Sampling to Achieve the Goal: An Age-aware Remote Markov Decision Process

Aimin Li*‡, Shaohua Wu*†, Gary C.F. Lee‡, Xiaomeng Chen*, and Sumei Sun‡, *Fellow, IEEE*
*Harbin Institute of Technology (Shenzhen), China
‡Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore
†The Department of Boradband Communication, Peng Cheng Laboratory, Shenzhen, China
E-mail: {liaimin,23s052026}@stu.hit.edu.cn; hitwush@hit.edu.cn; {Gary_Lee,sunsm}@i2r.a-star.edu.sg

***Abstract*—Age of Information (AoI) has been recognized as an important metric to measure the freshness of information. Central to this consensus is that minimizing AoI can enhance the freshness of information, thereby facilitating the accuracy of subsequent decision-making processes. However, to date the direct causal relationship that links AoI to the utility of the decision-making process is unexplored. To fill this gap, this paper proposes a sampling-control co-design problem, referred to as an *age-aware remote Markov Decision Process* (MDP) problem, to explore this unexplored relationship. Our framework revisits the sampling problem in [1] with a refined focus: moving from AoI penalty minimization to directly optimizing goal-oriented remote decision-making process under random delay. We derive that the *age-aware remote MDP problem* can be reduced to a standard MDP problem without delays, and reveal that treating AoI solely as a metric for optimization is not optimal in achieving remote decision making. Instead, AoI can serve as important *side information* to facilitate remote decision making.**

***Index Terms*—Age of Information, Markov Decision Process, Goal-oriented Communications, Remote Communication-Control Co-Design.**

I. Introduction

Markov Decision Process (MDP) has been a general framework for treating the sequential stochastic control problem [2], [3], and has been applied as an efficient theoretical framework for healthcare management, transportation scheduling, industrial production and automation, response and rescue systems, financial modeling, and *etc*. [4]. Typically, a standard MDP framework assumes immediate access to the current state information, and the decision maker chooses actions based on the available *delay-free* state of the system to achieve a specific goal. This idealization, however, may not hold in many practical scenarios. For instance, in a remote healthcare management system, the monitored patient's condition might be delayed for subsequent healthcare operations. In the Industrial Internet of Things, the transmission of critical safety data to the decision center might be subject to various network delays. These highlight the need for extending the standard MDP to the MDP with observation delays [5], [6].

There are two types of MDP that consider the observation delay, termed deterministic delayed MDP (DDMDP) [5] and stochastic delayed MDP (SDMDP) [6]. The DDMDP introduces a constant observation delay d to the standard MDP framework. At any given time t, the decision-maker accesses the time-varying data as $O(t) = X_{t-d}$. The main result of the DDMDP problem is its reducibility to a standard MDP without delays through *state augmentation*, as detailed by Altman and Nain [5]. The SDMDP extends DDMDP by treating the observation delay not as a static constant but as a random variable D following a given distribution $\Pr(D = d)$, with $O(t) = X_{t-D}$. In 2003, V. Katsikopoulos and E. Engelbrecht showed that an SDMDP is also reducible to a standard MDP problem without delay [6]. Thus, it becomes clear to solve a SDMDP problem by solving its equivalent standard MDP.

TABLE I
Comparisons of Time-Lag MDPs

Type	Observation	Reference
MDP Without Delay	$O(t) = X_t$	[3]
DDMDP	$O(t) = X_{t-d}$	[5]
SDMDP	$O(t) = X_{t-D}$	[6]
Age-Aware Remote MDP	$O(t) = X_{t-\Delta(t)}$	This Work

However, the above time-lag MDPs, where the observation delay follows a given distribution, potentially assumes that the state is sampled and transmitted to the decision maker at every time slot[1]. This setup presumes that the system can transmit every state information without encountering any *backlog*. In practice, constantly sampling and transmitting may result in infinitely accumulated packets in the queue, resulting in severe congestion. This motivates the need for queue control and adaptive sampling policy design in the network [1], [7]–[11], where Age of Information (AoI) has emerged as an important indicator [12]–[15]. Currently, AoI has been applied in a wide range of applications such as queue control [16]–[19], remote estimation [8], [20]–[24], and communications & network design [25]–[33]. Suppose the i-th sample is generated at time S_i and is delivered at the receiver at time D_i, AoI is defined as a *sawtooth piecewise function*:

$$\Delta(t) = t - S_i, D_i \leq t < D_{i+1}, \forall i \in \mathbb{N}, \tag{1}$$

as shown in Fig. 1. From this definition, the most recently available information at the receiver at time slot t is $O(t) =$

This work has been supported in part by the National Key Research and Development Program of China under Grant no. 2020YFB1806403, and in part by the Guangdong Basic and Applied Basic Research Foundation under Grant no. 2022B1515120002, and in part by the Major Key Project of PCL under Grant no. PCL2024A01.

[1] In this case, each state $X_i, \forall i \in \{0, 1, ...n\}$ are all sampled and forwarded to the decision maker. The observation delay D is i.i.d random variable and is independent of the sampling policy.

2024 IEEE Information Theory Workshop (ITW) | 979-8-3503-4893-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/ITW61385.2024.10806969

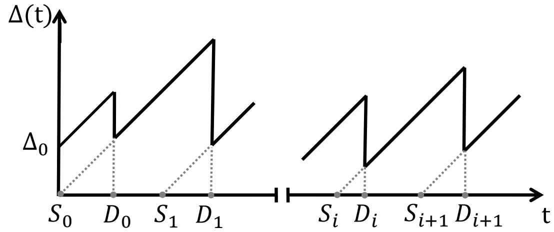

Fig. 1. Evolution of the age $\Delta(t)$ over time.

