Asymptotically Optimal Lagrangian Priority Policy for Deadline Scheduling with Processing Rate Limits

Liangliang Hao, Yunjian Xu, and Lang Tong

***Abstract*—We study the deadline scheduling problem for multiple deferrable jobs that arrive in a random manner and are to be processed before individual deadlines. The processing of the jobs is subject to a time-varying limit on the total processing rate at each stage. We formulate the scheduling problem as a restless multi-armed bandit (RMAB) problem. Relaxing the scheduling problem into multiple independent single-arm scheduling problems, we define the *Lagrangian priority value* as the greatest tax under which it is optimal to activate the arm. We propose a Lagrangian priority policy which processes jobs in the order of their Lagrangian priority values, and establish its asymptotic optimality as the system scales. Numerical results show that the proposed Lagrangian priority policy achieves 22%-49% higher average reward than the classical Whittle index policy (that does not take into account the processing rate limits).**

***Index Terms*—Electric vehicle charging, Deadline scheduling, Restless multi-armed bandit (RMAB), Dynamic programming, Index policy.**

I. Introduction

We study the deadline scheduling of multiple deferrable jobs by a service provider with multiple processors, in the presence of stochastic processing costs, random job arrivals, and time-varying processing rate limits. A newly arrived job requests a certain amount of service (*e.g.*, energy or data processing) to be fulfilled by a user-prescribed deadline. The amount of resource available for processing these deferrable demands is constrained by a time-varying limit. The objective is to minimize the long-term (expected) total cost, consisting of job processing costs and the non-completion penalty when a job is not completed by its deadline.

The deadline scheduling problem is motivated by applications such as the charging of a large number of electric vehicles (EV) in a charging service center where the number of EVs that can be charged simultaneously is limited by the capacity of the transformer and the available local renewable generations [2]–[5]. Another example is the scheduling of processors in cloud services where the submitted jobs have different priorities and deadlines and the number of processors (available for processing on-demand jobs) varies in time [6], [7]. In these applications, the varying constraints on the available processing rate pose significant theoretical and algorithmic challenges.

The deadline scheduling problem is a sequential decision process that can be formulated as a stochastic dynamic program (DP). However, the formulated DP is intractable due to the well-known *curse of dimensionality [8]*: the system state space grows exponentially with the number of jobs, making it intractable to solve the deadline scheduling problem as a general DP.

In this paper, we consider an important class of scheduling algorithms, which are based on some ranking mechanism using priority values that can be computed with polynomial complexity. The classic index policies derive priority values, referred to as indices, from the individual decision processes (rather than the joint decision process in DP), thus reducing the computation complexity to be linear with the number of processors. By prioritizing jobs based on their individual properties, index policies are in general suboptimal with a few exceptions. Most celebrated are the Gittins index policy for the multi-armed bandit problem [9] and the Whittle index policy for the restless multi-armed bandit (RMAB) problem [10].

A. Related works

There exists an extensive literature on the deadline scheduling for multiple processors (for a survey, see [11]). Some relevant applications include scheduling of EV charging in power system [4], task scheduling in multi-core processors, [12], [13], transmission of packets in communication systems [14], [15], and inventory allocation in retail revenue management [16]. Similar to the present work, index policies have been adopted to explore deadline scheduling problems through a RMAB formulation [5], [16]. The indexability of a RMAB with bi-dimensional state (perishable product lifetime and inventory level) is established in [16] for a setting with constant processing cost. The authors of [5] established the asymptotic optimality of the Whittle index policy in the light traffic regime. In the aforementioned literature, the processing rate limit is assumed to be constant or hold on average over the entire scheduling horizon, whereas this work incorporates time-varying hard constraints on the total processing rate in each period.

L. Hao and Y. Xu are with the Department of Mechanical and Automation Engineering, the Chinese University of Hong Kong, Hong Kong SAR. (*Corresponding author: Yunjian Xu*). L. Tong is with the School of Electrical and Computer Engineering, Cornell University, Ithaca, NY 14853, USA. Emails: {llhao, yjxu}@mae.cuhk.edu.hk, lt35@cornell.edu.

This research is supported in part by project 4720249 of the Shun Hing Institute of Advanced Engineering, the Chinese University of Hong Kong, and General Research Fund (GRF) project 14200720, the University Grants Committee (UGC) of Hong Kong. The work of Lang Tong is supported in part by the National Science Foundation under Award 1816397 and 1809830. A preliminary version of this work that contains a subset of the results was submitted for presentation at the 2020 American Control Conference (ACC) [1].

The literature on deadline scheduling with time-varying processing capacity is less extensive, and none leads to any form of optimality. A combined EDF (Earliest Deadline First) and LLF (Least Laxity First) policy is proposed in [17] for online deadline scheduling with constant rewards. In [18], based on the primal-dual paradigm, an approximation algorithm is proposed for scheduling of non-preemptive jobs under time-varying processing rate constraints. In this paper, we consider a different setting with preemptive jobs, stochastic job arrivals and processing costs.

We formulate the stochastic deadline scheduling problem as a RMAB, which is a weakly coupled dynamic program (WCDP) consisting of multiple subproblems that are independent of each other except for a set of linking constraints on the controls [19]–[22]. The *Lagrangian relaxation* technique has been widely adopted to relax the linking constraints and to decouple the WCDP into multiple independent subproblems, which yield the dual bound for the original WCDP [21], [23]–[26].

Bandit [27], [28] and restless bandit [29], [30] approaches have been widely adopted to compute efficient control policies for resource sharing in wireless communications and demand response. The authors of [31] established the asymptotic optimality of the Whittle index policy for Markov-modulated restless bandits. In the restless bandit model of aforementioned works, the total number of activated arms (in each period) is either constant or below a time-average limit, whereas our model incorporates a time-varying hard constraint on the total number of activated arms in each period.

In a more closely related work, the authors of [32] develop an asymptotically optimal policy for the general RMAB problem that builds on the notion of WCDP introduced in [21]. A major contribution of [32] is a Lagrange based index policy with tie-breaking rule that is shown to be asymptotically optimal in the regime originally considered by Whittle, even when the RMAB is not indexable. The work in [32] assumes a *constant rate constraint* and *i.i.d.* reward process among arms. In this work, we consider a different setting with deferrable jobs, time-varying processing rate constraints, and coupled processing costs.

B. Summary of results

The main contribution of this work is three-fold. First, we construct a RMAB model to explore a general *stochastic deadline scheduling* problem with random job arrivals and time-varying hard constraints on the total processing rate at each stage. At each stage, a binary decision is made on whether to process each job. At each stage, the processing of a job achieves a time-varying reward that is uniform among all jobs. At the deadline of each job, a non-completion penalty incurs if the job is not fully processed. The objective is to properly schedule the processing of each job (before its deadline) to maximize the expected difference between processing reward and non-completion penalty. We establish closed form expressions of the Lagrangian priority value, and present an efficient algorithm to compute the Lagrangian priority values (with linear complexity with respect to the number of arms).

Second, we propose a Lagrangian priority policy (that processes jobs in the order of their Lagrangian priority values) with a new randomized tie-breaking rule. As the Lagrangian priority values of multiple jobs (at different states) may be the same, an arbitrary tie-breaking destroys asymptotic optimality. The proposed tie-breaking rule marks a main difference between the Lagrangian priority policy proposed here and that in [32].[1] Even under the conditions that the processing rate constraint is constant and the reward processes are *i.i.d.*, the proposed Lagrangian priority policy does not reduce to that in [32].

Third, we establish the asymptotic optimality for the proposed Lagrangian priority policy (under deterministic, time-varying processing cost), as the number of processors, job arrival rate, and the processing rate limit simultaneously increase to infinity. In particular, we show that the proposed tie-breaking rule enables the Lagrangian priority policy to imitate the optimal randomized policy (for the Lagrangian relaxation). Consequently, the gap between the expected reward per arm of the proposed policy and its upper bound (achievable by the optimal randomized policy) converges to zero as the system scales large. Note that the asymptotic optimality applies to both the light and heavy traffic regimes. We provide a simple example demonstrating that the proposed tie-breaking rule is crucial for the established asymptotic optimality result, as the Lagrangian priority policy with uniform tie-breaking rule may lead to suboptimal decisions.

Numerical results show that the Lagrangian priority policy with the proposed tie-breaking rule achieves 22%-49% higher average reward than the Whittle index policy that does not take into account the time-varying processing rate limit constraints. Compared with uniform tie-breaking, simulation results demonstrate that the proposed tie-breaking rule improves the average reward by 6%-12%. The proposed Lagrangian priority policy (with randomized tie-breaking) is shown to improve the expected reward achieved by the index based policy proposed in [32] by 1.9%-5.9% when the number of processors is in the range of 50 to 200.

The rest of the paper is organized as follows. In Section II, we formulate the EV charging scheduling problem as a dynamic program. In Section III, we discuss the Lagrangian relaxation that enables the decomposition of the formulated dynamic program. In Section IV, we define the Lagrangian priority value, establish the close form of the Lagrangian priority value, and derive a recursive expression for the Lagrangian priority value under deterministic processing cost. We propose a new tie-breaking rule for the Lagrangian priority policy which is shown to be asymptotic optimal in Section V. In Section VI we present some numerical results to compare the proposed Lagrangian priority policy against state-of-the-art algorithms, e.g., the Whittle index policy. Finally, we make some brief concluding remarks in Section VII.

[1] For detailed discussion on the key differences between these two tie-breaking approaches, please refer to Remark 3.

II. Problem Formulation

In Section II-A, we introduce the problem settings and assumptions. In Section II-B, we formulate the stochastic deadline scheduling problem with time-varying processing rate constraints as a Markov decision process.

A. The deadline scheduling problem

We study the stochastic deadline scheduling of multiple electric vehicles with time-varying processing rate limits, random job arrivals and processing cost. In presenting the problem formulation of stochastic deadline scheduling, we use EV charging as a concrete example.

1) **Scheduling horizon**: We consider a discrete-time model, where stage is indexed by $t \in \{1, \dots, T\}$.
2) **Processors**: The system has I processors (e.g., EV chargers) labelled by $i \in \{1, \dots, I\}$. The processing rate of all processors is constant and normalized to 1.
3) **Action**: We let a_t^i denote the action of the i^{th} processor at stage t. It equals 1 if the job i is processed[2], and 0 if not.
4) **Time-varying processing rate limits**: At stage t, at most $m_t \in \mathbb{Z}^+$ processors can be simultaneously activated, i.e.,

$$\sum_{i=1}^{I} a_t^i \leq m_t, \; \forall\, t. \tag{1}$$

$\{m_t\}$ is an exogenous deterministic process. In the context of EV charging, the processing rate limit can reflect the (time-varying) power capacity available for electric vehicle charging in a power distribution system.

5) **Job arrival and departure**: At the beginning of period t, the probability that a new job (with a deadline d and energy demand e) arrives at processor i is $Q^i(d, e)$. We use $Q^i(t, 0)$ to denote the probability that no job arrives at processor i at the beginning of period t. Each processor can process only one job at each time, and an occupied processor ignores any newly arrived jobs.

Once a job arrives at an unoccupied processor i, its (integer) energy demand e_t^i and deadline d^i are revealed to the operator, with $0 \leq e_t^i \leq E^i$ and $d^i \leq T+1$. Job i leaves processor i at the beginning of stage d^i. The job arrivals to different processors are mutually independent. At stage t, job i's state is a two-dimensional vector (e_t^i, d^i). The job can be served during periods $\{t, t+1, \dots, d-1\}$. We assume that it is possible to finish a new job before its departure, *i.e.*, $t + e \leq d$.

6) **Processing reward**: A processor receives a reward of $1 - c_t$ if it processes the job, where $c_t \in \mathbb{R}$ is the processing cost at stage t. We assume that the evolution of c_t is an exogenous Markov process and is independent of the job arrivals as well as the operator's charging decisions. At stage t, the transition probability of the processing cost is given by $\mathbb{P}_t^c(c_{t+1}|c_t)$.

A non-completion penalty $g(e_{d^i}^i)$ incurs if a job i's demand is not fulfilled by its deadline, where g is a quadratic, increasing function, and $e_{d^i}^i$ denotes the remaining energy demand at job i's deadline.

B. Dynamic programming formulation

We are now ready to formulate the problem as a dynamic program (DP) by introducing its states, action sets, state transitions, and reward.

1) System state: As illustrated in Fig. 1, the state of job i (e.g., an EV) at stage t consists of its deadline d^i and its remaining demand e_t^i at stage t:

$$x_t^i \doteq (d^i, e_t^i) \in \mathcal{X}^i, \quad \forall\, i, \,\forall\, t, \tag{2}$$

where $\mathcal{X}^i = \{1, 2, \dots, T+1\} \times \{0, 1, \dots, E^i\}$. We let $(t, 0)$ denote the state of an unoccupied processor at stage t.

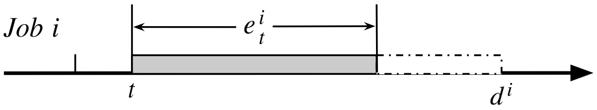


Fig. 1. State of job i at stage t: e_t^i is the remaining demand, and d^i is the deadline (departure time).

The system state consists of the states of all the I processors, and the current processing cost:

$$\boldsymbol{s}_t \doteq [x_t^1, \dots, x_t^I, c_t] \in \mathcal{S}, \quad \forall\, t, \tag{3}$$

where $\mathcal{S} = \mathcal{X}^1 \times \cdots \times \mathcal{X}^I \times \mathcal{X}^c$ is the system state space, $\mathcal{X}^c$ represents the state space of the processing cost.

2) Actions: Let

$$\boldsymbol{a}_t \doteq [a_t^1, \dots, a_t^I] \in \{0, 1\}^I, \quad \forall\, t, \tag{4}$$

denote the action vector at stage t. For the sake of notational convenience, a positive action is allowed for an unoccupied processor (without any job), causing no reward or penalty. The feasible action set under system state $\boldsymbol{s}_t$ is

$$\mathcal{A}_t(\boldsymbol{s}_t) \doteq \{\boldsymbol{a}_t : \text{Eqs. (1), (4)}\}, \quad \forall\, t. \tag{5}$$

A policy $\pi = \{\pi_1, \pi_2, \dots, \pi_T\}$ is a sequence of Markovian decision rules such that $\pi_t(\boldsymbol{s}_t) \in \mathcal{A}_t(\boldsymbol{s}_t)$, for all $\boldsymbol{s}_t \in \mathcal{S}$.

