Lagrange Index based Scheduling for Minimizing Age of Updates from Heterogeneous Sources

Aniket Mukherjee, Chandramani Singh, Joy Kuri
Department of Electronic Systems Engineering, Indian Institute of Science, Bangalore-560012, India
Email: {maniket, chandra, kuri}@iisc.ac.in

***Abstract*—Modern sensing systems generate heterogeneous updates ranging from small status packets to large data objects. We study a single-hop wireless uplink network where sensors generate updates at will, each consisting of a sensor dependent number of packets. Under a strict medium-access constraint and non-preemptive (no-switching) transmissions, decision stages become action-dependent and stochastic. We formulate the problem as a restless multi-armed bandit (RMAB) with semi-Markov decision process (SMDP) dynamics and develop a Lagrange index based heuristic for minimizing weighted average AoI cost. For the weighted AoI setting, we utilize the structural properties of the heuristic to enable efficient index computation. Numerical results demonstrate consistent performance gains over existing non-preemptive scheduling policies, providing a practical solution for heterogeneous freshness-aware systems.**

***Index Terms*—Age of information, Semi-Markov decision process, Restless multiarmed bandit, Lagrange indices.**

I. Introduction

Modern sensing systems consist of spatially distributed sensors that continuously monitor physical processes and transmit status updates to a central controller or edge server [1]. These updates are often heterogeneous in size, ranging from small scalar measurements to large data objects such as images, video frames, or LIDAR scans. In slotted wireless systems (eg., Wi-Fi 6 and 5G), such updates may span multiple transmission slots, and medium-access constraints limit simultaneous transmissions. As sensing networks increasingly support real-time monitoring, autonomous systems, and cyber-physical applications, efficient scheduling of heterogeneous updates becomes critical to maintaining timely situational awareness [2].

The Age of Information (AoI) quantifies information freshness by measuring the time elapsed since the most recently generated update was successfully delivered [2]–[4]. Unlike traditional delay or throughput metrics, AoI directly captures the timeliness of the received information, making it particularly suitable for sensing and monitoring applications where stale data can degrade estimation accuracy, control performance or tracking reliability. Although AoI is a theoretical metric, its practical value has been demonstrated experimentally; for example, [5] shows that an AoI-aware WiFi middleware significantly improves information freshness and tracking accuracy in UAV networks compared to standard WiFi-UDP/TCP.

We consider a set of heterogeneous sources connected to a common receiver through an unreliable wireless channel that allows only one source to transmit at a time (see Figure 1). The sources generate updates of different sizes, which must be delivered to the receiver so as to minimize the long term weighted average age of updates at the receiver. A non-preemptive (non-switching) transmission discipline is employed where a scheduled source keeps the channel until it successfully transmits all the packet pertaining to its update.

We formulate the problem as a Semi-Markov Decision Process (SMDP). However, this problem suffers from the curse of dimensionality. To address it, we treat the problem as a Restless Multi-Armed Bandit (RMAB) and develop a Lagrange index based heuristic. To the best of our knowledge, this is the first work to study RMAB formulation of weakly coupled SMDPs and to develop scalable index policies for them. Our heuristic scheduling policy substantially outperforms the existing solutions.

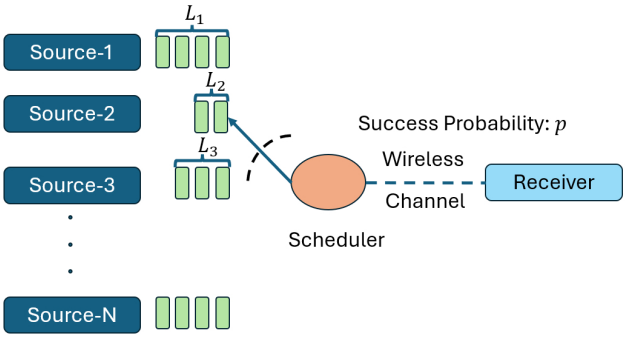


Fig. 1: N sources with heterogeneous update sizes, connected to a receiver through an unreliable wireless channel.

A. Related Work

We now briefly discuss the related work on (a) AoI minimization and on (b) Lagrange index based heuristics for RMAB problems.

1) AoI Minimization: The concept of AoI was introduced in [3], and a comprehensive overview of its extensions and practical relevance was provided in [6]. The papers concerned with AoI minimization can be categorized as in Table I. There are two dimensions: (a) Whether all source updates are of equal length or not, and (b) whether one can interrupt transmission of an update from a source before it is complete, and "switch" to transmitting another source's update ["switching"], or not ["non-switching"]. [7] studied a wireless network in which a

ISBN 978-3-903176-82-9 © 2026 IFIP

base station generated single-packet updates and, in each slot, selected a user for transmission over an unreliable channel with user-dependent success probabilities p_i (constant over time). The objective was to minimize the long-term average AoI. The greedy policy was shown to be optimal when $p_i = p$ for all i. For the asymmetric case, a Whittle index and a max-weight policy were proposed. In [8], for equal update lengths, the same model as in [7] was considered, but with a nonlinear function of AoI as the objective. The authors established indexability and proposed a Whittle index–based policy. In reality update lengths can be different and we consider differing update lengths in our work.

For heterogeneous sources with reliable channels, [9] considered an AoI model in which each source requires a fixed processing time, followed by a fixed transmission time under a non-preemptive service discipline. They established indexability and proposed a Whittle index policy. The problem was formulated as a restless multi-arm bandit and decoupled into single-source Markov Decision Process (MDP). In the resulting single-source MDP, only two actions were considered: when the tagged source was scheduled, the AoI reset to the total processing and transmission time; when it was not scheduled, the AoI increased by one. In heterogeneous systems, this assumption is restrictive, since the cost when the tagged source was not scheduled depends on service duration of the scheduled source.

In [10] and [11], the same uplink model as ours (Figure 1) was considered. [10] considers policies that allowed switching and suggested a suboptimal policy to minimize the average AoI. For the non-preemptive setting, [11] proposed a stationary randomized policy, referred to as the No-Switching Randomized Policy (NSRP).

TABLE I: Related Works

	Equal length update	**Unequal length update**
Switching		[11], [10]
Non-switching	[7], [8]	[11], [9]

2) Lagrange Indices for RMABs: Lagrange index policies for restless multi-armed bandits have been studied through linear programming relaxations in [12], [13], where the dual decomposition yields index-type policies and asymptotic optimality was established for large-scale systems with underlying Markov Decision Process (MDP) dynamics. The framework was further extended in [14] to settings with unknown transition kernels, preserving asymptotic optimality via learning-based methods. Our setting differs fundamentally in that we consider restless bandits under an SMDP formulation induced by non-preemptive service with random completion times, rather than discrete-time MDP dynamics. While we adopt a similar Lagrange relaxation principle, we do not establish asymptotic optimality for the resulting policy under SMDP and instead use it as a computationally tractable heuristic for the heterogeneous AoI problem.

None of the above works considers RMAB formulation and index-based policies for weakly coupled SMDPs. In the context of AoI minimization for heterogeneous sources and unreliable channels, existing work is limited to state-independent stationary policies that are understandably sub-optimal.

B. Our Contributions

Following are our main contributions.