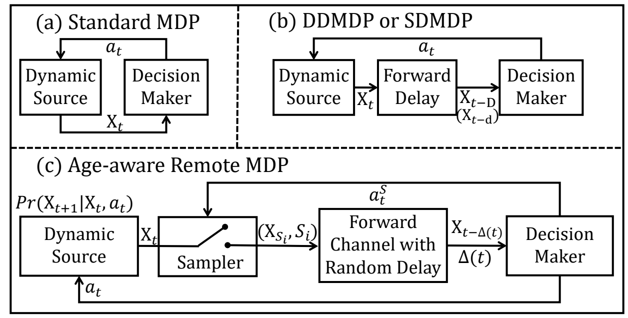

Fig. 2. Comparisons among standard MDP, DDMDP, SDMDP, and *age-aware remote MDP*.

$X_{t-\Delta(t)}$. Different from the DDMDP and SDMDP where the time lag is a constant d or an i.i.d random variable D, with $O(t) = X_{t-d}$ or $O(t) = X_{t-D}$, the time lag in the practical network with queue and finite transmission rate is a controlled random process $\Delta(t)$, with $O(t) = X_{t-\Delta(t)}$. See Table I and Fig. 2 for the comparisons of different time-lag MDPs.

Motivated by the above, this work enriches the time-lag MDP family by treating the observation delay not as a stationary stochastic variable d or D, but as a dynamically controlled stochastic process, represented by *age* $\Delta(t)$ defined in (1). We refer to this problem as *age-aware remote MDP* problem[2], where AoI serves as no longer a typical indicator but important *side information* for goal-oriented remote decision making. Our main result is that *age-aware remote MDP*, like DDMDP [5] and SDMDP [6], can also be reduced to a standard MDP problem with a constraint. We design efficient algorithms to solve this type of standard MDP with a constraint. *To the best of our knowledge, this is the first work that introduces AoI into the time-lag MDP family, and the first work that explores AoI's new role as side information to facilitate remote decision making under random delay.*

II. System Model and Problem Formulation

We consider a time-slotted *age-aware remote MDP* problem illustrated in Fig. 2(c). Let $X_t \in \mathcal{S}$ be the controlled source at time slot t. The evolution of the source is a Markov decision process, characterized by the transition probability $\Pr(X_{t+1}|X_t, a_t)$[3], where $a_t \in \mathcal{A}$ represents the action taken by the remote decision maker to control the source in the desired way. The sampler conducts the sampling action $a_t^S \in \{0,1\}$, with $a_t^S = 1$ representing the sampling action and $a_t^S = 0$ otherwise. Consider the random channel delay of the i-th packet as $Y_i \in \mathcal{Y} \subseteq \mathbb{N}^+$, which is independent of the source X_t and is bounded $\max[Y_i] < \infty$. The sampling times $S_0, S_1, \cdots$ shown in Fig. 1 record the time stamp with $a_t^S = 1$,

$$S_i = \max\{t \,\big|\, t \le D_i, a_t^S = 1\}, \tag{2}$$

where the initial state of the system is $S_0 = 0$ and $\Delta(0) = \Delta_0$.

At the sampling time $S_i, \forall i \in \mathbb{N}$, the state X_{S_i} along with the corresponding time stamp S_i are encapsulated into a packet (X_{S_i}, S_i), which is transmitted to a remote decision maker. Upon the reception of the packet (X_{S_i}, S_i) during $t \in [D_i, D_{i+1})$, the *observation history* at the decision maker is $\{(X_{S_j}, S_j) : j \le i\}$. By employing (1), this sequence can be equivalently represented as $\{(X_{k-\Delta(k)}, \Delta(k)) : k \le t\}$.

Denote $\mathcal{I}_t = \{(X_{k-\Delta(k)}, \Delta(k), a_{k-1}^S, a_{k-1}) : k \le t\}$ as the *observation history* and *action history* (or simply *history*) available to the decision maker up to time t. The decision maker is tasked with determining both the sampling action and the controlled action $(a_t^S, a_t) \in \{0,1\} \times \mathcal{A}$ at each time slot t by leveraging the *history* $\mathcal{I}_t$. A decision policy of the decision maker is defined as a mapping from the *history* to a distribution over the joint action space $\{0,1\} \times \mathcal{A}$, denoted by $\pi_t : \mathcal{I}_t \to \mathrm{Prob}(\mathcal{A} \times \{0,1\})$. Similar to [1], we assume that the sampling policy satisfies two conditions:

- $i)$ No sample is taken when the channel is busy, *i.e.*, $S_{i+1} \ge D_i$, or equivalently,
$$S_{i+1} - Z_i = D_i \text{ with } Z_i \ge 0, i \in \mathbb{N}, \tag{3}$$
with Z_i representing the sampling waiting time. From this assumption, we can also obtain that $D_i = S_i + Y_i, \forall i \in \mathbb{N}$.
- $ii)$ The inter-sample times $G_i = S_{i+1} - S_i$ is a *regenerative process* [35, Section 6.1]: There is a sequence $0 \le g_1 < g_2 < \cdots$ of almost surely finite random integers such that the post-g_j process $\{G_{g_j+i}\}_{i\in\mathbb{N}}$ has the same distribution as the post-g_1 process $\{G_{g_1+i}\}_{i\in\mathbb{N}}$ and is independent of the pre-g_j process $\{G_{g_1+i}\}_{i\in\{0,1,\cdots k_j-1\}}$. Condition $ii)$ implies that, almost surely [4]
$$\lim_{i\to\infty} S_i = \infty, \quad \lim_{i\to\infty} D_i = \infty. \tag{4}$$

In addition, we assume that the controlled policy satisfies the following condition: the action a_t is updated only upon the delivery of a sample X_{S_i}[5], *i.e.*,

$$a_t = A_i, D_i \le t \le D_{i+1}, i \in \mathbb{N}, \tag{5}$$

where $A_i \in \mathcal{A}$ is the updated controlled action upon the delivery of packet (X_{S_i}, S_i). We consider bounded cost func-

[2]In [34], the term *remote MDP* was first proposed as a pathway to pragmatic or goal-oriented communications. Our paper follows this research and focuses on the effect from *age*, hence the term *age-aware remote MDP*.