3) System dynamics: The state evolution of job i depends on its current state x_t^i and the action a_t^i:

$$x_{t+1}^i = \begin{cases} (d^i, (e_t^i - a_t^i)^+), & \text{if } d^i > t+1, \\ (d', e') \text{ with Prob. } Q^i(d', e'), & \text{if } d^i = t+1, \end{cases} \tag{6}$$

where $(x)^+ = \max(0, x)$. When job i with $d^i = t+1$ reaches its deadline, at the beginning of period $t+1$ job i is removed and a new job (d', e') arrives with probability $Q^i(d', e')$.

4) Reward: The immediate reward r_t^i of the i^{th} processor is determined by the current state x_t^i, action a_t^i, and processing cost c_t:

$$r_t^i(x_t^i, c_t, a_t^i) = \begin{cases} a_t^i(1 - c_t), & \text{if } d^i > t+1, \; e_t^i > 0, \\ a_t^i(1 - c_t) - g\left((e_t^i - a_t^i)^+\right), & \\ & \text{if } d^i = t+1, \; e_t^i > 0, \\ 0, & \text{otherwise}. \end{cases} \tag{7}$$

[2] With slight abuse of notation, we refer to the job (e.g., an EV) processed by the i^{th} processor (e.g., an EV charger) as job i.

For a job that has not reached its deadline and has remaining demand at stage $t+1$, a reward of $1-c_t$ is obtained if job i is processed. When $d^i = t+1$, the job will leave at the beginning of $t+1$, and a non-completion penalty $g\left((e_t^i - a_t^i)^+\right)$ incurs if job i is not fulfilled by its deadline d^i. For an unoccupied processor (with state $(t, 0)$), the reward it achieves at stage t is always 0. The immediate reward at stage t is given by

$$R_t(\boldsymbol{s}_t, \boldsymbol{a}_t) = \sum_{i=1}^{I} r_t^i(x_t^i, a_t^i, c_t),$$

where the system state $\boldsymbol{s}_t$ is defined in (3).

We assume that r_t^i is bounded, *i.e.*,

$$|r_t^i(x_t^i, a_t^i, c_t)| \leq C_r, \quad \forall\, x_t^i \in \mathcal{X}^i, c_t \in \mathcal{X}^c, \tag{8}$$

and as a result, R_t must be bounded too.

5) Dynamic programming formulation: We formulate the scheduling problem as a T-stage DP. We use $J_t^\pi(\boldsymbol{s}_t)$ to denote the reward-to-go function achieved by policy π:

$$J_t^\pi(\boldsymbol{s}_t) = R_t(\boldsymbol{s}_t, \pi_t(\boldsymbol{s}_t)) + \mathbb{E}[J_{t+1}^\pi(\boldsymbol{s}_{t+1}) \mid \boldsymbol{s}_t, \pi_t(\boldsymbol{s}_t)].$$

We use $J_t(\boldsymbol{s}_t)$ to denote the optimal reward-to-go function under system state $\boldsymbol{s}_t$, and the Bellman's equation yields

$$J_t(\boldsymbol{s}_t) = \max_{\boldsymbol{a}_t \in \mathcal{A}_t(\boldsymbol{s}_t)} \left\{R_t(\boldsymbol{s}_t, \boldsymbol{a}_t) + \mathbb{E}[J_{t+1}(\boldsymbol{s}_{t+1}) \mid \boldsymbol{s}_t, \boldsymbol{a}_t]\right\}. \tag{9}$$

Since $d^i \leq T+1$ for all i, all jobs leave before stage $T+1$, the terminal cost $J_{T+1}(\boldsymbol{s}_{T+1}) = 0$. We say a policy π^* is optimal if and only if $J_1^{\pi^*}(\boldsymbol{s}_1) = J_1(\boldsymbol{s}_1)$ for all initial states $\boldsymbol{s}_1$.

Remark 1: Similar to the deadline scheduling problem considered in [5], the formulated problem can be viewed as a special case of the restless multi-armed bandit (RMAB) problem. Each processor (which is able to process at most one job at a time) can be regarded as an arm, the state of which (x_t^i) may change under both the active and passive actions. The reward achieved by arm i at stage t is given in (7). We note that the reward achieved at different arms are coupled through the processing cost c_t.

The key difference between the formulated problem and the model adopted in [5] lies in constraint (1), which restricts the total processing power to be less than or equal to the available power capacity in the power distribution system m_t [33]. Note also that, unlike in [5], we consider a finite horizon RMAB, which requires significantly different treatment. ■

III. Lagrangian relaxation and decomposition

In Section III-A, we decouple the DP formulated in (9) using Lagrangian relaxation and decomposition [21]. In Section III-B, we adopt a linear programming (LP) approach [34] to solve the relaxed problem and obtain an upper bound (for all feasible policies satisfying (5)).

A. Lagrangian relaxation

Let $\boldsymbol{\lambda}_{t,T} = [\lambda_t, \ldots, \lambda_T]$ represent the vector of Lagrange multipliers from stage t to T with $\lambda_t \geq 0$. The Lagrangian relaxation considers

$$\begin{aligned} L_t(\boldsymbol{s}_t, \boldsymbol{\lambda}_{t,T}) = \max_{\boldsymbol{a}_t \in \{0,1\}^I} &\Big\{\mathbb{E}[L_{t+1}(\boldsymbol{s}_{t+1}, \boldsymbol{\lambda}_{t+1,T}) \mid \boldsymbol{s}_t, \boldsymbol{a}_t] \\ &+ \sum_{i=1}^{I} r_t^i(x_t^i, c_t, a_t^i) + \lambda_t(m_t - \sum_{i=1}^{I} a_t^i)\Big\}, \ \forall\, 1 \leq t \leq T, \end{aligned} \tag{10}$$

where the constraint on maximum processing rate (1) is relaxed and $L_{T+1}(\boldsymbol{s}_{T+1}) = 0$ at the terminal stage $T+1$.

In view of the separable structure of $L_t(\boldsymbol{s}_t, \boldsymbol{\lambda}_{t,T})$, we reformulate (10) based on the following proposition.

Proposition 1 ([21]): The Lagrangian DP (10) can be decomposed as the sum of I independent single-arms' reward-to-go plus an offset term with respect to $\boldsymbol{\lambda}_{t,T}$, *i.e.*,

$$L_t(\boldsymbol{s}_t, \boldsymbol{\lambda}_{t,T}) = \sum_{i=1}^{I} V_t^i(x_t^i, c_t, \boldsymbol{\lambda}_{t,T}) + \sum_{\tau=t}^{T} \lambda_\tau m_\tau,$$

where

$$\begin{aligned} V_t^i(x_t^i, c_t, \boldsymbol{\lambda}_{t,T}) = \max_{a_t^i \in \{0,1\}} &\Big\{r_t^i(x_t^i, c_t, a_t^i) - \lambda_t a_t^i \\ &+ \mathbb{E}[V_{t+1}^i(x_{t+1}^i, c_{t+1}, \boldsymbol{\lambda}_{t+1,T}) \mid x_t^i, c_t, a_t^i]\Big\} \end{aligned} \tag{11}$$

and $V_{T+1}^i(x_{T+1}^i, c_{T+1}) = 0$.

Since all arms have independent job arrivals, it follows from Proposition 1 that the Lagrangian dual problem can be written by

$$L_1^*(\boldsymbol{s}_1, \boldsymbol{\lambda}_{1,T}^*) = \min_{\boldsymbol{\lambda}_{1,T} \geq \boldsymbol{0}} \sum_{i=1}^{I} V_1^i(x_1^i, c_1, \boldsymbol{\lambda}_{1,T}) + \sum_{t=1}^{T} \lambda_t m_t, \tag{12}$$

where $\boldsymbol{\lambda}_{1,T}^* = [\lambda_1^*, \ldots, \lambda_T^*]$ is the optimal solution and $L_1^*(\boldsymbol{s}_1, \boldsymbol{\lambda}_{1,T}^*)$ is the optimal objective value under the initial state $\boldsymbol{s}_1$. Problem (12) is easier to solve, and can provide an upper bound on the total expected reward, *i.e*,

$$L_1^*(\boldsymbol{s}_1, \boldsymbol{\lambda}_{1,T}^*) \geq J_1(\boldsymbol{s}_1).$$

Therefore, to establish the asymptotic optimality of a policy, it is sufficient to show that its expected total reward converges to $L_1^*(\boldsymbol{s}_1, \boldsymbol{\lambda}_{1,T}^*)$.

Problem (12) is equivalent to the following PLP (primal linear program), which enables us to solve the problem through the LP approach.

$$\begin{aligned} \mathit{PLP}: \quad & \min_{\{\lambda_t, x_t^i, c_t, \boldsymbol{\lambda}_{t,T}\}} \sum_{i=1}^{I} V_1^i(x_1^i, c_1, \boldsymbol{\lambda}_{1,T}) + \sum_{t=1}^{T} \lambda_t m_t \\ s.t. \quad & \lambda_t \geq 0, \ \forall\, t, \\ & \sum_{x_{t+1}^i} \sum_{c_{t+1}} \mathbb{P}_t^i(x_{t+1}^i | x_t^i, 1) \mathbb{P}_t^c(c_{t+1} | c_t) V_{t+1}^i(x_{t+1}^i, c_{t+1}, \boldsymbol{\lambda}_{t+1,T}) \\ & \leq V_t^i(x_t^i, c_t, \boldsymbol{\lambda}_{t,T}) - r_t^i(x_t^i, c_t, 1) + \lambda_t, \ \forall\, t,\ i,\ x_t^i,\ c_t, \\ & \sum_{x_{t+1}^i} \sum_{c_{t+1}} \mathbb{P}_t^i(x_{t+1}^i | x_t^i, 0) \mathbb{P}_t^c(c_{t+1} | c_t) V_{t+1}^i(x_{t+1}^i, c_{t+1}, \boldsymbol{\lambda}_{t+1,T}) \\ & \leq V_t^i(x_t^i, c_t, \boldsymbol{\lambda}_{t,T}) - r_t^i(x_t^i, c_t, 0), \ \ \forall\, t,\ i,\ x_t^i,\ c_t, \end{aligned} \tag{13}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2021.3049340, IEEE Transactions on Automatic Control

where $\mathbb{P}^i_t(x^i_{t+1} \mid x^i_t, a^i_t)$ denotes the transition probability from x^i_t to x^i_{t+1} under action a^i_t. The optimal decision variables $V(\cdot)$ of the PLP are equal to the reward-to-go functions at stage $t = 1$ expressed in (11), and the linear program in (13) is equivalent to the dynamic problem in (12) [35].

B. Optimal randomized policy for the dual linear problem

We adopt the linear programming approach (cf. Chapter 3 of [34]) to solve Problem (12). We define the occupation measure $\rho^i_t(x^i_t, c_t, a^i_t)$ as the probability of visiting the state pair (x^i_t, c_t) and taking the action a^i_t at stage t [34]:

$$\sum\nolimits_{a^i_{t+1}} \rho^i_{t+1}(x^i_{t+1}, c_{t+1}, a^i_{t+1}) = \\ \sum\nolimits_{x^i_t} \sum\nolimits_{c_t} \sum\nolimits_{a^i_t} \mathbb{P}^i_t(x^i_{t+1}|x^i_t, a^i_t)\mathbb{P}^c_t(c_{t+1}|c_t)\rho^i_t(x^i_t, c_t, a^i_t). \tag{14}$$

The dual of the *PLP* is the following dual linear program (DLP):

$$DLP: \max_{\rho^i_t(x^i_t, c_t, a^i_t)} \sum_{t=1}^{T}\sum_{i=1}^{I}\sum_{a^i_t}\sum_{x^i_t}\sum_{c_t} \rho^i_t(x^i_t, c_t, a^i_t) r^i_t(x^i_t, c_t, a^i_t)$$

$$\text{s.t.} \quad (14), \ \sum_{a^i_1} \rho^i_1(x^i_1, c_1, a^i_1) = 1, \ \forall\, i, \tag{15a}$$

$$\sum_{i=1}^{I}\sum_{x^i_t}\sum_{c_t} \rho^i_t(x^i_t, c_t, 1) \le m_t, \ \forall\, t, \tag{15b}$$

$$\rho^i_t(x^i_t, c_t, a^i_t) \ge 0, \ \forall\, i, \, t, \, x^i_t \in \mathcal{X}^i, \, c_t \in \mathcal{X}^c, \, a^i_t. \tag{15c}$$

The constraint in (15a) requires that the probability of visiting the initial state pair (x^i_1, c_1) equals 1. For $t > 1$, multiple state pairs can be possibly visited, and the visiting probability $\sum_{a^i_t} \rho^i_t(x^i_t, c_t, a^i_t)$ for any single state pair (x^i_t, c_t) can be less than 1. Constraints (14) and (15a) ensure that the sum of occupation measures must equal 1 for each arm, *i.e.*,

$$\sum_{x^i_t}\sum_{c_t}\sum_{a^i_t} \rho^i_t(x^i_t, c_t, a^i_t) = 1, \ \ \forall\, t.$$

The Lagrangian dual problem (12) is equivalent to the PLP, and the DLP is the dual of PLP. The Lagrangian relaxation in Section III-A essentially replaces constraint (1) with the constraint in (15b) that requires the processing rate limit constraint to hold in expectation over a randomized policy.

The optimal objective value of *DLP* equals $L^*_1(\boldsymbol{s}_t, \boldsymbol{\lambda}^*_{1,T})$, where $\boldsymbol{\lambda}^*_{1,T}$ is the optimal dual variable associated with the constraint (15b). Based on the optimal solution of the DLP, we define an optimal randomized (Markovian) policy $\phi^* = \{\phi^{i*}_t\}^{I,\ \ T}_{i=1,t=1}$, with

$$\phi^{i*}_t(1 \mid x^i_t, c_t) = \frac{\rho^{i^*}_t(x^i_t, c_t, 1)}{\rho^{i^*}_t(x^i_t, c_t, 0) + \rho^{i^*}_t(x^i_t, c_t, 1)} \tag{16}$$

being the probability of taking an active action when the state (of arm i) x^i_t and the processing cost state c_t are visited by policy ϕ^*. Here, $\rho^{i^*}_t(x^i_t, c_t, a^i_t)$ is the optimal solution of the *DLP*. A state pair (x^i_t, c_t) with $\rho^{i^*}_t(x^i_t, c_t, 0) + \rho^{i^*}_t(x^i_t, c_t, 1) = 0$ is visited by policy ϕ^* with zero probability. The randomized policy ϕ^* will be useful for the tie-breaking of the proposed Lagrangian priority policy (cf. Section IV-B).

