entropy

MDPI

Article

A UoI-Optimal Policy for Timely Status Updates with Resource Constraint

Lehan Wang, Jingzhou Sun, Yuxuan Sun, Sheng Zhou * and Zhisheng Niu

Beijing National Research Center for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China; wang-lh19@mails.tsinghua.edu.cn (L.W.); sunjz18@mails.tsinghua.edu.cn (J.S.); sunyuxuan@tsinghua.edu.cn (Y.S.); niuzhs@tsinghua.edu.cn (Z.N.)

* Correspondence: sheng.zhou@tsinghua.edu.cn

Abstract: Timely status updates are critical in remote control systems such as autonomous driving and the industrial Internet of Things, where timeliness requirements are usually context dependent. Accordingly, the Urgency of Information (UoI) has been proposed beyond the well-known Age of Information (AoI) by further including context-aware weights which indicate whether the monitored process is in an emergency. However, the optimal updating and scheduling strategies in terms of UoI remain open. In this paper, we propose a UoI-optimal updating policy for timely status information with resource constraint. We first formulate the problem in a constrained Markov decision process and prove that the UoI-optimal policy has a threshold structure. When the context-aware weights are known, we propose a numerical method based on linear programming. When the weights are unknown, we further design a reinforcement learning (RL)-based scheduling policy. The simulation reveals that the threshold of the UoI-optimal policy increases as the resource constraint tightens. In addition, the UoI-optimal policy outperforms the AoI-optimal policy in terms of average squared estimation error, and the proposed RL-based updating policy achieves a near-optimal performance without the advanced knowledge of the system model.

Keywords: age of information; constrained Markov decision process; reinforcement learning; context-awareness; timely status updates

Citation: Wang, L.; Sun, J.; Sun, Y.; Zhou, S.; Niu, Z. A UoI-Optimal Updating Policy for Timely Status Information with Resource Constraint. *Entropy* **2021**, *23*, 1084. https://doi.org/10.3390/e23081084

Academic Editors: Yin Sun and Anthony Ephremides

Received: 20 June 2021
Accepted: 16 August 2021
Published: 20 August 2021

1. Introduction

With the development of 5G and the Internet of Things (IoT), requirements for wireless communication have shifted from merely providing communication channels to covering the entire process of various IoT applications, e.g., autonomous vehicle [1] and virtual reality (VR) [2], where sensing, communication, computation, and control form a closed loop. Therefore, in addition to the communication delay, it is necessary to consider the information delay counted from the generation of the state information to the execution, namely the timeliness of information. For this purpose, Age of Information (AoI) has been proposed, which is defined as the time elapsed since the generation time of the latest received packets [3]. Due to its concise definition and clear physical meaning, AoI has been widely used for the design of scheduling and updating policies in remote estimation [4–6] and wireless communication networks [7–12]. Most existing works focus on optimizing average AoI or peak age. In [13], the authors claim that minimizing average age cannot satisfy the requirements for ultra-reliable low-latency communication (URLLC) and study the tail distribution of AoI. The violation probability for peak age is derived in [14] and the stationary distribution of AoI is studied in [15].

Nevertheless, the AoI still has some limitations. First, it fails to measure the nonlinear performance degradation caused by information staleness. In [16–19], nonlinear age penalty functions were introduced to solve this problem. Meanwhile, the Age of Synchronization (AoS) [20] and Age of Incorrect Information (AoII) [21] are defined to associate information freshness with the content of information. AoS is the time elapsed since the information at

the receiver becomes desynchronized with the actual status of the monitored process. AoII is defined as the product of an increasing time penalty function and a penalty function of the estimation error. In addition, the status of heterogeneous data sources may change at different rates. A fast-changing process may require information with a lower age. However, age is independent of the changing rate and thus is not proper in the cases when heterogeneous data sources are jointly considered. To solve this problem, weighted age was introduced in [22,23] to distinguish important monitored processes. In [24], the metric based on information theory is proposed as a replacement of the time-based metric, AoI, to characterize the changing rate. In [5], the authors claim that minimizing age is not equivalent to minimizing the estimation error in a remote estimation problem and propose an effective age to solve this problem [25].

Practical systems (e.g., V2X-communication systems) may have different requirements for information freshness with different contexts. The context refers to all environmental factors that affect the requirement for information freshness. Therefore, resources should be reserved for frequent status updates in emergency to ensure safety.

However, the timeliness metrics mentioned above pay no attention to the significance of context information. To solve this problem, Urgency of Information (UoI) has been proposed in [26–28] to measure the influence of inaccurate information on performance under different contexts. To be specific, UoI uses a time-variant context-aware weight $\omega(t)$ to distinguish different contexts. A higher $\omega(t)$ indicates that the system is in more urgent situations (e.g., when a vehicle is approaching an intersection or overtaking) and therefore requires frequent updates. For example, when a vehicle passes through an intersection, the context-aware weight increases as the distance between the vehicle and the center of the intersection decreases. Meanwhile, the estimation error $Q(t)$ is introduced to measure the information inaccuracy, which is defined as the difference between the actual status and the estimated status at the receiver. The larger the absolute value of $Q(t)$ is, the less accurate the estimated status is. Therefore, UoI is defined as the product of context-aware weight and a cost function of the estimation error $Q(t)$:

$$F(t) = \omega(t)\delta(Q(t)). \tag{1}$$

In discrete-time systems, the estimation error $Q(t)$ is:

$$Q(t) = \sum_{\tau=g(t)}^{t-1} A(\tau), \tag{2}$$

where $g(t)$ is the generation time of the latest status update at the receiver and $A(t)$ is the increment in estimation error in time slot t. Specifically, if the context-aware weight is time-invariant (i.e., $\omega(t) = 1$), and $A(t) = 1$ as well as $\delta(Q(t)) = Q(t)$, UoI is the same as AoI. If the context-aware weight is process-dependent, UoI can represent weighted age. If the cost function $\delta(Q(t))$ is nonlinear, UoI can represent the nonlinear age penalty function. For example, when the outdated information is worthless, e.g., the information is about sales that expire after some time [29], then the shifted unit step cost function $\delta(Q(t)) = u(Q(t) - \tau)$, $\tau > 0$ is recommended. For the unit step function, $u(x) = 1$ when $x \geq 0$ and otherwise $u(x) = 0$.

In this work, we considered a single-user remote monitoring system, and the objective was to find an updating policy minimizing the average UoI over time under the constraint on average update frequency. To solve this problem, Refs. [27,30] proposed update-index-based adaptive schemes with Lyapunov optimization but did not conduct a theoretical analysis of their optimality. In addition, the constrained Markov decision process (CMDP) formulation was only used in the simulation for a numerically solved benchmark. Based on the existing works, in this paper, we theoretically analyzed the structure of the UoI-optimal policy and focused on how to derive an updating policy in an unknown environment.

The main contributions of this paper are summarized as follows.

- In contrast to [27,30], we assumed that the context-aware weight is a first-order irreducible positive recurrent Markov process or independent and identically distributed (i.i.d.) over time. We formulated the updating problem as a CMDP problem and proved the single threshold structure of the UoI-optimal policy. We then derived the policy through LP with the threshold structure and discussed the conditions that the monitored process needs to satisfy for the threshold structure.
- When the distributions of the context-aware weight and the increment in estimation error were unknown, we used model-based RL method to learn the state transitions of the whole system and derive a near-optimal RL-based updating policy.
- Simulations were conducted to verify the theoretical analysis of the threshold structure and show the near-optimal performance of the RL-based updating policy. The results indicate that: (i) the update thresholds decrease when the maximum average update frequency becomes large; (ii) the update threshold for emergency can actually be larger than that for ordinary states when the probability of transferring from emergency to ordinary states tends to 1.

The rest of this paper is organized as follows. The system model and the problem formulation are described in Section 2. In Section 3, we obtain the CMDP formulation of the problem with the given distribution of context-aware weight and prove the threshold structure of the UoI-optimal policy. The proposed model-based RL updating policy is obtained in Section 4. In Section 5, the simulation results are shown and discussed while the conclusions are drawn in Section 6.

2. System Model and Problem Formulation

In this paper, we considered a remote monitoring system, in which a fusion center collects the status information (e.g., current location, velocity, information of surrounding) from a vehicle of interest via a wireless channel with limited resources, as shown in Figure 1. The whole system is considered as a discrete-time system and the status can be generated at will. Due to the limitations on the wireless resources and energy supply, there is a constraint on the average update frequency of the vehicle. The update decision in time slot t is denoted by $U(t) \in \{0,1\}$, where $U(t) = 1$ means that the vehicle decides to transmit the current status to the center, and $U(t) = 0$ denotes that the vehicle decides to stay idle.

The wireless channel is assumed as a block fading channel with successful transmission probability p_s. Let $S(t) \in \{0,1\}$ be the state of the channel. $S(t) = 0$ represents that the channel is in deep fading, and no packet can be successfully transmitted. $S(t) = 1$ means the packets can be successfully transmitted to the center through the channel. If the center receives an update, then $U(t)S(t) = 1$ and an ACK will be sent to the vehicle.

Let $x(t)$ and $\hat{x}(t)$ denote the current status of the monitored vehicle and the estimated status of the vehicle at the center, and $Q(t) = x(t) - \hat{x}(t)$ denotes the estimation error. Similar to [26], we further assume that the time period of a packet transmission is less than a time slot and the estimation at the center equals the latest status information received by the center. This estimation scheme is easy to implement, theoretically tractable and has been proven to be an optimal policy that can minimize the average squared error of status estimation in a remote estimation system under energy constraints when the monitored process is a Wiener process [31]. Then, the recurrence relation of the estimation error $Q(t)$ is:

$$Q(t+1) = (1 - U(t)S(t))Q(t) + A(t). \tag{3}$$

Equation (3) indicates that the estimation error will be the amount of variation of the monitored process from the generation time of the latest received status to the current time. The increment $A(t)$ represents the variation of the monitored process. For example, when $A(t)$ follows a Gaussian distribution with a mean of zero and variance of σ^2, represented by $N(0, \sigma^2)$, the monitored status follows a Wiener process. When $A(t)$ takes values from $\{0, 1, -1\}$ with a probability of $\{1 - 2p_{rw}, p_{rw}, p_{rw}\}$, where $0 < p_{rw} < \frac{1}{2}$, then the status of the monitored source will be a one-dimensional random walk. In this paper, we

assumed that the monitored status of the vehicle is a Wiener process and $A(t)$ is i.i.d. over time. However, the increment in estimation error during a single slot cannot be infinite in practical systems. Therefore, in contrast to [27,30], we assumed that increment $A(t)$ obeys a truncated Gaussian distribution, i.e., the probability density function (PDF) of $A(t)$ is:

$$f_{A(t)}(a) = \frac{\frac{1}{\sigma}\phi\left(\frac{a-\mu}{\sigma}\right)}{\Phi\left(\frac{A_{max}-\mu}{\sigma}\right) - \Phi\left(\frac{-A_{min}-\mu}{\sigma}\right)}, \tag{4}$$

where μ and σ are the expectation and standard deviation of increment $A(t)$. ϕ and Φ are the PDF and the cumulative distribution function (CDF) of standard normal distribution. We also assumed $A(t) \in [-A_{min}, A_{max}]$, $A_{max} = A_{min} > 0$ and $\mu = 0$.