[3]For short-hand notations, we use the transition probability matrix $\mathbf{P}_a$ to encapsulate the dynamics of the source given an action $a_t = a$.

[4]This assumption also implies that the waiting time Z_i is bounded, belonging to a subset of nature numbers with $Z_i \in \mathcal{Z} \subseteq \mathbb{N}$.

[5]Throughout the time interval $t \in [D_i, D_{i+1})$, the decision maker's observations are fixed and consist only of $(X_{S_j}, S_j) : j \le i$. Thus, it is assumed that the chosen actions remain constant in this period. The possibility of varying these actions within such intervals will be our future work.

tion $\mathcal{C}(X_t, a_t) < \infty$, which represents the immediate cost incurred when action a_t is taken in state X_t. Under the above assumptions, the objective of the system is to design the optimal policies at each time slot $\pi_0, \pi_1, \pi_2 \cdots$ to minimize the *long-term average cost* [6]:

$$\mathcal{P}1: \inf_{\pi_{0:\infty}} \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} \mathcal{C}(X_t, a_t)\right]. \tag{6}$$

This problem aims at determining the distribution of joint sampling and controlled actions (a_t^S, a_t) based on the history $\mathcal{I}_t$, such that the *long-term average cost* is minimized.

III. Optimal Sampling and Remote Decision Making Policy Under Random Delay

A. Sufficient Statistics of History

The policy π_t is a mapping from $\mathcal{I}_t$ to the action provability space. One challenge to solving π_t is that the *history* space $\mathcal{I}_t$ explodes exponentially as t increases. This motivates us to compress and abstract information that is necessary for the optimal decision process. Specifically, we will analyze the *sufficient statistics* in this subsection.

Definition 1. *A sufficient statistics of $\mathcal{I}_t$ is a function $S_t(\mathcal{I}_t)$, such that $\min_{a_{t:T}} \mathbb{E}\left[\sum_{k=t}^{T} \mathcal{C}(X_k, a_k)|\mathcal{I}_t\right] = \min_{a_{t:T}} \mathbb{E}\left[\sum_{k=t}^{T} \mathcal{C}(X_k, a_k)|S_t(\mathcal{I}_t)\right]$ holds for any $T > t$.*

The above definition implies that the decision making based on the *sufficient statistics* $H_t(\mathcal{I}_t)$ can achieve an equivalent performance as that dependent on $\mathcal{I}_t$. The following lemma introduces an important *sufficient statistics* of $\mathcal{I}_t$.

Lemma 1. *In our problem, during the interval $t \in [D_i, D_{i+1})$, $\mathcal{G}_i = (X_{S_i}, Y_i, A_{i-1}) \in \mathcal{S}\times\mathcal{Y}\times\mathcal{A}$ is a sufficient statistics of $\mathcal{I}_t$. Besides, determining the optimal sampling actions a_t^S under condition (3) is equivalent to determining the optimal sampling time S_{i+1}, or the optimal waiting time Z_i.*

Proof. See the full version [36, Appendix A] for the proof. ■

With Lemma 1, determining the optimal policy $\pi_t : \mathcal{I}_t \to \text{Prob}(\mathcal{A}\times\{0,1\})$ for $\mathcal{P}1$ becomes equivalent to solving for an alternative policy $\phi_t : \mathcal{S}\times\mathcal{Y}\times\mathcal{A} \to \text{Prob}(\mathbb{N}\times\mathcal{A})$, which is a mapping from $\mathcal{G}_i$ to a distribution of (Z_i, A_i). The problem $\mathcal{P}1$ is thus rewritten as

$$\mathcal{P}2: \inf_{\phi_{0:\infty}} \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} \mathcal{C}(X_t, a_t)\right]. \tag{7}$$

B. Simplification of Problem $\mathcal{P}2$

As G_i is a *regenerative process* and $\lim_{i\to\infty} D_i = \infty$, we can rewrite the objective function in problem $\mathcal{P}2$ as

$$\begin{aligned}
&\limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} \mathcal{C}(X_t, a_t)\right] \\
&= \lim_{D_n\to\infty} \frac{\mathbb{E}\left[\sum_{t=1}^{D_n} \mathcal{C}(X_t, a(t))\right]}{\mathbb{E}[D_n]} \\
&= \lim_{n\to\infty} \frac{\sum_{i=0}^{n-1}\mathbb{E}\left[\sum_{t=D_i}^{D_{i+1}-1} \mathcal{C}(X_t, A_i)\right]}{\sum_{i=0}^{n-1}\mathbb{E}[D_{i+1}-D_i]} \\
&= \lim_{n\to\infty} \frac{\sum_{i=0}^{n-1}\mathbb{E}\left[\sum_{t=D_i}^{D_{i+1}-1} \mathcal{C}(X_t, A_i)\right]}{\sum_{i=0}^{n-1}\mathbb{E}[Y_{i+1}+Z_i]}.
\end{aligned} \tag{8}$$

