs, which means that a decision to send is more profitable for a higher difference of the AoIs.

We continue with a lemma about the cost $cost(\pi_{TB}(n))$ associated with the strategy $\pi_{TB}(n)$. This lemma is used to find the costwise optimal value for the threshold n and to derive the costwise optimal threshold-based strategy TB-Opt. Afterwards, risk is considered in a second lemma, where we provide a term for the frequency of the appearance of risky states during the strategy's execution. Combining both lemmas, one is able to find a value for the threshold n optimizing the cost under a given risk constraint. To apply the lemmas, the risk constraint has to be given in terms of a maximal frequency for the appearance of risky states.

Lemma 1. *The average cost of the strategy $\pi_{TB}(n)$ is given by*

$$cost(\pi_{TB}(n)) = \frac{\alpha(\sum_{k=1}^{n} p_k(a(n) - \frac{k(k-1)}{2}) + \sum_{k=n+1}^{\infty} p_k(k + a(1))) + \frac{\beta\cdot\nu}{p}}{\frac{1-p}{p} + 1 + \frac{1-\lambda}{\lambda} + \sum_{k=1}^{n-1} p_k \cdot (n-k)},$$

$$a(n) := \frac{n-2}{\lambda} + \frac{n-2}{p} + \frac{(n-2)(n-1)}{2} + \frac{1}{\lambda^2} + \frac{1}{\lambda p} + \frac{1}{p^2},$$

$$p_k := (1-\lambda)^{k-1}(1-p)^{k-1} - (1-\lambda)^k(1-p)^k.$$

Proof. The average cost $cost(\pi_{TB}(n))$ of $\pi_{TB}(n)$ depends on the required transmission energy ν and the AoI_{Rx} at every time step. To find $cost(\pi_{TB}(n))$, we consider the periods between two successful transmissions. These periods have an average period length $l \in \mathbb{R}$, which is measured in time steps. During a period, the strategy chooses to send for $m \in \mathbb{R}$ times on average. The average energy cost can then be written as $e := \nu \cdot m \cdot l^{-1}$. Finding the average sum A per period of the AoI_{Rx} weighted by the period length allows to write the AoI-cost as $A \cdot l^{-1}$. Combining AoI-cost and e, we get

$$cost(\pi_{TB}(n)) = \alpha(l^{-1}\cdot A) + \beta e = l^{-1}\cdot(\alpha\cdot A + \beta\cdot\nu\cdot m). \tag{7}$$

It remains to find l, m and A.

The number of average sending attempts per period m is

$$m = p\sum_{i=0}^{\infty}(i+1)(1-p)^i = \frac{1}{p}. \tag{8}$$

The average length l of a period depends on the value r_0 of AoI_{Rx} at the period's first time step. The probability for r_0 to be 1 is given by p_1, i.e., $\mathbb{P}(r_0 = 1) = p_1$. This results from the fact that r_0 will only become larger than 1, if the last period ended with at least one failed attempt to transmit. Also, during this last time step of the last period, there must not arrive a new status update. Generalizing the case $r_0 = 1$ to $r_0 = k$ for arbitrary k results in $\mathbb{P}(r_0 = k) = p_k$. If in a given period $r_0 \geq n$, the sender will decide to send as soon as a new status update arrives. This results in an average period length of

$$\begin{aligned} l[r_0 \geq n] &= \lambda \sum_{j=0}^{\infty} (1-\lambda)^j p \sum_{i=0}^{\infty} (1-\lambda)^i (i+j+1) \\ &= \frac{1-p}{p} + 1 + \frac{1-\lambda}{\lambda}. \end{aligned}$$

Otherwise, the sender will wait for $(n - r_0)$ time steps before sending newly arrived updates:

$$l[r_0 < n] = (n - r_0) + l[r_0 \geq n].$$

Including the relevant probabilities results in

$$\begin{aligned} l &= \sum_{k=1}^{\infty} p_k \cdot l[r_0 = k] \\ &= \frac{1-p}{p} + 1 + \frac{1-\lambda}{\lambda} + \sum_{k=1}^{n-1} p_k \cdot (n-k). \end{aligned} \tag{9}$$

