From Freshness to Effectiveness: Goal-Oriented Sampling for Remote Decision Making

Aimin Li, Shaohua Wu, Gary C.F. Lee, and Sumei Sun, *Fellow, IEEE*

***Abstract*—Data freshness, measured by Age of Information (AoI), is highly relevant in networked applications such as Vehicle to Everything (V2X), smart health systems, and Industrial Internet of Things (IIoT). However, freshness alone does not always equate to utility in decision-making. In decision-critical settings, some *stale* data may be more valuable than *fresh* updates. Motivated by this, we move beyond AoI-centric policies and investigate how data *staleness* affects remote decision-making effectiveness under random delay and limited communication resources. To this end, we propose AR-MDP, an Age-aware Remote Markov Decision Process framework, which co-designs optimal sampling and remote decision-making under a sampling frequency constraint and random delay. To efficiently solve this problem, we design a new *two-stage* hierarchical algorithm, namely Quick Bellman-Linear-Program (QUICKBLP), where the first stage involves solving the Dinkelbach root of a Bellman variant and the second stage involves solving a streamlined linear program (LP). For the tricky first stage, we propose a new One-layer Primal-Dinkelbach Synchronous Iteration (ONEPDSI) method, which overcomes the *re-convergence* and *non-expansive divergence* present in existing *per-sample* multi-layer algorithms. Through rigorous convergence analysis of our proposed algorithms, we establish that the worst-case optimality gap in ONEPDSI exhibits exponential decay with respect to iteration K at a rate of $\mathcal{O}(\frac{1}{R^K})$. Through *sensitivity analysis*, we derive a threshold for the sampling frequency, beyond which additional sampling does not yield further gains in decision-making. Simulation results validate our analyses.**

***Index Terms*—Age of Information, Value of Information, Markov Decision Process, Remote Decision Making, Goal-Oriented Communications, Effective Communications**

An earlier version of this work was presented in part by IEEE Information Theory Workshop (IEEE ITW) 2024 [1].

Aimin Li is with the Guangdong Provincial Key Laboratory of Aerospace Communication and Networking Technology, Harbin Institute of Technology (Shenzhen), Shenzhen 518055, China. This work is accomplished in part during his visit at Institute for Infocomm Research, Agency for Science, Technology and Research, 136832, Singapore (e-mail: liaimin@stu.hit.edu.cn).

Shaohua Wu is with the Guangdong Provincial Key Laboratory of Aerospace Communication and Networking Technology, Harbin Institute of Technology (Shenzhen), Shenzhen 518055, China, and also with the Peng Cheng Laboratory, Shenzhen 518055, China (e-mail: hitwush@hit.edu.cn).

Gary C.F. Lee is with the Institute for Infocomm Research, Agency for Science, Technology and Research, 138632, Singapore (e-mail: gary_lee@i2r.a-star.edu.sg).

S. Sun is with the Institute for Infocomm Research, Agency for Science, Technology and Research, 138632, Singapore (e-mail: sunsm@i2r.a-star.edu.sg).

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.

I. INTRODUCTION

Age of Information (AoI) is a crucial metric for evaluating information freshness in status update systems, garnering broad attention from both academia and industry [2], [3]. Currently, AoI has been applied in a wide range of applications such as queue control [4]–[13], source coding [14], [15], remote estimation [16]–[26], and network design [27]–[38] (see [39] for a comprehensive review). Central to this field is the question: "*How can we minimize the Age of Information?*" The conventional wisdom in AoI optimization lies in an intuitively compelling yet mathematically non-trivial heuristic: "*fresher information holds greater value*". This heuristic finds validation across real-world applications. In Internet of Vehicles (IoV) systems, *timely* status updates are essential for enabling safety-critical driving maneuvers. In financial markets, access to *first-hand* information directly impacts the effectiveness of trading decisions. These examples empirically demonstrate that minimizing AoI can improve estimation accuracy or enhance subsequent information-driven decision-making.

A significant challenge in the field lies in the lack of unified analytical frameworks that link information *freshness* to its *effectiveness* in real-time decision-making. In many scenarios, freshness alone does not determine how beneficial an update is to the downstream decision-making task. Instead, the *effectiveness* of information may depend on multiple interrelated factors beyond AoI, including the semantic content of transmitted packets and the underlying dynamics of the monitored source [40]–[43]. This recognition has driven the development of various AoI variants. One approach introduces nonlinear AoI penalties, implemented through both empirical configurations [44]–[47] and theoretically derived functions [48]–[51]. These nonlinear formulations aim to quantify the loss resulting from information *staleness*. Additionally, researchers have proposed and optimized various heuristic metrics, including Age of Synchronization (AoS) [52], Age of Incorrect Information (AoII) [53]–[55], Age of Changed Information (AoCI) [56], and Age of Collected Information [57], [58]. These metrics customize time-related penalization from a wider perspective than what can be captured with age, particularly in applications involving rapidly evolving source dynamics. For remote estimation, mean square estimation error (MSEE) [16] and context-aware Urgency of Information (UoI) [59] are leveraged to penalize the real-time reconstruction distortion. Despite these advances, the relationship between AoI and decision-making performance is not fully characterized.

Specifically, existing studies often optimize communication metrics (e.g., AoI or AoII) without explicitly modeling how delayed information influences sequential decision-making outcomes. This motivates the question: *How does delayed or outdated information affect the quality of remote decisions, and how should communication policies adapt accordingly?*

Several works have proposed heuristic approaches to characterize this relationship. In [60], Dong *et al.* introduced Age upon Decisions (AuD), which measures the time elapsed between data generation and its use in decision-making, where the decision epoch follows a stochastic distribution. In a similar vein, [61] proposed Age of Actuation (AoA). In [62]–[64], Cost of Actuation Error (CoAE) was proposed to penalize *distortion-induced* error actuation. In this setting, a penalty $C_{i,j}$ is incurred when the true system state is i while the remote controller makes decisions based on an estimated value $\hat{X}_t = j$. This line of work primarily focuses on the estimation of a discrete-time Markov chain (DTMC), and quantifies the semantic mismatch between state and inferred action due to delayed or lossy communication. In [65], three types of decisions: correct decisions, incorrect decisions, and missed decisions are assigned different time-cost functions. A new metric termed Penalty upon Decision (PuD) was proposed. In [41] and [66], a tensor-based metric termed Goal-oriented Tensor (GoT) was proposed as a unified framework for existing metrics. However, while prior work has advanced communication optimization through a variety of metrics, it often overlooks how decision-making systems actually function when operating with potentially *stale* information. Bridging this gap requires a framework that explicitly models how stochastic queuing delays influence decision quality in dynamic systems.

In this paper, we aim to examine how *stale* information impacts remote stochastic decision-making under communication constraints. To this end, we propose the *Age-aware Remote Markov Decision Process*[1] (AR-MDP), a comprehensive framework that jointly optimizes sampling and sequential decision-making, with a specific purpose to achieve goal-oriented effective decisions. The relationship between sampling and decision-making exhibits inherent *bidirectional coupling*. Sampling decisions affect the freshness of information available for the remote decision maker, which can result in unsatisfactory decision outcomes. Conversely, decision-making processes affect the stochastic evolution of the source system, which in turn impacts the effectiveness of content-driven goal-oriented sampling mechanisms. To decouple these two processes and achieve optimal decision-making under random delay and a sampling frequency constraint, we formulate the problem as a constrained partially observable semi-Markov decision process, where AoI no longer serves as a typical indicator, but as side information that informs delay-aware decision-making. We design efficient algorithms to solve this problem. *To the best of our knowledge, this is among the first attempts to treat AoI as dynamic side information in remote decision-making systems, and to systematically integrate it into a formal decision-theoretic framework.*

[1]In [67], the term *remote MDP* was first proposed as a pathway to pragmatic or goal-oriented communications. Our paper focuses on the communication delay and introduces the *age* to enhance remote decision-making to achieve a certain goal, hence the term *age-aware remote MDP*.

II. Related Work and Our Novelty

A. Sampling Under Random Delay

The results in this paper contribute to the optimal sampling design under random delay. In real-world network environments, communication channels inevitably experience random delays due to various factors: network handover, congestion, variable sample sizes, and packet retransmissions. These fundamental characteristics have driven research into developing optimal sampling policies under random delay. Existing literature has focused on optimizing three key aspects: i) **information freshness** [68]–[73]; ii) **remote estimation** [16], [17], [25], [74]–[76]; and iii) **remote inference** [77]–[80] under random delay. A particularly noteworthy and *counter-intuitive* finding in this field reveals that optimal sampling may require the source to *deliberately wait* before submitting a new sample to the channel, challenging the conventional wisdom of throughput-optimal *zero-wait* sampling policy.

Information Freshness: In the seminal work [68], Sun *et al.* derived an AoI-optimal sampling policy under random delay. This paper revealed that under a maximum rate constraint, the AoI-optimal sampling follows a threshold structure, where sampling is activated only when the current AoI exceeds a specific threshold determinable through a low-complexity bisection search method. In [69], the optimal sampling policy for a non-linear monotonic function of AoI was designed. Tang *et al.* [70] extended this framework to scenarios with unknown delay statistics, employing stochastic approximation methods to determine the AoI-optimal sampling policy under unknown delay statistics. Further advancements were made by Pan *et al.* [71], who developed AoI-optimal sampling policies under unreliable transmission with random two-way delay. Most recently, Liyanaarachchi and Ulukus extended [68] by incorporating random ACK delay, demonstrating that sampling before receiving acknowledgment can potentially achieve superior AoI performance [72]. In [73], Peng *et al.* designed optimal sampling policies that achieve minimal Age of Changed Information (AoCI) [81]—a metric capable of detecting source changes—under known and unknown delay statistics. Most recently, Chen *et al.* [55] derived AoII-optimal sampling policies under random delay.

Remote Estimation: The theoretical foundations of remote estimation-oriented sampling under random delay were established through [16], [17]. These studies developed an optimal sampling policy for the Wiener process that minimizes the mean square estimation error (MSEE) while adhering to sampling frequency constraints. Their research revealed that the optimal sampling policy exhibits a threshold structure, where sampling is initiated only when the real-time MSEE surpasses a predetermined threshold. Building upon this foundation, Ornee *et al.* [25] expanded this theoretical framework by investigating MSEE-optimal sampling for the Ornstein-Uhlenbeck (OU) process, a stationary Gauss-Markov process,

TABLE I
COMPARISONS OF TIME-LAG MDPS

Type	Observation	Reference
Standard MDP	$O(t) = X_t$	[82]
DDMDP	$O(t) = X_{t-d}$	[83]
SDMDP	$O(t) = X_{t-D}$	[84]
Age-Aware Remote MDP	$O(t) = X_{t-\Delta(t)}$	This Work

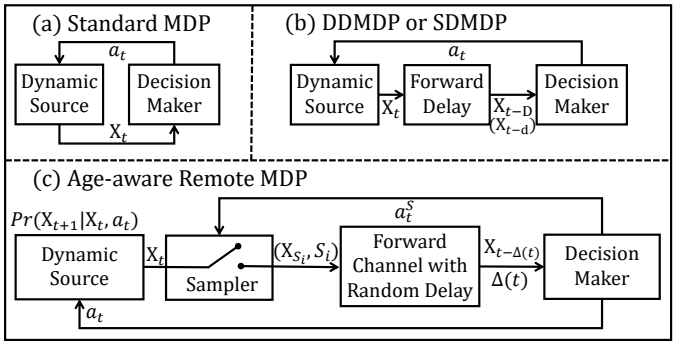


Fig. 1. Comparisons among standard MDP, DDMDP, SDMDP, and AR-MDP.

under random delay. In [74] and [75], the MSEE-optimal sampling policies are derived for the Wiener process and the OU process under unknown delay statistics. In [76], Chen *et al.* derived the optimal sampling policy that achieves minimum uncertainty of information under random delay, where UoI is defined as the conditional entropy of the source at the receiver given the observation history [51]—mathematically expressed as $H(X_t|\mathcal{I}_t)$, with X_t representing the source state at time t and $\mathcal{I}_t$ denoting the available observation history at the receiver.

Remote Inference: Recent research has revealed insights into remote inference performance and its relationship with information freshness metrics. In [77], [78], Shisher *et al.* demonstrated that the loss function in remote inference may not be monotonic in terms of the age of the samples (features) used, if the source sequence is not Markovian. Upon making this remarkable observation, the authors developed policies that allow selection of aged samples from the buffer, rather than the freshest one. This was termed the "*selection-from-buffer*" model. In [79], a learning and communication co-design problem that jointly optimizes feature length selection and transmission scheduling is proposed. In [80], Ari *et al.* expanded previous works by incorporating time-varying statistics of random delay and delayed feedback, developing optimal sampling policies to minimize long-term inference error within the "*selection-from-buffer*" model. All these works reveal that the remote inference utility may not be a monotonic function in terms of AoI. Together, these works demonstrate that in remote inference settings, the utility of a sample is not necessarily a monotonic function of its age.

To the best of our knowledge, the optimal sampling policy for ***Remote Decision Making*** *under random delay remains an open research problem, which we address in this paper.*

B. Decision-Making over Stale Status

The proposed AR-MDP in this paper also enriches the family of time-lag MDP, whose focus is on making decisions based on *stale* status. As illustrated in Table I and Fig. 1, two primary types of MDPs address observation delay at the decision maker: deterministic delayed MDP (DDMDP) [83] and stochastic delayed MDP (SDMDP) [84]. 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 [83]. 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 [84]. Thus, it becomes clear to solve an SDMDP problem by solving its equivalent standard MDP.

However, the above time-lag MDPs, where the observation delay follows a given distribution (DDMDP can be regarded as a special type of SDMDP), potentially assume that the state is sampled and transmitted to the decision maker *at every time slot*[2]. This setup presumes that the system can transmit every state update 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 [22], [68], [70], [85]–[87], where Age of Information (AoI) serves as a key performance indicator. Suppose the i-th sample is generated at time S_i and is delivered at the receiver at time D_i. Then, in a time slotted system, AoI is defined as:

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

as shown in Fig. 2. From this definition, the most recently available information at the receiver at time slot t is $O(t) = X_{t-\Delta(t)}$. In slotted time, AoI evolves deterministically between deliveries and resets to the realized delivery delay; specifically

$$\Delta(t+1) = \begin{cases} \Delta(t) + 1, & \text{if no delivery occurs at } t+1, \\ Y_i, & \text{if a delivery occurs at } t+1. \end{cases} \quad (2)$$

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 effective delay in AR-MDP is *sampling-dependent* (through S_i) and coupled with random transmission delay, so the time lag is policy-dependent rather than an exogenous fixed/i.i.d. variable[3].

[2]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 an i.i.d. random variable and is independent of the sampling policy.

[3]While the AoI evolves deterministically between successful updates (i.e., linearly increases), its reset events depend on both the sampling actions and the stochastic delay process. Thus, it can be viewed as a process indirectly governed by the sampling actions.

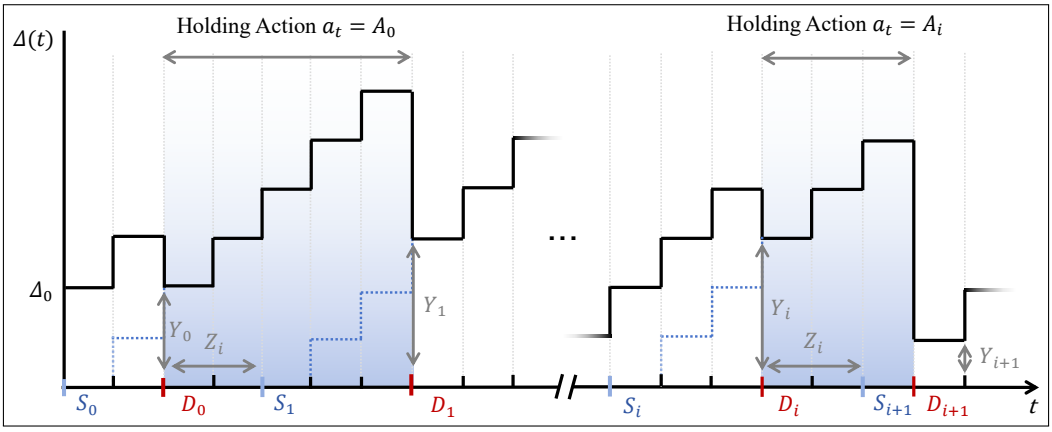


Fig. 2. AoI evolution in slotted time. The i-th sample is generated at S_i and delivered at D_i with random delay Y_i. The control is updated only upon delivery and is held constant until the next delivery. Shaded areas indicate action-holding intervals, during which the delivered observation remains X_{S_i} for $t \in [D_i, D_{i+1})$ while the staleness $\Delta(t)$ increases linearly.

To the best of our knowledge, the design of optimal remote decision-making in the presence of sampling-dependent stochastic observation delays, as captured by the AoI process, remains an unexplored research direction, which we address in this paper by studying AR-MDP.

C. The Novelty of Our Work

- **System Model**: This paper proposes AR-MDP, a novel theoretical framework integrating optimal sampling and decision-making under random delay. Differing from prior sampling designs which often treat information as an end in itself—optimizing for freshness, accuracy, or estimation quality under random delays [16], [17], [25], [68]–[80], we treat information as a **means to action**, where its contribution is defined not by how precise it is, but by how well it enables timely and effective decisions. This goes beyond distortion-based formulations such as the Cost of Actuation Error (CoAE) [62]–[64], [88], by embedding the staleness-induced impact directly into the sequential decision-making process. Different from classical time-lag MDPs (e.g., DDMDP/SDMDP) that model the delay as an *exogenous* quantity where either a fixed constant d or a random variable D independent of the sampling policy, our AR-MDP adopts a different remote-decision information structure in which AoI $\Delta(t)$ is an *endogenous* staleness process shaped jointly by controllable sampling or waiting decisions and random delay Y_i. Under the sample-and-hold information structure, we establish an exact fixed-dimensional sufficient statistic and an embedded lifted MDP (Lemma 1), enabling tractable average-cost analysis.
- **Methodology Design:** We design QUICKBLP, a computationally efficient single-layer algorithm that addresses limitations associated with iterative re-convergence common in multi-layer per-sample algorithms, *e.g.*, [76, Algorithm 1], [89, Algorithm 1], and [90, Section IV.C]. This algorithm is designed based on a key analytical insight that the optimal solution exhibits a threshold structure and can be obtained through a two-stage process. The first stage determines the *Dinkelbach root* of a Bellman variant, for which we develop ONEPDSI, a *Cauchy sequence* that converges asymptotically to the root without requiring *re-convergence*. The second stage involves finding the *Dinkelbach root* of a *per-sample* constrained Markov Decision Process (CMDP), which traditionally necessitates multiple CMDP solutions and suffers from *re-convergence*. We resolve this challenge by proving that the *Dinkelbach root* can be explicitly calculated through the optimal value of a linear program (LP), enabling direct solution to the root through a single LP solution. To the best of our knowledge, QUICKBLP is the first framework to tackle constrained partially observable SMDPs **through a streamlined single-layer flow**, fundamentally improving efficiency by eliminating re-convergence overhead.
- **Theoretical Rigor and Convergence:** We significantly advance our previous work [1] by resolving critical convergence limitations. To overcome inherent divergence risks, we design two novel algorithms in this paper: Bisec-τ-RVI and ONEPDSI. Although convergence analyses often rely on the Banach Contraction Mapping Theorem [91, Theorem 6] [71], [90], this method falls short in capturing the behavior of our models. Our approach departs from this tradition, providing rigorous proofs that both algorithms guarantee efficient *exponential convergence* in *worst-case* settings. These guarantees reinforce the reliability and efficiency of our algorithms in practical applications.

D. Notations

The main notations throughout this paper are summarized in Table II.

TABLE II
SUMMARY OF NOTATIONS

Symbol	Description
X_t	System state at time slot t
a_t	Response action taken by the decision maker at time t
u_t	Sampling action at time slot t
π_t	History-dependent policy
ϕ_t	State-dependent policy
$\boldsymbol{\psi}$	Policy composed by $\phi_{0:\infty}$
$\boldsymbol{\psi}^\lambda$	Policy in Problem 4 induced by parameter λ
$\mathcal{H}_t(\cdot)$	Sufficient statistics function
$\Delta(t)$	AoI at time t
S_i	Time slot when the i-th sample is taken
D_i	Time slot when the i-th sample is delivered
Y_i	Random delivery delay between S_i and D_i
Z_i	Waiting time to sample the i-th sample
A_i	Holding-action taken at the i-th epoch
G_i	Lifted State of Lifted MDP at epoch i
$C(x,a)$	Cost function given state x and action a
λ	Dinkelbach parameter
$\mathscr{P}_{\text{MDP}}(\lambda)$	Transformed MDP tuple give parameter λ
ϕ_λ^*	Optimal policy of MDP $\mathscr{P}_{\text{MDP}}(\lambda)$
$U(\lambda)$	Optimal value of Problem 4
$V^\star(\cdot;\lambda)$	Optimal relative value function given λ
$U_K(\lambda), V_K(\gamma;\lambda)$	RVI values in (25)
$\tilde{U}_K, \tilde{V}_K(\gamma;\lambda)$	τ-RVI values in (28)
$e_{\text{U}}^{(K)}(\lambda)$,$e_{\text{V}}^{(K)}(\cdot;\tau,\lambda)$	Relative errors for the K-iteration in τ-RVI given λ
κ	A parameter for ONEPDSI
$e_\rho^{(K)}$,$e_{\text{W}}^{(K)}(\cdot;\kappa)$	Relative errors for the K-iteration in ONEPDSI given κ
$W^\star(\cdot)$	Variables in fixed-point equations (38)
ρ_K,$\widetilde{W}_K(\cdot)$	ONEPDSI values in (43)
$f_{\max}$	Maximum average sampling rate
$H(\lambda; f_{\max})$	Optimal value of Problem 5
θ	Lagrangian multiplier
$\mathcal{L}(\boldsymbol{\psi};\theta,\lambda,f_{\max})$	Lagrangian function
$\Upsilon(\theta,\lambda;f_{\max})$	Lagrangian dual function
$d(\lambda;f_{\max})$	Optimal value of Problem 6
$Q^\star(f_{\max})$	Optimal value of Problem 7
$\theta_\lambda^\star$	Optimal variable θ for fixed λ in Problem 6
$\mathcal{Q}^\lambda$	Long-term average cost given policy ϕ_λ^*
$\mathcal{F}^\lambda$	Long-term average sampling rate given policy ϕ_λ^*
$\mathcal{F}^{\lambda^+}$	Right limit of $\mathcal{F}^\lambda$
$\mathcal{F}^{\lambda^-}$	Left limit of $\mathcal{F}^\lambda$
$f_{\max}^T$	Sapling frequency threshold
$\mathbf{P}_a$	Transition probability matrix given action a
$\mathbf{P}^n$	n-step transition probability matrix given action a
$\mathbf{P}_{i\times j}$	the (i,j)-th entry of the transition matrix $\mathbf{P}$
$\rho^\star$	Optimal value of Problem 3; Root of $U(\lambda)$
$h^\star$	Optimal value of Problem 1; Root of $H(\lambda;f_{\max})$
$\pi_a(\cdot)$	Stationary distribution over states under $\mathbf{P}_a$

III. SYSTEM MODEL AND PROBLEM FORMULATION

We consider a time-slotted[4] *age-aware remote MDP* problem illustrated in Fig. 1(c). Let $X_t \in \mathcal{S}$ be the controlled source of interest at time slot t. The evolution of the source is a Markov decision process, characterized by the transition probability $\Pr(X_{t+1}|X_t, a_t)$[5], where $a_t \in \mathcal{A}$ represents the controlled action taken by the remote decision maker to control the source in the desired way. We assume that both the state space $\mathcal{S}$ and the action space $\mathcal{A}$ are finite. This finite setting is a common starting point in infinite-horizon average-cost MDP literature [92], [93] and allows the proposed AR-MDP to inherit well-established *optimality* and *convergence results*. We view AR-MDP as a stepping stone toward more general models, and refer readers to [93, Chap. 4.6] for approaches that can be used to extend our framework to infinite spaces.

The sampler conducts the sampling action $u_t \in \{0,1\}$, with $u_t = 1$ representing the sampling action and $u_t = 0$ otherwise. Let S_i be the sampling time of the i-th delivered packet, and D_i be the corresponding delivery slot. The random channel delay of the i-th packet is denoted 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 $S_0, S_1, \cdots$ record the time stamp when $u_t = 1$, given by

$$S_i = \max\{t \in \mathbb{N} \,|\, t \le D_i, u_t = 1\}, \quad \forall i \in \mathbb{N}, \tag{3}$$

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

Since the system is time-slotted, the AoI evolves in discrete steps. Specifically, for any slot t, let $i(t) \triangleq \max\{i : D_i \le t\}$ be the index of the most recently delivered update. Then $\Delta(t) = t - S_{i(t)}$. Hence, upon a delivery at $t = D_i$, the AoI drops to

$$\Delta(D_i) = D_i - S_i = Y_i, \tag{4}$$

which is *not necessarily zero* unless the channel delay is zero. Moreover, during the interval $t \in [D_i, D_{i+1})$, no new observation is delivered and the age $\Delta(t)$ drifts with $\Delta(t+1) = \Delta(t) + 1$. The discrete-time AoI dynamics are shown in Fig. 2.

At the sampling time $S_i, \forall i \in \mathbb{N}$, the state X_{S_i} along with the corresponding time stamp S_i is encapsulated into a packet (X_{S_i}, S_i), which is transmitted to a remote decision maker. Upon receipt of the packet (X_{S_i}, S_i) at delivery time slot D_i, the *observation history* at the decision maker is $\{(X_{S_j}, S_j, D_j) : j \le i\}, t \in [D_i, D_{i+1})$. By employing (1), for any time slot $t \in \mathbb{N}$, the freshest available sample at the decision maker is $(X_{t-\Delta(t)}, t - \Delta(t))$. As t is known to the decision maker, the *observation history* up to time slot t is equivalently expressed by AoI, given as: $\{(X_{k-\Delta(k)}, \Delta(k)) : k \le t, k \in \mathbb{N}\}$.

A. Protocol and Assumptions

Similar to [68], we impose the following assumptions in sampling:

(S1) A new sample cannot be generated until the previous sample has been delivered. Specifically,

$$S_{i+1} = D_i + Z_i, \quad Z_i \ge 0, \quad i \in \mathbb{N}, \tag{5}$$

where Z_i is the sampling waiting time after the delivery at D_i. Consequently, the delivery time satisfies $D_i = S_i + Y_i$ for all $i \in \mathbb{N}$, where Y_i is the random delay of the i-th sample. Moreover, under (S1) and the delivery timeline, the observation is piecewise constant between two consecutive deliveries. Specifically, for any $t \in [D_i, D_{i+1})$ we have $t - \Delta(t) = S_i$, and hence $O(t) = X_{t-\Delta(t)} = X_{S_i}, D_i \le t < D_{i+1}$, while the staleness $\Delta(t)$ increases deterministically within the interval.

[4]As the proposed AR-MDP is formulated as an extension of the discrete-time MDP framework, a time-slotted system model is employed to maintain structural consistency with MDP. Therefore, all key variables including transmission delay Y_i, AoI $\Delta(k)$, and sampling time S_i are accordingly defined and evolve over discrete time slots.

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

(S2) The inter-sample times $G_i = S_{i+1} - S_i$ form a regenerative process [94, Section 6.1]. Hence, almost surely[6],

$$\lim_{i\to\infty} S_i = \infty, \qquad \lim_{i\to\infty} D_i = \infty. \tag{6}$$

In addition, we adopt a *holding-action* paradigm consistent with the timeline in Fig. 2:

(A1) The controlled action is updated only upon the delivery of a sample. That is, upon the delivery time D_i, the decision maker selects an updated action $A_i \in \mathcal{A}$ based on the available history, and then holds it constant until the next delivery:

$$a_t = A_i, \quad D_i \le t < D_{i+1}, \quad i \in \mathbb{N}. \tag{7}$$

As a result, the underlying state continues to evolve according to the controlled Markov kernel, i.e., $X_{t+1} \sim P(\cdot \mid X_t, A_i)$ for all $t \in [D_i, D_{i+1})$.

Remark 1. *The modeling assumption (A1) is widely used in networked and remote control systems where the controller can revise its command only when a new measurement is received; see, e.g., [95]–[97]. Typical real-world systems that follow the holding-action paradigm include:*

- ***Robotic manipulation and teleoperation:*** *In robotic teleoperation under constant communication delays, the controller continuously updates the motion command based on the current local state and the most recently received delayed remote state, effectively maintaining past remote information during delay periods [95].*
- ***Multi-UAV or vehicular coordination:*** *each autonomous agent maintains its last chosen coordination strategy (e.g., formation control gain or following distance target) until updated state information is received from neighboring agents [96].*
- ***Industrial supervisory control:*** *in process plants or smart grids, a supervisory controller holds the previously assigned operation mode (e.g., heater on/off state, pump flow rate setpoint) during communication gaps between control center and local devices [97].*

Remark 2 (Beyond holding actions)**.** *The holding-action rule (7) induces an epoch-based decision structure: between two consecutive deliveries, no new observation is received and the applied action is fixed, and the state continues to evolve according to the controlled Markov kernel under the held action. This structure is essential for compressing the growing history into a finite-dimensional sufficient statistic (Lemma 1) and for constructing a finite-state lifted MDP amenable to our average-cost optimality analysis and algorithms*[7].

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

[7]If (A1) is relaxed so that the decision maker can adapt a_t within $t \in [D_i, D_{i+1})$ based on the deterministic AoI drift, then the sufficient statistic generally becomes belief-based: the conditional distribution of X_t depends on the staleness and the within-epoch action sequence. This leads to a (belief-)MDP/POMDP formulation with either a continuous state space (belief simplex) or a significantly enlarged epoch action space (age-indexed action plans). We leave this non-holding extension as future work.

B. Joint Sampling and Decision-Making Policy

Let $\mathcal{I}_t = \{(X_{k-\Delta(k)}, \Delta(k), u_{k-1}, a_{k-1}) : k \le t\}$ denote the history available to the decision maker up to time t. Under the protocol in (5) and (7), the slot-level actions (u_t, a_t) are induced by the epoch-level decisions $\{(A_i, Z_i)\}$ made at delivery epochs. A possibly randomized decision policy is a sequence of mappings from the history to a distribution over the joint action space $\{0,1\} \times \mathcal{A}$:

$$\pi_t : \mathcal{I}_t \to \mathcal{P}(\{0,1\} \times \mathcal{A}), \tag{8}$$

where $\mathcal{P}(\cdot)$ is a *simplex* space which represents the probability that an action is taken.

An epoch-based policy is defined as

$$\phi : \mathcal{S} \times \mathcal{Y} \times \mathcal{A} \to \mathcal{P}(\mathcal{A} \times \mathcal{Z}), \tag{9}$$

which induces the slot-level policy π_t via $S_{i+1} = D_i + Z_i$ and $a_t = A_i$ on $[D_i, D_{i+1})$. For completeness, a slot-level policy can be written as $\pi_t : \mathcal{I}_t \to \mathcal{P}(\{0,1\} \times \mathcal{A})$, which in our setting is induced by ϕ. (See Lemma 1).

We consider a bounded cost function $\mathcal{C}(X_t, a_t) < \infty$, 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 joint sampling and decision policies at each time slot, *i.e.*, $\pi_0, \pi_1, \pi_2 \cdots$, to minimize the *long-term average cost*, subject to a *long-term average sampling frequency constraint*:

Problem 1 (*Joint Design of Sampling and Decision Processes under Sampling Frequency Constraint*)**.**

$$\inf_{\pi_{0:\infty}} \limsup_{T\to\infty} \frac{1}{T} \mathbb{E}_{\pi_{0:\infty}} \left[\sum_{t=1}^{T} \mathcal{C}(X_t, a_t) \right] \tag{10a}$$

$$\text{s.t.} \ \liminf_{N\to\infty} \frac{1}{N} \mathbb{E}_{\pi_{0:\infty}} \left[\sum_{i=0}^{N-1} (S_{i+1} - S_i) \right] \ge \frac{1}{f_{max}}, \tag{10b}$$

where π_t is the joint sampling and decision-making policy defined by (8), f_{max} represents the maximum allowed sampling frequency, and the expectation $\mathbb{E}_{\pi_{0:\infty}}$ is taken over the stochastic processes $(X_1, X_2, \cdots)$ and $(Y_0, Y_1, \cdots)$ under given policies $\pi_{0:\infty}$.

In practice, the overhead for information updates will increase with the average sampling frequency. Hence, Problem 1 represents a tradeoff between remote decision-making utility and communication overhead. Since *age* $\Delta(k)$ is available at the decision maker as *side information* to facilitate more informed decision-making, we call this problem an age-aware remote MDP problem. This problem aims at determining the distribution of joint sampling and controlled actions (u_t, a_t) based on the history $\mathcal{I}_t$, such that the long-term average cost subject to the sampling frequency constraint is minimized.

We remark that we will study Problem 1 both without (see Section IV and V) and with (see Section VI and VII) the sampling frequency constraint. To distinguish these two problems, we use $h^\star$ to denote the optimal value of the problem with the rate constraint and $\rho^\star$ to denote the optimal value of the problem without the sampling frequency constraint.

C. Sufficient Statistics of History

In principle, the joint policy π_t maps the growing history $\mathcal{I}_t$ to a distribution over (u_t, a_t), which leads to the classical *curse of history* and makes direct dynamic programming intractable.

More importantly, under delayed observations, optimal decisions generally depend not only on the latest delivered sample value but also on its *staleness*. In our time-slotted model, the staleness at delivery epochs is $\Delta(D_i) = Y_i$. Within the interval $t \in [D_i, D_{i+1})$, the freshest delivered sample remains X_{S_i} and the AoI drifts deterministically as $\Delta(t) = Y_i + (t - D_i)$, i.e., the time since sampling equals $t - S_i = \Delta(t)$. This determines how many Markov transitions have occurred since the sampled state. Consequently, even conditioned on the same delivered sample X_{S_i} and held actions, different Y_i (and hence different $\Delta(t)$ in the epoch $t \in [D_i, D_{i+1})$) induce different conditional distributions (beliefs) of the current state.

Crucially, under the holding-action rule (7), no new observation is delivered between two consecutive deliveries and the action is kept constant. This induces an epoch-based decision structure, under which the growing history can be compressed into a finite-dimensional sufficient statistic. The resulting statistic, given in Lemma 1, serves as the state of the lifted MDP and enables our subsequent analysis and algorithms.

A sufficient statistic is defined as follows.

Definition 1. *A sufficient statistic of $\mathcal{I}_t$ is a function $\mathcal{H}_t(\mathcal{I}_t)$, such that*

$$\min_{a_{t:T}} \mathbb{E}\left[\sum_{k=t}^{T} \mathcal{C}(X_k, a_k) | \mathcal{I}_t\right] = \min_{a_{t:T}} \mathbb{E}\left[\sum_{k=t}^{T} \mathcal{C}(X_k, a_k) | \mathcal{H}_t(\mathcal{I}_t)\right] \tag{11}$$

holds for any $T > t$.

This definition suggests that decision-making that leverages the *sufficient statistics* $\mathcal{H}_t(\mathcal{I}_t)$ can achieve the same performance as using the complete history $\mathcal{I}_t$. Thus, the compression $\mathcal{H}_t(\mathcal{I}_t)$ is *sufficient* for the agent to implement an optimal action, enabling the design of efficient policies that maintain optimality while overcoming the *curse of history*. In particular, the time stamp carried by each packet makes the staleness observable at delivery: upon receiving (X_{S_i}, S_i) at time D_i, the decision maker can compute $\Delta(D_i) = D_i - S_i = Y_i$. This staleness information enters the sufficient statistic and therefore the optimal delivery-epoch policy. The following Lemma establishes efficient *sufficient statistics* of the *history* $\mathcal{I}_t$ in a piecewise manner.

Lemma 1. *(Sufficient Statistics). 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 statistic of $\mathcal{I}_t$. In addition, determining the optimal sampling actions u_t under condition (5) is equivalent to determining the optimal sampling time S_{i+1}, or the optimal waiting time Z_i.*

Proof. See Appendix A. ∎

Lemma 1 indicates that solving Problem 1 over the history-dependent slot-level policies

$$\pi_t : \mathcal{I}_t \to \mathcal{P}(\{0, 1\} \times \mathcal{A}) \tag{12}$$

is equivalent to solving over delivery-epoch policies that depend only on $\mathcal{G}_i$:

$$\phi_i : \mathcal{S} \times \mathcal{Y} \times \mathcal{A} \to \mathcal{P}(\mathcal{A} \times \mathcal{Z}), \quad \forall i \in \mathbb{N}, \tag{13}$$

which maps the sufficient statistic $\mathcal{G}_i = (X_{S_i}, Y_i, A_{i-1})$ to a distribution over the joint epoch actions $(A_i, Z_i) \in \mathcal{A} \times \mathcal{Z}$. Consequently, the AoI at delivery, equivalently the realized delay $Y_i = \Delta(D_i)$, is an explicit argument of the optimal policy via $\mathcal{G}_i$. This provides a concrete AoI-related conclusion: even for the same delivered sample value X_{S_i}, different staleness values Y_i can induce different beliefs and thus the optimal epoch sampling and decision-making may depend on Y_i. Because the Cartesian product $\mathcal{S} \times \mathcal{Y} \times \mathcal{A}$ does not grow with time, this reformulation avoids the exponential growth of the original information set $\mathcal{I}_t$. From assumptions (S1) and (S2), we reformulate Problem 1 as:

Problem 2 (*From History-Dependent to State-Dependent Policy*)**.**

$$\inf_{\boldsymbol{\psi}} \limsup_{N\to\infty} \frac{\mathbb{E}_{\boldsymbol{\psi}}\left[\sum_{i=0}^{N-1} \sum_{t=D_i}^{D_{i+1}-1} \mathcal{C}(X_t, a_t)\right]}{\mathbb{E}_{\boldsymbol{\psi}}[D_N]} \tag{14a}$$

$$s.t. \ \liminf_{N\to\infty} \frac{1}{N} \mathbb{E}_{\boldsymbol{\psi}}\left[\sum_{i=1}^{N} (S_{i+1} - S_i)\right] \geq \frac{1}{f_{max}}, \tag{14b}$$

Here, we use $\boldsymbol{\psi}$ to denote the policy $\{\phi_i\}_{i=0}^{\infty}$.

IV. Optimal Sampling Without Rate Constraint: A Two-Layer Persample Solution

In this section, we address the unconstrained problem to determine $\rho^\star$. A series of theoretical results are developed through a divide-and-conquer approach. In subsection IV-A, we rewrite the unconstrained problem into a *non-linear fractional program*. By utilizing a *Dinkelbach-like* method [99], we transform the *non-linear fractional programming* into an infinite horizon MDP problem given a *Dinkelbach parameter*. The problem then becomes finding the *Dinkelbach parameter* such that the optimal value of the standard MDP is zero, which is equivalently a root-finding problem. To search for the root, in subsection IV-B we review a typical *two-layer nested* algorithm, namely *Bisec-RVI* [1, Algorithm 2][8]. We note that the inner-layer *Relative Value Iteration* (RVI) algorithm of *Bisec-RVI* may suffer from divergence and thus propose an improved *Bisec-τ-RVI* algorithm to achieve provable convergence.