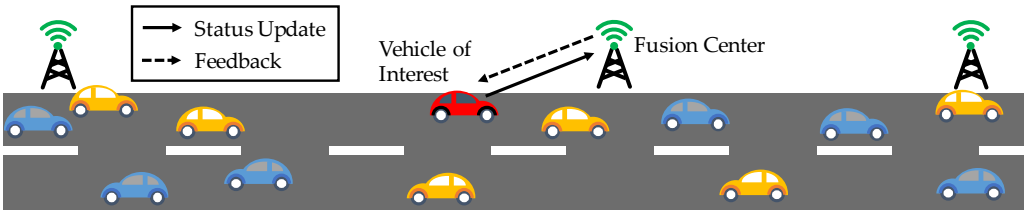

Figure 1. Remote control and monitoring model. The vehicle of interest is shown in red.

Meanwhile, the scheduling policy of information updates should also be related to the situation and environment of the system. For example, when the system is in an emergency, it should be very sensitive to the accuracy and the delay of the status information, thus the status should be updated more frequently. Therefore, our objective is to find a policy telling the vehicle whether to transmit status information or not in each slot for a minimum average UoI over time under the constraint:

$$\begin{aligned} &\min_{U(t)} \limsup_{T\to\infty} \frac{1}{T} E\left[\sum_{t=0}^{T-1} w(t)Q(t)^2\right] \\ &s.t. \limsup_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} E[U(t)] \le \rho, \end{aligned} \tag{5}$$

where $\omega(t) > 0$ is the context-aware weight, which is independent with $Q(t)$. $\rho \in (0, 1]$ is the maximum average update frequency. The cost function of the estimation error used here is $\delta(Q(t)) = (Q(t))^2$, which is inspired by the squared error of status estimation.

3. Scheduling with CMDP-Based Approach

In this section, we start by formulating problem (5) into a constrained Markov decision process (CMDP) with assumptions on the distribution of the context-aware weight. We will prove the threshold structure of the UoI-optimal updating policy and derive the optimal policy through a linear programming (LP) formulation.

3.1. Constrained Markov Decision Process Formulation

In the remote monitoring system, the context may be related to the distance between adjacent vehicles/mobile devices, the unexpected maneuver of the neighboring vehicles, etc. In [32], the authors prove that whether the distance between two mobile wireless devices with Ornstein–Uhlenbeck mobility is less than a certain threshold follows a first-order Markov process. When the two devices are closer, they are more interested in each other's status information, communication and computing resources to facilitate cooperation, share resources, and avoid collisions. At this time, the transmission of status information is more urgent than when the two devices are far apart. As for the unexpected maneuver of the neighboring vehicles, it is very challenging to find a proper formulation. Instead, we assumed that such emergencies occur independently in each slot according

to a certain probability. Therefore, in contrast to [27,30], we assumed that the context-aware weight $\omega(t)$ is i.i.d. over time or a first-order irreducible positive recurrent Markov process and formulated the problem (5) as a CMDP problem. The irreducible positive recurrent Markov formulation guarantees the existence of the UoI-optimal policy (see Appendix A). In this section, we will first focus on the situation where $\omega(t)$ is a first-order Markov process:

- State space: The state of the vehicle in slot t, denoted by $s(t) = (Q(t), \omega(t))$, includes the current estimation error and the context-aware weight. Then, we discretize $Q(t)$ with the step size $\Delta_Q > 0$, i.e., the estimation error $Q(t) \in \mathbb{Q} = \{0, \pm\Delta_Q, \pm 2\Delta_Q, \cdots, \pm n\Delta_Q, \cdots\}$. For example, when $Q(t) \in [n\Delta_Q - \frac{1}{2}\Delta_Q, n\Delta_Q + \frac{1}{2}\Delta_Q)$, its value will be taken as $n\Delta_Q$. The smaller the step size Δ_Q, the smaller the performance degradation caused by discretization. In addition, the value set of the context-aware weight is denoted by $\mathbb{W}$. Then, the state space $\mathbb{S} = \{\mathbb{Q} \times \mathbb{W}\}$ is thus countable but infinite.
- Action space: At each slot, the vehicle can take two actions, namely $U(t) \in \mathbb{U} = \{0, 1\}$, where $U(t) = 1$ denotes the vehicle deciding to transmit updates in slot t and $U(t) = 0$ denotes the vehicle deciding to wait.
- Probability transfer function: After taking action U at state $s = (Q, \omega)$, the next state is denoted by $s' = (Q', \omega')$. When the vehicle decides not to transmit or the transmission fails, the probability of the estimation error transferring from Q to Q' is written as $\Pr\{Q' - Q = a\} = p_a$. Due to the discretization of the estimation error, the increment $a \in \mathbb{A} = \{0, \pm\Delta_Q, \pm 2\Delta_Q, \cdots, \pm A_m\}$, where $A_m = \lfloor \frac{A_{max}}{\Delta_Q} \rfloor \Delta_Q > 0$. In addition, $p_a = F_A(a + \frac{1}{2}\Delta_Q) - F_A(a - \frac{1}{2}\Delta_Q)$, where $F_A(a)$ is the CDF of increment $A(t)$. In addition, the probability of the context-aware weight transferring from ω to ω' is written as $\Pr\{\omega \to \omega'\} = p_{\omega\omega'}$. Based on the assumption that the context-aware weight $\omega(t)$ is independent with the estimation error $Q(t)$, then the probability of the state transferring from $s = (Q, \omega)$ to $s' = (Q', \omega')$ given action U is:

$$\begin{aligned} &\Pr\{s \to s'|U\} = \Pr\{(Q, \omega) \to (Q', \omega')|U\} \\ &= \begin{cases} p_{\omega\omega'} p_{Q'-Q} & , U = 0, \\ p_{\omega\omega'}((1 - p_s)p_{Q'-Q} + p_s p_{Q'-0}) & , U = 1. \end{cases} \end{aligned} \tag{6}$$

- One-step cost: The cost caused by taking action U in state (Q, ω) is:

$$C(Q, \omega, U) = \omega Q^2, \tag{7}$$

while the one-step updating penalty only depends on the chosen action:

$$D(Q, \omega, U) = U. \tag{8}$$

The average cost caused under a certain policy π is the average UoI, which is defined as $\bar{C}^\pi$ and the average updating penalty under π is defined as $\bar{D}^\pi$. We aimed to find the UoI-optimal policy which minimizes the average cost under the resources constraint. Therefore, problem (5) can be formulated into the following CMDP problem:

$$\begin{aligned} &\min_\pi \bar{C}^\pi = \lim_{T\to\infty} \frac{1}{T} \mathbb{E}_\pi \left[\sum_{t=1}^{T} C(Q(t), \omega(t), U(t)) \right] \\ &s.t.\ \bar{D}^\pi = \lim_{T\to\infty} \frac{1}{T} \mathbb{E}_\pi \left[\sum_{t=1}^{T} D(Q(t), \omega(t), U(t)) \right] \leq \rho. \end{aligned} \tag{9}$$

3.2. Threshold Structure of the Optimal Policy

We start from some basic definitions in [33] and show the properties of problem (9).

Definition 1. *A stationary deterministic policy is a policy that takes the same action whenever in a given state $s = (Q, \omega)$, while a stationary randomized policy chooses to update or not in state s with a certain probability.*

Theorem 1. *There exists an optimal stationary randomized policy for problem (9). The optimal policy is a probabilistic combination of two stationary deterministic policies. The two deterministic policies only differ on at most one state and each policy minimizes the unconstrained cost in (10) with a different Lagrange multiplier λ:*

$$L_{\lambda}^{\pi} = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}_{\pi}\left[\sum_{t=1}^{T}[C(Q(t),\omega(t),U(t)) + \lambda D(Q(t),\omega(t),U(t))]\right]. \tag{10}$$

Proof of Theorem 1. The proof is shown in Appendix A. □

We denote the optimal policy that minimizes the unconstrained cost in (10) with a given λ by $\pi^{\star}$ and the cost obtained under policy $\pi^{\star}$ by $L_{\lambda}^{\pi^{\star}}$, namely $L_{\lambda}^{\pi^{\star}} = \min_{\pi} L_{\lambda}^{\pi}$. Then, there exists a differential cost function $V(Q,\omega)$ that satisfies the Bellman Equation [34]:

$$\begin{aligned} V(Q,\omega) + L_{\lambda}^{\pi^{\star}} = \min\Bigg\{ & C(Q,\omega,1) + \lambda D(Q,\omega,1) \\ & + (1-p_s)\sum_{\omega'\in\mathbb{W}} p_{\omega\omega'} \sum_{a=-A_m}^{A_m} p_a V(Q+a,\omega') + p_s \sum_{\omega'\in\mathbb{W}} p_{\omega\omega'} \sum_{a=-A_m}^{A_m} p_a V(a,\omega'), \\ & C(Q,\omega,0) + \sum_{\omega'\in\mathbb{W}} p_{\omega\omega'} \sum_{a=-A_m}^{A_m} p_a V(Q+a,\omega')\Bigg\}. \end{aligned} \tag{11}$$

To solve problem (5), we first prove that with a given λ, the optimal stationary deterministic policy $\pi^{\star}$ has a threshold structure. We then introduce a discounted problem with a discount factor α and the discounted cost starting from state (Q,ω) under a certain policy π is:

$$\begin{aligned} J_{\alpha,\pi}(Q,\omega) = \lim_{T\to\infty} \mathbb{E}_{\pi}\Bigg[\sum_{t=0}^{T} \alpha^t [C(Q(t),\omega(t),U(t)) \\ + \lambda D(Q(t),\omega(t),U(t))] \mid (Q(0)=Q,\omega(0)=\omega)\Bigg]. \end{aligned} \tag{12}$$

Denote the minimum cost starting from state (Q,ω) by $V_{\alpha}(Q,\omega) = \min_{\pi} J_{\alpha,\pi}(Q,\omega)$. Then, we have:

$$\begin{aligned} V_{\alpha}(Q,\omega) = \min\Bigg\{ & C(Q,\omega,1) + \lambda D(Q,\omega,1) + (1-p_s)\alpha \sum_{\omega'\in\mathbb{W}} p_{\omega\omega'} \sum_{a=-A_m}^{A_m} p_a V_{\alpha}(Q+a,\omega') \\ & + p_s\alpha \sum_{\omega'\in\mathbb{W}} p_{\omega\omega'} \sum_{a=-A_m}^{A_m} p_a V_{\alpha}(a,\omega'), C(Q,\omega,0) \\ & + \alpha \sum_{\omega'\in\mathbb{W}} p_{\omega\omega'} \sum_{a=-A_m}^{A_m} p_a V_{\alpha}(Q+a,\omega')\Bigg\}. \end{aligned} \tag{13}$$

Define $\Delta(Q,\omega)$ as the difference between the value functions by taking the two different actions $U = 0, 1$, meaning that:

$$\begin{aligned}\Delta(Q,\omega) &= C(Q,\omega,0) + \alpha \sum_{\omega' \in \mathbb{W}} p_{\omega\omega'} \sum_{a=-A_m}^{A_m} p_a V_\alpha\big(Q+a,\omega'\big) \\ &- C(Q,\omega,1) - \lambda D(Q,\omega,1) - p_s\alpha \sum_{\omega' \in \mathbb{W}} p_{\omega\omega'} \sum_{a=-A_m}^{A_m} p_a V_\alpha\big(a,\omega'\big) \\ &- (1-p_s)\alpha \sum_{\omega' \in \mathbb{W}} p_{\omega\omega'} \sum_{a=-A_m}^{A_m} p_a V_\alpha\big(Q+a,\omega'\big) \\ &= p_s\alpha \sum_{\omega' \in \mathbb{W}} p_{\omega\omega'} \sum_{a=-A_m}^{A_m} p_a \big\{ V_\alpha\big(Q+a,\omega'\big) - V_\alpha\big(a,\omega'\big) \big\} - \lambda. \end{aligned} \tag{14}$$

Define $\sum_{a=-A_m}^{A_m} p_a V_\alpha(Q+a,\omega)$ as a function $f_\alpha(Q,\omega)$. Then we will prove that for $\forall |Q_1| < |Q_2|$, we have $f_\alpha(Q_1,\omega) < f_\alpha(Q_2,\omega)$. To this end, we first prove the following Lemma 1.