IV. Lagrangian Priority Policy

In Section IV-A, we introduce the Lagrangian priority value, establish its close form, and derive a recursive formula for the computation of Lagrangian priority values (in the case with deterministic processing cost). In Section IV-B, we propose a new tie-breaking rule for the Lagrangian priority policy, which is crucial for the asymptotic optimality of the Lagrangian priority policy (to be established in the next section).[3]

A. Lagrangian priority value

To develop the priority value of arm i, we regard the current Lagrange multiplier λ_t in (11) as a *tax* on activation. In the following λ_t-tax reward maximization problem, we fix the future Lagrange multipliers $\boldsymbol{\lambda}^*_{t+1,T}$ (the tail part of optimal solution of (12) ranging from $t+1$ to T) to approximate a single arm i's optimal reward-to-go:

$$\begin{aligned} &V^i_t(x^i_t, c_t, \lambda_t, \boldsymbol{\lambda}^*_{t+1,T}) \\ &\quad = \max_{a^i_t \in \{0,1\}} \Big\{ \mathbb{E}[V^i_{t+1}(x^i_{t+1}, c_{t+1}, \boldsymbol{\lambda}^*_{t+1,T}) \mid x^i_t, c_t, a^i_t] \\ &\qquad\qquad + r^i_t(x^i_t, c_t, a^i_t) - a^i_t\lambda_t \Big\}, \end{aligned} \tag{17}$$

where the objective is to maximize the total accumulative reward over the remaining time horizon.

Fixing the multiplier vector $\boldsymbol{\lambda}^*_{t+1,T}$, we let $\Delta^i_t(x^i_t, c_t, \lambda_t, \boldsymbol{\lambda}^*_{t+1,T})$ denote the difference in the rewards achieved by the active action and the deactivated action on the right-hand-side of (17). The **Lagrangian priority value** at state (x^i_t, c_t) is defined as the maximum tax under which makes the two actions in (17) equally attractive at period t, *i.e.*,

$$v^i_t(x^i_t, c_t) \doteq \sup\nolimits_{\lambda_t} \ \{\lambda_t : \Delta^i_t(x^i_t, c_t, \lambda_t, \boldsymbol{\lambda}^*_{t+1,T}) = 0\}. \tag{18}$$

Different from the classic Whittle index, the computation of an arm's Lagrangian priority value involves the states of other arms, through the Lagrange multipliers $\boldsymbol{\lambda}^*_{t+1,T}$ that are optimal for the relaxed dual problem (12). However, the Lagrange multiplier vector $\boldsymbol{\lambda}^*_{1,T}$ is predetermined by solving Problem (13) and is the same for all arms. Since all arms are decoupled due to the Lagrangian relaxation, fixing $\boldsymbol{\lambda}^*_{t+1,T}$ in (17), an arm's Lagrangian priority value depends only on its own state. We note that the interpretation of the defined Lagrangian priority value is analogous to that of the Whittle index (proposed in [10]), which is the (minimum) subsidy on deactivation that makes the two actions equally attractive at a certain state.

Based on (18), the following theorem establishes the closed-form expressions of the Lagrangian priority value for the formulated RMAB.

Theorem 1: The Lagrangian priority value is given by (for arm i at stage t):

$$\begin{aligned} &v^i_t(x^i_t, c_t) \\ &= r^i_t(x^i_t, c_t, 1) + \mathbb{E}[V^i_{t+1}(x^i_{t+1}, c_{t+1}, \boldsymbol{\lambda}^*_{t+1,T}) \mid x^i_t, c_t, 1] \\ &\quad - r^i_t(x^i_t, c_t, 0) - \mathbb{E}[V^i_{t+1}(x^i_{t+1}, c_{t+1}, \boldsymbol{\lambda}^*_{t+1,T}) \mid x^i_t, c_t, 0]. \end{aligned} \tag{19}$$

[3] The importance of tie-breaking to establish asymptotic results for the Whittle index policy has been observed in existing literature [29], [36].

The expression of the Lagrangian priority values in (19) follows from the fact that the right-hand-side of (17) is linear in λ_t. Please refer to Appendix A for detailed proof.

Theorem 1 shows that given the Lagrange multiplier vector $\boldsymbol{\lambda}^*_{1,T}$ (predetermined by Problem (13)), each arm i has a **unique** Lagrangian priority value $v^i_t(x^i_t, c_t)$ that depends only on its current state (x^i_t, c_t). Similar to [5], the Lagrangian priority value $v^i_t(x^i_t, c_t)$ is a scalar reflecting the relative attractiveness of activating job i under state x^i_t and processing cost c_t at stage t. To compute the Lagrangian priority value for arm i according to (19), we need to solve Problem (17) using backward induction. We propose Algorithm 1 to calculate the Lagrangian priority value.

Algorithm 1 Computation of arm i's Lagrangian priority value

INPUT: $\boldsymbol{\lambda}^*_{1,T}$ by solving (13).
OUTPUT: $\{v^i_t(x^i_t, c_t)\}$, $\forall\, i, t, x^i_t, c_t$.

```
for $i = 1, \ldots, I$ do
    Preassign $V^i_{T+1}(x^i_{T+1}, c_{T+1}) = 0$
    for $t = T, \ldots, 1$ do
        for $x^i_t \in \mathcal{X}^i, c_t \in \mathcal{X}^c$ do
            Compute $\mathbb{E}[V^i_{t+1}(x^i_{t+1}, c_{t+1}, \boldsymbol{\lambda}^*_{t+1,T}) \mid x^i_t, c_t, a^i_t]$
            Compute $v^i_t(x^i_t, c_t)$ according to (19)
            Compute and store $V^i_t(x^i_t, c_t, \boldsymbol{\lambda}^*_{t,T})$ by (11)
        end for
    end for
end for
```

Algorithm 1 requires a backward induction process from stage T to stage 1. In (19), the value of current tax λ_t only influences current action a^i_t and state transitions of arm i from x^i_t to x^i_{t+1}. As future Lagrange multipliers are fixed as $\boldsymbol{\lambda}^*_{t+1,T}$, the reward-to-go function $V^i_{t+1}(x^i_{t+1}, c_{t+1}, \boldsymbol{\lambda}^*_{t+1,T})$ is independent of λ_t, and the Lagrangian priority value can be obtained using the results of backward induction at prior stage $t+1$ (directly through (19)).

The Lagrangian relaxation introduced in Section III-A decouples all arms in the PLP (13). Hence, the size of the DLP grows linearly with the number of arms I. Further, the computational complexity of backward inductions in Algorithm 1 increases linearly with the number of arms I. In contrast, the complexity of solving the original RMAB problem without relaxation increases exponentially with the number of arms [37].

Remark 2: Different from the Lagrangian index (defined under any feasible Lagrange multiplier associated with the coupling constraint (1)) proposed in [38], our Lagrangian priority value in (18) is defined under a set of special Lagrange multipliers that are optimal solutions of the relaxed dual problem (12). We note that similar Lagrangian priority values are defined in [32] for a RMAB problem with a time-invariant limit on the total processing rate at each stage.

When the processing cost is deterministic, the Lagrangian priority values are given in closed-form recursive formula in the following theorem.

Theorem 2: Suppose that the processing cost is deterministic. We have the following recursive expressions for the Lagrangian priority values:

$$v^i_t(d, e, c_t) = \begin{cases} 0, & \text{if } d = t \text{ or } e = 0, \\ 1 - c_t + g(e) - g(e-1), & \\ & \text{if } d = t+1, e \geq 1. \end{cases} \tag{20}$$

Otherwise, we have

$$v^i_t(d, e, c_t) = \begin{cases} A - c_t + c_{t+1}, & \text{if } A > \lambda^*_{t+1}, \\ B - c_t + c_{t+1}, & \text{if } B < \lambda^*_{t+1}, \\ \lambda^*_{t+1} - c_t + c_{t+1}, & \text{if } A \leq \lambda^*_{t+1} \leq B, \end{cases} \tag{21}$$

where $A = v^i_{t+1}(d, e-1, c_{t+1})$, $B = v^i_{t+1}(d, e, c_{t+1})$.

The proof of Theorem 2 is given in Appendix B.

B. Tie-breaking rule for the Lagrangian priority policy

The *Lagrangian priority policy* sorts the arms in descending order based on their Lagrangian priority values (uniquely defined in (18)), and activates up to m_t arms with non-negative priority values. There is a positive probability that the some of the Lagrangian priority values are equal. It turns out that that the tie-breaking rule can affect the asymptotic optimality of the Lagrangian priority policy. In what follows, we propose a tie-breaking rule that is based on the optimal randomized policy ϕ^* introduced in (16) [32], [38].

Definition 1: The *Lagrangian priority policy* $\bar{\pi}$ with random tie-breaking: First sort all arms in descending order based on the following modified (random) Lagrangian priority values

$$\bar{v}^i_t(x^i_t, c_t) = \begin{cases} v^i_t(x^i_t, c_t) + \delta, & \text{with probability } \eta^i_t(x^i_t, c_t), \\ v^i_t(x^i_t, c_t) - \delta, & \text{with probability } 1 - \eta^i_t(x^i_t, c_t), \end{cases} \tag{22}$$

where $\eta^i_t(x^i_t, c_t) = \phi^{i*}_t(1 \mid x^i_t, c_t)$ (defined in Eq. (16)) if (x^i_t, c_t) is visited by ϕ^*, otherwise $\eta^i_t(x^i_t, c_t) = 0.5$. Here, δ is a small positive constant such that (22) does not change the order of arms with different Lagrangian priority values. Second, activate up to m_t arms with non-negative modified priority values, and uniformly breaks the ties among modified priority values. ■

The key idea of the proposed Lagrangian priority policy is to randomize Lagrangian priority values to mimic the solution of the dual linear program in Section III-B. As we will show in the next section, the proposed tie-breaking rule is crucial for the asymptotic optimality of the proposed policy $\bar{\pi}$.

Remark 3: The authors of [32] develop a similar tie-breaking approach that allocates resource among tied arms (with the same Lagrangian priority value) in proportion to their occupation measures $\{\rho^{i^*}_t(x^i_t, c_t, 1)\}$ (i.e., optimal solutions to the DLP in Eq. (15)). Different from the proportional allocation adopted in [32], the proposed approach applies randomized priority values to determine the priority among tied arms, and therefore avoids the rounding of fractional actions into binary actions. Further, the randomized priority values defined in (22) are determined by the **conditional** probability that each state is activated by the optimal randomized policy ϕ^*, under the condition that the state is

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2021.3049340, IEEE Transactions on Automatic Control

visited by ϕ^* (cf. Eq. (16)), whereas the tie-breaking rule proposed in [32] depends on the **unconditional** probability that each state is visited and activated by ϕ^*, $\rho_t^{i^*}(x_t^i, c_t, 1)$.

V. Asymptotic optimality

For the case with deterministic processing costs, we now establish the asymptotic optimality of the proposed Lagrangian priority policy, as the total number of processors I, the maximum processing rates $\{m_t\}$, and the average number of job arrivals (per period) increase simultaneously to infinity. The asymptotic regime is consistent with that considered by Whittle [10].

Let $\beta \in \mathbb{Z}^+$ denote the scaling parameter for the original system defined in Section II. In the augmented system with parameter β, the system states are given by

$$\boldsymbol{s}_t(\beta) = [x_t^1(\beta), \ldots, x_t^j(\beta), \ldots, x_t^{\beta I}(\beta), c_t],$$

where $x_t^j(\beta)$ is the state of arm j in the augmented system. In the augmented system with parameter β, let $\mathcal{G}^i(\beta)$ denote the set of arms in group i, in which all the β arms share the same initial state at stage $t = 1$, as well as the same (probabilistic) distributions on job arrivals and job initial states, with the ith arm of the original system. The original system has I arms, and therefore an augmented system has I such groups of arms.

At each stage, the action vector must respect the following time-varying processing rate constraint:

$$\sum_{j=1}^{\beta I} a_t^j \leq \beta m_t, \quad \forall t.$$

Let $J_1^{\bar{\pi}}[\boldsymbol{s}_1(\beta)]$ denote the total reward collected by the proposed policy $\bar{\pi}$ (cf. Definition 1). We have

$$\beta L_1^*(\boldsymbol{s}_1, \boldsymbol{\lambda}_{1,T}^*) \geq J_1^{\bar{\pi}}[\boldsymbol{s}_1(\beta)],$$

where $L_1^*(\boldsymbol{s}_1, \boldsymbol{\lambda}_{1,T}^*)$ is the total expected reward resulting from the Lagrangian dual problem (cf. Eq. (12)). The main result of this section will establish the convergence of $J_1^{\bar{\pi}}[\boldsymbol{s}_1(\beta)]$ to the corresponding upper bound achieved in the Lagrangian dual problem as $\beta \to \infty$ (cf. Theorem 3).

Due to the stochasticity of job arrivals and job initial states, different arms in $\mathcal{G}^i(\beta)$ can visit different states at stage $t > 1$ under any given policy. Under the proposed policy $\bar{\pi}$, let $\bar{\mathcal{O}}_t^i(x, \beta) \subseteq \mathcal{G}^i(\beta)$ denote the set of arms at system state x at stage t, *i.e.*,

$$\bar{\mathcal{O}}_t^i(x, \beta) \doteq \{j \in \mathcal{G}^i(\beta) \mid x_t^j(\beta) = x\}, \; \forall \, x \in \mathcal{X}^i, 1 \leq i \leq I.$$

Note that $\cup_{x \in \mathcal{X}^i} \bar{\mathcal{O}}_t^i(x, \beta) = \mathcal{G}^i(\beta)$. We further define

$$\begin{aligned} &\bar{\mathcal{N}}_t^i(x, 1, \beta) \\ &\doteq \left\{ j \in \bar{\mathcal{O}}_t^i(x, \beta) \mid \text{the jth element of } \bar{\pi}_t(\boldsymbol{s}_t(\beta)) \text{ is } 1 \right\} \end{aligned}$$

as the set of arms in $\bar{\mathcal{O}}_t^i(x, \beta)$ activated by policy $\bar{\pi}$, and $\bar{\mathcal{N}}_t^i(x_t^i, 0, \beta)$ as its complementary set in $\bar{\mathcal{O}}_t^i(x_t^i, \beta)$.