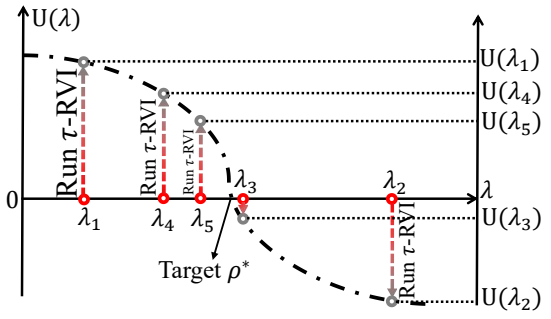
A. A Reformulation of Problem 2

From the action-holding assumption (A1), we can rewrite the objective function in Problem 2 as:

$$\limsup_{N\to\infty} \frac{\mathbb{E}_{\boldsymbol{\psi}}\left[\sum_{i=0}^{N-1} \sum_{t=D_i}^{D_{i+1}-1} \mathcal{C}(X_t, a_t)\right]}{\mathbb{E}_{\boldsymbol{\psi}}[D_N]} \tag{15a}$$

$$= \lim_{n\to\infty} \frac{\sum_{i=0}^{N-1} \mathbb{E}_{\boldsymbol{\psi}}\left[\sum_{t=D_i}^{D_{i+1}-1} \mathcal{C}(X_t, A_i)\right]}{\sum_{i=0}^{N-1} \mathbb{E}_{\boldsymbol{\psi}}\left[D_{i+1} - D_i\right]} \tag{15b}$$

[8]The idea of the two-layer nested algorithm has been applied in [20], [68], [100] and [44] to achieve Age-optimal and Mean Square Error (MSE)-optimal sampling and scheduling.



Step	$(\lambda_\downarrow, \lambda_\uparrow)$	Midpoint λ_{M}	$\mathrm{U}(\lambda_\downarrow)$	$\mathrm{U}(\lambda_\uparrow)$	$\mathrm{U}(\lambda_{\mathrm{M}})$
1	(λ_1,λ_2)	λ_3	>0	<0	<0
2	(λ_1,λ_3)	λ_4	>0	<0	>0
3	(λ_4,λ_3)	λ_5	>0	<0	>0

Fig. 3. Bisection search to find the root of the implicit function $U(\lambda)$. The implicit function $U(\lambda)$ is approximated using a value iteration approach. The interval containing the root, denoted by $(\lambda_\downarrow, \lambda_\uparrow)$, is halved at each outer-layer iteration, and this process eventually converges to the unique root $\rho^\star$.

$$= \lim_{\mathrm{n}\to\infty} \frac{\sum_{i=0}^{n-1} \mathbb{E}_{\boldsymbol{\psi}} \left[\sum_{t=D_i}^{D_{i+1}-1} \mathcal{C}(X_t, A_i)\right]}{\sum_{i=0}^{n-1} \mathbb{E}_{\boldsymbol{\psi}} \left[Y_{i+1} + Z_i\right]}. \tag{15c}$$

We can then express the *unconstrained* version of Problem 2 as the epoch-by-epoch variant:

Problem 3 (*Epoch-by-Epoch Reformulation*)**.**

$$\rho^\star \triangleq \inf_{\boldsymbol{\psi}} \lim_{\mathrm{N}\to\infty} \frac{\sum_{i=0}^{N-1} \mathbb{E}_{\boldsymbol{\psi}} \left[\sum_{t=D_i}^{D_{i+1}-1} \mathcal{C}(X_t, A_i)\right]}{\sum_{i=0}^{N-1} \mathbb{E}_{\boldsymbol{\psi}} \left[Y_{i+1} + Z_i\right]}. \tag{16}$$

Problem 3 is a *non-linear fractional program*, which is challenging due to its fractional nature. To simplify this problem, we consider the following sequential decision process with Dinkelbach parameter $\lambda \geq 0$:

Problem 4 (*Standard Infinite-Horizon Sequential Decision Process with Dinkelbach Parameter* λ)**.**

$$U(\lambda) \triangleq \inf_{\boldsymbol{\psi}^\lambda} \lim_{\mathrm{n}\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \left\{ \mathbb{E}\left[\sum_{t=D_i}^{D_{i+1}-1} \mathcal{C}(X_t, A_i)\right] - \lambda \mathbb{E}\left[Z_i + Y_{i+1}\right] \right\}. \tag{17}$$

By similarly applying the Dinkelbach-like method for *non-linear fractional programming* [99], we can obtain the following lemma:

Lemma 2. *The following assertions hold:*
(i). $\rho^\star \gtreqless \lambda$ if and only if $U(\lambda) \gtreqless 0$.
(ii). When $U(\lambda) = 0$, the policy solutions to Problem 4 are equivalent to those of Problem 3.
(iii). $U(\lambda) = 0$ has a unique root, with $U(\rho^\star) = 0$.

Proof. See Appendix B. ■

This key lemma enables the formulation of the following two-layer nested approach to determine $\rho^\star$.

Algorithm 1: *Two-layer approaches for* $\rho^\star$

Input: Tolerence $\epsilon > 0$, MDP $\mathscr{P}_{\mathrm{MDP}}(\lambda)$
Initialization: $\lambda_\uparrow = \min_a \sum_{s\in\mathcal{S}} \pi_a(s) \cdot \mathcal{C}(s,a)$, $\lambda_\downarrow = \min_{s,a} \mathcal{C}(s,a)$;
while $\lambda_\uparrow - \lambda_\downarrow \geq \epsilon$ **do**
  $\lambda = (\lambda_\uparrow + \lambda_\downarrow)/2$;
  Run RVI (Iteration 1) or τ-RVI (Iteration 2) to solve $\mathscr{P}_{\mathrm{MDP}}(\lambda)$ and calculate $U(\lambda)$;
  if $U(\lambda) > 0$ **then**
    $\lambda_\downarrow = \lambda$;
  else
    $\lambda_\uparrow = \lambda$;
Output: $\rho^\star = \lambda$

B. Two-layer Approaches: Bisec-RVI and Bisec-τ-RVI

Following Lemma 2, solving Problem 3 reduces to solving Problem 4 to determine the optimal value $U(\lambda)$ while simultaneously finding the unique root $\rho^\star$ of the implicit function $U(\lambda)$, which can be solved using a two-layer nested algorithm as shown in Fig. 3 and Algorithm 1. In the inner layer, value iteration approaches are applied to approximate the optimal value of Problem 4, $U(\lambda)$, by resolving the reformulated MDP $\mathscr{P}_{\mathrm{MDP}}(\lambda)$ detailed in subsection IV-B1. In the outer layer, a bisection search algorithm approximates the unique root $\rho^\star$. Conventionally, the RVI algorithm is applied in the inner layer to iteratively approximate $U(\lambda)$ [1], [20], but it proves to be divergent in our formulation (see Fig. 2). To address this limitation, we propose τ-RVI in this paper and present a rigorous convergence analysis. The condition and the rationale of the convergence will be discussed in this section.

1) Inner-Layer MDP Given Dinkelbach Parameter λ

To solve the inner layer sequential decision process in Problem 4, we reformulate it as an equivalent standard infinite-horizon MDP. A standard MDP is typically described by a *quadruple*: the state space, the action space, the transition probability, and the cost function. This subsection details this *quadruple*. The MDP with Dinkelbach parameter λ is denoted as $\mathscr{P}_{\mathrm{MDP}}(\lambda)$:

- **State Space**: the state of the equivalent MDP is the sufficient statistics $\mathcal{G}_i = (X_{S_i}, Y_i, A_{i-1}) \in \mathcal{S} \times \mathcal{Y} \times \mathcal{A}$, as established in Lemma 1.
- **Action Space**: the action space of the MDP is composed of 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 action that controls the source.
- **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 established in (18), whose detailed proof is given in Appendix C:

$$\begin{aligned} &\Pr(\mathcal{G}_{i+1} = (s', \delta', a') | \mathcal{G}_i = (s, \delta, a), Z_i, A_i) \\ &= \Pr(Y_{i+1} = \delta') \cdot [\mathbf{P}_a^\delta \cdot \mathbf{P}_{A_i}^{Z_i}]_{s\times s'} \cdot \mathbb{1}\{a' = A_i\}, \end{aligned} \tag{18}$$

where $[\mathbf{P}]_{s\times s'}$ denotes the element located at the s-th row and s'-th column of the matrix $\mathbf{P}$.

- **Cost Function**: the cost function is typically a real-valued function over the state space and the action space. We denote it as $g(\mathcal{G}_i, Z_i, A_i)$ and will show how to tailor the cost function to establish equivalence with Problem 4.

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

$$g(\mathcal{G}_i, Z_i, A_i; \lambda) \triangleq q(\mathcal{G}_i, Z_i, A_i) - \lambda f(Z_i), \tag{19}$$

where

$$q(\mathcal{G}_i, Z_i, A_i) \triangleq \mathbb{E}\left[\sum_{s' \in \mathcal{S}} \left[\sum_{t=0}^{Z_i + \mathrm{Y}_{i+1} - 1} \mathbf{P}_{A_{i-1}}^{Y_i} \cdot \mathbf{P}_{A_i}^{t}\right]_{X_{S_i} \times s'} \cdot \mathcal{C}(s', A_i)\right], \tag{20}$$

with the expectation $\mathbb{E}$ taken over the random delay Y_{i+1}, and

$$f(Z_i) \triangleq \mathbb{E}\left[Y_{i+1}\right] + Z_i. \tag{21}$$

Then Problem $\mathscr{P}_{\mathrm{MDP}}(\lambda)$ is equivalent to Problem 4.

Proof. See Appendix D. ∎

In what follows, we refer to the MDP with transition dynamics $\mathbf{P}_a$ as the *primal MDP*, and denote by $\mathscr{P}_{\mathrm{MDP}}(\lambda)$ the transformed (or *lifted*) MDP. The next theorem provides a simple sufficient condition under which the lifted MDP $\mathscr{P}_{\mathrm{MDP}}(\lambda)$ is *unichain*[9], thereby ensuring the existence of an optimal stationary deterministic policy.

Theorem 1 (Sufficient condition for a unichain lifted MDP)**.** *Let $\mathcal{S}$, $\mathcal{A}$, and $\mathcal{Y}$ be the finite state, action, and delay sets of the primal MDP. Suppose there exist a state $s^\star \in \mathcal{S}$, an integer $m \geq 1$, and a constant $\epsilon > 0$ such that, for every initial $(s, a) \in \mathcal{S} \times \mathcal{A}$ and for every admissible length-m sequence $\{A_t, Z_t, \delta_t\}_{t=0}^{m-1}$:*

$$\left[\mathbf{P}_a^{\delta_0} \mathbf{P}_{A_0}^{Z_0} \cdots \mathbf{P}_{A_{m-2}}^{\delta_{m-1}} \mathbf{P}_{A_{m-1}}^{Z_{m-1}}\right]_{s \times s^\star} \geq \epsilon. \tag{22}$$

Then, for every *stationary deterministic policy $\pi : \mathcal{S} \times \mathcal{Y} \times \mathcal{A} \to \mathcal{A} \times \mathcal{Z}$, the Markov chain on $\mathcal{S} \times \mathcal{Y} \times \mathcal{A}$ induced by π has a single recurrent class (all other states are at most transient). In particular, the lifted MDP $\mathscr{P}_{\mathrm{MDP}}(\lambda)$ is* unichain.

Proof. See Appendix E. ∎

If the lifted MDP $\mathscr{P}_{\mathrm{MDP}}(\lambda)$ is a *unichain*, an optimal *stationary deterministic* policy exists, and one can establish the following *Average Cost Optimality Equations* (ACOE) [101, Eq. 4.1]:

$$\textbf{ACOE}: V^\star(\gamma; \lambda) + U(\lambda) = \min_{A_i, Z_i} \Big\{ g(\gamma, Z_i, A_i; \lambda) + \mathbb{E}[V^\star(\gamma'; \lambda) \,|\gamma, Z_i, A_i] \Big\}, \forall \gamma \in \mathcal{S} \times \mathcal{Y} \times \mathcal{A}, \tag{23}$$

where $V^\star(\gamma; \lambda) \in \mathbb{R}$ is the optimal value function and $U(\lambda) \in \mathbb{R}$ is the optimal long-term average value of $\mathscr{P}_{\mathrm{MDP}}(\lambda)$ in (17). By solving $U(\lambda)$ and $V^\star(\gamma; \lambda)$ for $\gamma \in \mathcal{S} \times \mathcal{Y} \times \mathcal{A}$ from the ACOE (23), we can obtain the optimal *stationary deterministic* sampling and remote decision-making policy for $\mathscr{P}_{\mathrm{MDP}}(\lambda)$, defined as: $\phi_\lambda^* : \mathcal{S} \times \mathcal{Y} \times \mathcal{A} \to \mathcal{Z} \times \mathcal{A}$.

$$\phi_\lambda^*(\gamma) = \arg\min_{Z_i, A_i} \Big\{ g(\gamma, Z_i, A_i; \lambda) + \mathbb{E}[V^\star(\gamma'; \lambda) \,|\gamma, Z_i, A_i] \Big\}, \forall \gamma \in \mathcal{S} \times \mathcal{Y} \times \mathcal{A}. \tag{24}$$

The ACOE in (23) represents a series of nonlinear equations, which is mathematically intractable to solve explicitly. One can resort to the typical Dynamic Programming (DP)-like RVI algorithm [102] to iteratively generate the sequences $\{U_K(\lambda)\}^{K \in \mathbb{N}^+}$ and $\{V_K(\gamma; \lambda)\}_{\gamma \in \mathcal{S} \times \mathcal{Y} \times \mathcal{A}}^{K \in \mathbb{N}^+}$ that conditionally converge asymptotically to the solutions of the ACOE.

Iteration 1. *(RVI Algorithm [102]). For a given λ, the RVI starts with a given initial value $\{V_0(\gamma; \lambda)\}_{\gamma \in \mathcal{S} \times \mathcal{Y} \times \mathcal{A}}$ and computes $U_{K+1}(\lambda)$ and $V_{K+1}(\gamma; \lambda)$ iteratively:*

$$U_{K+1}(\lambda) = \min_{A_i, Z_i} \Big\{ g(\gamma^r, Z_i, A_i; \lambda) + \mathbb{E}\big[V_K(\gamma'; \lambda) \mid \gamma^r, Z_i, A_i\big] \Big\}, \tag{25a}$$

$$V_{K+1}(\gamma; \lambda) = \min_{A_i, Z_i} \Big\{ g(\gamma, Z_i, A_i; \lambda) + \mathbb{E}\big[V_K(\gamma'; \lambda) \mid \gamma, Z_i, A_i\big] \Big\} - U_{K+1}(\lambda), \quad \forall \gamma \in \mathcal{S} \times \mathcal{Y} \times \mathcal{A}. \tag{25b}$$

where $\gamma^r \in \mathcal{S} \times \mathcal{Y} \times \mathcal{A}$ is an arbitrarily chosen reference state and the conditional expectation $\mathbb{E}$ is taken with respect to the conditional distribution of the next state γ' given the current state and the current action. The iterative process continues until a predefined convergence criterion is satisfied.

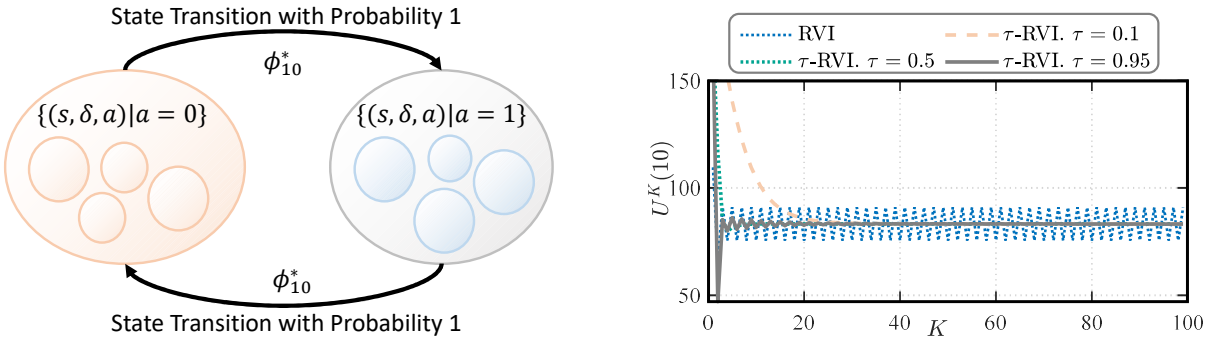
However, in our specific context, the convergence of the RVI algorithm is not necessarily guaranteed (see Example 1). To further investigate this issue and develop a convergent alternative, we present sufficient conditions for its convergence in the following Lemma.

Lemma 4. *(Restatement of [103, Proposition 4.3.2]). If an MDP satisfies the following conditions:*
(a) the MDP is a unichain MDP;
(b) the optimal stationary policy for the MDP yields an aperiodic transition probability matrix;
then the sequences $\{U_K(\lambda)\}_{K \in \mathbb{N}}$ and $\{V_K(\gamma; \lambda)\}_{K \in \mathbb{N}}$ will converge to the solution to the ACOE (23):

$$\lim_{K \to \infty} U_K(\lambda) = U(\lambda)$$
$$\lim_{K \to \infty} V_K(\gamma; \lambda) = V^\star(\gamma; \lambda), \forall \gamma \in \mathcal{S} \times \mathcal{Y} \times \mathcal{A}. \tag{26}$$

In the transformed MDP $\mathscr{P}_{\mathrm{MDP}}(\lambda)$, condition (a) in Lemma 4 is ensured by Theorem 1. Nevertheless, the following **counter-example** demonstrates that condition (b) does not necessarily hold in $\mathscr{P}_{\mathrm{MDP}}(\lambda)$. Specifically, even if the *primal MDP* characterized by $\langle \mathcal{S}, \mathcal{A}, \mathbf{P}_a, \mathcal{C} \rangle$ is *aperiodic*, the transformed MDP $\mathscr{P}_{\mathrm{MDP}}(\lambda)$ may still exhibit *periodicity*, which will cause the RVI algorithm to diverge.

[9] An MDP is said to be *unichain* if, under any stationary policy, the induced Markov chain has a single recurrent class (with all other states being transient).



(a) Ping-pong manner under policy ϕ_{10}^* **(b)** $U^K(10)$ vs. Iteration K

Fig. 4. Algorithmic behavior of RVI versus τ-RVI: Divergence mechanisms and comparative performance.

Example 1. *(A divergence example of RVI). Consider the parameter setup described in Appendix H where the delay is constant with $p=0$. In this case, the optimal policy for the transformed MDP $\mathscr{P}_{MDP}(10)$ is:*

$$\phi_{10}^*(\overbrace{0}^{X_{S_i}},\overbrace{10}^{Y_i},\overbrace{0}^{A_{i-1}})=(\overbrace{0}^{Z_i},\overbrace{1}^{A_i}), \tag{27a}$$

$$\phi_{10}^*(\overbrace{1}^{X_{S_i}},\overbrace{10}^{Y_i},\overbrace{0}^{A_{i-1}})=(\overbrace{0}^{Z_i},\overbrace{1}^{A_i}), \tag{27b}$$

$$\phi_{10}^*(\overbrace{0}^{X_{S_i}},\overbrace{10}^{Y_i},\overbrace{1}^{A_{i-1}})=(\overbrace{0}^{Z_i},\overbrace{0}^{A_i}), \tag{27c}$$

$$\phi_{10}^*(\overbrace{1}^{X_{S_i}},\overbrace{10}^{Y_i},\overbrace{1}^{A_{i-1}})=(\overbrace{0}^{Z_i},\overbrace{0}^{A_i}). \tag{27d}$$

*The optimal stationary policy in (27) induces the sub-state sequence $\{A_i\}_{i\in\mathbb{N}^+}$ to alternate in a $(0,1,0,1,\dots)$ ping-pong manner, as shown in Fig. 4-**(a)**. This alternating behavior results in the RVI oscillations as shown in Fig. 4-**(b)**.*

Lemma 4 and Example 1 show that the RVI algorithm [102] may not asymptotically converge to $U(\lambda)$, and the reason is the *periodic* nature inherent in the transformed MDP $\mathscr{P}_{\text{MDP}}(\lambda)$. Consequently, the existing two-layer Bisec-RVI Algorithm (e.g., [1, Algorithm 1], [76, Algorithm 1], and [89, Algorithm 1]) cannot reliably determine the root $\rho^\star$. To circumvent this problem, we propose a new iterative approach, namely τ-RVI, in Iteration 2. This approach eliminates the need for condition (b) in Lemma 4 but guarantees rigorous convergence.

Iteration 2. *(τ-RVI Algorithm). For a given λ and a parameter $0<\tau\leq 1$, the τ-RVI iteratively generate sequences $\{\tilde{U}_K(\lambda)\}^{K\in\mathbb{N}^+}$ and $\{\tilde{V}_K(\gamma;\lambda)\}^{K\in\mathbb{N}^+}_{\gamma\in\mathcal{S}\times\mathcal{Y}\times\mathcal{A}}$ with a starting initial value $\{\tilde{V}_0(\gamma;\lambda)\}_{\gamma\in\mathcal{S}\times\mathcal{Y}\times\mathcal{A}}$.*

$$\tilde{U}_{K+1}(\lambda)=\min_{A_i,Z_i}\Big\{g(\gamma^r,Z_i,A_i;\lambda) + \tau\,\mathbb{E}\big[\tilde{V}_K(\gamma';\lambda)\mid\gamma^r,Z_i,A_i\big]\Big\}, \tag{28a}$$

$$\tilde{V}_{K+1}(\gamma;\lambda)=(1-\tau)\,\tilde{V}_K(\gamma;\lambda) + \min_{A_i,Z_i}\Big\{g(\gamma,Z_i,A_i;\lambda) + \tau\,\mathbb{E}\big[\tilde{V}_K(\gamma';\lambda)\mid\gamma,Z_i,A_i\big]\Big\} - \tilde{U}_{K+1}(\lambda),\forall\gamma\in\mathcal{S}\times\mathcal{Y}\times\mathcal{A}, \tag{28b}$$

where $\gamma^r\in\mathcal{S}\times\mathcal{Y}\times\mathcal{A}$ is a predefined fixed reference state with initial condition $\tilde{V}_0(\gamma^r;\lambda)=0$.

Remark 3. *If $\tau=1$, then τ-RVI reduces to RVI.*

Fig. 4-**(b)** illustrates the convergence of τ-RVI across various values of the parameter τ, compared to the standard RVI algorithm [102]. The results demonstrate that τ-RVI overcomes oscillatory behavior encountered by RVI, with convergence rates depending on the selected τ values. A rigorous convergence analysis of τ-RVI will be presented in in section IV-C.

2) Outer-Layer Bisection and Bounds on $\rho^\star$

In the outer layer of Algorithm 1, the search interval $(\lambda_\downarrow,\lambda_\uparrow)$ is bisected on a slow time scale to approximate the root of $U(\lambda)$. This process relies on the uniqueness of the root of $U(\lambda)$ established in Lemma 2. For the bisection search process, the complexity mainly depends on the initialization of the search interval $(\lambda_\downarrow,\lambda_\uparrow)$, which requires establishing upper and lower bounds on $\rho^\star$. To address this, we establish the bounds on $\rho^\star$ to initialize the bisection search:

Lemma 5 (Upper and Lower Bounds on $\rho^\star$)**.** *The lower bound of $\rho^\star$ can be defined by the minimum value of the cost function, given by*

$$\rho^\star\geq\min_{s,a}\mathcal{C}(s,a). \tag{29}$$

The upper bound of $\rho^\star$ can be defined by the minimum stationary cost achievable under a constant action, given by

$$\rho^\star\leq\min_a\sum_{s\in\mathcal{S}}\pi_a(s)\cdot\mathcal{C}(s,a), \tag{30}$$

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

Proof. See Appendix F. ∎

C. Convergence of τ-RVI

In this subsection, we present convergence results for τ-RVI in Iteration 2. We theoretically show that the generated

sequences in τ-RVI will asymptotically approach the solution to the ACOE (23). To quantify the convergence behavior, we define the *relative error* for the K-th iteration value $\tilde{U}_K(\lambda)$ with respect to the ACOE solution $U(\lambda)$ as:

$$e_{\mathrm{U}}^{(K)}(\lambda) \triangleq |\tilde{U}_K(\lambda) - U(\lambda)|. \tag{31}$$

Similarly, define the relative error for the K-th iteration value $\tilde{V}_K(\gamma;\lambda)$ with respect to the ACOE solution $V^\star(\gamma;\lambda)$ as:

$$e_{\mathrm{V}}^{(K)}(\gamma;\tau,\lambda) \triangleq \left|\tilde{V}_K(\gamma;\lambda) - \frac{V^\star(\gamma;\lambda)}{\tau}\right|. \tag{32}$$

The main convergence results for τ-RVI are summarized in Theorem 2 below.

Theorem 2. *(Convergence of τ-RVI). If the MDP $\mathscr{P}_{MDP}(\lambda)$ is a unichain MDP, then the following limits hold true:*
(i). For $\forall\lambda\in\mathbb{R}$,

$$\lim_{K\to\infty} e_{\mathrm{U}}^{(K)}(\lambda) = 0. \tag{33}$$

(ii). For $\forall\lambda\in\mathbb{R}$, $\forall 0<\tau<1$, and $\forall\gamma\in\mathcal{S}\times\mathcal{Y}\times\mathcal{A}$,

$$\lim_{K\to\infty} e_{\mathrm{V}}^{(K)}(\gamma;\tau,\lambda) = 0. \tag{34}$$

Proof Sketch. The motivation behind (28) is to formulate an alternative MDP problem, denoted as $\widetilde{\mathscr{P}_{\mathrm{MDP}}}(\lambda)$, by eliminating potential *periodicity* in the transition probabilities in (18). We denote the transition probability from state i to state j, given action a, as $p_{ij}(a)$ in $\mathscr{P}_{\mathrm{MDP}}(\lambda)$, and as $\widetilde{p_{ij}}(a)$ in $\widetilde{\mathscr{P}_{\mathrm{MDP}}}(\lambda)$. The alternative transition probability $\widetilde{p_{ij}}(a)$ is defined as:

$$\widetilde{p_{ij}}(a) \triangleq \begin{cases} \tau p_{ij}(a), & \text{if } i\neq j \\ 1-\tau+\tau p_{ii}(a), & \text{if } i=j. \end{cases} \tag{35}$$

It is easy to verify that $\sum_j \widetilde{p_{ij}}(a) = 1$ and $\widetilde{p_{ii}}(a) > 0$ for $\forall i \in \mathcal{S}\times\mathcal{Y}\times\mathcal{A}$, and thus the new MDP $\widetilde{\mathscr{P}_{\mathrm{MDP}}}(\lambda)$ is *aperiodic*. Then, we can formulate the ACOE for the alternative $\widetilde{\mathscr{P}_{\mathrm{MDP}}}(\lambda)$:

$$\tilde{V}^\star(\gamma;\lambda)+\tilde{U}(\lambda) = \min_{A_i,Z_i}\Big\{g(\gamma,Z_i,A_i;\lambda)+ \sum_{\gamma'}\widetilde{p_{\gamma\gamma'}}(Z_i,A_i)\tilde{V}^\star(\gamma';\lambda)\Big\}, \forall\gamma\in\mathcal{S}\times\mathcal{Y}\times\mathcal{A}, \tag{36}$$

which can be solved by the traditional RVI with guaranteed convergence because of its *unichain* and *aperiodic* property. Then, comparing (23) and (36), we can establish the relationship that $\tau\tilde{V}^\star(\gamma;\lambda) = V^\star(\gamma;\lambda)$ for $\forall\gamma\in\mathcal{S}\times\mathcal{Y}\times\mathcal{A}$ and $\tilde{U}(\lambda)=U(\lambda)$. See Appendix I for the detailed proof. ■

The next theorem characterizes an upper bound on the *relative error* of the τ-RVI Algorithm, whose proof is included in Appendix I.

Theorem 3. *(Upper Bound of Relative Error). If the MDP $\mathscr{P}_{MDP}(\lambda)$ is a unichain MDP, then up to iteration K, the relative error $e_{\mathrm{U}}^{(K)}(\lambda)$ is upper bounded above by*

$$e_{\mathrm{U}}^{(K)}(\lambda) \le \frac{\tau M(1-\epsilon)^{(K-1)/L}}{1-(1-\epsilon)^{1/L}} = \mathcal{O}\left(\frac{1}{R^K}\right), \tag{37}$$

where M is a scaling factor, L is defined by (129). The term $R=\frac{1}{(1-\epsilon)^{1/L}}>1$ captures the asymptotic convergence rate.

Theorem 3 demonstrates that the upper bound of the *relative error* decreases **exponentially** with respect to the number of iterations K. This indicates that the proposed method exhibits faster convergence, as the number of inner iterations required to achieve a given accuracy δ is at most **logarithmic**, *i.e.*, $K\le\mathcal{O}(\log(1/\delta))$.

V. Optimal Sampling Without Rate Constraint: A One-Layer Primal-Dinkelbach Approach

The two-layer algorithm requires repeatedly executing the *computation-intensive* τ-RVI to evaluate $U(\lambda)$ at the fast time scale and update λ on the slow timescale. This results in high complexity since each update of λ in the outer-layer search necessitates running the inner layer value iterations. To address this, we develop efficient *one-layer* iterations in this section that eliminate the need for outer-layer bisection search. The key idea here is to treat the constraint $U(\rho^\star)=0$ as an intrinsic condition within the Markov Decision Process (MDP) $\mathscr{P}_{\mathrm{MDP}}(\rho^\star)$, allowing us to directly establish the *optimality equations*, which take the form of *fixed-point equations*, as shown in [1]. In this work, we show through an example that the *fixed-point operation* in [1] may not converge due to the *non-contractive* nature of the operator (see Section V-A). To address divergence, we develop a new one-layer iterative algorithm that guarantees provable convergence to the *fixed point* (see Section V-B).

A. Fixed-Point Equations and Iterations

We demonstrate here that the root finding process on $U(\lambda)$ is equivalent to solving the following non-linear equations, which we will then show to be *fixed-point equations*.

Theorem 4. *Solving Problem $\mathscr{P}_{\mathrm{MDP}}(\rho^\star)$ with $U(\rho^\star)=0$ is equivalent to solving the following nonlinear equations:*

$$\begin{cases} W^\star(\gamma) = \min\limits_{A_i,Z_i}\Big\{g(\gamma,A_i,Z_i;\rho^\star)+\mathbb{E}[W^\star(\gamma')\,|\gamma,Z_i,A_i]\Big\}, \\ \qquad\qquad\qquad\qquad \gamma\in\mathcal{S}\times\mathcal{Y}\times\mathcal{A}, \\ \rho^\star = \min\limits_{A_i,Z_i}\left\{\dfrac{q(\gamma^{\mathrm{r}},A_i,Z_i)+\mathbb{E}[W^\star(\gamma')\,|\gamma^{\mathrm{r}},A_i,Z_i]}{f(Z_i)}\right\}, \end{cases} \tag{38}$$

where $\gamma^{\mathrm{r}}\in\mathcal{S}\times\mathcal{Y}\times\mathcal{A}$ is the reference state and can be arbitrarily chosen.

Proof. See Appendix G. ■

In (38), there are $|\mathcal{S}|\times|\mathcal{Y}|\times|\mathcal{A}|+1$ variables, *i.e.*, $\rho^\star$ and $W^\star(\gamma),\gamma\in\mathcal{S}\times\mathcal{Y}\times\mathcal{A}$, which are matched by an equal number of equations. Solving non-linear equations is generally challenging, however, we establish that the equations (38) are *fixed-point equations*. Let $\mathbf{W}^\star$ denote the vector consisting of $W^\star(\gamma)$ for all $\gamma\in\mathcal{S}\times\mathcal{Y}\times\mathcal{A}$, (38) can be succinctly represented as follows:

$$\begin{cases} \mathbf{W}^\star = T(\mathbf{W}^\star,\rho^\star) \\ \rho^\star = H(\mathbf{W}^\star), \end{cases} \tag{39}$$

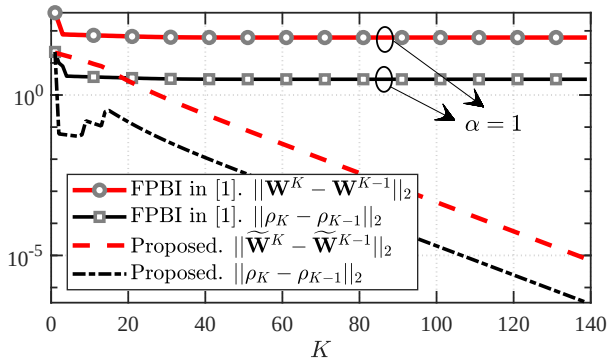


Fig. 5. Convergence comparison between FPBI [1, Algorithm 2] and ONEPDSI (Iteration 3). FPBI becomes *non-expansive* and thus diverges. The proposed ONEPDSI converges.

where $T(\cdot) : \mathbb{R}^{|\mathcal{S}\times\mathcal{Y}\times\mathcal{A}|} \times \mathbb{R}^{+} \to \mathbb{R}^{|\mathcal{S}\times\mathcal{Y}\times\mathcal{A}|}$ is a non-linear operator corresponding to the first equations of (39) and $H(\cdot) : \mathbb{R}^{|\mathcal{S}\times\mathcal{Y}\times\mathcal{A}|} \to \mathbb{R}^{+}$ is an operator corresponding to the second equation of (39). Substituting the second equation $\rho^\star = H(\mathbf{W}^\star)$ into the first equation of (39) yields $\mathbf{W}^\star = T(\mathbf{W}^\star, H(\mathbf{W}^\star))$. Define a new operator G as $G(\mathbf{W}) \triangleq T(\mathbf{W}, H(\mathbf{W}))$, as implied by (39), we have the *fixed-point equation*:

$$\mathbf{W}^\star = G(\mathbf{W}^\star). \tag{40}$$

Then, as was done in [1, Algorithm 2], we can conduct the following *fixed-point iterations* to asymptotically approximate the solution to (38):

$$\begin{cases} \mathbf{W}^{K+1} = G(\mathbf{W}^K), \\ \rho_{K+1} = H(\mathbf{W}^{K+1}). \end{cases} \tag{41}$$

Typically, the convergence of fixed-point iterations is established using the *Banach Contraction Mapping Theorem* [91, Theorem 6]. According to this theorem, if the operator $Q(\cdot) : \mathbb{R}^{|\mathcal{S}\times\mathcal{Y}\times\mathcal{A}|} \to \mathbb{R}^{|\mathcal{S}\times\mathcal{Y}\times\mathcal{A}|}$ is a *contraction mapping*, *i.e.*, for any $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{|\mathcal{S}\times\mathcal{Y}\times\mathcal{A}|}$, there exists a constant $0 < \alpha < 1$ such that

$$||G(\mathbf{W}_1) - G(\mathbf{W}_2)||_2 \le \alpha ||\mathbf{W}_1 - \mathbf{W}_2||_2, \tag{42}$$

then the operator G has a unique fixed point, and the fixed-point iteration described in (41) converges to the solution $\mathbf{W}^\star$, where $\mathbf{W}^\star = G(\mathbf{W}^\star)$.

However, the following **counter-example** shows that the operator $G : \mathbb{R}^{|\mathcal{S}\times\mathcal{Y}\times\mathcal{A}|} \to \mathbb{R}^{|\mathcal{S}\times\mathcal{Y}\times\mathcal{A}|}$ may be a *non-expansive mapping*[10] rather than a *contraction mapping*, which results in potential divergence.

Example 2. *(Divergence Example of FPBI [1, Algorithm 2]). Consider the parameter setup described in Appendix H where the delay is $p = 0$. In this case, the fixed-point iteration starting from $\mathbf{W}^0 = \mathbf{0}$ will diverge, as shown in Fig. 5.*

[10] *Non-expansive mapping* indicates $\alpha = 1$ in (42). Fig. 5 demonstrates that the fix-point operation in (41) may be *non-expansive*.

B. Primal-Dinkelbach Synchronous Iteration (Proposed ONEPDSI)

To overcome the divergence limitations of FPBI, we develop a novel one-layer iterative approach that guarantees rigorous convergence to the solution of the *fixed-point equations* (38). The details of the iteration are given as:

Iteration 3. *(ONEPDSI): For a given $0 < \kappa < 1$, we can iteratively generate sequences $\{\rho_K\}^{K\in\mathbb{N}^+}$ and $\{\widetilde{W}_K(\gamma)\}^{K\in\mathbb{N}^+}_{\gamma\in\mathcal{S}\times\mathcal{Y}\times\mathcal{A}}$ with a starting initial value $\{\widetilde{W}_0(\gamma)\}_{\gamma\in\mathcal{S}\times\mathcal{Y}\times\mathcal{A}}$:*

$$\rho_{K+1} = \min_{A_i,Z_i} \left\{ \frac{q(\gamma^r, Z_i, A_i) - \kappa\widetilde{W}_K(\gamma^r)\cdot\mathbb{E}[Y_i]}{f(Z_i)} + \frac{\kappa\mathbb{E}\left[\widetilde{W}_K(\gamma') \mid \gamma^r, Z_i, A_i\right]\cdot\mathbb{E}[Y_i]}{f(Z_i)} \right\} + \widetilde{W}_K(\gamma^r), \tag{43a}$$

$$\widetilde{W}_{K+1}(\gamma) = \min_{A_i,Z_i} \left\{ \frac{q(\gamma, Z_i, A_i) - \kappa\widetilde{W}_K(\gamma)\cdot\mathbb{E}[Y_i]}{f(Z_i)} + \frac{\kappa\mathbb{E}\left[\widetilde{W}_K(\gamma') \mid \gamma, Z_i, A_i\right]\cdot\mathbb{E}[Y_i]}{f(Z_i)} \right\} + \widetilde{W}_K(\gamma) - \rho_{K+1}, \quad \forall\gamma\in\mathcal{S}\times\mathcal{Y}\times\mathcal{A}. \tag{43b}$$

where $\gamma^r \in \mathcal{S}\times\mathcal{Y}\times\mathcal{A}$ is a fixed reference state with an initial condition $\widetilde{W}_0(\gamma^r) = 0$.

In what follows, we present a comprehensive convergence analysis demonstrating that ONEPDSI ensures robust and provable convergence.

C. Convergence of ONEPDSI

In this subsection, we theoretically demonstrate that the sequences in ONEPDSI will approach the solution to (38). To quantify the convergence, we define the *relative error* of ρ_K at iteration K as:

$$e_\rho^{(K)} = |\rho_K - \rho^\star|. \tag{44}$$

Meanwhile, define the *relative error* of the sequence $\widetilde{W}_K(\gamma)$ at K-th iteration as

$$e_{\mathrm{W}}^{(K)}(\gamma;\kappa) = \left|\widetilde{W}_K(\gamma) - \frac{W^\star(\gamma)}{\kappa\cdot\mathbb{E}[Y_i]}\right|. \tag{45}$$

The following theorem demonstrates the convergence of ONEPDSI.