Lemma 1. *For a given discount factor α and a fixed context-aware weight ω, the value function for Q equals the value function for $-Q$, namely:*

$$V_\alpha(Q,\omega) = V_\alpha(-Q,\omega).$$

Proof of Lemma 1. The Lemma is proven by induction. Define $V_\alpha^{(k)}(Q,\omega)$ as the value function obtained after the k^{th} iteration. Assume that for $\forall Q$, we have: $V_\alpha^{(k)}(Q,\omega) = V_\alpha^{(k)}(-Q,\omega)$. If action U is taken in the k^{th} iteration, then the expected discounted cost is defined as $J_{\alpha,U}^{(k)}(Q,\omega)$. Therefore, $V_\alpha^{(k+1)}(Q,\omega) = \min_U J_{\alpha,U}^{(k)}(Q,\omega)$. We have:

$$\begin{aligned} J_{\alpha,0}^{(k)}(Q,\omega) &= C(Q,\omega,0) + \alpha \sum_{\omega' \in \mathbb{W}} p_{\omega\omega'} \sum_{a=-A_m}^{A_m} p_a V_\alpha^{(k)}\big(Q+a,\omega'\big) \\ &= \omega(-Q)^2 + \alpha \sum_{\omega' \in \mathbb{W}} p_{\omega\omega'} \sum_{a=-A_m}^{A_m} p_a V_\alpha^{(k)}\big(-Q-a,\omega'\big) \\ &= C_X(-Q,\omega,0) + \alpha \sum_{\omega' \in \mathbb{W}} p_{\omega\omega'} \sum_{a=-A_m}^{A_m} p_a V_\alpha^{(k)}\big(-Q+a,\omega'\big) = J_{\alpha,0}^{(k)}(-Q,\omega). \end{aligned} \tag{15}$$

Similarly, we can further prove that $J_{\alpha,1}^{(k)}(Q,\omega) = J_{\alpha,1}^{(k)}(-Q,\omega)$. Notice that the value function obtained in $(k+1)^{\text{th}}$ iteration is obtained by: $V_\alpha^{(k+1)}(Q,\omega) = \min_U J_{\alpha,U}^{(k)}(Q,\omega)$, and for any action U, $J_{\alpha,U}^{(k)}(Q,\omega) = J_{\alpha,U}^{(k)}(-Q,\omega)$. Thus, $V_\alpha^{(k+1)}(Q,\omega) = V_\alpha^{(k+1)}(-Q,\omega)$. By letting $k \to \infty$, $V_\alpha^{(k)}(Q,\omega) \to V_\alpha(Q,\omega)$. Hence, $V_\alpha(Q,\omega) = V_\alpha(-Q,\omega)$. □

Lemma 2. *For a given discount factor α and a fixed context-aware weight ω, function $f_\alpha(Q,\omega)$ for Q increases monotonically with the absolute value of Q, namely: for $\forall |Q_1| < |Q_2|$, $f_\alpha(Q_1,\omega) < f_\alpha(Q_2,\omega)$.*

Proof of Lemma 2. Using the induction method, we first assume that for $\forall |Q_1| < |Q_2|$, we have $f_\alpha^{(k)}(Q_1,\omega) < f_\alpha^{(k)}(Q_2,\omega)$. Therefore:

$$\begin{aligned} J_{\alpha,0}^{(k)}(Q_1,\omega) &= C(Q_1,\omega,0) + \alpha \sum_{\omega' \in \mathbb{W}} p_{\omega\omega'} \sum_{a=-A_m}^{A_m} p_a V_\alpha^{(k)}(Q_1+a,\omega') \\ &= \omega Q_1^2 + \alpha \sum_{\omega' \in \mathbb{W}} p_{\omega\omega'} f_\alpha^{(k)}(Q_1,\omega') \\ &< C(Q_2,\omega,0) + \alpha \sum_{\omega' \in \mathbb{W}} p_{\omega\omega'} f_\alpha^{(k)}(Q_2,\omega') \\ &= J_{\alpha,0}^{(k)}(Q_2,\omega). \end{aligned} \tag{16}$$

Similarly, we can obtain $J_{\alpha,1}^{(k)}(Q_1,\omega) < J_{\alpha,1}^{(k)}(Q_2,\omega)$. Meanwhile, $V_\alpha^{(k+1)}(Q,\omega) = \min_U J_{\alpha,U}^{(k)}(Q,\omega)$, then we have $V_\alpha^{(k+1)}(Q_1,\omega) < V_\alpha^{(k+1)}(Q_2,\omega)$, for $\forall |Q_1| < |Q_2|$. Obviously, if we want to use induction to complete the proof of Lemma 2, we have to prove that: $f_\alpha^{(k+1)}(Q_1,\omega) < f_\alpha^{(k+1)}(Q_2,\omega)$, for $\forall |Q_1| < |Q_2|$. To simplify the proof, it is assumed that $Q_2 > Q_1 > 0$. The discussion will be divided into the following three situations.

- When $A_m \le |Q_1|$, then $|Q_1 + a| < |Q_2 + a|$, for $\forall a \in [-A_m, A_m]$, we can derive that:

$$\begin{aligned} f_\alpha^{(k+1)}(Q_1,\omega) &= \sum_{a=-A_m}^{A_m} p_a V_\alpha^{(k+1)}(Q_1+a,\omega) \\ &< \sum_{a=-A_m}^{A_m} p_a V_\alpha^{(k+1)}(Q_2+a,\omega) = f_\alpha^{(k+1)}(Q_2,\omega). \end{aligned} \tag{17}$$

- When $A_m > |Q_2|$, there exists an increment $a' \in \mathcal{A}' = \{a | a \in [-A_m, -\frac{1}{2}(Q_1+Q_2)\}$, such that $|Q_1 + a'| > |Q_2 + a'|$, and $V_\alpha^{(k+1)}(Q_1+a',\omega') > V_\alpha^{(k+1)}(Q_2+a',\omega')$. Notice that $-Q_1 - a' \in (\frac{1}{2}(Q_2 - Q_1), A_m - Q_1]$ and $Q_2 + a \in [-A_m + Q_2, A_m + Q_2]$, then $p_{-Q_1-a'-Q_2} V_\alpha^{(k+1)}(-Q_1-a',\omega)$ is a term in the summation $f_\alpha^{(k+1)}(Q_2,\omega)$, namely $\sum_{a=-A_m}^{A_m} p_a V_\alpha(Q+a,\omega)$. Similarly, $p_{-Q_2-a'-Q_1} V_\alpha^{(k+1)}(-Q_2-a',\omega)$ is a term in the summation $f_\alpha^{(k+1)}(Q_1,\omega)$. We further define $\mathcal{A}'' = \{a | a = -Q_1 - Q_2 - a'\}$, since $-Q_1 - Q_2 - a' \in \left(-\frac{1}{2}(Q_1+Q_2), A_m - Q_1 - Q_2\right]$, then $A' \cap A'' = \varnothing$.
  Furthermore, the probability of the estimation error transferring from Q_1 to $-Q_2 - a'$, i.e., $p_{-Q_2-a'-Q_1}$ equals $p_{-Q_1-a'-Q_2}$, the probability of the estimation error transferring from Q_2 to $-Q_1 - a'$. Since $-a' \in (\frac{1}{2}(Q_1+Q_2), A_m]$, then $|a'| > |-Q_1-Q_2-a'|$. According to our assumption of the increment, we can prove that for any $a' \in \mathcal{A}'$, $p_{a'} < p_{-Q_1-Q_2-a'}$. Then, we can derive:

$$\begin{aligned} & f_\alpha^{(k+1)}(Q_1,\omega) - f_\alpha^{(k+1)}(Q_2,\omega) \\ &= \sum_{a\in\mathcal{A}'} p_a V_\alpha^{(k+1)}(Q_1+a,\omega) + \sum_{a\in\mathcal{A}''} p_a V_\alpha^{(k+1)}(Q_1+a,\omega) \\ &- \sum_{a\in\mathcal{A}'} p_a V_\alpha^{(k+1)}(Q_2+a,\omega) - \sum_{a\in\mathcal{A}''} p_a V_\alpha^{(k+1)}(Q_2+a,\omega) + M(Q_1,Q_2) \\ &= \sum_{a\in\mathcal{A}'} p_a \{V_\alpha^{(k+1)}(Q_1+a,\omega) - V_\alpha^{(k+1)}(Q_2+a,\omega)\} \\ &+ \sum_{a\in\mathcal{A}'} p_{-Q_1-Q_2-a} \{V_\alpha^{(k+1)}(Q_2+a,\omega) - V_\alpha^{(k+1)}(Q_1+a,\omega)\} + M(Q_1,Q_2) \\ &= \sum_{a\in\mathcal{A}'} \left(p_a - p_{-Q_1-Q_2-a}\right) \{V_\alpha^{(k+1)}(Q_1+a,\omega) - V_\alpha^{(k+1)}(Q_2+a,\omega)\} + M(Q_1,Q_2) < 0, \end{aligned} \tag{18}$$

where $M(Q_1,Q_2) = \sum_{a\notin\mathcal{A}'\cup\mathcal{A}''} p_a \left(V_\alpha^{(k+1)}(Q_1+a,\omega) - V_\alpha^{(k+1)}(Q_2+a,\omega)\right) < 0$.

- When $|Q_2| > A_m > |Q_1|$, since $a' \in [-A_m, -\frac{1}{2}(Q_1+Q_2))$, we only need to consider the case when $A_m > \frac{1}{2}(Q_1+Q_2)$, in this case $-Q_1 - a' > \frac{1}{2}(Q_2 - Q_1) > Q_2 - A_m$. Therefore, $p_{-Q_1-a'-Q_2} V_\alpha^{(k+1)}(-Q_1-a',\omega)$ is a term in the summation $f_\alpha^{(k+1)}(Q_2,\omega)$. Similarly, we can also prove that $f_\alpha^{(k+1)}(Q_1,\omega) < f_\alpha^{(k+1)}(Q_2,\omega)$ when $|Q_2| > A_m > |Q_1|$.