1) We model the problem of scheduling of updates from heterogeneous sources to minimize the long run weighted average age of updates at the receiver (Section II). We pose it as a average cost SMDP problem (Section II-A).
2) We propose an approach to frame general weakly coupled SMDPs as RMABs. Using this formulation, we develop a Lagrange index based heuristic for weakly coupled SMDPs (Section III).
3) We apply the above heuristic to the update scheduling problem. In particular, we argue that the policies suggested for the decoupled single source problems are *one step lookahead policies* [15, Section 4.4] and provide an iterative algorithm to obtain these. We use solutions to the decoupled problems to obtain Lagrange indices, and subsequently, a heuristic for the original problem (Section V). Our numerical results show that the proposed heuristic noticeably outperforms the known solutions (Section VI).

II. System Model

We consider a discrete-time sensing and communication system with N sources (e.g., sensors) indexed by $1, \cdots, N$. Each source generates updates that have to be communicated to a common receiver through an unreliable wireless channel. We consider scheduling of update transmissions so as to minimize the average age of updates at the receiver. We now formally describe the update generation and communication processes, age evolution at the receiver, and the optimal scheduling problem.

Update generation: Source i's updates are L_i packet long. A source can take multiple slots to transmit an update. When a source is scheduled to transmit, it generates a new update unless it has an unfinished update transmission.

Update transmission: The communication channel allows only one source to transmit at a time. When scheduled, a source transmits one packet per slot. Owing to channel unreliability, each packet transmission succeeds with probability $p \in (0, 1]$ and fails with probability $1 - p$. Clearly, source i requires at least L_i slots to transmit an update. We consider non-preemptive update transmissions; a scheduled source keeps the channel until it successfully transmits all the packet pertaining to its update.

Let $\tau_k^{(i)}$ denote the time when kth successful update delivery of source i is accomplished. Moreover, let $X_k^{(i)}$ denote the number of slots needed for delivery of this update. From the above discussion, the generation time of this packet is $\tau_k^{(i)} - X_k^{(i)}$. Note that because of the packet generation model (generate at will) assumed, if $L_i = 1$ then $X_k^{(i)} = 1$ for all k and if $L_i \geq 2$ then $X_k^{(i)}, k \geq 1$ are i.i.d. random variables. In particular,

$$\mathbb{P}(X_k^{(i)} = l | \text{first packet is successfully sent}) = p_l^{(i)}$$

where, if $L_i = 1$ then $p_1^{(i)} = 1$, and if $L_i \geq 2$ then

$$p_l^{(i)} = \begin{cases} \binom{l-2}{L_i-2} p^{L_i-1}(1-p)^{l-L_i} & \text{if } l \geq L_i, \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

The above equation follows from the observation that after successful transmission of the first packet, the remaining $L_i - 1$ successful packet transmissions require a negative binomial number of additional slots. It follows that $\mathbb{E}[X_k^{(i)}] = \frac{L_i-(1-p)}{p}$.

Age of updates: For any source, the age of updates, also referred to as the AoI, at the receiver is defined as the time elapsed since generation of the last received update. Let $\bar{v}_i(t)$ denote the age of updates of source i at the receiver at time t. Clearly,

$$\bar{v}_i(t) = \begin{cases} X_k^{(i)} & \text{if } t = \tau_k^{(i)}, \\ \bar{v}_i(t-1) + 1 & \text{otherwise.} \end{cases}$$

The expected long term weighted average cost is given as

$$\lim_{T\to\infty} \frac{1}{T} \mathbb{E}\left[\sum_{t=1}^{T} \alpha_i \bar{v}_i(t)\right].$$

Optimal Scheduling problem: Update transmissions of various sources must be scheduled so as to minimize the above cost. Notice that a source must be selected for update transmission after either of the following events.

1) An update is successfully delivered.
2) Transmission of the first packet of an update fails.

Clearly, the intervals between successive decision instants are random variables. We frame the above scheduling problem as a SMDP as described below.

A. SMDP Formulation

Let t_k, $k \geq 1$ denote the successive decision instants. Let $v_i(k)$ denote the age of updates of source i at the receiver at t_k; $v(k) := (v_1(k), \cdots, v_N(k)) \in \mathbb{Z}_+^N$ represents the state at t_k. Further, let $a(k) := (a_1(k), \cdots, a_N(k)) \in \{0,1\}^N$ denote the scheduling decision at t_k; $a_i(k) = 1$ if source i is chosen for update transmission at t_k and $a_i(k) = 0$ otherwise. The scheduling constraint prescribes that exactly one source be chosen at each t_k, i.e., $\sum_{i=1}^N a_i(k) = 1 \quad \forall k \geq 1$. Let $\Delta(k) := t_{k+1} - t_k, k \geq 1$ denote the gaps between successive decision instants. Given action $a(k)$ with $a_i(k) = 1$, $\Delta(k) = 1$ if the first packet's transmission at t_k fails and $\Delta(k) \geq L_i$ otherwise. More specifically, $\Delta(k)$ is distributed as follows.

$$\mathbb{P}(\Delta(k) = l | a_i(k) = 1) = p p_l^{(i)} + (1-p)\mathbb{1}(l = 1). \quad (2)$$

It can be easily checked that $\mathbb{E}[\Delta(k)|a_i(k) = 1] = L_i$. More generally, the expected value of the interval $\Delta(k)$, denoted as $S(a(k))$, is given as follows.

$$S(a(k)) \triangleq \mathbb{E}[\Delta(k)] = \sum_{j=1}^{N} a_j(k) L_j.$$

Further, given $v(k)$ and $a(k)$ with $a_i(k) = 1$, $v_i(k+1) = v_i(k)+1$ if the first packet's transmission at t_k fails and $v_i(k+1) = \Delta(k)$ otherwise. In either case, $v_j(k+1) = v_j(k) + \Delta(k)$ for all $j \neq i$. See Figure 2 for an illustration. We thus see that $v_j(k+1) = v_j(k) + 1$ for all j with probability $1-p$, and $v_i(k+1) = l$ and $v_j(k+1) = v_j(k) + l$ for all $j \neq i$ with probability $p p_l^{(i)}$. Clearly, $\bar{v}_i(t), t \geq 1$ is a SMDP. We refer to t_k as the kth stage of this SMDP.

Now, we describe the single stage cost associated with the above SMDP. The weighted update-age cost incurred over $\{t_k, \cdots, t_{k+1} - 1\}$, also referred to as the kth stage cost, is a function of the state $v(k)$, action $a(k)$ and the interval length $\Delta(k)$. Note that the ages of updates for all the sources increment by one at successive slots between t_k and t_{k+1}. Hence the kth stage cost associated with source i equals

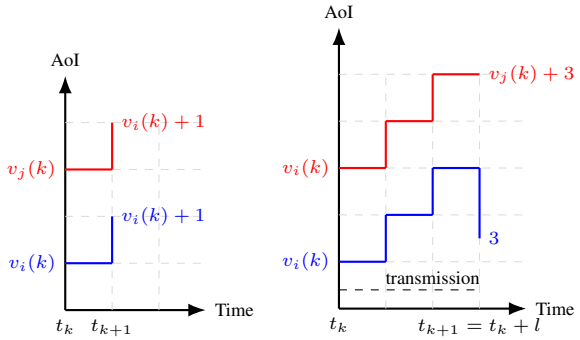
$$\alpha_i \sum_{s=0}^{\tau(k)-1} (v_i(k) + s) = \alpha_i \left(\Delta(k) v_i(k) + \frac{\Delta(k)(\Delta(k)-1)}{2}\right).$$

Consequently, the expected kth stage cost associated with source i, denoted as $g_i(v_i(k), a(k))$, is given by

$$g_i(v_i(k), a(k)) = \alpha_i \sum_{j=1}^{N} a_j(k)(v_i(k) L_j + w(L_j)). \quad (3)$$

where

$$w(L_j) := \mathbb{E}\left[\frac{\Delta(k)(\Delta(k)-1)}{2} l \middle| a_j(k) = 1\right] = \frac{L_j(L_j-1)}{2p}. \quad (4)$$



(a) $\Delta(k) = 1$ (failure) (b) $\Delta(k) = 3$ (success)

Fig. 2: State transition from t_k to t_{k+1} given that source i is scheduled at t_k. We assume $L_i = 2$.