Theorem 5. *(Convergence of ONEPDSI). If the transformed MDP is unichain and $0 < \kappa < 1$, then the ONEPDSI in (43) is convergent, with:*

$$\begin{aligned} &\lim_{k\to\infty} e_\rho^{(K)} = 0, \\ &\lim_{k\to\infty} e_{\mathrm{W}}^{(K)}(\gamma) = 0, \forall\gamma\in\mathcal{S}\times\mathcal{Y}\times\mathcal{A}. \end{aligned} \tag{46}$$

Proof. See Appendix J. ■

The next theorem characterizes an upper bound on the *relative error* of the proposed ONEPDSI, whose proof is

provided in Appendix J:

Theorem 6. *(Upper Bound of Relative Error). If the MDP $\mathscr{P}_{MDP}(\lambda)$ is a unichain MDP, then up to iteration K, the relative error $e_\rho^{(K)}$ is upper bounded above by*

$$e_\rho^{(K)} \leq \frac{\tau M(1-\epsilon)^{(K-1)/L}}{1-(1-\epsilon)^{1/L}} = \mathcal{O}\left(\frac{1}{R^K}\right), \tag{47}$$

where M is a scaling factor, L is defined by (182). The term $R = \frac{1}{(1-\epsilon)^{1/L}}$ captures the asymptotic convergence rate.

Theorem 6 demonstrates that the upper bound of the *relative error* decreases **exponentially** with respect to the number of iterations K. This indicates that the number of inner iterations required to achieve a given optimality gap δ is at most **logarithmic**, *i.e.*, $K \leq \mathcal{O}\left(\log(1/\delta)\right)$.

VI. Optimal Sampling With Rate Constraint: A Typical Three-Layer Approach

In this section, we investigate the optimal sampling and decision-making policy that minimizes the long-term average cost under a sampling frequency constraint, as formulated in Problem 1. Our goal is to derive this optimal policy and its corresponding value $h^\star$.

A. Lagrangian Dual Techniques

Following the steps in (14a)-(17) and applying Dinkelbach's method for non-linear fractional programming as in [99] and [20, Lemma 2], the problem of determining the optimal policy for Problem 1 is equivalent to solving the following alternative problem given the *Dinkelbach* parameter λ:

Problem 5 (*Standard Infinite-Horizon Constrained Markov Decision Process (CMDP) with Dinkelbach Parameter λ*)**.**

$$H(\lambda; f_{\max}) \triangleq$$
$$\inf_{\boldsymbol{\psi}} \lim_{n\to\infty} \frac{1}{n}\sum_{i=0}^{n-1}\left\{\mathbb{E}_{\boldsymbol{\psi}}\left[\sum_{t=D_i}^{D_{i+1}-1}\mathcal{C}(X_t,A_i)\right] - \lambda\mathbb{E}_{\boldsymbol{\psi}}\left[Z_i+Y_{i+1}\right]\right\}$$
$$s.t. \lim_{T\to\infty}\frac{1}{T}\mathbb{E}_{\boldsymbol{\psi}}\left[\sum_{i=1}^{T}(S_{i+1}-S_i)\right] \geq \frac{1}{f_{max}}, \tag{48}$$

The following lemma characterizes the relationship between $h^\star$ and the optimal value of Problem 5. The proof follows directly from modifying the policies in [1, Appendix B] to satisfy the sampling frequency constraint, and is therefore omitted.

Lemma 6. *The following assertions hold:*
(i). $h^\star \gtreqless \lambda$ if and only if $H(\lambda; f_{\max}) \gtreqless 0$.
(ii). When $H(\lambda; f_{\max}) = 0$, the solutions to Problem 5 coincide with those of Problem 1.
(iii). $H(\lambda; f_{\max}) = 0$ has a unique root, and the root is $h^\star$.

With Lemma 6 in hand, solving Problem 1 is equivalent to solving for the root of the implicit function $H(\lambda; f_{\max})$. To obtain $H(\lambda; f_{\max})$ given λ, we first transform the CMDP into an unconstrained Lagrangian MDP problem. Specifically, define the *Lagrange Function* as:

$$\mathcal{L}(\boldsymbol{\psi};\theta,\lambda,f_{\max}) = \frac{\theta}{f_{\max}} + \lim_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1}\mathbb{E}_{\boldsymbol{\psi}}\left\{\sum_{t=D_i}^{D_{i+1}-1}\mathcal{C}\left(X_t,A_i\right) - (\lambda+\theta)\left(Z_i+Y_{i+1}\right)\right\}, \tag{49}$$

where $\theta \geq 0$ is the *Lagrangian multiplier*. Let the *Lagrangian Dual Function* defined as:

$$\Upsilon(\theta,\lambda;f_{\max}) \triangleq \inf_{\boldsymbol{\psi}} \mathcal{L}(\boldsymbol{\psi};\theta,\lambda,f_{\max}). \tag{50}$$

Since an optimal *stationary deterministic* policy exists for the MDP problem $\inf_{\boldsymbol{\psi}} \mathcal{L}(\boldsymbol{\psi};\theta,\lambda,f_{\max})$ (as indicated in Theorem 1), we can use a short-hand notation ϕ to denote $\boldsymbol{\psi}$, and the *Lagrangian Dual Problem* of Problem 5 is:

Problem 6 (*Lagrangian Dual Problem*)**.**

$$d(\lambda; f_{\max}) = \max_{\theta\geq 0}\Upsilon(\theta,\lambda;f_{\max}), \tag{51}$$

where $\Upsilon(\theta,\lambda;f_{\max}) = \inf_\phi \mathcal{L}(\phi;\theta,\lambda,f_{\max})$.

The *weak duality principle* [104, Chapter 5.2.2] implies that $d(\lambda; f_{\max})$ is a lower bound of $H(\lambda; f_{\max})$, *i.e.*, $d(\lambda; f_{\max}) \leq H(\lambda; f_{\max})$. In the following lemma, we establish the conditions where the *strong duality* holds true and thus $d(\lambda; f_{\max}) = H(\lambda; f_{\max})$. Under these conditions, it is sufficient to solve $d(\lambda; f_{\max})$ to obtain $H(\lambda; f_{\max})$.

Lemma 7. *(Restatement of [104, Chapter 5.5.3]) The duality gap between Problem 5 and Problem 6 is zero,* i.e., *$d(\lambda; f_{\max}) = H(\lambda; f_{\max})$, if and only if for any given λ, we can find $\phi^\star_{\lambda+\theta^\star_\lambda}$ and $\theta^\star_\lambda$ such that the Karush–Kuhn–Tucker (KKT) conditions are satisfied:*

$$\theta^\star_\lambda \geq 0, \tag{52a}$$
$$\phi^\star_{\lambda+\theta^\star_\lambda} = \arg\min_\phi \lim_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1}\mathbb{E}_\phi\left\{\sum_{t=D_i}^{D_{i+1}-1}\mathcal{C}\left(X_t,A_i\right) - (\lambda+\theta^\star_\lambda)\left(Z_i+Y_{i+1}\right)\right\} + \frac{\theta^\star_\lambda}{f_{\max}}, \tag{52b}$$
$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1}\mathbb{E}_{\phi^\star_{\lambda+\theta^\star_\lambda}}\left[Z_i+Y_{i+1}\right] \geq \frac{1}{f_{\max}}, \tag{52c}$$
$$\theta^\star_\lambda\left\{\lim_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1}\mathbb{E}_{\phi^\star_{\lambda+\theta^\star_\lambda}}\left[Z_i+Y_{i+1}\right] - \frac{1}{f_{\max}}\right\} = 0. \tag{52d}$$

By leveraging Lemma 7, we reformulate the constrained problem as an unconstrained Problem 6. Given a fixed *Dinkelbach* parameter λ, the goal of Problem 6 is to determine the *saddle point* $(\phi^\star_{\lambda+\theta^\star_\lambda}, \theta^\star_\lambda)$ of the function $\mathcal{L}(\phi;\theta,\lambda)$. The inner layer of Problem 6 is a standard MDP problem with fixed *Lagrangian multiplier* $\theta \geq 0$ and *Dinkelbach parameter* λ, while the outer layer seeks the optimal *Lagrangian multiplier* $\theta^\star_\lambda$ that maximizes the *Lagrangian Dual Function* (50) under the KKT conditions (52a)-(52d).

B. Three-Layer Solutions and the Structure of Optimal Policies

In this subsection, we propose a **three-layer** algorithm outlined in Algorithm 2. The basic framework of this algorithm is inspired by [90, Section IV.C]. This algorithm consists of inner, middle, and outer layers, whose implementations are detailed below.

1) Inner-Layer: A Standard MDP Given λ and θ

For any given θ and λ, the inner layer $\inf_{\phi} \mathcal{L}(\phi; \theta, \lambda, f_{\max})$ is a standard unconstrained infinite horizon MDP as defined in IV-B1, which is denoted by $\mathscr{P}_{\text{MDP}}(\theta+\lambda)$ with the optimal value $U(\lambda+\theta)$. This problem can be efficiently solved using the τ-RVI algorithm, as outlined in Iteration 2. Notably, [90, Section IV.C] applies a value iteration process in the inner layer, while we apply the τ-RVI to ensure the rigorous convergence in our context.

2) Middle-Layer: Update Lagrangian multiplier θ Given λ

Given a *Dinkelbach parameter* λ, the middle layer involves solving $\max_{\theta\geq 0} \Upsilon(\theta, \lambda; f_{\max})$ in Problem 6 to accurately approximate $d(\lambda; f_{\max})$, which searches for the optimal $\theta^{\star}_{\lambda}$ and its corresponding optimal policy $\phi^{\star}_{\lambda+\theta^{\star}_{\lambda}}$ that satisfy the KKT conditions in Lemma 7. Different from [90, Section IV.C] where the sub-gradient method is employed to update the Lagrangian multiplier leveraging the convexity of the Lagrangian Dual function, we conduct a monotonicity analysis and explicitly state the conditions for the optimal multiplier. We begin by introducing the following notations that help clarify the role of the multiplier θ given a fixed λ. Specifically,

$$\mathcal{Q}^{\lambda+\theta} \triangleq \limsup_{\mathrm{n}\to\infty} \frac{1}{n}\sum_{i=0}^{n-1} \mathbb{E}_{\phi^{\star}_{\lambda+\theta}}\left[\sum_{t=D_i}^{D_{i+1}-1} \mathcal{C}(X_t, A_i)\right], \tag{53}$$

$$\mathcal{F}^{\lambda+\theta} \triangleq \liminf_{\mathrm{n}\to\infty} \frac{1}{n}\sum_{i=0}^{n-1} \mathbb{E}_{\phi^{\star}_{\lambda+\theta}}\left[Z_i + Y_{i+1}\right], \tag{54}$$

where $\phi^{\star}_{\lambda+\theta}$ is the optimal *stationary deterministic* policy for the unconstrained problem $\mathscr{P}_{\text{MDP}}(\lambda+\theta)$ given in IV-B1. The subsequent Lemma presents key properties of $\mathcal{F}^{\lambda+\theta}$, $\mathcal{Q}^{\lambda+\theta}$, $U(\lambda+\theta)$, and $\Upsilon(\theta, \lambda; f_{\max})$ with respect to θ, which are useful for the explicit solutions.

Lemma 8. *(A variant of [105, Lemma 3.1]) The following assertions hold true:*

(i) $U(\lambda+\theta)$ is non-increasing with respect to θ;

(ii) $\mathcal{Q}^{\lambda+\theta}$ and $\mathcal{F}^{\lambda+\theta}$ are non-decreasing functions with respect to θ;

(iii) $\mathcal{Q}^{\lambda+\theta}$ and $\mathcal{F}^{\lambda+\theta}$ are both step functions with respect to θ;

(iv) If $\mathcal{F}^{\lambda+\theta} \geq 1/f_{\max}$, $\Upsilon(\theta, \lambda; f_{\max})$ is non-increasing with respect to θ.

Proof. Part (i) and part (ii) are supported by [105, Lemma 3.1]. Part (iii) is supported by [105, Lemma 3.2]. See Appendix K for the detailed proof of part (iv). ■

Lemma 8 leads to the following corollary, which provides an explicit solution for the optimal value $\theta^{\star}_{\lambda}$ in Problem 6 that satisfies the KKT conditions (52a)-(52d).

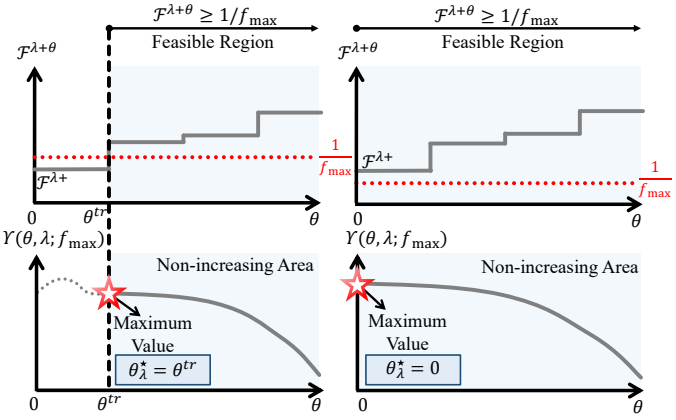


Fig. 6. Illustrations of Case (i) and case (ii) in Corollary 1. In this figure, $\mathcal{F}^{\lambda+\theta}$ is a non-increasing step function with respect to θ, as Lemma 8-(ii) and Lemma 8-(iii) indicate. In addition, $\Upsilon(\theta, \lambda; f_{\max})$ is non-increasing with respect to θ if $\mathcal{F}^{\lambda+\theta} \geq 1/f_{\max}$, as Lemma 8-$(iv)$ indicates.

Corollary 1. *Denote $\mathcal{F}^{\lambda^+}$ as the right limit*[11] *of $\mathcal{F}^{\lambda}$:*

$$\mathcal{F}^{\lambda^+} = \lim_{\Delta\lambda\to 0} \mathcal{F}^{\lambda+\Delta\lambda}, \tag{55}$$

The following assertions hold true:

(i). If $\mathcal{F}^{\lambda^+} \geq 1/f_{\max}$, then $\theta^{\star}_{\lambda} = 0$;

(ii). If $\mathcal{F}^{\lambda^+} < 1/f_{\max}$, then $\theta^{\star}_{\lambda}$ is equal to a positive break point $\theta^{tr} > 0$, which satisfies:

$$\mathcal{F}^{(\theta^{tr}+\lambda)^-} < \frac{1}{f_{\max}} \leq \mathcal{F}^{(\theta^{tr}+\lambda)^+}, \tag{56}$$

Proof Sketch. If $\mathcal{F}^{\lambda+} \geq \frac{1}{f_{\max}}$, which is illustrated in the right panel of Fig. 6, the function $\Upsilon(\theta, \lambda; f_{\max})$ is non-increasing with θ for $\theta \geq 0$. In this case, the maximum value of $\Upsilon(\theta, \lambda; f_{\max})$ is obtained at $\theta^{\star}_{\lambda} = 0$. If $\mathcal{F}^{\lambda+} < \frac{1}{f_{\max}}$, the feasible region under the KKT condition (52c) is shown in the left panel of Fig. 6 with $\theta \geq \theta^{\text{tr}}$. In this feasible region, the function $\Upsilon(\theta, \lambda; f_{\max})$ is non-increasing in θ. Therefore, the maximum value of $\Upsilon(\theta, \lambda; f_{\max})$ is obtained at $\theta^{\star}_{\lambda} = \theta^{\text{tr}}$. A detailed proof is provided in Appendix L. ■

Having established Corollary 1, our remaining task is to identify the threshold value θ^{tr} that satisfies (56) under the condition that $\mathcal{F}^{\lambda^+} < 1/f_{\max}$. We here introduce two algorithms for searching this threshold:

- **Bisection Search**: Given that $\mathcal{F}^{\lambda+\theta}$ is non-decreasing in θ, we can apply a *bisection search* to gradually converge on the threshold θ^{tr}. This method is a classical approach for locating thresholds in monotonic functions.
- **Intersection Search** [106, Algorithm 2]: To improve the efficiency of solving constrained MDPs, [106] further exploits the piece-wise linear and concave (PWLC) structure Lagrangian cost function under finite state and action spaces. Unlike traditional bisection methods that locate the zero-crossing of a single monotonic function, *intersection search* identifies the intersection point of two tangents to the Lagrangian curve, thereby accelerating the

[11] Since $\mathcal{F}^{\lambda}$ is a step function, it does not necessarily follow that $\mathcal{F}^{\lambda^+} = \mathcal{F}^{\lambda}$. Specifically, when λ is a break point, we have $\mathcal{F}^{\lambda^+} > \mathcal{F}^{\lambda}$.

search for the optimal Lagrange multiplier θ^{tr} with fewer iterations. We refer readers to [106, Section V.A] for a detailed discussion of this algorithm.

Once this threshold value is obtained, setting $\theta^\star_\lambda = \theta^{\text{tr}}$ ensures that the KKT conditions (52a)–(52c) are satisfied. What remains is to determine the optimal policy $\phi_{\lambda+\theta^\star_\lambda}$ that guarantees the KKT condition (52d). The structure of this optimal policy is provided in the following theorem.

Theorem 7. *(Structure of the optimal policy) The following assertions hold true:*

(i). If $\mathcal{F}^{\lambda^+} \geq 1/f_{\max}$, the optimal policy is a stationary deterministic policy $\phi^\star_\lambda$ determined in (24);

(ii). If $\mathcal{F}^{\lambda^+} < 1/f_{\max}$ and $\mathcal{F}^{(\lambda+\theta^\star_\lambda)^+} = 1/f_{\max}$, the optimal policy is a stationary deterministic policy $\phi^\star_{(\lambda+\theta^\star_\lambda)}$;

(iii). If $\mathcal{F}^{\lambda^+} < 1/f_{\max}$ and $\mathcal{F}^{(\lambda+\theta^\star_\lambda)^+} > 1/f_{\max}$, the optimal policy is a random mixture of two stationary deterministic policy $\phi^\star_{(\lambda+\theta^\star_\lambda)^+}$ and $\phi^\star_{(\lambda+\theta^\star_\lambda)^-}$:

$$\phi^\star_{\lambda+\theta^\star_\lambda}(\gamma) = \begin{cases} \phi^\star_{(\lambda+\theta^\star_\lambda)^+}(\gamma), & w.p.\ \eta \\ \phi^\star_{(\lambda+\theta^\star_\lambda)^-}(\gamma), & w.p.\ 1-\eta \end{cases}, \text{ for } \forall \gamma \in \mathcal{S}\times\mathcal{Y}\times\mathcal{A}, \tag{57}$$

where η is a randomization factor given by

$$\eta = \frac{\frac{1}{f_{\max}} - \mathcal{F}^{\lambda+\theta^\star_\lambda}}{\mathcal{F}^{(\lambda+\theta^\star_\lambda)^+} - \mathcal{F}^{\lambda+\theta^\star_\lambda}}. \tag{58}$$

Proof. See Appendix M. ∎

With Theorem 7, we can determine the optimal policy once the optimal $\theta^\star_\lambda$ is obtained. In this way, the optimal value $d(\lambda; f_{\max})$ is determined as:

$$\begin{aligned} d(\lambda; f_{\max}) &= \Upsilon(\theta^\star_\lambda, \lambda; f_{\max}) \\ &= \begin{cases} U(\lambda) & \text{if } \mathcal{F}^{\lambda^+} \geq \frac{1}{f_{\max}} \\ U(\lambda+\theta^\star_\lambda) + \frac{\theta^\star_\lambda}{f_{\max}} & \text{if } \mathcal{F}^{\lambda^+} < \frac{1}{f_{\max}} \end{cases}. \end{aligned} \tag{59}$$

The following subsection aims at searching the root of $d(\lambda; f_{\max})$.

3) Outer-Layer: Update Dinkelbach parameter λ

With the approximation of $d(\lambda; f_{\max})$ in hand, the outer layer updates λ in a *bisection-search* fashion by leveraging Lemma 6 to finally approach the root $h^\star$ such that $d(h^\star; f_{\max}) = 0$. The flow of the algorithm is demonstrated in Algorithm 2.

Algorithm 2: *A Three-layer Algorithm (**A Variant of [90, Section IV.C]**)*

Input: Tolerence $\epsilon_1, \epsilon_2 > 0$, MDP $\mathscr{P}_{\text{MDP}}(\lambda)$, maximum sampling frequency $f_{\max}$

Initialization: sufficiently large $\lambda_\uparrow$, and $\lambda_\downarrow = \min_{s,a} \mathcal{C}(s,a)$;
// Outer-Layer Bisection Search on Dinkelbach parameter λ
while $\lambda_\uparrow - \lambda_\downarrow \geq \epsilon_1$ **do**
  $\lambda = (\lambda_\uparrow + \lambda_\downarrow)/2$;
  Run τ-RVI to solve $\mathscr{P}_{\text{MDP}}(\lambda)$ and calculate $\mathcal{F}^{\lambda^+}$ and $U(\lambda)$;
  if $\mathcal{F}^{\lambda^+} \geq 1/f_{\max}$ **then**
    $d(\lambda; f_{\max}) = U(\lambda)$;
  else
    Initialization: sufficiently large $\theta_\uparrow$, and $\theta_\downarrow = 0$;
    // Middle-Layer Bisection Search on Lagrangian multiplier θ
    while $\theta_\uparrow - \theta_\downarrow \geq \epsilon_2$ **do**
      $\theta = (\theta_\uparrow + \theta_\downarrow)/2$;
      // Inner-Layer MDP Given λ and θ
      Run τ-RVI to solve $\mathscr{P}_{\text{MDP}}(\lambda+\theta)$ and calculate $\mathcal{F}^{(\lambda+\theta)^+}$ and $U(\lambda+\theta)$;
      if $\mathcal{F}^{(\lambda+\theta)^+} \geq 1/f_{\max}$ **then**
        $\theta_\uparrow = \theta$;
      else
        $\theta_\downarrow = \theta$;
    $d(\lambda; f_{\max}) = U(\lambda+\theta) + \frac{\theta}{f_{\max}}$;
  if $d(\lambda; f_{\max}) > 0$ **then**
    $\lambda_\uparrow = \lambda$;
  else
    $\lambda_\downarrow = \lambda$;
Output: $h^\star = \lambda$

VII. Optimal Sampling With Rate Constraint: One-Layer Iteration is All You Need

In the three-layer algorithm, each update of λ or θ necessitates solving the inner-layer MDP using the τ-RVI algorithm. This process incurs a high computational complexity due to the repeated execution of the τ-RVI algorithm required to iteratively search and optimize the parameters λ and θ.

A. QuickBLP

In this section, we design a one-layer *two-stage* hierarchical algorithm namely, QuickBLP, which reduces the computational complexity by explicitly solving for the root $h^\star$. Rather than relying on iterative *bisection search* to find the root $h^\star$, our approach leverages a direct structural exploration of the solution space, allowing us to bypass the need for multiple τ-RVI executions. Consequently, the QuickBLP solves this problem in two stages. The first stage solves a Bellman variant and the second stage explicitly expresses the root $h^\star$ as a function of the solution to an LP problem, thus directly obtaining the root $h^\star$ in a more computationally efficient manner. The main structural results are summarized in the following theorem.

Theorem 8. *(Structural Results of the root $h^\star$) The following assertions hold true:*

(i). If the root of $U(\rho)$, denoted as $\rho^\star$, satisfies $\mathcal{F}^{(\rho^\star)^-} \geq 1/f_{\max}$, then the optimal value of Problem 1 is $h^\star = \rho^\star$, and the optimal policy for Problem 1 is $\phi^\star_{\rho^\star}$, as defined in (24);

(ii). If $\mathcal{F}^{(\rho^\star)^-} < 1/f_{\max}$, the optimal value of Problem 1

Algorithm 3: ***Proposed*** *One-layer* QUICKBLP

Input: MDP $\mathscr{P}_{\text{MDP}}(\lambda)$ and maximum sampling frequency $f_{\max}$

Run ONEPDSI in Iteration 3 to obtain $\rho^\star$;
Calculate the left limit $\mathcal{F}^{(\rho^\star)^-}$;
if $\mathcal{F}^{(\rho^\star)^-} \geq 1/f_{\max}$ **then**
  $h^\star = \rho^\star$; // Case (i) of Theorem 8
else
  Solve LP Problem 7 to obtain $Q^\star(f_{\max})$;
  Calculate the root $h^\star = f_{\max} \cdot Q^\star(f_{\max})$;
  // Case (ii) of Theorem 8

Output: $h^\star$

is

$$h^\star = f_{\max} \cdot Q^\star(f_{\max}), \tag{60}$$

where $Q^\star(f_{\max})$ is the optimal value of the following Linear Programming:

Problem 7 (*Linear Programming Reformulation*)**.**

$$Q^\star(f_{\max}) = \min_{\mathbf{x}} \sum_{\gamma,z,a} q(\gamma,z,a)x(\gamma,z,a) \tag{61a}$$

$$\text{s.t.} \quad \sum_{\gamma,z,a} f(z)x(\gamma,z,a) = \frac{1}{f_{\max}}, \tag{61b}$$

$$\sum_{z,a} x(\gamma',z,a) = \sum_{\gamma,z,a} p(\gamma'|\gamma,z,a)x(\gamma,z,a), \quad \forall \gamma' \in \mathcal{S}\times\mathcal{Y}\times\mathcal{A}, \tag{61c}$$

$$\sum_{\gamma,z,a} x(\gamma,z,a) = 1, \tag{61d}$$

$$x(\gamma,z,a) \geq 0, \quad \forall \gamma \in \mathcal{S}\times\mathcal{Y}\times\mathcal{A}, z \in \mathcal{Z}, a \in \mathcal{A}, \tag{61e}$$

and the corresponding optimal policy is a randomized policy given by:

$$\phi^\star(\gamma) = (z,a), \text{w.p. } \frac{x(\gamma,z,a)}{\sum_{\zeta,\alpha} x(\gamma,\zeta,\alpha)}, \quad \forall \gamma \in \mathcal{S}\times\mathcal{Y}\times\mathcal{A}, z \in \mathcal{Z}, a \in \mathcal{A}. \tag{62}$$

Proof. See Appendix N. ∎

Theorem 8 leads to a one-layer two-stage algorithm presented in Algorithm 3. In the first stage, the algorithm solves the unconstrained MDP $\mathscr{P}_{\text{MDP}}(\lambda)$ and determines the root of $U(\lambda)$, denoted by $\rho^\star$. This root is obtained by implementing ONEPDSI in Iteration 3. If the condition specified in part (i) of Theorem 8 holds true, then the root $h^\star$ is immediately found as $h^\star = \rho^\star$. If the condition of part (i) of Theorem 8 is not met, the algorithm proceeds to the second stage. Here, the LP problem[12] Problem 7 is solved. The solution of this LP provides the optimal value $Q^\star(f_{\max})$. Finally, the root $h^\star$ is explicitly determined as $h^\star = f_{\max} \cdot Q^\star(f_{\max})$. Compared to the previous Algorithm 2, this newly proposed algorithm eliminates the need for multiple executions of the τ-RVI for searching $h^\star$ and $\theta^\star_{h^\star}$.

[12]Once an LP problem is established, it can be solved to global optimality using well-established algorithms such as the *simplex* and modern *primal–dual interior-point* methods. For problems whose variable set is large, mature decomposition techniques such as *column generation* allow solving the LP without ever forming the entire matrix, while preserving exact optimality.

B. Polynomial Complexity Results

In this subsection, we analyze that both the methods proposed in this paper are **polynomial** in time and space complexity. For the three-layer Algorithm 2, the computational cost of a single τ-RVI iteration is $\mathcal{O}(|\mathcal{Z}||\mathcal{S}|^2|\mathcal{Y}|^2|\mathcal{A}|^3)$, and according to Theorem 3, the τ-RVI requires at most $\mathcal{O}\left(\log(1/\delta)\right)$ iterations to reach accuracy δ. Consequently, the time complexity of the three-layer Algorithm 2 is

$$\mathcal{O}\left(\log(\frac{1}{\epsilon_1})\log(\frac{1}{\epsilon_2})\log(\frac{1}{\delta})|\mathcal{Z}||\mathcal{S}|^2|\mathcal{Y}|^2|\mathcal{A}|^3\right), \tag{63}$$

which is **polynomial** in all problem parameters. In terms of space complexity, the algorithm stores one value vector $\tilde{V}_K(\gamma;\lambda)$ of size $\mathcal{O}(|\mathcal{S}\times\mathcal{Y}\times\mathcal{A}|)$ together with the sparse transition structure of size $\mathcal{O}(|\mathcal{Z}||\mathcal{S}|^2|\mathcal{Y}|^2|\mathcal{A}|^3)$. However, the overall computational cost can be dominated by the outer layers: the three-layer algorithm performs a large number of nested loops to reach accuracies ϵ_1, ϵ_2 and δ. When these tolerances are set small, the number of outer loops can be very high, and the resulting run time, though still polynomial, might become impractical.

In contrast, the proposed QUICKBLP addresses this challenge by eliminating the outer-loop search. In the first stage, ONEPDSI incurs $\mathcal{O}(|\mathcal{Z}||\mathcal{S}|^2|\mathcal{Y}|^2|\mathcal{A}|^3)$ per iteration and converges in $\mathcal{O}\big(\log(1/\delta)\big)$ iterations (Theorem 6), *i.e.*, the complexity of ONEPDSI is polynomial:

$$\mathcal{O}\left(\log(\frac{1}{\delta})|\mathcal{Z}||\mathcal{S}|^2|\mathcal{Y}|^2|\mathcal{A}|^3\right). \tag{64}$$

If the stopping condition of Theorem 8 is satisfied, the algorithm terminates, and no LP needs to be solved. Only when this condition is not met does the second stage become active, in which the LP Problem 7 with $n = |\mathcal{Z}| \times |\mathcal{S}| \times |\mathcal{Y}| \times |\mathcal{A}|^2$ variables and $m = |\mathcal{S}\times\mathcal{Y}\times\mathcal{A}|+2$ equality constraints must be solved. The LP is handled via the Column Generation (CG) method [107] to decompose the Master Problem (MP) into some manageable Restricted Master Problems (RMP). Each RMP is solved using a primal–dual interior-point method [104, Chap. 11]. Since the occupancy-measure LP admits optimal basic feasible solutions, the solution vector $\mathbf{x}$ can be chosen with at most m nonzeros; hence, at CG iteration t the RMP involves $n_t \leq m$ active columns. A primal–dual interior-point method requires $\mathcal{O}(\sqrt{m}\log(1/\epsilon))$ iterations [108, Theo. 3.1], each dominated by solving a sparse Schur system with n_t variables and m constraints, whose factorization cost $T_{\text{fact}}(m,n_t)$ ranges from $\mathcal{O}(m^{1.5})$ to $\mathcal{O}(m^2)$. As a result, the total cost for solving the LP via CG and primal–dual interior-point methods can be upper bounded by:

$$\mathcal{O}\left(\sum_{t=1}^{q} \sqrt{m}\log(1/\epsilon)\cdot T_{\text{fact}}(m,n_t)\right), \tag{65}$$

where $q \leq m$ is the number of CG iterations and $\sum_{t=1}^{q} \leq$

m^2. This yields a worst-case complexity between $\mathcal{O}(m^{3.5})$ to $\mathcal{O}(m^4)$. The space complexity to solve the LP is dominated by storing and factorizing the Schur complement system of order m. Since each RMP contains $n_t \leq m$ active columns, the overall memory requirement is $\mathcal{O}(m)$ to $\mathcal{O}(m^2)$, depending on the sparsity structure of the constraints.

Discussion: Both the algorithms achieve **polynomial** complexity. The three-layer algorithm admits a lower-degree polynomial bound in theory but suffers from heavy nested searches, which can make the runtime large in practice. In contrast, QUICKBLP adopts a two-stage design: in many instances the first stage alone suffices, and the second-stage LP is only occasionally invoked. Although the LP stage has a higher *worst-case* bound, the combination of CG and sparse interior-point methods makes the practical complexity often much lower than the theoretical worst case. As a result, QUICKBLP is typically far more efficient in real problem instances.

C. Approximate LP: Scaling to Large Spaces

In large-scale systems, the growth of the state space $\mathcal{S}$, the action space $\mathcal{A}$, and the delay space $\mathcal{Y}$ renders optimally solving the problem intractable. A key advantage of QUICKBLP is that its linear programming formulation naturally accommodates approximate linear programming (ALP), making it well suited for scaling to large state–action–delay spaces. Following [109], we first adopt a linear approximation of the occupation measure $x(\gamma, z, a)$ to reduce the number of variables:

$$\begin{aligned} x(\gamma, z, a) &\approx \hat{x}(\gamma, z, a; \boldsymbol{\theta}) \\ &\triangleq \mu_0(\gamma, z, a) + \sum_{i=1}^{d} \theta_i \psi_i(\gamma, z, a) \\ &= \mu_0(\gamma, z, a) + \boldsymbol{\theta}\boldsymbol{\psi}^T, \end{aligned} \tag{66}$$

where $\mu_0(\gamma, z, a)$ is a fixed baseline function, $\{\psi_i(\gamma, z, a)\}_{i=1}^d$ are predefined basis functions, and $\boldsymbol{\theta} \in \mathbb{R}^d$ is the parameter vector to be optimized. This reduces the number of variables from $n = |\mathcal{Z}| \times |\mathcal{S}| \times |\mathcal{Y}| \times |\mathcal{A}|^2$ to feature dimension d. Substituting this approximation into the LP Problem 7 yields the following ALP:

Problem 8 (*Linear Programming Approximation*)**.**

$$Q^\star(f_{\max}) \approx \min_{\boldsymbol{\theta}} \sum_{\gamma,z,a} q(\gamma, z, a)(\mu_0(\gamma, z, a) + \boldsymbol{\theta}\boldsymbol{\psi}^T) \tag{67a}$$

$$\text{s.t.} \quad \sum_{\gamma,z,a} f(z)\hat{x}(\gamma, z, a; \boldsymbol{\theta}) = \frac{1}{f_{\max}}, \tag{67b}$$

$$\sum_{z,a} \hat{x}(\gamma', z, a; \boldsymbol{\theta}) = \sum_{\gamma,z,a} p(\gamma'|\gamma, z, a)\hat{x}(\gamma, z, a; \boldsymbol{\theta}), \quad \forall \gamma' \in \mathcal{S} \times \mathcal{Y} \times \mathcal{A}, \tag{67c}$$

$$\sum_{\gamma,z,a} \hat{x}(\gamma, z, a; \boldsymbol{\theta}) = 1, \tag{67d}$$

$$\hat{x}(\gamma, z, a; \boldsymbol{\theta}) \geq 0, \quad \forall \gamma \in \mathcal{S} \times \mathcal{Y} \times \mathcal{A}, z \in \mathcal{Z}, a \in \mathcal{A}, \tag{67e}$$

The above ALP reduces the number of decision variables from $n = |\mathcal{Z}| \times |\mathcal{S}| \times |\mathcal{Y}| \times |\mathcal{A}|^2$ to a fixed feature dimension d, but the number of equality constraints still scales with the sizes of $\mathcal{S}$, $\mathcal{Y}$, and $\mathcal{A}$. This residual dependence on the system size limits the scalability of the approximation unless the constraints are further relaxed.

To achieve scalability, the hard constraints can be relaxed into soft penalties in the objective function [109]. Specifically, the objective function is augmented with weighted ℓ_1 violations of the constraints, yielding a surrogate loss. This reformulation converts the constrained LP into an *unconstrained stochastic convex optimization problem* over the low-dimensional parameter $\boldsymbol{\theta} \in \mathbb{R}^d$, which can be solved efficiently using *stochastic subgradient descent*. As a result, the overall complexity scales only with the feature dimension d, becoming independent of $\mathcal{S}$, $\mathcal{Y}$, and $\mathcal{A}$, and thus achieves true scalability.

D. Sensitivity Analysis and Sampling Frequency Threshold

To better understand the relationship between $h^\star$ and the maximum rate constraint parameter $f_{\max}$, we conduct a *sensitivity analysis* in this subsection to explore how $h^\star$ depends on the maximum rate constraint with parameter $f_{\max}$. The key result is stated in the following Lemma, whose proof can be found in Appendix O.

Theorem 9. *(Sensitivity Analysis). Define $f_{\max}^T$ as: $f_{\max}^T \triangleq \frac{1}{\mathcal{F}(\rho^\star)^-}$, the following assertions hold true:*
(i). If $f_{\max} \geq f_{\max}^T$, $h^\star$ is independent of $f_{\max}$;
(ii). If $f_{\max} < f_{\max}^T$, $h^\star = f_{\max} Q^\star(f_{\max})$ is monotonically non-increasing with $f_{\max}$, with the derivative given as: $\frac{\mathrm{d}h^\star}{\mathrm{d}f_{\max}} = U(\lambda^\star)$, where $\lambda^\star = \arg\max_{\lambda} \{\lambda + f_{\max} U(\lambda)\}$.

Furthermore, the condition that distinguishes the two cases in Lemma 9 directly establishes the following corollary, which characterizes the phenomenon where additional sampling does not contribute to further decision-making performance.

Corollary 2. *(Sampling frequency threshold) When the sampling frequency $f_{\max}$ exceeds or equals the threshold $f_{\max}^T$, i.e., $f_{\max} \geq f_{\max}^T$, any further increase in sampling frequency provides no improvement in decision performance.*

VIII. Simulation Results

A. Comparing Benchmark

1) Sampling Policies

In this paper, the sampling policy that is co-designed with the goal-oriented remote decision-making policy to minimize the *long-term average cost* in Problem 1 is referred to as "*Goal-oriented sampling*". We compare this sampling policy with the following benchmarks:

- *Uniform sampling*: The sampling is activated periodically with a constant sampling interval $d \in \mathbb{N}^+$. In this case, the sampling process follows $S_{i+1} - S_i = d$ for $\forall i$ (see [4, Section VI] for a detailed system description of uniform sampling with queuing and random service delay). Since the sampling interval d is limited to integer values in the discrete-time MDP setting, the corresponding sampling frequency $1/d$ also takes discrete values. Hence, the simulation yields a performance curve comprising discrete points, as shown in Fig. 10.

- *Zero-wait sampling*: An update is transmitted once the previous update is delivered, *i.e.*, $Z_i = 0$ for $\forall i$. This policy achieves the maximum throughput. The sampling frequency for *zero-wait* sampling is feasible only if $f_{\max} \geq 1/\mathbb{E}[Y_i]$.
- *Constant-wait sampling*: Waiting before transmitting is reported to be a good alternative for information *freshness* [110]. Here we consider $Z_i = z$, where $z \in \mathbb{N}^+$ is a constant waiting time. The sampling frequency for *constant-wait sampling* is $1/(z + \mathbb{E}[Y_i])$, which imposes a feasibility condition: $f_{\max} \geq 1/(z + \mathbb{E}[Y_i])$.
- *AoI-optimal sampling*: The *AoI-optimal* policy determines Z_i by [68, Theorem 4], which is a threshold-based policy $Z_i = \max(0, \beta - Y_i)$, where β is the solution to the following equations:

$$\mathbb{E}\left[Y + z(Y)\right] = \max\left(\frac{1}{f_{\max}}, \frac{\mathbb{E}[(Y + z(Y))^2]}{2\beta}\right) \quad (68)$$

with $z(Y) = \max(0, \beta - Y)$. [68, Algorithm 2] presents a low-complexity algorithm to solve β.