According to Lemma 1, the conclusions above can be easily generalized to the cases without the condition $Q_2 > Q_1 > 0$. Finally, by letting $k \to \infty$, $V_\alpha^{(k+1)}(Q,\omega) \to V_\alpha(Q,\omega)$, therefore: $f_\alpha^{(k+1)}(Q,\omega) \to f_\alpha(Q,\omega)$. Hence: $f_\alpha(Q_1,\omega) < f_\alpha(Q_2,\omega)$. □

Remark 1. *Lemma 2 holds when $f_A(a)$, i.e., the PDF of increment $A(t)$ satisfies the following conditions:*

- $f_A(a) = f_A(-a)$, $\mu = 0$;
- $f_A(a_2) \le f_A(a_1), \forall a_2 \ge a_1 \ge 0$.

Then, with Lemmas 1 and 2, we can prove the threshold structure of the optimal stationary deterministic policy which minimizes L_λ^π in (10).

Theorem 2. *For a given λ, the optimal stationary deterministic policy which minimizes L_λ^π in (10) has a threshold structure when the context-aware weight is a first-order irreducible positive recurrent Markov process.*

Proof of Theorem 2. Let $s_\alpha^*(Q,\omega)$ denote the optimal action which minimizes the discounted cost $V_\alpha(Q,\omega)$ at state (Q,ω). If the optimal action $s_\alpha^*(Q,\omega) = 1$, then the vehicle will transmit its status update to the center at state (Q,ω) and $\Delta(Q,\omega) \ge 0$. Thus, we have:

$$\Delta(Q,\omega) = p_s\alpha \sum_{\omega' \in \mathbb{W}} p_{\omega\omega'} \sum_{a=-A_m}^{A_m} p_a \{ V_\alpha(Q+a,\omega') - V_\alpha(a,\omega') \} - \lambda \ge 0. \tag{19}$$

According to Lemma 2, for any $|Q'| > |Q|$, $\Delta(Q',\omega)$ can be lower bounded by

$$\begin{aligned} \Delta(Q',\omega) &= p_s\alpha \sum_{\omega' \in \mathbb{W}} p_{\omega\omega'} \sum_{a=-A_m}^{A_m} p_a \{ V_\alpha(Q'+a,\omega') - V_\alpha(a,\omega') \} - \lambda \\ &\ge p_s\alpha \sum_{\omega' \in \mathbb{W}} p_{\omega\omega'} \sum_{a=-A_m}^{A_m} p_a \{ V_\alpha(Q+a,\omega') - V_\alpha(a,\omega') \} - \lambda \ge 0. \end{aligned} \tag{20}$$

If $\Delta(Q,\omega) > 0$, then for any states with $|Q'| > |Q|$, the optimal policy is to transmit the status to the center. If $\Delta(Q,\omega) < 0$, then for any states with $|Q'| < |Q|$, the optimal action is not to transmit. In addition, the optimal policy will not be choosing to wait in all the slots. Therefore, for each context-aware weight ω, there must be a threshold $\tau_\omega \ge 0$. For any state (Q,ω) with $|Q| > \tau_\omega$, the optimal choice is to transmit the status update. We can then conclude that for a given weight ω, the optimal policy with a discount factor α has a threshold structure.

Let $\{\alpha_1, \alpha_2, \cdots, \alpha_k\}$ denote a sequence of discount factors and α_k converges to 1 when $k \to \infty$. Then, the optimal deterministic policy for $\alpha = 1$ will also converge to the optimal policy with a discount factor which is less than 1 [35]. Similar derivation is also applied in [12]. Therefore, we can prove the threshold structure of the optimal stationary deterministic policy which minimizes L_λ^π. □

Similarly, when the context-aware weight is i.i.d. over time, we can obtain the following theorem:

Theorem 3. *For a given λ, the optimal stationary deterministic policy which minimizes L_λ^π in (10) has a threshold structure when the context-aware weight is i.i.d. over time. The thresholds are the same for each state of the context-aware weight.*

Proof of Theorem 3. If the context-aware weight is i.i.d. over time, then we have:

$$\Delta(Q,\omega) = p_s\alpha \sum_{\omega'\in\mathbb{W}} p_{\omega'} \sum_{a=-A_m}^{A_m} p_a\{V_\alpha(Q+a,\omega') - V_\alpha(a,\omega')\} - \lambda = \Delta(Q), \tag{21}$$

where p_ω is the probability of the value of the context-aware weight being in state ω. Therefore, in this case, the state will be reduced to one dimension and the thresholds will be the same for all the states of the context-aware weight. □

According to Theorems 2 and 3, we proved the threshold structure of the two stationary deterministic policies that compose the UoI-optimal policy. Since the UoI-optimal policy for problem (9) is a probabilistic combination of two deterministic policies with threshold structures, we can finally draw the conclusion that the UoI-optimal policy also has a threshold structure.

3.3. Numerical Solution of Optimal Strategy

Based on Theorem 2, we only need to consider the policy that chooses to update with a probability of 1 in state (Q,ω), for $\forall|Q| \geq Q_{\max} = \max_\omega \tau_\omega$. Let $\mu_{Q,\omega}$ denote the probability that the state of the vehicle is (Q,ω). $y_{Q,\omega}$ denotes the probability that the state is (Q,ω) and the vehicle chooses to transmit an update. Therefore, we have:

Theorem 4. *When the context-aware weight is a first-order irreducible positive recurrent Markov process, the UoI-optimal policy can be derived by solving the following LP problem:*

$$\{\mu^*_{Q,\omega}, y^*_{Q,\omega}\} = \arg\min_{\{\mu_{Q,\omega}, y_{Q,\omega}\}} \sum_{\omega\in\mathbb{W}} \sum_{Q=-Q_{\max}}^{Q_{\max}} \omega Q^2 \mu_{Q,\omega}, \tag{22a}$$

$$s.t. \sum_{\omega\in\mathbb{W}} \sum_{Q=-Q_{\max}}^{Q_{\max}} \mu_{Q,\omega} = 1, \tag{22b}$$

$$\sum_{\omega\in\mathbb{W}} \sum_{Q=-Q_{\max}}^{Q_{\max}} y_{Q,\omega} \leq \rho, \tag{22c}$$

$$y_{Q,\omega} \leq \mu_{Q,\omega}, \forall Q,\omega, \tag{22d}$$

$$0 \leq y_{Q,\omega} \leq 1, 0 \leq \mu_{Q,\omega} \leq 1, \forall Q,\omega, \tag{22e}$$

$$\mu_{Q,\omega} = \sum_{\omega'\in\mathbb{W}} \sum_{Q'=-Q_{\max}}^{Q_{\max}} y_{Q',\omega'} p_s p_Q p_{\omega\omega'} + \sum_{\omega'\in\mathbb{W}} \sum_{Q'=-Q_{\max}}^{Q_{\max}} (\mu_{Q',\omega'} - y_{Q',\omega'} p_s) p_{Q'-Q} p_{\omega\omega'}. \tag{22f}$$

Proof of Theorem 4. We first derive the average UoI $\bar{C}^\pi$ as a function of $\mu_{Q,\omega}$ and $y_{Q,\omega}$. The vehicle is in state (Q,ω) and produces a cost of $C(Q,\omega,u) = \omega Q^2$ with a probability of $\mu_{Q,\omega}$. Therefore, the average UoI is:

$$\sum_{\omega\in\mathbb{W}} \sum_{Q=-Q_{\max}}^{Q_{\max}} \omega Q^2 \mu_{Q,\omega}. \tag{23}$$

As for the constraints, (22b) means that the sum of the probabilities of all the states should be 1. To explain (22c), we note that $y_{Q,\omega}$ is the probability of the vehicle being in state (Q,ω) and choosing to transmit the update, then the expectation of a one-step updating penalty for state (Q,ω) in (8) is $\mu_{Q,\omega}$. Therefore, the constraint on average update frequency $\bar{D}^\pi$ can be illustrated by

$$\sum_{\omega \in \mathbb{W}} \sum_{Q=-Q_{\max}}^{Q_{\max}} \mu_{Q,\omega} \leq \rho. \tag{24}$$

Then, we introduce $\xi_{Q,\omega} \in [0,1]$ to represent that the probability of the vehicle choosing to transmit updates in state (Q,ω) and (22d) can be obtained by the fact that $y_{Q,\omega} = \mu_{Q,\omega}\xi_{Q,\omega}$, while (22e) is derived by the nature of probability.

The right-hand side of (22f) can be viewed as two terms. The first term is the sum of transition probability from all the states to state (Q,ω) when the vehicle chooses to update and the transmission of status is successful. The second term is the sum of transition probability from all the states to state (Q,ω) when the transmission is failed or the vehicle chooses to wait. Therefore, we can prove that the optimal solution of problem (5) equals the solution of the LP problem. □

When $\omega(t)$ is i.i.d. over time, we can also obtain the UoI-optimal policy through the LP problem proposed in Theorem 4 and only need to use $p_{\omega'}$ as a replacement of $p_{\omega\omega'}$.

4. Scheduling in Unknown Contexts

To make decisions, the UoI-optimal updating policy obtained in Section 3 still needs the distributions of the context-aware weight $\omega(t)$, the increment $A(t)$ and the successful transmission probability, which may not be available in advance or may change over time in most practical systems. To solve this problem, we will assume that the distribution of the context-aware weight is not pre-determined and the vehicle has to learn it. In this section, we use the reinforcement learning (RL) algorithm to learn the dynamic of the context and the characteristic of the wireless channel.

To solve this problem, we turn to the model-based RL framework proposed in [36]. We only consider the cases when the UoI-optimal policy has a threshold structure. This assumption makes the optimal policy based on the truncated state space equal the optimal policy of the original problem.

We use the 3-tuple (s,s',U) to formulate the proposed RL-based updating policy. The states in the current slot and next slot are denoted by s and s', respectively. U denotes the action chosen in the current slot. The settings of the discretized state space and the action space are the same as the settings proposed in Section 3.1. The smaller the step size used in the discretization is, the closer our results are to those in continuous state space. In addition, the selection of the step size only affects the accuracy of the update threshold. Therefore, the performance loss caused by discretization can be reduced by choosing a smaller step size.