We can formulate the scheduling problem as a long-term average cost control problem. A stationary admissible policy is a mapping from the state space $\mathbb{Z}_+^N$ to the set of N-dimensional standard unit vectors. So, given the initial state, any policy π induces a set of actions $a(k) \in \{0,1\}^N, k \geq 1$ such that

$$\sum_{j=1}^{N} a_j(k) = 1 \quad \forall k \geq 1. \quad (5)$$

Using Markov renewal reward theorem [16, Theorem 7.5], the long-term expected average-cost associated with policy π is given by

$$\lim_{K\to\infty} \frac{\mathbb{E}_\pi\left[\sum_{k=0}^{K}\sum_{i=1}^{N} g_i(v_i(k), a(k))\right]}{\mathbb{E}_\pi\left[\sum_{k=0}^{K}\sum_{i=1}^{N} a_i(k) L_i\right]} \quad (6)$$

The optimal control problem aims to minimize this cost over all the admissible policies. Evidently, this problem suffers from curse of dimensionality and is non-viable even for moderate number of sources.

We address this issue via posing the scheduling problem as a RMAB problem with the sources being treated as the arms, and proposing a Lagrange index based heuristic. We

can see that the update age evolution of different sources are weakly coupled through the actions at the decision instants as in RMAB problems. However, unlike a classical RMAB setting where states constitute a controlled Markov chain, state evolution in our case is a controlled semi-Markov chain. Here, the decoupled problems associated with different arms are also SMDPs. To the best of our knowledge, RMAB formulation of SMDPs has not been considered so far. In the next section, we discuss how the RMAB framework can be extended for general weakly-coupled SMDPs and also propose a Lagrange index based heuristic.

III. RMAB Formulation of SMDPs

In this section, we consider general weakly-coupled SMDPs, treat them as RMABs and develop a Lagrange index based heuristic. We retain most of the notation in Section II-A. In particular, we consider N arms with $v_i(k) \in \mathcal{V}_i$ being the state of arm i and $a_i(k) \in \{0,1\}$ being the action associated with arm i at the kth stage. As before, the actions are coupled as $\sum_{i=1}^{N} a_i(k) = 1$ for all $k \geq 1$. Here, $v(k)$ and $a(k)$ represent state and action vectors at the kth stage. Given $a(k)$, the states of different arms, i.e., $v_i(k), k \geq 1$ evolve independently. In particular, for all $k \geq 1$, given $v_i(k)$ and $a(k)$, $v_i(k+1)$ is independent of $v_j(k), j \neq i$ and previous state and action vectors. For any i, L_i represents the sojourn time of the joint SMDP in the kth stage given $a_i(k) = 1$. Finally, $g_i(v_i, a)$ represents the single stage cost associated with arm i given that it is in state i and action a is taken. The total single state cost for state-action pair (v, a) is given by $\sum_{i=1}^{N} g_i(v_i, a)$. We do not specify the state transition probabilities as these are not needed for the following discussion.

An admissible policy $\pi : \prod_{i=1}^{N} \mathcal{V}_i \rightarrow \{0,1\}^N$ induces a sequence of actions $a(k), k \geq 1$ that satisfy (5). Let Π be the set of all admissible policies. We aim to minimize the long-term average expected cost, given by (6), over all $\pi \in \Pi$. Constraint (5) which couples the individual SMDPs necessitates that these be solved together rendering an optimal solution intractable. Below, we propose a novel relaxation of (5) that facilitates decoupling of individual SMDPs.

A. A Relaxed Problem

We adopt the standard relaxation used in classical RMABs [12]–[14]. Let $N(T)$ denote the number of decision stages up to time T. Constraint (5) which ensures that exactly one arm be played at each stage also implies that the long term expected fractions of slots during which different arms are played add up to one. More formally, (5) implies that

$$\lim_{T\to\infty} \frac{1}{T}\mathbb{E}_\pi \left[\sum_{k=0}^{N(T)} \sum_{i=1}^{N} a_i(k) S(a(k)) \right] = 1. \tag{7}$$

Note that the expected sojourn times $\mathbb{E}_\pi \left[\sum_{i=1}^{N} a_i(k) L_i\right]$ are bounded by $\max_i L_i$. Hence, by Markov renewal reward theorem [15], [16], (7) is equivalent to

$$\lim_{K\to\infty} \frac{\mathbb{E}_\pi \left[\sum_{k=0}^{K} \sum_{i=1}^{N} a_i(k) S(a(k))\right]}{\mathbb{E}_\pi \left[\sum_{k=0}^{K} S(a(k))\right]} = 1. \tag{8}$$

The relaxed problem therefore becomes

$$\min_\pi \lim_{K\to\infty} \frac{\mathbb{E}_\pi \left[\sum_{k=0}^{K} \sum_{i=1}^{N} g_i(v_i(k), a(k))\right]}{\mathbb{E}_\pi \left[\sum_{k=0}^{K} S(a(k))\right]} \tag{9}$$

over all π for which the the induced actions $a(k), k \geq 0$ satisfy (8). Constraint (8) allows multiple arms to be played in a slot. This relaxation enables a Lagrangian formulation that decomposes into N independent single-arm control problems.

The Dual Problem: Introducing a Lagrange multiplier $\lambda \in \mathbb{R}$ associated with constraint (8), the Langrangian can be written as

$$\lim_{K\to\infty} \frac{\mathbb{E}_\pi \left[\sum_{k=0}^{K} \sum_{i=1}^{N} (g_i(v_i(k), a(k)) + \lambda a_i(k) S(a(k)))\right]}{\mathbb{E}_\pi \left[\sum_{k=0}^{K} S(a(k))\right]} - \lambda$$

So, the dual problem is

$$\max_{\lambda\in\mathbb{R}} \left(\min_\pi \sum_{i=1}^{N} \tilde{\mathcal{L}}_i(\pi, \lambda) - \lambda \right). \tag{10}$$

where

$$\tilde{\mathcal{L}}_i(\pi, \lambda) \coloneqq \lim_{K\to\infty} \frac{\mathbb{E}_\pi \left[\sum_{k=0}^{K} \left(g_i(v_i(k), a(k)) + \lambda\, a_i(k)\, S(a(k)) \right) \right]}{\mathbb{E}_\pi \left[\sum_{k=0}^{K} S(a(k))\right]}. \tag{11}$$

We note that (10) does not decouple. However, a lower bound to (10) can be obtained as follows

$$\max_{\lambda\in\mathbb{R}} \left(\sum_{i=1}^{N} \min_\pi \tilde{\mathcal{L}}_i(\pi, \lambda) - \lambda \right). \tag{12}$$

Now we propose a method to solve (12).