Remember A is the average sum per period of the AoI at the receiver weighted by the period length. Now, we evaluate A for two cases: $r_0 \leq n$ and $r_0 > n$. For $r_0 \leq n$ we get

$$\begin{aligned} A[r_0 \leq n] = \lambda \sum_{j=0}^{\infty} (1-\lambda)^j p \sum_{i=0}^{\infty} ((1-\lambda)^i \\ \cdot \left(\frac{(n+i+j)(n+i+j+1)}{2} - \frac{r_0(r_0-1)}{2}\right)) \\ = \frac{n-2}{\lambda} + \frac{n-2}{p} + \frac{(n-2)(n-1)}{2} \\ + \frac{1}{\lambda^2} + \frac{1}{\lambda p} + \frac{1}{p^2} - \frac{r_0(r_0-1)}{2}. \end{aligned}$$

Note that $a(n) = A[r_0 = 1]$. For $r_0 > n$ we then get

$$A[r_0 > n] = r_0 + A[r_0 = 1] = r_0 + a(n).$$

As for l, we include the relevant probabilities to obtain

$$A = \sum_{k=1}^{n} p_k (a(n) - \frac{k(k-1)}{2}) + \sum_{k=n+1}^{\infty} p_k (k + a(1)). \tag{10}$$

Combining A from (10), l from (9) and m from (8) as indicated in (7) yields the result of Lemma 1. □

To find the threshold for TB-Opt, the resulting term for $cost(\pi_{TB}(n))$ from Lemma 1 can be easily minimized in n. This is because the corresponding function in n is convex in the considered parameter space.

Note that TB-Opt is designed in a risk-neutral manner. Risk is now included into a threshold-based strategy by using a lower threshold than the costwise optimal threshold. To this end, we provide an expression for the frequency of the appearance of *risky states* with high AoI_{Rx} during the strategy's execution. Given a risk constraint in terms of a maximal frequency for risky states, this expression can be used to find a sufficiently low threshold. Conversely, given a threshold, this expression can be used to quantify the arising risk. To make the notion of a frequency precise, we use the following definition.

Definition 2. *For a sequence of random variables $(s_i)_{i=1,2,...}$, the frequency f_A of an event A is defined as*

$$f_A := \mathbb{E}[\lim_{m \to \infty} \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{s_i \in A}].$$

For the threshold based strategy $\pi_{TB}(n)$, risky states with $\text{AoI}_{\text{Rx}} = k \geq \zeta$ appear with the frequency f_k given by the following lemma. Note that the lemma holds for all $\text{AoI}_{\text{Rx}} = k \geq n$, where n is the strategy's transmission threshold.

Lemma 3. *For the strategy $\pi_{TB}(n)$, the frequency f_k of an AoI at the receiver of $k > n$ is given by*

$$f_k = \frac{p_k + P_n \cdot (\sum_{r_0=1}^{n} p_{r_0}) + \sum_{r_0=n+1}^{k-1} p_{r_0} P_{r_0}}{l},$$

where

$$P_{r_0} := 1 - p\lambda \sum_{j=0}^{k-r_0-1} \sum_{j=0}^{i} (1-\lambda)^j (1-p)^{i-j}$$

and l is as in Eq. (9).

Proof. As in the proof of Lemma 1, we will use the concept of periods. A period ranges from one successful transmission to the next. The average length of a period is given by l as found in Eq. (9). Next, we want to find the probability P_{r_0} that the AoI_{Rx} will be equal to k at some time step in a given period. Note that in every period, the event $\text{AoI}_{\text{Rx}} = k$ will appear at most once. Whether the event $\text{AoI}_{\text{Rx}} = k$ appears depends on the first value r_0 of AoI_{Rx} in the respective period. If $r_0 > k$, AoI_{Rx} will not take the value k in this period ($P_{r_0} = 0$). If $r_0 = k$ in the first time step of a period, the event $\text{AoI}_{\text{Rx}} = k$ appears in this period ($P_{r_0} = 1$). If $r_0 < n$, the sender waits until $\text{AoI}_{\text{Rx}} = n$, which means that the probability for an AoI_{Rx} of k in periods with $r_0 < n$ is the same as in periods starting with an AoI_{Rx} of $r_0 = n$ ($P_{r_0} = P_n$). Then, by using p_{r_0} from the previous proof, we get that

$$f_k = \frac{1}{l} \sum_{r_0=1}^{\infty} p_{r_0} P_{r_0} = \frac{1}{l} (p_k + P_n \cdot (\sum_{r_0=1}^{n} p_{r_0}) + \sum_{r_0=n+1}^{k-1} p_{r_0} P_{r_0}).$$

