Contextual Status Updating: How to Maximize Situational Awareness?

Tasmeen Zaman Ornee, *Member, IEEE,* Md Kamran Chowdhury Shisher, *Member, IEEE,* Clement Kam, *Member, IEEE,* and Yin Sun, *Senior Member, IEEE*

***Abstract*—In this study, we investigate contextual status updating, where a centralized monitor pulls status updates from multiple agents monitoring several safety-critical situations (e.g., carbon monoxide density in forest fire detection, machine safety in industrial automation, road safety, etc). Based on the received updates, multiple estimators determine the current safety-critical situations. Due to transmission errors and limited communication resources, the received updates may not be timely, resulting in the possibility of misunderstanding the current situation. In particular, if a dangerous situation is misinterpreted as safe, the safety risk is high. Therefore, we aim to solve a multi-sensor, multi-channel transmission scheduling problem that minimizes the loss due to the unawareness of potential danger. Due to communication resource constraints, the scheduling problem can be formulated as a Restless Multi-armed Bandit (RMAB) which exhibits a complicated state space. By leveraging the sufficient statistic of the history, we are able to significantly reduce the state space which only consists of the latest received observation and its Age of Information (AoI). We further simplify our problem by characterizing the optimal estimator and modeling the loss as generalized conditional entropy. This illustrates that frequent status updating is necessary when the received observation is near the safety boundary. We provide an asymptotically optimal low-complexity scheduling algorithm to solve the RMAB that does not need to satisfy any indexability. Our results hold for general loss functions and both reliable and unreliable channels. Numerical evidence shows that our scheduling policy achieves higher performance gain over periodic updating, randomized policy, and Maximum Age First (MAF) policy.**

***Index Terms*—safety, age of information, Markov decision process, estimation.**

I. Introduction

REAL-time decision-making capabilities are essential in safety-critical systems in various domains [2]. For example, in a health monitoring system, precise tracking of the glucose level or heart rate is crucial to facilitate prompt precautionary measures [3]. In disaster monitoring, it is important to swiftly track any consistent changes in temperature or humidity, as they could indicate a possible disaster [4]. These scenarios highlight the need for monitoring systems to access accurate and up-to-date information about remote systems. Any misunderstanding of the system state can lead to severe consequences.

One challenge to efficiently utilize the state information in real-time is the limited capacity of the communication medium. Conversely, some information may have more crucial content than others and hence need more attention. For example, in autonomous driving, a self-driving car deciding to change lanes need to prioritize real-time information about vehicles in the adjacent lane over those in its current lane.

In this context, we consider a pull-based status updating system where multiple agents monitor the status of different safety-critical situations. A central scheduler requests updates from agents whenever it is uncertain of their safety situations. Based on the received updates, multiple estimators estimate the safety situation of the agents. Due to transmission errors, the received updates may not be fresh. One performance metric that characterizes data freshness is the *Age of Information (AoI)* or simply *Age* [5]. Let $U(t)$ be the generation time of the freshest received observation by time t. AoI, as a function of t, is defined as $\Delta(t) = t - U(t)$ which exhibits a linear growth with time t and drops down to a smaller value whenever a fresher observation is delivered. However, AoI only captures the timeliness of the information, but not its significance. Hence, solely relying on AoI-based decision-making is not sufficient, particularly, in safety-critical scenarios where misunderstanding about the situation can lead to significant performance loss. This motivated us to explore beyond AoI-based decision-making by incorporating the information conveyed by *all* of the received signals in the decision-making.

To address this issue, we formulate a multi-agent, multi-channel scheduling problem that minimizes the performance loss due to the unawareness of potential danger while satisfying a channel resource constraint. The formulated problem is a Restless Multi-armed Bandit (RMAB) where the state space comprises the history of all received packets, their AoI values, the scheduling decisions, and the delivery indicators up to the current time t. However, the large state space in RMAB often poses curse of dimensionality and significant computational challenges. One may wonder how to reduce the state space to obtain a computationally efficient solution. Therefore, in this study, we seek the answers of two questions: *(1) How to simplify the state space to reduce the overall complexity of the problem?* and *(2) How to design an efficient transmission*

A part of this manuscript was presented in *IEEE MILCOM Workshop on QuAVoI*, Boston, MA, 2023 [1].

T. Z. Ornee is with the Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH, 43210 USA (email: ornee.1@osu.edu).

M. K. C. Shisher is with the Department of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, 47907 USA (email: mshisher@purdue.edu).

T. Z. Ornee and M. K. C. Shisher were Ph.D. students in the Department of Electrical and Computer Engineering at Auburn University, Auburn, AL, 36830 USA.

C. Kam is with the U.S. Naval Research Laboratory, Washington, DC 20375 USA (e-mail: clement.kam@nrl.navy.mil).

Y. Sun is with the Department of Electrical and Computer Engineering, Auburn University, Auburn, AL, 36830 USA e-mail: (yzs0078@auburn.edu).

scheduling policy to maximize situational awareness? The contributions of our study are as follows:

- We significantly reduce the state space by leveraging the sufficient statistic of the history. While belief states are commonly used as sufficient statistics in the literature [6]–[15], they can still result in a high-dimensional state space. We find that the latest received observation and its age also form a sufficient statistic for estimating each agent's safety situation (see Theorem 1). Consequently, we replace the belief MDP framework with a Markov Decision Process (MDP) that uses the latest received observation and its age as states. This alternative formulation is crucial because the belief MDP approach leads to a quadratic increase in the size of the state space with AoI [7], [9], [15], [16], while our method demonstrates only a linear increase (see Section IX-A). This results in a substantial computational advantage.
- Our characterization of the optimal estimator reveals an interesting connection between the loss of situational awareness and the concept of generalized conditional entropy. We prove that the expected loss due to the unawareness of potential danger given the latest received observation and its age is a generalized conditional entropy (see Lemma 1). This characterization also facilitates the design of an efficient scheduling policy. Using this information-theoretic analysis, we show that frequent updates are necessary when the received observations are near safety boundaries (high uncertainty, low situational awareness). Conversely, when observations are far from the boundaries (low uncertainty, high situational awareness), less frequent updates are sufficient (see Section IX-B).
- We develop an asymptotically optimal Maximum Gain First policy (see Algorithm 2 and Theorem 2) by solving the multi-sensor, multi-channel transmission scheduling problem which is formulated as an RMAB. We utilize constraint relaxation and the Lagrangian method to decompose the original problem into multiple separated Markov Decision Processes (MDPs) and solve each MDP by dynamic programming [17]. Most of the prior works [6], [14], [18]–[20] in RMAB have utilized Whittle index policy that requires an indexability condition to satisfy. In our problem, the indexability condition is difficult to establish due to (i) complicated state transitions, (ii) unreliable channels, and (iii) general loss functions. The benefit of our Maximum Gain First policy is that no indexability condition is required to satisfy. Our results hold for general loss functions and both reliable and unreliable channels.
- Numerical results illustrate that our scheduling policy achieves significant performance gain compared to periodic updating policy, randomized policy, and Maximum Age First policy (see Section X). Because our policy utilizes the knowledge of the latest received signal, it illustrates good performance compared to the other policies that ignore this knowledge.

II. Related Work

AoI in context-aware updating: Minimization of linear and non-linear functions of AoI has been extensively studied in literature [21]–[27]. One limitation of AoI is that it only captures the timeliness of the information while neglecting the actual influence of the conveyed information. To address this issue, several performance metrics were introduced in conjunction with AoI [16], [28]–[36]. In [29], the concept of Age of Incorrect Information (AoII) was introduced which is characterized as a function of both age and estimation error. In [28], Age of Synchronization (AoS) was considered along with AoI to measure the freshness of a local cache. Urgency of Information (UoI) was proposed in [30] that captures the context-dependence of the status information along with AoI. Version AoI was introduced in [31] which represents how many versions are outdated at the receiver compared to the transmitter. An AoI at Query (QAoI) metric was investigated in [32], [34], [35] to capture the freshness only when required in a pull-based communication system. Value of Information (VoI), defined by the Shannon mutual information was investigated in [36]. In [33], the authors studied the cost of actuation error which is a goal-oriented measure to capture the costs associated with decisions.

In addition, several research papers studied information-theoretic measures to evaluate the impact of information freshness along with information content [16], [20], [21], [36]–[40]. In [21], [36]–[38], the authors employed Shannon's mutual information to quantify the information carried by received data messages regarding the current signal at the source and used Shannon's conditional entropy to measure the uncertainty about the current signal. Based on the studies of [21], [36]–[38], the authors in [16] utilized Uncertainty of Information (UoI) by using the Shannon's conditional entropy. In [20], [39], [40], a generalized conditional entropy associated with a loss function L, or L-conditional entropy $H_L(Y_t|, \Delta(t), X_{t-\Delta(t)})$ was utilized, where Y_t is the true state of the source and $X_{t-\Delta(t)}$ is the observed value. Building upon the insights of [20], [39], [40], we utilized L-conditional entropy $H_L(Y_t|X_{t-\Delta(t)} = x, \Delta(t) = \delta)$ given both the AoI δ and the knowledge of the received observation x to measure the impact of the AoI and the information content in remote estimation and prediction. Compared to [20], [39], [40], we consider a signal-aware scheduling scheme (the decision-maker utilizes the knowledge of the signal value) with the goal of minimizing the performance loss caused by situational unawareness where [20], [39], [40] focused on signal-agnostic scenario (the decision-maker does not utilize signal value).

AoI in Sampling and Scheduling: There exist numerous papers on AoI-based sampling and scheduling [6], [14], [16], [19]–[21], [24], [25], [40]–[45]. Authors in [44] studied an AoI minimization problem under a pulling model that considers replicated requests to the server. In [21], sampling policies for optimizing non-linear AoI functions were studied. In [16], the authors proposed a Whittle index-based scheduling policy to minimize the UoI modeled as Shanon entropy. Optimal scheduling policies for both single and multi-source systems were studied and a Whittle index policy was proposed for

multi-source case in [20]. The optimal sampling policies for Gauss-Markov processes were studied in [24], [25], [45] where the estimation error becomes a monotonic function of age in signal-agnostic scenarios and the associated problem for minimizing age-penalty functions were reported. A Whittle index policy for continuous-time Gauss Markov processes for both signal-aware and signal-agnostic scenarios was reported in [19]. Besides Whittle index-based policies that require an indexability condition, non-indexable scheduling policies were also studied in [14], [15], [40]–[42]. In this paper, because of the complicated nature of state transition along with erasure channels, indexability is very difficult to establish. However, we provide a "Maximization Gain First Policy" developed in [15], [40]. The comparison of our model with [15], [40] is that our model considers erasure channels, signal observation, and general loss functions. Our scheduling policy is designed for a pull-based communication model where the scheduling decisions are based on the latest received observation and its AoI. We further proved that the developed policy is asymptotically optimal.