We display details about the proposed RL-based updating policy in Algorithm 1. At the beginning of episode k, we randomly decide whether to explore or exploit. $l \in [0,1]$ represents the trade-off between exploration and exploitation during the following episode. A larger l means a higher frequency of exploration and vice versa. If the algorithm chooses to explore during this episode, a random policy $\pi_{rand}(s)$ will be used, i.e., we randomly choose to update or not in each state to find more valuable actions. If the algorithm chooses to exploit, then we have to obtain the probability transfer functions $\tilde{p}_k(s'|s,U)$ for each state transmission pair. In Algorithm 1, $N(s,U)$ and $N(s,U,s')$ represent the number of occurrences of state–action pair s,U and state transition from s to s' given action U, respectively. Based on the assumption that the optimal policy has a threshold structure, the policy $\pi(k)$ which can minimize the average UoI with the estimated probability transfer functions, can be directly solved through the LP problem proposed in Theorem 4. Then, the vehicle will use policy π_k to derive state–action pairs and the state transitions in the following $\lceil L_k \rceil$ slots. Here, $L > 0$ is defined to control the number of state transitions observed in each episode. At the end of each episode, the model will be updated according to the state–action pairs and the state transitions observed during the episode. Finally,

after K episodes, the algorithm will output the RL-based updating policy $\pi^{\star}(s)$, which is derived based on $\tilde{p}_K(s'|s,U)$.

Algorithm 1 RL-based Updating Policy

Input: $l \in [0,1], L > 0, K > 0$

for episodes $k = 1, 2, \ldots, K$ **do**
  Set $L_k = L\sqrt{k}, \epsilon_k = l/\sqrt{k}$, uniformly draw $\alpha \in [0,1]$.
  if $\alpha < \epsilon_k$ **then**
    Set $\pi_k(s) = \pi_{rand}(s)$,
  else
    for each state $s, s' \in \mathbb{S}$ and $U \in \mathbb{U}$ **do**
      if $N(s,U) > 0$ **then**
        Let $\tilde{p}_k(s'|s,U) = N(s,U,s')/N(s,u)$,
      else
        $\tilde{p}_k(s'|s,U) = 1/|\mathbb{S}|$.
      end if
    end for
    obtain policy $\pi_k(s)$ by solving the estimated CMDP
  end if
  Randomly choose an initial state $s(1)$.
  for slots $t = 1, 2, \ldots, \lceil L_k \rceil - 1$ **do**
    Choose action $U(t)$ as $\pi_k(s(t))$.
    Observe the next state $s(t+1)$.
    $N(s(t), U(t), s(t+1)) = N(s(t), U(t), s(t+1)) + 1$.
    $N(s(t), U(t)) = N(s(t), U(t)) + 1$.
    $s(t) \leftarrow s(t+1)$.
  end for
end for
obtain policy $\pi^{\star}(s)$ by solving the estimated CMDP based on $\tilde{p}_k(s'|s,U)$, $s, s' \in \mathbb{S}, U \in \mathbb{U}$.

Output: output the RL-based updating policy $\pi^{\star}(s)$

5. Simulation Results and Discussion

5.1. Simulation Setup

To facilitate the simulation, we consider the case where the context-aware weight of the vehicle only has two different states: the 'normal' state and 'urgent' state. The 'normal' state means that the vehicle is in ordinary situations and the significance of accuracy of status information is relatively low. We set $\omega(t)$ as 1 in 'normal' state while $\omega(t)$ is set as a constant much larger than 1, ω_e, in 'urgent' state to show that the vehicle is in emergencies. Two different distributions of the context-aware weight are taken into consideration to conform to the assumptions about $\omega(t)$ used in Section 3.1:

1. The context-aware weight $\omega(t)$ has the first-order Markov property. The state transition diagram of $\omega(t)$ is shown in Figure 2 and $\omega(t)$ is irreducible and positive recurrent. p_1 is the probability of the context-aware weight transferring from the normal state to the urgent state, while p_2 is the probability of the weight transferring from the urgent state to the normal state;
2. The context-aware weight $\omega(t)$ is i.i.d. over time. The probability of the weight being in the urgent state and the normal state are denoted by p_h and p_l, respectively.

As for the increment $A(t)$, A_{max} is set to a large enough positive number to simplify the simulations.

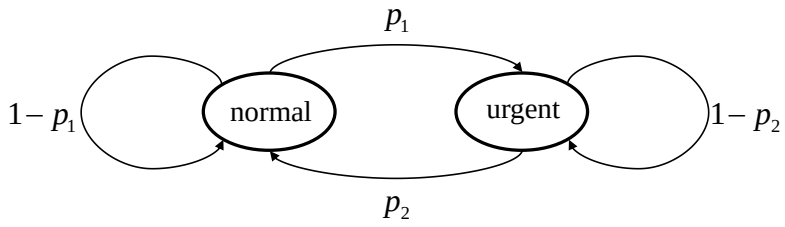

Figure 2. The state transition diagram of $\omega(t)$.

5.2. Numerical Results

Figure 3 shows the structure of the UoI-optimal updating policy. For the discretization of the estimation error, the step size used is 1. It can be seen that under the two different distributions of the context-aware weight mentioned above, the optimal updating policies all have threshold structures. Especially when the context-aware weight is i.i.d. over time, Figure 3b shows that thresholds for all the states of the context-aware weight are the same, which matches well with theoretical analysis. From Figure 3c, we can find that the UoI-optimal policy also has threshold structure when increment $A(t)$ obeys a uniform distribution $Unif(-3,3)$, which verifies Remark 1. We then simulate the UoI-optimal policy under the contexts with more states to show the policy is generic. We consider a three-state context-aware weight which takes value from $\omega_1 = 1, \omega_2 = 50, \omega_3 = 100$. The state transition matrix P_3 of the three-state context-aware weight is:

$$P_3 = \begin{bmatrix} 0.997 & 0.002 & 0.001 \\ 0.02 & 0.97 & 0.01 \\ 0.2 & 0.1 & 0.7 \end{bmatrix}, \tag{25}$$

where the j-th element on the i-th row indicates the probability that the context transfers from state ω_i to state ω_j. The numerical results (Figure 3d) show that when the context-aware weight has more states, the UoI-optimal policy still has a threshold structure, which verifies our theoretical results.

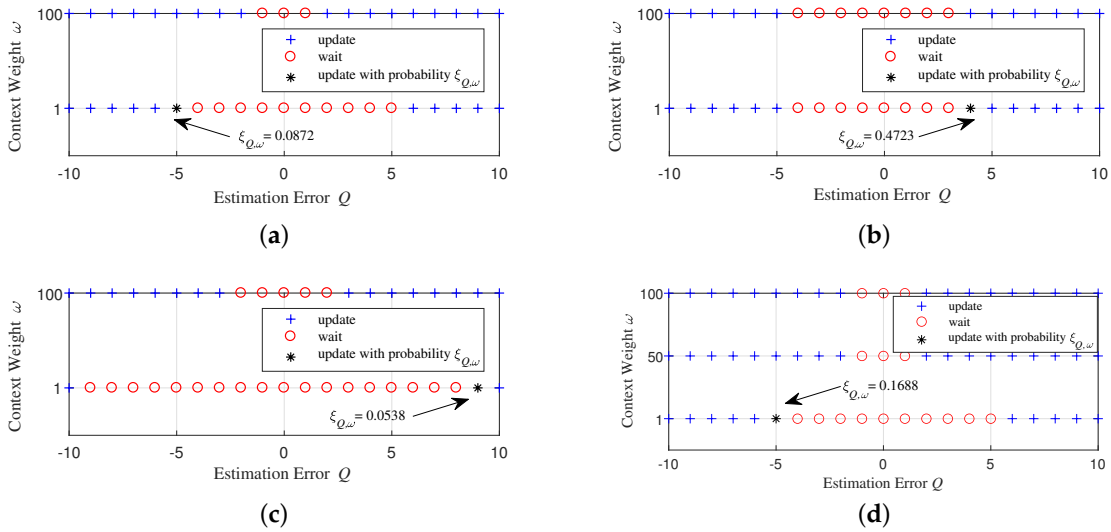

Figure 3. Threshold structure of the UoI-optimal updating policy when: (**a**) the context-aware weight is a first-order Markov process, $\rho = 0.05, p_1 = 0.001, p_2 = 0.01, p_s = 0.9, \sigma^2 = 1, \omega_e = 100$; (**b**) the context-aware weight is i.i.d. over time, $\rho = 0.05, p_l = 0.999, p_h = 0.001, p_s = 0.9, \sigma^2 = 1, \omega_e = 100$; (**c**) the context-aware weight is a first order Markov process, $\rho = 0.05, p_1 = 0.001, p_2 = 0.01, p_s = 0.9, \omega_e = 100$, increment in the estimation error during one slot, i.e., $A(t) \sim Unif(-3,3)$, for $\forall t$; and (**d**) the context-aware weight is a three-state first-order Markov process, which takes value from $\omega_1 = 1, \omega_2 = 50, \omega_3 = 100$ and evolves according to the state transition matrix P_3, $\rho = 0.05, p_s = 0.9, \sigma^2 = 1$.

Then, we will focus on the results obtained when the context-aware weight is a first-order irreducible positive recurrent Markov process, as shown in Figure 2. Figure 4 shows the average UoI of the UoI-optimal policy, the AoI-optimal policy derived by CMDP, the RL-based updating policy, and the update-index-based adaptive scheme [27]. In the RL-based updating policy, $L = 8000$, $l = 1$ and $K = 50$. All the numerical results of the RL-based policy are averaged over 100 runs.

First of all, the UoI-optimal policy can only be obtained based on advanced information about the system dynamics. However, the RL-based policy achieves near-optimal without knowing the system dynamics, which indicates that Algorithm 1 learns relatively accurate probability transfer functions from the observed state–action pairs and state transitions during the training.

Secondly, according to Figure 4, the AoI-optimal policy yields a much higher UoI than the three UoI-based policies, namely the UoI-optimal policy, the RL-based updating policy, and the update-index-based adaptive scheme. On the one hand, AoI is one special case of UoI. When the context-aware weight $\omega(t) = 1$, the increment $A(t) = 1$, and the cost function $\delta(Q(t)) = Q(t)$, then UoI equals AoI. Therefore, the AoI-optimal policy ignores the fact that different contexts have different requirements for information freshness. In the proposed UoI-based updating policies, different contexts have different policies and update thresholds, while the AoI-optimal updating policies for different contexts are the same. On the other hand, Figure 5 reveals that the AoI-optimal policy leads to a much higher estimation error, which results in worse performance in terms of UoI. The AoI-optimal policy is an oblivious policy, which is independent of the monitored process. Since AoI increases linearly with time, the AoI-optimal policy can only minimize the linear performance degradation in terms of time. However, the UoI-based policies (the cost function $\delta(Q(t)) = (Q(t))^2$) considered in this paper are process-dependent, which are called non-oblivious policies, and can benefit from both age and process realization [37]. These policies can directly minimize the nonlinear impact exerted by information staleness and the gap between the actual status and the estimated status.

Thirdly, our updating policies outperform the update-index-based adaptive scheme [27] in terms of UoI. Under the adaptive scheme, the vehicle will derive an update index as a function of the current estimation error and the context-aware weight for the next slot. If the index is larger than the adaptive update threshold, then the vehicle is supposed to transmit its status information to the center. If the vehicle transmits an update in slot t, then the adaptive threshold will increase in the next slot; otherwise, the adaptive threshold will decrease. The adaptive scheme will cause an overuse of the resource in 'urgent' states and lead to the fact that the vehicles cannot receive resources in 'normal' states. However, the UoI-optimal policy and the trained RL-updating policy are fixed schemes, which can avoid the extremely unbalanced resource allocation between the two contexts and achieve better performance.