We still need to find P_{r_0} for $r_0 \in \{n, ..., k-1\}$. In a given period starting with an AoI_{Rx} of $r_0 \in \{n, ..., k-1\}$, k will appear if and only if it takes at least $k - r_0$ time steps until the next successful transmission. We will now find Σ_{r_0}, which is the sum of all the probabilities for faster successful transmissions. Subtracting Σ_{r_0} from 1 results in P_{r_0}.

Algorithm 1: Q-learning + risky states (Q+RS)

Data: simulator for $\mathcal{M}$, starting state s_0, no. of time steps N, actions $a_1, ..., a_k$, real learning rates $(\alpha_i)_{i\in\{1,...,n\}}$, discount factor γ, initial ϵ, decay factor δ, set of risky states $\mathcal{R}$, risk-factor ρ

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

1 $Q^{(0)} \leftarrow (0, ..., 0)$
2 $s_t \leftarrow s_0$
3 **for** *i=1,...,N* **do**
4 sample a random action a ϵ-greedy
5 update ϵ as $\epsilon \leftarrow \delta \cdot \epsilon$
6 sample next state s_{t+1} and cost C_{t+1} using the simulator for $\mathcal{M}$
7 **if** $s_{t+1} \in \mathcal{R}$ **then**
8 $C_{t+1} \leftarrow \rho \cdot C_{t+1}$
9 **end**
10 **for** $(s', a') \in \mathcal{S} \times \mathcal{A}$ **do**
11 **if** $s' = s_t$ & $a' = a$ **then**
12 $V(s_{t+1}) \leftarrow \max_{a_j=a_1,...,a_k} Q^{(i-1)}(s_{t+1}, a_j)$
$Q^{(i)}(s', a') \leftarrow$
$(1-\alpha_i)Q^{(i-1)}(s', a') + \alpha_i(C_{t+1} + \gamma V(s_{t+1}))$
13 **else**
14 $Q^{(i)}(s', a') \leftarrow Q^{(i-1)}(s', a')$
15 **end**
16 **end**
17 $s_t \leftarrow s_{t+1}$
18 **end**
19 **return** $Q^{(N)}$

As $r_0 \geq n$, the sender chooses to send immediately as soon as a new status update arrives. The necessity for a new status update results in a factor λ in Σ_{r_0}. The transmission is successful with a probability of p, which is the second necessary factor for Σ_{r_0}. In the remaining $k - r_0 - 1$ time steps, the sender could either wait for the new update, resulting in an additional factor $(1 - \lambda)$, or fail to send, resulting in an additional factor $(1 - p)$. Adding all possible sequences of waiting and failing resulting in successful transmissions before AoI_{Rx} reaches k, we get

$$\Sigma_{r_0} = p\lambda \sum_{j=0}^{k-r_0-1} \sum_{j=0}^{i} (1-\lambda)^j (1-p)^{i-j}.$$

Subtracting Σ_{r_0} from 1 results in P_{r_0} as in Lemma 3. □

V. The Q-Learning Based Approach

In this section, we present the risk-sensitive learning algorithm Q+RS, which combines Q-learning and the notion of *risky states*. This is achieved by modifying the costs associated with each time step by adding a penalty for risky states. Q+RS does not depend on a-priori knowledge of the system parameters. Q+RS is also not limited to small sets of possible AoIs as the value iteration approach in [12], because in contrast to value iteration, the number of performed machine operations does not grow in the size of the state space. We apply ϵ-greedy tabular Q-learning to the MDP in Sec. III.

The pseudo code for Q+RS is given in Algorithm 1. The algorithm iteratively approximates the so-called Q-value of each state-action pair, i.e., the pair's expected future cost. The resulting approximations after N iterations are called $Q^{(N)}$-values. The initial approximations $Q^{(0)}$ are set to be 0. After initial operations (lines 1-2), the algorithm works in an iterative fashion (l. 3-18). The ϵ-greedy strategy used during learning chooses a random action with a probability of ϵ and the action with the lowest estimated Q-value with a probability of $(1 - \epsilon)$ (l. 4). During learning, ϵ is reduced by multiplying it by a decay factor $\delta \in (0, 1)$ after every iteration (l. 5). The Q-value update from traditional Q-learning (l. 10-16) is used with an additional manipulation of the time step's cost C_{t+1} (l. 7-9) in case of risky states. To weight current and future costs (see l. 12), Q-learning uses a discount factor γ, which we here introduce as a hyperparameter. From the resulting $Q^{(N)}$-values, a strategy is constructed by choosing the action with the lowest $Q^{(N)}$-value in each state.