Then Problem $\mathcal{P}2$ can be rewritten as

$$\mathcal{P}3: h^* \triangleq \inf_{\phi_{0:\infty}} \lim_{n\to\infty} \frac{\sum_{i=0}^{n-1}\mathbb{E}\left[\sum_{t=D_i}^{D_{i+1}-1} \mathcal{C}(X_t, A_i)\right]}{\sum_{i=0}^{n-1}\mathbb{E}[Y_{i+1}+Z_i]}. \tag{9}$$

To solve Problem $\mathcal{P}3$, we consider the following problem with parameter $\lambda \geq 0$:

$$\begin{aligned}
&\mathcal{P}4: U(\lambda) \triangleq \\
&\inf_{\phi_{0:\infty}} \lim_{n\to\infty} \frac{1}{n}\sum_{i=0}^{n-1}\left\{\mathbb{E}\left[\sum_{t=D_i}^{D_{i+1}-1} \mathcal{C}(X_t, A_i)\right] - \lambda\mathbb{E}[Z_i+Y_{i+1}]\right\},
\end{aligned} \tag{10}$$

By similarly applying Dinkelbach's method as in [37] and [20, Lemma 2], we obtain the following lemma:

Lemma 2. *The following assertions hold:*
(i). $h^* \gtreqless \lambda$ if and only if $U(\lambda) \gtreqless 0$.
(ii). When $U(\lambda) = 0$, the solutions to Problem $\mathcal{P}4$ coincide with those of Problem $\mathcal{P}3$.
(iii). $U(\lambda) = 0$ has a unique root, and the root is h^*.

Proof. See the full version [36, Appendix B] for the proof. ■

Following Lemma 2, solving Problem $\mathcal{P}3$ can be simplified to solving Problem $\mathcal{P}4$ under $U(\lambda) = 0$. Our remaining goal is to solve $\mathcal{P}4$ and search h^* such that $U(h^*) = 0$.

C. Reformulate Problem $\mathcal{P}4$ as a Standard MDP

In this subsection, we formulate Problem $\mathcal{P}4$ as a standard infinite horizon MDP problem. We introduce the state space, action space, transition probability, and cost function of the MDP problem in this subsection specifically. This MDP problem with parameter λ is denoted as $\mathscr{P}_{\text{MDP}}(\lambda)$:

- **State Space**: the state of the equivalent MDP is the sufficient statistics $\mathcal{G}_i = (X_{S_i}, Y_i, A_{i-1}) \in \mathcal{S}\times\mathcal{Y}\times\mathcal{A}$.
- **Action Space**: the actions space of the MDP composed by the tuple $(Z_i, A_i) \in \mathcal{Z}\times\mathcal{A}$, where Z_i is the sampling waiting time and A_i is the controlled actions.
- **Transition Probability**: The transition probability is defined by $\Pr(\mathcal{G}_{i+1}|\mathcal{G}_i, Z_i, A_i)$. We have the transition probability as (see [36, Appendix C] for the proof):

$$\begin{aligned}
&\Pr(\mathcal{G}_i = (s', \delta', a')|\mathcal{G}_i = (s, \delta, a), Z_i, A_i) \\
&= \Pr(Y_{i+1} = \delta')\cdot[\mathbf{P}_a^{\delta}\cdot\mathbf{P}_{a_i}^{Z_i}]_{s\times s'}\cdot\mathbb{1}\{a' = A_i\},
\end{aligned} \tag{11}$$

- **Cost Function**: the cost function is typically a real-valued function over the state space and the action space. We denote the cost function as $g(\mathcal{G}_i, Z_i, A_i)$, and next

[6]In Theorem 1, we will discuss the unichain property of the MDP, therefore the problem is independent of the initial state X_0.

show that we can tailor the cost function to establish an identical standard MDP of Problem $\mathcal{P}4$.

Lemma 3. *If the cost function is defined by*

$$g(\mathcal{G}_i, Z_i, A_i; \lambda) \triangleq q(\mathcal{G}_i, Z_i, A_i) - \lambda f(Z_i), \tag{12}$$

$$\text{where} \qquad f(Z_i) = Z_i + \mathbb{E}\left[Y_{i+1}\right], \tag{13}$$

$$q(\mathcal{G}_i, Z_i, A_i) = q(X_{S_i}, Y_i, A_{i-1}, Z_i, A_i)$$
$$= \mathbb{E}\left[\sum_{s' \in \mathcal{S}} \left[\sum_{t=0}^{Z_i + Y_{i+1} - 1} \mathbf{P}_{A_{i-1}}^{Y_i} \cdot \mathbf{P}_{A_i}^{t}\right]_{X_{S_i} \times s'} \cdot \mathcal{C}(s', A_i)\right], \tag{14}$$

then Problem $\mathscr{P}_{\mathrm{MDP}}(\lambda)$ is identical to Problem $\mathcal{P}4$.

Proof. See the full version [36, Appendix D] for the proof. ■

Our remaining focus is to solve $\mathscr{P}_{\mathrm{MDP}}(\lambda)$ and seek a value h^* such that $U(h^*) = 0$.

D. Existence of Optimal Stationary Deterministic Policy

We examine the sufficient conditions required for the existence of a *stationary deterministic* policy within $\mathscr{P}_{\mathrm{MDP}}(\lambda)$. Our main result is described in the following theorem:

Theorem 1. *If an MDP characterized by finite state space $\mathcal{S}$, finite action space $\mathcal{A}$, and transition probability $\mathbf{P}_a$ is a unichain, then: (i) the transformed Age-aware remote MDP $\mathscr{P}_{\mathrm{MDP}}(\lambda)$ is also a unichain; (ii) an optimal stationary deterministic policy exists for $\mathscr{P}_{\mathrm{MDP}}(\lambda)$.*

Proof. See the full version [36, Appendix E] for the proof. ■

E. Numerical Solutions

We propose two algorithms to solve the infinite-horizon MDP and seek the parameter h^* such that $U(h^*) = 0$.