Figure 6 shows the influence of the maximum average update frequency ρ and the context weight for emergency, ω_e, on update threshold of UoI-optimal policy. In order to obtain more accurate results, the step size used here is 0.25. The solid curves show update thresholds for the normal state while the dashed curves show update thresholds for the urgent state. When the constraint on update resources is strict, the update thresholds fall faster. Furthermore, a larger ω_e results in a lower update threshold for the urgent state and a higher threshold for the normal state. This phenomenon indicates that the value of ω_e means the tolerance of estimation error in the emergency. When $\rho < 0.1$, the influence of ω_e on the update threshold for the normal state is larger than the urgent state. For the cases where the maximum average update frequency is relatively large, ω_e has little effect on update thresholds for both normal state and urgent state.

Figure 7 shows that the update thresholds also depend on the dynamic of context-aware weight when the weight has first-order Markov property. When p_2 is approaching $1 - p_1$, the gap between update thresholds for the urgent state and the normal state becomes smaller for the context-aware weight which tends to be i.i.d. over time. When $p_2 = 1$, the

update threshold for the urgent state exceeds the threshold for the normal state. Therefore, the update threshold for the urgent state is not necessarily lower than the update threshold for the normal state.

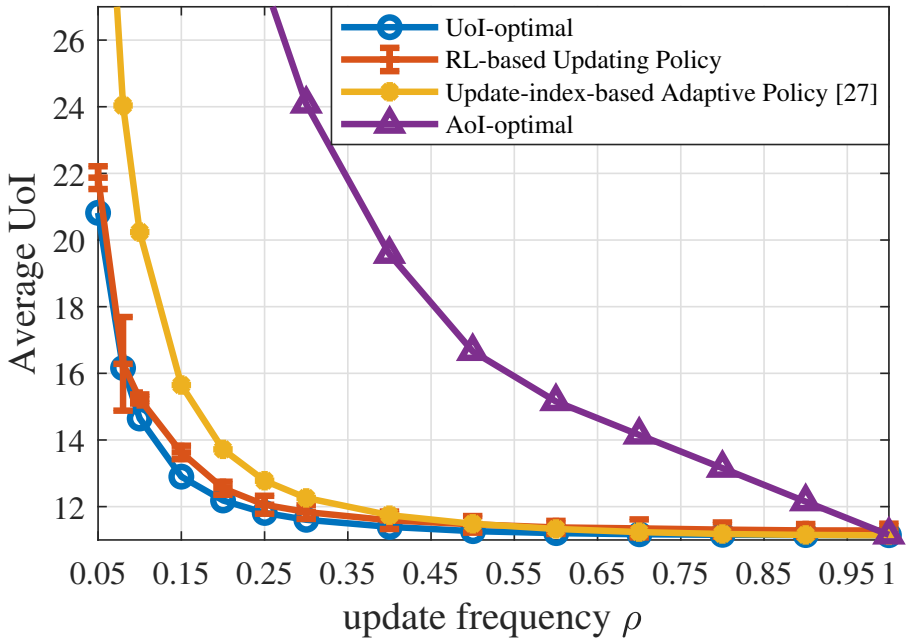

Figure 4. Average UoI of the UoI-optimal updating policy, the RL-based updating policy, the update-index-based adaptive scheme [27], and the AoI-optimal updating policy when $p_1 = 0.001$, $p_2 = 0.01, p_s = 0.9, \sigma^2 = 1, \omega_e = 100$.

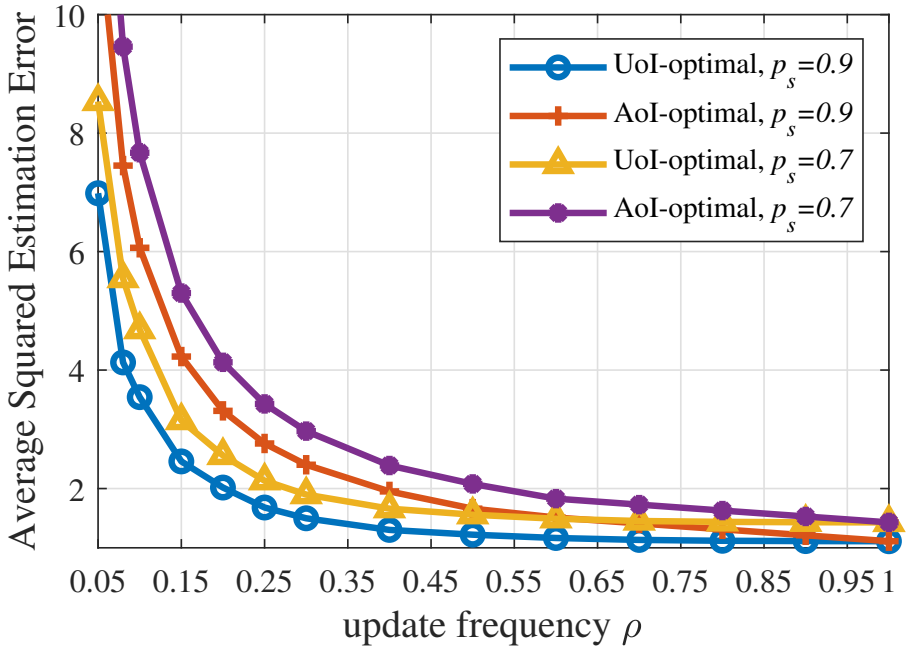

Figure 5. Average squared estimation error of the UoI-optimal updating policy and the AoI-optimal updating policy when $p_1 = 0.001, p_2 = 0.01, \sigma^2 = 1, \omega_e = 100$.

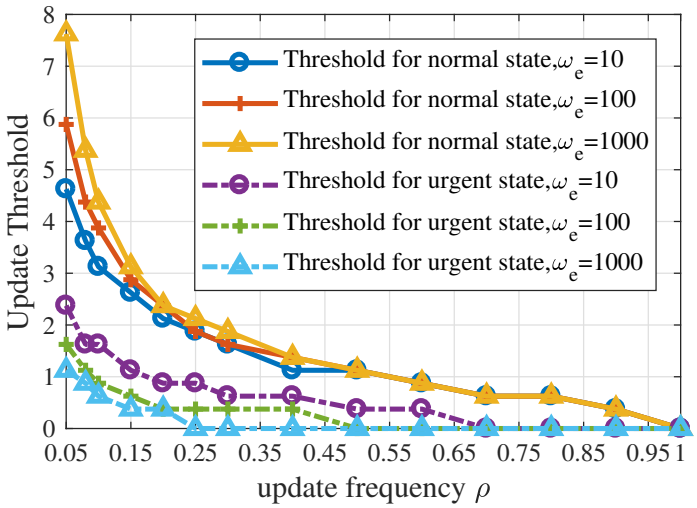

Figure 6. Update thresholds of the UoI-optimal updating policy with different values of ω_e when $p_1 = 0.001, p_2 = 0.01, p_s = 0.9, \sigma^2 = 1$.

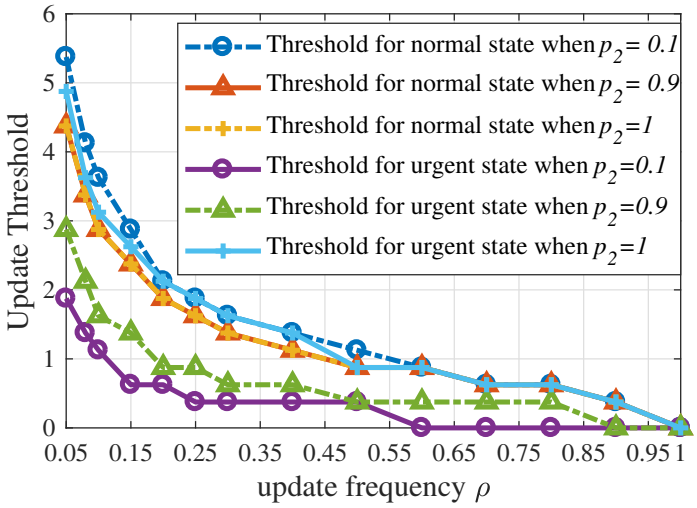

Figure 7. Update thresholds of the UoI-optimal updating policy with different values of p_2 when $p_1 = 0.01, p_s = 0.9, \sigma^2 = 1, \omega_e = 100$.

Figure 8 shows the performance of the RL-based updating policy with different values of L. According to Algorithm 1, the number of state transitions observed in episode k is $\lceil L\sqrt{k} \rceil$. Therefore, L denotes the number of state transitions observed during the whole learning process. Generally speaking, a larger L reduces the randomness of the performance and achieves a better UoI. The performance of the RL-based updating policy depends on the accuracy of the model obtained through training, namely whether the estimated probability transfer function of the system is accurate. A larger L means that the algorithm can collect more data or state transitions and obtain a more accurate model.

Figure 9 shows the influence of the number of episodes, i.e., K, on the performance of the RL-based updating policy. A larger K leads to a lower average UoI and smaller randomness over 100 runs. On the one hand, the more episodes and the more data the algorithm observes, the more accurate the model obtained will be and the better the performance of the updating policy will be. On the other hand, the value of K is the number of iterations for the policy obtained through the estimated CMDP. The policy $\pi_k(s)$ used in episode k is derived based on the state–action pairs and the state transitions observed in the previous $k-1$ episodes. Therefore, more frequent iterations of the updating policy can obtain more valuable state–action pairs and better performance.

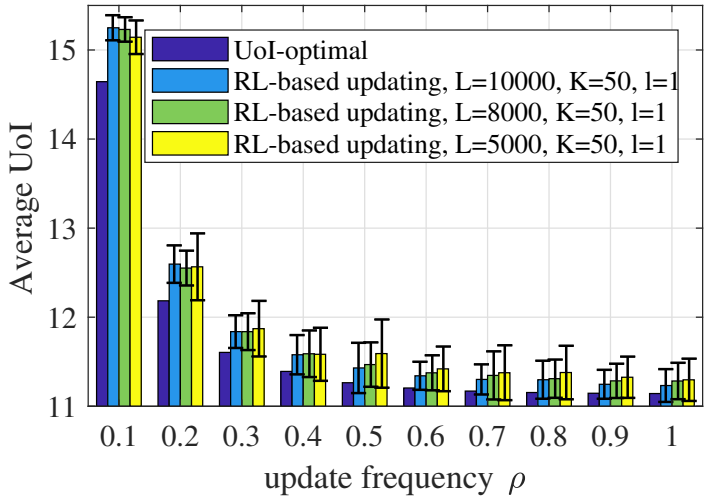

Figure 8. Average UoI of the RL-based updating policy with different values of L when $p_1 = 0.001, p_2 = 0.01, \sigma^2 = 1, \omega_e = 100$.