1) Inner Minimization Problem: For a fixed λ, the inner minimization problem is equivalent to

$$\min_\pi \tilde{\mathcal{L}}_i(\pi, \lambda)$$

which decouples into N SMDPs associated with the N arms. A policy π for the ith arm's problem is a mapping from $\mathcal{V}_i$ to $\{0,1\}^N$. The collection of these single-flow solutions provides the minimizing policy for the current λ. We now focus on the corresponding single-arm problem. In particular, for a fixed multiplier λ, we study the optimal control of an individual arm i under the modified expected per-stage cost $g_i(v_i(k), a(k)) + \lambda a_i(k) \tilde{S}(a(k))$.

The single-arm average-cost optimality equation can be written as

$$h_i(v_i) = \min \left\{ Q_{i,i}(v_i),\ \min_{j\neq i} Q_{i,j}(v_i) \right\}, \tag{13}$$

where

$$Q_{i,k}(v_i) = \begin{cases} g_i(v_i, \mathbf{e}_i) + (\lambda - \theta(\lambda)) S(\mathbf{e}_i) + \mathbb{E}[h_i(V_i') \mid v_i, \mathbf{e}_i], k = i, \\ g_i(v_i, \mathbf{e}_k) - \theta(\lambda) S(\mathbf{e}_k) + \mathbb{E}[h_i(V_i') \mid v_i, \mathbf{e}_k], k \neq i. \end{cases}$$

Here, $\mathbf{e}_k$ denotes the unit scheduling vector that selects source k. In (13), $h_i(v)$ is the bias (relative value) function and $\theta(\lambda)$ is the optimal average cost for the single-flow problem. The Bellman equation contains N actions corresponding to activating each arm. Solving (13) for a fixed λ yields the

optimal policy for source i under the Lagrangian relaxation. The assumptions on $g_i(v, \mathbf{e}_i)$ such that the solution to (13) exists is as per [17], [18].

2) Dual Update: Note that the dual problem is a concave maximization problem. The resulting dual function is then maximized over λ using standard scalar ascent methods (e.g., subgradient ascent), yielding an optimal multiplier $\lambda^\star$, which is then used to update the dual variable. Iterating this procedure produces $\lambda^\star$ and a corresponding policy. To update the dual variable, we compute the average fraction of time the i-th arm is activated under this optimal policy. This is obtained by solving an auxiliary average-cost Bellman equation where the per-stage cost equals 1 when arm i is activated and 0 otherwise. Let $\mu_i(\lambda)$ denote the resulting long-run activation fraction of arm i under the optimal policy for multiplier λ. The dual variable is then updated using bisection method.

B. Lagrange Indices

Upon convergence to $\lambda^\star$, we compute, for each arm i, the Lagrange index

$$\gamma_i(v_i) = Q_{i,i}(v_i) - \min_{j \neq i} Q_{i,j}(v_i). \tag{14}$$

Heuristic for the Original Problem: The resulting Lagrange index policy selects the arm

$$m^* = \operatorname*{argmin}_i \gamma_i(v_i). \tag{15}$$

This constitutes the Lagrange index policy for restless multi-armed bandits for SMDP. In Section IV we discuss about the AoI minimization problem stated in (6).

IV. Single source AoI minimization

We now return to the weighted AoI minimization problem introduced in Section II. Under Lagrangian relaxation, the original multi-source problem decouples into independent single-source subproblems as per Section III. We focus on one such subproblem corresponding to a fixed source index i.

A. Average-Cost Bellman Equation (SMDP)

Fix a multiplier $\lambda \in \mathbb{R}$. For the weighted AoI cost, the expected cumulative cost incurred over a decision stage when scheduling source j admits the closed-form expressions given in (4)–(3). The corresponding average-cost optimality equation is given by [17]:

$$h_i(v_i) = (1-p)\, h_i(v_i+1) + \min\Big\{Q_{i,i}(v_i),\ \min_{j \neq i} Q_{i,j}(v_i)\Big\}. \tag{16}$$

The term $(1-p)h_i(v_i+1)$ represents the continuation value under transmission failure, which occurs with probability $1-p$ regardless of the chosen action. In that case, the AoI deterministically increases by one. Since this transition is action-independent, the corresponding term appears outside the minimization. The action-dependent terms are

$$Q_{i,j}(v_i) \triangleq \begin{cases} g_i(v_i, \mathbf{e}_i) + (\lambda - \theta(\lambda))L_i + \sum_{l \geq L_i} pp_l^{(i)} h_i(l), & j = i, \\ g_i(v_i, \mathbf{e}_j) - \theta(\lambda)L_j + \sum_{l \geq L_j} pp_l^{(j)} h_i(v_i + l), & j \neq i. \end{cases}$$

Here, $\mathbf{e}_k$ denotes the unit scheduling vector that selects source k. The scalar $\theta(\lambda)$ denotes the optimal average cost of the single-source problem for a fixed multiplier λ. The multiplier λ can be interpreted as a penalty you pay to associated with scheduling source i.

We establish a few structural properties of the $h_i(.)$ function. In Lemma 1, we show the monotonicity property of the Bellman equation. In Lemma 2 we reduce the action space from N to 2 actions. Theorem 1 shows that the 2 action Bellman equation has a threshold structure using *one step lookahead policies*.

Lemma 1 (Monotonicity of the value function)**.** *The one-step cost $g_i(v, \mathbf{a})$ is nondecreasing in v for every feasible action $\mathbf{a}$. Consequently, any solution $h_i(\cdot)$ of* (16) *is nondecreasing in v, i.e.,*

$$v_1 \geq v_2 \quad \Longrightarrow \quad h_i(v_1) \geq h_i(v_2).$$

Proof. See Appendix A □

We now show that the inner minimization over $j \neq i$ in (16) reduces to a single dominant competing action.

Lemma 2 (Dominant competing action)**.** *Fix a source i and multiplier λ, and define*

$$m(i) = \arg\min_{j \neq i} L_j.$$

Then, for all $v_i \in \mathbb{Z}_+$ and all $j \neq i$,

$$Q_{i,m(i)}(v_i) \leq Q_{i,j}(v_i).$$

That is, among all competing sources, it is optimal to select the one with the smallest update length.

Proof. See [19, Appendix B] □

Lemma 2 formalizes the structural property that, whenever source i is not scheduled, it is optimal to schedule the "fastest" competing source, i.e., one with smallest update length L_j. Intuitively, this minimizes the time during which the age of flow i continues to grow.

As a consequence of Lemma 2, the inner minimization over $j \neq i$ in (16) is achieved by the single index $m(i)$. Accordingly, the Bellman equation simplifies to a reduced two-action form.

$$h_i(v_i) = (1-p)\, h_i(v_i+1) + \min\Big\{Q_{i,i}(v_i),\ Q_{i,m(i)}(v_i)\Big\} \tag{17}$$

where $Q_{i,i}$ and $Q_{i,m(i)}$ are given in (16). We show that the Bellman equation (17) admits a threshold structure: there exists a threshold $T_i(\lambda)$ such that it is optimal to schedule source i when $v_i \geq T_i(\lambda)$ and the competing source $m(i)$ otherwise.

B. Computing the threshold and average cost

In this section we suggest a numerical method to compute the optimal threshold of the single source problem.