2) Remote Decision Making

We consider the following decision-making policies:

- *Co-design with goal-oriented sampling*: This decision-making policy is obtained by solving Problem 1. For conciseness, we refer to the co-design approach simply as "*goal-oriented sampling*".
- *Myopic decision policy* π_{myopic}: At each time step, the remote decision maker selects an action based on the most recent (possibly outdated) observation $X_{t-\Delta(t)}$. The selected action aims to minimize the instantaneous cost and is formally given by

$$a_t = \pi_{\text{myopic}}(X_{t-\Delta(t)}) = \arg\min_a \mathcal{C}(X_{t-\Delta(t)}, a), \quad (69)$$

where $\mathcal{C}(\cdot, \cdot)$ denotes the one-step cost function. This policy is termed *myopic* because it excludes any prediction or planning mechanisms; instead, decisions are made greedily to minimize the one-step cost based on the most recently available observation.
- *Long-term optimal decision policy* $\pi^\star$: The long-term optimal policy thinks *ahead* to minimize the long-term average cost over time. We compute it by solving the ACOE of the primal MDP using the RVI algorithm. The resulting policy employed at the remote decision maker $\pi^\star : \mathcal{S} \to \mathcal{A}$ selects an action based on the latest available observation $X_{t-\Delta(t)}$, according to $a_t = \pi^\star(X_{t-\Delta(t)})$. *Unless otherwise stated, this long-term optimal decision policy is employed as the default downstream decision-making policy under all benchmark sampling strategies. Throughout the paper, each baseline (e.g., "zero-wait", "AoI-optimal", "constant-wait") refers to a composite strategy that combines the corresponding sampling policy with this long-term optimal decision policy.*

B. Simulation Parameter Setup

1) Primal MDP

As a case study, we specifically consider a benchmark setup for clarity and insight. In this setup, the sizes of the state space and the action space of the *primal MDP* are both 2, consisting of states s_0, s_1 and actions a_0, a_1[13]. The *primal MDP* tuple is detailed in Appendix H, and is visualized by transition diagram given in Fig. 11.

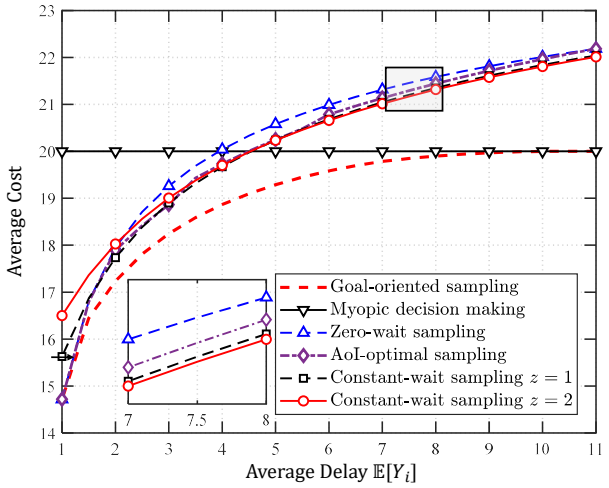


Fig. 7. Average cost vs. Average delay $\mathbb{E}[Y_i]$ under binary delay model ($Y_{\max} = 11$), with parameter p controlling the average delay. All baseline sampling policies are paired with the corresponding *long-term optimal decision policy*.

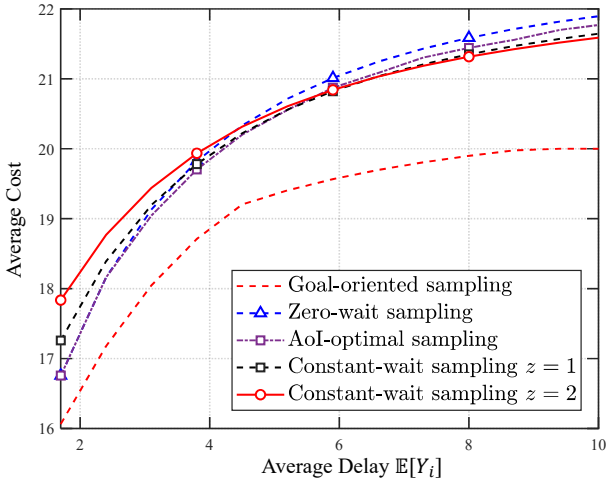


Fig. 8. Average cost vs. Average delay $\mathbb{E}[Y_i]$ under binary delay model ($p = 0.3$), with $Y_{\max}$ controlling the average delay. All baseline sampling policies are paired with the corresponding *long-term optimal decision policy*.

In the primal MDP, the corresponding *myopic* decision-making policy is

$$\pi_{\text{myopic}}(s_0) = a_0, \pi_{\text{myopic}}(s_1) = a_0, \quad (70)$$

indicating a constant action regardless of state. Consequently, this decision-making policy does not utilize state information at the remote decision maker, and thus the value of information transmission is *null* in this context.

[13]The binary state space is broadly representative of many real-world systems. For instance, it can model the **occurrence or absence** of critical events such as *fires, industrial failures, security breaches, or anomalies in equipment status*. The binary action space can be interpreted as whether or not to apply a control intervention in response to the critical event, e.g. extinguishing a detected fire or repairing a malfunctioning piece of equipment.

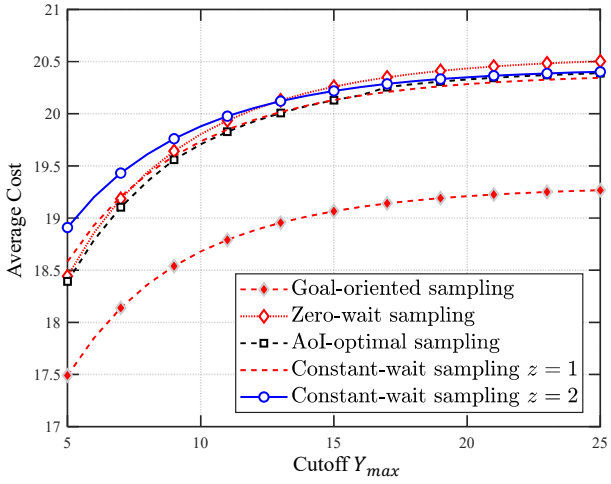


Fig. 9. Average cost vs. Cutoff $Y_{\max}$ under truncated geometric delay model with $q = 0.3$. All baseline sampling policies are paired with the corresponding *long-term optimal decision policy*.

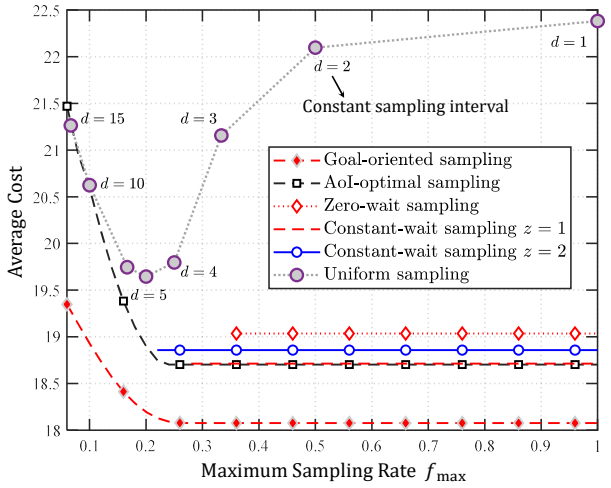


Fig. 10. Average cost vs. Maximum sampling frequency $f_{\max}$. All baseline sampling policies are paired with the corresponding *long-term optimal decision policy*.

The *long-term optimal decision* policy is solved by using RVI algorithm, and the solution is given as:

$$\pi^\star(s_0) = a_1, \pi^\star(s_1) = a_0, \tag{71}$$

which clearly depends on the system state. In this case, the remote decision maker must rely on the potentially *stale* state to select the "*right*" action, highlighting the utility of information transmission under the decision-making policy.

2) Finite-support Memory-less Delays

Throughout, we restrict attention to delay distributions with finite support. It is also practically justified, as many communication and control systems enforce a maximum admissible delay (via timeouts, buffer limits, or QoS deadlines). As a representative example, we consider a binary-valued random delay model where each delay Y_i takes value 1 with probability p and $Y_{\max}$ with probability $1-p$, i.e., $\Pr(Y_i = 1) = p, \Pr(Y_i = Y_{\max}) = 1-p, \forall i \in \mathbb{Z}^+$ (cf. [100]). This binary-delay formulation enables tractable analysis while capturing

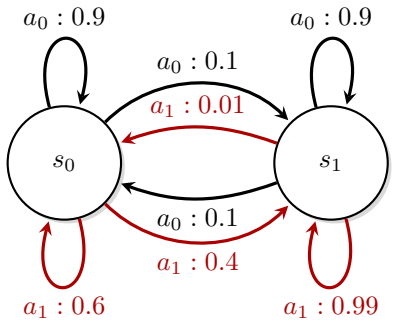


Fig. 11. Simulation Setup: Transition diagram of the primal MDP given in Appendix H.

TABLE III
RELATIVE COST REDUCTION (%) OF GOAL-ORIENTED SAMPLING COMPARED WITH BASELINES AT SELECTED AVERAGE DELAYS.

$\mathbb{E}[Y_i]$	Zero-wait	AoI-optimal	Const-wait $z = 2$
1.7	4.18%	4.18%	9.98%
5.9	6.23%	6.85%	6.09%
8.0	7.18%	7.83%	6.66%
14.3	10.11%	9.87%	8.76%

key aspects of stochastic delay.

Importantly, the proposed policy maps (X_{S_i}, Y_i, A_{i-1}) to actions, so the solution complexity scales with $|\mathcal{Y}|$; a finite-support assumption keeps the policy solution computationally manageable. The analysis extends to *any* finite-support distribution without memory, including truncated variants of heavy-tailed models[14]. Specifically, a geometric channel can be approximated by a *truncated geometric* delay with parameter $q \in (0,1)$ and cutoff $Y_{\max}$:

$$\Pr(Y = y) = \frac{q(1-q)^{y-1}}{1-(1-q)^{Y_{\max}}}, \quad y = 1, \ldots, Y_{\max}. \tag{72}$$

C. Discussions

Fig. 7 and Fig. 8 illustrate the performance comparisons between various sampling and decision-making policies under different average delays $\mathbb{E}[Y_i]$. In Fig. 7, the average delay is adjusted by adjusting p, with the maximum delay fixed at $Y_{\max} = 11$. In Fig. 8, the average delay is increased by enlarging $Y_{\max}$ while keeping $p = 0.3$ fixed. Across both scenarios, the proposed goal-oriented sampling strategy consistently outperforms all baseline methods in terms of minimizing the average cost. Table III reports the relative cost reduction $\eta_b = (C_b - \rho^\star)/C_b \times 100\%$ of the proposed goal-oriented sampling compared with representative baselines at several average delay levels $\mathbb{E}[Y_i]$, where C_b is the average cost of baseline b. The average delay is controlled by adjusting $Y_{\max}$. For small delays, goal-oriented sampling achieves about 4% reduction over zero-wait and AoI-optimal policies and nearly 10% over the constant-wait policy. As the delay

[14]Beyond truncation, handling *unbounded-support* delays (e.g., geometric channels) directly would require additional conditions to ensure average-cost optimality; a systematic treatment is an interesting direction for future work.

increases to moderate levels ($\mathbb{E}[Y_i] = 5.9$–8), the improvement grows to 6–8%, and at large delays ($\mathbb{E}[Y_i] = 14.3$) the gain further increases to about 10% against zero-wait and AoI-optimal. The performance gain primarily arises from the fact that the proposed sampling policy is explicitly designed to optimize decision effectiveness rather than communication metrics alone. Unlike conventional AoI-optimal sampling policy, which treats information freshness and control decisions as separate optimization problems, goal-oriented sampling allocates sampling opportunities more judiciously, prioritizing updates that are most relevant to improving decision outcomes. In contrast, AoI-optimal sampling may transmit timely but less informative data, resulting in suboptimal decision performance despite lower AoI.

A key observation from Fig. 7 is that the *myopic decision-making policy*, which selects a fixed action regardless of state information, yields a constant cost independent of delay. This myopic policy serves as a delay-agnostic baseline in the sense that it does not explicitly condition its decision rule on AoI beyond the last received state; the gap to the proposed goal-oriented sampling quantifies the benefit of explicitly leveraging AoI in decision-making. This highlights an insight: in scenarios where the control policy is insensitive to informed state information, enhancing the communication channel (e.g., by reducing delay) may not lead to any improved goal-oriented decision-making performance. Moreover, Fig. 7 demonstrates that the performance gap between the proposed *goal-oriented sampling* policy and the *myopic decision-making* policy gradually narrows as the average delay $\mathbb{E}[Y_i]$ increases. This convergence reflects a fundamental limitation in decision-making under communication constraints: when the delay becomes excessively large, the received state information becomes so outdated that it no longer provides reliable guidance for current decisions. As a result, the optimal goal-oriented policy degenerates into state-independent behavior in high-delay regimes, where the influence of timely and fresh information on decision performance becomes negligible.

Fig. 9 illustrates the average cost achieved by various sampling strategies under a truncated geometric delay model with parameter $q = 0.3$. The x-axis represents the cutoff parameter $Y_{\max}$, which limits the maximum possible delivery delay. As $Y_{\max}$ increases, the truncated delay distribution gradually approaches the standard (non-truncated) geometric distribution, and the resulting average cost curves converge accordingly. Across the entire range of $Y_{\max}$, the proposed goal-oriented sampling strategy consistently outperforms all baselines, achieving significantly lower average cost—especially under large delay cutoffs.

Fig. 10 illustrates the relationship between average cost and the maximum sampling frequency $f_{\max}$. The proposed goal-oriented sampling consistently achieves the lowest cost across all values of $f_{\max}$. As $f_{\max}$ increases, the performance of both *goal-oriented sampling* and *AoI-optimal* sampling improves gradually. Notably, the *goal-oriented sampling* curve aligns with the *sensitivity analysis* presented in Theorem 9, which predicts that the average cost initially decreases with $f_{\max}$, then saturates to a constant value.

In contrast, the uniform sampling policy in Fig. 10 exhibits a *U-shaped* cost curve. As $f_{\max}$ decreases, its performance converges to that of AoI-optimal sampling policy. This convergence highlights a key insight: under sparse sampling constraints, the transmitted information becomes too outdated to support effective state-dependent decision-making, resulting in uniformly poor performance. Conversely, when the sampling interval $d = 1/f_{\max}$ becomes small, the system begins to accumulate a *queuing backlog*, resulting in increased delivery delays. This queuing-induced staleness reduces the timeliness of information at the decision-maker, thereby degrading goal-oriented decision-making performance.

IX. Conclusion and Future Directions

In this paper, we have proposed a new remote MDP problem, namely AR-MDP in the time-lag MDP framework. Specifically, AoI, typically an optimization indicator for information *freshness*, is incorporated into AR-MDP both as a controllable random processing delay and as critical side information to support remote decision-making. To investigate the fundamental trade-off between communication constraint and the decision-making effectiveness, we considered both *sampling frequency constraint* and *random delay* in this problem and developed low-complexity *one-layer* algorithms, OnePDSI and QuickBLP to solve this problem efficiently. Through theoretical analysis and experiments, we reveal how communication-induced information *staleness* can negatively impact remote decision-making performance.

Future directions. Several promising directions remain open for future exploration. First, while this work focuses on finite-state, finite-action, and finite bounded-delay systems, extending the framework to handle *continuous state and action spaces*, as well as delay distributions with infinite support, remains a challenging yet important direction. In particular, finiteness underlies several key steps of the present analysis, including the existence of stationary average-cost optimal policies and the associated optimality arguments on the lifted MDP, the finite-dimensional occupation-measure LP formulation in the constrained case, and the convergence and error guarantees of the proposed algorithms. Extending to infinite spaces will therefore require replacing these finite-dimensional tools, for example via discretization/quantization, or by invoking average-cost MDP/CMDP theory on Borel spaces under additional regularity conditions. Second, the current model assumes that a sample, once sent, cannot be interrupted. Enabling *preemptive or adaptive transmission mechanisms* may enhance the system responsiveness. Third, relaxing the action-holding assumption and incorporating *AoI-triggered decision-making mechanisms*, where sampling or control actions are explicitly adapted based on real-time AoI values, may allow for more efficient utilization of limited communication resources. Moreover, while the proposed algorithms offer a favorable balance between optimality and complexity under the considered setting, future work could explore *scalable reinforcement learning techniques* that generalize to unknown

systems.

APPENDIX A
PROOF OF LEMMA 1

A. Proof of Sufficient Statistics

We consider the finite-horizon value function

$$\begin{aligned} J_t(\mathcal{I}_t) &\triangleq \max_{a_{t:T}} \mathbb{E}^{a_{t:T}}\left[\sum_{k=t}^{T} \mathcal{C}(X_k, a_k) | \mathcal{I}_t\right] \\ &= \max_{a_{t:T}} \sum_{k=t}^{T} \sum_{s' \in \mathcal{S}} \mathcal{C}(s', a_k) \cdot \overset{a_{t:T}}{\Pr}(X_k = s' \mid \mathcal{I}_t), \end{aligned} \tag{73}$$

where $\Pr^{a_{t:T}}$ (and $\mathbb{E}^{a_{t:T}}$) denotes the probability measure (and expectation) *induced by* the chosen action sequence $a_{t:T}$. Because X_k is a controlled Markov process, the filtering distribution at time k given $\mathcal{I}_t$ depends on $\left(X_{t-\Delta(t)}, \Delta(t), a_{t-\Delta(t):t-1}\right)$ and not on earlier history, hence

$$\begin{aligned} &\overset{a_{t:T}}{\Pr}\left(X_k \in \cdot \,\middle|\, \mathcal{I}_t\right) \\ &= \overset{a_{t:T}}{\Pr}\left(X_k \in \cdot \,\middle|\, X_{t-\Delta(t)}, \Delta(t), a_{t-\Delta(t):t-1}\right). \end{aligned} \tag{74}$$

For $t \in [D_i, D_{i+1})$ we have $t - \Delta(t) = S_i$ by (1) and $Y_i = D_i - S_i$, so

$$a_{S_i:t-1} = \left(a_{S_i:D_i-1},\, a_{D_i:t-1}\right) = (A_{i-1},\, a_{D_i:t-1}), \tag{75}$$

and

$$\overset{a_{t:T}}{\Pr}\left(X_k \in \cdot \,\middle|\, \mathcal{I}_t\right) = \overset{a_{t:T}}{\Pr}\left(X_k \in \cdot \mid X_{S_i}, Y_i, A_{i-1}, a_{D_i:t-1}\right). \tag{76}$$

As assumed by (7), the control is held constant on $[D_i, D_{i+1})$,

$$a_{D_i:t-1} = (a_t, \ldots, a_t), \text{for } t \in [D_i, D_{i+1}), \tag{77}$$

so the term $a_{D_i:t-1}$ is completely determined by a_t. We can therefore *absorb* its effect into the induced measure and establish,

$$\begin{aligned} J_t(\mathcal{I}_t) &= \max_{a_{t:T}} \sum_{k=t}^{T} \sum_{s' \in \mathcal{S}} \mathcal{C}(s', a_k)\ \overset{a_{t:T}}{\Pr}\left(X_k = s' \,\middle|\, \mathcal{G}_i\right) \\ &= \max_{a_{t:T}}\ \mathbb{E}^{a_{t:T}}\left[\sum_{k=t}^{T} \mathcal{C}(X_k, a_k) \,\middle|\, \mathcal{G}_i\right] \ =\ J_t(\mathcal{G}_i). \end{aligned} \tag{78}$$

From Definition 1 and we know that $\mathcal{G}_i = (X_{S_i}, Y_i, A_{i-1})$ is a sufficient statistics of $\mathcal{I}_t$ for $t \in [D_i, D_{i+1})$.

B. Determining u_t is equivalent to Determining Z_i

By (3) and (5), within $[D_i, D_{i+1})$ the sampling control u_t determines the next sampling epoch via

$$S_{i+1} \ =\ \inf\{\tau \geq D_i : a_\tau^S = 1\} \ =\ D_i + Z_i, \tag{79}$$

with $Z_i \in \mathbb{N}$ the waiting time. Conversely, any choice of S_{i+1} (equivalently Z_i) induces a unique sequence $\{a_\tau^S\}_{\tau \in [D_i, D_{i+1})}$ by setting $a_\tau^S = 0$ for $\tau \in [D_i, S_{i+1})$ and $a_{S_{i+1}}^S = 1$ (and then proceeding to the next interval). Thus the mapping between $\{a_\tau^S\}_{\tau \in [D_i, D_{i+1})}$ and S_{i+1} (or Z_i) is one-to-one, and optimizing over one is equivalent to optimizing over the other.

APPENDIX B
PROOF OF LEMMA 2

A. Proof of Part (i)

1) $\rho^\star \leq \lambda \iff U(\lambda) \leq 0$

If $\rho^\star \leq \lambda$, there exists a policy $\pi = (Z_0, A_0, Z_1, A_1, \cdots)$ such that

$$\lim_{\mathrm{n}\to\infty} \frac{\sum_{i=0}^{n-1} \mathbb{E}_\pi\left[\sum_{t=D_i}^{D_{i+1}-1} \mathcal{C}(X_t, A_i)\right]}{\sum_{i=0}^{n-1} \mathbb{E}_\pi\left[Y_{i+1} + Z_i\right]} \leq \lambda, \tag{80}$$

which is equivalent to

$$\lim_{\mathrm{n}\to\infty} \frac{\frac{1}{n}\sum_{i=0}^{n-1} \mathbb{E}_\pi\left[\sum_{t=D_i}^{D_{i+1}-1} \mathcal{C}(X_t, A_i)\right] - \lambda \mathbb{E}_\pi\left[Y_{i+1} + Z_i\right]}{\frac{1}{n}\sum_{i=0}^{n-1} \mathbb{E}_\pi\left[Y_{i+1} + Z_i\right]} \leq 0. \tag{81}$$

Since $Y_i > 0$ and $0 \leq Z_i < \infty$, we have that $\lim_{\mathrm{n}\to\infty} \frac{1}{n}\sum_{i=0}^{n-1} \mathbb{E}_\pi\left[Y_{i+1} + Z_i\right]$ always exists, satisfying

$$0 < \lim_{\mathrm{n}\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \mathbb{E}_\pi\left[Y_{i+1} + Z_i\right] < \infty. \tag{82}$$

Thus, we have that the numerator of (81) satisfies

$$\lim_{\mathrm{n}\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \mathbb{E}_\pi\left[\sum_{t=D_i}^{D_{i+1}-1} \mathcal{C}(X_t, A_i)\right] - \lambda \mathbb{E}_\pi[Z_i + Y_{i+1}] \leq 0. \tag{83}$$

This implies that the infimum of the left hand side of (83) is also at most 0, *i.e*, $U(\lambda) \leq 0$.

On the contrary, when $U(\lambda) \leq 0$, we can know that there exists a policy $\pi = (Z_0, A_0, Z_1, A_1, \cdots)$ that satisfies (83). As (82) always holds, we can easily obtain that (80) holds. Note that $\rho^\star$ is the infimum of the left hand side of (80), we have

$$\rho^\star \leq \lim_{\mathrm{n}\to\infty} \frac{\sum_{i=0}^{n-1} \mathbb{E}_\pi\left[\sum_{t=D_i}^{D_{i+1}-1} \mathcal{C}(X_t, a_i)\right]}{\sum_{i=0}^{n-1} \mathbb{E}_\pi\left[Y_{i+1} + Z_i\right]} \leq \lambda. \tag{84}$$

2) $\rho^\star > \lambda \iff U(\lambda) > 0$

If $\rho^\star > \lambda$, we have that for any policy $(Z_0, A_0, Z_1, A_1, \cdots)$, the following inequality always holds

$$\lim_{\mathrm{n}\to\infty} \frac{\sum_{i=0}^{n-1} \mathbb{E}\left[\sum_{t=D_i}^{D_{i+1}-1} \mathcal{C}(X_t, A_i)\right]}{\sum_{i=0}^{n-1} \mathbb{E}\left[Y_{i+1} + Z_i\right]} > \lambda. \tag{85}$$

Since (82) holds, we have that for any policy π,

$$\lim_{\mathrm{n}\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \mathbb{E}_\pi\left[\sum_{t=D_i}^{D_{i+1}-1} \mathcal{C}(X_t, A_i)\right] - \lambda \mathbb{E}_\pi[Z_i + Y_{i+1}] > 0. \tag{86}$$

Since (86) holds for any policies, it follows that the infimum value of the left-hand side (LHS) of (86) is also greater than 0, implying that $\rho^\star > \lambda$.

When $U(\lambda) > 0$, it is established that condition (86) is satisfied for any policy sequence $(Z_0, A_0, Z_1, A_1, \cdots)$. Given that (82) always holds, it follows directly that (85) holds for any policies, implying that the infimum of the LHS of (85) is also greater than λ, *i.e.*, $\rho^\star > \lambda$.

B. Proof of Part (ii)

By Part (i), $U(\lambda) = 0$ iff $\lambda = \rho^\star$. Let $\lambda = \rho^\star$. If $\pi^\star$ is optimal for Problem 4, then

$$\lim_{\mathrm{n}\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \mathbb{E}\left[\sum_{t=D_i}^{D_{i+1}-1} \mathcal{C}(X_t, A_i)\right] - \lambda(Z_i + Y_{i+1}) = 0, \quad (87)$$

which implies that

$$\lim_{\mathrm{n}\to\infty} \frac{\sum_{i=0}^{n-1} \mathbb{E}_{\pi^*}\left[\sum_{t=D_i}^{D_{i+1}-1} \mathcal{C}(X_t, A_i)\right]}{\sum_{i=0}^{n-1} \mathbb{E}_{\pi^*}\left[Y_{i+1} + Z_i\right]} = \lambda. \quad (88)$$

Note that $\rho^\star = \lambda$, we have that for the policy π^*,

$$\rho^\star = \lim_{\mathrm{n}\to\infty} \frac{\sum_{i=0}^{n-1} \mathbb{E}\left[\sum_{t=D_i}^{D_{i+1}-1} \mathcal{C}(X_t, A_i)\right]}{\sum_{i=0}^{n-1} \mathbb{E}\left[Y_{i+1} + Z_i\right]}, \quad (89)$$

which infers that policy π^* is also the optimal policy of Problem 3.

C. Proof of Part (iii)

From Part (i), we know that proving Part (iii) is equivalent to prove that $U(\lambda)$ is monotonically non-increasing in terms of λ, *i.e.*, for any $\Delta\lambda > 0$, $U(\lambda+\Delta\lambda) \le U(\lambda)$. This is verified by the following inequalities:

$$\begin{aligned}
&U(\lambda + \Delta\lambda) = \\
&\inf_{\phi_{0:\infty}} \lim_{\mathrm{n}\to\infty} \frac{1}{n}\sum_{i=0}^{n-1} \left\{ \mathbb{E}\left[\sum_{t=D_i}^{D_{i+1}-1} \mathcal{C}(X_t, A_i)\right] - (\lambda + \Delta\lambda)\mathbb{E}\left[Z_i + Y_{i+1}\right] \right\} \\
&= \inf_{\phi_{0:\infty}} \left\{ \lim_{\mathrm{n}\to\infty} \frac{1}{n}\sum_{i=0}^{n-1} \left\{ \mathbb{E}\left[\sum_{t=D_i}^{D_{i+1}-1} \mathcal{C}(X_t, A_i)\right] - \lambda\mathbb{E}\left[Z_i + Y_{i+1}\right] \right\} \right. \\
&\left. - \lim_{\mathrm{n}\to\infty} \frac{1}{n}\sum_{i=0}^{n-1} \Delta\lambda\mathbb{E}[Z_i + Y_{i+1}] \right\} \\
&\le \inf_{\phi_{0:\infty}} \lim_{\mathrm{n}\to\infty} \frac{1}{n}\sum_{i=0}^{n-1} \left\{ \mathbb{E}\left[\sum_{t=D_i}^{D_{i+1}-1} \mathcal{C}(X_t, A_i)\right] - \lambda\mathbb{E}\left[Z_i + Y_{i+1}\right] \right\} \\
&= U(\lambda).
\end{aligned} \quad (90)$$

Thus, we have that $\lambda = \rho^\star$ is the unique root of $U(\lambda) = 0$.

Appendix C
Transition Probability of $\mathscr{P}_{\mathrm{MDP}}(\lambda)$

Recall that $\mathcal{G}_i = (X_{S_i}, Y_i, A_{i-1})$ and $\mathcal{G}_{i+1} = (X_{S_{i+1}}, Y_{i+1}, A_i)$. We assume $\{Y_i\}$ are i.i.d. and independent of the source process $\{X_t\}$ and of the actions. We derive $\Pr(\mathcal{G}_{i+1} \mid \mathcal{G}_i, Z_i, A_i)$ by computing the three marginals $\Pr(X_{S_{i+1}} \mid \mathcal{G}_i, Z_i, A_i)$, $\Pr(Y_{i+1} \mid \mathcal{G}_i, Z_i, A_i)$, and $\Pr(A_i \mid \mathcal{G}_i, Z_i, A_i)$.

A. $\Pr\left(X_{S_{i+1}} | \mathcal{G}_i, Z_i, A_i\right)$

Condition on $\mathcal{G}_i = (s, \delta, a)$, where $\delta = Y_i$ and $a = A_{i-1}$. From $t = S_i$ to $t = D_i - 1$ the action is constant and equal to $A_{i-1} = a$, so the source evolves for δ steps under transition matrix $\mathbf{P}_a$, yielding a δ-step kernel $\mathbf{P}_a^\delta$. From $t = D_i$ to $t = S_{i+1} - 1$ the action is A_i, so the source further evolves for Z_i steps under $\mathbf{P}_{A_i}$, i.e., kernel $\mathbf{P}_{A_i}^{Z_i}$ (with the convention $\mathbf{P}_{A_i}^0 = I$). By the semigroup property, we have

$$\Pr\big(X_{S_{i+1}} = s' \mid \mathcal{G}_i = (s, \delta, a), Z_i, A_i\big) = \left[\mathbf{P}_a^\delta \mathbf{P}_{A_i}^{Z_i}\right]_{s \times s'}. \quad (91)$$

B. $\Pr(Y_{i+1} | \mathcal{G}_i, Z_i, A_i)$

Since $\{Y_i\}$ are i.i.d. and independent of $(\mathcal{G}_i, Z_i, A_i)$,

$$\Pr(Y_{i+1} = \delta' \mid \mathcal{G}_i = (s, \delta, a), Z_i, A_i) = \Pr(Y_{i+1} = \delta'). \quad (92)$$

C. $\Pr(A_i | \mathcal{G}_i, Z_i, A_i)$

The third component of $\mathcal{G}_{i+1}$ is the record of the action taken during $[D_i, S_{i+1})$, namely A_i. Hence,

$$\Pr\left(A_i = a' | \mathcal{G}_i = (s, \delta, a), Z_i, A_i\right) = \mathbb{1}\{a' = A_i\}. \quad (93)$$

D. Product form and the transition kernel

By the stated independence, Y_{i+1} is independent of $X_{S_{i+1}}$ given $(\mathcal{G}_i, Z_i, A_i)$, and the value of A_i is deterministic under the conditioning. Therefore,

$$\begin{aligned}
&\Pr(\mathcal{G}_{i+1} = (s', \delta', a') \mid \mathcal{G}_i = (s, \delta, a), Z_i, A_i) \\
&\quad = \Pr(Y_{i+1} = \delta') \cdot \left[\mathbf{P}_a^\delta \mathbf{P}_{A_i}^{Z_i}\right]_{s \times s'} \cdot \mathbb{1}\{a' = A_i\},
\end{aligned} \quad (94)$$

which is exactly the transition probability stated in (18).

Appendix D
Proof of Lemma 3

Recall that $\mathcal{G}_i = (X_{S_i}, Y_i, A_{i-1})$, $Z_i = S_{i+1} - D_i$, $D_{i+1} = S_{i+1} + Y_{i+1}$, and $\{Y_i\}$ are i.i.d. and independent of the source process $\{X_t\}$ and of the actions. With the per-epoch cost in (19), we have

$$\begin{aligned}
&\inf_{\phi_{0:\infty}} \limsup_{n\to\infty} \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} g(\mathcal{G}_i, A_i, Z_i; \lambda)\right] \\
&= \inf_{\phi_{0:\infty}} \limsup_{T\to\infty} \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} q(\mathcal{G}_i, A_i, Z_i) - \lambda\mathbb{E}[Z_i + Y_{i+1}]\right].
\end{aligned} \quad (95)$$

Hence it suffices to show

$$\mathbb{E}\left[\sum_{t=D_i}^{D_{i+1}-1} \mathcal{C}(X_t, A_i)\right] = \mathbb{E}[q(\mathcal{G}_i, Z_i, A_i)]. \quad (96)$$

For each epoch i, the expectation $\mathbb{E}\left[\sum_{t=D_i}^{D_{i+1}-1} \mathcal{C}(X_t, A_i)\right]$ can be decomposed as (97). From $t = S_i$ to $t = D_i - 1$ the action is $A_{i-1} = a$, i.e., δ steps under transition matrix $\mathbf{P}_a$, yielding $\mathbf{P}_a^\delta$. From $t = D_i$ to any $t \in \{D_i, \ldots, D_{i+1} - 1\}$, the action is A_i, i.e., $(t - D_i)$ further steps under $\mathbf{P}_{A_i}$, yielding $\mathbf{P}_{A_i}^{t-D_i}$ (with $\mathbf{P}_u^0 = I$). By the semigroup property,

$$\Pr(X_t = s' \mid \mathcal{G}_i = (s, \delta, a), Z_i, A_i) = \left[\mathbf{P}_a^\delta \mathbf{P}_{A_i}^{t-D_i}\right]_{s \times s'}. \quad (98)$$

$$\mathbb{E}\left[\sum_{t=D_i}^{D_{i+1}-1}\mathcal{C}(X_t,A_i)\right]=\mathbb{E}\left[\mathbb{E}\left[\sum_{t=D_i}^{D_i+Z_i+Y_{i+1}-1}\mathcal{C}(X_t,A_i)\,\middle|\,\mathcal{G}_i,Z_i,A_i\right]\right]$$
$$=\mathbb{E}\left[\mathbb{E}_{Y_{i+1}}\left[\sum_{t=D_i}^{D_i+Z_i+Y_{i+1}-1}\sum_{s'\in\mathcal{S}}\mathcal{C}(s',A_i)\,\Pr(X_t=s'\mid\mathcal{G}_i,Z_i,A_i)\right]\right]. \tag{97}$$

Substitute (98) into (97) and reindex $\tau=t-D_i\in\{0,\dots,Z_i+Y_{i+1}-1\}$, then exchange the sums:

$$\mathbb{E}\left[\sum_{t=D_i}^{D_{i+1}-1}\mathcal{C}(X_t,A_i)\right]$$
$$=\mathbb{E}\left[\sum_{s'\in\mathcal{S}}\mathcal{C}(s',A_i)\,\mathbb{E}_{Y_{i+1}}\left[\sum_{\tau=0}^{Z_i+Y_{i+1}-1}\left[\mathbf{P}_{A_{i-1}}^{Y_i}\mathbf{P}_{A_i}^{\tau}\right]_{X_{S_i}\times s'}\right]\right]. \tag{99}$$

By the definition in (20), the right-hand side equals $\mathbb{E}[q(\mathcal{G}_i,Z_i,A_i)]$, which proves (96). Therefore, we establish:

$$\limsup_{n\to\infty}\frac{1}{n}\mathbb{E}\left[\sum_{i=0}^{n-1}\big(q(\mathcal{G}_i,Z_i,A_i)-\lambda f(Z_i)\big)\right]=$$
$$\limsup_{n\to\infty}\frac{1}{n}\mathbb{E}\left[\sum_{i=0}^{n-1}\left(\sum_{t=D_i}^{D_{i+1}-1}\mathcal{C}(X_t,A_i)-\lambda\left(Z_i+\mathbb{E}[Y_{i+1}]\right)\right)\right], \tag{100}$$

which is exactly the objective of Problem 4. Hence $\mathscr{P}_{\mathrm{MDP}}(\lambda)$ is equivalent to Problem 4.

Appendix E
Proof of Theorem 1

Fix an arbitrary stationary deterministic policy $\pi:\mathcal{S}\times\mathcal{Y}\times\mathcal{A}\to\mathcal{A}\times\mathcal{Z}$ and an arbitrary initial lifted state $\mathcal{G}_0=(s_0,\delta_0,a_0)\in\mathcal{S}\times\mathcal{Y}\times\mathcal{A}$. Denote by K^π the one-step transition kernel of the lifted chain under π, and by $(K^\pi)^m$ its m-step kernel. Define the *common small set*

$$\mathcal{V}\triangleq\{s^\star\}\times\mathcal{Y}\times\mathcal{A}\subseteq\mathsf{S}\times\mathcal{Y}\times\mathcal{A}.$$

We claim that there is a policy-independent one-block minorization:

$$[(K^\pi)^m]_{(s,\delta,a)\times\mathcal{V}}\ \geq\ \epsilon\qquad\forall(s,\delta,a)\in\mathcal{S}\times\mathcal{Y}\times\mathcal{A}. \tag{101}$$

To see this, fix (s,δ,a) and unfold m steps. Along any admissible length-m sequence $\{(A_t,Z_t,\delta_t)\}_{t=0}^{m-1}$ generated by the policy π with linkage constraints $a_{t+1}=A_t$, the s-component transition from s to $s^\star$ equals the matrix product on the left of (22). By the hypothesis (22), this probability is at least ϵ, uniformly over all such sequences. Since $\mathcal{C}$ does not restrict the (δ,a) components at time m, we obtain

$$[(K^\pi)^m]_{(s,\delta,a)\times\mathcal{V}}\geq\epsilon>0, \tag{102}$$

which proves (101).

Now apply the Markov property on block times $\{0,m,2m,\dots\}$. For every $n\in\mathbb{N}$, and the policy π

$$\Pr\Big\{\mathcal{G}_{km}\notin\mathcal{C}\text{ for }k=1,\dots,n\ \Big|\ \mathcal{G}_0=(s,\delta,a)\Big\}\ \leq\ (1-\epsilon)^n. \tag{103}$$

Letting $n\to\infty$ shows that the hitting time $\tau_{\mathcal{C}}\triangleq\inf\{k\geq1:\mathcal{G}_{km}\in\mathcal{C}\}$ is almost surely finite. Restarting from any $y\in\mathcal{C}$ and repeating the same argument yields that the chain visits $\mathcal{C}$ infinitely often with probability one, uniformly over the initial state and the policy π.

We now establish uniqueness of the recurrent class. Suppose, by contradiction, that there exist two disjoint recurrent classes $\mathcal{R}_1,\mathcal{R}_2\subseteq\mathsf{S}\times\mathcal{Y}\times\mathcal{A}$ under π. Recurrent classes are closed, and starting from any $x\in\mathcal{R}_j$ the chain returns to $\mathcal{R}_j$ infinitely often almost surely. But from any starting point the chain also visits $\mathcal{C}$ infinitely often almost surely, hence $\mathcal{C}\subseteq\mathcal{R}_j$ for $j=1,2$, which implies $\mathcal{R}_1\cap\mathcal{R}_2\supseteq\mathcal{C}\neq\varnothing$, which is a contradiction. Therefore, there is at most one recurrent class. Since the state space is finite, at least one recurrent class exists, and uniqueness follows. This proves that the lifted MDP is unichain in the sense stated.