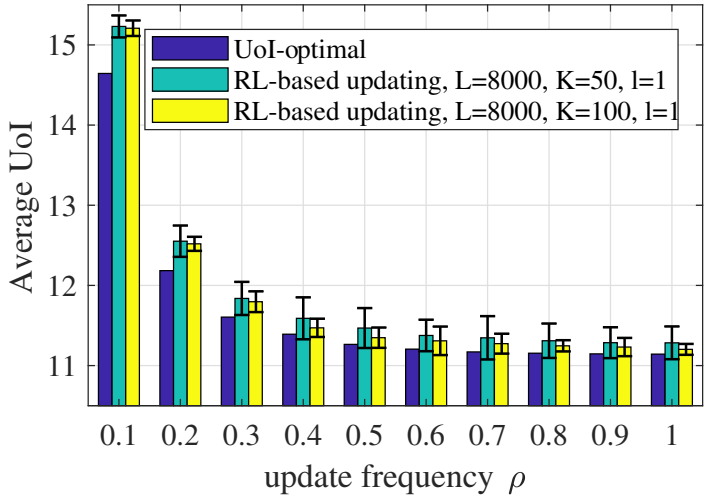

Figure 9. Average UoI of the RL-based updating policy with different values of K when $p_1 = 0.001, p_2 = 0.01, \sigma^2 = 1, \omega_e = 100$.

6. Conclusions

In this work, we studied how to minimize the performance degradation caused by outdated information in terms of UoI, which is a new metric jointly considering context and information freshness. We proved that the UoI-optimal updating policy for the considered single-user remote monitoring system has a single threshold structure. Then, the policy was obtained through linear programming by assuming that the state transition probability of the system is known in advance. In unknown contexts, we further used a reinforcement learning algorithm to learn the dynamics of the system. Simulations verified the threshold structure of the UoI-optimal policies and showed that the update thresholds decrease as the maximum average update frequency increases. In addition, a larger context-aware weight in emergencies resulted in a lower update threshold for urgent states. However, since the state transition probability also influenced the update thresholds, the update threshold for emergencies was not necessarily higher than the update threshold for normal states, especially when the probability of transferring from urgent states to normal states tended towards 1. Furthermore, the numerical results showed that the proposed RL-based updating policy achieved a near-optimal performance without advanced knowledge of the system model.

In fact, determining the context-aware weight in practical systems, where the models of the context are often very complicated and difficult to obtain in advance, remains open. As for future work, we plan to use deep RL algorithms to learn the models of the context variation. We believe that UoI can provide a new performance metric for information timeliness measurement in the future V2X scenario. In addition, we believe the proposed UoI metric and the context-aware scheduling policy can shed some light on low-latency and ultra-reliable wireless communication in the future 5G/6G systems.

Author Contributions: Conceptualization, L.W.; formal analysis, L.W.; methodology, L.W.; software, L.W.; supervision, S.Z. and Z.N.; writing—original draft, L.W.; writing—review and editing, J.S., Y.S., S.Z. and Z.N. All authors have read and agreed to the published version of the manuscript.

Funding: This work is sponsored in part by the National Key R&D Program of China No. 2020YFB1806605, by the Nature Science Foundation of China (No. 62022049, No. 61871254, No. 61861136003), by the China Postdoctoral Science Foundation No. 2020M680558, and Hitachi Ltd.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AoI	Age of Information
AoII	Age of Incorrect Information
AoS	Age of Synchronization
CDF	Cumulative Distribution Function
CMDP	Constrained Markov Decision Process
CoUD	Cost of Update Delay
i.i.d.	Independent and Identically Distributed
IoT	Internet of Things
LP	Linear Programming
PDF	Probability Density Function
RL	Reinforcement Learning
UoI	Urgency of Information
URLLC	Ultra-Reliable Low-Latency Communication
VR	Virtual Reality
V2X	Vehicle to Everything

Appendix A. Proof of Theorem 1

Given a state $s = (Q, \omega) \in \mathbb{S}$ and a nonempty subset of the state space, $\mathbb{G} \subset \mathbb{S}$, let $\mathcal{R}(s, \mathbb{G})$ denote the class of policies θ such that the probability $P^{\theta}(s(t) \in \mathbb{G}$ for some $t \geq 1 | s(0) = s) = 1$ and the expected time $m_{s,\mathbb{G}}(\theta)$ of the first passage from s to $\mathbb{G}$ under policy θ is finite. Then, let $\mathcal{R}^{\star}(s, \mathbb{G})$ denote the class of policies θ such that the expected average UoI $c_{s,\mathbb{G}}(\theta)$ and the expected transmission cost $d_{s,\mathbb{G}}(\theta)$ of the first passage from s to $\mathbb{G}$ are finite and $\theta \in \mathcal{R}(s, \mathbb{G})$. To prove Theorem 1, we then introduce Assumptions A1–A5 in [33]:

Assumption A1. *For all* $b > 0$, *the set* $\mathbb{G}(b) \triangleq \{s |$ *there exists an action* U *such that* $C(s, U) + D(s, U) \leq b\}$ *is finite.*

Assumption A2. *There exists a stationary deterministic policy* π *that induces a Markov chain with the following properties: the state space consists of a single (nonempty) positive recurrent class* $\mathbb{R}^{\pi}$ *and a set* $\mathbb{T}^{\pi}$ *of transient states such that* $\pi \in \mathcal{R}^{\star}(s, \mathbb{R}^{\pi})$, *for any* $s \in \mathbb{T}^{\pi}$, *and both the average UoI* $\bar{C}^{\pi}$ *and the average transmission cost* $\bar{D}^{\pi}$ *on* $\mathbb{R}^{\pi}$ *are finite.*

Assumption A3. *Given any two states $s, s' \in \mathbb{S}$ and $s \neq s'$, there exists a policy π (a function of s and s') such that $\pi \in \mathcal{R}^\star(s, \{s'\})$.*

Assumption A4. *If a stationary deterministic policy has at least one positive recurrent state, then it has a single positive recurrent class, and this class contains the state (Q, ω) with $Q = 0$.*

Assumption A5. *There exists a policy π such that the average UoI $\bar{C}^\pi < \infty$ and average transmission cost $\bar{D}^\pi < \rho$.*

Furthermore, the problem (9) has the following property:

Lemma A1. *Assumptions A1–A5 hold for problem (9).*

Proof of Lemma A1. First of all, we focus on the cases where the context-aware weight is assumed as a first-order irreducible positive recurrent Markov process:

- Assumption A1: In this problem, $C(s, U)$ is the UoI at state s, namely $C(Q, \omega, U) = \omega Q^2$. $D(s, U)$ is 1 if the vehicle chooses to transmit its status and $D(s, U)$ is 0 otherwise, namely $D(Q, \omega, U) = U$. Therefore, Assumption A1 holds, for any $b > 0$, the number of states (Q, ω) with $\omega Q^2 \leq b$ is finite.
- Assumption A2: Due to the current high-level wireless communication technology, we reasonably assumed that the successful transmission probability p_s is relatively close to 1. Based on the assumptions mentioned above, the Markov chain of context-aware weight obviously satisfies Assumption A2. Define the probability of the context-aware weight transferring from ω to ω' in k steps for the first time as $P_{\omega,\omega',k}$. Then, we consider the policy $\pi(Q, \omega) = 1$ for all $(Q, \omega) \in \mathbb{S}$, namely this policy chooses to transmit in all the states.
  Since the evolution of the context-aware weight is independent with the evolution of the estimation error and the updating policy. Therefore, we first focused on the estimation error, which can be formulated as a one-dimensional irreducible Markov chain with state space $\mathbb{Q} = \{0, \pm\Delta_Q, \pm 2\Delta_Q, \cdots, \pm n\Delta_Q, \cdots\}$. We denote the set of states which can transfer to state Q in a single step by $\mathcal{Z}_Q$. The probability of the estimation error transferring from state Q to state Q' at the k-th step without an arrival to state $Q = 0$ is defined as $P'_{Q,Q',k}$. Obviously, $\sum_{Q' \in \mathbb{Q}} P'_{Q,Q',k} < (1 - p_s)^k$. Then, the probability of the first passage from state $Q(Q \neq 0)$ to 0 taking $k + 1$ steps is $\sum_{Q' \notin \mathcal{Z}_0} P'_{Q,Q',k} p_s + \sum_{Q' \in \mathcal{Z}_0} P'_{Q,Q',k}(p_s + (1 - p_s)p_{0-Q'}) < (1 - p_s)^k$, where $p_{0-Q'}$ is the probability that the increment in estimation error is $-Q'$. Therefore, the expected time of the first passage from $Q(Q \neq 0)$ to 0 is finite.
  For state $Q = 0$, the estimation error will stay in this state in the next step with a probability of $p_s + p_{0-0}$ and will first return to state $Q = 0$ in the second transition with a probability smaller than $(1 - p_s - p_{0-0})$. Then, starting from state $Q = 0$, the estimation error will first return to state $Q = 0$ in the $k + 1$-th ($k > 2$) step will be smaller than $(1 - p_s - p_{0-0})(1 - p_s)^{k-1}$. Therefore, we can prove that state $Q = 0$ is a positive recurrent state, and $\mathbb{R}^\pi_Q = \{Q = 0\}$ is a positive recurrent class of the induced Markov chain of the estimation error. Furthermore, for any states in $\mathbb{T}^\pi_Q = \mathbb{Q} \backslash \mathbb{R}^\pi_Q$, the expected time of the first passage from the state in $\mathbb{T}^\pi_Q$ to state $Q = 0$ under π is finite and the probability of the states in $\mathbb{T}^\pi_Q$ not getting to state $Q = 0$ in k steps is smaller than $(1 - p_s)^k$.
  Define the probability of state Q transferring to state Q' in k steps for the first time as $P_{Q,Q',k}$. Then, the probability of state (Q, ω) transferring to state (Q', ω') in k steps for the first time is $P_{Q,Q',k} P_{\omega,\omega',k}$. Since $\sum_{k=1}^\infty P_{Q,Q',k} k < \infty$ and $\sum_{k=1}^\infty P_{\omega,\omega',k} k < \infty$, then $\sum_{k=1}^\infty P_{Q,Q',k} P_{\omega,\omega',k} k < \infty$. Therefore, the set of states $\mathbb{R}^\pi = \{(Q, \omega) | Q \in \mathbb{R}^\pi_Q, \omega \in \mathbb{W}\}$ is a positive recurrent class. Similarly, we can prove that $\mathbb{T}^\pi = \mathbb{S} \backslash \mathbb{R}^\pi$ satisfies Assumption A2. Finally, $\bar{D}^\pi = 1 < \infty$, $\bar{C}^\pi = E[\omega] \frac{1}{p_s} \sigma^2 < \infty$.