• *Bisec-MRVI*: The first one is a *two-layer* algorithm. The outer layer is based on the *bisection-search* method: the search interval $(\lambda_{\downarrow}^{(k)}, \lambda_{\uparrow}^{(k)})$ is iteratively narrowed down by half until the interval can closely approximate the value h^* such that $U(h^*) = 0$. The internal layer utilizes a *modified Relative Value Iteration* (MRVI) [38, Eq. 4.72] to compute the value $U(\lambda)$ by resolving the MDP $\mathscr{P}_{\mathrm{MDP}}(\lambda)$ [7]. A similar *two-layer bisection-based* method has been introduced in [1] and [20] to achieve Age-optimal or Mean Square Error (MSE)-optimal sampling. We note the complexity of *Bisec-RVI* algorithm is dependent on the initialization of the search interval $(\lambda_{\downarrow}, \lambda_{\uparrow})$, and thus establish an upper and lower bound of h^* here:

Lemma 4. *The lower bound of h^* is given by*

$$h^* \geq \min_{s,a} \mathcal{C}(s, a). \tag{15}$$

The upper bound of h^* is given by

$$h^* \leq \min_{a} \sum_{s \in \mathcal{S}} \pi_a(s) \cdot \mathcal{C}(s, a), \tag{16}$$

where $\pi_a(s)$ represents the stationary distribution of state s, corresponding to the transition probability matrix $\mathbf{P}_a$.

Proof. See the full version [36, Appendix F] for the proof. ■

The *Bisec-MRVI* algorithm is in Algorithm 1, where the details of the *MRVI* is demonstrated in [38, Eq. 4.72].

Algorithm 1: *Bisec-MRVI* algorithm

```
Input: Tolerence ϵ > 0, MDP 𝒫_MDP(λ)
Initialization: λ↑ = min_a Σ_{s∈S} π_a(s) · C(s, a),
    λ↓ = min_{s,a} C(s, a) ;
while λ↑ − λ↓ ≥ ϵ do
    λ = (λ↑ + λ↓)/2;
    Run MRVI to solve 𝒫_MDP(λ) and calculate U(λ);
    if U(λ) > 0 then
        λ↓ = λ;
    else
        λ↑ = λ;
Output: h* = λ
```

• *Fixed-Point-Based Iteration (FPBI)*: The *two-layer Bisec-MRVI* algorithm requires repeatedly executing the MRVI algorithm in the inner layer, which is *computation-intensive*. This motivates us to propose a *one-layer* algorithm, termed FPBI algorithm, to avoid the computation overhead. Unlike *Bisec-MRVI*, The *one-layer* FPBI algorithm treats $U(h^*) = 0$ as a constraint within the Markov Decision Process $\mathscr{P}_{\mathrm{MDP}}(h^*)$. Our main result is the following equivalent equations.

Theorem 2. *Solving the root finding problem of $U(\lambda)$ is equivalent to solving the following nonlinear equations:*

$$\begin{cases} W^*(\gamma) = \min\limits_{A_i, Z_i} \left\{ g(\gamma, A_i, Z_i; h^*) + \mathbb{E}[W^*(\gamma') \,|\gamma, Z_i, A_i] \right\} \\ \textit{for } \gamma \in \mathcal{S} \times \mathcal{Y} \times \mathcal{A}, \\ h^* = \min\limits_{A_i, Z_i} \left\{ \dfrac{q\left(\gamma^{\mathrm{ref}}, A_i, Z_i\right) + \mathbb{E}[W^*(\gamma') \,|\gamma^{\mathrm{ref}}, A_i, Z_i]}{f\left(Z_i\right)} \right\}, \end{cases} \tag{17}$$

where $\gamma^{\mathrm{ref}} \in \mathcal{S} \times \mathcal{Y} \times \mathcal{A}$ can be arbitrarily chosen.

Proof. See the full version [36, Appendix G] for the proof. ■

In (17), there are $|\mathcal{S}| \times |\mathcal{Y}| \times |\mathcal{A}| + 1$ variables and an equal number of non-linear equations. In what follows we reveal (17) forms a set of *fixed-point equations*, which can be effectively solved through *fixed-point iterations* [39].

Let $\mathbf{W}^*$ denote the vector consisting of $W^*(\gamma)$ for all $\gamma \in \mathcal{S} \times \mathcal{Y} \times \mathcal{A}$, (17) can be succinctly represented as:

$$\begin{cases} \mathbf{W}^* = T(\mathbf{W}^*, h^*) \\ h^* = H(\mathbf{W}^*). \end{cases} \tag{18}$$

Then, substituting the second equation $h^* = H(\mathbf{W}^*)$ into the first equation of (18) yields $\mathbf{W}^* = T(\mathbf{W}^*, H(\mathbf{W}^*))$. Define $Q(\mathbf{W}) \triangleq T(\mathbf{W}, H(\mathbf{W}))$ for simplification. Consequently, as implied by (18), we have the *fixed-point equation*:

$$\mathbf{W}^* = Q(\mathbf{W}^*), \tag{19}$$

[7]Different from a standard RVI algorithm, the MRVI algorithm ensures that the algorithm converges under a weaker condition that the Markov chain under a given stationary policy is *periodic* [38, Proposition 4.3.4].