For the original system with deterministic costs $\{c_t\}_{t=1}^T$, the optimal occupation measure $\rho_t^{i^*}(x, c_t, a)$ of the *DLP* (defined in (15)) reduces to $\rho_t^{i^*}(x, a)$, the probability of visiting state x and taking an optimal action a. In the augmented system with parameter β, we let $\rho_t^{i^*}(x, a)$ denote the optimal occupation measures of all arms in the set $\bar{\mathcal{O}}_t^i(x, \beta)$.

In the following lemma, we establish an important link between the proposed policy $\bar{\pi}$ and the optimal randomized policy ϕ^*.

Lemma 1: Suppose that the processing costs are deterministic. The ratios of activated and deactivated arms in any $\mathcal{G}^i(\beta)$ under the proposed policy $\bar{\pi}$ converge to the probabilities that the optimal randomized policy ϕ^* takes active and passive actions, respectively, *i.e.*,

$$\lim_{\beta \to \infty} \frac{\mathbb{E}^{\bar{\pi}}\left[|\bar{\mathcal{N}}_t^i(x, 1, \beta)|\right]}{\beta} = \rho_t^{i^*}(x, 1), \; \forall \, x \in \mathcal{X}^i, i, t, \quad \text{(23a)}$$

$$\lim_{\beta \to \infty} \frac{\mathbb{E}^{\bar{\pi}}\left[|\bar{\mathcal{N}}_t^i(x, 0, \beta)|\right]}{\beta} = \rho_t^{i^*}(x, 0), \; \forall \, x \in \mathcal{X}^i, i, t, \quad \text{(23b)}$$

where the expectation is over the stochasticity in job arrivals, job initial states, and the random priority values (defined in (22)) adopted by the policy $\bar{\pi}$.

The proof of Lemma 1 is given in Appendix C. The result in (23) suggests that the proposed policy $\bar{\pi}$ mimics the optimal randomized policy for the DLP (15). It enables us to prove the asymptotic optimality for the proposed policy $\bar{\pi}$ in the following theorem.

Theorem 3: Suppose that the processing costs are deterministic. The Lagrangian priority policy $\bar{\pi}$ with random tie-breaking (proposed in Definition 1) is asymptotically optimal as β grows large, *i.e.*,

$$\lim_{\beta \to \infty} \frac{\beta L_1^*(\boldsymbol{s}_1, \boldsymbol{\lambda}_{1,T}^*) - J_1^{\bar{\pi}}[\boldsymbol{s}_1(\beta)]}{\beta} = 0,$$

where $L_1^*(\boldsymbol{s}_1, \boldsymbol{\lambda}_{1,T}^*)$ is an upper bound on the total expected reward (resulting from the Lagrange relaxation in Eq. (12)).

The proof is given in Appendix D. Theorem 3 shows that the expected reward per arm achieved by the proposed policy $\bar{\pi}$ converges to its corresponding upper bound.

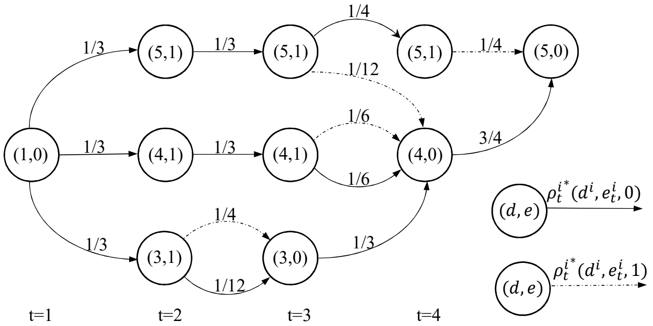


Fig. 2. Transitions of the optimal occupation measures under policy ϕ^*. Each node represents a state $x = (d, e)$, and the weight on a solid (dash) edge starting from node x represents $\rho_t^{i^*}(x, 0)$ ($\rho_t^{i^*}(x, 1)$, respectively).

In the following example, we present a simple case study with $T = 4$ to demonstrate that the tie-breaking rule in (22) is crucial for the asymptotic optimality results established in Lemma 1 and Theorem 3.

Example 1: Consider a heavy traffic scenario with $m_t/I = 0.25$, *i.e.*, at most 1/4 of arms can be simultaneously activated

TABLE I
STATE EVOLUTION AND ACTIONS TAKEN UNDER THE PROPOSED POLICY $\bar{\pi}$.

	t=2			t=3		t=4
x_t^i	(5,1)	(4,1)	(3,1)	(5,1)	(4,1)	(5,1)
Lagrangian priority value v	0.6	0.6	0.6	1	1	0.6
$\phi_t^{i*}(1 \mid x_t^i, c_t)$	0	0	3/4	1/4	1/2	1
Random Lagrangian priority value $\bar{v}$	0.6-δ	0.6-δ	0.6+δ with Prob. 3/4 0.6-δ with Prob. 1/4	1+δ with Prob. 1/4 1-δ with Prob. 3/4	1+δ with Prob. 1/2 1-δ with Prob. 1/2	0.6+δ

at each stage. The deterministic costs at the four stages are $\boldsymbol{c}_{1,4} = [0.3, 0.7, 0.3, 0.7]$, and the penalty function is $g(x) = 0.3x^2$. All processors are unoccupied at $t = 1$. Jobs arrive at $t = 2$ with three equally possible initial states, {(5,1), (4,1), (3,1)}. Fig. 2 shows the transitions of the optimal occupation measures (solution to the DLP in (15)). Through Eq. (16), these optimal occupation measures lead to the probability of taking active actions under the optimal randomized policy ϕ^* (as listed in the third row of Table I). With the Lagrangian priority values (computed by Theorem 2) and the actions taken by policy ϕ^*, via Eq. (22), we obtain the random Lagrangian priority values for the proposed policy $\bar{\pi}$ (on the fourth row of Table I).

We take the state $(5, 1)$ at $t = 3$ as an example to demonstrate that Lemma 1 holds. When β is sufficiently large, according to the law of large numbers, approximately $1/4$ of the βI arms with a priority value $1+\delta$ are activated by policy $\bar{\pi}$, in which approximately $1/6$ are at state $(4, 1)$ and the other approximately $1/12$ are at state $(5, 1)$. Since the probability of visiting state $(5, 1)$ is 1/3 and the activation ratio under policy $\bar{\pi}$ (the left hand of (23a)) approximately equals $1/12$, the deactivation ratio (the left hand of (23b)) is approximately $1/4$. From Fig. 2, ${\rho_3^i}^*(5, 1, 1) = 1/12$ and ${\rho_3^i}^*(5, 1, 0) = 1/4$. That is, Lemma 1 holds for the state $(5, 1)$ at $t = 3$. Indeed, as β grows large, the (random) actions taken by policy $\bar{\pi}$ converges to the optimal randomized actions (listed in the third row of Table I), which establishes the asymptotic optimality result in Theorem 3.

It is interesting to note that the tie-breaking rule (specified in (22)) is crucial for the asymptotic optimality of policy $\bar{\pi}$, as the two states share the same Lagrangian priority value at $t = 3$ (cf. the second row in Table I). ■

VI. NUMERICAL RESULTS

In this section, we numerically compare the performance of the Lagrangian priority policy $\bar{\pi}$ with the proposed tie-breaking rule (based on the randomized Lagrangian priority values proposed in (22)), a Lagrangian priority policy with uniform tie-breaking, the Whittle index policy, and an upper bound resulting from the Lagrange relaxation (12).

The Whittle index is defined as [10]:

$$\begin{aligned} & w_t^i(x_t^i, c_t) \doteq \\ & \inf_v \{ r_t^i(x_t^i, c_t, 1) + \mathbb{E}[V_{t+1}^i(x_{t+1}^i, c_{t+1}, v\mathbf{1}_{t+1,T}) | x_t^i, c_t, 1] \\ & = r_t^i(x_t^i, c_t, 0) + \mathbb{E}[V_{t+1}^i(x_{t+1}^i, c_{t+1}, v\mathbf{1}_{t+1,T}) | x_t^i, c_t, 0] + v \}, \end{aligned} \tag{24}$$

where $\mathbf{1}_{t+1,T}$ is a $T - t$ dimensional unit vector consisting of all one elements. The authors of [5] have established the indexability for the Whittle indices $\{w_t^i(x_t^i, c_t)\}$, which are computed through the binary search. The Whittle index policy charges up to m_t EVs with the highest (non-negative) Whittle index. The upper bound (on per arm reward) is obtained by solving the relaxed Problem (12).

We obtained 27 days' real-time hourly electricity prices from PJM (https://www.pjm.com/markets-and-operations/energy.aspx) in February, 2019. Following [39], we train the price data as a Markov chain to represent the real-time processing cost, where the transition probability is estimated using the frequency of price changes over states. We also obtain the daily power supply surplus data from the California Independent System Operator (http://www.caiso.com/Pages/default.aspx). The data of surplus supply are normalized and rounded into integers to represent the time-varying charging capacity m_t.

In our simulation, each stage lasts for 4 hours. The entire scheduling horizon is 4 days, i.e., $T = 24$. The arrivals of EVs among I chargers follow a binomial distribution with a constant arrival rate γ. We assume that all chargers are unoccupied at the beginning of $t = 1$, and consider the case with time-varying, deterministic processing costs. This setting is practical as in some deregulated electricity markets, the EV charging stations can pay day-ahead hourly prices which are settled one day before the real time.

For each scenario in Figs. 3-4, we simulate the 24-period decision horizon for 5000 times and compute the average reward per arm. In Figs. 3-4, the vertical error bar associated with each scenario denotes the 95% confidence interval of the (realized) per arm reward.

A. Simulation in heavily overload regime

We consider a heavy traffic regime. We increase the number of chargers I from 50 to 4000 while keeping $\gamma = 0.3$ and $\mathbb{E}(m_t)/I = 0.205$ constant. The non-completion penalty function is given by $g(x) = 0.3x^2$.

We observe from Fig. 3 that the Lagrangian priority policy with the proposed tie-breaking rule is asymptotically optimal as I grows large, when the ratio $E(m_t)/I$ is fixed. The proposed Lagrangian priority policy (with randomized tie-breaking) improves the expected reward achieved by the Lagrangian priority policy (with uniform tie-breaking) and the Whittle index policy by $6-8\%$ and $22-24\%$, respectively. The performance gap between the proposed Lagrangian priority policy and the index policy in [32] decreases from 5.9% to 2.2% when I increases from 50 to 200. This gap is mainly due to the rounding loss resulting from the proportional allocation proposed in [32]. We note that the performance gap between these two policies shrinks to zero as the system scale grows to infinity.

B. Simulation in slightly overload regime

Fig. 4 shows the simulation result in a slight overloaded scenario, where the charging facilities occasionally fail to meet all the charging demands before their deadlines. The operator suffers tougher non-completion penalties with $g(x) = 0.5x^2$.

In Fig. 4, the two Lagrangian priority value based policies significantly outperform the Whittle index policy, by $45-49\%$ (with the proposed tie-breaking rule) and $33-36\%$ (with uniform tie-breaking) in per arm reward, respectively. The performance gap between the proposed Lagrangian priority policy and the index policy in [32] is 4.9% when $I = 50$, and the gap gradually decreases to 1.9% as I increases to 200.

Different from the Whittle index which does not take into account future charging capacity constraints $\{m_\tau\}_{\tau=t+1}^T$, the proposed Lagrangian priority values are computed with future Lagrange multipliers fixed as $\boldsymbol{\lambda}^*_{t+1,T}$, which are the optimal solutions of the Lagrange relaxation (12), and are associated with the processing rate limit constraint (1) through *complementary slackness*.

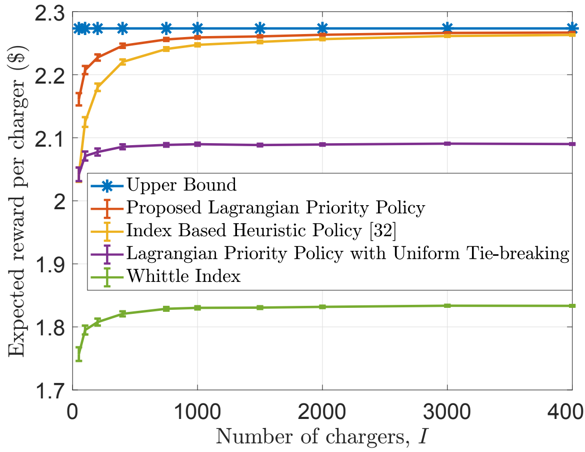


Fig. 3. Performance comparison under (deterministic) time-varying processing cost with $\bar{E} = 3$, $g(x) = 0.3x^2$, $\gamma = 0.3$, $\mathbb{E}(m_t)/I = 0.205$, $\sigma(m_t)/I = 0.057$.

VII. Conclusion

We consider the stochastic deadline scheduling for multiple (randomly arrival) deferrable jobs under time-varying processing rate limits. We formulate the scheduling problem as a restless multi-armed bandit (RMAB) problem. We relax the formulated RMAB problem to decompose it into multiple independent single-arm scheduling problems. We define the *Lagrangian priority value* (associated with an arm at a stage) as the maximum price the operator would like to pay for consuming one unit of energy to activate the arm. We develop closed-form expressions and efficient computational approaches for the Lagrangian priority values.

We propose an priority policy that processes unfinished jobs in the order of their (randomized) Lagrangian priority values.

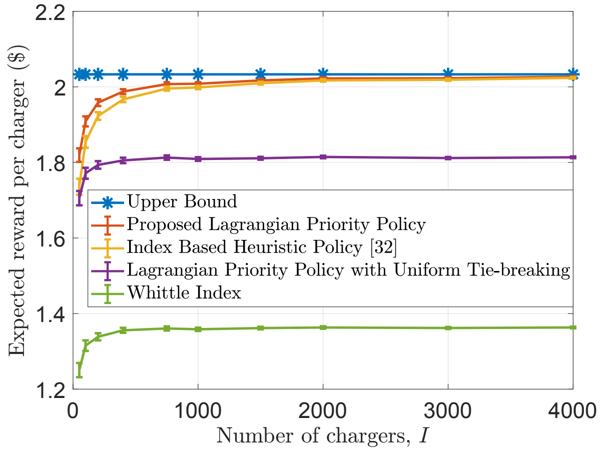


Fig. 4. Performance comparison under deterministic processing costs with $\bar{E} = 4$, $g(x) = 0.5x^2$, $\gamma = 0.18$, $\mathbb{E}(m_t)/I = 0.195$, $\sigma(m_t)/I = 0.057$.