Appendix F
Proof of Lemma 5

A. Proof of $\rho^\star\geq\min_{s,a}\mathcal{C}(s,a)$

For all t and any policy, $\mathcal{C}(X_t,a_t)\geq\min_{s,a}\mathcal{C}(s,a)$ almost surely. Hence

$$\rho^\star=\inf_{\phi_{0:\infty}}\limsup_{T\to\infty}\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\mathcal{C}(X_t,a_t)\right]$$
$$\geq\inf_{\phi_{0:\infty}}\limsup_{T\to\infty}\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\min_{s,a}\mathcal{C}(s,a)\right]=\min_{s,a}\mathcal{C}(s,a). \tag{104}$$

B. Proof of $\rho^\star\leq\min_a\sum_{s\in\mathcal{S}}\pi_a(s)\cdot\mathcal{C}(s,a)$

Restrict to the subclass of policies that fix the action to a constant $a\in\mathcal{A}$. Then $\{X_t\}$ evolves as a time-homogeneous Markov chain with transition matrix $\mathbf{P}_a$. Assume that $\mathbf{P}_a$ admits a stationary distribution π_a and that the chain is ergodic so that time averages converge to stationary expectations. By the Markov chain ergodic theorem,

$$\lim_{T\to\infty}\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\mathcal{C}(X_t,a)\right]=\sum_{s\in\mathcal{S}}\pi_a(s)\,\mathcal{C}(s,a). \tag{105}$$

Therefore, for each a,

$$\rho^\star \leq \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\mathcal{C}(X_t,a)\right] = \sum_s \pi_a(s)\,\mathcal{C}(s,a), \tag{106}$$

and minimizing over a yields the stated upper bound.

Appendix G
Proof of Theorem 4

From [93, Proposition 7.4.1], we know that for any λ, the optimal value of Problem 4, which is $U(\lambda)$, is the same for all initial states and some values $V^*(\gamma;\lambda), \gamma \in \mathcal{S}\times\mathcal{Y}\times\mathcal{A}$ and satisfies the following Bellman equation:

$$V^*(\gamma;\lambda) + U(\lambda) = \min_{A_i,Z_i}\left\{g(\gamma,A_i,Z_i;\lambda) + \mathbb{E}[V^*(\gamma';\lambda)\,|\gamma,A_i,Z_i]\right\}, \tag{107}$$

Substituting $\lambda=\rho^\star$ and $U(\rho^\star)=0$ into the Bellman equation,

$$V^*(\gamma;\rho^\star) = \min_{A_i,Z_i}\left\{g(\gamma,A_i,Z_i;\rho^\star) + \mathbb{E}[V^*(\gamma';\rho^\star)\,|\gamma,A_i,Z_i]\right\}, \tag{108}$$

Similar to the RVI algorithm, we introduce the *relative value function* defined as

$$W^*(\gamma) \triangleq V^*(\gamma;\rho^\star) - V^*(\gamma^{\text{ref}};\rho^\star), \tag{109}$$

where γ^{ref} is called *reference state* and can be arbitrarily chosen from space $\mathcal{S}\times\mathcal{Y}\times\mathcal{A}$. Then, substituting (109) into (108) yields

$$W^*(\gamma) = \min_{A_i,Z_i}\left\{g(\gamma,A_i,Z_i;\rho^\star) + \mathbb{E}[W^*(\gamma')\,|\gamma,Z_i,A_i]\right\}, \tag{110}$$

Applying $\gamma=\gamma^{\text{ref}}$ in (109) and (110) leads to

$$W^*(\gamma^{\text{ref}}) = 0, \forall\gamma\in\mathcal{S}\times\mathcal{Y}\times\mathcal{A}, \tag{111}$$

and

$$W^*(\gamma^{\text{ref}}) = \min_{A_i,Z_i}\left\{g(\gamma^{\text{ref}},A_i,Z_i;\rho^\star) + \mathbb{E}[W^*(\gamma')\,|\gamma^{\text{ref}},Z_i,A_i]\right\}, \tag{112}$$

Then, substituting (111) and $g(\gamma^{\text{ref}},A_i,Z_i;\rho^\star) = q(\gamma^{\text{ref}},A_i,Z_i) - \rho^\star\cdot f(Z_i)$ into (112) yields (113).

Because $f(Z_i)>0$, we have that (113) holds only if

$$\min_{A_i,Z_i}\left\{\frac{q(\gamma^{\text{ref}},A_i,Z_i)+\mathbb{E}[W^*(\gamma')\,|\gamma^{\text{ref}},Z_i,A_i]}{f(Z_i)} - \rho^\star\right\} = 0. \tag{114}$$

Moving $\rho^\star$ to the RHS of (114) yields

$$\rho^\star = \min_{A_i,Z_i}\left\{\frac{q(\gamma^{\text{ref}},A_i,Z_i)+\mathbb{E}[W^*(\gamma')\,|\gamma^{\text{ref}},Z_i,A_i]}{f(Z_i)}\right\}. \tag{115}$$

We thus accomplish the proof.

Appendix H
The Primal MDP

We consider the following parameter setup:

- The *state space* is a binary space: $\mathcal{S}=\{s_0,s_1\}$.
- The *action space* is a binary space: $\mathcal{A}=\{a_0,a_1\}$.
- The *transition probability matrix* of X_t is

$$\mathbf{P}_{a_0} = \begin{bmatrix}0.9 & 0.1\\ 0.1 & 0.9\end{bmatrix}, \mathbf{P}_{a_1} = \begin{bmatrix}0.6 & 0.4\\ 0.01 & 0.99\end{bmatrix}. \tag{116}$$

- The cost function $\mathcal{C}(X_t,a_t)$ is given as

$$\begin{aligned}&\mathcal{C}(s_0,a_0)=40, \mathcal{C}(s_0,a_1)=60,\\ &\mathcal{C}(s_1,a_0)=0, \mathcal{C}(s_1,a_1)=20.\end{aligned} \tag{117}$$

Appendix I
Proof of Theorem 2 and Theorem 3

The proof is divided into two parts. First, we prove in Section I-A that the limits $\lim_{K\to\infty}\tilde{U}_K(\lambda)$ and $\lim_{K\to\infty}\tilde{V}_K(\gamma;\lambda), \forall\gamma\in\mathcal{S}\times\mathcal{Y}\times\mathcal{A}$ are both finite; Second, we explicitly establish that $\lim_{K\to\infty}\tilde{U}_K(\lambda) = U(\lambda)$ and $\lim_{K\to\infty}\tilde{V}_K(\gamma;\lambda) = V^\star(\gamma;\lambda)/\tau$ in Section I-B.

A. Convergence of (28)

Denote $Z^{(K)}(\gamma)$, $A^{(K)}(\gamma)$ as the waiting time and controlled action that achieves the minimum in the K-th relation:

$$(A^{(K)}(\gamma),Z^{(K)}(\gamma)) = \arg\min_{A_i,Z_i}\left\{g(\gamma,Z_i,A_i;\lambda) + \tau\mathbb{E}[\tilde{V}_K(\gamma';\lambda)\,|\gamma,Z_i,A_i]\right\}. \tag{118}$$

Define $\widetilde{\mathbf{V}}_K(\lambda)\in\mathbb{R}^{|\mathcal{S}\times\mathcal{Y}\times\mathcal{A}|}$ as the column vector formed by stacking the values $\tilde{V}_K(\gamma;\lambda)$ for all $\gamma\in\mathcal{S}\times\mathcal{Y}\times\mathcal{A}$, and let $\mathbf{e}$ an all-one vector of the same dimension. Similarly, define $\mathbf{g}_K(\lambda)$ as the column vector composed of the immediate costs $g(\gamma,Z^{(K)}(\gamma),A^{(K)}(\gamma);\lambda)$ arranged under the same indexing scheme. Let $\mathbf{P}(K)\in\mathbb{R}^{|\mathcal{S}\times\mathcal{Y}\times\mathcal{A}|\times|\mathcal{S}\times\mathcal{Y}\times\mathcal{A}|}$ denote the transition probability matrix where the (i,j)-th entry corresponds to the transition probability from γ_i to γ_j under the control $(Z^{(K)}(\gamma_i),A^{(K)}(\gamma_i))$, with γ_i,γ_j indexed according to the same fixed ordering of $\mathcal{S}\times\mathcal{Y}\times\mathcal{A}$. Similarly, let $\widetilde{\mathbf{P}(K)}\in\mathbb{R}^{|\mathcal{S}\times\mathcal{Y}\times\mathcal{A}|\times|\mathcal{S}\times\mathcal{Y}\times\mathcal{A}|}$ be defined using the modified transition probabilities $\widetilde{p_{\gamma_i\gamma_j}}(Z^{(K)}(\gamma),A_K(\gamma))$. Under this notation, τ-

$$\begin{aligned}0 &= \min_{A_i,Z_i}\left\{g(\gamma^{\text{ref}},A_i,Z_i;\rho^\star)+\mathbb{E}[W^*(\gamma')\,|\gamma^{\text{ref}},Z_i,A_i]\right\} = \min_{A_i,Z_i}\left\{q(\gamma^{\text{ref}},A_i,Z_i)-\rho^\star\cdot f(Z_i)+\mathbb{E}[W^*(\gamma')\,|\gamma^{\text{ref}},Z_i,A_i]\right\}\\ &= \min_{A_i,Z_i}\left\{f(Z_i)\cdot\left(\frac{q(\gamma^{\text{ref}},A_i,Z_i)+\mathbb{E}[W^*(\gamma')\,|\gamma^{\text{ref}},Z_i,A_i]}{f(Z_i)} - \rho^\star\right)\right\},\end{aligned} \tag{113}$$

RVI in (28) can be equivalently written in vector form as:

$$\begin{aligned}\tilde{\mathbf{V}}_{K+1}(\lambda) = (1-\tau)\tilde{\mathbf{V}}_K(\lambda) + \mathbf{g}_K(\lambda) \\ + \tau\mathbf{P}(K)\tilde{\mathbf{V}}_K(\lambda) - \tilde{U}_{K+1}(\lambda)\mathbf{e}.\end{aligned} \tag{119}$$

From (118), the pair $A^{(K)}(\gamma), Z^{(K)}(\gamma)$ is selected to minimize the following objective

$$g(\gamma, Z_i, A_i; \lambda) + \tau\mathbb{E}[\tilde{V}_K(\gamma'; \lambda)\,|\gamma, Z_i, A_i]. \tag{120}$$

This implies that the chosen action $A^{(K)}(\gamma), Z^{(K)}(\gamma)$ at the K-th iteration yields an objective value no greater than that obtained by any other actions. In vector form, this yields the following inequality:

$$\mathbf{g}_K(\lambda) + \tau\mathbf{P}(K)\tilde{\mathbf{V}}_K(\lambda) \leq \mathbf{g}_t(\lambda) + \tau\mathbf{P}(t)\tilde{\mathbf{V}}_K(\lambda), \forall K, t \geq 0, \tag{121}$$

which can be combined with the update rule in (119) to derive the following upper bound:

$$\begin{aligned}\tilde{\mathbf{V}}_{K+1}(\lambda) \leq (1-\tau)\tilde{\mathbf{V}}_K(\lambda) + \mathbf{g}_{K-1}(\lambda) \\ +\tau\mathbf{P}(K-1)\tilde{\mathbf{V}}_K(\lambda) - \tilde{U}_{K+1}(\lambda)\mathbf{e}.\end{aligned} \tag{122}$$

Similarly, applying the update equation (119) at iteration K and letting $t = K$ yields:

$$\begin{aligned}\tilde{\mathbf{V}}_K(\lambda) \leq (1-\tau)\tilde{\mathbf{V}}_{K-1}(\lambda) + \mathbf{g}_K(\lambda) + \\ \tau\mathbf{P}(K)\tilde{\mathbf{V}}_{K-1}(\lambda) - \tilde{U}_K(\lambda)\mathbf{e}.\end{aligned} \tag{123}$$

Let us define the difference between successive relative value function iterates as

$$\widetilde{\Delta_K\mathbf{V}}(\lambda) \triangleq \tilde{\mathbf{V}}_{K+1}(\lambda) - \tilde{\mathbf{V}}_K(\lambda). \tag{124}$$

Subtracting (123) from (122) yields recursive inequalities that characterize the evolution of the value difference:

$$\begin{aligned}\widetilde{\Delta_K\mathbf{V}}(\lambda) \leq (1-\tau)\widetilde{\Delta_{K-1}\mathbf{V}}(\lambda) + \\ \tau\mathbf{P}(K-1)\widetilde{\Delta_{K-1}\mathbf{V}}(\lambda) + (\tilde{U}_K(\lambda) - \tilde{U}_{K+1}(\lambda))\mathbf{e},\end{aligned} \tag{125a}$$

$$\begin{aligned}\widetilde{\Delta_K\mathbf{V}}(\lambda) \geq (1-\tau)\widetilde{\Delta_{K-1}\mathbf{V}}(\lambda) + \\ \tau\mathbf{P}(K)\widetilde{\Delta_{K-1}\mathbf{V}}(\lambda) + (\tilde{U}_K(\lambda) - \tilde{U}_{K+1}(\lambda))\mathbf{e}.\end{aligned} \tag{125b}$$

From (35), we know that the matrix $\widetilde{\mathbf{P}(K)}$ satisfies:

$$\widetilde{\mathbf{P}(K)} = (1-\tau)\mathbf{I} + \tau\mathbf{P}(K), \text{ for } K \geq 1. \tag{126}$$

Substituting (126) into (125) yields:

$$\widetilde{\Delta_K\mathbf{V}}(\lambda) \geq \widetilde{\mathbf{P}(K)}\widetilde{\Delta_{K-1}\mathbf{V}}(\lambda) + (\tilde{U}_K(\lambda) - \tilde{U}_{K+1}(\lambda))\mathbf{e}, \tag{127a}$$

$$\widetilde{\Delta_K\mathbf{V}}(\lambda) \leq \widetilde{\mathbf{P}(K-1)}\widetilde{\Delta_{K-1}\mathbf{V}}(\lambda) + (\tilde{U}_K(\lambda) - \tilde{U}_{K+1}(\lambda))\mathbf{e}. \tag{127b}$$

By recursively applying these inequalities over L iterations, we derive the following bounds:

$$\begin{aligned}\widetilde{\Delta_K\mathbf{V}}(\lambda) \geq \prod_{t=K}^{K-L+1}\widetilde{\mathbf{P}(t)}\widetilde{\Delta_{K-L}\mathbf{V}}(\lambda) + \\ (\tilde{U}_{K+1-L}(\lambda) - \tilde{U}_{K+1}(\lambda))\mathbf{e},\end{aligned} \tag{128a}$$

$$\begin{aligned}\widetilde{\Delta_K\mathbf{V}}(\lambda) \leq \prod_{t=K-1}^{K-L}\widetilde{\mathbf{P}(t)}\widetilde{\Delta_{K-L}\mathbf{V}}(\lambda) + \\ (\tilde{U}_{K+1-L}(\lambda) - \tilde{U}_{K+1}(\lambda))\mathbf{e}.\end{aligned} \tag{128b}$$

Since the transformed MDP is a *unichain* and the transition probability matrix $\widetilde{\mathbf{P}(K)}$ holds *aperiodic* for $\forall K \geq 1$ (as verified in the proof sketch), we have that there exists a positive integer L, a constant $\epsilon > 0$, and a state $\gamma^\star$ such that:

$$\left[\prod_{t=K}^{K-L+1}\widetilde{\mathbf{P}(t)}\right]_{\gamma\times\gamma^\star} \geq \epsilon, \quad \forall\gamma \in \mathcal{S}\times\mathcal{Y}\times\mathcal{A}, \tag{129a}$$

$$\left[\prod_{t=K-1}^{K-L}\widetilde{\mathbf{P}(t)}\right]_{\gamma\times\gamma^\star} \geq \epsilon, \quad \forall\gamma \in \mathcal{S}\times\mathcal{Y}\times\mathcal{A} \tag{129b}$$

In the following, we establish the existence of finite limits by analyzing two distinct cases.

1) Case 1: $\gamma^\star = \gamma^{\mathrm{r}}$

If $\gamma^\star = \gamma^{\mathrm{r}}$, we can derive the inequality (130) from (128b), as shown at the top of this page, where the transition from (130a) to (130b) uses the fact that for all K, the following holds:

$$\tilde{V}_K(\gamma^{\mathrm{r}}; \lambda) = (1-\tau)^K\tilde{V}_0(\gamma^{\mathrm{r}}; \lambda) = 0, \quad \forall K \in \mathbb{N}; \tag{131}$$

given the initialization $\tilde{V}_{K-1}(\gamma^{\mathrm{r}}; \lambda) = 0$. Inequality (130c) uses the uniform bound:

$$\begin{aligned}\tilde{V}_{K-L+1}(\gamma; \lambda) - \tilde{V}_{K-L}(\gamma; \lambda) \leq \\ \max_\gamma\{\tilde{V}_{K-L+1}(\gamma; \lambda) - \tilde{V}_{K-L}(\gamma; \lambda)\},\end{aligned} \tag{132}$$

and inequality (130d) follows from the definition in (129b):

$$\sum_{\gamma'\neq\gamma^\star}\left[\prod_{t=K-1}^{K-L}\widetilde{\mathbf{P}(t)}\right]_{\gamma\times\gamma'} = 1 - \left[\prod_{t=K-1}^{K-L}\widetilde{\mathbf{P}(t)}\right]_{\gamma\times\gamma^\star} \leq 1-\epsilon, \tag{133}$$

and the fact that the product term is *non-negative*:

$$\begin{aligned}\max_\gamma\left\{\tilde{V}_{K-L+1}(\gamma; \lambda) - \tilde{V}_{K-L}(\gamma; \lambda)\right\} \\ \geq \tilde{V}_{K-L+1}(\gamma^{\mathrm{r}}; \lambda) - \tilde{V}_{K-L}(\gamma^{\mathrm{r}}; \lambda) = 0.\end{aligned} \tag{134}$$

Since (130) holds for $\forall\gamma \in \mathcal{S}\times\mathcal{Y}\times\mathcal{A}$, we can bound the maximum increment:

$$\begin{aligned}\max_\gamma\left\{\tilde{V}_{K+1}(\gamma; \lambda) - \tilde{V}_K(\gamma; \lambda)\right\} \\ \leq (1-\epsilon)\max_\gamma\left\{\tilde{V}_{K-L+1}(\gamma; \lambda) - \tilde{V}_{K-L}(\gamma; \lambda)\right\} \\ + \tilde{U}_{K+1-L}(\lambda) - \tilde{U}_{K+1}(\lambda).\end{aligned} \tag{135}$$

In a similar manner, using (128a), we obtain the lower bound shown in (136) at the top of the next page, where inequality (136a) establishes by $\gamma^\star = \gamma^{\mathrm{r}}$ and

$$\begin{aligned}\tilde{V}_{K-L+1}(\gamma'; \lambda) - \tilde{V}_{K-L}(\gamma'; \lambda) \geq \\ \min_\gamma\left\{\tilde{V}_{K-L+1}(\gamma; \lambda) - \tilde{V}_{K-L}(\gamma; \lambda)\right\};\end{aligned} \tag{137}$$

$$
\begin{aligned}
&\tilde{V}_{K+1}(\gamma;\lambda)-\tilde{V}_K(\gamma;\lambda)\\
\leq&\tilde{U}_{K+1-L}(\lambda)-\tilde{U}_{K+1}(\lambda)+\sum_{\gamma'}\left[\prod_{t=K-1}^{K-L}\widetilde{\mathbf{P}(t)}\right]_{\gamma\times\gamma'}\times\left(\tilde{V}_{K-L+1}(\gamma';\lambda)-\tilde{V}_{K-L}(\gamma';\lambda)\right) &\text{(130a)}\\
=&\tilde{U}_{K+1-L}(\lambda)-\tilde{U}_{K+1}(\lambda)+\sum_{\gamma'\neq\gamma^{\mathrm{r}}}\left[\prod_{t=K-1}^{K-L}\widetilde{\mathbf{P}(t)}\right]_{\gamma\times\gamma'}\times\left(\tilde{V}_{K-L+1}(\gamma';\lambda)-\tilde{V}_{K-L}(\gamma';\lambda)\right) &\text{(130b)}\\
\leq&\tilde{U}_{K+1-L}(\lambda)-\tilde{U}_{K+1}(\lambda)+\max_{\gamma}\left\{\tilde{V}_{K-L+1}(\gamma;\lambda)-\tilde{V}_{K-L}(\gamma;\lambda)\right\}\times\sum_{\gamma'\neq\gamma^\star}\left[\prod_{t=K-1}^{K-L}\widetilde{\mathbf{P}(t)}\right]_{\gamma\times\gamma'} &\text{(130c)}\\
\leq&\tilde{U}_{K+1-L}(\lambda)-\tilde{U}_{K+1}(\lambda)+(1-\epsilon)\max_{\gamma}\left\{\tilde{V}_{K-L+1}(\gamma;\lambda)-\tilde{V}_{K-L}(\gamma;\lambda)\right\},\quad\forall\gamma\in\mathcal{S}\times\mathcal{Y}\times\mathcal{A}. &\text{(130d)}
\end{aligned}
$$

$$
\begin{aligned}
&\tilde{V}_{K+1}(\gamma;\lambda)-\tilde{V}_K(\gamma;\lambda)\\
\geq&\tilde{U}_{K+1-L}(\lambda)-\tilde{U}_{K+1}(\lambda)+\sum_{\gamma'}\left[\prod_{t=K}^{K-L+1}\widetilde{\mathbf{P}(t)}\right]_{\gamma\times\gamma'}\times\left(\tilde{V}_{K-L+1}(\gamma';\lambda)-\tilde{V}_{K-L}(\gamma';\lambda)\right) &\text{(136a)}\\
=&\tilde{U}_{K+1-L}(\lambda)-\tilde{U}_{K+1}(\lambda)+\sum_{\gamma'\neq\gamma^{\mathrm{r}}}\left[\prod_{t=K}^{K-L+1}\widetilde{\mathbf{P}(t)}\right]_{\gamma\times\gamma'}\times\left(\tilde{V}_{K-L+1}(\gamma';\lambda)-\tilde{V}_{K-L}(\gamma';\lambda)\right) &\text{(136b)}\\
\geq&\tilde{U}_{K+1-L}(\lambda)-\tilde{U}_{K+1}(\lambda)+\min_{\gamma}\left\{\tilde{V}_{K-L+1}(\gamma;\lambda)-\tilde{V}_{K-L}(\gamma;\lambda)\right\}\times\sum_{\gamma'\neq\gamma^\star}\left[\prod_{t=K}^{K-L+1}\widetilde{\mathbf{P}(t)}\right]_{\gamma\times\gamma'} &\text{(136c)}\\
\geq&\tilde{U}_{K+1-L}(\lambda)-\tilde{U}_{K+1}(\lambda)+(1-\epsilon)\min_{\gamma}\left\{\tilde{V}_{K-L+1}(\gamma;\lambda)-\tilde{V}_{K-L}(\gamma;\lambda)\right\},\quad\forall\gamma\in\mathcal{S}\times\mathcal{Y}\times\mathcal{A}. &\text{(136d)}
\end{aligned}
$$

$$
\begin{aligned}
&\max_{\gamma}\left\{\tilde{V}_{K+1}(\gamma;\lambda)-\tilde{V}_K(\gamma;\lambda)\right\}-\min_{\gamma}\left\{\tilde{V}_{K+1}(\gamma;\lambda)-\tilde{V}_K(\gamma;\lambda)\right\}\\
&\leq(1-\epsilon)\left(\max_{\gamma}\left\{\tilde{V}_{K-L+1}(\gamma;\lambda)-\tilde{V}_{K-L}(\gamma;\lambda)\right\}-\min_{\gamma}\left\{\tilde{V}_{K-L+1}(\gamma;\lambda)-\tilde{V}_{K-L}(\gamma;\lambda)\right\}\right).
\end{aligned}
\tag{141}
$$

and (136b) establishes because of (129b):

$$
\sum_{\gamma'\neq\gamma^\star}\left[\prod_{t=K}^{K-L+1}\widetilde{\mathbf{P}(t)}\right]_{\gamma\times\gamma'}=1-\left[\prod_{t=K}^{K-L+1}\widetilde{\mathbf{P}(t)}\right]_{\gamma\times\gamma^\star}\overset{(a)}{\leq}1-\epsilon, \tag{138}
$$

and the fact that the product term is *non-positive*:

$$
\begin{aligned}
&\min_{\gamma}\left\{\tilde{V}_{K-L+1}(\gamma;\lambda)-\tilde{V}_{K-L}(\gamma;\lambda)\right\}\\
&\leq\tilde{V}_{K-L+1}(\gamma^{\mathrm{r}};\lambda)-\tilde{V}_{K-L}(\gamma^{\mathrm{r}};\lambda)=0.
\end{aligned}
\tag{139}
$$

Since inequality (136) holds for $\forall\gamma$, we can establish:

$$
\begin{aligned}
&\min_{\gamma}\left\{\tilde{V}_{K+1}(\gamma;\lambda)-\tilde{V}_K(\gamma;\lambda)\right\}\\
&\geq(1-\epsilon)\min_{\gamma}\left\{\tilde{V}_{K-L+1}(\gamma;\lambda)-\tilde{V}_{K-L}(\gamma;\lambda)\right\}\\
&\quad+\tilde{U}_{K+1-L}(\lambda)-\tilde{U}_{K+1}(\lambda).
\end{aligned}
\tag{140}
$$

Subtracting (140) from (135) directly yields (141) at the top of the next page. Iterating (141) yields that for some $M>0$ and all $K\geq 1$, we have

$$
\begin{aligned}
&\max_{\gamma}\left\{\tilde{V}_{K+1}(\gamma;\lambda)-\tilde{V}_K(\gamma;\lambda)\right\}\\
&-\min_{\gamma}\left\{\tilde{V}_{K+1}(\gamma;\lambda)-\tilde{V}_K(\gamma';\lambda)\right\}\\
&\leq M(1-\epsilon)^{K/L}.
\end{aligned}
\tag{142}
$$

Therefore, the relative difference between $\tilde{V}_{K+1}(\gamma;\lambda)$ and $\tilde{V}_K(\gamma;\lambda)$ is upper bounded by

$$
\begin{aligned}
&\left|\tilde{V}_{K+1}(\gamma;\lambda)-\tilde{V}_K(\gamma;\lambda)\right|\\
&\leq\max_{\gamma}\left\{\tilde{V}_{K+1}(\gamma;\lambda)-\tilde{V}_K(\gamma;\lambda)\right\}-\\
&\quad\min_{\gamma}\left\{\tilde{V}_{K+1}(\gamma';\lambda)-\tilde{V}_K(\gamma;\lambda)\right\}\\
&\leq M(1-\epsilon)^{K/L},\forall\gamma\in\mathcal{S}\times\mathcal{Y}\times\mathcal{A}.
\end{aligned}
\tag{143}
$$

This indicates that $\{\tilde{V}_K(\gamma;\lambda)\}_{K\in\mathbb{N}^+}$ forms a *Cauchy sequence*. Specifically, for $\forall T>1$, the following holds:

$$
|\tilde{V}_{K+T}(\gamma;\lambda)-\tilde{V}_K(\gamma;\lambda)|
$$

$$\tilde{U}_{K+1}(\lambda) = g(\gamma^{\mathrm{r}}, Z^{(K)}(\gamma^{\mathrm{r}}), A^{(K)}(\gamma^{\mathrm{r}}); \lambda) + \tau \left[\mathbf{P}_{A^{(K)}(\gamma^{\mathrm{r}}), Z^{(K)}(\gamma^{\mathrm{r}})}\right]_{N(\gamma^{\mathrm{r}}),:} \times \tilde{\mathbf{V}}_K(\lambda) \tag{147a}$$

$$\leq g(\gamma^{\mathrm{r}}, Z^{(K-1)}(\gamma^{\mathrm{r}}), A^{(K-1)}(\gamma^{\mathrm{r}}); \lambda) + \tau \left[\mathbf{P}_{A^{(K-1)}(\gamma^{\mathrm{r}}), Z^{(K-1)}(\gamma^{\mathrm{r}})}\right]_{N(\gamma^{\mathrm{r}}),:} \times \tilde{\mathbf{V}}_K(\lambda). \tag{147b}$$

$$\tilde{U}_{K}(\lambda) = g(\gamma^{\mathrm{r}}, Z^{(K-1)}(\gamma^{\mathrm{r}}), A^{(K-1)}(\gamma^{\mathrm{r}}); \lambda) + \tau \left[\mathbf{P}_{A^{(K-1)}(\gamma^{\mathrm{r}}), Z^{(K-1)}(\gamma^{\mathrm{r}})}\right]_{N(\gamma^{\mathrm{r}}),:} \times \tilde{\mathbf{V}}_{K-1}(\lambda) \tag{148a}$$

$$\leq g(\gamma^{\mathrm{r}}, Z^{(K)}(\gamma^{\mathrm{r}}), A^{(K)}(\gamma^{\mathrm{r}}); \lambda) + \tau \left[\mathbf{P}_{A^{(K)}(\gamma^{\mathrm{r}}), Z^{(K)}(\gamma^{\mathrm{r}})}\right]_{N(\gamma^{\mathrm{r}}),:} \times \tilde{\mathbf{V}}_K(\lambda). \tag{148b}$$

$$\leq \sum_{t=0}^{T-1} |\tilde{V}_{K+t+1}(\gamma;\lambda) - \tilde{V}_{K+t}(\gamma;\lambda)| \tag{144a}$$

$$\leq M \sum_{t=0}^{T-1} (1-\epsilon)^{\frac{K+t}{L}} = \frac{M(1-\epsilon)^{K/L}(1-(1-\epsilon)^{T/L})}{1-(1-\epsilon)^{1/L}}, \tag{144b}$$

where inequality (144a) follows from the triangle inequality.

Letting $T \to \infty$, (144b) yields

$$|\tilde{V}_K(\gamma;\lambda) - \tilde{V}_\infty(\gamma;\lambda)| \leq \frac{M(1-\epsilon)^{K/L}}{1-(1-\epsilon)^{1/L}}, \tag{145}$$

which confirms that $\tilde{V}_K(\gamma;\lambda)$ converges to a bounded value $\tilde{V}_\infty(\gamma;\lambda)$ as $K \to \infty$, given any γ.

We next prove that $\tilde{U}_K(\lambda)$ also converges to a bounded value as $K \to \infty$. The update rule in (28a) can be rewritten into a vector form as:

$$\tilde{U}_{K+1}(\lambda) = \min_{A_i, Z_i} \left\{ g(\gamma^{\mathrm{r}}, Z_i, A_i; \lambda) + \tau [\mathbf{P}_{A_i, Z_i}]_{N(\gamma^{\mathrm{r}}),:} \times \tilde{\mathbf{V}}_K(\lambda) \right\}, \tag{146}$$

where $[\mathbf{P}_{a,z}]_{N(\gamma^{\mathrm{r}}),:}$ denotes the row vector formed by stacking the transition probabilities $\{p_{\gamma^{\mathrm{r}}\gamma'}(a,z)\}$ for all $\gamma' \in \mathcal{S} \times \mathcal{Y} \times \mathcal{A}$, arranged according to the same index as that of $\tilde{\mathbf{V}}_K(\lambda)$. Here, $N(\gamma^{\mathrm{r}})$ denotes the index of the reference state γ^{r}. By substituting the optimal control pair $(A^{(K)}(\gamma^{\mathrm{r}}), Z^{(K)}(\gamma^{\mathrm{r}}))$, as defined in (118), into the right-hand side of (146), we obtain the upper bound of $\tilde{U}_{K+1}$ given in (147), which is at the top of the next page. Similarly, the $\tilde{U}_K(\lambda)$ can be upper bounded as shown in (148), which is at the top of the next page. Subtracting (148a) from (147b) yields:

$$\tilde{U}_{K+1}(\lambda) - \tilde{U}_K(\lambda) \leq \tau \left[\mathbf{P}_{A^{(K-1)}(\gamma^{\mathrm{r}}), Z^{(K-1)}(\gamma^{\mathrm{r}})}\right]_{N(\gamma^{\mathrm{r}}),:} \times \left(\tilde{\mathbf{V}}_K(\lambda) - \tilde{\mathbf{V}}_{K-1}(\lambda)\right). \tag{149}$$

Subtracting (147a) from (148b) yields:

$$\tilde{U}_{K+1}(\lambda) - \tilde{U}_K(\lambda) \geq \tau \left[\mathbf{P}_{A^{(K)}(\gamma^{\mathrm{r}}), Z^{(K)}(\gamma^{\mathrm{r}})}\right]_{N(\gamma^{\mathrm{r}}),:} \times \left(\tilde{\mathbf{V}}_K(\lambda) - \tilde{\mathbf{V}}_{K-1}(\lambda)\right). \tag{150}$$

Combining the upper bound in (149) and the lower bound in (150), and applying the inequality $y \leq x \leq z \Rightarrow |x| \leq \max\{|y|, |z|\}$, we obtain (151) at the top of the next page.

The two terms inside the maximum operator in (151) can each be bounded as follows:

$$\left|\tau \left[\mathbf{P}_{A^{(K)}(\gamma^{\mathrm{r}}), Z^{(K)}(\gamma^{\mathrm{r}})}\right]_{N(\gamma^{\mathrm{r}}),:} \times \left(\tilde{\mathbf{V}}_K(\lambda) - \tilde{\mathbf{V}}_{K-1}(\lambda)\right)\right| \leq \tau \left[\mathbf{P}_{A^{(K)}(\gamma^{\mathrm{r}}), Z^{(K)}(\gamma^{\mathrm{r}})}\right]_{N(\gamma^{\mathrm{r}}),:} \times \left|\tilde{\mathbf{V}}_K(\lambda) - \tilde{\mathbf{V}}_{K-1}(\lambda)\right| \tag{152a}$$

$$\leq \tau M (1-\epsilon)^{\frac{K-1}{L}}, \tag{152b}$$

$$\left|\tau \left[\mathbf{P}_{A^{(K-1)}(\gamma^{\mathrm{r}}), Z^{(K-1)}(\gamma^{\mathrm{r}})}\right]_{N(\gamma^{\mathrm{r}}),:} \times \left(\tilde{\mathbf{V}}_K(\lambda) - \tilde{\mathbf{V}}_{K-1}(\lambda)\right)\right| \leq \tau \left[\mathbf{P}_{A^{(K)}(\gamma^{\mathrm{r}}), Z^{(K)}(\gamma^{\mathrm{r}})}\right]_{N(\gamma^{\mathrm{r}}),:} \times \left|\tilde{\mathbf{V}}_K(\lambda) - \tilde{\mathbf{V}}_{K-1}(\lambda)\right| \tag{153a}$$

$$\leq \tau M (1-\epsilon)^{\frac{K-1}{L}}, \tag{153b}$$

where (152a) and (153a) follow from the *triangle inequality*; (152b) and (153b) use the contraction bound established in (143), along with the fact that sum of the vector $\tau \left[\mathbf{P}_{A^{(K-1)}(\gamma^{\mathrm{r}}), Z^{(K-1)}(\gamma^{\mathrm{r}})}\right]_{N(\gamma^{\mathrm{r}}),:}$ is 1. Substituting (152) and (153) into (151), we conclude that

$$\left|\tilde{U}_{K+1}(\lambda) - \tilde{U}_K(\lambda)\right| \leq \tau M (1-\epsilon)^{\frac{K-1}{L}}. \tag{154}$$

This demonstrates that $\{\tilde{U}_K(\lambda)\}_{K \in \mathbb{N}^+}$ forms a *Cauchy sequence*. In particular, for $\forall T > 1$, we have:

$$|\tilde{U}_{K+T}(\lambda) - \tilde{U}_K(\lambda)|$$

$$\leq \sum_{t=0}^{T-1} |\tilde{U}_{K+t+1}(\lambda) - \tilde{U}_{K+t}(\lambda)| \tag{155a}$$

$$\leq \tau M \sum_{t=0}^{T-1} (1-\epsilon)^{\frac{K+t-1}{L}} \tag{155b}$$

$$= \frac{\tau M (1-\epsilon)^{(K-1)/L} \left(1-(1-\epsilon)^{T/L}\right)}{1-(1-\epsilon)^{1/L}}, \tag{155c}$$

where inequality (155a) follows from the *triangle inequality*, and (155b) follows directly from the bound in (154). Taking

$$\left|\tilde{U}_{K+1}(\lambda)-\tilde{U}_K(\lambda)\right| \leq$$
$$\max\left\{\left|\tau\left[\mathbf{P}_{A^{(K-1)}(\gamma^{\mathrm{r}}),Z^{(K-1)}(\gamma^{\mathrm{r}})}\right]_{N(\gamma^{\mathrm{r}}),:}\times\left(\tilde{\mathbf{V}}_K(\lambda)-\tilde{\mathbf{V}}_{K-1}(\lambda)\right)\right|,\left|\tau\left[\mathbf{P}_{A^{(K)}(\gamma^{\mathrm{r}}),Z^{(K)}(\gamma^{\mathrm{r}})}\right]_{N(\gamma^{\mathrm{r}}),:}\times\left(\tilde{\mathbf{V}}_K(\lambda)-\tilde{\mathbf{V}}_{K-1}(\lambda)\right)\right|\right\}. \tag{151}$$

the limit as $T\to\infty$, we have

$$\begin{aligned}&\left|\tilde{U}_K(\lambda)-\tilde{U}_\infty(\lambda)\right|\\&\leq\lim_{T\to\infty}\frac{\tau M(1-\epsilon)^{(K-1)/L}\left(1-(1-\epsilon)^{T/L}\right)}{1-(1-\epsilon)^{1/L}}\\&=\frac{\tau M(1-\epsilon)^{(K-1)/L}}{1-(1-\epsilon)^{1/L}}.\end{aligned}\tag{156}$$

This confirms that $\tilde{U}_K(\lambda)$ converges to a bounded limiting value, denoted by $\tilde{U}_\infty(\lambda)$.

2) Case 2: $\gamma^\star\neq\gamma^r$

In *Case 1*, we established that when $\gamma^\star=\gamma^{\mathrm{r}}$, both sequences $\{\tilde{V}_K(\gamma;\lambda)\}_{K\in\mathbb{N}^+}$ and $\{\tilde{U}_K(\lambda)\}_{K\in\mathbb{N}^+}$ constitute *Cauchy sequences*, and thus converge to bounded values. Here we extend the result to the more general case where the strict condition $\gamma^\star=\gamma^{\mathrm{r}}$ is relaxed. This generalization introduces significant analytical challenges, as key inequalities, specifically (130c) and (136c), no longer hold under the relaxed assumption. To overcome this challenge, we introduce an *auxiliary iteration sequence* in the following.