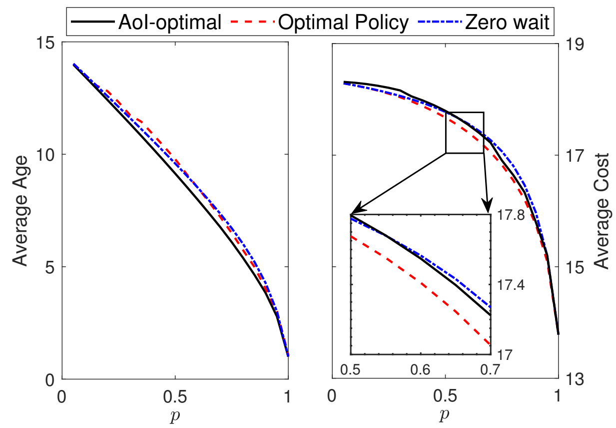

Fig. 3. Average age and average cost vs. p with *i.i.d* random delay Y_i, where $\Pr(Y_i = 1) = p$ and $\Pr(Y_i = 10) = 1 - p$.

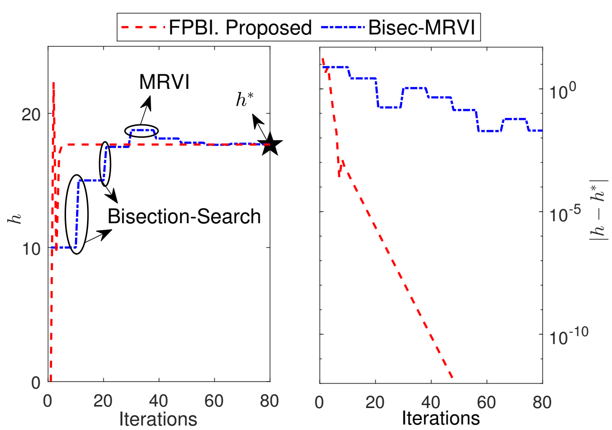

Fig. 4. *Bisec-MRVI* vs. *FPBI*, where $p = 0.5$.

which can be solved by *fixed-point iteration* in Algorithm 2.

Algorithm 2: FPBI for Solving (17)

Input: MDP $\mathscr{P}_{\mathrm{MDP}}(\lambda)$, Tolerence ϵ;
Initialization: $\mathbf{W}^*_{(0)} = \mathbf{0}$, $\mathbf{W}^*_{(-1)} = \infty$, $k = 0$;
Choose γ^{ref} arbitrarily;
while $||\tilde{V}^k_{\pi_A}(\mathbf{w}) - \tilde{V}^{k-1}_{\pi_A}(\mathbf{w})|| \geq \epsilon$ **do**
$k = k + 1$;
$h^*_{(k)} = H(\mathbf{W}^*_{(k-1)})$;
$\mathbf{W}^*_{(k)} = T(\mathbf{W}^*_{(k-1)}, h^*_{(k)})$;
// See (17) and (18) for $T(\cdot)$ and $H(\cdot)$.
Output: $h^* = h^*_{(k)}$

IV. Simulation Results

In this paper, the policy that minimizes the *long-term average cost* in Problem $\mathcal{P}1$ is referred to as "optimal policy". We compare it with the following two benchmarks:

- *Zero wait*: An update is transmitted once the previous update is delivered, *i.e.*, $Z_i = 0$ for $\forall i$. This policy achieves the minimum delay and maximum throughput. The controlled action a_i is *age-aware* and is obtained by substituting $Z_i = 0$ into the RHS of (17) and similarly implementing the *fixed-point iteration*, implying *age-aware optimal control* under *zero-wait* sampling.
- *AoI-optimal*: The *AoI-optimal* policy determines Z_i by [1, Theorem 4], which is a threshold-based policy $Z_i = \max(0, \beta - Y_i)$, where β is numerically solved by [1, Algorithm 2]. The controlled action a_i is *age-aware* and is obtained by fixing the *AoI-optimal* sampling in (17) and similarly implementing the *fixed-point iteration*, implying *age-aware optimal control* under *AoI-optimal* sampling.

As a case study, we consider the parameters detailed in [36, Appendix H] for the simulation setup.

Fig. 3 compares the benchmarks with "optimal policy" in terms of average age and cost. The left panel demonstrates that the *AoI-optimal* policy consistently achieves the lowest *age*. However, the right panel reveals a *counterintuitive* result: the *AoI-optimal* policy does not necessarily lead to the best decision-making performance. This suggests that the *value-of-information* transcends mere age freshness; it is also shaped by the specific goal of the receiver and the semantic content of the information [40]–[48]. The "optimal policy" consistently yields the lowest average cost.

Fig. 4 compares *FPBI* with *Bisec-MRVI*, where the latter is the benchmark. This benchmark is inspired by the *two-layer* solution frameworks by [1], [20] and [49]. The left panel records the trajectories of h across algorithms iterations, illustrating that both *Bisec-MRVI* and *FPBI* converge to the optimal h^*. For *Bisec-MRVI*, updates to h are deponent on the *outer-layer bisection search*, which occur only upon the convergence of the *inter-layer* MRVI. In contrast, our developed *one-layer FPBI* eliminates the need for the *outer-layer bisection search* by fixing $U(h^*) = 0$ and directly reestablishing *fixed point equations* for $\mathscr{P}_{\mathrm{MDP}}(h^*)$ under this condition. The right panel records the distances to h^* across algorithms iterations, which demonstrated that *FPBI* algorithm achieves faster convergence compared *Bisec-MRVI*.