For multiple arms with equal Lagrangian priority values, we propose a new tie-breaking rule that adjusts the Lagrangian priority value of an arm (by adding a small amount) with the probability that the optimal randomized policy (for the Lagrange relaxation of the original problem) takes the active action at the state of the arm. The proposed tie-breaking rule ensures the asymptotic optimality of the proposed Lagrangian priority policy, for the case with (time-varying) deterministic processing costs and stochastic job arrivals. Numerical results show that the proposed Lagrangian priority policy achieves 22%-49% higher average reward than the Whittle index policy.

Appendix A
Proof of Theorem 1

We let

$$\begin{aligned}&\Delta_t^i(x_t^i, c_t, \lambda_t, \boldsymbol{\lambda}^*_{t+1,T})\\ \doteq\, &h_t^i(x_t^i, c_t, \lambda_t) + H_{t+1}^i(x_{t+1}^i, c_{t+1}, \boldsymbol{\lambda}^*_{t+1,T}),\end{aligned}$$

where

$$h_t(x_t^i, c_t, \lambda_t) \doteq r_t^i(x_t^i, c_t, 1) - \lambda_t - r_t^i(x_t^i, c_t, 0)$$

is the immediate reward difference at period t, and $H_{t+1}^i(x_{t+1}^i, c_{t+1}, \boldsymbol{\lambda}^*_{t+1,T})$ is the difference in the reward-to-go functions on the right-hand-side of (17).

We next show that $\Delta_t^i(x_t^i, c_t, \lambda_t, \boldsymbol{\lambda}^*_{t+1,T})$ is continuous and linearly decreasing in λ_t, for all given x_t^i and c_t. It is straightforward to check that the immediate reward difference $h_t^i(x_t^i, c_t, \lambda_t)$ is linear and decreasing in λ_t. Since future Lagrangian multipliers are fixed as $\boldsymbol{\lambda}^*_{t+1,T}$ in (17), $H_{t+1}^i(x_{t+1}^i, c_{t+1}, \boldsymbol{\lambda}^*_{t+1,T})$ is independent from λ_t, *i.e.*, $\partial H_{t+1}^i(\cdot)/\partial \lambda_t = 0$. Hence, $\Delta_t^i(x_t^i, c_t, \lambda_t, \boldsymbol{\lambda}^*_{t+1,T})$ is continuous and linearly decreasing in $\lambda_t, \forall x_t^i, c_t$.

By (8), $h_t^i(x_t^i, c_t, \lambda_t)$ and $H(\cdot)$ are bounded. The continuity and linearity of $\Delta_t^i(x_t^i, c_t, \lambda_t, \boldsymbol{\lambda}^*_{t+1,T})$ implies that there exists a unique threshold $\bar{\lambda}_t$ such that $\Delta_t^i(x_t^i, c_t, \lambda_t, \boldsymbol{\lambda}^*_{t+1,T}) = 0$. Under state (x_t^i, c_t), it is optimal to activate the arm under tax $\lambda_t < \bar{\lambda}_t$, and to deactivate the arm under tax $\lambda_t > \bar{\lambda}_t$. By definition in (18), $\bar{\lambda}_t(x_t^i, c_t)$ is the Lagrangian priority value, which is the greatest tax under which it is optimal to activate the arm (cf. (17)). By the monotonicity of $\Delta_t^i(x_t^i, c_t, \lambda_t, \boldsymbol{\lambda}_{t+1,T}^*)$ in λ_t, the Lagrangian priority value is the reward difference under two actions in (17), which yields the result in (19).

Appendix B
Proof of Theorem 2

We first prove (20). When there is no job for processor i ($x_t^i = (t,0)$), or processor i has completed a job ($x_t^i = (d,0)$), the two actions lead to no difference in terms of reward. Thus, $v_t^i(d,e,c_t) = 0$.

If $x_t^i = (t+1, e)$ with $e \geq 1$, *i.e.*, the job will leave at the beginning of the next period, the current reward difference resulting from the two different actions is $1 - c_t + g(e) - g(e-1)$, and the expected reward-to-go resulting from the two actions are the same in (19). Hence, $v_t^i(t+1,e,c_t) = 1 - c_t + g(e) - g(e-1)$.

In the rest of this proof, we establish the recursive expression (21) for the case with $d \geq t+2$.

Step 1. When $d = t+2$, with $e = 1$ and deterministic processing costs, the Lagrangian priority value is

$$\begin{aligned} v_t^i(t+2,1,c_t) &= 1 - c_t + V_{t+1}^i(t+2,0,c_{t+1},\boldsymbol{\lambda}_{t+1,T}^*) \\ &\quad - V_{t+1}^i(t+2,1,c_{t+1},\boldsymbol{\lambda}_{t+1,T}^*) \\ &= 1 - c_t \\ &\quad - \max\{1 - c_{t+1} - \lambda_{t+1}^*, -g(1)\}, \end{aligned} \tag{25}$$

where the first equality follows from (19) and $\mathbb{P}_t^c(c_{t+1} \mid c_t) = 1$, and the second equality follows from (11). It follows from (20) that $A = 0$, and $B = 1 - c_{t+1} + g(1)$.

Since $A = 0$ and $\lambda_{t+1}^* \geq 0$, we only need to consider the last two cases in (21). If $B < \lambda_{t+1}^*$, the second term inside the maximization operator in (25) is greater, and $v_t^i(t+2,1,c_t) = B - c_t + c_{t+1}$. Otherwise, $v_t^i(t+2,1,c_t) = -c_t + c_{t+1} + \lambda_{t+1}^*$. The recursion expression (21) holds.

When $e \geq 2$, according to (11) and (19), the Lagrangian priority value is

$$\begin{aligned} &v_t^i(t+2,1,c_t) \\ &= 1 - c_t - \max\left\{1 - c_{t+1} - \lambda_{t+1}^* - g(e-1), -g(e)\right\} \\ &\quad + \max\left\{1 - c_{t+1} - \lambda_{t+1}^* - g(e-2), -g(e-1)\right\}. \end{aligned} \tag{26}$$

We have $A = 1 - c_{t+1} + g(e-1) - g(e-2)$, and $B = 1 - c_{t+1} + g(e) - g(e-1)$, directly from (20). By the convexity of the penalty function $g(\cdot)$ we have $B \geq A$. If $A > \lambda_{t+1}^*$, the first terms inside the two maximization operators in (26) are greater, and $v_t^i(t+2,1,c_t) = A - c_t + c_{t+1}$. If $B < \lambda_{t+1}^*$, the second terms inside the two maximization operators in (26) are greater, and $v_t^i(t+2,1,c_t) = B - c_t + c_{t+1}$. Otherwise, $v_t^i(t+2,1,c_t) = -c_t + c_{t+1} + \lambda_{t+1}^*$. The result in (21) holds.

Step 2. When $D > t+2$, with $e = 1$, according to (11) and (19), the Lagrangian priority value is

$$\begin{aligned} &v_t^i(d,1,c_t) \\ &= 1 - c_t + V_{t+1}^i(d,0,c_{t+1},\boldsymbol{\lambda}_{t+1,T}^*) \\ &\quad - \max\big\{1 - c_{t+1} - \lambda_{t+1}^* + V_{t+2}^i(d,0,c_{t+2},\boldsymbol{\lambda}_{t+2,T}^*), \\ &\qquad\qquad V_{t+2}^i(d,1,c_{t+2},\boldsymbol{\lambda}_{t+2,T}^*)\big\}. \end{aligned} \tag{27}$$

By (20), $A = 0$. According to (19),

$$\begin{aligned} B = 1 - c_{t+1} &+ V_{t+2}^i(d,0,c_{t+2},\boldsymbol{\lambda}_{t+2,T}^*) \\ &- V_{t+2}^i(d,1,c_{t+2},\boldsymbol{\lambda}_{t+2,T}^*). \end{aligned}$$

By analogy to the analysis for (25), if $B < \lambda_{t+1}^*$, the second term inside the maximization operator of (27) is greater, and $v_t^i(d,1,c_t) = B - c_t + c_{t+1}$. Otherwise, we have $v_t^i(d,1,c_t) = -c_t + c_{t+1} + \lambda_{t+1}^*$. The induction in (21) holds.

The proof for the case with $e \geq 2$ is more complicated, and requires the following lemma.

Lemma 2: The Lagrangian priority value $v_t^i(d,e,c_t)$ is increasing in e, and the reward-to-go function $V_t^i(d,e,c_t,\boldsymbol{\lambda}_{t,T}^*)$ in (11) is concave in e for any $t < d$.

Proof: We prove Lemma 2 by backward induction.

Step 1. We first show that Lemma 2 holds at the terminal stage of a job when $t = d-1$.

According to (20), $v_t^i(d,e,c_t) = 1 - c_t + g(e) - g(e-1)$, for any $e > 0$. It follows immediately that $v_t^i(d,e',c_t) \geq v_t^i(d,e,c_t)$ for any $e' \geq e > 0$ by the convexity of the penalty function.

According to (19), $v_t^i(d,e,c_t) \geq \lambda_t^*$ implies that the optimal action in (11) (with optimal Lagrange multipliers $\boldsymbol{\lambda}_{t,T}^*$) is to activate. We have

$$\begin{aligned} &V_t^i(d,e+2,c_t,\boldsymbol{\lambda}_{t,T}^*) + V_t^i(d,e,c_t,\boldsymbol{\lambda}_{t,T}^*) \\ &-2V_t^i(d,e+1,c_t,\boldsymbol{\lambda}_{t,T}^*) \\ &= \begin{cases} 2g(e) - g(e+1) - g(e-1), \\ \qquad \text{if } v_t^i(d,e,c_t) \geq \lambda_t^*, \\ g(e) - g(e+1) - 1 + c_t + \lambda_t^*, \\ \qquad \text{if } v_t^i(d,e+2,c_t) \geq v_t^i(d,e+1,c_t) \\ \qquad\qquad \geq \lambda_t^* \geq v_t^i(d,e,c_t), \\ g(e+1) - g(e) + 1 - c_t - \lambda_t^*, \\ \qquad \text{if } v_t^i(d,e+2,c_t) \geq \lambda_t^* \\ \qquad\qquad \geq v_t^i(d,e+1,c_t) \geq v_t^i(d,e,c_t), \\ 2g(e+1) - g(e+2) - g(e), \\ \qquad \text{if } \lambda_t^* \geq v_t^i(d,e+2,c_t), \end{cases} \end{aligned}$$

where the first and last cases are non-positive due to the convexity of the penalty function, the second and third cases are non-positive following the definition of $v_t^i(d,e+1,c_t)$. Thus, $V_t^i(d,e,c_t,\boldsymbol{\lambda}_{t,T}^*)$ is concave in e.

Step 2. Suppose that

$$v_{t+1}^i(d,e+1,c_{t+1}) \geq v_{t+1}^i(d,e,c_{t+1}), \quad \forall e > 0,$$

and that $V_{t+1}^i(d,e,c_{t+1},\boldsymbol{\lambda}_{t+1,T}^*)$ is concave in e. We will show that these properties also hold at stage t.

The Lagrangian priority value is increasing with $e \geq 1$ at stage t, i.e.,

$$\begin{aligned} &v_t^i(d,e+1,c_t) - v_t^i(d,e,c_t) \\ &= 2V_{t+1}^i(d,e,c_{t+1},\boldsymbol{\lambda}_{t+1,T}^*) - V_{t+1}^i(d,e+1,c_{t+1},\boldsymbol{\lambda}_{t+1,T}^*) \\ &\quad -V_{t+1}^i(d,e-1,c_{t+1},\boldsymbol{\lambda}_{t+1,T}^*) \\ &\geq 0, \end{aligned}$$

where the equality follows from (19), and the inequality follows from the condition that $V_{t+1}^i(d,e,c_{t+1},\boldsymbol{\lambda}_{t+1,T}^*)$ is concave in e.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2021.3049340, IEEE Transactions on Automatic Control

By analogy to the analysis in Step 1, we have

$$\begin{aligned}
&V_t^i(d,e+2,c_t,\boldsymbol{\lambda}_{t,T}^*)+V_t^i(d,e,c_t,\boldsymbol{\lambda}_{t,T}^*)\\
&-2V_t^i(d,e+1,c_t,\boldsymbol{\lambda}_{t,T}^*)\\
=&\begin{cases}
V_{t+1}^i(d,e+1,c_{t+1},\boldsymbol{\lambda}_{t+1,T}^*)\\
+V_{t+1}^i(d,e-1,c_{t+1},\boldsymbol{\lambda}_{t+1,T}^*)\\
-2V_{t+1}^i(d,e,c_{t+1},\boldsymbol{\lambda}_{t+1,T}^*),\\
\qquad\qquad\text{if } v_t^i(d,e,c_t)\ge\lambda_t^*,\\
V_{t+1}^i(d,e+1,c_{t+1},\boldsymbol{\lambda}_{t+1,T}^*)\\
-V_{t+1}^i(d,e,c_{t+1},\boldsymbol{\lambda}_{t+1,T}^*)-1+c_t+\lambda_t^*\\
\qquad\text{if } v_t^i(d,e+2,c_t)\ge v_t^i(d,e+1,c_t)\\
\qquad\qquad\ge\lambda_t^*\ge v_t^i(d,e,c_t),\\
V_{t+1}^i(d,e+1,c_{t+1},\boldsymbol{\lambda}_{t+1,T}^*)\\
-V_{t+1}^i(d,e,c_{t+1},\boldsymbol{\lambda}_{t+1,T}^*)+1-c_t-\lambda_t^*,\\
\qquad\qquad\text{if } v_t^i(d,e+2,c_t)\ge\lambda_t^*\\
\qquad\ge v_t^i(d,e+1,c_t)\ge v_t^i(d,e,c_t),\\
V_{t+1}^i(d,e+2,c_{t+1},\boldsymbol{\lambda}_{t+1,T}^*)\\
-2V_{t+1}^i(d,e+1,c_{t+1},\boldsymbol{\lambda}_{t+1,T}^*)\\
+V_{t+1}^i(d,e,c_{t+1},\boldsymbol{\lambda}_{t+1,T}^*)\\
\qquad\qquad\text{if } \lambda_t^*\ge v_t^i(d,e+2,c_t),
\end{cases}
\end{aligned}$$

where the first and last cases are non-positive due to the concavity of the reward-to-go function at $t+1$, and the second and third cases are non-positive because of the definition of $v_t^i(d,e,c_t)$. Thus, $V_t^i(d,e,c_t,\boldsymbol{\lambda}_{t,T}^*)$ is concave in e.