Restless Bandits with Belief States: RMAB problems are a well-established framework for studying sequential decision-making problems. Our problem focuses on an RMAB where each arm observes a finite state Markov process. There exists numerous closely related studies to our setting that focused on solving RMAB [6]–[15]. In [6], a Whittle index policy of a class of RMAB problems for dynamic multiaccess channels was studied. [8] considered the design of a flow scheduling policy for the Gilbert-Elliot model. [7] considered time-correlated fading channels with ARQ feedback. Opportunistic scheduling of multiple *i.i.d.* channels is studied in [10]. In [12], a multi-user wireless downlink has been studied for unknown channel states. [11] developed an index heuristic for a multi-class queuing system with increasing convex holding cost rates. [13] studied multi-user scheduling for ON-OFF Markov chains with random delayed ARQ feedback. All of these studies tackle the problem by considering a belief MDP formulation (or POMDP formulation) using belief states (probability distribution over all possible states) [6]–[9], [11]–[15]. Such formulations render the state space uncountable and leads to the curse of dimensionality [6], [10]–[12]. The difference between the formulation in [6]–[15] and our problem is that we do not need to utilize belief states. By utilizing the sufficient statistic of the history observation (the latest received observation and the corresponding age value), we obtain a significantly smaller state space. Consequently, our framework achieves more computational efficiency compared to existing approaches involving belief states [6]–[15].

III. Model and Problem Formulation

A. System Model

We consider a time-slotted pull-based status-updating system as depicted in Figure 1, where a central scheduler pulls the statuses (e.g., location of a car on the road, images captured by a camera installed in a UAV, joint angles of a robotic arm within a factory environment, etc) of N agents (e.g. cars, UAVs, robotic arms, etc) to monitor their safety levels

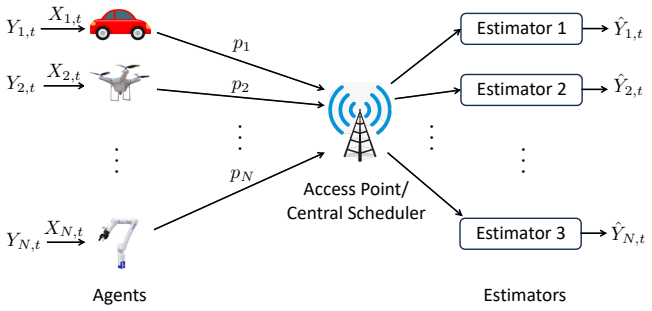

Fig. 1: A multi-agent, multi-channel safety monitoring system.

$$\cdots \leftrightarrow X_{n,t-1} \leftrightarrow X_{n,t} \leftrightarrow X_{n,t+1} \leftrightarrow \cdots$$
$$\cdots \quad\; Y_{n,t-1} \quad\; Y_{n,t} \quad\; Y_{n,t+1} \quad\; \cdots$$

Fig. 2: Relationship between $X_{n,t}$ and $Y_{n,t}$.

(e.g., safe, cautious, dangerous). At every time slot t, N estimators estimate the safety levels based on the available information. Let $X_{n,t} \in \mathcal{X}_n$ be the status of agent n at time t. We assume that $X_{n,t}$ is independent of the statuses of other agents, i.e., $X_{n,t}$ is independent of $X_{n',k}$ for all $n \neq n'$ and $k = 0, 1, 2, \ldots$. Let $Y_{n,t} \in \mathcal{Y}_n$ quantify the safety level of agent n. For each time slot t, $Y_{n,t} \leftrightarrow X_{n,t} \leftrightarrow X_{n,t-1} \leftrightarrow X_{n,t-2} \leftrightarrow \cdots$ and $Y_{n,t} \leftrightarrow X_{n,t} \leftrightarrow X_{n,t+1} \leftrightarrow X_{n,t+2} \leftrightarrow \cdots$ are two Markov chains, as illustrated in Figure 2. An example of this relationship is $Y_{n,t} = f(X_{n,t})$, where $f : \mathcal{X}_n \to \mathcal{Y}_n$ is a function mapping the state space of $X_{n,t}$ to the state space of $Y_{n,t}$.

Status update packets of the N agents are transmitted through M unreliable wireless channels. Upon receiving a pull request from the central scheduler, agent n submits a time-stamped status message $(X_{n,t}, t)$ to one wireless channel. It takes one-time slot to transmit the message to the receiver. Due to wireless channel fading, transmissions of the status messages are subject to errors. Let p_n be the probability of a successful transmission from agent n, irrespective of the selected wireless channel.

Our system consists of N estimators. At time slot t, estimator n infers the safety level $Y_{n,t}$ by utilizing the history information available at the receiver. The loss due to the unawareness of potential associated with agent n is characterized by the function $L_n : \mathcal{Y}_n \times \mathcal{Y}_n \to \mathbb{R}$, where $L_n(y, \hat{y})$ is the incurred loss if $Y_{n,t} = y$ is the actual safety level and $\hat{y}$ is estimated value of the safety level. Let $\mu_n(t) \in \{0, 1\}$ be the decision variable to request a packet from agent n at time-slot t which can be stated as follows:

$$\mu_n(t) = \begin{cases} 1, \text{ if agent } n \text{ is scheduled in time-slot } t, \\ 0, \text{ otherwise.} \end{cases}$$

Because of channel erasures, all requested packets may not be delivered successfully. Let $\gamma_n(t) \in \{0, 1\}$ be the delivery indicator of agent n at time-slot t, where $\gamma_n(t) = 1$ represents a successful delivery from agent n with probability p_n. The events of a successful transmission from each agent n, when scheduled, are independent of each other.

Due to transmission errors, the delivered information may not be fresh and is represented by $X_{n,t-\Delta_n(t)}$ that is generated at time $t-\Delta_n(t)$. The time difference $\Delta_n(t)$ between current time t and the generation time $t-\Delta_n(t)$ is usually called the *age of information (AoI)* [5], which represents the staleness of the status update from the n-th agent. The AoI evolution is given by

$$\Delta_n(t+1)=\begin{cases}1, & \text{if } \mu_n(t)=1 \text{ and } \gamma_n(t)=1,\\ \Delta_n(t)+1, & \text{otherwise.}\end{cases}$$

The information available for each estimator n is then given by

$$\mathbf{H}_{n,t}=\left(X_{n,\tau-\Delta_n(\tau)},\gamma_n(\tau),\mu_n(\tau),\Delta_n(\tau)\right)_{\tau=0}^{t}, \quad (1)$$

which represents the set of all historically received packets, the decision variables, the delivery indicators, and the AoI values of agent n at time-slot t.

Based on the latest available information, the n-th estimator estimates the safety level $Y_{n,t}$ and outputs $\hat{y}=\phi_n(\mathbf{H}_{n,t})\in\mathcal{Y}_n$, where $\phi_n(\cdot)$ is the optimal estimator of the underlying data distribution, given by

$$\phi_n(\mathbf{H}_{n,t})=\underset{\hat{y}\in\mathcal{Y}_n}{\operatorname{argmin}}\,\mathbb{E}[L_n(Y_{n,t},\hat{y})|\mathbf{H}_{n,t}]. \quad (2)$$

B. Loss Model for Situational Awareness

If the *dangerous* situation is wrongly estimated as *safe*, the loss will be significantly high due to the huge risk of damage. Conversely, if the *safe* situation is wrongly estimated as *dangerous*, the loss will be small. This is because even if the estimation is incorrect, the risk of damage is small. An example of such a loss function is presented below.

Example 1. *Consider a scenario with three possible safety situations: dangerous, cautious, safe, where* L_n(*dangerous, safe*) $=1000, L_n$(*safe, dangerous*) $=5, L_n$(*cautious, safe*) $=10, L_n$(*safe, cautious*) $=1$.

The loss for perfect estimation is zero, i.e., L_n(*safe, safe*) $=$ L_n(*cautious, cautious*) $=L_n$(*danger, danger*) $=0$.

It might look surprising that the loss associated with a dangerous situation, i.e., L_n(*danger, danger*) is 0. However, this is because even when the actual safety level is *dangerous*, the situational awareness is good due to perfect estimation.[1] The specific loss values and safety situations used in Example 1 are not fixed. Based on different application contexts, the loss values can be adjusted for multiple safety situations.

Notably, our findings can be applied to general loss functions including the well-known loss functions such as 0-1 loss, quadratic loss, logarithmic loss, etc (See Appendix A for definitions). The key advantage of $L_n(\cdot,\cdot)$ is its ability to address safety concerns that arise from the unawareness of potential danger which the existing loss functions fail to capture. Based upon this insights of $L_n(\cdot,\cdot)$, we design a scheduling policy that prioritizes safety by dynamically adjusting update frequencies: frequent updates while uncertain of the states (unaware of the situation), and less frequent updates while certain of the states (aware of the situation).

[1] In this paper, we are interested in maximizing situational awareness and not optimizing the control policy. Hence, we consider *L(dangerous, dangerous)* $=$ *L(cautious, cautious)*$=0$.

C. Scheduling Policy and Problem Formulation

Let $\pi=(\mu_n(0),\mu_n(1),\ldots)_{n=1}^{N}$ denote a scheduling policy, where $\mu_n(t)\in\{0,1\}$ is the decision variable to schedule agent n at every time slot t. The information available at the beginning of time-slot t will remain the same until the beginning of time-slot $t+1$. If agent n is scheduled for transmission at time slot t, then $\mu_n(t)=1$; otherwise $\mu_n(t)=0$. Let Π denote the set of all causal scheduling policies in which every decision is made by using the current and historical information available at the scheduler.

Our goal is to find an optimal scheduling policy that minimizes the time-average sum of the expected loss of the N agents due to the unawareness of potential danger. The scheduling problem is formulated as

$$\mathsf{L}_{\text{opt}}=\inf_{\pi\in\Pi}\limsup_{T\to\infty}\sum_{n=1}^{N}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_{\pi}\big[L_n(Y_{n,t},\phi_n(\mathbf{H}_{n,t}))\big] \quad (3)$$

$$\text{s.t.}\sum_{n=1}^{N}\mu_n(t)\le M,\mu_n(t)\in\{0,1\},t=0,1,\ldots, \quad (4)$$

where L_{opt} is the optimum value of the problem (3)-(4) and $\phi_n(\cdot,\cdot)$ is defined in (2). Because our system consists of M channels, $\sum_{n=1}^{N}\mu_n(t)\le M$ is required to hold for all time t.

Problem (3)-(4) is a Restless Multi-armed Bandit (RMAB) because the AoI process associated with each agent n continues to evolve regardless of whether agent n is selected for transmission [46]. Solving RMAB problems and finding optimal solutions are significantly challenging. Moreover, the expanding state $(\mathbf{H}_{n,t})_{n=1}^{N}$ makes the problem (3)-(4) more complicated. However, we are able to reduce the state space of problem (3)-(4) by utilizing the sufficient statistic of the history $(\mathbf{H}_{n,t})_{n=1}^{N}$. The details are provided in Section IV.

IV. Characterization of the Optimal Estimator

In order to simplify problem (3)-(4), we leverage the sufficient statistic of the history [17]. By using sufficient statistic, we can significantly reduce the state-space of problem (3)-(4).