Theorem 1 (Threshold structure and characterization)**.** *Fix a source index i and multiplier $\lambda \in \mathbb{R}$. Consider the corresponding two-action SMDP Bellman equation* (17)*, where the competing action is denoted by $m \neq i$. Then the optimal policy is of threshold type: there exists a threshold $T_i(\lambda) \in \mathbb{Z}_{\geq 0}$ such that*

$$\pi_i^\star(v_i) = \begin{cases} \mathbf{e}_m, & L_i \leq v_i < T_i(\lambda), \\ \mathbf{e}_i, & v_i \geq T_i(\lambda). \end{cases} \tag{18}$$

$T_i(\lambda)$ *is given by*

$$T_i(\lambda) = \left\lceil \frac{\theta(\lambda)}{\alpha_i} - \frac{L_m - 1}{2p} - \frac{L_i}{p} \right\rceil. \tag{19}$$

Proof. See Appendix B □

Theorem 1 shows that the threshold $T_i(\lambda)$ is determined by the (unknown) average cost $\theta(\lambda)$. For a fixed source index i, we now derive a second relation between $\theta(\lambda)$ and $T_i(\lambda)$ that enables efficient computation of both quantities for a fixed multiplier λ.

From Theorem 1 and (17), we can write for all $v_i \geq L_i$ as

$$h_i(v_i) = f_{i,1}(v_i) - \theta(\lambda) f_{i,2}(v_i), \tag{20}$$

where $f_{i,1}$ and $f_{i,2}$ are functions independent of $\theta(\lambda)$. In particular, for the region $v_i \geq T_i(\lambda)$, we have

$$f_{i,1}(v_i) = \frac{\alpha_i L_i}{p}\, v_i, \qquad f_{i,2}(v_i) = \frac{L_i}{p}. \tag{21}$$

Next, consider the region $L_i \leq v_i < T_i(\lambda)$, where the threshold policy schedules the competing source $m \neq i$. Define $c(v_i) \coloneqq T_i(\lambda) - v_i$, so $c(v_i) \geq 1$ in this region. The Bellman equation (17) reduces to

$$h_i(v_i) = (1-p)\, h_i(v_i+1) + (\alpha_i v_i - \theta(\lambda)) L_m + \alpha_i w(L_m) + \sum_{l \geq L_m} p p_l^{(m)}\, h_i(v_i + l), \tag{22}$$

Since $T_i(\lambda)$ is the threshold, the term $h_i(v_i+l)$ must be treated differently depending on whether $v_i + l < T_i(\lambda)$ or $v_i + l \geq T_i(\lambda)$. For $l \geq c(v)$ we have $v_i + l \geq T_i(\lambda)$, so $h_i(v_i + l)$ is affine. Substituting into the Bellman equation and collecting the coefficients of $\theta(\lambda)$ gives the recursions for $f_{i,1}(\cdot)$ and $f_{i,2}(\cdot)$ in the region $L_i \leq v_i < T_i(\lambda)$.

$$f_{i,1}(v_i) = \alpha_i v_i L_m + \alpha_i w(L_m) + (1-p)\, f_{i,1}(v_i+1) + \sum_{L_m \leq l \leq c(v)} p_l^{(m)}\, f_{i,1}(v_i + l) + \sum_{l > c(v)} p_l^{(m)}\, \frac{\alpha_i L_i}{p}(v_i + l), \tag{23}$$

and

$$f_{i,2}(v_i) = L_m + (1-p)\, f_{i,2}(v_i+1) + \sum_{L_k \leq l \leq c(v)} p_l^{(m)}\, f_{i,2}(v_i + l) + \sum_{l > c(v)} p_l^{(k)}\, \frac{L_i}{p}. \tag{24}$$

From equation (21) and the recursions (23)–(24) we uniquely determine $f_{i,1}(v_i)$ and $f_{i,2}(v_i)$ for all $v_i \geq L_i$. Finally, we obtain a closed-form expression for $\theta(\lambda)$ using (20);

$$\theta(\lambda) = \frac{p\Big(\lambda L_i + p \sum_{l \geq L_i} p_l^{(i)} f_{i,1}(l) + \alpha_i w(L_i)\Big) + \alpha_i L_i (1-p)}{p^2 \sum_{l \geq L_i} p_l^{(i)} f_{i,2}(l)}. \tag{25}$$

The equations (19) and (25) define an implicit relationship between the threshold $T_i(\lambda)$ and the unknown average cost $\theta(\lambda)$. We compute $(T_i(\lambda), \theta_\lambda)$ via a fixed-point iteration as shown in Algorithm 1.

Lemma 3 (Monotonicity in λ)**.** *For each source i, the threshold $T_i(\lambda)$ and the corresponding average cost $\theta(\lambda)$ are nondecreasing functions of the multiplier λ. That is, for any $\lambda_1 \leq \lambda_2$,*

$$T_i(\lambda_1) \leq T_i(\lambda_2) \quad \textit{and} \quad \theta(\lambda_1) \leq \theta(\lambda_2).$$

Proof. See [19, Appendix D] □

Algorithm 1: Fixed-point iteration to compute $T_i(\lambda)$

Input: $i, \lambda, m \neq i$, (p, α_i, L_i, L_m), $\beta \in (0,1)$, $\theta(\lambda)^{(0)}$, $\varepsilon > 0$
Output: $T_i(\lambda)$ and $\theta(\lambda)$
$n \leftarrow 0$;
repeat
  $T_i^{(n)} \leftarrow \left\lceil \frac{\theta(\lambda)^{(n)}}{\alpha_i} - \frac{L_m - 1}{2p} - \frac{L_i}{p} \right\rceil$;
  Compute $f_{i,1}(\cdot), f_{i,2}(\cdot)$ using (23)–(24);
  $\bar{\theta}(\lambda)^{(n)} \leftarrow \theta(\lambda)\Big(T_i^{(n)}\Big)$ using (25);
  $\theta(\lambda)^{(n+1)} \leftarrow \beta\, \theta(\lambda)^{(n)} + (1-\beta)\, \bar{\theta}(\lambda)^{(n)}$;
  $n \leftarrow n + 1$;
until $|\theta\lambda^{(n+1)} - \theta(\lambda)^{(n)}| \leq \varepsilon$;
$T_i(\lambda) \leftarrow T_i^{(n-1)}$;
$\theta(\lambda) \leftarrow \theta(\lambda)^{(n)}$;

V. Lagrange Index policy

A. Dual update via average activation fractions

For a fixed multiplier λ, the dual problem (12) decomposes into N independent single-source SMDPs, each solved as in Section IV. To solve the outer maximization over λ in (12), we update λ using bisection method. The derivative of the Lagrangian with respect to λ is given by $(\sum_{i=1}^{N} \mu_i(\lambda) - 1)$, where $\mu_i(\lambda)$ denotes the long-run fraction of time (in slots) during which source i is scheduled under the single-flow optimal policy for multiplier λ.

To compute $\mu_i(\lambda)$, we consider the policy π_i^* in (18) and evaluate the long run fraction of time for which the action i is chosen. This is done as follows; The one-step cost function is modified to

$$\hat{g}_i(v_i, \pi_i^*(v)) = \begin{cases} 1, & v_i \geq T_i(\lambda) \\ 0, & L_i \leq v_i < T_i(\lambda) \end{cases} \tag{26}$$

and we consider the policy evaluation equation

$$A_i^{\pi_i^*}(v_i) = (\hat{g}_i(v_i, \pi_i^*(\lambda)) - \mu_i(v))\, L_i + \mathbb{E}[A_i^{\pi_i^*}(v_i')] \tag{27}$$

where $A_i^{\pi_i^*}(\cdot)$ is the bias function and v_i' is the next state to which the process transitions.