V. Conclusion

In this paper, we have proposed a new remote MDP problem in the time-lag MDP framework, termed *age-aware remote MDP*. Specifically, AoI, typically an optimization indicator to ensure information *freshness*, has been introduced into the remote MDP problem as a specific category of controlled random processing delay and as important *side information* to enhance remote decision-making. The main result of this work is that the age-aware remote MDP can be reformulated into a standard, *delay-free* MDP. Under such an equivalent problem, we have established sufficient conditions for the existence of an optimal *stationary deterministic* policy. Additionally, we have developed a low-complexity *one-layer* algorithm to effectively solve this remote MDP problem. We revealed that the age-optimal policy, which ensures the freshest information, does not necessarily achieve the best remote decision making. In contrast, we can design goal-oriented sampling policy that directly optimize remote decision making.

References

[1] Y. Sun, E. Uysal, R. D. Yates, C. E. Koksal, and N. B. Shroff, "Update or wait: How to keep your data fresh," *IEEE Trans. Inf. Theory*, vol. 63, no. 11, pp. 7492–7508, 2017.

[2] R. A. Howard, *Dynamic programming and markov processes*. Cambridge, MA; MIT Press, 1960.

[3] R. Bellman, "Dynamic programming," *Science*, vol. 153, no. 3731, pp. 34–37, 1966.

[4] R. J. Boucherie and N. M. Van Dijk, *Markov decision processes in practice*. Cham, Switzerland: Springer, 2017.

[5] E. Altman and P. Nain, "Closed-loop control with delayed information," *Perf. Eval. Rev.*, vol. 14, pp. 193–204, 1992.

[6] K. V. Katsikopoulos and S. E. Engelbrecht, "Markov decision processes with delays and asynchronous cost collection," *IEEE Trans. Autom. Control.*, vol. 48, no. 4, pp. 568–574, 2003.

[7] R. D. Yates, "Lazy is timely: Status updates by an energy harvesting source," in *IEEE Proc. ISIT*, 2015, pp. 3008–3012.

[8] A. Arafa, K. Banawan, K. G. Seddik, and H. V. Poor, "Sample, quantize, and encode: Timely estimation over noisy channels," *IEEE Trans. Commun.*, vol. 69, no. 10, pp. 6485–6499, 2021.

[9] H. Tang, Y. Chen, J. Wang, P. Yang, and L. Tassiulas, "Age optimal sampling under unknown delay statistics," *IEEE Trans. Inf. Theory*, vol. 69, no. 2, pp. 1295–1314, 2023.

[10] J. Pan, A. M. Bedewy, Y. Sun, and N. B. Shroff, "Optimal sampling for data freshness: Unreliable transmissions with random two-way delay," *IEEE/ACM Trans. Netw.*, vol. 31, no. 1, pp. 408–420, 2023.

[11] B. Zhou and W. Saad, "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.

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

[13] A. Kosta, N. Pappas, and V. Angelakis, "Age of Information: A new concept, metric, and tool," *Found. Trends Netw.*, vol. 12, no. 3, pp. 162–259, 2017.

[14] R. D. Yates, Y. Sun, D. R. Brown, S. K. Kaul, E. H. Modiano, and S. Ulukus, "Age of Information: An Introduction and Survey," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 5, pp. 1183–1210, 2021.

[15] J. Cao, X. Zhu, S. Sun, Z. Wei, Y. Jiang, J. Wang, and V. K. Lau, "Toward industrial metaverse: Age of Information, latency and reliability of short-packet transmission in 6G," *IEEE Wirel. Commun.*, vol. 30, no. 2, pp. 40–47, 2023.

[16] M. Costa, M. Codreanu, and A. Ephremides, "On the age of information in status update systems with packet management," *IEEE Trans. Inf. Theory*, vol. 62, no. 4, pp. 1897–1910, 2016.

[17] O. Dogan and N. Akar, "The multi-source probabilistically preemptive M/PH/1/1 Queue With packet errors," *IEEE Trans. Commun.*, vol. 69, no. 11, pp. 7297–7308, 2021.

[18] R. D. Yates and S. K. Kaul, "The Age of Information: Real-time status updating by multiple sources," *IEEE Trans. Inf. Theory*, vol. 65, no. 3.

[19] C. Kam, S. Kompella, G. D. Nguyen, and A. Ephremides, "Effect of message transmission path diversity on status age," *IEEE Trans. Inf. Theory*, vol. 62, no. 3, pp. 1360–1374, 2015.

[20] Y. Sun, Y. Polyanskiy, and E. Uysal, "Sampling of the Wiener process for remote estimation over a channel with random delay," *IEEE Trans. Inf. Theory*, vol. 66, no. 2, pp. 1118–1135, 2019.

[21] T. Z. Ornee and Y. Sun, "Sampling for remote estimation through queues: Age of Information and beyond," in *IEEE Proc. WiOPT*, 2019, pp. 1–8.

[22] H. Tang, Y. Sun, and L. Tassiulas, "Sampling of the wiener process for remote estimation over a channel with unknown delay statistics," in *Proc. ACM MobiHoc*, 2022, p. 51–60.

[23] C.-H. Tsai and C.-C. Wang, "Unifying AoI minimization and remote estimation—Optimal sensor/controller coordination with random two-way delay," *IEEE/ACM Trans. Netw.*, vol. 30, no. 1, pp. 229–242, 2021.

[24] T. Z. Ornee and Y. Sun, "Sampling and remote estimation for the Ornstein-Uhlenbeck process through queues: Age of Information and beyond," *IEEE/ACM Trans. Netw.*, vol. 29, no. 5, pp. 1962–1975, 2021.