We have proved Lemma 2 by backward induction. ■

We now apply Lemma 2 to obtain the recursion expression (21). By Lemma 2, $B\ge A$. It is therefore sufficient to consider the three cases in (21). According to (11) and (19), the Lagrangian priority value is

$$\begin{aligned}
v_t^i(d,e,c_t)=&1-c_t+\max\big\{V_{t+2}^i(d,e-1,c_{t+2},\boldsymbol{\lambda}_{t+2,T}^*),\\
&1-c_{t+1}-\lambda_{t+1}^*+V_{t+2}^i(d,e-2,c_{t+2},\boldsymbol{\lambda}_{t+2,T}^*)\big\}\\
&-\max\big\{V_{t+2}^i(d,e,c_{t+2},\boldsymbol{\lambda}_{t+2,T}^*),1-c_{t+1}\\
&-\lambda_{t+1}^*+V_{t+2}^i(d,e-1,c_{t+2},\boldsymbol{\lambda}_{t+2,T}^*)\big\}.
\end{aligned}\tag{28}$$

According to (19),

$$\begin{aligned}
A=&1-c_{t+1}+V_{t+2}^i(d,e-2,c_{t+2},\boldsymbol{\lambda}_{t+2,T}^*)\\
&-V_{t+2}^i(d,e-1,c_{t+2},\boldsymbol{\lambda}_{t+2,T}^*),
\end{aligned}$$

and

$$\begin{aligned}
B=&1-c_{t+1}+V_{t+2}^i(d,e-1,c_{t+2},\boldsymbol{\lambda}_{t+2,T}^*)\\
&-V_{t+2}^i(d,e,c_{t+2},\boldsymbol{\lambda}_{t+2,T}^*).
\end{aligned}$$

By analogy to the analysis for (26), $A>\lambda_{t+1}^*$ implies that the second terms inside the maximization operators in (28) are greater, and (28) can be simplified as $v_t^i(d,e,c_t)=A-c_t+c_{t+1}$. When $B<\lambda_{t+1}^*$, the first terms inside the maximization operators in (28) are greater, and thus the Lagrangian priority value is $v_t^i(d,e,c_t)=B-c_t+c_{t+1}$. In the last case with $A\le\lambda_{t+1}^*\le B$, in (28) the first term inside the first maximization operator and the second term inside the second maximization operator are greater. We then have $v_t^i(d,e,c_t)=-c_t+c_{t+1}+\lambda_{t+1}^*$. Hence, the induction expression (21) of the Lagrangian priority value holds for the case with $d>t+2$.

We have proved Theorem 2.

Appendix C
Proof of Lemma 1

Before proving Lemma 1 through mathematical induction, we first establish the following lemma that will be useful throughout the proof.

Recall that $\mathcal{G}^i(\beta)$ is the set of arms in group i and $\bar{\mathcal{O}}_t^i(x,\beta)$ is the set of arms in $\mathcal{G}^i(\beta)$ at system state x at stage t, resulting from policy $\bar{\pi}$. We let $\phi_t^{i*}(1\mid x)$ denote the optimal probability of activating an arm in the set $\bar{\mathcal{O}}_t^i(x,\beta)$ (cf. Eq. (16)).

Lemma 3: For any set $\mathcal{G}^i(\beta)$, if $x\in\mathcal{X}^i$ is visited by policy ϕ^* with positive probability, for any $j\in\bar{\mathcal{O}}_t^i(x,\beta)$ we have

$$\phi_t^{i*}(1\mid x)=\begin{cases}1, & \text{if } v_t^j[x_t^j(\beta),c_t]>\lambda_t^*,\\ 0, & \text{if } v_t^i[x_t^j(\beta),c_t]<\lambda_t^*,\end{cases}$$

and $\phi_t^{i*}(1\mid x)\in[0,1]$ if $v_t^i[x_t^j(\beta),c_t]=\lambda_t^*$.

Proof: Consider arm i in the original system. We first write the dual constraint (cf. Eq. (13)) associated with $\rho_t^{i*}(x,1)$ for any $x\in\mathcal{X}^i$ under deterministic costs,

$$\begin{aligned}
&V_t^i(x,c_t,\boldsymbol{\lambda}_{t,T})+\lambda_t^*\\
&\ge r_t^i(x,c_t,1)+\sum_{x'\in\mathcal{X}^i}\mathbb{P}_t^i(x'|x,1)V_{t+1}^i(x',c_{t+1},\boldsymbol{\lambda}_{t+1,T}),
\end{aligned}\tag{29a}$$

and the dual constraint associated with $\rho_t^{i*}(x_t^i,0)$,

$$\begin{aligned}
&V_t^i(x,c_t,\boldsymbol{\lambda}_{t,T})\\
&\ge r_t^i(x,c_t,0)+\sum_{x'\in\mathcal{X}^i}\mathbb{P}_t^i(x'|x,0)V_{t+1}^i(x',c_{t+1},\boldsymbol{\lambda}_{t+1,T}).
\end{aligned}\tag{29b}$$

It follows from the definition of the Lagrangian priority value in (19) that for any $x\in\mathcal{X}^i$, if $v_t^i(x,c_t)>\lambda_t^*$, then the constraint (29b) is non-binding, and thus $\rho_t^{i*}(x,0)=0$ due to the complementary slackness. By (16), we have $\phi_t^{i*}(1\mid x)=1$. Similar argument applies to the case with $v_t^i(x,c_t)<\lambda_t^*$. When $v_t^i(x,c_t)=\lambda_t^*$, we have $\phi_t^{i*}(1\mid x)\in[0,1]$ by definition.

Since $v_t^j[x_t^j(\beta),c_t]=v_t^i(x,c_t)$ for any $j\in\bar{\mathcal{O}}_t^i(x,\beta)$ and $x\in\mathcal{X}^i$, Lemma 3 holds for any arm $j\in\mathcal{G}^i(\beta)$. The above argument applies to $\mathcal{G}^i(\beta)$ for any $1\le i\le I$. ■

We start the proof of Lemma 1 by showing that Eqs. (23a) and (23b) in Lemma 1 hold at $t=1$.

A. $t=1$

Note that all arms in the set $\mathcal{G}^i(\beta)$ have the same initial state x_1^i at $t=1$. We have

$$|\bar{\mathcal{O}}_1^i(x_1^i,\beta)|=\beta,\ \forall\, i,\tag{30a}$$

$$\rho_1^{i*}(x_1^i,1)+\rho_1^{i*}(x_1^i,0)=1,\ \forall\, i.\tag{30b}$$

We only need to establish (23a) for the initial state x_1^i of the set $\mathcal{G}^i(\beta)$, which implies (23b) according to (30a) and (30b). We will prove this result in two steps.

Step 1. We first show that (23a) holds for any set $\mathcal{G}^i(\beta)$ with $\phi_1^{i*}(1\mid x_1^i)>0$. By Eqs. (16) and (30b), we have $\rho_1^{i*}(x_1^i,1)=\phi_1^{i*}(1\mid x_1^i)$. According to Lemma 3, if $\rho_1^{i*}(x_1^i,1)=1$, then

$$v_1^j[x_1^j(\beta),c_1]\ge\lambda_1^*,\ \forall\, j\in\mathcal{G}^i(\beta),$$

and if $\rho_1^{i*}(x_1^i, 1) \in (0,1)$, then

$$v_1^j[x_1^j(\beta), c_1] = \lambda_1^*, \ \forall\, j \in \mathcal{G}^i(\beta).$$

Due to the law of large numbers, when $\beta \to \infty$, the ratio of arms (with the randomized Lagrangian priority values greater than λ_1^* according to (22)) in $\mathcal{G}^i(\beta)$ converges to $\rho_1^{i*}(x_1^i, 1)$, and the ratio of arms (with the randomized Lagrangian priority values less than λ_1^*) converges to $1 - \rho_1^{i*}(x_1^i, 1)$.

We consider two cases on the value of λ_1^*.

- If $\lambda_1^* = 0$, due to the feasibility of $\rho_1^{i*}(x_1^i, 1)$, *i.e.*,

  $$m_1 - \sum\nolimits_i \rho_1^{i*}(x_1^i, 1) \geq 0,$$

  the ratio of activated arms (with the randomized Lagrangian priority values greater than 0) in $\mathcal{G}^i(\beta)$ under policy $\bar{\pi}$ converges to $\rho_1^{i*}(x_1^i, 1)$, as β increases to infinity. Hence, (23a) holds for any set $\mathcal{G}^i(\beta)$ with $\lambda_1^* = 0$.
- If $\lambda_1^* > 0$, due to the complementary slackness, *i.e.*, $\lambda_1^*[m_1 - \sum_i \rho_1^{i*}(x_1^i, 1)] = 0$, we have

  $$m_1 = \sum_i \rho_1^{i*}(x_1^i, 1). \tag{31}$$

  The ratio of activated arms (with the randomized Lagrangian priority values greater than λ_1^*) in $\mathcal{G}^i(\beta)$ converges to $\rho_1^{i*}(x_1^i, 1)$. This establishes (23a) for any set $\mathcal{G}^i(\beta)$ with $\lambda_1^* > 0$.

Step 2. We show that (23a) holds for any set $\mathcal{G}^i(\beta)$ with $\phi_1^{i*}(1 \mid x_1^i) = 0$. By (16), we have $\rho_1^{i*}(x_1^i, 1) = 0$. According to Lemma 3, $v_1^j[x_1^j(\beta), c_1] \leq \lambda_1^*$ for any $j \in \mathcal{G}^i(\beta)$. Due to (22), the random Lagrangian priority values of all arms in $\mathcal{G}^i(\beta)$ are less than λ_1^*.

We consider the following two cases on the value of λ_1^*.

- If $\lambda_1^* = 0$, all arms are deactivated by the policy $\bar{\pi}$. As a result, (23a) holds for any set $\mathcal{G}^i(\beta)$ with $\lambda_1^* = 0$.
- If $\lambda_1^* > 0$, when $\beta \to \infty$, the ratio of activated arms in $\mathcal{G}^i(\beta)$ converges to 0. This establishes (23a) for any set $\mathcal{G}^i(\beta)$ with $\lambda_1^* > 0$.

Combining the results established in the above two steps, we conclude that Lemma 1 holds for the initial state x_1^i of any set $\mathcal{G}^i(\beta)$.

B. $t > 1$.

Suppose that Lemma 1 holds at stage t, we will establish Lemma 1 at stage $t+1$ in the following four steps.

Step 1. We show that for any set $\mathcal{G}^i(\beta)$, the probability of visiting any state $x' \in \mathcal{X}^i$ at stage $t+1$ under policy $\bar{\pi}$ converges to the corresponding visiting probability under the optimal randomized policy ϕ^* (cf. (16)), *i.e.*, for any i and $x' \in \mathcal{X}^i$,

$$\lim_{\beta\to\infty} \frac{\mathbb{E}^{\bar{\pi}}[|\bar{\mathcal{O}}_{t+1}^i(x', \beta)|]}{\beta} = \rho_{t+1}^{i*}(x', 1) + \rho_{t+1}^{i*}(x', 0). \tag{32}$$

For a given set $\mathcal{G}^i(\beta)$, we let $\mathbb{P}_t^i(x' \mid x, a)$ denote the transition probability from state x at stage t to state x' at stage $t+1$ under action a (cf. Section III-B). We have

$$\begin{aligned}
&\lim_{\beta\to\infty} \frac{\mathbb{E}^{\bar{\pi}}[|\bar{\mathcal{O}}_{t+1}^i(x', \beta)|]}{\beta} \\
=& \lim_{\beta\to\infty} \frac{\sum\limits_{x\in\mathcal{X}^i} \mathbb{P}_t^i(x' \mid x, 1)\mathbb{E}^{\bar{\pi}}[|\bar{\mathcal{N}}_t^i(x, 1, \beta)|]}{\beta} \\
&+ \lim_{\beta\to\infty} \frac{\sum\limits_{x\in\mathcal{X}^i} \mathbb{P}_t^i(x' \mid x, 0)\mathbb{E}^{\bar{\pi}}[|\bar{\mathcal{N}}_t^i(x, 0, \beta)|]}{\beta} \\
=& \sum_{x\in\mathcal{X}^i} \mathbb{P}_t^i(x' \mid x, 1)\rho_t^{i^*}(x, 1) + \sum_{x\in\mathcal{X}^i} \mathbb{P}_t^i(x' \mid x, 0)\rho_t^{i^*}(x, 0) \\
=& \rho_{t+1}^{i*}(x', 1) + \rho_{t+1}^{i*}(x', 0), \ \forall\, x' \in \mathcal{X}^i, i,
\end{aligned}$$

where the first equality is from enumerating all paths of visiting x' at $t+1$ under policy $\bar{\pi}$, the second equality follows from the assumption that Lemma 1 holds at stage t, and the last equality is from (14) and $\mathbb{P}_t^c(c_{t+1} \mid c_t) = 1$ (deterministic processing costs). We have established (32) for any set $\mathcal{G}^i(\beta)$ and $x' \in \mathcal{X}^i$.