- Assumption A3: Define $P_{Q,\min} = \min_{Q'} p_{Q-Q'}$, $P_{Q,max} = \max_{Q'} p_{Q-Q'}$. Consider the policy $\pi'(Q,\omega) = 0$ for all states $(Q,\omega) \in \mathbb{S}$, notably that this policy chooses not to transmit in any states. Similarly, we first focus on the Markov chain of estimation error. Starting from state Q, the probability of transferring to state Q' in $k+1$-th $(k \geq 2)$ steps for the first time is smaller than $(1 - p_{Q'-Q})P_{Q',max}(1 - P_{Q',min})^{k-1}$. Then, the expected time of the first passage from state Q to state Q' under policy π' is finite. Similarly, since the Markov chain of context-aware weight is irreducible positive recurrent and independent with the updating policy, we can therefore prove that the expected time of the first passage from state (Q,ω) to state (Q',ω') under policy π' is finite.
- Assumption A4: For the Markov chain of the estimation error, any state will return to state $Q = 0$ if a successful transmission occurs. For the policy without transmission, namely $\pi'(Q,\omega) = 0$, state $Q = 0$ still exists in only one positive recurrent class. For each positive recurrent class containing state $Q = 0$, we can prove that there is only one positive recurrent class. Since the Markov chain of the context-aware weight is irreducible positive recurrent, we can similarly prove Assumption A4 .
- Assumption A5: The policy π_ρ that updates the status with a probability of $\rho - \delta$ satisfies Assumption A5. Here, δ is a small positive number. Under this policy, $\bar{D}^\pi = \rho - \delta < \rho$ and $\bar{C}^\pi = E[\omega]\frac{1}{p_s(\rho-\delta)}\sigma^2 < \infty$.

Similarly, we can prove that Assumptions A1–A5 also holds for problem (9) when the context-aware weight is i.i.d. over time. □

Since Assumptions A1–A5 hold for problem (9), then according to Theorem 2.5 in [33], there exists an optimal stationary randomized policy for problem (9). Meanwhile, the optimal policy is a probabilistic combination of two stationary deterministic policies which only differ on at most one state.

Furthermore, according to Lemma 3.9 in [33], the two stationary deterministic policies each optimize the unconstrained cost in (10) with a different λ.

References

1. Talak, R.; Karaman, S.; Modiano, E. Speed limits in autonomous vehicular networks due to communication constraints. In Proceedings of the 2016 IEEE 55th Conference on Decision and Control (CDC), Las Vegas, NV, USA, 12–14 December 2016; pp. 4998–5003.
2. Hou, I.H.; Naghsh, N.Z.; Paul, S.; Hu, Y.C.; Eryilmaz, A. Predictive Scheduling for Virtual Reality. In Proceedings of the IEEE INFOCOM 2020-IEEE Conference on Computer Communications, Toronto, ON, Canada, 6–9 July 2020; pp. 1349–1358.
3. Kaul, S.; Yates, R.; Gruteser, M. Real-time status: How often should one update? In Proceedings of the 2012 Proceedings IEEE INFOCOM, Orlando, FL, USA, 25–30 March 2012; pp. 2731–2735.
4. Sun, Y.; Polyanskiy, Y.; Uysal-Biyikoglu, E. Remote estimation of the Wiener process over a channel with random delay. In Proceedings of the 2017 IEEE International Symposium on Information Theory (ISIT), Aachen, Germany, 25–30 June 2017; pp. 321–325.
5. Sun, Y.; Polyanskiy, Y.; Uysal, E. Sampling of the wiener process for remote estimation over a channel with random delay. *IEEE Trans. Inf. Theory* **2019**, *66*, 1118–1135. [CrossRef]
6. Jiang, Z.; Zhou, S. Status from a random field: How densely should one update? In Proceedings of the 2019 IEEE International Symposium on Information Theory (ISIT), Paris, France, 7–12 July 2019; pp. 1037–1041.
7. Bedewy, A.M.; Sun, Y.; Singh, R.; Shroff, N.B. Optimizing information freshness using low-power status updates via sleep-wake scheduling. In Proceedings of the Twenty-First International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing, New York, NY, USA, 11–14 October 2020; pp. 51–60.
8. Ceran, E.T.; Gündüz, D.; György, A. Average age of information with hybrid ARQ under a resource constraint. *IEEE Trans. Wirel. Commun.* **2019**, *18*, 1900–1913. [CrossRef]
9. Sun, J.; Jiang, Z.; Krishnamachari, B.; Zhou, S.; Niu, Z. Closed-form Whittle's index-enabled random access for timely status update. *IEEE Trans. Commun.* **2019**, *68*, 1538–1551. [CrossRef]
10. Yates, R.D.; Kaul, S.K. Status updates over unreliable multiaccess channels. In Proceedings of the 2017 IEEE International Symposium on Information Theory (ISIT), Aachen, Germany, 25–30 June 2017; pp. 331–335.
11. Sun, J.; Wang, L.; Jiang, Z.; Zhou, S.; Niu, Z. Age-Optimal Scheduling for Heterogeneous Traffic with Timely Throughput Constraints. *IEEE J. Sel. Areas Commun.* **2021**, *39*, 1485–1498. [CrossRef]
12. Tang, H.; Wang, J.; Song, L.; Song, J. Minimizing age of information with power constraints: Multi-user opportunistic scheduling in multi-state time-varying channels. *IEEE J. Sel. Areas Commun.* **2020**, *38*, 854–868. [CrossRef]

13. Abdel-Aziz, M.K.; Samarakoon, S.; Liu, C.F.; Bennis, M.; Saad, W. Optimized age of information tail for ultra-reliable low-latency communications in vehicular networks. *IEEE Trans. Commun.* **2019**, *68*, 1911–1924. [CrossRef]
14. Devassy, R.; Durisi, G.; Ferrante, G.C.; Simeone, O.; Uysal-Biyikoglu, E. Delay and peak-age violation probability in short-packet transmissions. In Proceedings of the 2018 IEEE International Symposium on Information Theory (ISIT), Vail, CO, USA, 17–22 June 2018; pp. 2471–2475.
15. Inoue, Y.; Masuyama, H.; Takine, T.; Tanaka, T. A general formula for the stationary distribution of the age of information and its application to single-server queues. *IEEE Trans. Inf. Theory* **2019**, *65*, 8305–8324. [CrossRef]
16. Sun, Y.; Uysal-Biyikoglu, E.; Yates, R.D.; Koksal, C.E.; Shroff, N.B. Update or wait: How to keep your data fresh. *IEEE Trans. Inf. Theory* **2017**, *63*, 7492–7508. [CrossRef]
17. Zheng, X.; Zhou, S.; Jiang, Z.; Niu, Z. Closed-form analysis of non-linear age of information in status updates with an energy harvesting transmitter. *IEEE Trans. Wirel. Commun.* **2019**, *18*, 4129–4142. [CrossRef]
18. Kosta, A.; Pappas, N.; Ephremides, A.; Angelakis, V. Non-linear age of information in a discrete time queue: Stationary distribution and average performance analysis. In Proceedings of the ICC 2020—2020 IEEE International Conference on Communications (ICC), Dublin, Ireland, 7–11 June 2020; pp. 1–6.
19. Kosta, A.; Pappas, N.; Ephremides, A.; Angelakis, V. The cost of delay in status updates and their value: Non-linear ageing. *IEEE Trans. Commun.* **2020**, *68*, 4905–4918. [CrossRef]
20. Zhong, J.; Yates, R.D.; Soljanin, E. Two freshness metrics for local cache refresh. In Proceedings of the 2018 IEEE International Symposium on Information Theory (ISIT), Vail, CO, USA, 17–22 June 2018; pp. 1924–1928.
21. Maatouk, A.; Kriouile, S.; Assaad, M.; Ephremides, A. The age of incorrect information: A new performance metric for status updates. *IEEE/ACM Trans. Netw.* **2020**, *28*, 2215–2228. [CrossRef]
22. Kadota, I.; Sinha, A.; Uysal-Biyikoglu, E.; Singh, R.; Modiano, E. Scheduling policies for minimizing age of information in broadcast wireless networks. *IEEE/ACM Trans. Netw.* **2018**, *26*, 2637–2650. [CrossRef]
23. Song, J.; Gunduz, D.; Choi, W. Optimal scheduling policy for minimizing age of information with a relay. *arXiv* **2020**, arXiv:2009.02716.
24. Sun, Y.; Cyr, B. Information aging through queues: A mutual information perspective. In Proceedings of the 2018 IEEE 19th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), Kalamata, Greece, 25–28 June 2018; pp. 1–5.
25. Kam, C.; Kompella, S.; Nguyen, G.D.; Wieselthier, J.E.; Ephremides, A. Towards an effective age of information: Remote estimation of a markov source. In Proceedings of the IEEE INFOCOM 2018—IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Honolulu, HI, USA, 15–19 April 2018; pp. 367–372.
26. Zheng, X.; Zhou, S.; Niu, Z. Context-aware information lapse for timely status updates in remote control systems. In Proceedings of the 2019 IEEE Global Communications Conference (GLOBECOM), Waikoloa, HI, USA, 9–13 December 2019; pp. 1–6.
27. Zheng, X.; Zhou, S.; Niu, Z. Beyond age: Urgency of information for timeliness guarantee in status update systems. In Proceedings of the 2020 2nd IEEE 6G Wireless Summit (6G SUMMIT), Levi, Finland, 17–20 March 2020; pp. 1–5.
28. Zheng, X.; Zhou, S.; Niu, Z. Urgency of Information for Context-Aware Timely Status Updates in Remote Control Systems. *IEEE Trans. Wirel. Commun.* **2020**, *19*, 7237–7250. [CrossRef]
29. Ioannidis, S.; Chaintreau, A.; Massoulié, L. Optimal and scalable distribution of content updates over a mobile social network. In Proceedings of the IEEE INFOCOM 2009, Rio de Janeiro, Brazil, 19–25 April 2009; pp. 1422–1430.
30. Wang, L.; Sun, J.; Zhou, S.; Niu, Z. Timely Status Update Based on Urgency of Information with Statistical Context. In Proceedings of the 2020 32nd IEEE International Teletraffic Congress (ITC 32), Osaka, Japan, 22–24 September 2020; pp. 90–96.
31. Nayyar, A.; Başar, T.; Teneketzis, D.; Veeravalli, V.V. Optimal strategies for communication and remote estimation with an energy harvesting sensor. *IEEE Trans. Autom. Control* **2013**, *58*, 2246–2260. [CrossRef]
32. Cika, A.; Badiu, M.A.; Coon, J.P. Quantifying link stability in Ad Hoc wireless networks subject to Ornstein-Uhlenbeck mobility. In Proceedings of the ICC 2019—2019 IEEE International Conference on Communications (ICC), Shanghai, China, 20–24 May 2019; pp. 1–6.
33. Sennott, L.I. Constrained average cost Markov decision chains. *Probab. Eng. Inf. Sci.* **1993**, *7*, 69–83. [CrossRef]
34. Bertsekas, D.P. *Dynamic Programming and Optimal Control*; Athena Scientific: Belmont, MA, USA, 2000.
35. Sennott, L.I. Average cost optimal stationary policies in infinite state Markov decision processes with unbounded costs. *Oper. Res.* **1989**, *37*, 626–633. [CrossRef]
36. Liu, B.; Xie, Q.; Modiano, E. Rl-qn: A reinforcement learning framework for optimal control of queueing systems. *arXiv* **2020**, arXiv:2011.07401.
37. Chen, X.; Liao, X.; Bidokhti, S.S. Real-time Sampling and Estimation on Random Access Channels: Age of Information and Beyond. *arXiv* **2020**, arXiv:2007.03652.