Q-learning in its original form is risk-neutral in the sense that it optimizes costs in the MDP without considering any risk-measure. In contrast to this original form of Q-learning, we include risk-sensitivity by modifying the cost function c in the MDP. This approach trivially inherits all convergence properties of Q-learning.

The usage of a modified cost function c can be naturally combined with the notion of *risky states*. This is achieved by multiplying costs for transitions to *risky states* by a risk factor $\rho > 1$. $\rho = 1$ would result in the original MDP, while $\rho < 1$ would result in risk-seeking strategies. The modified cost function as implemented in Algorithm 1 is defined as

$$c_{\mathcal{R}}(s, a, s') := (\mathbb{1}_{s' \notin \mathcal{R}} + \rho \cdot \mathbb{1}_{s' \in \mathcal{R}}) \cdot c(s, a, s'). \tag{11}$$

VI. Simulation Results

A. Reference schemes

This section contains numerical results for the evaluation of the proposed optimal threshold-based strategy TB-Opt as well as of Q-learning using risky states Q+RS. We compare our results with two reference schemes. The first reference is a random strategy choosing to wait or to send both with a probability of 0.5 and independently of the current state. As a second reference, we use traditional risk-neutral Q-learning. We omit the value iteration approach introduced in [12] in the comparison, as otherwise, it would have become necessary to introduce a limit for the AoI. This limit could be chosen large enough to not influence the results, e.g. if it is never reached by the AoI during the simulations. However, this is not feasible due to the high computational complexity of value iteration.

B. Simulation setup

To simulate the system, we fix the parameters for transmission energy $\nu := 1$ and the channel's successful transmission probability $p := 0.9$. The weights in the cost function are set to $\alpha = 1$ and $\beta = 3$. These weights are chosen, such that the costs arising from a transmission attempt are high enough that it is costwise reasonable for the sender to decide for the wait action in some time steps. The default update arrival probability is set to $\lambda = 0.5$. λ is varied in one of the experiments to values between $\lambda = 0.1$ and $\lambda = 0.9$. For Q-learning based strategies, we use $N := 100.000$ time steps for learning and a discount factor $\gamma := 0.7$. As risk-factor, we

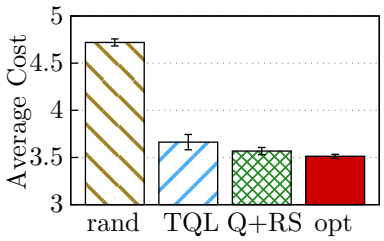

Fig. 2: Avg costs with std deviation error bars using the random strategy, tabular Q-learning, Q+RS and TB-Opt

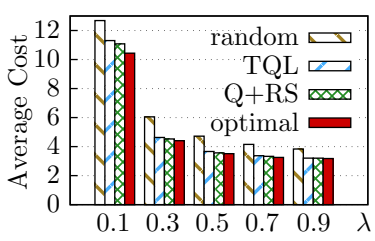

Fig. 3: Avg costs for different λ using the random strategy, tabular Q-learning, Q+RS and TB-Opt

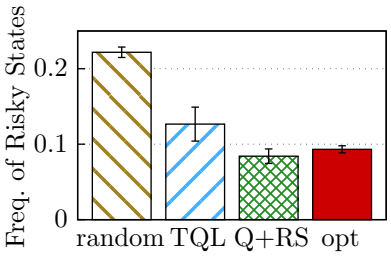

Fig. 4: Avg frequency of risky states visited by the random strategy, tabular Q-learning, Q+RS and TB-Opt

choose $\rho = 2$ and as risk-threshold, we use $\zeta = 5$. Initially, $\epsilon = 0.9$, the decay factor is set to $\delta = 0.999$.