Theorem 1. *IF* $X_{n,t}$ *is a Markov chain and for each time slot* t, $Y_{n,t}\leftrightarrow X_{n,t}\leftrightarrow X_{n,t-1}\leftrightarrow X_{n,t-2}\leftrightarrow\cdots$ *and* $Y_{n,t}\leftrightarrow X_{n,t}\leftrightarrow X_{n,t+1}\leftrightarrow X_{n,t+2}\leftrightarrow\cdots$ *are two Markov chains. Then,* $(\Delta_n(t),X_{n,t-\Delta_n(t)})$ *is a sufficient statistic of* $\mathbf{H}_{n,t}$ *for estimating* $Y_{n,t}$.

Proof. See Appendix B. □

Using Theorem 1, the RMAB (3)-(4) can be reformulated as the following simpler problem:

$$\mathsf{L}_{\text{opt}}=\inf_{\pi\in\Pi}\limsup_{T\to\infty}\sum_{n=1}^{N}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_{\pi}\big[L_n(Y_{n,t},f_n(\Delta_n(t),X_{n,t-\Delta_n(t)}))\big] \quad (5)$$

$$\text{s.t.}\sum_{n=1}^{N}\mu_n(t)\le M,\mu_n(t)\in\{0,1\},t=0,1,\ldots, \quad (6)$$

where $f_n(\cdot,\cdot)$ is given by

$$f_n(\delta,x)=\underset{\hat{y}\in\mathcal{Y}_n}{\operatorname{argmin}}\,\mathbb{E}[L_n(Y_{n,t},\hat{y})|\Delta_n(t)=\delta,X_{n,t-\Delta_n(t)}=x]. \tag{7}$$

By this, we obtain RMAB (5)-(6) with a reduced state space, where the latest received observation $X_{n,t-\Delta_n(t)}$ of agent n and its AoI $\Delta_n(t)$ at time slot t are the state of the n-th restless bandit.

Problem (5)-(6) can be further simplified by using the concept of generalized conditional entropy [47], [48] or specifically, the L-conditional entropy [49]. For a random variable Y, the L-entropy is given by

$$H_L(Y)=\min_{a\in\mathcal{A}}\mathbb{E}_{Y\sim P_Y}[L(Y,a)]. \tag{8}$$

Let a_Y be the optimal solution to (8), or specifically the optimal estimator associated with the random variable Y (also called a Bayes estimator [47]). Then, the L-conditional entropy of Y given $X=x$ can be defined as [47]–[49]

$$H_L(Y|X=x)=\min_{a\in\mathcal{A}}\mathbb{E}_{Y\sim P_{Y|X=x}}[L(Y,a)]. \tag{9}$$

Lemma 1. *It holds that*

$$H_L(Y_{n,t}|X_{n,t-\delta}=x)=\min_{\hat{y}\in\mathcal{Y}_n}\mathbb{E}_{Y\sim P_{Y_{n,t}|X_{n,t-\delta}=x}}[L_n(Y,\hat{y})], \tag{10}$$

where $H_L(Y_{n,t}|X_{n,t-\delta}=x)$ is the generalized conditional entropy of $Y_{n,t}$ given the latest received observation $X_{n,t-\delta}=x$ at time slot t which is generated δ times ago.

Proof. See Appendix C. □

By using Lemma 1, problem (5)-(6) can be written as

$$\mathsf{L}_{\text{opt}}=\inf_{\pi\in\Pi}\limsup_{T\to\infty}\sum_{n=1}^{N}\mathbb{E}_\pi\left[\frac{1}{T}\sum_{t=0}^{T-1}H_L(Y_{n,t}|\Delta_n(t),X_{n,t-\Delta_n(t)})\right] \tag{11}$$

$$\text{s.t.}\sum_{n=1}^{N}\mu_n(t)\le M,\mu_n(t)\in\{0,1\},t=0,1,\ldots. \tag{12}$$

V. Restless-Multi Armed Bandit Formulation

Even with the reduced state space, solving RMAB (11)-(12) remains a significant challenge. A Whittle index policy is known to be an efficient approach to solving RMAB problems which requires to satisfy a condition called indexability [46], [50]. A key challenge in solving problem (11)-(12) is that indexability is very difficult to establish. This difficulty arises due to the following reasons: (i) The state of each bandit of RMAB (11)-(12) exhibits a complicated transition, (ii) the transmission channels are unreliable, and (iii) the expected penalty associated with each bandit can be non-monotonic function of the AoI while most of the previous studies considered monotonic penalty functions of AoI [18], [21], [51], [52]. Hence, (11)-(12) is a more challenging problem than the problems studied in [15], [16], [18], [21], [51], [52]. However, we are able to develop a Maximum Gain First policy that does not need to satisfy indexability.

We solve problem (11)-(12) in three-steps: (i) We first relax constraint (12) and utilize Lagrangian dual decomposition to decompose the original problem into separated per-bandit problems; (ii) Next, we develop a Maximum Gain First policy that does not need to satisfy any indexability condition; (iii) Finally, we prove that the developed policy is asymptotically optimal.

1) Relaxation and Lagrangian Decomposition: In standard RMAB problems, the constraint (12) needs to be satisfied with equality, i.e., exactly M bandits are activated at any time slot t. However, in our problem, constraint (12) activates less than M bandits at any time t. Following [53, Section 5.1.1], [19, Section IV-A], we introduce M additional *dummy bandits* that never change state and therefore, they incur no cost. If a *dummy bandit* is activated, it occupies one channel but does not incur any cost. Let $\mu_0(t)\in\{1,2,\ldots,M\}$ be the number of *dummy bandits* that are activated at time slot t. After incorporating these *dummy bandits*, the RMAB (11)-(12) can be expressed as

$$\mathsf{L}_{\text{opt}}=\inf_{\pi\in\Pi}\limsup_{T\to\infty}\sum_{n=1}^{N}\mathbb{E}_\pi\left[\frac{1}{T}\sum_{t=0}^{T-1}H_L(Y_{n,t}|\Delta_n(t),X_{n,t-\Delta_n(t)})\right] \tag{13}$$

$$\text{s.t.}\sum_{n=0}^{N}\mu_n(t)=M,\mu_0(t)\in\{1,2,\ldots,M\},t=0,1,\ldots, \tag{14}$$

$$\mu_n(t)\in\{0,1\},n=1,2,\ldots,t=0,1,\ldots, \tag{15}$$

which is an RMAB with an equality constraint. Because the *dummy bandits* never change state, problem (11)-(12) and (13)-(14) are equivalent.

Next, we follow the standard relaxation and Lagrangian decomposition procedure for RMAB [46] and relax the constraint (14) and obtain the following relaxed problem:

$$\mathsf{L}_{\text{opt}}=\inf_{\pi\in\Pi}\limsup_{T\to\infty}\sum_{n=1}^{N}\mathbb{E}_\pi\left[\frac{1}{T}\sum_{t=0}^{T-1}H_L(Y_{n,t}|\Delta_n(t),X_{n,t-\Delta_n(t)})\right], \tag{16}$$

$$\text{s.t.}\limsup_{T\to\infty}\sum_{n=0}^{N}\mathbb{E}_\pi\left[\frac{1}{T}\sum_{t=0}^{T-1}\mu_n(t)\right]=M, \tag{17}$$

$$\mu_0(t)\in\{1,2,\ldots,M\},t=0,1,\ldots, \tag{18}$$

$$\mu_n(t)\in\{0,1\},n=1,2,\ldots,t=0,1,\ldots. \tag{19}$$

The relaxed constraint (19) needs to be satisfied on average, instead of satisfying at every time slot t. To solve the relaxed problem (16)-(19), we take a dual cost $\lambda\ge0$ (also known as Lagrange multiplier) for the relaxed constraint. The dual problem is given by

$$\sup_{\lambda\ge0}\bar{L}(\lambda), \tag{20}$$

where

$$\bar{L}(\lambda) = \inf_{\pi\in\Pi} \limsup_{T\to\infty} \mathbb{E}_{\pi}\Bigg[\sum_{n=1}^{N}\frac{1}{T}\sum_{t=0}^{T-1} H_L(Y_{n,t}|\Delta_n(t), X_{n,t-\Delta_n(t)}) + \lambda\Bigg(\sum_{n=0}^{N}\mu_n(t) - M\Bigg)\Bigg]. \quad (21)$$

The term $\frac{1}{T}\sum_{t=0}^{T-1}\sum_{n=0}^{N}\lambda M$ in (21) does not depend on policy π and hence can be removed. For a given λ, problem (21) can be decomposed into $(N+1)$ separated sub-problems and each sub-problem associated with agent n is formulated as

$$\bar{L}_n(\lambda) = \inf_{\pi_n\in\Pi_n} \limsup_{T\to\infty} \mathbb{E}_{\pi_n}\Bigg[\frac{1}{T}\sum_{t=0}^{T-1} H_L(Y_{n,t}|\Delta_n(t), X_{n,t-\Delta_n(t)}) + \lambda\mu_n(t)\Bigg], \quad (22)$$

where $\bar{L}_n(\lambda)$ is the optimum value of (22), $\pi_n = (\mu_n(0), \mu_n(1), \ldots)$ is the sub-scheduling policy for agent n, and Π_n is the set of all causal sub-scheduling policies of agent n. Problem (22) is a per-bandit problem associated with bandit n. On the other hand, the sub-problem associated with the *dummy bandits* is given by

$$\bar{L}_0(\lambda) = \inf_{\pi_0\in\Pi_0} \limsup_{T\to\infty} \mathbb{E}_{\pi_0}\Bigg[\frac{1}{T}\sum_{t=0}^{T-1}\lambda\mu_0(t)\Bigg], \quad (23)$$

where $\bar{L}_0(\lambda)$ is the optimum value of (23), $\pi_0 = \{\mu_0(t), t = 0, 1, \ldots\}$, and Π_0 is the set of all causal activation policies π_0.

VI. Maximum Gain First Policy

For a given transmission cost λ, the per-bandit problem (22) can be cast as an average-cost infinite horizon MDP with state $(\Delta_n(t), X_{n,t-\Delta_n(t)})$. The state associated with each bandit n is the latest received observation $X_{n,t-\Delta_n(t)}$ and its AoI $\Delta_n(t)$ at time slot t. The action is defined by $\mu_n(t) \in \{0, 1\}$ which denotes the scheduling decision for agent n at every time slot t. Each MDP associated with each bandit n has two actions: active and passive. If a packet from agent n is requested and submitted to a channel at time slot t, then restless bandit n takes an active action at time slot t; otherwise, bandit n is made passive at time slot t.

If bandit n is not scheduled for transmission (i.e., $\mu_n(t) = 0$), then the AoI will increase by 1, i.e., $\Delta_n(t) = \Delta_n(t-1)+1$ and the observation $X_{n,t-\Delta_n(t)}$ will remain in the same state. If bandit n is scheduled for transmission (i.e., $\mu_n(t) = 1$) and the transmission succeeds with probability p_n, then the AoI will drop to 1, i.e., $\Delta_n(t) = 1$ and the observation will change to a new received value $X_{n,t-1}$; otherwise, if transmission fails with probability $1 - p_n$, then the AoI will increase by 1, i.e., $\Delta_n(t) = \Delta_n(t-1) + 1$ and the observation will remain the same.