[25] A. Li, S. Wu, J. Jiao, N. Zhang, and Q. Zhang, "Age of Information with Hybrid-ARQ: A Unified Explicit Result," *IEEE Trans. Commun.*, vol. 70, no. 12, pp. 7899–7914, 2022.

[26] H. Pan, T.-T. Chan, V. C. Leung, and J. Li, "Age of Information in Physical-layer Network Coding Enabled Two-way Relay Networks," *IEEE Trans. Mob. Comput.*, 2022.

[27] M. Xie, Q. Wang, J. Gong, and X. Ma, "Age and energy analysis for LDPC coded status update with and without ARQ," *IEEE Internet Things J.*, vol. 7, no. 10, pp. 10 388–10 400, 2020.

[28] S. Meng, S. Wu, A. Li, J. Jiao, N. Zhang, and Q. Zhang, "Analysis and optimization of the HARQ-based Spinal coded timely status update system," *IEEE Trans. Commun.*, vol. 70, no. 10, pp. 6425–6440, 2022.

[29] J. Cao, X. Zhu, Y. Jiang, Z. Wei, and S. Sun, "Information age-delay correlation and optimization with Finite Block Length," *IEEE Trans. Commun.*, vol. 69, no. 11, pp. 7236–7250, 2021.

[30] Y. Long, W. Zhang, S. Gong, X. Luo, and D. Niyato, "AoI-aware scheduling and trajectory optimization for multi-UAV-assisted wireless networks," in *IEEE GLOBECOM*, 2022, pp. 2163–2168.

[31] H. Feng, J. Wang, Z. Fang, J. Chen, and D. Do, "Evaluating AoI-centric HARQ protocols for UAV networks," *IEEE Trans. Commun.*, vol. 72, no. 1, pp. 288–301, 2024.

[32] H. Pan, T. Chan, V. C. M. Leung, and J. Li, "Age of Information in physical-layer network coding enabled two-way relay networks," *IEEE Trans. Mob. Comput.*, vol. 22, no. 8, pp. 4485–4499, 2023.

[33] Y. Long, S. Zhao, S. Gong, B. Gu, D. Niyato, and X. Shen, "AoI-aware sensing scheduling and trajectory optimization for multi-UAV-assisted wireless backscatter networks," *IEEE Trans. Veh. Technol.*, pp. 1–16, 2024.

[34] D. Gündüz, F. Chiariotti, K. Huang, A. E. Kalør, S. Kobus, and P. Popovski, "Timely and massive communication in 6G: Pragmatics, learning, and inference," *IEEE BITS Inf. Theory Mag.*, vol. 3, no. 1, pp. 27–40, 2023.

[35] P. J. Haas, *Stochastic petri nets: Modelling, stability, simulation*. Springer Science & Business Media, 2006.

[36] A. Li, S. Wu, G. C. F. Lee, X. Cheng, and S. Sun, "Sampling to achieve the goal: An age-aware remote markov decision process," *arXiv*, 2024. [Online]. Available: https://arxiv.org/abs/2405.02042

[37] W. Dinkelbach, "On nonlinear fractional programming," *Management science*, vol. 13, no. 7, pp. 492–498, 1967.

[38] D. Bertsekas, *Dynamic Programming and Optimal Control: Volume II*. Athena scientific, 2012, vol. 2.

[39] F. B. Hildebrand, *Introduction to numerical analysis*. Courier Corporation, 1987.

[40] Y. Wang, S. Wu, C. Lei, J. Jiao, and Q. Zhang, "A review on wireless networked control system: The communication perspective," *IEEE Internet Things J.*, 2023.

[41] N. Pappas and M. Kountouris, "Goal-oriented communication for real-time tracking in autonomous systems," in *IEEE Proc. ICAS*, 2021, pp. 1–5.

[42] S. Meng, S. Wu, J. Zhang, J. Cheng, H. Zhou, and Q. Zhang, "Semantics-empowered space-air-ground-sea integrated network: New paradigm, frameworks, and challenges," *IEEE Commun. Surv. Tutorials*, pp. 1–1, 2024.

[43] M. Kountouris and N. Pappas, "Semantics-empowered communication for networked intelligent systems," *IEEE Commun. Mag.*, vol. 59, no. 6, pp. 96–102, 2021.

[44] E. Uysal, O. Kaya, A. Ephremides, J. Gross, M. Codreanu, P. Popovski, M. Assaad, G. Liva, A. Munari, B. Soret, T. Soleymani, and K. H. Johansson, "Semantic communications in networked systems: A data significance perspective," *IEEE Netw.*, vol. 36, no. 4, pp. 233–240, 2022.

[45] C. Ari, M. K. C. Shisher, E. Uysal, and Y. Sun, "Goal-oriented communications for remote inference with two-way delay," *arXiv*, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2311.11143

[46] A. Li, S. Wu, S. Meng, R. Lu, S. Sun, and Q. Zhang, "Toward goal-oriented semantic communications: New metrics, framework, and open challenges," *IEEE Wirel. Commun.*, pp. 1–8, 2024.

[47] A. Li, S. Wu, S. Sun, and J. Cao, "Goal-oriented tensor: Beyond age of information towards semantics-empowered goal-oriented communications," *IEEE Trans. Commun.*, pp. 1–1, 2024.

[48] M. Salimnejad, M. Kountouris, and N. Pappas, "Real-time reconstruction of markov sources and remote actuation over wireless channels," *IEEE Trans. Commun.*, pp. 1–1, 2024.

[49] Y. Sun and B. Cyr, "Sampling for data freshness optimization: Non-linear age functions," *J. Commun. Netw.*, vol. 21, no. 3, pp. 204–219, 2019.