Before proceeding to the second step, we introduce some useful notations. According to the randomized tie-breaking rule in (22), different arms in the set $\bar{\mathcal{O}}_t^i(x, \beta)$ can have different Lagrangian priority values, though they are at the same state x. Let

$$\bar{\mathcal{Y}}_t^i(x, 1, \beta) \doteq \{j \in \bar{\mathcal{O}}_t^i(x, \beta) \mid \bar{v}_t^j[x_t^j(\beta), c_t] \geq \lambda_t^*\}, \tag{33}$$

be the set of arms with the randomized Lagrangian priority values (cf. (22)) greater or equal to the current optimal Lagrange multiplier, and $\bar{\mathcal{Y}}_t^i(x, 0, \beta)$ as its complementary in $\bar{\mathcal{O}}_t^i(x, \beta)$. We have

$$\begin{aligned}
&\bar{\mathcal{Y}}_t^i(x, 1, \beta) \cup \bar{\mathcal{Y}}_t^i(x, 0, \beta) = \bar{\mathcal{O}}_t^i(x, \beta) \\
=& \bar{\mathcal{N}}_t^i(x, 1, \beta) \cup \bar{\mathcal{N}}_t^i(x, 0, \beta), \ \forall\, x \in \mathcal{X}^i, i.
\end{aligned} \tag{34}$$

Step 2. We prove that for any set $\mathcal{G}^i(\beta)$ the ratios of arms in the sets $\bar{\mathcal{Y}}_{t+1}^i(x, 1, \beta)$ and $\bar{\mathcal{Y}}_{t+1}^i(x, 0, \beta)$ converge to $\rho_{t+1}^{i*}(x, 1)$ and $\rho_{t+1}^{i*}(x, 0)$, respectively. That is,

$$\lim_{\beta\to\infty} \frac{\mathbb{E}^{\bar{\pi}}[|\bar{\mathcal{Y}}_{t+1}^i(x, 1, \beta)|]}{\beta} = \rho_{t+1}^{i*}(x, 1), \ \forall\, x \in \mathcal{X}^i, i, \tag{35a}$$

$$\lim_{\beta\to\infty} \frac{\mathbb{E}^{\bar{\pi}}[|\bar{\mathcal{Y}}_{t+1}^i(x, 0, \beta)|]}{\beta} = \rho_{t+1}^{i*}(x, 0), \ \forall\, x \in \mathcal{X}^i, i. \tag{35b}$$

We only need to prove (35a), which implies (35b) due to Eqs. (32) and (34).

For any state $x \in \mathcal{X}^i$ that is visited by policy ϕ^* with zero probability, we have $\rho_{t+1}^{i*}(x, 1) = \rho_{t+1}^{i*}(x, 0) = 0$. Therefore, both sides of (35a) equal 0.

We now show that (35a) holds when $\phi_{t+1}^{i*}(1 \mid x) = 1$. For any $j \in \bar{\mathcal{O}}_{t+1}^i(x, \beta)$, we have

$$\bar{v}_{t+1}^j[x_{t+1}^j(\beta), c_{t+1}] > v_{t+1}^j[x_{t+1}^j(\beta), c_{t+1}] \geq \lambda_{t+1}^*, \tag{36}$$

where the first inequality follows from (22) and the second inequality follows from Lemma 3. Eq. (36) implies that $\bar{\mathcal{Y}}_{t+1}^i(x, 1, \beta) = \bar{\mathcal{O}}_{t+1}^i(x, \beta)$. When $\phi_{t+1}^{i*}(1 \mid x) = 1$, we have $\rho_{t+1}^{i*}(x, 0) = 0$ by (16). Therefore, (35a) follows from (32).

We next establish (35a) when $\phi^{i*}_{t+1}(1 \mid x) = 0$. Following a similar argument for (36), we have

$$\bar{v}^j_{t+1}[x^j_{t+1}(\beta), c_{t+1}] < v^j_{t+1}[x^j_{t+1}(\beta), c_{t+1}] \le \lambda^*_{t+1}, \qquad (37)$$

for any $j \in \bar{\mathcal{O}}^i_{t+1}(x,\beta)$. (37) implies that $\bar{\mathcal{Y}}^i_{t+1}(x,1,\beta) = \emptyset$. We have $\rho^{i*}_{t+1}(x,1) = 0$ by (16). Hence, (35a) holds due to (32).

Finally, we show that (35a) holds when $\phi^{i*}_{t+1}(1 \mid x) \in (0,1)$. By Lemma 3, the Lagrangian priority value $v^j_{t+1}[x^j_{t+1}(\beta), c_{t+1}] = \lambda^*_{t+1}$ for any $j \in \bar{\mathcal{O}}^i_{t+1}(x,\beta)$. Since the randomized Lagrangian priority values defined in (22) are *i.i.d* for all arms in the set $\bar{\mathcal{Y}}^i_{t+1}(x,1,\beta)$, $|\bar{\mathcal{Y}}^i_{t+1}(x,1,\beta)|$ is the sum of *i.i.d* Bernoulli random variables, which follows the binomial distribution, *i.e.*,

$$|\bar{\mathcal{Y}}^i_{t+1}(x,1,\beta)| \sim B\big(|\bar{\mathcal{O}}^i_{t+1}(x,\beta)|, p_s\big), \quad \forall\, x \in \mathcal{X}^i, i. \quad (38)$$

According to the tie-breaking rule (22), $p_s = \phi^{i*}_{t+1}(1 \mid x)$ when $\phi^{i*}_{t+1}(1 \mid x) \in (0,1)$. We have

$$\begin{aligned}\lim_{\beta\to\infty} \frac{\mathbb{E}^{\bar\pi}[|\bar{\mathcal{Y}}^i_{t+1}(x,1,\beta)|]}{\beta} &= \lim_{\beta\to\infty} \frac{\mathbb{E}^{\bar\pi}[|\bar{\mathcal{O}}^i_{t+1}(x,\beta)|]\phi^{i*}_{t+1}(1 \mid x)}{\beta} \\ &= \rho^{i*}_{t+1}(x,1),\ \forall\, x \in \mathcal{X}^i,\end{aligned}$$

where the second equality follows from (16) and (32).

We conclude that (35a) and (35b) hold for any $x \in \mathcal{X}^i$ and any set $\mathcal{G}^i(\beta)$.

Step 3. In this step, we show that

$$\begin{aligned}&\sum_i \sum_{x\in\mathcal{X}^i} \big||\bar{\mathcal{Y}}^i_{t+1}(x,1,\beta)| - |\bar{\mathcal{N}}^i_{t+1}(x,1,\beta)|\big| \\ \le &\sum_i \sum_{x\in\mathcal{X}^i} \big||\bar{\mathcal{Y}}^i_{t+1}(x,1,\beta)| - \beta\rho^{i*}_{t+1}(x,1)\big|.\end{aligned} \qquad (39)$$

Step 3.1. We first establish (39) for the case with $\lambda^*_{t+1} = 0$.

According to Definition 1, arms in the set $\bar{\mathcal{N}}^i_{t+1}(x,1,\beta)$ must also be in the set $\bar{\mathcal{Y}}^i_{t+1}(x,1,\beta)$, *i.e.*,

$$\bar{\mathcal{N}}^i_{t+1}(x,1,\beta) \subseteq \bar{\mathcal{Y}}^i_{t+1}(x,1,\beta), \quad \forall\, i,\, x \in \mathcal{X}^i. \qquad (40)$$

As a result, we have

$$\begin{aligned}&\sum_i \sum_{x\in\mathcal{X}^i} \big||\bar{\mathcal{Y}}^i_{t+1}(x,1,\beta)| - |\bar{\mathcal{N}}^i_{t+1}(x,1,\beta)|\big| \\ &= \begin{cases} 0, & \text{if } \sum\limits_i \sum\limits_{x\in\mathcal{X}^i} |\bar{\mathcal{Y}}^i_{t+1}(x,1,\beta)| \le \beta m_{t+1}, \\ \sum\limits_i \sum\limits_{x\in\mathcal{X}^i} |\bar{\mathcal{Y}}^i_{t+1}(x,1,\beta)| - \beta m_{t+1}, & \text{otherwise,}\end{cases}\end{aligned} \qquad (41)$$

where the first case follows from (40) and the fact that every arm in the set $\bar{\mathcal{Y}}^i_{t+1}(x,1,\beta)$ will be activated by policy $\bar\pi$, the second case follows from (40) and the fact that policy $\bar\pi$ activates βm_{t+1} arms. According to (41), we have

$$\begin{aligned}&\sum_i \sum_{x\in\mathcal{X}^i} \big||\bar{\mathcal{Y}}^i_{t+1}(x,1,\beta)| - |\bar{\mathcal{N}}^i_{t+1}(x,1,\beta)|\big| \\ \le &\Big|\sum_i \sum_{x\in\mathcal{X}^i} |\bar{\mathcal{Y}}^i_{t+1}(x,1,\beta)| - \beta \sum_i \sum_{x\in\mathcal{X}^i} \rho^{i*}_{t+1}(x,1)\Big| \\ \le &\sum_i \sum_{x\in\mathcal{X}^i} \big||\bar{\mathcal{Y}}^i_{t+1}(x,1,\beta)| - \beta\rho^{i*}_{t+1}(x,1)\big|,\end{aligned} \qquad (42)$$

where the first inequality holds due to the feasibility of $\rho^{i*}_{t+1}(x,1)$, *i.e.*, $\beta\sum_i\sum_{x\in\mathcal{X}^i}\rho^{i*}_{t+1}(x,1) \le \beta m_{t+1}$, and the second inequality follows from the triangle inequality. We have established the result in (39) for the case with $\lambda^*_{t+1} = 0$.

Step 3.2. We now show that (39) holds when $\lambda^*_{t+1} > 0$.

If $\sum_i\sum_{x\in\mathcal{X}^i}|\bar{\mathcal{Y}}^i_{t+1}(x,1,\beta)| \le \beta m_{t+1}$, then for any $\mathcal{G}^i(\beta)$, every arm in $\bar{\mathcal{Y}}^i_{t+1}(x,1,\beta)$ is activated by policy $\bar\pi$, *i.e.*, $\bar{\mathcal{Y}}^i_{t+1}(x,1,\beta) \subseteq \bar{\mathcal{N}}^i_{t+1}(x,1,\beta)$ for any $x \in \mathcal{X}^i$ and any set $\mathcal{G}^i(\beta)$. Since policy $\bar\pi$ activates no more than βm_{t+1} arms, we have

$$\begin{aligned}&\sum_i \sum_{x\in\mathcal{X}^i} \big||\bar{\mathcal{Y}}^i_{t+1}(x,1,\beta)| - |\bar{\mathcal{N}}^i_{t+1}(x,1,\beta)|\big| \\ &= \sum_i \sum_{x\in\mathcal{X}^i} |\bar{\mathcal{N}}^i_{t+1}(x,1,\beta)| - \sum_i \sum_{x\in\mathcal{X}^i} |\bar{\mathcal{Y}}^i_{t+1}(x,1,\beta)| \\ &\le \beta m_{t+1} - \sum_i \sum_{x\in\mathcal{X}^i} |\bar{\mathcal{Y}}^i_{t+1}(x,1,\beta)|.\end{aligned} \qquad (43)$$

If $\sum_i\sum_{x\in\mathcal{X}^i}|\bar{\mathcal{Y}}^i_{t+1}(x,1,\beta)| > \beta m_{t+1}$, (40) still holds and we obtain same result with the second case of (41).

For the case with $\lambda^*_{t+1} > 0$, we therefore have

$$\begin{aligned}&\sum_i \sum_{x\in\mathcal{X}^i} \big||\bar{\mathcal{Y}}^i_{t+1}(x,1,\beta)| - |\bar{\mathcal{N}}^i_{t+1}(x,1,\beta)|\big| \\ \le &\Big|\sum_i \sum_{x\in\mathcal{X}^i} |\bar{\mathcal{Y}}^i_{t+1}(x,1,\beta)| - \beta m_{t+1}\Big| \\ = &\Big|\sum_i \sum_{x\in\mathcal{X}^i} |\bar{\mathcal{Y}}^i_{t+1}(x,1,\beta)| - \beta \sum_i \sum_{x\in\mathcal{X}^i} \rho^{i*}_{t+1}(x,1)\Big| \\ \le &\sum_i \sum_{x\in\mathcal{X}^i} \big||\bar{\mathcal{Y}}^i_{t+1}(x,1,\beta)| - \beta\rho^{i*}_{t+1}(x,1)\big|,\end{aligned}$$

where the first inequality follows from (43) and the second case of (41), and the equality follows from the complementary slackness when $\lambda^*_{t+1} > 0$, *i.e.*, $\beta\sum_i\sum_{x\in\mathcal{X}^i}\rho^{i*}_{t+1}(x,1) = \beta m_{t+1}$. We have obtained the result in (39) when $\lambda^*_{t+1} > 0$.

We conclude that (39) holds.

Step 4. In the final step, we establish the desired results of Lemma 1 at $t+1$. By (38), for any $\mathcal{G}^i(\beta)$ and $x \in \mathcal{X}^i$, the variance of $|\bar{\mathcal{Y}}^i_{t+1}(x,1,\beta)|$ is no more than its mean [40], *i.e.*,

$$\sigma^2\big[|\bar{\mathcal{Y}}^i_{t+1}(x,1,\beta)|\big] \le \mathbb{E}\big[|\bar{\mathcal{Y}}^i_{t+1}(x,1,\beta)|\big],\ \forall\, x \in \mathcal{X}^i,\ i. \quad (44)$$

We then have

$$\begin{aligned}&\sum_i \sum_{x\in\mathcal{X}^i} \mathbb{E}^{\bar\pi}\Big[\big||\bar{\mathcal{Y}}^i_{t+1}(x,1,\beta)| - |\bar{\mathcal{N}}^i_{t+1}(x,1,\beta)|\big|\Big] \\ \le &\sum_i \sum_{x\in\mathcal{X}^i} \mathbb{E}^{\bar\pi}\Big[\big||\bar{\mathcal{Y}}^i_{t+1}(x,1,\beta)| - \beta\rho^{i*}_{t+1}(x,1)\big|\Big] \\ \le &\sum_i \sum_{x\in\mathcal{X}^i} \sqrt{\mathbb{E}^{\bar\pi}\Big[\big(|\bar{\mathcal{Y}}^i_{t+1}(x,1,\beta)| - \beta\rho^{i*}_{t+1}(x,1)\big)^2\Big]} \\ \le &\sum_i \sum_{x\in\mathcal{X}^i} \sqrt{\mathbb{E}^{\bar\pi}[|\bar{\mathcal{Y}}^i_{t+1}(x,1,\beta)|]} \\ &+ \sum_i \sum_{x\in\mathcal{X}^i} \Big|\mathbb{E}^{\bar\pi}[|\bar{\mathcal{Y}}^i_{t+1}(x,1,\beta)|] - \beta\rho^{i*}_{t+1}(x,1)\Big|,\end{aligned} \qquad (45)$$

where the first inequality follows from (39), the second inequality follows from the Cauchy-Schwarz inequality

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2021.3049340, IEEE Transactions on Automatic Control

($\mathbb{E}\big[|X|\big] \leq \sqrt{\mathbb{E}[X^2]}$), and the last inequality follows from (44) and the concavity of the square root function.