From (27), we can write for all $v \geq L_i$,

$$A_i^{\pi_i^*}(v_i) = A_{i,1}(v_i) - \mu_i(\lambda)\, A_{i,2}(v_i), \tag{28}$$

where $A_{i,1}$ and $A_{i,2}$ are independent of $\mu_i(\lambda)$. Substituting (28) into (27) we get linear recursions for $A_{i,1}$ and $A_{i,2}$ over the region $L_i \leq v_i < T_i(\lambda)$, with boundary conditions for $v_i \geq T_i(\lambda)$ given by, $A_{i,1}(v_i) = A_{i,2}(v_i) = \frac{L_i}{p}$. Finally, we get the activation fraction in closed form

$$\mu_i(\lambda) = \frac{\sum_{l \geq L_i} p_l^{(i)} A_{i,1}(l)}{\sum_{l \geq L_i} p_l^{(i)} A_{i,2}(l)}. \tag{29}$$

Having computed $\mu_i(\lambda)$ policies for a fixed multiplier λ, we update λ to solve the outer maximization in (10) using bisection method as given in Algorithm 2.

Remark 1 (On existence of an exact multiplier)**.** *In general, there may not exist a multiplier $\lambda^\star$ such that*

$$\sum_{i=1}^{N} \mu_i(\lambda^\star) = 1.$$

This is due to the discrete nature of the actions, which makes the mapping $\lambda \mapsto \sum_{i=1}^{N} \mu_i(\lambda)$ piecewise constant and possibly discontinuous. Consequently, the relaxed constraint may not be met with equality for any single value of λ. Nevertheless, there exist multipliers $\lambda_-^\star$ and $\lambda_+^\star$ such that

$$\sum_{i=1}^{N} \mu_i(\lambda_-^\star) < 1 \qquad \textit{and} \qquad \sum_{i=1}^{N} \mu_i(\lambda_+^\star) > 1,$$

which ensures that the dual optimality condition is satisfied. The resulting multiplier is therefore dual optimal. However, since the feasible set of the original problem does not satisfy a strict interior condition, a nonzero dual gap may exist between the primal and dual problems.

The activation fractions satisfy a monotonicity property: as λ increases, each $\mu_i(\lambda)$ decreases, which makes bisection algorithm particularly effective.

B. Lagrange index policy

Having obtained the optimal multiplier $\lambda^\star$, we define the Lagrange index for each source i at state v_i as

$$\gamma_i(v_i) \triangleq Q_{i,i}(v_i) - Q_{i,m(i)}(v_i), \tag{30}$$

which measures the incremental cost incurred by serving source i over competing action. The resulting scheduling policy selects, at each decision instant, the source with the smallest Lagrange index given in (15). For classical restless bandit problems formulated as discrete-time MDPs, Lagrange index policies are known to be asymptotically optimal as N increases [13]. Such guarantees have not been established for semi-Markov decision processes (SMDPs). We adopt the Lagrange index policy as a heuristic and evaluate it numerically, where it consistently outperforms the NSRP policy in [11].

Algorithm 2: Compute Lagrange index

Input: $\{L_i\}_{i=1}^N$, p, λ_{low}, λ_{high}, ε
Output: $\lambda^\star$, $\{\gamma_i(\cdot)\}_{i=1}^N$
repeat
  $\lambda \leftarrow \dfrac{\lambda_{\text{low}} + \lambda_{\text{high}}}{2}$;
  Compute $T_1(\lambda), \ldots, T_N(\lambda)$ using Algorithm 1;
  Obtain $\mu_1(\lambda), \ldots, \mu_N(\lambda)$ using (29);
  if $\sum_{i=1}^N \mu_i(\lambda) - 1 < 0$ **then**
    $\lambda_{\text{high}} \leftarrow \lambda$;
  else
    $\lambda_{\text{low}} \leftarrow \lambda$;
until $|\lambda_{high} - \lambda_{low}| < \varepsilon$;
$\lambda^\star \leftarrow \dfrac{\lambda_{\text{low}} + \lambda_{\text{high}}}{2}$;
for $i = 1$ **to** N **do**
  $\gamma_i(\cdot) \leftarrow Q_{i,i}(\cdot) - Q_{i,j}(\cdot)$;
return $\lambda^\star$, $\{\gamma_i(\cdot)\}_{i=1}^N$;

VI. Numerical Results

We compare the proposed Lagrange index policy with NSRP [11], Greedy, Scaled Greedy, and the Whittle index policy. The Greedy policy selects the flow with the largest AoI, while Scaled Greedy selects $\arg\max_i \alpha_i v_i$. Under NSRP, a probability mass function over sources is obtained by solving the optimization problem in [11], and a source is selected randomly according to this distribution. For reliable channels ($p = 1$), we additionally compare with the Whittle index policy for heterogeneous sources derived in [9]. Figure 3 compares the NSRP, Greedy, Scaled Greedy, and proposed Lagrange index policies as the channel reliability varies. As p increases, all policies improve; however, the Lagrange index policy consistently achieves the lowest long-run average weighted AoI across the range of reliabilities.

Figures 4, 5, and 6 examine the impact of update-length heterogeneity, system scaling, and weight asymmetry, respectively. In each case, the left subplots (Figures 4a, 5a, and 6a) correspond to unreliable channels ($p < 1$) and compare NSRP, Lagrange index, Greedy, and Scaled Greedy policies, where the proposed policy consistently performs best. The right subplots (Figures 4b, 5b, and Figure 6b corresponds to the reliable channel case ($p = 1$) and includes the Whittle index policy. In this regime, the proposed Lagrange index policy consistently outperforms or closely matches the Whittle index policy across all considered scenarios, demonstrating its effectiveness even in settings where Whittle indices are available. Importantly, while the Whittle index policy is currently characterized only for the reliable channel case, the Lagrange index framework naturally extends to unreliable channels ($p < 1$), where no Whittle index characterization is yet available.

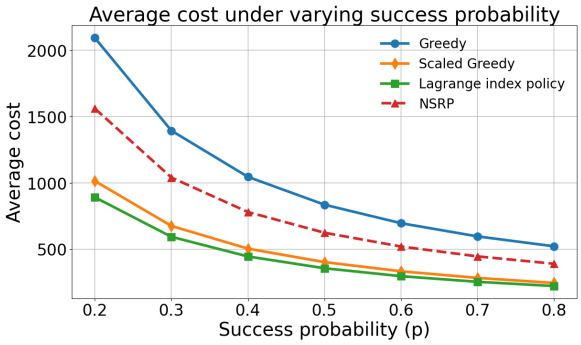
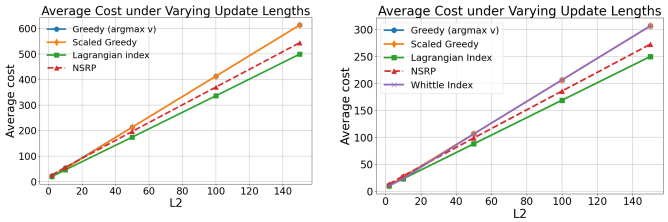


Fig. 3: Simulation results under varying channel reliability. The network has $N = 10$ sources split evenly into two classes: Class 1 with $(L_i, \alpha_i) = (2, 5)$ and Class 2 with $(L_i, \alpha_i) = (50, 1)$. The common channel reliability varies over $p \in \{0.2, 0.3, \ldots, 0.8\}$.