We solve (22) by using dynamic programming [17]. The Bellman optimality equation for the MDP in (22) is

$$h_{n,\lambda}(\delta, x) = \min_{\mu\in\{0,1\}} Q_{n,\lambda}(\delta, x, \mu), \quad (24)$$

where $h_{n,\lambda}(\delta, x)$ is the relative-value function of the average-cost MDP and $Q_{n,\lambda}(\delta, x, \mu)$ is the relative action-value function defined as

$$Q_{n,\lambda}(\delta, x, \mu) = \begin{cases} q_n(\delta, x) - \bar{L}_n(\lambda) + h_{n,\lambda}(\delta+1, x), & \text{if } \mu = 0, \\ q_n(\delta, x) - \bar{L}_n(\lambda) + (1-p_n)h_{n,\lambda}(\delta+1, x) & \\ \quad + p_n\mathbb{E}[h_{n,\lambda}(1, X_{n,0})|X_{n,-\delta} = x] + \lambda, & \text{otherwise}, \end{cases} \quad (25)$$

where $q_n(\delta, x)$ is given by

$$q_n(\delta, x) = H_L(Y_{n,t}|X_{n,t-\delta} = x), \quad (26)$$

and (25) holds because $X_{n,t}$ is a time-homogeneous Markov chain. The relative value function $h_{n,\lambda}(\delta, x)$ can be computed by using relative value iteration algorithm for average-cost MDP [17].

Let $\pi^*_{n,\lambda} = (\mu^*_{n,\lambda}(1), \mu^*_{n,\lambda}(2), \ldots)$ be an optimal solution to (22). The optimal decision at time slot t for bandit n is given by

$$\mu^*_{n,\lambda}(t) = \operatorname*{argmin}_{\mu\in\{0,1\}} Q_{n,\lambda}(\Delta_n(t), X_{n,t-\Delta_n(t)}, \mu), \quad (27)$$

where the dual cost is iteratively updated using the stochastic dual sub-gradient ascent method with step size $\beta > 0$ [54]:

$$\lambda(j+1) = \lambda(j) + \frac{\beta}{j}\Bigg(\sum_{n=0}^{N}\mu^*_{n,\lambda(j)}(j) - M\Bigg), \quad (28)$$

for j-th iteration. Let λ^* be the optimal dual cost to problem (20) to which $\lambda(t)$ converges. Then, we can apply $(\pi_{n,\lambda^*})_{n=0}^{N}$ for the relaxed problem (16)-(19). But applying this policy to the problem (13)-(14) may violate the constraint (14).

Definition 1 (Gain Index). Following [15], [40], we define the "gain" $\alpha_n(\delta, x)$ as

$$\alpha_n(\delta, x) = Q_{n,\lambda^*}(\delta, x, 0) - Q_{n,\lambda^*}(\delta, x, 1). \quad (29)$$

If $Q_{n,\lambda^*}(\delta, x, 0) > Q_{n,\lambda^*}(\delta, x, 1)$, i.e., $\alpha_n(\delta, x) > 0$, it is optimal to schedule bandit n. If $Q_{n,\lambda^*}(\delta, x, 0) < Q_{n,\lambda^*}(\delta, x, 1)$, i.e., $\alpha_n(\delta, x) < 0$, it is optimal to not to schedule bandit n.

Substituting (25) into (29), we get

$$\alpha_n(\delta, x) = p_n\Big(h_{n,\lambda^*}(\delta+1, x) - \mathbb{E}[h_{n,\lambda^*}(1, X_{n,0})|X_{n,-\delta} = x]\Big) - \lambda^*. \quad (30)$$

Lemma 2. *The following assertions are true:*
(i) $\lambda^* \geq 0$.
(ii) For the dummy bandits, it holds that $\alpha_0(\delta, x) = -\lambda^*$.

Proof. See Appendix D. □

The Algorithm for solving (13)-(14) is provided in Algorithm 1 which activates the M bandits with the highest "gain" index at any time slot t. As stated in Lemma 2, each dummy

Algorithm 1 Maximum Gain First Policy for Solving (13)-(14)

1: At time $t = 0$:
2: Input $\alpha_n(\delta, x)$ in (29) for every bandit $n = 1, 2, \ldots, N$.
3: Input $\alpha_0(\delta, x) = -\lambda^*$ for M dummy bandits using Lemma 2.
4: For all time $t = 0, 1, \ldots$,
5: Update $(\Delta_n(t), X_{n,t-\Delta_n(t)})$ for all bandits $n = 0, 1, \ldots, N$.
6: Update current "gain" $\alpha_n(\Delta_n(t), X_{n,t-\Delta_n(t)})$ for all bandits $n = 0, 1, \ldots, N$.
7: Choose M bandits with the highest "gain".

bandit has a "gain" index of $\alpha_0(\delta, x) = -\lambda^*$. Consequently, if a bandit n (for $n = 1, 2, ..., N$) possesses a "gain" smaller than $-\lambda^*$, denoted as $\alpha_n(\delta, x) < -\lambda^*$, it will remain inactive.

Next, we return to the RMAB (11)-(12). Because Lemma 2 implies that the dummy bandits have a "gain" smaller than $-\lambda^*$, we can obtain Algorithm 2 for solving (11)-(12) from Algorithm 1. The Algorithm for solving (11)-(12) is provided in Algorithm 2 that selects at most M bandits having the highest "gain" at time slot t.

Because (11)-(12) is equivalent to (3)-(4), Algorithm 2 exhibits the same performance for both (11)-(12) and (3)-(4). We prove the asymptotic optimality of the Maximum Gain First policy in Section VII.

VII. Asymptotic Optimality

In this section, we demonstrate that the "Maximum Gain First Policy" in Algorithm 2 is asymptotically optimal as the number of agents increases while maintaining the ratio between the number of agents and the number of channels are fixed. Let π_{gain} denote the policy presented in Algorithm 2. Following the standard practice, we consider a set of bandits to be in the same class if they share identical penalty functions and transition probabilities.

Definition 2 (Asymptotic optimality). Consider a "base" system with N bandits, M channels, and M dummy bandits. Let $\mathsf{L}^r_{\text{gain}}$ represent the long-term average cost under policy π_{gain} for a system that includes rN bandits, rM channels, and rM dummy bandits with $N+1$ class of bandits including one dummy bandit class with $r \in \mathbb{Z}^+$. The policy π_{gain} is asymptotically optimal if $\mathsf{L}^r_{\text{gain}} = \mathsf{L}^r_{\text{opt}}$ as the number of bandits in each class increases by r times.

We denote by $V^n_{\delta,x}(t)$ be the fraction of class n bandits in state (δ, x) at time t. Then, we define

$$v^n_{\delta,x} = \limsup_{T\to\infty} \sum_{t=0}^{T-1} \frac{1}{T}\mathbb{E}[V^n_{\delta,x}(t)]. \tag{31}$$

We further use the vectors $\mathbf{V}^n(t)$ and $\mathbf{v}^n$ to contain $V^n_{\delta,x}(t)$ and $v^n_{\delta,x}$, respectively for all $\delta = 1, 2, \ldots, \delta_{\text{bound}}$ and $x \in \mathcal{X}$. Truncated AoI space will have little to no impact if δ_{bound} is very large. This is because $L_n(\delta, x)$ for any $x \in \mathcal{X}$ achieves a stationary point as the AoI δ increases to a large value.

Algorithm 2 Maximum Gain First Policy for Solving (11)-(12)

1: At time $t = 0$:
2: Input $\alpha_n(\delta, x)$ in (29) for every bandit $n = 1, 2, \ldots, N$.
3: For all time $t = 0, 1, \ldots$,
4: Update $(\Delta_n(t), X_{n,t-\Delta_n(t)})$ for all bandits $n = 1, 2, \ldots, N$.
5: Update current "gain" $\alpha_n(\Delta_n(t), X_{n,t-\Delta_n(t)})$ for all bandits $n = 1, 2, \ldots, N$.
6: Choose at most M bandits with the highest "gain" that satisfies $\alpha_n(\delta, x) > -\lambda^*$.

For a policy π, we can have the following mapping

$$\Psi_\pi((\mathbf{v}^n)_{n=1}^N) = \mathbb{E}_\pi[(\mathbf{V}^n(t+1))_{n=1}^N | (\mathbf{V}^n(t))_{n=1}^N = (\mathbf{v}^n)_{n=1}^N]. \tag{32}$$

We define the t-th iteration of the maps $\Psi_{\pi,t\geq 0}(\cdot)$ as follows

$$\Psi_{\pi,0}((\mathbf{v}^n)_{n=1}^N) = (\mathbf{v}^n)_{n=1}^N, \tag{33}$$

$$\Psi_{\pi,t+1}((\mathbf{v}^n)_{n=1}^N) = \Psi_\pi(\Psi_{\pi,t}((\mathbf{v}^n)_{n=1}^N)). \tag{34}$$

Now, we are ready to define a global attractor condition.

Definition 3. Uniform Global attractor. An equilibrium point $(\mathbf{v}^n_{\text{opt}})_{n=1}^N$ given by the optimal solution of (11)-(12) is a uniform global attractor of $\Psi_{\pi,t\geq 0}(\cdot)$, i.e., for all $\epsilon > 0$, there exists $T(\epsilon)$ such that for all $t \geq T(\epsilon)$ and for all $(\mathbf{v}^n_{\text{opt}})_{n=1}^N)$, one has $||\Psi_{\pi,t}((\mathbf{v}^n)_{n=1}^N) - (\mathbf{v}^n_{\text{opt}})_{n=1}^N))||_1 \leq \epsilon$.

Theorem 2. *Under the uniform global attractor condition in Definition 3, the policy π_{gain} is asymptotically optimal.*

Proof. See Appendix E. □

Unlike the Whittle Index policy, we do not need to establish any indexability condition in the Maximum Gain First policy. However, this is still asymptotically optimal.

VIII. Special Case: Single-source, single-channel

Let us consider a special case with $N = M = 1$, where the system has one source and one channel. Then, problem (11)-(12) reduces to

$$\mathsf{L}_{1,\text{opt}} = \inf_{\pi_1\in\Pi_1} \limsup_{T\to\infty} \mathbb{E}_{\pi_1}\left[\frac{1}{T}\sum_{t=0}^{T-1} H_L(Y_{1,t}|\Delta_1(t), X_{1,t-\Delta_1(t)})\right]. \tag{35}$$

Problem (35) is an MDP that can be solved by Dynamic programming [17]. The optimal policy associated with agent 1 satisfies the following Bellman optimality equation:

$$\begin{aligned} J_1(\delta, x) =& H_L(Y_{1,\delta}|X_{1,0} = x) - \mathsf{L}_{1,\text{opt}} \\ &+ \min\{J_1(\delta+1, x), (1-p_1)J_1(\delta+1, x) \\ &+ p_1\mathbb{E}[J_1(1, X_{1,0})|X_{1,-\delta} = x]\}, \end{aligned} \tag{36}$$

where $J_1(\delta, x)$ is the value function associated with state (δ, x) and $\mathsf{L}_{1,\text{opt}}$ is the optimal value of (35).