Iteration 4. *(Auxiliary Iteration Sequence). For a given λ and a parameter $0<\tau\leq 1$, the auxiliary iteration sequence iteratively generate sequences $\{\bar{U}_K(\lambda)\}^{K\in\mathbb{N}^+}$ and $\{\bar{V}_K(\gamma;\lambda)\}^{K\in\mathbb{N}^+}_{\gamma\in\mathcal{S}\times\mathcal{Y}\times\mathcal{A}}$ with a starting initial value $\{\bar{V}_0(\gamma;\lambda)\}_{\gamma\in\mathcal{S}\times\mathcal{Y}\times\mathcal{A}}$.*

$$\begin{aligned}\bar{U}_{K+1}(\lambda)=\min_{A_i,Z_i}\Big\{&g(\gamma^\star,Z_i,A_i;\lambda)+\\&\tau\mathbb{E}[\bar{V}_K\left(\gamma';\lambda\right)|\gamma^\star,Z_i,A_i]\Big\},\end{aligned}\tag{157a}$$

$$\begin{aligned}\bar{V}_{K+1}(\gamma;\lambda)=(1-\tau)\bar{V}_K(\gamma;\lambda)+\min_{A_i,Z_i}\Big\{&g(\gamma,Z_i,A_i;\lambda)\\&+\tau\mathbb{E}[\bar{V}_K\left(\gamma';\lambda\right)|\gamma,Z_i,A_i]\Big\}-\bar{U}_{K+1}(\lambda),\\&\forall\gamma\in\mathcal{S}\times\mathcal{Y}\times\mathcal{A},\end{aligned}\tag{157b}$$

where the initial condition satisfies that $\bar{V}_0(\gamma;\lambda)=\tilde{V}_0(\gamma;\lambda)$ for $\forall\gamma\in\mathcal{S}\times\mathcal{Y}\times\mathcal{A}$.

A key property of the generated *auxiliary iteration sequence* is that there exists an $M>0$ such that for all $K\geq 1$,

$$\begin{aligned}&\max_\gamma\left\{\bar{V}_{K+1}(\gamma;\lambda)-\bar{V}_K(\gamma;\lambda)\right\}\\&-\min_\gamma\left\{\bar{V}_{K+1}(\gamma;\lambda)-\bar{V}_K(\gamma';\lambda)\right\}\\&\leq M(1-\epsilon)^{K/L},\end{aligned}\tag{158}$$

where the proof follows an analogous approach to that of inequality (142) in *Case 1* and thus we omit the details here. This demonstrates that the sequences $\{\bar{V}_K(\gamma;\lambda)\}_{K\in\mathbb{N}^+}$ also forms a *Cauchy sequence*. We next describe the relationship between $\{\tilde{V}_K(\gamma;\lambda)\}_{K\in\mathbb{N}^+}$ and $\{\bar{V}_K(\gamma;\lambda)\}_{K\in\mathbb{N}^+}$. Compare (157) with (28), the relationship between $\bar{V}_K(\gamma;\lambda)$ and $\bar{V}_K(\gamma;\lambda)$ can be established by:

$$\tilde{V}_K(\gamma;\lambda)=\bar{V}_K(\gamma;\lambda)+\Psi\left(\bar{\mathbf{V}}_{K-1}(\lambda)\right),\tag{159}$$

where $\bar{\mathbf{V}}_K(\lambda)$ is a row vector consisted of $\bar{V}_K(\gamma;\lambda)$ for all $\gamma\in\mathcal{S}\times\mathcal{Y}\times\mathcal{A}$, arranged by the same index scheme as that of $\tilde{V}_K(\gamma)$. The function $\Psi\left(\bar{\mathbf{V}}_{K-1}(\lambda)\right)$ is defined as

$$\begin{aligned}&\Psi\left(\bar{\mathbf{V}}_{K-1}(\lambda)\right)\triangleq\\&\min_{A_i,Z_i}\left\{g(\gamma^\star,Z_i,A_i;\lambda)+\tau\mathbb{E}[\bar{V}_{K-1}\left(\gamma';\lambda\right)|\gamma^\star,Z_i,A_i]\right\}\\&-\min_{A_i,Z_i}\left\{g(\gamma^{\mathrm{r}},Z_i,A_i;\lambda)+\tau\mathbb{E}[\bar{V}_{K-1}\left(\gamma';\lambda\right)|\gamma^{\mathrm{r}},Z_i,A_i]\right\}.\end{aligned}\tag{160}$$

Similarly, the relationship between $\bar{V}_{K+1}(\gamma;\lambda)$ and $\tilde{V}_{K+1}(\gamma;\lambda)$ can be established by

$$\tilde{V}_{K+1}(\gamma;\lambda)=\bar{V}_{K+1}(\gamma;\lambda)+\Psi(\bar{\mathbf{V}}_K(\lambda))\tag{161}$$

Subtracting (159) from (161) yields:

$$\begin{aligned}&\tilde{V}_{K+1}(\gamma;\lambda)-\tilde{V}_K(\gamma;\lambda)\\&=\bar{V}_{K+1}(\gamma;\lambda)-\bar{V}_K(\gamma;\lambda)+\Psi(\bar{\mathbf{V}}_K(\lambda))-\Psi(\bar{\mathbf{V}}_{K-1}(\lambda)),\end{aligned}\tag{162}$$

Thus, by applying the max and min operators to both sides of (162), we obtain the following:

$$\begin{aligned}&\max_\gamma\{\tilde{V}_{K+1}(\gamma;\lambda)-\tilde{V}_K(\gamma;\lambda)\}=\\&\max_\gamma\{\bar{V}_{K+1}(\gamma;\lambda)-\bar{V}_K(\gamma;\lambda)\}+\Psi(\bar{\mathbf{V}}_K;\lambda)-\Psi(\bar{\mathbf{V}}_{K-1}(\lambda)),\end{aligned}\tag{163a}$$

$$\begin{aligned}&\min_\gamma\{\tilde{V}_{K+1}(\gamma;\lambda)-\tilde{V}_K(\gamma;\lambda)\}=\\&\min_\gamma\{\bar{V}_{K+1}(\gamma;\lambda)-\bar{V}_K(\gamma;\lambda)\}+\Psi(\bar{\mathbf{V}}_K(\lambda))-\Psi(\bar{\mathbf{V}}_{K-1}(\lambda)).\end{aligned}\tag{163b}$$

By subtracting equation (163b) from equation (163a), we obtain the key identity presented in (164), which is at the top of this page. Combining (164) with (158), it follows that the sequence $\tilde{V}_K(\gamma;\lambda)$ satisfies:

$$\begin{aligned}&\max_\gamma\left\{\tilde{V}_{K+1}(\gamma;\lambda)-\tilde{V}_K(\gamma;\lambda)\right\}\\&-\min_\gamma\left\{\tilde{V}_{K+1}(\gamma;\lambda)-\bar{V}_K(\gamma';\lambda)\right\}\leq M(1-\epsilon)^{K/L}.\end{aligned}\tag{165}$$

Given (165), and by applying the same reasoning used in Case 1 ((143)–(156)), we can show that both $\tilde{V}_K(\gamma;\lambda)$ and $\tilde{U}_K(\lambda)$ form *Cauchy sequences*. Thus, they also converge to bounded values. ■

$$\begin{aligned}&\max_{\gamma}\left\{\tilde{V}_{K+1}(\gamma;\lambda)-\tilde{V}_K(\gamma;\lambda)\right\}-\min_{\gamma}\left\{\tilde{V}_{K+1}(\gamma;\lambda)-\tilde{V}_K(\gamma;\lambda)\right\}\\&=\max_{\gamma}\left\{\bar{V}_{K+1}(\gamma;\lambda)-\bar{V}_K(\gamma;\lambda)\right\}-\min_{\gamma}\left\{\bar{V}_{K+1}(\gamma;\lambda)-\bar{V}_K(\gamma;\lambda)\right\}.\end{aligned}\tag{164}$$

B. Convergence Direction

In this subsection, we prove that the convergent bounded values $\tilde{U}_\infty(\lambda)$ and $\tilde{V}_\infty(\gamma;\lambda)$ constitute a solution to the ACOE (23). Given the convergence of $\tilde{U}_K(\lambda)$ and $\tilde{V}_K(\gamma;\lambda)$, we can take $K\to\infty$ into the τ-RVI (28), and establish that:

$$\begin{aligned}&\tilde{V}_\infty(\gamma;\lambda)+\tilde{U}_\infty(\lambda)=\min_{A_i,Z_i}\left\{g(\gamma,Z_i,A_i;\lambda)+\right.\\&\left.\sum_{\gamma'}\widetilde{p_{\gamma\gamma'}}(Z_i,A_i)\tilde{V}_\infty(\gamma';\lambda)\right\},\forall\gamma\in\mathcal{S}\times\mathcal{Y}\times\mathcal{A}.\end{aligned}\tag{166}$$

Compare (166) with (36) and we have that

$$\begin{aligned}\tilde{U}_\infty(\lambda)&=\tilde{U}(\lambda),\\\tilde{V}_\infty(\gamma;\lambda)&=\tilde{V}^\star(\gamma;\lambda),\forall\gamma\in\mathcal{S}\times\mathcal{Y}\times\mathcal{A}.\end{aligned}\tag{167}$$

Having established the equivalence between $\tilde{U}_\infty(\lambda)$ and $\tilde{U}(\lambda)$, as well as between $\tilde{V}_\infty(\gamma;\lambda)$ and $\tilde{V}^\star(\gamma;\lambda)$, we now proceed to derive the relationships between $\tilde{U}(\lambda)$ and $U(\lambda)$, as well as between $\tilde{V}_\infty(\gamma;\lambda)$ and $V^\star(\gamma;\lambda)$. By substituting (35) into (36), we obtain:

$$\begin{aligned}&\tilde{V}^\star(\gamma;\lambda)+\tilde{U}(\lambda)=\min_{A_i,Z_i}\left\{g(\gamma,Z_i,A_i;\lambda)+\right.\\&\left.\sum_{\gamma'}\tau p_{\gamma\gamma'}(Z_i,A_i)\tilde{V}^\star(\gamma';\lambda)+(1-\tau)\tilde{V}^\star(\gamma;\lambda)\right\},\\&\qquad\qquad\forall\gamma\in\mathcal{S}\times\mathcal{Y}\times\mathcal{A}.\end{aligned}\tag{168}$$

This expression can be reformulated into a more concise form:

$$\begin{aligned}&\tau\tilde{V}^\star(\gamma;\lambda)+\tilde{U}(\lambda)=\min_{A_i,Z_i}\left\{g(\gamma,Z_i,A_i;\lambda)+\right.\\&\left.\mathbb{E}\left[\tau\tilde{V}^\star(\gamma';\lambda)|Z_i,A_i\right]\right\},\text{for }\forall\gamma\in\mathcal{S}\times\mathcal{Y}\times\mathcal{A}.\end{aligned}\tag{169}$$

Comparing (169) with the ACOE (23), we observe that $\tilde{U}(\lambda)$ and $\tau\tilde{V}^\star(\gamma;\lambda),\gamma\in\mathcal{S}\times\mathcal{Y}\times\mathcal{A}$ are solutions to (23). This leads to the following relationship:

$$\begin{aligned}\tilde{U}(\lambda)&=U(\lambda),\\\tau\tilde{V}^\star(\gamma;\lambda)&=V^\star(\gamma;\lambda),\gamma\in\mathcal{S}\times\mathcal{Y}\times\mathcal{A}.\end{aligned}\tag{170}$$

By substituting (170) into (167) and invoking (156), we obtain

$$e_{\mathrm{U}}^{(K)}(\lambda)\le\frac{\tau M(1-\epsilon)^{(K-1)/L}}{1-(1-\epsilon)^{1/L}},\tag{171}$$

which completes the proof of Theorem 3.

Furthermore, letting $K\to\infty$ on both sides of (145) and (156), we have

$$\begin{aligned}0\le\lim_{K\to\infty}e_{\mathrm{U}}^{(K)}(\lambda)\le\lim_{K\to\infty}\frac{\tau M(1-\epsilon)^{(K-1)/L}}{1-(1-\epsilon)^{1/L}}=0,\\0\le\lim_{K\to\infty}e_{\mathrm{V}}^{(K)}(\gamma;\tau,\lambda)\le\lim_{K\to\infty}\frac{M(1-\epsilon)^{K/L}}{1-(1-\epsilon)^{1/L}}=0.\end{aligned}\tag{172}$$

Therefore, by the squeeze theorem, both error terms converge to zero, which completes the proof of Theorem 2. ■

Appendix J
Proof of Theorem 5

The proof proceeds in two distinct parts to establish the desired result. First, in section J-A, we demonstrate that both $\lim_{K\to\infty}\rho_K$ and $\lim_{K\to\infty}\widetilde{W}_K(\gamma),\forall\gamma\in\mathcal{S}\times\mathcal{Y}\times\mathcal{A}$ exist and are finite for all $\gamma\in\mathcal{S}\times\mathcal{Y}\times\mathcal{A}$. Then, we explicitly establish that $\lim_{K\to\infty}\rho_K=\rho^\star$ and $\lim_{K\to\infty}\widetilde{W}_K(\gamma)=\frac{W^\star(\gamma)}{\kappa\cdot\mathbb{E}[Y_i]}$ in Section J-B.

A. Convergence of (43)

Denote $Z^{(K)}(\gamma)$, $A^{(K)}(\gamma)$ as the waiting time and controlled action that achieves the minimum in the K-th relation given in (173) at the top of this page. Let $\widetilde{\mathbf{W}}_K\in\mathbb{R}^{|\mathcal{S}\times\mathcal{Y}\times\mathcal{A}|}$ as the column vector formed by stacking the values $\widetilde{W}_K(\gamma)$ for all $\gamma\in\mathcal{S}\times\mathcal{Y}\times\mathcal{A}$, and define $\mathbf{q}_K\in\mathbb{R}^{|\mathcal{S}\times\mathcal{Y}\times\mathcal{A}|}$ as the column vector composed $q(\gamma,Z^{(K)}(\gamma),A^{(K)}(\gamma))$ arranged under the same indexing scheme. Similarly, let $\mathbf{f}_K\in\mathbb{R}^{|\mathcal{S}\times\mathcal{Y}\times\mathcal{A}|}$ be a column vector consisting of $f(Z^{(K)}(\gamma))$. Let $\mathbf{P}(K)\in\mathbb{R}^{|\mathcal{S}\times\mathcal{Y}\times\mathcal{A}|\times|\mathcal{S}\times\mathcal{Y}\times\mathcal{A}|}$ denote a stochastic matrix where the (i,j)-th entry is given by $p_{\gamma_i\gamma_j}(Z^{(K)}(\gamma_i),A^{(K)}(\gamma_i))$. Finally, we denote by $\oslash$ the element-wise *Hadamard division* operator between vectors, defined as:

$$\mathbf{b}\oslash\mathbf{a}\triangleq\left[\frac{b_1}{a_1},\frac{b_2}{a_2},\cdots,\frac{b_n}{a_n}\right]^{\mathrm{T}},\tag{174}$$

where $\mathbf{b}=[b_1,\cdots,b_n]$ and $\mathbf{a}=[a_1,\cdots,a_n]$. With the above notations, the recursive relation for $\widetilde{W}_K(\gamma)$ in (43) can be compactly expressed in vector form as:

$$\begin{aligned}\widetilde{\mathbf{W}}_{K+1}=\widetilde{\mathbf{W}}_K+\mathbf{q}_K\oslash\mathbf{f}_K-\kappa\mathbb{E}[Y_i]\cdot\widetilde{\mathbf{W}}_K\oslash\mathbf{f}_K\\+\kappa\mathbb{E}[Y_i]\cdot\mathbf{P}(K)\widetilde{\mathbf{W}}_K\oslash\mathbf{f}_K-h_{K+1}\mathbf{e}.\end{aligned}\tag{175}$$

Since $A^{(K)}(\gamma)$ and $Z^{(K)}(\gamma)$ are chosen to minimize the right-hand side of (173), it follows that for all $K,t\ge0$,

$$\begin{aligned}&\mathbf{q}_K\oslash\mathbf{f}_K-\kappa\mathbb{E}[Y_i]\cdot\widetilde{\mathbf{W}}_K\oslash\mathbf{f}_K+\kappa\mathbb{E}[Y_i]\cdot\mathbf{P}(K)\widetilde{\mathbf{W}}_K\oslash\mathbf{f}_K\\&\le\mathbf{q}_t\oslash\mathbf{f}_t-\kappa\mathbb{E}[Y_i]\cdot\widetilde{\mathbf{W}}_K\oslash\mathbf{f}_t+\kappa\mathbb{E}[Y_i]\cdot\mathbf{P}(t)\widetilde{\mathbf{W}}_K\oslash\mathbf{f}_t.\end{aligned}\tag{176}$$

Applying this inequality in (175) yields the upper bound:

$$\begin{aligned}\widetilde{\mathbf{W}}_{K+1}\le\widetilde{\mathbf{W}}_K+\mathbf{q}_{K-1}\oslash\mathbf{f}_{K-1}-\kappa\mathbb{E}[Y_i]\cdot\widetilde{\mathbf{W}}_K\oslash\mathbf{f}_{K-1}\\+\kappa\mathbb{E}[Y_i]\cdot\mathbf{P}(K-1)\widetilde{\mathbf{W}}_K\oslash\mathbf{f}_{K-1}-\rho_{K+1}\mathbf{e}.\end{aligned}\tag{177}$$

Similarly, the recursive relationship between $\widetilde{\mathbf{W}}_K$ and $\widetilde{\mathbf{W}}_{K-1}$ can be established leveraging (176):

$$\begin{aligned}\widetilde{\mathbf{W}}_K&=\widetilde{\mathbf{W}}_{K-1}+\mathbf{q}_{K-1}\oslash\mathbf{f}_{K-1}-\kappa\mathbb{E}[Y_i]\cdot\widetilde{\mathbf{W}}_{K-1}\oslash\mathbf{f}_{K-1}\\&\quad+\kappa\mathbb{E}[Y_i]\cdot\mathbf{P}(K-1)\widetilde{\mathbf{W}}_{K-1}\oslash\mathbf{f}_{K-1}-\rho_K\mathbf{e}&\text{(178a)}\\&\le\widetilde{\mathbf{W}}_{K-1}+\mathbf{q}_K\oslash\mathbf{f}_K-\kappa\mathbb{E}[Y_i]\cdot\widetilde{\mathbf{W}}_{K-1}\oslash\mathbf{f}_K\\&\quad+\kappa\mathbb{E}[Y_i]\cdot\mathbf{P}(K)\widetilde{\mathbf{W}}_{K-1}\oslash\mathbf{f}_K-\rho_K\mathbf{e}.&\text{(178b)}\end{aligned}$$

$$(A^{(K)}(\gamma), Z^{(K)}(\gamma)) \triangleq \underset{A_i, Z_i}{\arg\min} \left\{ \frac{q(\gamma, Z_i, A_i) - \kappa \widetilde{W}_K(\gamma) \cdot \mathbb{E}[Y_i] + \kappa \mathbb{E}\left[\widetilde{W}_K(\gamma')|\gamma, Z_i, A_i\right] \cdot \mathbb{E}[Y_i]}{f(Z_i)} \right\}. \tag{173}$$

$$\begin{aligned} \widetilde{\Delta_K \mathbf{W}} &\leq \widetilde{\Delta_{K-1} \mathbf{W}} - \kappa \mathbb{E}[Y_i] \cdot \widetilde{\Delta_{K-1} \mathbf{W}} \oslash \mathbf{f}_{K-1} + \kappa \mathbb{E}[Y_i] \cdot \mathbf{P}(K-1) \widetilde{\Delta_{K-1} \mathbf{W}} \oslash \mathbf{f}_{K-1} + (\rho_K - \rho_{K+1}) \mathbf{e} \\ &= \left(\mathbf{I} - \operatorname{diag}\left(\frac{\kappa \mathbb{E}[Y_i]}{\mathbf{f}_{K-1}} \right) + \mathbf{P}(K-1) \operatorname{diag}\left(\frac{\kappa \mathbb{E}[Y_i]}{\mathbf{f}_{K-1}} \right) \right) \widetilde{\Delta_{K-1} \mathbf{W}} + (\rho_K - \rho_{K+1}) \mathbf{e}, \end{aligned} \tag{180a}$$

$$\begin{aligned} \widetilde{\Delta_K \mathbf{W}} &\geq \widetilde{\Delta_{K-1} \mathbf{W}} - \kappa \mathbb{E}[Y_i] \cdot \widetilde{\Delta_{K-1} \mathbf{W}} \oslash \mathbf{f}_{K} + \kappa \mathbb{E}[Y_i] \cdot \mathbf{P}(K) \widetilde{\Delta_{K-1} \mathbf{W}} \oslash \mathbf{f}_{K} + (\rho_K - \rho_{K+1}) \mathbf{e} \\ &= \left(\mathbf{I} - \operatorname{diag}\left(\frac{\kappa \mathbb{E}[Y_i]}{\mathbf{f}_{K}} \right) + \mathbf{P}(K) \operatorname{diag}\left(\frac{\kappa \mathbb{E}[Y_i]}{\mathbf{f}_{K}} \right) \right) \widetilde{\Delta_{K-1} \mathbf{W}} + (\rho_K - \rho_{K+1}) \mathbf{e}. \end{aligned} \tag{180b}$$

By subtracting (178) from (177) and defining:

$$\widetilde{\Delta_K \mathbf{W}} \triangleq \widetilde{\mathbf{W}}_{K+1} - \widetilde{\mathbf{W}}_K, \tag{179}$$

we obtain the recursive relations in (180) at the top of the next page. For short-hand notations, we define $\widetilde{\mathbf{P}_\kappa(K)}$ as:

$$\widetilde{\mathbf{P}_\kappa(K)} \triangleq \mathbf{I} - \operatorname{diag}\left(\frac{\kappa \mathbb{E}[Y_i]}{\mathbf{f}_K} \right) + \mathbf{P}(K) \operatorname{diag}\left(\frac{\kappa \mathbb{E}[Y_i]}{\mathbf{f}_K} \right). \tag{181}$$

In the following lemma, we show that the introduced matrix $\widetilde{\mathbf{P}}_\kappa(K)$ is *aperiodic* and *stochastic*.

Lemma 9. *If $0 < \kappa < 1$ and $\mathbf{P}(K)$ forms a unichain, the matrix $\widetilde{\mathbf{P}_\kappa(K)}$ is an aperiodic unichain stochastic matrix,* i.e., *for any $0 < \epsilon < 1$, there exists a positive integer L and a state $\gamma^\star$ satisfying:*

$$\left[\prod_{t=K}^{K-L+1} \widetilde{\mathbf{P}_\kappa(t)} \right]_{\gamma \times \gamma^\star} \geq \epsilon, \quad \forall \gamma \in \mathcal{S} \times \mathcal{Y} \times \mathcal{A}, \tag{182a}$$

$$\left[\prod_{t=K-1}^{K-L} \widetilde{\mathbf{P}_\kappa(t)} \right]_{\gamma \times \gamma^\star} \geq \epsilon, \quad \forall \gamma \in \mathcal{S} \times \mathcal{Y} \times \mathcal{A}. \tag{182b}$$

Proof. See Appendix P. ∎

By substituting (181) into (180a) and (180b), we obtain

$$\widetilde{\Delta_K \mathbf{W}} \geq \widetilde{\mathbf{P}_\kappa(K-1)} \widetilde{\Delta_{K-1} \mathbf{W}} + (\rho_K - \rho_{K+1}) \mathbf{e}, \tag{183a}$$

$$\widetilde{\Delta_K \mathbf{W}} \leq \widetilde{\mathbf{P}_\kappa(K)} \widetilde{\Delta_{K-1} \mathbf{W}} + (\rho_K - \rho_{K+1}) \mathbf{e}. \tag{183b}$$

Upon iterating (183a) and (183b) for L successive steps, we obtain the following lower and upper bounds:

$$\widetilde{\Delta_K \mathbf{W}} \geq \prod_{t=K}^{K-L+1} \widetilde{\mathbf{P}_\kappa(t)} \widetilde{\Delta_{K-L} \mathbf{W}} + (\rho_{K+1-L} - \rho_{K+1}) \mathbf{e}, \tag{184a}$$

$$\widetilde{\Delta_K \mathbf{W}} \leq \prod_{t=K-1}^{K-L} \widetilde{\mathbf{P}_\kappa(t)} \widetilde{\Delta_{K-L} \mathbf{W}} + (\rho_{K+1-L} - \rho_{K+1}) \mathbf{e}. \tag{184b}$$

In the following, we establish the existence of finite limits by analyzing two distinct cases.

1) Case 1: $\gamma^\star = \gamma^{\mathrm{r}}$

If $\gamma^\star = \gamma^{\mathrm{r}}$, by reformulating (184b) in its scalar form, we can obtain the inequality (185). where (185b) follows from substituting $\gamma = \gamma^{\mathrm{r}}$ into (43b), which yields:

$$\widetilde{W}_{K+1}(\gamma^{\mathrm{r}}) = \widetilde{W}_K(\gamma^{\mathrm{r}}). \tag{186}$$

Iterating (186) yields

$$\widetilde{W}_K(\gamma^{\mathrm{r}}) = \widetilde{W}_0(\gamma^{\mathrm{r}}) = 0, \forall K \geq 1, \tag{187}$$

where (187) holds due to the initialization $\widetilde{W}_0(\gamma^{\mathrm{r}}) = 0$ in OnePDSI. Inequality (185c) is derived by noting that $\gamma^\star = \gamma^{\mathrm{r}}$ and applying the following bound:

$$\begin{aligned} &\widetilde{W}_{K-L+1}(\gamma) - \widetilde{W}_{K-L}(\gamma) \leq \\ &\max_{\gamma} \left\{ \widetilde{W}_{K-L+1}(\gamma) - \tilde{W}_{K-L}(\gamma) \right\}; \end{aligned} \tag{188}$$

Inequality (185d) is established by observing that:

$$\begin{aligned} &\max_{\gamma} \left\{ \widetilde{W}_{K-L+1}(\gamma) - \tilde{W}_{K-L}(\gamma) \right\} \\ &\geq \widetilde{W}_{K-L+1}(\gamma^{\mathrm{r}}) - \tilde{W}_{K-L}(\gamma^{\mathrm{r}}) = 0, \end{aligned} \tag{189}$$

and a reformulation of (182b):

$$\begin{aligned} \sum_{\gamma' \neq \gamma^\star} \left[\prod_{t=K-1}^{K-L} \widetilde{\mathbf{P}_\kappa(t)} \right]_{\gamma \times \gamma'} &= 1 - \left[\prod_{t=K-1}^{K-L} \widetilde{\mathbf{P}_\kappa(t)} \right]_{\gamma \times \gamma^\star} \\ &\leq 1 - \epsilon, \end{aligned} \tag{190}$$

Since inequality (185) holds for $\forall \gamma \in \mathcal{S} \times \mathcal{Y} \times \mathcal{A}$, we can rewrite it as

$$\begin{aligned} &\max_{\gamma} \left\{ \widetilde{W}_{K+1}(\gamma) - \widetilde{W}_K(\gamma) \right\} \\ &\leq (1 - \epsilon) \max_{\gamma} \left\{ \widetilde{W}_{K-L+1}(\gamma) - \widetilde{W}_{K-L}(\gamma) \right\} \\ &\quad + \rho_{K+1-L} - \rho_{K+1}. \end{aligned} \tag{191}$$

Similar to the derivation of (185), we can obtain inequality (192) from (183a), as shown at the top of the next page. The derivation proceeds through several key steps: First, inequality (192a) is established by expanding the vector form of (183a) into its scalar representation. Then, (192b) follows directly from (187). To establish (192c), we use the condition $\gamma^\star = \gamma^{\mathrm{r}}$

$$
\begin{aligned}
&\widetilde{W}_{K+1}(\gamma) - \widetilde{W}_K(\gamma) \\
&\leq \rho_{K+1-L} - \rho_{K+1} + \sum_{\gamma'} \left[\prod_{t=K-1}^{K-L} \widetilde{\mathbf{P}(t)} \right]_{\gamma \times \gamma'} \times \left(\widetilde{W}_{K-L+1}(\gamma') - \widetilde{W}_{K-L}(\gamma') \right) && \text{(185a)} \\
&= h_{K+1-L} - h_{K+1} + \sum_{\gamma' \neq \gamma^{\mathrm{r}}} \left[\prod_{t=K-1}^{K-L} \widetilde{\mathbf{P}(t)} \right]_{\gamma \times \gamma'} \times \left(\widetilde{W}_{K-L+1}(\gamma') - \widetilde{W}_{K-L}(\gamma') \right) && \text{(185b)} \\
&\leq h_{K+1-L} - h_{K+1} + \max_{\gamma} \left\{ \widetilde{W}_{K-L+1}(\gamma) - \widetilde{W}_{K-L}(\gamma) \right\} \times \sum_{\gamma' \neq \gamma^\star} \left[\prod_{t=K-1}^{K-L} \widetilde{\mathbf{P}(t)} \right]_{\gamma \times \gamma'} && \text{(185c)} \\
&\leq h_{K+1-L} - h_{K+1} + (1-\epsilon) \max_{\gamma} \left\{ \widetilde{W}_{K-L+1}(\gamma) - \widetilde{W}_{K-L}(\gamma) \right\}, \quad \forall \gamma \in \mathcal{S} \times \mathcal{Y} \times \mathcal{A}. && \text{(185d)}
\end{aligned}
$$

$$
\begin{aligned}
&\widetilde{W}_{K+1}(\gamma) - \widetilde{W}_K(\gamma) \\
&\geq \rho_{K+1-L} - \rho_{K+1} + \sum_{\gamma'} \left[\prod_{t=K}^{K-L+1} \widetilde{\mathbf{P}_\kappa(t)} \right]_{\gamma \times \gamma'} \times \left(\widetilde{W}_{K-L+1}(\gamma') - \widetilde{W}_{K-L}(\gamma') \right) && \text{(192a)} \\
&= \rho_{K+1-L} - \rho_{K+1} + \sum_{\gamma' \neq \gamma^{\mathrm{r}}} \left[\prod_{t=K}^{K-L+1} \widetilde{\mathbf{P}_\kappa(t)} \right]_{\gamma \times \gamma'} \times \left(\widetilde{W}_{K-L+1}(\gamma') - \widetilde{W}_{K-L}(\gamma') \right) && \text{(192b)} \\
&\geq \rho_{K+1-L} - \rho_{K+1} + \min_{\gamma} \left\{ {}_{K-L+1}(\gamma;\lambda) - \widetilde{W}_{K-L}(\gamma;\lambda) \right\} \times \sum_{\gamma' \neq \gamma^\star} \left[\prod_{t=K}^{K-L+1} \widetilde{\mathbf{P}(t)} \right]_{\gamma \times \gamma'} && \text{(192c)} \\
&\geq \rho_{K+1-L} - \rho_{K+1} + (1-\epsilon) \min_{\gamma} \left\{ \widetilde{W}_{K-L+1}(\gamma) - \rho_{K-L}(\gamma;\lambda) \right\}, \quad \forall \gamma \in \mathcal{S} \times \mathcal{Y} \times \mathcal{A}. && \text{(192d)}
\end{aligned}
$$

---

along with the inequality:

$$
\begin{aligned}
&\widetilde{W}_{K-L+1}(\gamma') - \widetilde{W}_{K-L}(\gamma') \geq \\
&\min_{\gamma} \left\{ \widetilde{W}_{K-L+1}(\gamma) - \widetilde{W}_{K-L}(\gamma) \right\};
\end{aligned} \tag{193}
$$

Inequality (192d) follows from two key observations. First,

$$
\begin{aligned}
&\min_{\gamma} \left\{ \widetilde{W}_{K-L+1}(\gamma) - \tilde{W}_{K-L}(\gamma) \right\} \\
&\leq \widetilde{W}_{K-L+1}(\gamma^{\mathrm{r}}) - \tilde{W}_{K-L}(\gamma^{\mathrm{r}}) = 0.
\end{aligned} \tag{194}
$$

And second:

$$
\sum_{\gamma' \neq \gamma^\star} \left[\prod_{t=K}^{K-L+1} \widetilde{\mathbf{P}(t)} \right]_{\gamma \times \gamma'} = 1 - \left[\prod_{t=K}^{K-L+1} \widetilde{\mathbf{P}(t)} \right]_{\gamma \times \gamma^\star} \overset{(a)}{\leq} 1 - \epsilon, \tag{195}
$$

where the inequality follows from (182a). Given that inequality (192) holds for all $\gamma \in \mathcal{S} \times \mathcal{Y} \times \mathcal{A}$, we can rewrite it in terms of the minimal difference across states:

$$
\begin{aligned}
&\min_{\gamma} \left\{ \widetilde{W}_{K+1}(\gamma) - \widetilde{W}_K(\gamma) \right\} \\
&\geq (1-\epsilon) \min_{\gamma} \left\{ \widetilde{W}_{K-L+1}(\gamma) - \widetilde{W}_{K-L}(\gamma) \right\} \\
&\quad + \rho_{K+1-L} - \rho_{K+1}.
\end{aligned} \tag{196}
$$

Finally, by subtracting (196) from (191), we have (197) at the top of the next page. By iterating (197), we can conclude that for some constant M and any $K \geq 1$:

$$
\begin{aligned}
&\max_{\gamma} \left\{ \widetilde{W}_{K+1}(\gamma) - \widetilde{W}_K(\gamma) \right\} - \min_{\gamma'} \left\{ \widetilde{W}_{K+1}(\gamma) - \widetilde{W}_K(\gamma) \right\} \\
&\leq M(1-\epsilon)^{K/L}.
\end{aligned} \tag{198}
$$

This result allows us to establish an upper bound on the relative difference between $\widetilde{W}_{K+1}(\gamma)$ and $\widetilde{W}_K(\gamma)$:

$$
\begin{aligned}
&\left| \widetilde{W}_{K+1}(\gamma) - \widetilde{W}_K(\gamma) \right| \leq \\
&\max_{\gamma} \left\{ \widetilde{W}_{K+1}(\gamma) - \widetilde{W}_K(\gamma) \right\} - \min_{\gamma} \left\{ \widetilde{W}_{K+1}(\gamma) - \widetilde{W}_K(\gamma) \right\} && \text{(199a)} \\
&\leq M(1-\epsilon)^{K/L}, \quad \forall \gamma \in \mathcal{S} \times \mathcal{Y} \times \mathcal{A}. && \text{(199b)}
\end{aligned}
$$

This bound demonstrates that the sequence $\{\widetilde{W}_K(\gamma)\}_{K \in \mathbb{N}^+}$ constitutes a *Cauchy sequence*. Specifically, $\forall T > 1$, the following inequality holds:

$$
\begin{aligned}
&\left| \widetilde{W}_{K+T}(\gamma) - \widetilde{W}_K(\gamma) \right| \\
&\leq \sum_{t=0}^{T-1} \left| \widetilde{W}_{K+t+1}(\gamma) - \widetilde{W}_{K+t}(\gamma) \right| && \text{(200a)} \\
&\leq M \sum_{t=0}^{T-1} (1-\epsilon)^{\frac{K+t}{L}} = \frac{M(1-\epsilon)^{K/L}(1-(1-\epsilon)^{T/L})}{1-(1-\epsilon)^{1/L}}, && \text{(200b)}
\end{aligned}
$$

$$\begin{aligned}&\max_{\gamma}\left\{\widetilde{W}_{K+1}(\gamma)-\widetilde{W}_{K}(\gamma)\right\}-\min_{\gamma}\left\{\widetilde{W}_{K+1}(\gamma)-\widetilde{W}_{K}(\gamma)\right\}\\ \leq&(1-\epsilon)\left(\max_{\gamma}\left\{\widetilde{W}_{K-L+1}(\gamma)-\widetilde{W}_{K-L}(\gamma)\right\}-\min_{\gamma}\left\{\widetilde{W}_{K-L+1}(\gamma)-\widetilde{W}_{K-L}(\gamma)\right\}\right).\end{aligned} \tag{197}$$

$$\rho_{K+1}=\widetilde{W}_K(\gamma^{\mathrm{r}})+\min_{A_i,Z_i}\left\{\frac{q(\gamma^{\mathrm{r}},Z_i,A_i)-\kappa\widetilde{W}_K(\gamma^{\mathrm{r}})\cdot\mathbb{E}[Y_i]+\kappa\mathbb{E}\left[\widetilde{W}_K(\gamma')|\gamma^{\mathrm{r}},Z_i,A_i\right]\cdot\mathbb{E}[Y_i]}{f(Z_i)}\right\}$$

$$=\widetilde{W}_K(\gamma^{\mathrm{r}})+\frac{q(\gamma^{\mathrm{r}},Z^{(K)}(\gamma^{\mathrm{r}}),A^{(K)}(\gamma^{\mathrm{r}}))-\kappa\widetilde{W}_K(\gamma^{\mathrm{r}})\cdot\mathbb{E}[Y_i]+\kappa\mathbb{E}[Y_i]\cdot\left[\mathbf{P}_{A^{(K)}(\gamma^{\mathrm{r}}),Z^{(K)}(\gamma^{\mathrm{r}})}\right]_{N(\gamma^{\mathrm{r}}),:}\widetilde{\mathbf{W}}_K}{f(Z^{(K)}(\gamma^{\mathrm{r}}))} \tag{202a}$$

$$\leq\widetilde{W}_K(\gamma^{\mathrm{r}})+\frac{q(\gamma^{\mathrm{r}},Z^{(K-1)}(\gamma^{\mathrm{r}}),A^{(K-1)}(\gamma^{\mathrm{r}}))-\kappa\widetilde{W}_K(\gamma^{\mathrm{r}})\cdot\mathbb{E}[Y_i]+\kappa\mathbb{E}[Y_i]\cdot\left[\mathbf{P}_{A^{(K-1)}(\gamma^{\mathrm{r}}),Z^{(K-1)}(\gamma^{\mathrm{r}})}\right]_{N(\gamma^{\mathrm{r}}),:}\widetilde{\mathbf{W}}_K}{f(Z^{(K-1)}(\gamma^{\mathrm{r}}))}. \tag{202b}$$

$$\rho_{K}=\widetilde{W}_{K-1}(\gamma^{\mathrm{r}})+\min_{A_i,Z_i}\left\{\frac{q(\gamma^{\mathrm{r}},Z_i,A_i)-\kappa\widetilde{W}_{K-1}(\gamma^{\mathrm{r}})\cdot\mathbb{E}[Y_i]+\kappa\mathbb{E}\left[\widetilde{W}_{K-1}(\gamma')|\gamma^{\mathrm{r}},Z_i,A_i\right]\cdot\mathbb{E}[Y_i]}{f(Z_i)}\right\}$$

$$=\widetilde{W}_{K-1}(\gamma^{\mathrm{r}})+\frac{q(\gamma^{\mathrm{r}},Z^{(K-1)}(\gamma^{\mathrm{r}}),A^{(K-1)}(\gamma^{\mathrm{r}}))-\kappa\widetilde{W}_{K-1}(\gamma^{\mathrm{r}})\cdot\mathbb{E}[Y_i]+\kappa\mathbb{E}[Y_i]\cdot\left[\mathbf{P}_{A^{(K-1)}(\gamma^{\mathrm{r}}),Z^{(K-1)}(\gamma^{\mathrm{r}})}\right]_{N(\gamma^{\mathrm{r}}),:}\widetilde{\mathbf{W}}_{K-1}}{f(Z^{(K-1)}(\gamma^{\mathrm{r}}))} \tag{203a}$$

$$\leq\widetilde{W}_{K-1}(\gamma^{\mathrm{r}})+\frac{q(\gamma^{\mathrm{r}},Z^{(K)}(\gamma^{\mathrm{r}}),A^{(K)}(\gamma^{\mathrm{r}}))-\kappa\widetilde{W}_{K-1}(\gamma^{\mathrm{r}})\cdot\mathbb{E}[Y_i]+\kappa\mathbb{E}[Y_i]\cdot\left[\mathbf{P}_{A^{(K)}(\gamma^{\mathrm{r}}),Z^{(K)}(\gamma^{\mathrm{r}})}\right]_{N(\gamma^{\mathrm{r}}),:}\widetilde{\mathbf{W}}_{K-1}}{f(Z^{(K-1)}(\gamma^{\mathrm{r}}))}. \tag{203b}$$

The inequality (200a) is derived from the *triangle inequality*, while (200b) follows directly from (199b). Taking the limit $T\to\infty$ yields:

$$\left|\widetilde{W}_K(\gamma)-\widetilde{W}_\infty(\gamma)\right|$$

$$\leq\lim_{T\to\infty}\frac{M(1-\epsilon)^{K/L}(1-(1-\epsilon)^{T/L})}{1-(1-\epsilon)^{1/L}} \tag{201a}$$

$$=\frac{M(1-\epsilon)^{K/L}}{1-(1-\epsilon)^{1/L}}. \tag{201b}$$

This final result demonstrates that $\widetilde{W}_K(\gamma)$ will converge to a bounded value $\widetilde{W}_\infty(\gamma)$.