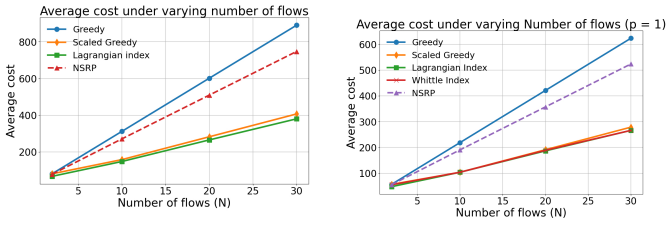
VII. Conclusion

In this paper, we developed a Lagrange index heuristic to minimize the weighted average AoI. The problem was formulated as a RMAB with SMDP dynamics. We proposed an heuristic based on the Lagrange index policy. We showed that the proposed policy outperforms both the NSRP in [11] and a scaled greedy policy, demonstrating that structural analysis translates into tangible performance gains. For the weighted AoI model, we established key structural properties, most notably threshold behavior and leveraged them to derive efficient algorithms to compute the Lagrange indices. Beyond the specific AoI setting, the proposed framework extends naturally to a broader class of RMAB problems with SMDP dynamics, thereby expanding the scope of index-based control beyond



(a) $p = 0.5$ (b) $p = 1$

Fig. 4: Average weighted AoI versus update-length heterogeneity. We consider $N = 2$ sources with $(L_1, \alpha_1) = (2, 5)$ and $(L_2, \alpha_2) = (L_2, 1)$, where $L_2 \in \{2, 10, 50, 100, 150\}$. The left panel corresponds to $p = 0.5$, where the greedy and scaled greedy policies achieve identical performance. The right panel corresponds to $p = 1$, where the greedy, scaled greedy, and Whittle index policies acheive identical performance.



(a) $p = 0.7$ (b) $p = 1$

Fig. 5: Average cost versus the number of sources. We consider two classes with an equal number of sources in each class. The total number of sources is $N \in \{2, 10, 20, 30\}$. Class–1 sources have $(L_1, \alpha_1) = (2, 5)$, and class–2 sources have $(L_2, \alpha_2) = (25, 1)$. The left panel corresponds to $p = 0.7$, and the right panel to $p = 1$. The Whittle index and Lagrange index policies achieve identical performance.

standard discrete-time formulations. Future work will consider systems with switching and alternative update-generation models, building on the structural foundations developed here.

REFERENCES

[1] M. Tubaishat and S. Madria, "Sensor networks: an overview," *IEEE Potentials*, vol. 22, no. 2, pp. 20–23, 2003.

[2] S. Kaul, M. Gruteser, V. Rai, and J. Kenney, "Minimizing age of information in vehicular networks," in *2011 8th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks*, pp. 350–358, 2011.

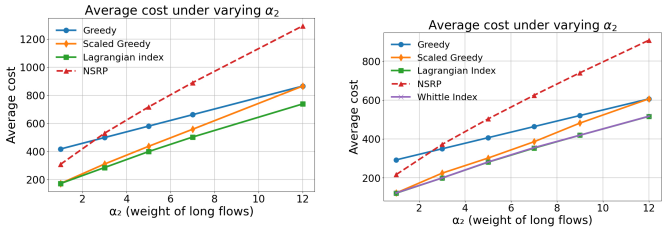
[3] S. Kaul, R. Yates, and M. Gruteser, "Real-time status: How often should one update?," in *2012 Proceedings IEEE INFOCOM*, pp. 2731–2735, 2012.

[4] A. Kosta, N. Pappas, and V. Angelakis, "Age of information: A new concept, metric, and tool," *Foundations and Trends in Networking*, vol. 12, pp. 162–259, 11 2017.

[5] V. Tripathi, I. Kadota, E. Tal, M. S. Rahman, A. Warren, S. Karaman, and E. Modiano, "Wiswarm: Age-of-information-based wireless networking for collaborative teams of uavs," in *IEEE INFOCOM 2023-IEEE Conference on Computer Communications*, pp. 1–10, IEEE, 2023.

[6] R. D. Yates, Y. Sun, D. R. Brown, S. K. Kaul, E. Modiano, and S. Ulukus, "Age of information: An introduction and survey," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 5, pp. 1183–1210, 2021.

[7] I. Kadota, A. Sinha, E. Uysal-Biyikoglu, R. Singh, and E. Modiano, "Scheduling policies for minimizing age of information in broadcast wireless networks," *IEEE/ACM Transactions on Networking*, vol. 26, no. 6, pp. 2637–2650, 2018.



(a) $p = 0.7$ (b) $p = 1$

Fig. 6: Average weighted AoI versus α_2. We consider $N = 10$ flows, with five class–1 sources having $(L_1, \alpha_1) = (2, 12)$ and five class–2 sources having $(L_2, \alpha_2) = (15, \alpha_2)$, where $\alpha_2 \in \{1, 3, 5, 7, 9\}$. The left panel corresponds to $p = 0.7$, and the right panel to $p = 1$. The Whittle index and Lagrange index policies achieve nearly identical performance.

[8] V. Tripathi and E. Modiano, "A whittle index approach to minimizing functions of age of information," *IEEE/ACM Transactions on Networking*, vol. 32, no. 6, pp. 5144–5158, 2024.

[9] V. Tripathi, L. Ballotta, L. Carlone, and E. Modiano, "Computation and communication co-design for real-time monitoring and control in multi-agent systems," in *2021 19th International Symposium on Modeling and Optimization in Mobile, Ad hoc, and Wireless Networks (WiOpt)*, pp. 1–8, 2021.

[10] B. Zhou and W. Saad, "Minimum age of information in the internet of things with non-uniform status packet sizes," *IEEE Transactions on Wireless Communications*, vol. 19, no. 3, pp. 1933–1947, 2020.

[11] Z. Zhao, V. Tripathi, and I. Kadota, "Optimizing age of information in networks with large and small updates," in *2025 23rd International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, pp. 1–8, 2025.

[12] I. M. Verloop, "Asymptotically optimal priority policies for indexable and nonindexable restless bandits," *The Annals of Applied Probability*, vol. 26, no. 4, pp. 1947–1995, 2016.

[13] 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*, vol. 49, p. 2468–2491, Nov. 2024.

[14] K. Avrachenkov, V. S. Borkar, and P. Shah, "Lagrangian index policy for restless bandits with average reward," *arXiv preprint arXiv:2412.12641*, 2024.

[15] D. P. Bertsekas, *Dynamic Programming and Optimal Control, Vol. II*. Athena Scientific, 3rd ed., 2007.

[16] S. M. Ross, *Applied Probability Models with Optimization Applications*. New York: Dover Publications, 1992.

[17] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. USA: John Wiley & Sons, Inc., 1st ed., 1994.

[18] L. I. Sennott, *Stochastic dynamic programming and the control of queueing systems*. John Wiley & Sons, 1998.

[19] A. Mukherjee, J. Kuri, and C. Singh, "Lagrange index based scheduling for minimizing age of updates from heterogeneous sources," *arXiv preprint arXiv:2604.18077*, 2026.