As explained in [17], the optimal value function can be derived by using value iteration and the sequence of value functions $J_{1,k}(\delta, x)$ can be written as

$$\begin{aligned} J_{1,k+1}(\delta, x) =& H_L(Y_{1,\delta}|X_{n,0} = x) - \mathsf{L}_{n,\text{opt}} + \\ & \min\{J_{1,k}(\delta+1, x), (1-p_1)J_{1,k}(\delta+1, x) \\ & + p_1 \mathbb{E}[J_{1,k}(1, X_{1,0})|X_{1,-\delta} = x]\}, \end{aligned} \tag{37}$$

which converges to $\lim_{k\to\infty} J_{1,k} = J_1$ for any $J_{1,0}$. After some rearrangements, we can write (37) as

$$\begin{aligned} J_{1,k+1}(\delta, x) =& H_L(Y_{1,\delta}|X_{1,0} = x) - \mathsf{L}_{1,\text{opt}} + \\ & J_{1,k}(\delta+1, x) + p_1 \min\{0, -J_{1,k}(\delta+1, x) \\ & + \mathbb{E}[J_{1,k}(1, X_{1,0})|X_{1,-\delta} = x]\}. \end{aligned} \tag{38}$$

Then, we have the following lemma which illustrates sending is beneficial at every k.

Lemma 3. *For any k, it holds that $J_{1,k}(\delta+1, x) \geq \mathbb{E}[J_{1,k}(1, X_{1,0})|X_{1,-\delta} = x]$.*

Lemma 3 states that the penalty for not sending at iteration step k is higher than sending. Therefore, taking the active action is beneficial to reduce the penalty. One interesting observation from (38) is that each time a packet is successfully delivered with probability p_1, a new piece of information about the sensor signal value is added with the existing information $(X_{1,t-\delta} = x)$ (see the term $\mathbb{E}[J_{1,k}(1, X_{1,0})|X_{1,t-\delta} = x]$ in (37)). This new information plays a crucial role in reducing the system penalty and hence benefits the system through sending. In this sequel, we introduce the following useful lemma which illustrates that more information reduces the L-conditional entropy.

Lemma 4. *For random variables X, Y, and Z, it holds that $H_L(Y|Z = z) \geq H_L(Y|X, Z = z)$, where*

$$H_L(Y|Z = z) = \min_{a\in\mathcal{A}} \mathbb{E}[L(Y, a)|Z = z], \tag{39}$$

$$H_L(Y|X, Z = z) = \sum_{x\in\mathcal{X}} P(X = x|Z = z) H_L(Y|X = x, Z = z). \tag{40}$$

Proof. See Appendix F. □

Given Lemma 4, we are ready to prove Lemma 3. The proof of Lemma 3 is provided in Appendix G.

Theorem 3. *For single-source, single-channel case, the optimal policy π_1 in (35) chooses the active action at every time slot t.*

Proof. See Appendix H. □

Theorem 3 states that for single-source, single-channel case, it is always better to send. Though the penalty $L_1(\delta, x)$ is not necessarily a monotonic function of the age, the insights obtained from Lemma 4 tell us that having additional information helps reduce the average penalty.

IX. Discussions

The solution of problem (3)-(4) yields two key insights: (i) Our methodology used to solve problem (3)-(4) significantly reduces the state space compared to existing methods involving belief states, and (ii) The structure of the penalty function reveals an interesting relationship between an agent's proximity to safety boundaries and the required update frequency: specifically, agents near safety boundaries require frequent updating, whereas agents far from the safety boundaries require less frequent updating. The details are provided below:

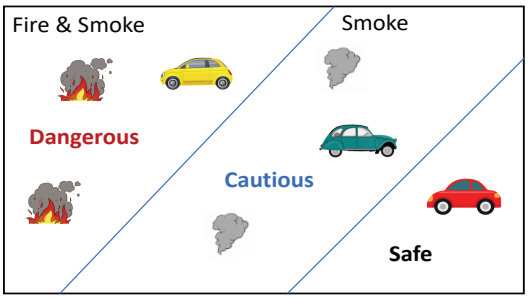

Fig. 3: Safety regions with three possible situations *dangerous, cautious,* and *safe*, where the event of fire indicates a *dangerous* situation, smoke indicates a *cautious* situation, and *safe*, otherwise.

A. Reduction in the size of the state-space and complexity

One major contribution of our study is the MDP formulation using the sufficient statistic of the history (i.e., the latest received observation $X_{n,t-\Delta_n(t)}$ and its AoI $\Delta_n(t)$) while existing studies utilized belief MDP or POMDP formulation. The belief states used in [6], [7], [9]–[12], [14], [15] help reduce the state space. However, the state space is still uncountable [6], [10]–[12]. Although some attempts have been made to make the state space countable under a positive recurrent assumption and using a sufficiently large truncated AoI value, the state space still exhibits a quadratic increase with the AoI [7], [9], [14], [15]. Specifically, for a truncated set $\{1, 2, \ldots, \tau\}$ of AoI values, the state space increases as $\tau \times |\mathcal{X}_n|^2$, where $X_{n,t} \in \mathcal{X}_n$ represents the n-th bandit process. The difference between the formulation in [7], [9], [14], [15] and problem (5)-(6) is that we do not need to utilize belief states. By utilizing the sufficient statistic of the history observation, i.e., the latest received observation and the corresponding AoI value, we obtain a significantly smaller state space which demonstrate linear growth with AoI, such as $\tau \times |\mathcal{X}_n|$. Consequently, the overall complexity achieved by our framework is $\mathcal{O}(|\mathcal{X}_n|)$, whereas, the existing approaches exhibit $\mathcal{O}(|\mathcal{X}_n|^2)$ complexity [7], [9], [14], [15].

B. Frequent Updates Near the Boundary

Analyzing the penalty function $q_n(\delta, x)$ for a given δ illustrates that at the boundary of each safety region, the scheduler should update frequently and at far from the boundary, the update can be less frequent.

To understand this characteristic, we do the following experiment: consider a safety-critical system where N agents (e.g., cars) are moving in a region illustrated in Figure 3. This region is equally divided into 400 positions and the received observation $X_{n,t}$ of agent n is represented by the

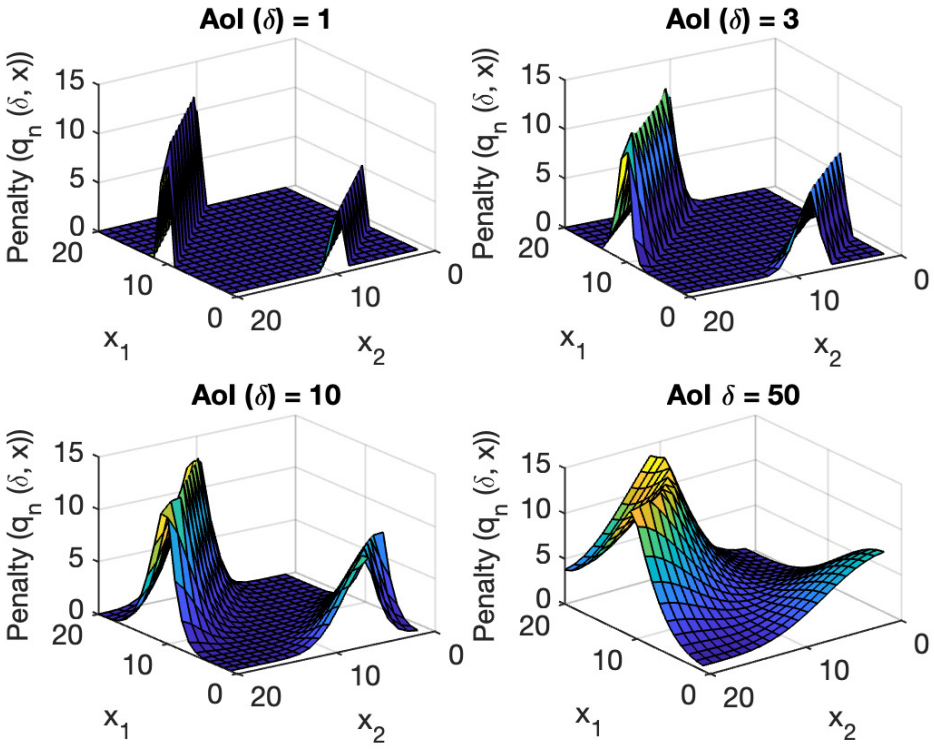

Fig. 4: Penalty $q_n(\delta, x)$ vs received observation x for fixed AoI δ.

position $x_n = (x_{n,1}, x_{n,2})$ at time t. The safety level $Y_{n,t}$ is divided into three regions: {*safe, cautious, dangerous*}. An agent n can randomly move in any of the four directions: *up, down, left*, and *right* with equal probability 0.2. If agent n is in the leftmost position, then moving left means it will stay in the same position, similar criteria are applied for the rightmost, upmost, and downmost positions. The losses considered in this experiment are: *L(cautious, safe)* $= 50$, *L(safe, cautious)* $= 1$, *L(dangerous, safe)* $= 200$, *L(safe, dangerous)* $= 1$, *L(dangerous, cautious)* $= 50$, *L(cautious, dangerous)*$= 10$, and *L(dangerous, dangerous)* $=$ *L(cautious, cautious)* $=$ *L(safe, safe)* $= 0$.

Figure 4 demonstrates the penalty $(q_n(\delta, x))$ vs received observation (x) graph for different AoI (δ) values. In this figure, when AoI is small, i.e., $\delta = 1$, the penalty is high only at the two boundary regions which implies that we need to update frequently if x is at any of the boundaries. Because near the boundaries, the uncertainty about the position of the agents is high, as a result, the scheduler has less situational awareness about the position of the agents. If x is far from the boundary, then the situational awareness is good, and less frequent updating does not harm the system performance. With increasing δ, the penalty graph spreads to the adjacent regions of the boundaries. Hence, the region that requires frequent updating is also increasing with increasing δ. This intuition is crucial for designing an efficient status updating policy that can maximize situational awareness and eventually minimize the loss due to the unawareness of potential danger.

X. Numerical Results

In this section, we evaluate the performance of the following policies:

- Periodic Updating: The agents generate updates at every time slot and store in a FIFO queue. Whenever a channel is available, an update from the queue is sent.
- Randomized Policy: If M channels are available, this policy randomly selects at most M agents.
- Maximum Age First (MAF) Policy: If M channels are available, this policy selects at most M agents with the highest AoI.
- Maximum Gain First (MGF) Policy: The policy provided in Algorithm 2.

We consider a safety-critical system with N agents navigating within a region. This region is uniformly divided into a 20 grid (20 rows and 20 columns). The observation $X_{n,t}$ for agent n at time t is represented by its position $x_n = (x_{n,1}, x_{n,2})$. The safety level $Y_{n,t}$ is categorized into three regions: safe, cautious, and dangerous. Rows 1 through 6 are designated as safe, rows 7 through 13 as cautious, and the remaining rows as dangerous. This setup differs from the one depicted in Figure 4, where all 400 positions represent distinct states, leading to a state space of size $\tau \times 400$, where τ is the truncated AoI value. This large state space significantly increases the complexity of determining the value function and gain index. To mitigate this, we simplify the representation by considering only the row information for safety assessment. Consequently, the total number of states is reduced from $\tau \times 400$ to $\tau \times 20$.