For the experiments displayed in Fig. 2 to 4, we take the average of 100 independent runs. In each run, we first train the Q-learning based approaches. We then use the resulting learned strategies and compare them with the reference schemes. In each run, we use 10.000 time steps per strategy for testing.

C. Numerical results

Figure 2 shows the average cost of our proposed strategies Q+RS and TB-Opt compared to the reference schemes. Error bars indicate standard deviations of the outcomes. While using knowledge of the system parameters, the optimal threshold-based strategy TB-Opt outperforms all other strategies and has the lowest standard deviation, but at the price of requiring a-priori knowledge of the parameters. Both Q-learning based strategies are able to perform close to TB-Opt in comparison to the random reference strategy. Risk-neutral Q-learning generates average costs 4.3% higher than that of the optimal threshold-based strategy. The strategy derived from Q+RS generates costs only 1.6% higher than that of TB-Opt and has a by 55% lower standard deviation than unmodified Q-learning. Comparing the costs of our strategies to traditional Q-learning, Q+RS is reducing the average cost by 2.6%, while TB-Opt is reducing it by 4.1%.

In Figure 3, we show the average cost for different update arrival probabilities ranging from $\lambda = 0.1$ to $\lambda = 0.9$. For greater λ, average costs are smaller due to smaller AoI costs. Our proposed strategies consistently outperform the reference schemes as they do for $\lambda = 0.5$.

Figure 4 shows the average frequency of the appearance of *risky states* for our approaches and the reference strategies. The strategy derived from Q+RS avoids those states actively and hence has a low frequency of 8.4% compared to 22.2% in the random case, 12.7% for risk-neutral Q-learning and 9.3% for TB-Opt. Note that the optimal threshold-based strategy TB-Opt is not optimizing the frequency of the appearance of risky states. Although TB-Opt was designed to minimize costs, it still is risk-sensitive, as it visits a low number of risky states compared to traditional Q-learning or the random strategy.

VII. Conclusions

In this work, we derive risk-sensitive strategies for a point-to-point wireless communication scenario with randomly arriving status updates. We measure risk using the notion of *risky states*. We first propose a threshold-based strategy and use offline optimization to find the costwise optimal threshold and derive a costwise optimal threshold-based strategy TB-Opt. By lowering this optimal threshold, the frequency of visited risky states decreases, leading to a risk-sensitive strategy. We provide expressions for cost and risk of the threshold-based strategy. Our second proposed strategy is based on the modified Q-learning algorithm Q+RS, where we add risk penalties to the cost function.

In simulations, we show that both of our proposed strategies outperform the reference schemes costwise and riskwise. Also, the standard deviations of the experiment outcomes are lower for our proposed strategies.

References

[1] 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] Kaul *et al.*, "Real-time status: How often should one update?" in *IEEE Int. Conf. Computer Commun. (Infocom)*, 2012, pp. 2731–2735.

[3] 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.

[4] Hu *et al.*, "Status update in IoT networks: Age-of-information violation probability and optimal update rate," *IEEE Internet Things J.*, vol. 8, no. 14, pp. 11 329–11 344, 2021.

[5] 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.

[6] Y. Wang *et al.*, "Age-optimal transmission policy for Markov source with differential encoding," in *IEEE Global Commun. Conf.*, 2020, pp. 1–6.

[7] B. Wang *et al.*, "When to preempt? Age of information minimization under link capacity constraint," *J. Commun. Networks*, vol. 21, no. 3, pp. 220–232, 2019.

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

[9] Devassy *et al.*, "Reliable transmission of short packets through queues and noisy channels under latency and peak-age violation guarantees," *IEEE J. Sel. A. Commun.*, vol. 37, no. 4, pp. 721–734, 2019.

[10] Song *et al.*, "Analysis of AoI violation probability in wireless networks," in *Int. Symp. Wireless Commun. Syst. (ISWCS)*, 2021, pp. 1–6.

[11] Hu *et al.*, "Asymptotically optimal arrival rate for IoT networks with AoI and peak AoI constraints," *IEEE Commun. Lett.*, vol. 25, no. 12, pp. 3853–3857, 2021.

[12] 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.

[13] Kiekenap *et al.*, "Energy-optimal short packet transmission for time-critical control," in *IEEE Veh. Tech. Conf. (VTC-Fall)*, 2021, pp. 01–06.