APPENDIX

A. *Proof of Lemma 1*

Proof. **Step 1: Convergence of RVI.**

Consider the single–flow SMDP under a fixed multiplier λ. Let $\pi^{(i)}$ denote the stationary policy that schedules flow i at every decision stage.

Under this policy, the AoI process $\{v(k)\}$ evolves as follows:

$$v \to \begin{cases} v+1, & \text{with probability } (1-p), \\ l \geq L_i, & \text{with probability } pp_l^{(i)}. \end{cases}$$

The state space is $\{L_i, L_i + 1, \dots\}$.

Since from any state $v \geq L_i$ there is positive probability of transitioning to any state $l \geq L_i$ with $pp_l^{(i)} > 0$, the induced Markov chain is irreducible on its state space.

We now establish positive recurrence using a Foster–Lyapunov drift argument. Consider the Lyapunov function $V(v) = v$. The conditional drift satisfies

$$\begin{aligned}\mathbb{E}[v(k+1) - v(k) \mid v(k) = v] &= (1-p)(1) + \sum_{l \geq L_i} pp_l^{(i)}(l - v) \\ &= (1-p) + L_i - (1-p) - vp.\end{aligned}$$

Since $\sum_{l \geq L_i} p_l^{(i)} = p$ and $\sum_{l \geq L_i} p_l^{(i)} l = L_i - (1-p)$, the drift simplifies to

$$\mathbb{E}[v(k+1) - v(k) \mid v(k) = v] = L_i - pv.$$

For sufficiently large v, the drift is strictly negative. Therefore, the Markov chain satisfies the Foster–Lyapunov condition and is positive recurrent. Hence the induced Markov Chain under the policy $\pi^{(i)}$ is unichain. Since the one–stage cost is nonnegative in v [18], the average cost under π is finite. Under this unichain condition and boundedness of the relative value differences, standard SMDP results imply that Relative Value Iteration (RVI) converges (up to an additive constant) to a solution h_i of (16) [18].

Step 2: Monotonicity of the iterates.

We prove by induction that

$$v_1 \leq v_2 \implies h^{(k)}(v_1) \leq h^{(k)}(v_2), \qquad \forall k.$$

Base case: $h^{(0)} \equiv 0$ is nondecreasing.

Induction step: Assume $h^{(k)}$ is nondecreasing. Let $v_1 \leq v_2$.

The stage cost $g_i(v, \mathbf{a})$ is nondecreasing in v. We verify monotonicity separately for the two types of actions.

Case 1: $j = i$.

$$Q_{i,i}^{h^{(k)}}(v) = g_i(v, \mathbf{e}_i) + (\lambda - \theta)L_i + \sum_{l \geq L_i} pp_l^{(i)}\, h^{(k)}(l).$$

The last two terms are independent of v. Since $g_i(v, \mathbf{e}_i)$ is nondecreasing in v, it follows that

$$v_1 \leq v_2 \implies Q_{i,i}^{h^{(k)}}(v_1) \leq Q_{i,i}^{h^{(k)}}(v_2).$$

Case 2: $j \neq i$.

$$Q_{i,j}^{h^{(k)}}(v) = g_i(v, \mathbf{e}_j) - \theta L_j + \sum_{l \geq L_j} pp_l^{(j)}\, h^{(k)}(v + l).$$

The term $-\theta L_j$ is constant in v. Since $g_i(v, \mathbf{e}_j)$ is nondecreasing in v and $v \mapsto v + l$ is increasing, the induction hypothesis implies

$$v_1 \leq v_2 \implies h^{(k)}(v_1 + l) \leq h^{(k)}(v_2 + l), \qquad \forall l.$$

Taking expectation preserves order, hence

$$Q_{i,j}^{h^{(k)}}(v_1) \leq Q_{i,j}^{h^{(k)}}(v_2).$$

Since $v \mapsto v'$ is increasing and $h^{(k)}$ is nondecreasing,

$$v_1 \leq v_2 \implies Q_{i,j}^{h^{(k)}}(v_1) \leq Q_{i,j}^{h^{(k)}}(v_2), \qquad \forall j.$$

Taking minima preserves order, hence

$$(\mathcal{T}h^{(k)})(v_1) \leq (\mathcal{T}h^{(k)})(v_2).$$

So,

$$h^{(k+1)}(v_1) \leq h^{(k+1)}(v_2).$$

Thus $h^{(k+1)}$ is nondecreasing. By induction, all iterates are nondecreasing. Since RVI converges pointwise (up to a constant), and monotonicity is preserved under limits, the limiting bias function h_i is nondecreasing. □

B. Proof of Theorem 1

Proof. Fix λ and flow index i. Let $m \neq i$ denote any competing action. We use a one–step look ahead argument and the optimality principle [15].

Step 1: Cost of scheduling flow i.

Suppose that at state v we schedule flow i, and thereafter continue scheduling flow i at every decision stage. The corresponding value function is given by

$$\begin{aligned}C_i(v) = (\alpha_i v + \lambda - \theta_\lambda)L_i + \alpha_i w(L_i) \\ + (1-p)C_i(v+1) + \sum_{l \geq L_i} pp_l^{(i)} C_i(l).\end{aligned}$$

Scheduling of only flow i admits the affine solution

$$C_i(x) = \frac{\alpha_i L_i}{p}\, x - \frac{\theta_\lambda L_i}{p} + A(\lambda), \qquad x \geq T_i(\lambda),$$

where $A(\lambda)$ is a constant.

Step 2: One–step look ahead to flow m.

Now consider a deviating policy that schedules flow m at state v for exactly one decision epoch and then switches permanently to scheduling flow i.

The corresponding value function is

$$\begin{aligned}C_m(v) = (\alpha_i v - \theta_\lambda)L_m + \alpha_i w(L_m) \\ + (1-p)C_i(v+1) + \sum_{l \geq L_m} pp_l^{(m)} C_i(v+l).\end{aligned}$$

Step 3: Comparison.

Scheduling flow i is optimal at state v if $C_i(v) - C_m(v) \leq 0$.

The common term $(1-p)C_i(v+1)$ cancels. Substituting the affine form of $C_i(\cdot)$ and simplifying yields a linear inequality in v of the form

$$v \geq \frac{\theta_\lambda}{\alpha_i} - \frac{L_m - 1}{2p} - \frac{L_i}{p}.$$

Define

$$T_i(\lambda) = \left\lceil \frac{\theta_\lambda}{\alpha_i} - \frac{L_m - 1}{2p} - \frac{L_i}{p} \right\rceil.$$

Then for all $v \geq T_i(\lambda)$,

$$C_i(v) \leq C_m(v).$$

Step 4: Invariance and optimality.

For $v \geq T_i(\lambda)$, if flow m is scheduled, the next state equals $v + 1$ with probability $(1 - p)$ or $v + l$ for some $l \geq L_m$. In all cases, the next state is at least v, and hence remains in the region $[T_i(\lambda), \infty)$.

Thus, once $v \geq T_i(\lambda)$, any one–step lookahead to action m cannot improve the cost, and the process remains in the region where scheduling i is better.

By the one–step deviation principle for average cost [15], no profitable deviation exists for $v \geq T_i(\lambda)$. we conclude that for all $v \geq T_i(\lambda)$, scheduling flow i is optimal, whereas for $v < T_i(\lambda)$ scheduling flow $m \neq i$ is optimal.

Hence the optimal policy is of threshold type.

□