We now establish that ρ_k is also a *Cauchy sequence* and consequently converges to a bounded value ρ_∞. Through equation (43b) and the formal definition of $(A^{(K)}(\gamma),Z^{(K)}(\gamma))$ provided in (173), we can derive the bound presented in (202), which is shown at the top of the next page. The establishment of (202a) follows from recasting relation (43b) into its vector form, where $[\mathbf{P}_{a,z}]_{N(\gamma^{\mathrm{r}}),:}$ denotes the row vector consisted of $\{p_{\gamma^{\mathrm{r}}\gamma'}(a,z)\}_{\gamma'\in\mathcal{S}\times\mathcal{Y}\times\mathcal{A}}$ and $N(\gamma^{\mathrm{r}})$ denotes the index of the reference state γ^{r}. Subsequently, (202b) is established through direct application of the definition of $(A^{(K)}(\gamma),Z^{(K)}(\gamma))$ as given in (173). Iteratively, we can also establish the bound on ρ_k as shown in (203), presented at the top of the next page. Subtracting (203a) from (202b) yields the upper bound of $\rho_{K+1}-\rho_K$ in (204), where (204b) is established by re-writing the righ-hand side of (204a) into a vector form and (204c) holds directly from (181). Similarly, by subtracting (203b) from (202a), we can establish the lower bound of $\rho_{K+1}-\rho_K$ in (205) in next page. Because $a\leq b\leq c\Rightarrow|b|\leq\max\{|a|,|c|\}$, combing (204) and (205) yields (206) in the next page, where the first and second terms in the maximum operator satisfy the following inequalities:

$$\left[\mathbf{P}_\kappa\widetilde{(K-1)}\right]_{N(\gamma^{\mathrm{r}}),:}\left|\widetilde{\mathbf{W}}_K-\widetilde{\mathbf{W}}_{K-1}\right|$$

$$\leq\left[\mathbf{P}_\kappa\widetilde{(K-1)}\right]_{N(\gamma^{\mathrm{r}}),:}M(1-\epsilon)^{\frac{K-1}{L}}\mathbf{e} \tag{207a}$$

$$=M(1-\epsilon)^{\frac{K-1}{L}}, \tag{207b}$$

$$\left[\widetilde{\mathbf{P}_\kappa(K)}\right]_{N(\gamma^{\mathrm{r}}),:}\left|\widetilde{\mathbf{W}}_K-\widetilde{\mathbf{W}}_K\right|$$

$$\leq\left[\mathbf{P}_\kappa\widetilde{(K-1)}\right]_{N(\gamma^{\mathrm{r}}),:}M(1-\epsilon)^{\frac{K-1}{L}}\mathbf{e} \tag{208a}$$

$$=M(1-\epsilon)^{\frac{K-1}{L}}, \tag{208b}$$

where (207a) and (208a) are established from the vector form of (199). (207b) and (208b) are established as the matrix $\widetilde{\mathbf{P}_\kappa(k)}$ is a stochastic matrix for $\forall K\geq1$:

$$\sum_{\gamma'}\left[\widetilde{\mathbf{P}_\kappa(k)}\right]_{\gamma\times\gamma'}$$

$$\rho_{K+1}-\rho_K \leq \frac{\left(f(Z^{K-1}(\gamma^{\mathrm{r}}))-\kappa\cdot\mathbb{E}[Y_i]\right)\cdot\left(\widetilde{W}_K(\gamma^{\mathrm{r}})-\widetilde{W}_{K-1}(\gamma^{\mathrm{r}})\right)+\kappa\mathbb{E}[Y_i]\cdot\left[\mathbf{P}_{A^{(K-1)}(\gamma^{\mathrm{r}}),Z^{(K-1)}(\gamma^{\mathrm{r}})}\right]_{N(\gamma^{\mathrm{r}}),:}\left(\widetilde{\mathbf{W}_K}-\widetilde{\mathbf{W}_{K-1}}\right)}{f(Z^{K-1}(\gamma^{\mathrm{r}}))} \tag{204a}$$

$$=\left[\mathbf{I}-\mathrm{diag}\left(\frac{\kappa\mathbb{E}[Y_i]}{\mathbf{f}_{K-1}}\right)+\mathbf{P}(K-1)\mathrm{diag}\left(\frac{\kappa\mathbb{E}[Y_i]}{\mathbf{f}_{K-1}}\right)\right]_{N(\gamma^{\mathrm{r}}),:}\times\left(\widetilde{\mathbf{W}_K}-\widetilde{\mathbf{W}_{K-1}}\right) \tag{204b}$$

$$=\left[\widetilde{\mathbf{P}_\kappa(K-1)}\right]_{N(\gamma^{\mathrm{r}}),:}\left(\widetilde{\mathbf{W}}_K-\widetilde{\mathbf{W}}_{K-1}\right). \tag{204c}$$

$$\rho_{K+1}-\rho_K \geq \frac{\left(f(Z^{K}(\gamma^{\mathrm{r}}))-\kappa\cdot\mathbb{E}[Y_i]\right)\cdot\left(\widetilde{W}_K(\gamma^{\mathrm{r}})-\widetilde{W}_K(\gamma^{\mathrm{r}})\right)+\kappa\mathbb{E}[Y_i]\cdot\left[\mathbf{P}_{A^{(K)}(\gamma^{\mathrm{r}}),Z^{(K)}(\gamma^{\mathrm{r}})}\right]_{N(\gamma^{\mathrm{r}}),:}\left(\widetilde{\mathbf{W}_K}-\widetilde{\mathbf{W}_{K-1}}\right)}{f(Z^{K}(\gamma^{\mathrm{r}}))} \tag{205a}$$

$$=\left[\mathbf{I}-\mathrm{diag}\left(\frac{\kappa\mathbb{E}[Y_i]}{\mathbf{f}_{K}}\right)+\mathbf{P}(K)\mathrm{diag}\left(\frac{\kappa\mathbb{E}[Y_i]}{\mathbf{f}_{K}}\right)\right]_{N(\gamma^{\mathrm{r}}),:}\times\left(\widetilde{\mathbf{W}_K}-\widetilde{\mathbf{W}_{K-1}}\right) \tag{205b}$$

$$=\left[\widetilde{\mathbf{P}_\kappa(K)}\right]_{N(\gamma^{\mathrm{r}}),:}\left(\widetilde{\mathbf{W}}_K-\widetilde{\mathbf{W}}_{K-1}\right). \tag{205c}$$

$$\begin{aligned}|\rho_{K+1}-\rho_K| &\leq \max\left\{\left|\left[\widetilde{\mathbf{P}_\kappa(K-1)}\right]_{N(\gamma^{\mathrm{r}}),:}\left(\widetilde{\mathbf{W}}_K-\widetilde{\mathbf{W}}_{K-1}\right)\right|,\left|\left[\widetilde{\mathbf{P}_\kappa(K)}\right]_{N(\gamma^{\mathrm{r}}),:}\left(\widetilde{\mathbf{W}}_K-\widetilde{\mathbf{W}}_{K-1}\right)\right|\right\}\\ &\leq \max\left\{\left[\widetilde{\mathbf{P}_\kappa(K-1)}\right]_{N(\gamma^{\mathrm{r}}),:}\left|\widetilde{\mathbf{W}}_K-\widetilde{\mathbf{W}}_{K-1}\right|,\left[\widetilde{\mathbf{P}_\kappa(K)}\right]_{N(\gamma^{\mathrm{r}}),:}\left|\widetilde{\mathbf{W}}_K-\widetilde{\mathbf{W}}_{K-1}\right|\right\}.\end{aligned} \tag{206}$$

$$\begin{aligned}&=\sum_{\gamma'}\left[\mathbf{I}-\mathrm{diag}\left(\frac{\kappa\mathbb{E}[Y_i]}{\mathbf{f}_k}\right)+\mathbf{P}(k)\mathrm{diag}\left(\frac{\kappa\mathbb{E}[Y_i]}{\mathbf{f}_k}\right)\right]_{\gamma\times\gamma'}\\ &=1-\frac{\kappa\mathbb{E}[Y_i]}{f\left(Z^{(k)}(\gamma)\right)}+\frac{\kappa\mathbb{E}[Y_i]}{f\left(Z^{(k)}(\gamma)\right)}\sum_{\gamma'}\left[\mathbf{P}(k)\right]_{\gamma\times\gamma'}=1.\end{aligned} \tag{209}$$

Substituting (207) and (208) into (206) yields:

$$|\rho_{K+1}-\rho_K|\leq M(1-\epsilon)^{\frac{K-1}{L}}, \tag{210}$$

indicating that the sequence $\{\rho_K\}_{K\in\mathbb{N}^+}$ forms a *Cauchy sequence*. Specifically, for $\forall T>1$:

$$|\rho_{K+T}-\rho_K|\leq\sum_{t=0}^{T-1}|\rho_{K+t+1}-\rho_{K+t}| \tag{211a}$$

$$\leq\tau M\sum_{t=0}^{T-1}(1-\epsilon)^{\frac{K+t-1}{L}} \tag{211b}$$

$$=\frac{\tau M(1-\epsilon)^{(K-1)/L}(1-(1-\epsilon)^{T/L})}{1-(1-\epsilon)^{1/L}}, \tag{211c}$$

where (211a) is established from the *triangle inequality* and inequality (211b) holds from (210). Taking the limit $T\to\infty$:

$$\begin{aligned}|\rho_K-\rho_\infty|&\leq\lim_{T\to\infty}\frac{\tau M(1-\epsilon)^{(K-1)/L}(1-(1-\epsilon)^{T/L})}{1-(1-\epsilon)^{1/L}}\\ &=\frac{M(1-\epsilon)^{K/L}}{1-(1-\epsilon)^{1/L}}.\end{aligned} \tag{212}$$

This indicates that ρ_K will converge to a bounded value ρ_∞.

2) Case 2: $\gamma^\star\neq\gamma^r$

The primary objective of this subsection is to demonstrate that the conclusion under $\gamma^\star=\gamma^{\mathrm{r}}$ remains valid even when if $\gamma^\star\neq\gamma^{\mathrm{r}}$. To prove this, we introduce an auxiliary iteration sequence in the following:

Iteration 5. *(Auxiliary Iteration Sequence): For a given $0<\kappa<1$, the auxiliary iteration generates sequence $\{\overline{\rho_K}\}_{K\in\mathbb{N}^+}$ and $\{\overline{W}_K(\gamma)\}_{\gamma\in\mathcal{S}\times\mathcal{Y}\times\mathcal{A}}^{K\in\mathbb{N}^+}$ with a starting initial value $\{\overline{W}_0(\gamma)\}_{\gamma\in\mathcal{S}\times\mathcal{Y}\times\mathcal{A}}$ by (213), where $\gamma^r\in\mathcal{S}\times\mathcal{Y}\times\mathcal{A}$ is a fixed reference state with an initial condition $\overline{W}_0(\gamma^r)=0$ and $\overline{W}_0(\gamma)=\widetilde{W}_0(\gamma)$ for $\forall\gamma\in\mathcal{S}\times\mathcal{Y}\times\mathcal{A}$.*

A key property of the auxiliary iteration sequence is that there exists an $M>0$ such that for all $K\geq1$,

$$\begin{aligned}&\max_\gamma\left\{\overline{W}_{K+1}(\gamma)-\overline{W}_K(\gamma)\right\}-\\ &\min_\gamma\left\{\overline{W}_{K+1}(\gamma)-\overline{W}_K(\gamma)\right\}\leq M(1-\epsilon)^{K/L}.\end{aligned} \tag{214}$$

Compare (213) with (43), the relationship between $\overline{W}_K(\gamma)$

$$\overline{\rho_{K+1}} = \overline{W}_K(\gamma^\star) + \min_{A_i,Z_i}\left\{\frac{q(\gamma^\star,Z_i,A_i) - \kappa\overline{W}_K(\gamma^\star)\cdot\mathbb{E}[Y_i] + \kappa\mathbb{E}\left[\overline{W}_K(\gamma')|\gamma^\star,Z_i,A_i\right]\cdot\mathbb{E}[Y_i]}{f(Z_i)}\right\}, \tag{213a}$$

$$\overline{W}_{K+1}(\gamma) = \overline{W}_K(\gamma) + \min_{A_i,Z_i}\left\{\frac{q(\gamma,Z_i,A_i) - \kappa\overline{W}_K(\gamma)\cdot\mathbb{E}[Y_i] + \kappa\mathbb{E}\left[\overline{W}_K(\gamma')|\gamma,Z_i,A_i\right]\cdot\mathbb{E}[Y_i]}{f(Z_i)}\right\} - \overline{\rho_{K+1}},$$
$$\gamma\in\mathcal{S}\times\mathcal{Y}\times\mathcal{A}, \tag{213b}$$

$$\begin{aligned}\Phi(\overline{\mathbf{W}}_{K-1}) =& \overline{W}_{K-1}(\gamma^\star) + \min_{A_i,Z_i}\left\{\frac{q(\gamma^\star,Z_i,A_i) - \kappa\overline{W}_{K-1}(\gamma^\star)\cdot\mathbb{E}[Y_i] + \kappa\mathbb{E}\left[\overline{W}_{K-1}(\gamma')|\gamma^\star,Z_i,A_i\right]\cdot\mathbb{E}[Y_i]}{f(Z_i)}\right\}\\ &-\overline{W}_{K-1}(\gamma^{\mathrm{r}}) - \min_{A_i,Z_i}\left\{\frac{q(\gamma^{\mathrm{r}},Z_i,A_i) - \kappa\overline{W}_{K-1}(\gamma^{\mathrm{r}})\cdot\mathbb{E}[Y_i] + \kappa\mathbb{E}\left[\overline{W}_{K-1}(\gamma')|\gamma^{\mathrm{r}},Z_i,A_i\right]\cdot\mathbb{E}[Y_i]}{f(Z_i)}\right\}.\end{aligned} \tag{216}$$

$$\begin{aligned}&\max_\gamma\left\{\widetilde{W}_{K+1}(\gamma) - \widetilde{W}_K(\gamma)\right\} - \min_{\gamma'}\left\{\widetilde{W}_{K+1}(\gamma) - \widetilde{W}_K(\gamma)\right\}\\ &= \max_\gamma\left\{\overline{W}_{K+1}(\gamma) - \overline{W}_K(\gamma)\right\} - \min_{\gamma'}\{\overline{W}_{K+1}(\gamma) - \overline{W}_K(\gamma).\end{aligned} \tag{220}$$

and $\widetilde{W}_K(\gamma)$ is established by:

$$\widetilde{W}_K(\gamma) = \overline{W}_K(\gamma) + \Phi(\overline{\mathbf{W}}_{K-1}), \tag{215}$$

where $\overline{\mathbf{W}}_K$ is a column vector consisted of $\overline{W}_K(\gamma)$ for all $\gamma\in\mathcal{S}\times\mathcal{Y}\times\mathcal{A}$ and $\Phi(\overline{\mathbf{W}}_{K-1})$ is given in (216) at the top of the next page. Meanwhile, we can establish

$$\widetilde{W}_{K+1}(\gamma) = \overline{W}_{K+1}(\gamma) + \Phi(\overline{\mathbf{W}}_K). \tag{217}$$

Subtracting (215) from (217) yields:

$$\begin{aligned}&\widetilde{W}_{K+1}(\gamma) - \widetilde{W}_K(\gamma) =\\ &\overline{W}_{K+1}(\gamma) - \overline{W}_K(\gamma) + \Phi(\overline{\mathbf{W}}_K) - \Phi(\overline{\mathbf{W}}_{K-1}),\end{aligned} \tag{218}$$

Thus, by applying the max and min operators to both sides of (218), we obtain the following:

$$\begin{aligned}\max_\gamma\left\{\widetilde{W}_{K+1}(\gamma) - \widetilde{W}_K(\gamma)\right\} =& \max_\gamma\left\{\overline{W}_{K+1}(\gamma) - \overline{W}_K(\gamma)\right\}\\ &+ \Phi(\overline{\mathbf{W}}_K) - \Phi(\widetilde{\mathbf{W}}_{K-1}),\end{aligned} \tag{219a}$$

$$\begin{aligned}\min_\gamma\left\{\widetilde{W}_{K+1}(\gamma) - \widetilde{W}_K(\gamma)\right\} =& \min_\gamma\left\{\overline{W}_{K+1}(\gamma) - \overline{W}_K(\gamma)\right\}\\ &+ \Phi(\overline{\mathbf{W}}_K) - \Phi(\widetilde{\mathbf{W}}_{K-1}),\end{aligned} \tag{219b}$$

Subtracting (219b) from (219a) yields the key equality in (220), which is shown at the top of the next page. Substituting (214) into (220), we have that the original sequence $\{\widetilde{W}_K(\gamma)\}_{K\in\mathbb{N}^+}$ is upper bounded by:

$$\begin{aligned}&\max_\gamma\left\{\widetilde{W}_{K+1}(\gamma) - \widetilde{W}_K(\gamma)\right\}\\ &- \min_\gamma\left\{\widetilde{W}_{K+1}(\gamma) - \widetilde{W}_K(\gamma)\right\}\\ &\le M(1-\epsilon)^{K/L}.\end{aligned} \tag{221}$$

With (221) in hand, it follows that $\{\widetilde{W}_K(\gamma)\}_{K\in\mathbb{N}^+}$ and $\{\rho_K(\lambda)\}_{K\in\mathbb{N}^+}$ are both *Cauchy sequences*, following a reasoning similar to that in *Case 1*: (186)-(212).

B. Convergence Direction

In this subsection, we establish the relationship between the convergent values and the solution to (38). As $\widetilde{W}_K(\gamma)$ and ρ_K are convergent, we have

$$\lim_{K\to\infty}\rho_K(\lambda) = \lim_{K\to\infty}\rho_{K+1}(\lambda) = \rho_\infty(\lambda), \tag{222a}$$

$$\lim_{K\to\infty}\widetilde{W}_{K+1}(\gamma) = \lim_{K\to\infty}\widetilde{W}_K(\gamma) = \widetilde{W}_\infty(\gamma), \forall\ \gamma, \tag{222b}$$

Substituting (222a) and (222b) into (43) yields (223), presented at the top of this page.

Since $f(Z_i) > 0$, from (223b) we have that

$$\begin{aligned}0 = \min_{A_i,Z_i}\Big\{&g(\gamma,Z_i,A_i;\rho_\infty) +\\ &\kappa\mathbb{E}[Y_i]\sum_{\gamma'}p(\gamma'|\gamma,Z_i,A_i)\widetilde{W}_\infty(\gamma') - \kappa\mathbb{E}[Y_i]\widetilde{W}_\infty(\gamma)\Big\},\\ &\forall\gamma\in\mathcal{S}\times\mathcal{Y}\times\mathcal{A},\end{aligned} \tag{224}$$

where $g(\gamma,Z_i,A_i;\rho_\infty) = q(\gamma,Z_i,A_i) - \rho_\infty\cdot f(Z_i)$. By moving the term $\kappa\mathbb{E}[Y_i]\cdot\widetilde{W}_\infty(\gamma)$ on the left-hand side of (224) to the left-hand side and rewriting the summation terms in (224) into an expectation form, we have

$$\begin{aligned}\kappa\mathbb{E}[Y_i]\cdot\widetilde{W}_\infty(\gamma) = \min_{A_i,Z_i}\Big\{&g(\gamma,Z_i,A_i;\rho_\infty)\\ &+\mathbb{E}\left[\kappa\mathbb{E}[Y_i]\cdot\widetilde{W}_\infty(\gamma')|\gamma,Z_i,A_i\right]\Big\}, \forall\gamma\in\mathcal{S}\times\mathcal{Y}\times\mathcal{A}.\end{aligned} \tag{225}$$

Meanwhile, substituting (187) into (223a) and rewriting the

$$\rho_\infty = \widetilde{W}_\infty(\gamma^{\mathrm{r}}) + \min_{A_i,Z_i}\left\{\frac{q(\gamma^{\mathrm{r}},Z_i,A_i) - \kappa\widetilde{W}_\infty(\gamma^{\mathrm{r}})\cdot\mathbb{E}[Y_i] + \kappa\mathbb{E}[Y_i]\sum_{\gamma'}p(\gamma'|\gamma^{\mathrm{r}},Z_i,A_i)\cdot\widetilde{W}_\infty(\gamma')}{f(Z_i)}\right\}, \tag{223a}$$

$$0 = \min_{A_i,Z_i}\left\{\frac{q(\gamma,Z_i,A_i) - \rho_\infty\cdot f(Z_i) + \kappa\mathbb{E}[Y_i]\sum_{\gamma'}p(\gamma'|\gamma,Z_i,A_i)\cdot\widetilde{W}_\infty(\gamma') - \kappa\widetilde{W}_\infty(\gamma)\cdot\mathbb{E}[Y_i]}{f(Z_i)}\right\}, \quad \forall\gamma\in\mathcal{S}\times\mathcal{Y}\times\mathcal{A}, \tag{223b}$$

summation terms in (223a) yields:

$$\rho_\infty = \min_{A_i,Z_i}\left\{\frac{q(\gamma^{\mathrm{r}},Z_i,A_i) + \mathbb{E}\left[\kappa\mathbb{E}[Y_i]\cdot\widetilde{W}_\infty(\gamma')|\gamma,Z_i,A_i\right]}{f(Z_i)}\right\}, \tag{226}$$

Compare (225) and (226) with (38), we have that $\left(\rho_\infty, \left\{\kappa\mathbb{E}[Y_i]\cdot\widetilde{W}_\infty(\gamma)\right\}_{\gamma\in\mathcal{S}\times\mathcal{Y}\times\mathcal{A}}\right)$ is a root of (38). We thus establish the equivalence:

$$\begin{aligned}\rho^\star &= \rho_\infty,\\ \kappa\mathbb{E}[Y_i]\cdot\widetilde{W}_\infty(\gamma) &= \widetilde{W}^\star(\gamma), \forall\gamma\in\mathcal{S}\times\mathcal{Y}\times\mathcal{A},\end{aligned} \tag{227}$$

Substituting (227) into (212) completes the proof of Theorem 6. Moreover, substituting (227) into (201) and (212) and letting $K\to\infty$, we obtain

$$\begin{aligned}0 \le \lim_{K\to\infty}|\rho^K - \rho^\star| &\le \lim_{K\to\infty}\frac{M(1-\epsilon)^{K/L}}{1-(1-\epsilon)^{1/L}} = 0,\\ 0 \le \lim_{K\to\infty}\left|\widetilde{W}_K(\gamma) - \frac{\widetilde{W}^\star(\gamma)}{\kappa\mathbb{E}[Y_i]}\right| &\le \lim_{K\to\infty}\frac{M(1-\epsilon)^{K/L}}{1-(1-\epsilon)^{1/L}} = 0.\end{aligned} \tag{228}$$

Therefore, by the squeeze theorem, both limits are zero, which completes the proof of Theorem 5.

Appendix K
Proof of Lemma 8

By substituting the definitions of $\mathcal{Q}^{\lambda+\theta}$ and $\mathcal{F}^{\lambda+\theta}$ from (53) and (54) into (49) and (50), we express $\Upsilon(\theta,\lambda;f_{\max})$ as:

$$\begin{aligned}\Upsilon(\theta,\lambda;f_{\max}) &= \mathcal{Q}^{\lambda+\theta} - (\lambda+\theta)\mathcal{F}^{\lambda+\theta} + \frac{\theta}{f_{\max}}\\ &= \mathcal{Q}^{\lambda+\theta} - (\lambda+\theta)(\mathcal{F}^{\lambda+\theta} - \frac{1}{f_{\max}}) - \frac{\lambda}{f_{\max}}.\end{aligned} \tag{229}$$

We now demonstrate that for a fixed value of λ, the function $\Upsilon(\theta,\lambda;f_{\max})$ is non-increasing with respect to θ, provided that $\mathcal{F}^{\lambda+\theta}\ge 1/f_{\max}$. For any $\Delta\theta\ge 0$, we can establish (230) at the top of the next page, where (230a) follows from expanding $\mathcal{Q}^{\lambda+\theta}$ and $\mathcal{F}^{\lambda+\theta}$ according to (53) and (54); and inequality (230d) is established by the optimality of policy $\phi^\star_{\theta+\Delta\theta+\lambda}$:

$$\mathcal{L}(\phi^\star_{\theta+\Delta\theta+\lambda};\theta+\Delta\theta,\lambda,f_{\max}) \le \mathcal{L}(\phi^\star_{\theta+\lambda};\theta+\Delta\theta,\lambda,f_{\max}). \tag{231}$$

This completes the proof.

Appendix L
Proof of Corollary 1

A. Case 1: $\mathcal{F}^{\lambda^+}\ge 1/f_{\max}$

From part (ii) of Lemma 8, which indicates the non-decreasing property of $\mathcal{F}^{\lambda+\theta}$, it turns out that if $\mathcal{F}^{\lambda^+}\ge 1/f_{\max}$, then for any $\theta\ge 0$,

$$\mathcal{F}^{\lambda+\theta} \ge \mathcal{F}^{\lambda^+} \ge 1/f_{\max}. \tag{232}$$

From this, it follows that the inequality $\mathcal{F}^{\lambda+\theta}\ge\frac{1}{f_{\max}}$ is always true, satisfying the KKT condition (52c). Additionally, under this assumption, part (iv) of Lemma 8 ensures that $\Upsilon(\theta,\lambda;f_{\max})$ is non-increasing for $\theta\ge 0$. Hence, the minimum value of $\Upsilon(\theta,\lambda;f_{\max})$ occurs at $\theta=0$:

$$\min_{\theta\ge 0}\Upsilon(\theta,\lambda;f_{\max}) = \Upsilon(0,\lambda;f_{\max}) = U(\lambda). \tag{233}$$

Note that $\theta^\star_\lambda = 0$ also satisfies the KKT condition (52c), we accomplish the proof under this case.

B. Case 2: $\mathcal{F}^{\lambda^+} < 1/f_{\max}$

On the other hand, consider the case where $\mathcal{F}^{\lambda^+} < 1/f_{\max}$. According to part (ii) of Lemma 8, there exists a threshold value θ^{tr} such that (56) holds. When $\theta\in[0,\theta^{\mathrm{tr}})$, it follows from (56) that $\mathcal{F}^{\lambda+\theta}\le\frac{1}{f_{\max}}$, which contradicts the KKT condition (52c), indicating that values of θ in this range do not satisfy the necessary optimality conditions. However, for $\theta\in[\theta^{\mathrm{tr}},\infty)$, we observe from (56) that the KKT condition (52c) is satisfied, as $\mathcal{F}^{\lambda+\theta}\ge\frac{1}{f_{\max}}$. Furthermore, from part (iv) of Lemma 8, $\Upsilon(\theta,\lambda;f_{\max})$ is non-increasing with respect to θ in this feasible region. Therefore, the minimum value of $\Upsilon(\theta,\lambda;f_{\max})$, under the KKT conditions occurs at the smallest θ for which the KKT condition (52c) holds, namely $\theta^\star_\lambda = \theta^{\mathrm{tr}}$.

Appendix M
Proof of Theorem 7

A. Case 1: $\mathcal{F}^{\lambda^+}\ge 1/f_{\max}$

From part (i) of Corollary 1, we know that the optimal Lagrangian multiplier satisfies $\theta^\star_\lambda = 0$. Substituting $\theta^\star_\lambda = 0$ into (52b) yields

$$\phi^\star_\lambda = \arg\min_{\phi}\lim_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1}\mathbb{E}_\phi\left\{\sum_{t=D_i}^{D_{i+1}-1}\mathcal{C}\left(X_t,A_i\right) - \lambda\left(Z_i+Y_{i+1}\right)\right\}. \tag{234}$$

In this case, the optimal policy is equivalent to the optimal policy determined in (24), which is stationary deterministic.

$$\Upsilon(\theta+\Delta\theta,\lambda;f_{\max})-\Upsilon(\theta,\lambda;f_{\max})$$

$$\overset{(a)}{=}\frac{\Delta\theta}{f_{\max}}+\lim_{\mathrm{n}\to\infty}\frac{1}{n}\sum_{i=0}^{n-1}\mathbb{E}_{\phi^{\star}_{\theta+\Delta\theta+\lambda}}\left[\sum_{t=D_i}^{D_{i+1}-1}\mathcal{C}(X_t,A_i)\right]-(\theta+\Delta\theta+\lambda)\lim_{\mathrm{n}\to\infty}\frac{1}{n}\sum_{i=0}^{n-1}\mathbb{E}_{\phi^{\star}_{\theta+\Delta\theta+\lambda}}\left[Z_i+Y_{i+1}\right] \tag{230a}$$

$$-\left(\lim_{\mathrm{n}\to\infty}\frac{1}{n}\sum_{i=0}^{n-1}\mathbb{E}_{\phi^{\star}_{\theta+\lambda}}\left[\sum_{t=D_i}^{D_{i+1}-1}\mathcal{C}(X_t,A_i)\right]-(\theta+\lambda)\lim_{\mathrm{n}\to\infty}\frac{1}{n}\sum_{i=0}^{n-1}\mathbb{E}_{\phi^{\star}_{\theta+\lambda}}\left[Z_i+Y_{i+1}\right]\right)$$

$$\leq\frac{\Delta\theta}{f_{\max}}-\Delta\theta\lim_{\mathrm{n}\to\infty}\frac{1}{n}\sum_{i=0}^{n-1}\mathbb{E}_{\phi^{\star}_{\theta+\lambda}}\left[Z_i+Y_{i+1}\right] \tag{230b}$$

$$=-\Delta\theta\left(\mathcal{F}^{\lambda+\theta}-\frac{1}{f_{\max}}\right) \tag{230c}$$

$$\leq 0, \tag{230d}$$

B. Case 2: $\mathcal{F}^{\lambda^+}<1/f_{\max}$ and $\mathcal{F}^{(\lambda+\theta^{\star}_{\lambda})^+}=1/f_{\max}$

As $\mathcal{F}^{\lambda^+}<1/f_{\max}$, we can establish from Corollary 1 that $\theta^{\star}_{\lambda}>0$. The condition $\mathcal{F}^{(\lambda+\theta^{\star}_{\lambda})^+}=1/f_{\max}$ leads to the fact that (52c) and (52d) are satisfied naturally, and thus the optimal policy is exactly the stationary deterministic policy $\phi^{\star}_{\lambda+\theta^{\star}_{\lambda}}$.

C. Case 3: $\mathcal{F}^{\lambda^+}<1/f_{\max}$ and $\mathcal{F}^{(\lambda+\theta^{\star}_{\lambda})^+}>1/f_{\max}$

(iii). For this case, it is proved in [111, Theorem 4.4] that the policy given in (57) and (58) is the optimal policy.

Appendix N
Proof of Theorem 8

A. Proof of Part (i)

Substituting (49), (53) and (54) into (50), we have

$$\Upsilon(\theta,\lambda;f_{\max})=\mathcal{Q}^{\lambda+\theta}-(\lambda+\theta)\mathcal{F}^{\lambda+\theta}+\frac{\theta}{f_{\max}}. \tag{235}$$

Let $\lambda=h^{\star}$ and $\rho=h^{\star}+\theta$, (235) turns to

$$\Upsilon(\rho-h^{\star},h^{\star};f_{\max})=\mathcal{Q}^{\rho}-\rho\mathcal{F}^{\rho}+\frac{\rho-h^{\star}}{f_{\max}}. \tag{236}$$

Note that $d(h^{\star};f_{\max})=U(h^{\star};f_{\max})=0$, the Problem 6 turns to the following equation:

$$0=\max_{\rho\geq h^{\star}}\left\{\underbrace{\mathcal{Q}^{\rho}-\rho\left(\mathcal{F}^{\rho}-\frac{1}{f_{\max}}\right)-\frac{h^{\star}}{f_{\max}}}_{\mathcal{J}(\rho)}\right\}. \tag{237}$$

Similar to the proof of part (iv) of Lemma 8, we can establish the following Lemma.

Lemma 10. *If $\mathcal{F}^{(\rho)^-}\geq 1/f_{\max}$, then $\mathcal{J}(\rho)$ is non-increasing with respect to ρ.*

We next employ a proof by contradiction to establish this result. Specifically, we consider the following three cases:

- *Case 1*: If $h^{\star}<\rho^{\star}$, we know that

$$\mathcal{J}(\rho^{\star})=\frac{\rho^{\star}-h^{\star}}{f_{\max}}>0. \tag{238}$$

This contradicts (237), as (237) is equivalent to:

$$\mathcal{J}(\rho)\leq 0,\forall\rho\geq h^{\star}. \tag{239}$$

- *Case 2*: If $h^{\star}=\rho^{\star}$, we have that $\mathcal{F}^{(h^{\star})^-}\geq 1/f_{\max}$. Since $\mathcal{F}^{\lambda}$ is non-decreasing with λ, as indicated in Lemma 8, we know that

$$\mathcal{F}^{(\rho)^-}\geq\frac{1}{f_{\max}},\forall\rho\geq h^{\star}, \tag{240}$$

which indicates that $\mathcal{J}(\rho)$ is non-increasing with respect to ρ, as shown in Lemma 10. As a result, we have that

$$\begin{aligned}0=\max_{\rho\geq h^{\star}}\{\mathcal{J}(\rho)\}&=\mathcal{J}(h^{\star})\\&=\mathcal{Q}^{h^{\star}}-h^{\star}\mathcal{F}^{h^{\star}}=U(h^{\star}).\end{aligned} \tag{241}$$

This establishes that $U(h^{\star})=0$. Since the root of $U(\lambda)$ is unique, as detailed in part (iii) of Lemma 2, it is sufficient to verify that $\rho^{\star}=h^{\star}$.

- *Case 3*: If $h^{\star}>\rho^{\star}$, we have that $\mathcal{F}^{(h^{\star})^-}\geq\mathcal{F}^{(\rho^{\star})^-}\geq 1/f_{\max}$ as $\mathcal{F}^{\lambda}$ is non-decreasing with λ (part (ii) of Lemma 8). Similar to *Case 2*, we know that $\mathcal{J}(\rho)$ is non-increasing with respect to ρ and thus $U(h^{\star})=0$. Since the root of $U(\lambda)$ is unique (part (iii) of Lemma 2), we have $\rho^{\star}=h^{\star}$. This contradicts the case where $h^{\star}>\rho^{\star}$.

As such, we establish that $h^{\star}=\rho^{\star}$.

B. Proof of Part (ii)

If $\mathcal{F}^{(\rho^{\star})^-}<1/f_{\max}$, *Case 1* in the proof of part (i) remains contradictory and thus $\rho^{\star}\geq h^{\star}$. Since $\mathcal{F}^{\lambda}$ is non-decreasing with λ, as indicated in Lemma 8, we know that

$$\mathcal{F}^{(h^{\star})^-}\leq\mathcal{F}^{(\rho^{\star})^-}<1/f_{\max}. \tag{242}$$

Lemma 11. *$\mathcal{J}(\rho)$ and $U(\rho)$ is uniformly absolutely continuous, with the derivative given as:*

$$\frac{\mathrm{d}U(\rho)}{\mathrm{d}\rho}=-\mathcal{F}^{\rho}, \tag{243a}$$

$$\frac{\mathrm{d}\mathcal{J}(\rho)}{\mathrm{d}\rho}=-\mathcal{F}^{\rho}+\frac{1}{f_{\max}}. \tag{243b}$$

Proof. (243a) is a restatement of [105, Lemma 3.1]. (243b) is a corollary of (243a). ■

As a result, the maximum value of $\mathcal{J}(\rho)$ is taken where $\mathcal{F}^{\rho} = \frac{1}{f_{\max}}$. We denote one of the root as

$$\rho^{\downarrow} = \inf\left\{\rho \geq h^{\star} : \mathcal{F}^{\rho} = \frac{1}{f_{\max}}\right\}. \tag{244}$$

The root $\rho^{\downarrow}$ always exists since:

$$\left.\frac{\mathrm{d}\mathcal{J}(\rho)}{\mathrm{d}\rho}\right|_{\rho=(h^{\star})^{-}} < 0, \tag{245}$$

and the monotonic property of $\mathcal{F}(\rho)$. As a result, the Problem 6 turns to the following problem with a stricter equation constraint:

Problem 9 (*CMDP with equation constraint*)**.**

$$\begin{aligned} &H(\lambda; f_{\max}) \triangleq \\ &\inf_{\phi} \lim_{\mathrm{n}\to\infty} \frac{1}{n}\sum_{i=0}^{n-1}\left\{\mathbb{E}_{\psi}\left[\sum_{t=D_i}^{D_{i+1}-1}\mathcal{C}(X_t, A_i)\right] - \lambda\mathbb{E}\left[Z_i + Y_{i+1}\right]\right\} \\ &s.t. \lim_{T\to\infty}\frac{1}{T}\mathbb{E}_{\phi}\left[\sum_{i=1}^{T}(Y_{i+1}+Z_i)\right] = \frac{1}{f_{max}}. \end{aligned} \tag{246}$$

As did in [92, Chapter 8.8], we can utilize the *dual relaxation* to reformulate Problem 9 as the following LP problem:

Problem 10 (*LP Transformation of Problem 9*)**.**

$$\begin{aligned} H(\lambda; f_{\max}) = \min_{\mathbf{x}} &\sum_{\gamma,z,a}(q(\gamma,z,a) - \lambda f(z))x(\gamma,z,a) \\ \text{s.t.} \quad &(61\text{b}) - (61\text{e}) \end{aligned} \tag{247}$$

Our remaining focus is to explicitly express the root of $H(\lambda; f_{\max})$ in Problem 10. By substituting the constraint (61b) into the objective function of Problem 10, this problem is decomposed as:

$$\begin{aligned} H(\lambda; f_{\max}) = \min_{\mathbf{x}} &\sum_{\gamma,z,a} q(\gamma,z,a)x(\gamma,z,a) - \frac{\lambda}{f_{\max}} \\ \text{s.t.} \quad &(61\text{b}) - (61\text{e}) \end{aligned} \tag{248}$$

Since $\lambda/f_{\max}$ is independent of the decision variable $\mathbf{x}$, let $\lambda = h^{\star}$ and $H(h^{\star}; f_{\max}) = 0$, we establish the following LP program:

Problem 11 (*Closed-form Root Solution*)**.**

$$\begin{aligned} \frac{h^{\star}}{f_{\max}} = \min_{\mathbf{x}} &\sum_{\gamma,z,a} q(\gamma,z,a)x(\gamma,z,a) \\ \text{s.t.} \quad &(61\text{b}) - (61\text{e}) \end{aligned} \tag{249}$$

Comparing Problem 11 with Problem 7 establishes the relationship $h^{\star} = f_{\max}\cdot Q^{\star}(f_{\max})$.