We then have

$$\begin{aligned}
&\lim_{\beta\to\infty} \frac{\sum_i \sum_{x\in\mathcal{X}^i} \mathbb{E}^{\bar{\pi}}\Big[\big||\bar{\mathcal{Y}}^i_{t+1}(x,1,\beta)| - |\bar{\mathcal{N}}^i_{t+1}(x,1,\beta)|\big|\Big]}{\beta} \\
\leq &\lim_{\beta\to\infty} \frac{\sum_i \sum_{x\in\mathcal{X}^i} \sqrt{\mathbb{E}^{\bar{\pi}}[|\bar{\mathcal{Y}}^i_{t+1}(x,1,\beta)|]}}{\beta} \\
&\quad + \lim_{\beta\to\infty} \frac{\sum_i \sum_{x\in\mathcal{X}^i} \Big|\mathbb{E}^{\bar{\pi}}\big[|\bar{\mathcal{Y}}^i_{t+1}(x,1,\beta)|\big] - \beta\rho^{i*}_{t+1}(x,1)\Big|}{\beta} \\
= &\, 0,
\end{aligned} \tag{46}$$

where the inequality follows from (45), and the equality follows from (35a).

By the dominant convergence theorem, Eq. (46) implies that

$$\lim_{\beta\to\infty} \frac{\mathbb{E}^{\bar{\pi}}\Big[\big||\bar{\mathcal{Y}}^i_{t+1}(x,1,\beta)| - |\bar{\mathcal{N}}^i_{t+1}(x,1,\beta)|\big|\Big]}{\beta} = 0, \forall x \in \mathcal{X}^i, i. \tag{47}$$

Combining (47) and (35a), we obtain

$$\lim_{\beta\to\infty} \frac{\mathbb{E}^{\bar{\pi}}\big[|\bar{\mathcal{N}}^i_{t+1}(x,1,\beta)|\big]}{\beta} = \rho^{i*}_{t+1}(x,1),\ \forall\, x \in \mathcal{X}^i, i, \tag{48}$$

which is the desired result in (23a). It is straightforward to check that (23b) holds by (32) and (48).

We have proved that Lemma 1 holds at stage $t+1$. By induction, Lemma 1 holds for any stage t.

Appendix D
Proof of Theorem 3

Given a scaling parameter β, the gap between the expected total rewards achieved in the Lagrangian dual problem and by the proposed policy $\bar{\pi}$ is

$$\begin{aligned}
&\beta L^*_1(\boldsymbol{s}_1, \boldsymbol{\lambda}^*_{1,T}) - J^{\bar{\pi}}_1[\boldsymbol{s}_1(\beta), \beta] \\
= &\sum_t \sum_i \sum_{x\in\mathcal{X}^i} \beta\big[\rho^{i*}_t(x,1) r^i_t(x,1) + \rho^{i*}_t(x,0) r^i_t(x,0)\big] \\
&\quad - \mathbb{E}^{\bar{\pi}}\Big\{\sum_t \sum_i \sum_{x\in\mathcal{X}^i} \big(|\bar{\mathcal{N}}^i_t(x,1,\beta)| r^i_t(x,1) \\
&\quad + |\bar{\mathcal{N}}^i_t(x,0,\beta)| r^i_t(x,0)\big)\Big\} \\
\leq &\sum_t \sum_i \sum_{x\in\mathcal{X}^i} C_r \Big|\beta\rho^{i*}_t(x,1) - \mathbb{E}^{\bar{\pi}}\big[|\bar{\mathcal{N}}^i_t(x,1,\beta)|\big]\Big| \\
&\quad + \sum_t \sum_i \sum_{x\in\mathcal{X}^i} C_r \Big|\beta\rho^{i*}_t(x,0) - \mathbb{E}^{\bar{\pi}}\big[|\bar{\mathcal{N}}^i_t(x,0,\beta)|\big]\Big|,
\end{aligned} \tag{49}$$

where the equality is from (15) and the re-enumeration of the rewards collected by policy $\bar{\pi}$, and C_r is the upper bound on the reward (cf. (8)).

It is straightforward to check that Theorem 3 follows from Lemma 1 and (49).

References

[1] L. Hao and Y. Xu, "Index policies for stochastic deadline scheduling with time-varying processing rate limits," *submitted to ACC2020*, 2019.

[2] T. Zhang, W. Chen, Z. Han, and Z. Cao, "Charging scheduling of electric vehicles with local renewable energy under uncertain electric vehicle arrival and grid power price," *IEEE Trans. Veh. Technol.*, vol. 63, no. 6, pp. 2600–2612, 2014.

[3] Q. Huang, Q. S. Jia, Z. Qiu, X. Guan, and G. Deconinck, "Matching EV charging load with uncertain wind: A simulation-based policy improvement approach," *IEEE Trans. Smart Grid*, vol. 6, no. 3, pp. 1425–1433, 2015.

[4] Y. Xu, F. Pan, and L. Tong, "Dynamic scheduling for charging electric vehicles: A priority rule," *IEEE Trans. Autom. Control*, vol. 61, no. 12, pp. 4094–4099, 2016.

[5] Z. Yu, Y. Xu, and L. Tong, "Deadline scheduling as restless bandits," *IEEE Trans. Autom. Control*, vol. 63, no. 8, pp. 2343–2358, 2018.

[6] Y. Ji, L. Tong, T. He, J. Tan, K. won Lee, and L. Zhang, "Improving multi-job mapreduce scheduling in an opportunistic environment," in *IEEE Sixth International Conference on Cloud Computing*, 2013.

[7] J. Vilaplana, F. Solsona, I. Teixido, J. Mateo, F. Abella, and J. Rius, "A queuing theory model for cloud computing," *The Journal of Supercomputing*, vol. 69, no. 1, pp. 492–507, 2014.

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

[9] J. C. Gittins, "Bandit processes and dynamic allocation indices," *J. Roy. Stat. Soc. Ser. B (Methodological)*, vol. 41, no. 2, pp. 148–164, 1979.

[10] P. Whittle, "Restless bandits: Activity allocation in a changing world," *J. Appl. Probab.*, vol. 25, no. A, pp. 287–298, 1988.

[11] R. I. Davis and A. Burns, "A survey of hard real-time scheduling for multiprocessor systems," *ACM Comput. Surv. (CSUR)*, vol. 43, no. 4, p. 35, 2011.

[12] M. L. Dertouzos and A. K. Mok, "Multiprocessor online scheduling of hard-real-time tasks," *IEEE Trans. Softw. Eng.*, vol. 15, no. 12, pp. 1497–1506, 1989.

[13] A. Saifullah, J. Li, K. Agrawal, C. Lu, and C. Gill, "Multi-core real-time scheduling for generalized parallel task models," *Real-Time Systems*, vol. 49, no. 4, pp. 404–435, 2013.

[14] R. Singh and P. R. Kumar, "Decentralized throughput maximizing policies for deadline-constrained wireless networks," in *Proc. IEEE 54st Conf. Decision and Control (CDC), Osaka, Japan, Dec.*, 2015, pp. 3759–3766.

[15] V. Raghunathan, V. Borkar, M. Cao, and P. R. Kumar, "Index policies for real-time multicast scheduling for wireless broadcast systems," in *Proc. of IEEE INFOCOM*, 2008, pp. 1570–1578.

[16] D. Graczová and P. Jacko, "Generalized restless bandits and the knapsack problem for perishable inventories," *Oper. Res.*, vol. 62, no. 3, pp. 696–711, 2014.

[17] S. Chen, T. He, H. Y. S. Wong, K.-W. Lee, and L. Tong, "Secondary job scheduling in the cloud with deadlines," in *Proc. IEEE IPDPSW, Anchorage, AK, USA, May.*, 2011, pp. 1009–1016.

[18] V. T. Chakaravarthy, A. R. Choudhury, S. Gupta, S. Roy, and Y. Sabharwal, "Improved algorithms for resource allocation under varying capacity," in *European Symposium on Algorithms*. Springer, 2014, pp. 222–234.

[19] N. Meuleau, M. Hauskrecht, K.-E. Kim, L. Peshkin, L. P. Kaelbling, T. L. Dean, and C. Boutilier, "Solving very large weakly coupled markov decision processes," in *AAAI/IAAI*, 1998, pp. 165–172.

[20] S. Takriti and J. R. Birge, "Lagrangian solution techniques and bounds for loosely coupled mixed-integer stochastic programs," *Oper. Res.*, vol. 48, no. 1, pp. 91–98, 2000.

[21] J. T. Hawkins, "A Langrangian decomposition approach to weakly coupled dynamic optimization problems and its applications," Ph.D. dissertation, Massachusetts Institute of Technology, 2003.

[22] D. Adelman and A. J. Mersereau, "Relaxations of weakly coupled stochastic dynamic programs," *Oper. Res.*, vol. 56, no. 3, pp. 712–727, 2008.

[23] H. Everett III, "Generalized lagrange multiplier method for solving problems of optimum allocation of resources," *Oper. Res.*, vol. 11, no. 3, pp. 399–417, 1963.

[24] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*. Academic press, 1982.

[25] D. Bertsekas, G. Lauer, N. Sandell, and T. Posbergh, "Optimal short-term scheduling of large-scale power systems," *IEEE Trans. Autom. Control*, vol. 28, no. 1, pp. 1–11, 1983.

[26] F. Ye, H. Zhu, and E. Zhou, "Weakly coupled dynamic program: Information and Lagrangian relaxations," *IEEE Trans. Autom. Control*, vol. 63, no. 3, pp. 698–713, 2017.

[27] U. Ayesta, M. Erausquin, and P. Jacko, "Resource-sharing in a single server with time-varying capacity," in *49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2011, pp. 377–383.

[28] A. Lesage-Landry and J. A. Taylor, "The multi-armed bandit with stochastic plays," *IEEE Trans. Autom. Control*, vol. 63, no. 7, pp. 2280–2286, 2018.

[29] U. Ayesta, M. Erausquin, and P. Jacko, "A modeling framework for optimizing the flow-level scheduling with time-varying channels," *Performance Evaluation*, vol. 67, no. 11, pp. 1014–1029, 2010.

[30] J. A. Taylor and J. L. Mathieu, "Index policies for demand response," *IEEE Trans. Power Systems*, vol. 29, no. 3, pp. 1287–1295, 2014.

[31] S. Duran and I. M. Verloop, "Asymptotic optimal control of markov-modulated restless bandits," *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, vol. 2, no. 1, pp. 1–25, 2018.

[32] W. Hu and P. Frazier, "An asymptotically optimal index policy for finite-horizon restless bandits," *arXiv preprint arXiv:1707.00205*, 2017.

[33] S. M. Ismael, S. H. A. Aleem, A. Y. Abdelaziz, and A. F. Zobaa, "State-of-the-art of hosting capacity in modern power systems with distributed generation," *Renew Energ.*, 2018.

[34] E. Altman, *Constrained Markov Decision Processes*. CRC Press, 1999.

[35] C. Derman, *Finite State Markovian Decision Processes*. Academic Press: New York, 1970.

[36] U. Ayesta, M. Erausquin, M. Jonckheere, and I. Verloop, "Scheduling in a random environment: stability and asymptotic optimality," *IEEE/ACM Transactions on Networking*, vol. 21, no. 2, pp. 258–271, 2012.

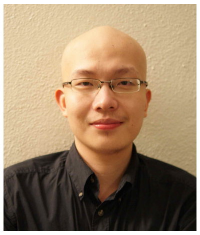
[37] C. H. Papadimitriou and J. N. Tsitsiklis, "The complexity of optimal queuing network control," *Mathematics of Operations Research*, vol. 24, no. 2, pp. 293–305, 1999.

[38] D. B. Brown and J. E. Smith, "Index policies and performance bounds for dynamic selection problems," Working paper, Tech. Rep., 2017.

[39] S. Kwon, Y. Xu, and N. Gautam, "Meeting inelastic demand in systems with storage and renewable sources," *IEEE Trans. Smart Grid*, vol. 8, no. 4, pp. 1619–1629, 2017.

[40] D. P. Bertsekas and J. N. Tsitsiklis, *Introduction to Probability*. Athena Scientific Belmont, MA, 2002, vol. 1.

Liangliang Hao received the B.S. degree in Electrical Engineering from Wuhan University, Wuhan, China, in 2016. He is currently pursuing the Ph.D. degree in the Mechanical and Automation Engineering at the Chinese University of Hong Kong (CUHK), Hong Kong SAR. His research interests include stochastic dynamic programming and its applications.



Yunjian Xu (S'06-M'10) received the B.S. and M.S. degrees in Electrical Engineering from Tsinghua University, Beijing, China, in 2006 and 2008, respectively, and the Ph.D. degree from the Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, in 2012.

Dr. Xu was a CMI (Center for the Mathematics of Information) postdoctoral fellow at the California Institute of Technology, Pasadena, CA, USA, in 2012-2013. Before joining the Chinese University of Hong Kong (CUHK) as an assistant professor, he was an assistant professor at the Singapore University of Technology and Design in 2013-2017. His research interests lie in stochastic optimal control, power system control and optimization, and mechanism design for electricity markets. Dr. Xu was a recipient of the MIT-Shell Energy Fellowship.

Lang Tong (F'05) is the Irwin and Joan Jacobs Professor in Engineering of Cornell University and the site director of Power Systems Engineering Research Center (PSERC). He received the B.E. degree from Tsinghua University in 1985, and M.S. and Ph.D. degrees in electrical engineering in 1987 and 1991, respectively, from the University of Notre Dame. He was a Postdoctoral Research Affiliate at the Information Systems Laboratory, Stanford University in 1991. He was the 2001 Cor Wit Visiting Professor at the Delft University of Technology and had held visiting positions at Stanford University and the University of California at Berkeley. Lang Tongs research is in the general area of statistical inference, communications, and complex networks. His current research focuses on inference, optimization, and economic problems in energy and power systems. He received the 1993 Outstanding Young Author Award from the IEEE Circuits and Systems Society, the 2004 best paper award from IEEE Signal Processing Society, and the 2004 Leonard G. Abraham Prize Paper Award from the IEEE Communications Society. He is also a coauthor of seven student paper awards. He received Young Investigator Award from the Office of Naval Research. He was a Distinguished Lecturer of the IEEE Signal Processing Society.