In this simulation, each agent n can move randomly in any of the eight cardinal and diagonal directions *up, up-left corner,*

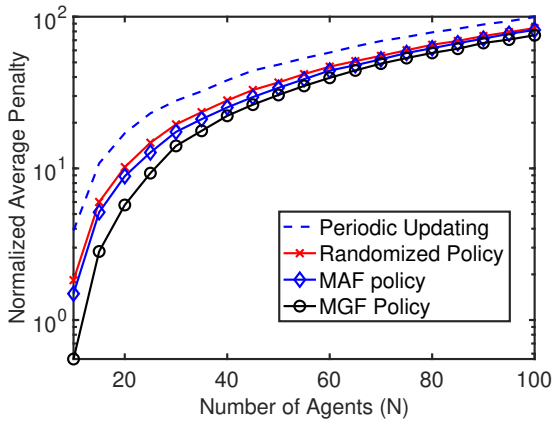

Fig. 5: Normalized average penalty vs Number of Agents (N) where Number of channel is $M = 1$ with success probability 0.95.

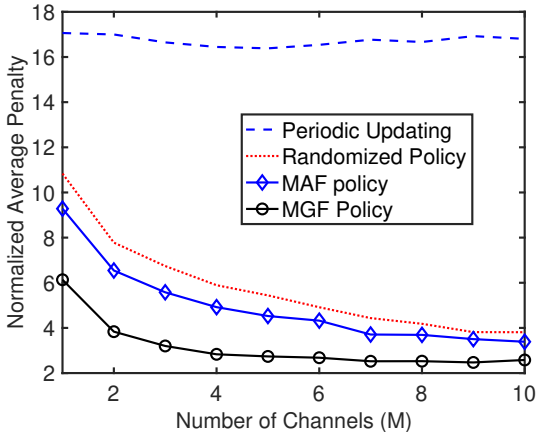

Fig. 6: Normalized average penalty vs Number of channels (M) where Number of agents are $N = 20$ with success probability 0.95.

up-right corner, down, down-left corner, down-right corner, left, and *right* with equal probability 0.105. Additionally, the agent does not move with probability 0.16. If an agent reaches a boundary, it stays in its current position. The loss incurred by agent n is given by $L_n(y, \hat{y}) = w_n L(y, \hat{y})$, where the weight w_n represents the importance of the agent n and the loss function L is defined as follows: *L(cautious, safe)* $= 50$, *L(safe, cautious)* $= 1$, *L(dangerous, safe)* $= 200$, *L(safe, dangerous)* $= 1$, *L(dangerous, cautious)* $= 50$, *L(cautious, dangerous)*$= 10$, and *L(dangerous, dangerous)* $=$ *L(cautious, cautious)* $=$ *L(safe, safe)* $= 0$. One agent is considered crucial with $w_1 = 10$, another 9 agents are considered less important with $w_n = 0$ for $n = 2, 3, \ldots, 10$, and the rest are assigned $w_n = 0.5n$. In our simulation, the success probability is 0.95.

Figure 5 illustrates the performance comparison of the four policies mentioned above as the number of sensors increases. We consider one erasure channel and the normalized average penalty (time-averaged penalty per agent) in Figure 5 is obtained by dividing time-average cost by the number of agents. From the figure, the MGF policy outperforms periodic updating, randomized policy, and MAF policy. The performance of periodic updating deteriorates as N increases due to growing queue lengths. In our simulation, we have used a buffer size of 20 for periodic updating. Furthermore, the randomized policy randomly selects one agent for sending updates and the MAF policy only utilizes the AoI in decision-making. In contrast, the MGF policy makes the decision in a smarter way by considering both the AoI and latest received observation, hence, shows better performance than the other three policies. The performance gain of the MGF policy is up to 7.03 times compared to periodic updating, up to 3.13 times compared to randomized policy, and up to 2.7 times compared to MAF policy.

Figure 6 illustrates the performance comparison of the four policies as the number of channels increases. We consider 20 agents in this simulation. With the increase of the number of available channels for sending updates, the performance of the policies are getting better. However, because of the intelligent decision strategy by utilizing the AoI and the latest received observation, the MGF policy outperforms the other three policies. The performance gain of the MGF policy is up to 6.52 times compared to periodic updating, up to 1.8 times compared to randomized policy, and up to 1.52 times compared to MAF policy.

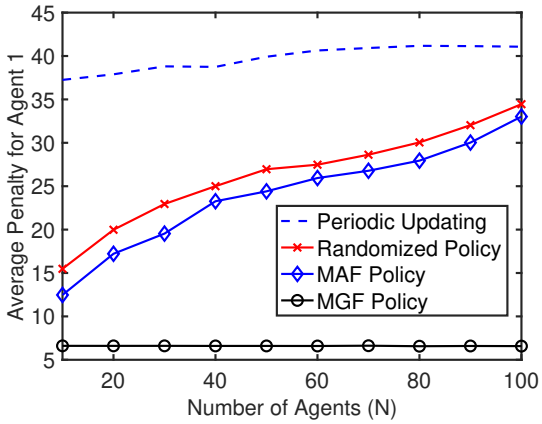

Fig. 7: Average penalty of Agent 1 vs Number of Agents (N) where Number of channels is $M = 2$ with success probability 0.95.

Figure 7 illustrates the time-averaged penalty incurred by Agent 1 as the number of agents N increases, with the number of channels M fixed at 2. In this scenario, we assign a weight of $w_1 = 10$ to Agent 1 and $w_n = 1$ to all other agents. This prioritizes Agent 1, making its safety the primary concern. Figure 7 reveals that the MGF policy achieves up to 5 times performance gain over MAF policy. Also, the time-averaged penalty for agent 1 remains constant under MGF policy as the number of agents increases. This suggests that the MGF policy allocates resources more effectively to the prioritized agent and MAF policy does not. This is because the AoI metric, used in MAF policy, does not explicitly consider agent importance, whereas the gain index, employed in MGF policy, does.

Figure 8 demonstrates that the gain ($\alpha_n(\delta, x)$) vs the received observation (x) follows the same pattern what we observed from Figure 4. In this simulation, we consider $N = 20$, $M = 10$, and $w_n = 1$ for all agents. In Figure 4, we observe that frequent updating is required at the safety boundary regions and with the increase of AoI, this region

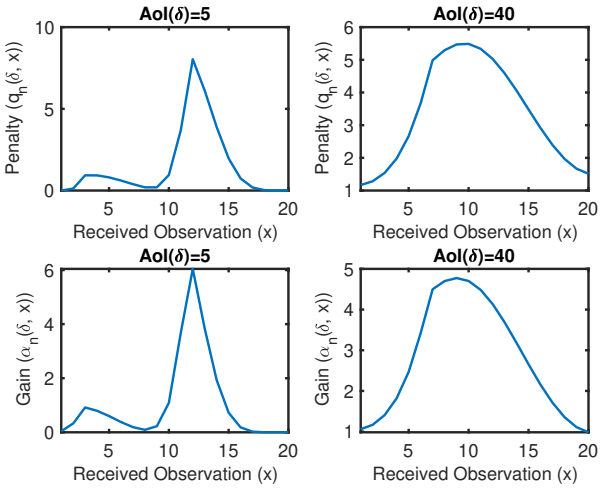

Fig. 8: Penalty $q_n(\delta, x)$ vs received observation x and Gain $\alpha_n(\delta, x)$ vs received observation x for fixed AoI δ. We consider number of agents $N = 20$, number of channels $M = 10$, success probability 0.95, and $w_n = 1$ for all agents

requiring frequent updates expands. Similar pattern is observed from Figure 8 that illustrates that $\alpha_n(\delta, x)$ is high at the safety boundaries and we need frequent updating.

XI. Conclusion

We address the importance of situational awareness in safety-critical systems. The developed scheduling policy requests updates more frequently when the received observation is near the safety boundaries. In future we will study systems where agents dynamically enter and exit at any time. Another interesting direction is to consider a finite time horizon problem where there is a termination state while encountering a danger.

References

[1] T. Z. Ornee, M. K. C. Shisher, C. Kam, and Y. Sun, "Context-aware status updating: Wireless scheduling for maximizing situational awareness in safety-critical systems," in *IEEE MILCOM*, 2023, pp. 194–200.

[2] A. Grau, M. Indri, L. L. Bello, and T. Sauter, "Industrial robotics in factory automation: From the early stage to the internet of things," in *IEEE IECON*, 2017, pp. 6159–6164.

[3] S. Abdulmalek, A. Nasir, W. A. Jabbar, M. A. Almuhaya, A. K. Bairagi, M. A.-M. Khan, and S.-H. Kee, "IoT-based healthcare-monitoring system towards improving quality of life: A review," in *Healthcare*, vol. 10, no. 10, 2022, p. 1993.

[4] M. Seenivasan, M. Arularasu, K. Senthilkumar, and R. Thirumalai, "Disaster prevention and control management in automation: a key role in safety engineering," *Procedia Earth and Planetary Science*, vol. 11, pp. 557–565, 2015.

[5] S. Kaul, R. D. Yates, and M. Gruteser, "Real-time status: How often should one update?" in *IEEE INFOCOM*, 2012.

[6] K. Liu and Q. Zhao, "Indexability of restless bandit problems and optimality of whittle index for dynamic multichannel access," *IEEE Trans. Inf. Theory*, vol. 56, no. 11, pp. 5547–5567, 2010.

[7] W. Ouyang, A. Eryilmaz, and N. B. Shroff, "Low-complexity optimal scheduling over time-correlated fading channels with arq feedback," *IEEE Trans. Mob. Comput.*, vol. 15, no. 9, pp. 2275–2289, 2016.

[8] P. Jacko and S. S. Villar, "Opportunistic schedulers for optimal scheduling of flows in wireless systems with arq feedback," in *IEEE ITC*, 2012, pp. 1–8.

[9] W. Ouyang, A. Eryilmaz, and N. B. Shroff, "Asymptotically optimal downlink scheduling over markovian fading channels," in *IEEE INFOCOM*, 2012, pp. 1224–1232.

[10] S. H. A. Ahmad, M. Liu, T. Javidi, Q. Zhao, and B. Krishnamachari, "Optimality of myopic sensing in multichannel opportunistic access," *IEEE Trans. Inf. Theory*, vol. 55, no. 9, pp. 4040–4050, 2009.

[11] P. Ansell, K. Glazebrook, J. Niño-Mora, and M. O'Keeffe, "Whittle's index policy for a multi-class queueing system with convex holding costs," *Math. Meth. Oper. Res.*, vol. 57, pp. 21–39, 04 2003.

[12] C.-p. Li and M. J. Neely, "Exploiting channel memory for multi-user wireless scheduling without channel measurement: Capacity regions and algorithms," in *ACM MobiHoc*, 2010, pp. 50–59.

[13] S. Murugesan, P. Schniter, and N. B. Shroff, "Multiuser scheduling in a markov-modeled downlink using randomly delayed arq feedback," *IEEE Trans. Inf. Theory*, vol. 58, no. 2, pp. 1025–1042, 2012.