Appendix O
Proof of Theorem 9

A. Part (i)

Part (i) holds naturally since $h^{\star} = \rho^{\star}$ when $\mathcal{F}^{(\rho^{\star})^{-}} \geq 1/f_{\max}$. In this case, as discussed in Section IV and Section V, $\rho^{\star}$ corresponds to the optimal value without a sampling frequency constraint. Consequently, $h^{\star}$ is independent of $f_{\max}$.

B. Part (ii)

When $\mathcal{F}^{(\rho^{\star})^{-}} < 1/f_{\max}$, we have that $h^{\star} = f_{\max}\cdot Q^{\star}(f_{\max})$. To analyze the relationship between $h^{\star}$ and $f_{\max}$, we need to conduct a comprehensive *sensitivity analysis* on the LP problem specified in Problem 7. To facilitate the *sensitivity analysis*, we first rewrite $h^{\star}$ as follows:

Problem 12 (*Reformulation of $h^{\star}$*)**.**

$$h^{\star} = f_{\max}Q^{\star}(f_{\max}) \tag{250a}$$

$$= \min_{\mathbf{x}} \sum_{\gamma,z,a} f_{\max}q(\gamma,z,a)x(\gamma,z,a) \tag{250b}$$

$$\text{s.t.} \sum_{\gamma,z,a} f_{\max}f(z)x(\gamma,z,a) = 1, \tag{250c}$$

$$(61\text{b}) - (61\text{e}) \tag{250d}$$

For easy notations, we define the feasible region that satisfies constraint (61b)-(61e) as $\mathcal{D} = \{\mathbf{x} : (61\text{b}) - (61\text{e})\}$. Our goal is to conduct a sensitivity analysis on the new LP problem in Problem 12, where the parameter affects both the objective (250b) and subjective (250c). By applying the Lagrangian dual technique to Problem 11, we establish the partial Lagrangian dual problem:

Problem 13 (*Lagrangian Dual of Problem 12*)**.**

$$\Theta^{\star} = \max_{\lambda}\left\{\lambda + f_{\max}\min_{\mathbf{x}\in\mathcal{D}}\sum_{\gamma,z,a}(q(\gamma,z,a) - \lambda f(z))x(\gamma,z,a)\right\}. \tag{251}$$

The *weak duality* principle establishes that $\Theta^{\star} \leq h^{\star}$, while the strong duality holds that $\Theta^{\star} = h^{\star}$. For LP problems, the *Staler's constraint qualification* is always satisfied [104, Chapter 5.2.3], which naturally ensures *strong duality*. Thus, for our LP problem, we have $\Theta^{\star} = h^{\star}$. By applying the inverse transformation approach as demonstrated in the transformation from Problem 9 to Problem 10, we obtain:

$$U(\lambda) = \min_{\mathbf{x}\in\mathcal{D}}\sum_{\gamma,z,a}(q(\gamma,z,a) - \lambda f(z))x(\gamma,z,a). \tag{252}$$

This transformation, combined with strong duality, converts Problem 13 into:

$$h^{\star} = \max_{\lambda}\left\{\underbrace{\lambda + f_{\max}U(\lambda)}_{g(\lambda)}\right\} \tag{253}$$

To analyze the derivative of $g(\lambda)$, we utilize the following lemma, which is a restatement of [105, Lemma 3.1].

Lemma 12. *(Restatement [105, Lemma 3.1]). $U(\lambda)$ is uniformly absolutely continuous, with the derivative given as:*

$$\frac{\mathrm{d}U(\lambda)}{\mathrm{d}\lambda} = -\mathcal{F}^{\lambda}. \tag{254}$$

Consequently, the derivative of $g(\lambda)$ is:

$$\frac{\mathrm{d}g(\lambda)}{\mathrm{d}\lambda} = \frac{\mathrm{d}(\lambda + f_{\max}U(\lambda))}{\mathrm{d}\lambda} = 1 - f_{\max}\mathcal{F}^{\lambda}. \tag{255}$$

Setting $\frac{\mathrm{d}g(\lambda)}{\mathrm{d}\lambda} = 0$ and by utilizing the monotonically non-decreasing property of $\mathcal{F}^{\lambda}$ in terms of λ (as established in Lemma 8-(ii)), we have that the optimal $\lambda^{\star}$ satisfies:

$$\mathcal{F}^{(\lambda^{\star})^{-}} < \frac{1}{f_{\max}} \leq \mathcal{F}^{(\lambda^{\star})^{+}} \tag{256}$$

Given that $\mathcal{F}^{(\rho^{\star})^{-}} < 1/f_{\max}$, and considering the monotonically non-decreasing property of $\mathcal{F}^{\lambda}$ with respect to λ, we establish that

$$\lambda^{\star} \geq \rho^{\star}. \tag{257}$$

Thus, from the monotonically non-decreasing property of $U(\lambda)$ in terms of λ (as established in Lemma 8-(i)), we have that

$$U(\lambda^{\star}) \leq U(\rho^{\star}) = 0. \tag{258}$$

As last, we conduct the *sensitivity analysis* as follows:

$$\frac{\mathrm{d}h^{\star}}{\mathrm{d}f_{\max}} = \frac{\mathrm{d}(\lambda^{\star} + f_{\max}U(\lambda^{\star}))}{\mathrm{d}f_{\max}} = U(\lambda^{\star}) \leq 0, \tag{259}$$

which accomplished the part (ii). ■

Appendix P
Proof of Lemma 9

First, in (209), it has been proved that the sum of each row of $\widetilde{\mathbf{P}_{\kappa}(k)} \in \mathbb{R}^{|\mathcal{S}\times\mathcal{Y}\times\mathcal{A}|\times|\mathcal{S}\times\mathcal{Y}\times\mathcal{A}|}$ is 1. Define $p_{\gamma\gamma'}(Z^{(k)}(\gamma), A^{(k)}(\gamma))$ as the transition probability that γ transitions to state γ' under actions $(Z^{(k)}(\gamma), A^{(k)}(\gamma))$ in $\mathbf{P}(k)$, and $\widetilde{p_{\gamma\gamma'}}(Z^{(k)}(\gamma), A^{(k)}(\gamma))$ as the corresponding element in matrix $\widetilde{\mathbf{P}_{\kappa}(k)}$, we have that

$$\widetilde{p_{\gamma\gamma'}}(Z^{(k)}(\gamma), A^{(k)}(\gamma)) = \begin{cases} \frac{\kappa\mathbb{E}[Y_i]p_{\gamma\gamma'}(Z^{(k)}(\gamma), A^{(k)}(\gamma))}{f(Z^{(k)}(\gamma))} & \text{if } \gamma \neq \gamma', \\ 1 - \frac{\kappa\mathbb{E}[Y_i] - \kappa\mathbb{E}[Y_i]\cdot p_{\gamma\gamma}(Z^{(k)}(\gamma), A^{(k)}(\gamma))}{f(Z^{(k)}(\gamma))} & \text{if } \gamma = \gamma'. \end{cases} \tag{260}$$

Note that $0 < \kappa < 1$, $Z^{(k)}(\gamma) \geq 0$ and $0 \leq p_{\gamma\gamma'}(Z^{(k)}(\gamma), A^{(k)}(\gamma)) \leq 1$, we have

$$0 \leq \frac{\kappa\mathbb{E}[Y_i]p_{\gamma\gamma'}(Z^{(k)}(\gamma), A^{(k)}(\gamma))}{f(Z^{(k)}(\gamma))} < \frac{\mathbb{E}[Y_i]}{\mathbb{E}[Y_i] + Z^{(k)}(\gamma)} \leq 1, \tag{261}$$

and

$$0 < \frac{\kappa\mathbb{E}[Y_i] - \kappa\mathbb{E}[Y_i]\cdot p_{\gamma\gamma}(Z^{(k)}(\gamma), A^{(k)}(\gamma))}{f(Z^{(k)}(\gamma))} < \frac{\mathbb{E}[Y_i](1 - p_{\gamma\gamma}(Z^{(k)}(\gamma), A^{(k)}(\gamma)))}{\mathbb{E}[Y_i] + Z^{(k)}(\gamma)} \leq 1. \tag{262}$$

By combining (261) and (262), we establish that the transition probabilities $\widetilde{p_{\gamma\gamma'}}(Z^{(k)}(\gamma), A^{(k)}(\gamma))$ satisfy the condition $0 \leq \widetilde{p_{\gamma\gamma'}}(Z^{(k)}(\gamma), A^{(k)}(\gamma)) \leq 1$. Given this and the result from (209), it follows that $\widetilde{\mathbf{P}_{\kappa}(k)}$ is indeed a stochastic matrix.

Second, we show that $\widetilde{\mathbf{P}_{\kappa}(k)}$ is always *aperiodic*. This is verified by the following inequality:

$$\begin{aligned} &\widetilde{p_{\gamma\gamma}}(Z^{(k)}(\gamma), A^{(k)}(\gamma)) \\ &= 1 - \frac{\kappa\mathbb{E}[Y_i] - \kappa\mathbb{E}[Y_i]p_{\gamma\gamma}(Z^{(k)}(\gamma), A^{(k)}(\gamma))}{Z^{(k)}(\gamma) + \mathbb{E}[Y_i]} \\ &> 1 - \frac{\mathbb{E}[Y_i] - \mathbb{E}[Y_i]p_{\gamma\gamma}(Z^{(k)}(\gamma), A^{(k)}(\gamma))}{\mathbb{E}[Y_i]} \\ &= p_{\gamma\gamma}(Z^{(k)}(\gamma), A^{(k)}(\gamma)) \geq 0. \end{aligned} \tag{263}$$

As the self-transition probability is always positive for each state γ, the chain is always *aperiodic*. At last, we show that if MDP characterized by $p_{\gamma\gamma'}(z, a)$ is a *unichain*, then the MDP characterized by $\widetilde{p_{\gamma\gamma'}}(z, a)$ is also a *unichain*, this in given in the following lemma.

Lemma 13. *If the MDP characterized by the transition probability $p_{\gamma\gamma'}(z, a)$ is a unichain, then the MDP characterized by $\widetilde{p_{\gamma\gamma'}}(z, a)$ is also a unichain.*

Proof. See Appendix Q. ■

According to Lemma 13, the Markov Decision Process (MDP) defined by $\widetilde{\mathbf{P}_{\kappa}(k)}$ is an *aperiodic unichain*, satisfying the condition established in [112]. This completes the proof.

Appendix Q
Proof of Lemma 13

Since the MDP is a *unichain*, it follows that for any policy $\phi : \mathcal{S}\times\mathcal{Y}\times\mathcal{A} \to \mathcal{Z}\times\mathcal{A}$, the markov chain constitutes some transient states and a single recurrent class. Let $\mathcal{X}_{\phi}$ denote the *recurrent class* of the Markov chain under policy ϕ, we have that if $\gamma' \in \mathcal{X}_{\phi}$ and $\gamma \in \mathcal{S}\times\mathcal{Y}\times\mathcal{A}$, there always exists a smallest positive integer m such that:

$$[\mathbf{P}_{\phi}^{n}]_{\gamma\times\gamma'} \triangleq p_{\phi,\gamma\gamma'}^{n} \begin{cases} > 0 & \text{if } n = m \\ = 0 & \text{if } n < m \end{cases}, \tag{264}$$

where $\mathbf{P}_{\phi} \in \mathbb{R}^{|\mathcal{S}\times\mathcal{Y}\times\mathcal{A}|\times|\mathcal{S}\times\mathcal{Y}\times\mathcal{A}|}$ is the transition probability matrix under policy ϕ, with entries $p_{\phi,\gamma\gamma'}^{1} = p_{\gamma\gamma'}(\phi(\gamma))$. Note that $p_{\phi,\gamma\gamma'}^{m-1} = 0$, we can apply the *Chapman–Kolmogorov equation* to establish:

$$\begin{aligned} p_{\phi,\gamma\gamma'}^{m} &= \sum_{\gamma_1} p_{\phi,\gamma\gamma_1}^{1} p_{\phi,\gamma_1\gamma'}^{m-1} \\ &= p_{\phi,\gamma\gamma}^{1} p_{\phi,\gamma\gamma'}^{m-1} + \sum_{\gamma_1\neq\gamma} p_{\phi,\gamma\gamma_1}^{1} p_{\phi,\gamma_1\gamma'}^{m-1} \\ &= \sum_{\gamma\neq\gamma_1} p_{\phi,\gamma\gamma_1}^{1} p_{\phi,\gamma_1\gamma'}^{m-1}. \end{aligned} \tag{265}$$

Iterating (265) for $m-1$ times yields:

$$p_{\phi,\gamma\gamma'}^{m} = \sum_{\substack{\gamma\neq\gamma_1,\gamma_2\neq\gamma_3, \\ \cdots,\gamma_{m-1}\neq\gamma'}} p_{\phi,\gamma\gamma_1}^{1} p_{\phi,\gamma_1\gamma_2}^{1} \cdots p_{\phi,\gamma_{m-1}\gamma'}^{1}. \tag{266}$$

Similarly, define $\widetilde{p_{\phi,\gamma\gamma'}^{m}}$ as the m-step transition probability for the Markov chain characterized by $\widetilde{p_{\gamma\gamma'}}(\phi(\gamma))$, we can apply

$$\widetilde{p^{n+1}_{\phi,\gamma'\gamma'}} = \widetilde{p^{1}_{\phi,\gamma'\gamma'}}\widetilde{p^{n}_{\phi,\gamma'\gamma'}} + \sum_{\gamma_1\neq\gamma'} \widetilde{p^{1}_{\phi,\gamma'\gamma_1}}\widetilde{p^{n}_{\phi,\gamma_1\gamma'}} \tag{272a}$$

$$= \widetilde{p^{1}_{\phi,\gamma'\gamma'}}\widetilde{p^{n}_{\phi,\gamma'\gamma'}} + \sum_{\gamma_1\neq\gamma'} \widetilde{p^{1}_{\phi,\gamma'\gamma_1}} \sum_{\gamma_2,\cdots,\gamma_n} \widetilde{p_{\phi,\gamma_1\gamma_2}}\widetilde{p_{\phi,\gamma_2\gamma_3}}\cdots\widetilde{p_{\phi,\gamma_n\gamma'}} \tag{272b}$$

$$= \widetilde{p^{1}_{\phi,\gamma'\gamma'}}\widetilde{p^{n}_{\phi,\gamma'\gamma'}} + \sum_{\gamma_1\neq\gamma',\gamma_2,\cdots,\gamma_n} \frac{\kappa\mathbb{E}[Y_i]\cdot p^1_{\phi,\gamma'\gamma_1}}{f(\phi(\gamma'))}\cdot\prod_{j\notin\mathcal{D}}\frac{\kappa\mathbb{E}[Y_i]\cdot p^1_{\phi,\gamma_j\gamma_{j+1}}}{f(\phi(\gamma_j))}\cdot\prod_{j\in\mathcal{D}}\left(1-\frac{\kappa\mathbb{E}[Y_i]-\kappa\mathbb{E}[Y_i]p^1_{\phi,\gamma_j\gamma_j}}{f(\phi(\gamma_j))}\right) \tag{272c}$$

$$= \widetilde{p^{1}_{\phi,\gamma'\gamma'}}\widetilde{p^{n}_{\phi,\gamma'\gamma'}} + \sum_{\gamma_1\neq\gamma',\gamma_2,\cdots,\gamma_n} \frac{(\kappa\mathbb{E}[Y_i])^{n-|\mathcal{D}+1|}\cdot\prod_{j\in\mathcal{D}}\left(1-\frac{\kappa\mathbb{E}[Y_i]-\kappa\mathbb{E}[Y_i]p^1_{\phi,\gamma_j\gamma_j}}{f(\phi(\gamma_j))}\right)}{f(\phi(\gamma'))\prod_{j\notin\mathcal{D}}f(\phi(\gamma_j))}\cdot p^1_{\phi,\gamma'\gamma_1}\prod_{j\notin\mathcal{D}}p^1_{\phi,\gamma_j\gamma_{j+1}}. \tag{272d}$$

the *Chapman–Kolmogorov equation* to have

$$\widetilde{p^{m}_{\phi,\gamma\gamma'}} = \sum_{\gamma_1,\gamma_2,\cdots,\gamma_{m-1}} \widetilde{p^{1}_{\phi,\gamma\gamma_1}}\widetilde{p^{1}_{\phi,\gamma_1\gamma_2}}\cdots\widetilde{p^{1}_{\phi,\gamma_{m-1}\gamma'}} \tag{267a}$$

$$\geq \sum_{\substack{\gamma\neq\gamma_1,\gamma_2\neq\gamma_3,\\ \cdots,\gamma_{m-1}\neq\gamma'}} \widetilde{p^{1}_{\phi,\gamma\gamma_1}}\widetilde{p^{1}_{\phi,\gamma_1\gamma_2}}\cdots\widetilde{p^{1}_{\phi,\gamma_{m-1}\gamma'}} \tag{267b}$$

$$= \frac{(\kappa\mathbb{E}[Y_i])^m}{f(\phi(\gamma))\prod_{i=1}^{m-1}f(\phi(\gamma_i))}\times \sum_{\substack{\gamma\neq\gamma_1,\gamma_2\neq\gamma_3,\\ \cdots,\gamma_{m-1}\neq\gamma'}} p^1_{\phi,\gamma\gamma_1}p^1_{\phi,\gamma_1\gamma_2}\cdots p^1_{\phi,\gamma_{m-1}\gamma'} \tag{267c}$$

$$= \frac{(\kappa\mathbb{E}[Y_i])^m\cdot p^m_{\phi,\gamma\gamma'}}{f(\phi(\gamma))\prod_{i=1}^{m-1}f(\phi(\gamma_i))} > 0, \tag{267d}$$

where (267c) is established by the first line of (260) and (267d) is obtained by substituting (266). As such, if γ' belongs to a recurrent class of the Markov chain characterized by $p^1_{\phi,\gamma\gamma'}$, it will also belong to a recurrent class of the Markov chain under $\widetilde{p^1_{\phi,\gamma\gamma'}}$, indicating that $\mathcal{X}_\phi$ is a recurrent class of the new chain $\widetilde{p^1_{\phi,\gamma\gamma'}}$.

We next show that if γ' is a transient state of the Markov chain $p^1_{\phi,\gamma\gamma'}$, it is also a transient state of the Markov chain $\widetilde{p^1_{\phi,\gamma\gamma'}}$. If γ' is a transient state of the Markov unichain $p^1_{\phi,\gamma\gamma'}$, it follows that

$$p^k_{\phi,\gamma'\gamma_1}p^n_{\phi,\gamma_1\gamma'} = 0, \text{for } \forall\gamma_1\neq\gamma', k,n\in\mathbb{N}^+, \tag{268}$$

otherwise, the chain will be a multichain or γ' is not a transient state. By leveraging the *Chapman–Kolmogorov equation* to expand $p^n_{\phi,\gamma_1\gamma'}$, we have

$$0 = p^k_{\phi,\gamma'\gamma_1}p^n_{\phi,\gamma_1\gamma'} = \sum_{\gamma_2,\cdots,\gamma_n} p^k_{\phi,\gamma'\gamma_1}p^1_{\phi,\gamma_1\gamma_2}p^1_{\phi,\gamma_2\gamma_3}\cdots p^1_{\phi,\gamma_n\gamma'} \geq 0. \tag{269}$$

From the *sandwich theorem*, we know that every element in the summand of (269) is zero:

$$p^k_{\phi,\gamma'\gamma_1}p^1_{\phi,\gamma_1\gamma_2}p^1_{\phi,\gamma_2\gamma_3}\cdots p^1_{\phi,\gamma_n\gamma'} = 0, \quad \text{for } \forall\gamma_1\neq\gamma',\gamma_2,\cdots,\gamma_n,k,n\in\mathbb{N}^+, \tag{270}$$

Meanwhile, we can leverage the *Chapman–Kolmogorov equation* to expand the $n+1$-step transition probability $\widetilde{p^{n+1}_{\phi,\gamma'\gamma'}}$ and obtain (272) at the top of the previous page, where the set $\mathcal{D}$ is defined as $\mathcal{D}\triangleq\{j|\gamma_j=\gamma_{j+1}, j=1,2,\cdots,n\}$, with $\gamma_{n+1}=\gamma'$ and (272c) is obtained by substituting (260). As (270) holds for all n, we know that

$$p^1_{\phi,\gamma'\gamma_1}\prod_{j\notin\mathcal{D}}p^1_{\phi,\gamma_j\gamma_{j+1}} = 0. \tag{272}$$

Applying (272) in (272) yields:

$$\widetilde{p^{n+1}_{\phi,\gamma'\gamma'}} = \widetilde{p^{1}_{\phi,\gamma'\gamma'}}\widetilde{p^{n}_{\phi,\gamma'\gamma'}}. \tag{273}$$

Iterating (273) yields

$$\widetilde{p^{n}_{\phi,\gamma'\gamma'}} = \left(\widetilde{p^{1}_{\phi,\gamma'\gamma'}}\right)^n = \left(1-\frac{\kappa\mathbb{E}[Y_i]-\kappa\mathbb{E}[Y_i]p^1_{\phi,\gamma'\gamma'}}{f(\phi(\gamma'))}\right)^n. \tag{274}$$

Because γ' is a transient state of the chain characterized by $p^1_{\phi,\gamma\gamma'}$, we have $p^1_{\phi,\gamma'\gamma'}<1$, and thus

$$\lim_{n\to\infty}\widetilde{p^n_{\phi,\gamma'\gamma'}} = 0, \tag{275}$$

which indicates that γ' is also a transient state of the new chain characterized by $\widetilde{p^1_{\phi,\gamma\gamma'}}$. We thus accomplish the proof.

REFERENCES

[1] A. Li, S. Wu, G. C. F. Lee, X. Chen, and S. Sun, "Sampling to achieve the goal: An age-aware remote markov decision process," in *Proc. IEEE ITW*, 2024, pp. 121–126.

[2] 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.

[3] 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.

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

[5] R. Talak and E. H. Modiano, "Age-delay tradeoffs in queueing systems," *IEEE Trans. Inf. Theory*, vol. 67, no. 3, pp. 1743–1758, 2020.

[6] 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.

[7] A. M. Bedewy, Y. Sun, and N. B. Shroff, "Minimizing the age of information through queues," *IEEE Trans. Inf. Theory*, vol. 65, no. 8, pp. 5215–5232, 2019.

[8] L. Huang and E. Modiano, "Optimizing age-of-information in a multi-class queueing system," in *Proc. IEEE ISIT*, 2015, pp. 1681–1685.

[9] 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.

[10] 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.

[11] 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.

[12] M. Moltafet, M. Leinonen, and M. Codreanu, "On the age of information in multi-source queueing models," *IEEE Trans. Commun.*, vol. 68, no. 8, pp. 5003–5017, 2020.

[13] N. Pappas, J. Gunnarsson, L. Kratz, M. Kountouris, and V. Angelakis, "Age of information of multiple sources with queue management," in *Proc. IEEE ICC*, 2015, pp. 5935–5940.

[14] J. Li and W. Zhang, "Asymptotically optimal joint sampling and compression for timely status updates: Age–distortion tradeoff," *IEEE Trans. Veh. Technol.*, vol. 74, no. 2, pp. 2338–2352, 2025.

[15] P. Mayekar, P. Parag, and H. Tyagi, "Optimal source codes for timely updates," *IEEE Trans. Inf. Theory*, vol. 66, no. 6, pp. 3714–3731, 2020.

[16] Y. Sun, Y. Polyanskiy, and E. Uysal-Biyikoglu, "Remote estimation of the wiener process over a channel with random delay," in *Proc. IEEE ISIT*. IEEE, 2017, pp. 321–325.

[17] 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.

[18] A. Mitra, J. A. Richards, S. Bagchi, and S. Sundaram, "Distributed state estimation over time-varying graphs: Exploiting the age-of-information," *IEEE Trans. Autom. Control.*, vol. 67, no. 12, pp. 6349–6365, 2021.

[19] X. Chen, X. Liao, and S. S. Bidokhti, "Real-time sampling and estimation on random access channels: Age of information and beyond," in *Proc. IEEE INFOCOM*, 2021, pp. 1–10.

[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 *Proc. IEEE WiOPT*, 2019, pp. 1–8.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] A. Arafa, K. Banawan, K. G. Seddik, and H. Vincent Poor, "Timely estimation using coded quantized samples," in *Proc. IEEE ISIT*, 2020, pp. 1812–1817.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] J. P. Mena and F. Núñez, "Age of information in IoT-based networked control systems: A MAC perspective," *Autom.*, vol. 147, p. 110652, 2023.

[32] 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.

[33] 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.

[34] 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.*, vol. 73, no. 10, pp. 15 440–15 455, 2024.

[35] 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.

[36] 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.

[37] H. Tang, J. Wang, L. Song, and J. Song, "Minimizing age of information with power constraints: Multi-user opportunistic scheduling in multi-state time-varying channels," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 5, pp. 854–868, 2020.

[38] Y. Long, J. Zhuang, S. Gong, B. Gu, J. Xu, and J. Deng, "Exploiting deep reinforcement learning for stochastic aoi minimization in multi-uav-assisted wireless networks," in *IEEE Proc. WCNC*, 2024, pp. 1–6.

[39] 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.

[40] 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.

[41] 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.*, vol. 31, no. 5, pp. 238–245, 2024.

[42] D. Gündüz, Z. Qin, I. E. Aguerri, H. S. Dhillon, Z. Yang, A. Yener, K. K. Wong, and C.-B. Chae, "Beyond transmitting bits: Context, semantics, and task-oriented communications," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 5–41, 2023.

[43] Z. Qin, L. Liang, Z. Wang, S. Jin, X. Tao, W. Tong, and G. Y. Li, "Ai empowered wireless communications: From bits to semantics," *Proceedings of the IEEE*, vol. 112, no. 7, pp. 621–652, 2024.

[44] 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.

[45] A. Kosta, N. Pappas, A. Ephremides, and V. Angelakis, "The cost of delay in status updates and their value: Non-linear ageing," *IEEE Trans. Commun.*, vol. 68, no. 8, pp. 4905–4918, 2020.

[46] J. Cho and H. Garcia-Molina, "Effective page refresh policies for web crawlers," *ACM Trans. Database Syst.*, vol. 28, no. 4, pp. 390–426, 2003.

[47] M. Bastopcu and S. Ulukus, "Information freshness in cache updating systems," *IEEE Trans. Wirel. Commun.*, vol. 20, no. 3, pp. 1861–1874, 2020.

[48] K. T. Truong and R. W. Heath, "Effects of Channel Aging in Massive MIMO systems," *J. Commun. Netw.*, vol. 15, no. 4, pp. 338–351, 2013.

[49] Y. Sun and B. Cyr, "Information aging through queues: A mutual information perspective," in *2018 IEEE 19th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2018, pp. 1–5.

[50] Z. Wang, M.-A. Badiu, and J. P. Coon, "A Framework for Characterising the Value of Information in Hidden Markov Models," *IEEE Trans. Inf. Theory*, vol. 9448, no. c, pp. 1–15, 2021.

[51] G. Chen, S. C. Liew, and Y. Shao, "Uncertainty-of-information scheduling: A restless multiarmed bandit framework," *IEEE Trans. Inf. Theory*, vol. 68, no. 9, pp. 6151–6173, 2022.

[52] J. Zhong, R. D. Yates, and E. Soljanin, "Two freshness metrics for local cache refresh," in *2018 IEEE Proceedings ISIT*, 2018, pp. 1924–1928.

[53] A. Maatouk, S. Kriouile, M. Assaad, and A. Ephremides, "The age of incorrect information: A new performance metric for status updates," *IEEE/ACM Trans. Netw.*, vol. 28, no. 5, pp. 2215–2228, 2020.

[54] E. Delfani and N. Pappas, "Optimizing information freshness in constrained iot systems: A token-based approach," *IEEE Trans. Commun.*, vol. 73, no. 8, pp. 5848–5863, 2025.

[55] Y. Chen and A. Ephremides, "Minimizing age of incorrect information over a channel with random delay," *IEEE/ACM Trans. Netw.*, 2024.

[56] X. Wang, W. Lin, C. Xu, X. Sun, and X. Chen, "Age of changed information: Content-aware status updating in the internet of things," *IEEE Trans. Commun.*, vol. 70, no. 1, pp. 578–591, 2022.

[57] L. Wang, J. Sun, Y. Sun, S. Zhou, Z. Niu, M. Jiang, and L. Geng, "Grouping-based cyclic scheduling under age of correlated information constraints," *IEEE Trans. Inf. Theory*, vol. 71, no. 3, pp. 2218–2244, 2025.

[58] Q. He, G. Dán, and V. Fodor, "Joint assignment and scheduling for minimizing age of correlated information," *IEEE/ACM Trans. Netw.*, vol. 27, no. 5, pp. 1887–1900, 2019.

[59] X. Zheng, S. Zhou, and Z. Niu, "Urgency of information for context-aware timely status updates in remote control systems," *IEEE Trans. Wirel. Commun.*, vol. 19, no. 11, pp. 7237–7250, 2020.

[60] Y. Dong, Z. Chen, S. Liu, P. Fan, and K. B. Letaief, "Age-upon-decisions minimizing scheduling in internet of things: To be random or to be deterministic?" *IEEE Internet Things J.*, vol. 7, no. 2, pp. 1081–1097, 2019.

[61] A. Nikkhah, A. Ephremides, and N. Pappas, "Age of actuation in a wireless power transfer system," in *IEEE INFOCOM INFOCOM WKSHPS*. IEEE, 2023, pp. 1–6.

[62] M. Kountouris and N. Pappas, "Semantics-Empowered Communication for Networked Intelligent Systems," *IEEE Commun. Mag.*, vol. 59, pp. 96–102, 2020.

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

[64] M. Salimnejad, M. Kountouris, and N. Pappas, "Real-time reconstruction of markov sources and remote actuation over wireless channels," *IEEE Trans. Commun.*, vol. 72, no. 5, pp. 2701–2715, 2024.

[65] P. Zou, A. Maatouk, J. Zhang, and S. Subramaniam, "How costly was that (in) decision?" in *Proc. IEEE WiOpt*. IEEE, 2023, pp. 278–285.

[66] A. Li, S. Wu, S. Sun, and J. Cao, "Goal-oriented tensor: Beyond age of information toward semantics-empowered goal-oriented communications," *IEEE Trans. Commun.*, vol. 72, no. 12, pp. 7689–7704, 2024.

[67] 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.

[68] 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.

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

[70] 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.

[71] 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.

[72] S. Liyanaarachchi and S. Ulukus, "The role of early sampling in age of information minimization in the presence of ACK delays," *IEEE Trans. Inf. Theory*, vol. 70, no. 9, pp. 6665–6678, 2024.

[73] F. Peng, X. Wang, and X. Chen, "Online learning of goal-oriented status updating with unknown delay statistics," *IEEE J. Sel. Areas Commun.*, 2024.

[74] H. Tang, Y. Sun, and L. Tassiulas, "Sampling of the wiener process for remote estimation over a channel with unknown delay statistics," *IEEE/ACM Trans. Netw.*, vol. 32, no. 3, pp. 1920–1935, 2024.

[75] Y. Chen, H. Tang, J. Wang, P. Yang, and L. Tassiulas, "Sampling for remote estimation of an ornstein-uhlenbeck process through channel with unknown delay statistics," *J. Commun. Networks*, vol. 25, no. 5, pp. 670–687, 2023.

[76] X. Chen, A. Li, and S. Wu, "Optimal sampling for Uncertainty-of-In formation minimization in a remote monitoring system," in *2024 IEEE Information Theory Workshop (ITW)*, 2024.

[77] M. K. C. Shisher, Y. Sun, and I.-H. Hou, "Timely communications for remote inference," *IEEE/ACM Trans. Netw.*, 2024.

[78] M. K. C. Shisher and Y. Sun, "How does data freshness affect real-time supervised learning?" in *Proc. ACM Mobihoc*, 2022, pp. 31–40.

[79] M. K. C. Shisher, B. Ji, I.-H. Hou, and Y. Sun, "Learning and communications co-design for remote inference systems: Feature length selection and transmission scheduling," *IEEE J. Sel. Areas Commun.*, 2023.

[80] C. Ari, M. K. C. Shisher, E. Uysal, and Y. Sun, "Goal-oriented communications for remote inference under two-way delay with memory," in *Proc. IEEE ISIT*, 2024, pp. 1179–1184.

[81] X. Wang, W. Lin, C. Xu, X. Sun, and X. Chen, "Age of changed information: Content-aware status updating in the internet of things," *IEEE Trans. Commun.*, vol. 70, no. 1, pp. 578–591, 2021.

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

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

[84] 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.

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

[86] 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.

[87] 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.

[88] E. Fountoulakis, N. Pappas, and M. Kountouris, "Goal-oriented policies for cost of actuation error minimization in wireless autonomous systems," *IEEE Communications Letters*, vol. 27, no. 9, pp. 2323–2327, 2023.

[89] A. M. Bedewy, Y. Sun, S. Kompella, and N. B. Shroff, "Optimal sampling and scheduling for timely status updates in multi-source networks," *IEEE Tran. Inf. Theory*, vol. 67, no. 6, pp. 4019–4034, 2021.

[90] M. Zhou, M. Zhang, H. H. Yang, and R. D. Yates, "Age-minimal CPU scheduling," in *Proc. IEEE INFOCOM*, 2024, pp. 401–410.

[91] S. Banach, "Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales," *Fundamenta mathematicae*, vol. 3, no. 1, pp. 133–181, 1922.

[92] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[93] D. Bertsekas, *Dynamic Programming and Optimal Control: Volume I*. Athena scientific, 2012, vol. 1.

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

[95] D. Lee and M. W. Spong, "Passive bilateral teleoperation with constant time delay," *IEEE Trans. Robotics*, vol. 22, no. 2, pp. 269–281, 2006.

[96] D. V. Dimarogonas and K. H. Johansson, "Event-triggered control for multi-agent systems," in *Proc. IEEE CDC*. IEEE, 2009, pp. 7131–7136.

[97] J. P. Hespanha, P. Naghshtabrizi, and Y. Xu, "Survey of recent results in networked control systems," *Proc. IEEE*, vol. 95, no. 1, pp. 138–162, 2007.

[98] A. Somani, N. Ye, D. Hsu, and W. S. Lee, "DESPOT: online POMDP planning with regularization," in *Proc. NeurIPS*, 2013, pp. 1772–1780.

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

[100] A. M. Bedewy, Y. Sun, S. Kompella, and N. B. Shroff, "Optimal sampling and scheduling for timely status updates in multi-source networks," *IEEE Trans. Inf. Theory*, vol. 67, no. 6, pp. 4019–4034, 2021.

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

[102] D. J. White, "Dynamic programming, markov chains, and the method of successive approximations," *J. Math. Anal. Appl*, vol. 6, no. 3, pp. 373–376, 1963.

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

[104] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[105] F. J. Beutler and K. W. Ross, "Optimal policies for controlled Markov chains with a constraint," *J. Math. Anal. Appl.*, vol. 112, no. 1, pp. 236–252, 1985.

[106] J. Luo and N. Pappas, "Semantic-aware remote estimation of multiple markov sources under constraints," *IEEE Trans. Commun. (Early Access)*, 2025.

[107] G. Desaulniers, J. Desrosiers, and M. M. Solomon, *Column Generation*. Springer Science & Business Media, 2006, vol. 5.

[108] J. Gondzio, "Interior point methods 25 years later," *Eur. J. Oper. Res.*, vol. 218, no. 3, pp. 587–601, 2012.

[109] A. Malek, Y. Abbasi-Yadkori, and P. Bartlett, "Linear programming for large-scale markov decision problems," in *IEEE Proc. ICML*. PMLR, 2014, pp. 496–504.

[110] Y. Sun, E. Uysal-Biyikoglu, 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.
[111] Y. Li, B. Yin, and H. Xi, "Finding optimal memoryless policies of pomdps under the expected average reward criterion," *Eur. J. Oper. Res.*, vol. 211, no. 3, pp. 556–567, 2011.
[112] D. Bertsekas and J. N. Tsitsiklis, *Introduction to probability*. Athena Scientific, 2008, vol. 1.

Aimin Li (Member, IEEE) received the B.S. degree (Best Thesis Award) and the Ph.D. degree (Awarded Best Dissertation Nomination) in electronic engineering from Harbin Institute of Technology (Shenzhen) in 2020 and 2025, respectively. From 2023 to 2024, he was a visiting researcher with the Institute for Infocomm Research (I^2R), Agency for Science, Technology, and Research (A*STAR), Singapore. His research interests include advanced channel coding techniques, information theory, and wireless communications. He has served as a reviewer for IEEE TIT, IEEE JSAC, IEEE TWC, IEEE TMC, IEEE TCOM, IEEE ISIT, among others, and as a session chair for IEEE Information Theory Workshop 2024 and IEEE Globecom 2024.

Shaohua Wu (Member, IEEE) received the Ph.D. degree in communication engineering from the Harbin Institute of Technology (HIT), Harbin, China, in 2009. From 2009 to 2011, he held a Postdoctoral position with the Department of Electronics and Information Engineering, Shenzhen Graduate School, HIT (Shenzhen), Shenzhen, China, where he has been an Associate Professor since 2012. He is also an Associate Professor of Peng Cheng Laboratory, Shenzhen. From 2014 to 2015, he was a Visiting Researcher with BBCR, University of Waterloo, Waterloo, ON, Canada. He holds more than 30 Chinese patents. He has authored or coauthored more than 100 papers in the above-mentioned areas. His current research interests include wireless image/video transmission, space communications, advanced channel coding techniques, and B5G wireless transmission technologies.

Gary C.F. Lee (Member, IEEE) received his B.S. degree in Electrical Engineering from Stanford University in 2016, his M.S. degree in Electrical Engineering from the Massachusetts Institute of Technology (MIT) in 2019, and his Ph.D. degree in Electrical Engineering from the Massachusetts Institute of Technology (MIT) in 2023. He is currently a Senior Scientist at the Institute for Infocomm Research (I2R), Agency for Science, Technology, and Research (A*STAR), Singapore. His current research interests are in wireless communications, machine learning, and signal processing.

Sumei Sun (Fellow, IEEE) is a Principal Scientist, Distinguished Institute Fellow, and Acting Executive Director of the Institute for Infocomm Research (I2R), Agency for Science, Technology, and Research (A*STAR), Singapore. She is also holding a joint appointment with the Singapore Institute of Technology, and an adjunct appointment with the National University of Singapore, both as a full professor. Her current research interests are in next-generation wireless communications, cognitive communications and networks, industrial internet of things, communications-computing-control integrative design, joint radar-communication systems, and signal intelligence. She is Editor-in-Chief of the IEEE Open Journal Vehicular Technology, and Steering Committee Chair of the IEEE Transactions on Machine Learning in Communications and Networking. She is also Member-at-Large of the IEEE Communications Society, a member of the IEEE Vehicular Technology Society Board of Governors (2022-2024), Fellow of the IEEE and the Academy of Engineering Singapore.