[14] Y. Chen and A. Ephremides, "Scheduling to minimize age of incorrect information with imperfect channel state information," *Entropy*, vol. 23, no. 12, p. 1572, 2021.

[15] G. Chen and S. C. Liew, "An index policy for minimizing the uncertainty-of-information of Markov sources," *arXiv preprint arXiv:2212.02752*, 2022.

[16] 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.

[17] D. P. Bertsekas *et al.*, "Dynamic programming and optimal control, 4th edition," *Belmont, MA: Athena Scientific*, vol. 1, 2011.

[18] V. Tripathi and E. Modiano, "A Whittle index approach to minimizing functions of age of information," in *IEEE Allerton*, 2019, pp. 1160–1167.

[19] T. Z. Ornee and Y. Sun, "A Whittle index policy for the remote estimation of multiple continuous Gauss-Markov processes over parallel channels," in *ACM MobiHoc*, 2023, p. 91–100.

[20] M. K. C. Shisher, Y. Sun, and I.-H. Hou, "Timely communications for remote inference," *IEEE/ACM Trans. Netw.*, 2024, in press.

[21] 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.

[22] 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.

[23] 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.

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

[25] 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, 2020.

[26] M. Klügel, M. H. Mamduhi, S. Hirche, and W. Kellerer, "AoI-Penalty minimization for networked control systems with packet loss," in *IEEE INFOCOM WKSHPS*, 2019, pp. 189–196.

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

[28] J. Zhong, R. D. Yates, and E. Soljanin, "Two freshness metrics for local cache refresh," in *IEEE ISIT*, 2018, pp. 1924–1928.

[29] 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, p. 2215–2228, oct 2020.

[30] 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.

[31] R. D. Yates, "The age of gossip in networks," in *IEEE ISIT*, 2021, pp. 2984–2989.

[32] J. Holm, A. E. Kalør, F. Chiariotti, B. Soret, S. K. Jensen, T. B. Pedersen, and P. Popovski, "Freshness on demand: Optimizing age of information for the query process," in *IEEE ICC*, 2021, pp. 1–6.

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

[34] M. E. Ildiz, O. T. Yavascan, E. Uysal, and O. T. Kartal, "Query age of information: Optimizing AoI at the right time," in *IEEE ISIT*, 2022, pp. 144–149.

[35] ——, "Pull or wait: How to optimize query age of information," *IEEE J. Sel. Areas Inf. Theory*, pp. 1–1, 2023.

[36] A. Kosta, N. Pappas, A. Ephremides, and V. Angelakis, "Age and value of information: Non-linear age case," in *IEEE ISIT*, 2017, pp. 326–330.

[37] Z. Wang, M.-A. Badiu, and J. P. Coon, "A framework for characterizing the value of information in hidden Markov models," *IEEE Trans. Inf. Theory*, vol. 68, no. 8, pp. 5203–5216, 2022.

[38] T. Soleymani, S. Hirche, and J. S. Baras, "Optimal self-driven sampling for estimation based on value of information," in *IEEE WODES*, 2016, pp. 183–188.

[39] M. K. C. Shisher, H. Qin, L. Yang, F. Yan, and Y. Sun, "The age of correlated features in supervised learning based forecasting," in *IEEE INFOCOM Workshops*, 2021, pp. 1–8.

[40] 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 Inf. Theory*, vol. 4, pp. 524–538, 2023.

[41] G. Xiong, X. Qin, B. Li, R. Singh, and J. Li, "Index-aware reinforcement learning for adaptive video streaming at the wireless edge," in *ACM MobiHoc*, 2022, pp. 81–90.

[42] Y. Zou, K. T. Kim, X. Lin, and M. Chiang, "Minimizing age-of-information in heterogeneous multi-channel systems: A new partial-index approach," in *ACM MobiHoc*, 2021, pp. 11–20.

[43] N. Akbarzadeh and A. Mahajan, "Restless bandits with controlled restarts: Indexability and computation of Whittle index," in *IEEE CDC*, 2019, pp. 7294–7300.

[44] F. Li, Y. Sang, Z. Liu, B. Li, H. Wu, and B. Ji, "Waiting but not aging: Optimizing information freshness under the pull model," *IEEE/ACM Trans. Netw.*, vol. 29, no. 1, pp. 465–478, 2021.

[45] T. Z. Ornee and Y. Sun, "Performance bounds for sampling and remote estimation of Gauss-Markov processes over a noisy channel with random delay," in *IEEE SPAWC*, 2021, pp. 1–5.

[46] P. Whittle, "Restless bandits: activity allocation in a changing world," *Journal of Applied Probability*, vol. 25A, pp. 287–298, 1988.

[47] A. P. Dawid, "Coherent measures of discrepancy, uncertainty and dependence, with applications to Bayesian predictive experimental design," *Department of Statistical Science, University College London*, vol. 139, 1998.

[48] F. Farnia and D. Tse, "A minimax approach to supervised learning," *NIPS*, vol. 29, 2016.

[49] M. K. C. Shisher and Y. Sun, "How does data freshness affect real-time supervised learning?" in *ACM MobiHoc*, 2022, pp. 31–40.

[50] R. R. Weber and G. Weiss, "On an index policy for restless bandits," *Journal of applied probability*, vol. 27, no. 3, pp. 637–648, 1990.

[51] A. Kosta, N. Pappas, A. Ephremides, and V. Angelakis, "Age and value of information: Non-linear age case," in *2017 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2017, pp. 326–330.

[52] 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 Transactions on Information Theory*, vol. 63, no. 11, pp. 7492–7508, 2017.

[53] I. M. Verloop, "Asymptotically optimal priority policies for indexable and nonindexable restless bandits," 2016.

[54] A. Nedic and A. Ozdaglar, "Subgradient methods in network resource allocation: Rate analysis," in *IEEE CISS*, 2008, pp. 1189–1194.

[55] N. Gast, B. Gaujal, and C. Yan, "Linear program-based policies for restless bandits: Necessary and sufficient conditions for (exponentially fast) asymptotic optimality," *Mathematics of Operations Research*, 2023.

APPENDIX A
DEFINITIONS OF 0-1 LOSS, QUADRATIC LOSS, LOGARITHMIC LOSS

0-1 loss: The 0-1 loss function assigns a loss of 0 for incorrect estimation of a random variable $Y = y$, and 1 otherwise. It is given by

$$L_{0-1}(y, \hat{y}) = \mathbf{1}\{y \neq \hat{y}\}, \tag{41}$$

where $\mathbf{1}\{y \neq \hat{y}\}$ is the indicator function for the event $\{y \neq \hat{y}\}$.

Quadratic loss: The quadratic loss function quantifies the error between the true value of a random variable $Y = y$ and and its estimated value $\hat{y}$ by calculating the square of their difference. It is given by

$$L_2(y, \hat{y}) = (y - \hat{y})^2. \tag{42}$$

Logarithmic loss: The log-loss function $L_{\log}(y, P_Y)$ is the negative log-likelihood of the true value $Y = y$ which is given by

$$L_{\log}(y, P_Y) = \log P_Y(y), \tag{43}$$

where the action $a = P_Y$ is a distribution of Y.

APPENDIX B
PROOF OF THEOREM 1

Let $J((\mathbf{H}_{n,t})_{n=1}^N) \in \mathbb{R}$ be the value function of the average cost MDP (3)-(4), which is given by

$$J((\mathbf{H}_{n,t})_{n=1}^N) = \min_{\pi\in\Pi} \sum_{k=t}^{\infty} \sum_{n=1}^{N} \mathbb{E}\left[L_n(Y_{n,t}, \phi_n(\mathbf{H}_{n,k})) - \mathsf{L}_{\text{opt}} \middle| \mathbf{H}_{n,t} \right]. \tag{44}$$

Because $Y_{n,t}$ is independent of $\mathbf{H}_{n,t}$ given $(\Delta_n(t), X_{n,t-\Delta_n(t)})$, we have

$$\begin{aligned} \phi_n(\mathbf{H}_{n,t}) &= \operatorname*{argmin}_{\hat{y}} \mathbb{E}[L(Y_{n,t}, \hat{y}) | \mathbf{H}_{n,t}] \\ &= \operatorname*{argmin}_{\hat{y}} \mathbb{E}[L(Y_{n,t}, \hat{y}) | \Delta_n(t), X_{n,t-\Delta_n(t)}] \\ &= f_n(\Delta_n(t), X_{n,t-\Delta_n(t)}), \end{aligned} \tag{45}$$

where (45) follows from (7). Hence, by substituting (45) into (44), we get

$$\begin{aligned} &J((\mathbf{H}_{n,t})_{n=1}^N) = \\ &\min_{\pi\in\Pi} \sum_{k=t}^{\infty} \sum_{n=1}^{N} \mathbb{E}\Big[L_n(Y_{n,t}, f_n(\Delta_n(k), X_{n,k-\Delta_n(k)})) - \mathsf{L}_{\text{opt}} \\ &\qquad\qquad \Big| \mathbf{H}_{n,t} \Big]. \end{aligned} \tag{46}$$

Furthermore, since $(\Delta_n(k), X_{n,k-\Delta_n(k)})$ is independent of $\mathbf{H}_{n,t}$ given $(\Delta_n(t), X_{n,t-\Delta_n(t)})$ for any $k \geq t$, we can write

$$\begin{aligned} &J((\mathbf{H}_{n,t})_{n=1}^N) \\ &= \min_{\pi\in\Pi} \sum_{k=t}^{\infty} \sum_{n=1}^{N} \mathbb{E}\Big[L_n(Y_{n,t}, f_n(\Delta_n(k), X_{n,k-\Delta_k(t)})) - \mathsf{L}_{\text{opt}} \\ &\qquad\qquad \Big| \mathbf{H}_{n,t} \Big] \\ &= \min_{\pi\in\Pi} \sum_{k=t}^{\infty} \sum_{n=1}^{N} \mathbb{E}\Big[L_n(Y_{n,t}, f_n(\Delta_n(k), X_{n,k-\Delta_k(t)})) - \mathsf{L}_{\text{opt}} \\ &\qquad\qquad \Big| \Delta_n(t), X_{n,t-\Delta_n(t)} \Big]. \end{aligned} \tag{47}$$

From (47), it is evident that the value function $J((\mathbf{H}_{n,t})_{n=1}^N)$ can be uniquely determined by $(\Delta_n(t), X_{n,t-\Delta_n(t)})_{n=1}^N$. Therefore, from [17, Chapter 4.3], Theorem 1 follows.

APPENDIX C
PROOF OF LEMMA 1

Because $f_n(\delta, x)$ in (7) is the optimal estimator, we can write the penalty function as

$$\begin{aligned}&\min_{\hat{y}\in\mathcal{Y}} \mathbb{E}_{Y\sim P_{Y_{n,t}|\Delta_n(t)=\delta, X_{n,t-\Delta_n(t)}=x}}[L_n(Y,\hat{y})]\\ =&\min_{\hat{y}\in\mathcal{Y}} \mathbb{E}_{Y\sim P_{Y_{n,t}|\Delta_n(t)=\delta, X_{n,t-\delta}=x}}[L_n(Y,\hat{y})].\end{aligned} \tag{48}$$

Because $Y_{n,t}$ is related to $X_{n,t}$ according to Figure 2 and is conditionally independent of $\Delta_n(t)$ given $X_{n,t-\Delta_n(t)}$, (48) can be written as

$$\begin{aligned}&\min_{\hat{y}\in\mathcal{Y}} \mathbb{E}_{Y\sim P_{Y_{n,t}|\Delta_n(t)=\delta, X_{n,t-\Delta_n(t)}=x}}[L_n(Y,\hat{y})]\\ =&\min_{\hat{y}\in\mathcal{Y}} \mathbb{E}_{Y\sim P_{Y_{n,t}|X_{n,t-\delta}=x}}[L_n(Y,\hat{y})].\end{aligned} \tag{49}$$

Following (9), the generalized conditional entropy $H_L(Y_{n,t}|X_{n,t-\delta}=x)$ given the latest received observation at time slot t which is generated δ times ago is given by

$$H_L(Y_{n,t}|X_{n,t-\delta}=x)=\min_{\hat{y}\in\mathcal{Y}} \mathbb{E}_{Y\sim P_{Y_{n,t}|X_{n,t-\delta}=x}}[L_n(Y,\hat{y})]. \tag{50}$$

Comparing (49) and (50), Lemma 1 follows.

APPENDIX D
PROOF OF LEMMA 2

Because λ represents the cost to activate a bandit, it is optimal to activate a dummy bandit only when $\lambda < 0$. In addition, when $\lambda < 0$, it is optimal to activate all bandits. Then, in (28), the constraint will be higher than 0, i.e., $\sum_{n=1}^{N}\mu_n(t) - M > 0$. Hence, $\lambda(j+1)$ will continue to increase in every iteration step j and will never converge. Conversely, when $\lambda \geq 0$, it is optimal not to activate the dummy bandit. Combining these two conditions illustrate that $\lambda^* \geq 0$.

On the other hand, from (30) of Definition 1 and the fact that the dummy bandits are activated only when $\lambda < 0$, we get $\alpha_0(\delta, x) = -\lambda$.

APPENDIX E
PROOF OF THEOREM 2

We first present a set of LP-based priority policies that achieve asymptotically optimality under the uniform global attractor condition in Definition 3. Subsequently, we demonstrate that π_{gain} belongs to this set of priority policies. Let $U_{\delta,x}^{n,\mu}(t)$ be the number of class n bandits in state δ, x taking action μ. We define $u_{\delta,x}^{n,\mu}$ as the expected number of class n bandits in state δ, x taking action μ, given by

$$u_{\delta,x}^{n,\mu} = \limsup_{T\to\infty}\sum_{t=0}^{T-1}\frac{1}{T}\mathbb{E}[U_{\delta,x}^{n,\mu}(t)]. \tag{51}$$

Let $\bar{u}_{\delta,x}^{n,\mu}$ be the optimal state action frequency of the Lagrangian problem (21) in which $\lambda = \lambda^*$ is the optimal dual Lagrangian multiplier.

Define the following sets

$$\mathcal{S}_+^n = \{(\delta,x) : \bar{u}_{\delta,x}^{n,1} > 0 \text{ and } \bar{u}_{\delta,x}^{n,0} = 0\}, \tag{52}$$

$$\mathcal{S}_0^n = \{(\delta,x) : \bar{u}_{\delta,x}^{n,1} > 0 \text{ and } \bar{u}_{\delta,x}^{n,0} > 0\}, \tag{53}$$

$$\mathcal{S}_-^n = \{(\delta,x) : \bar{u}_{\delta,x}^{n,1} < 0 \text{ and } \bar{u}_{\delta,x}^{n,0} > 0\}. \tag{54}$$

Definition 4. LP-based Priority Policies. [53], [55] We define a set $\Pi_{\text{LP-Priority}}$ that consists of priority policies that satisfy the following conditions:

i) A class-k bandit in state $(\delta_k, x_k) \in \mathcal{S}_+^k$ is given high priority than a class-j bandit in state $(\delta_j, x_j) \in \mathcal{S}_0^j$.

ii) A class-k bandit in state $(\delta_k, x_k) \in \mathcal{S}_0^k$ is given high priority than a class-j bandit in state $(\delta_j, x_j) \in \mathcal{S}_-^j$.

iii) If a class-k bandit is activated, then $\mu = 1$ is chosen; otherwise, $\mu = 0$ is chosen.

Lemma 5. *For any policy $\pi \in \Pi_{\text{LP-Priority}}$, if the uniform global attractor condition in Definition 3 is satisfied, then the policy is asymptotically optimal.*

By utilizing [55, Theorem 13], we can obtain Lemma 5. Following [55, Proposition 14] , for class-k bandits

1. For any class-k bandit, $(\delta_k, x_k) \in \mathcal{S}_+$ implies

$$\alpha_n(\delta_k, x_k) > 0. \tag{55}$$

2. For any class-k bandit, $(\delta_k, x_k) \in \mathcal{S}_0$ implies

$$\alpha_n(\delta_k, x_k) = 0. \tag{56}$$

3. For any class-k bandit, $(\delta_k, x_k) \in \mathcal{S}_-$ implies

$$\alpha_n(\delta_k, x_k) < 0. \tag{57}$$

Because the policy π_{gain} activates exactly M bandits with the highest gain and if a bandit k is activated, π_{gain} chooses $\mu = 1$, we can deduce form (55)-(57) and Definition 4 that π_{gain} belongs to $\Pi_{\text{LP-Priority}}$. This concludes the proof.

APPENDIX F
PROOF OF LEMMA 4

From the definition of L-conditional entropy in (9), we get that

$$\begin{aligned}&H_L(Y|Z=z)\\ =&\min_{a\in\mathcal{A}}\mathbb{E}[L(Y,a)|Z=z]\\ =&\min_{a\in\mathcal{A}}\sum_{y\in\mathcal{Y}}P(Y=y|Z=z)L(y,a)\\ =&\min_{a\in\mathcal{A}}\sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}}P(Y=y|X=x,Z=z)P(X=x|Z=z)L(y,a)\\ =&\min_{a\in\mathcal{A}}\sum_{x\in\mathcal{X}}P(X=x|Z=z)\sum_{y\in\mathcal{Y}}P(Y=y|X=x,Z=z)L(y,a)\\ \geq&\sum_{x\in\mathcal{X}}P(X=x|Z=z)\min_{a\in\mathcal{A}}\sum_{y\in\mathcal{Y}}P(Y=y|X=x,Z=z)L(y,a),\end{aligned} \tag{58}$$

where (58) holds because $\min(f(w)+g(w)) \geq \min f(w) + \min g(w)$ for all w. Continuing from (58), we get that

$$\begin{aligned} &H_L(Y|Z=z) \\ &\geq \sum_{x\in\mathcal{X}} P(X=x|Z=z) \min_{a\in\mathcal{A}} \mathbb{E}[L(Y,a)|X=x,Z=z]. \end{aligned} \tag{59}$$

Utilizing (39), we obtain that [47]–[49]

$$H_L(Y|X=x,Z=z) = \min_{a\in\mathcal{A}} \mathbb{E}[L(Y,a)|X=x,Z=z]. \tag{60}$$

Substituting (60) into (59) yields

$$\begin{aligned} &H_L(Y|Z=z) \\ &\geq \sum_{x\in\mathcal{X}} P(X=x|Z=z) H_L(Y|X=x,Z=z), \\ &= H_L(Y|X,Z=z), \end{aligned} \tag{61}$$

where (61) follows from (40). This completes the proof.

Appendix G
Proof of Lemma 3

Without loss of generality, we can assume that for all (δ, x), $J_{n,0}(\delta,x)=0$. At $k=0$, (37) becomes

$$J_{n,1}(\delta,x) = H_L(Y_{n,t}|\Delta_n(t)=\delta, X_{n,t-\Delta_n(t)}=x) - \mathsf{q}_{\text{opt}}. \tag{62}$$

In this case, a minimum value can be achieved by both sending and not sending. Hence, we can conclude that sending is beneficial at iteration step k=0. Next, at $k=1$, (37) becomes

$$\begin{aligned} J_{n,2}(\delta,x) =& H_L(Y_{n,\delta}|X_{n,0}=x) - \mathsf{L}_{n,\text{opt}} \\ &+ \min\{J_{n,1}(\delta+1,x), (1-p_n)J_{n,1}(\delta+1,x) \\ &+ p_n \mathbb{E}[J_{n,1}(1,X_{n,0})|X_{n,-\delta}=x]\}. \end{aligned} \tag{63}$$

Sending will be beneficial at $k=1$ if

$$J_{n,1}(\delta+1,x) \geq \mathbb{E}[J_{n,1}(1,X_{n,0})|X_{n,-\delta}=x]. \tag{64}$$

From the right-side term in (64), we get

$$\begin{aligned} &\mathbb{E}[J_{n,1}(1,X_{n,0})|X_{n,-\delta}] \\ &= \sum_{z\in\mathcal{X}} J_{n,1}(1,z) P(X_{n,0}=z|X_{n,-\delta}=x) \\ &= \sum_{z\in\mathcal{X}} H_L(Y_{n,t}|X_{n,0}=z) P(X_{n,0}=z|X_{n,-\delta}=x) - \mathsf{L}_{n,\text{opt}} \\ &= H_L(Y_{n,t}|X_{n,0}, X_{n,-\delta}=x) - \mathsf{L}_{n,\text{opt}}, \end{aligned} \tag{65}$$

where (65) holds from Lemma 4.

In (62), no new information is obtained because $J_{n,1}(\delta+1,x) = H_L(Y_{n,\delta+1}|X_{n,0}=x) - \mathsf{L}_{n,\text{opt}}$. However, in (65), a new piece of information is added. By comparing (62) and (65) and by utilizing Lemma 4, we can conclude that

Next, assume that this result holds up to iteration step k so that the following is true:

$$J_{n,k}(\delta+1,x) \geq \mathbb{E}[J_{n,k}(1,X_{n,0})|X_{n,-\delta}=x]\}. \tag{66}$$

By using the result in (66), we have to prove that for iteration step $k+1$, the following holds.

$$J_{n,k+1}(\delta+1,x) \geq \mathbb{E}[J_{n,k+1}(1,X_{n,0})|X_{n,-\delta}=x]\}. \tag{67}$$

The right side of (67) contains $H_L(Y_{n,k}|X_{n,0},X_{n,1},\ldots,X_{n,k},X_{n,-\delta}=x)$ which has more information than the left-side term. In the similar way, it can shown that all the terms in the right side of (67) has more information than left-side. By utilizing Lemma 4, we can conclude that sending is beneficial at iteration step $k+1$.

Hence, it is beneficial to send for all k if a channel is available. This concludes the proof.

Appendix H
Proof of Theorem 3

From Lemma 3, we get that sending is beneficial at every iteration step k. Hence, sending is beneficial at every time-slot. Therefore, Theorem 3 follows from Lemma 3.