Online Fair Allocation of Reusable Resources

QINGSONG LIU, University of Massachusetts Amherst, USA
MOHAMMAD HAJIESMAILI, University of Massachusetts Amherst, USA

Motivated by the emerging paradigm of resource allocation that integrates classical objectives, such as cost minimization, with societal objectives, such as carbon awareness, this paper proposes a general framework for the online fair allocation of reusable resources. Within this framework, an online decision-maker seeks to allocate a finite resource with capacity C to a sequence of requests arriving with unknown distributions of types, utilities, and resource usage durations. To accommodate diverse objectives, the framework supports multiple actions and utility types, and the goal is to achieve max-min fairness among utilities, i.e., maximize the minimum time-averaged utility across all utility types. Our performance metric is an (α, β)-competitive guarantee of the form: $\text{ALG} \geq \alpha \cdot \text{OPT}^* - O(T^{\beta-1}),\ \alpha, \beta \in (0, 1]$, where OPT^* and ALG are the time-averaged optimum and objective value achieved by the decision maker, respectively. We propose a novel algorithm that achieves a competitive guarantee of $(1 - O(\sqrt{\log C/C}), 2/3)$ under the bandit feedback. As resource capacity increases, the multiplicative competitive ratio term $1 - O(\sqrt{\log C/C})$ asymptotically approaches optimality. Notably, when the resource capacity exceeds a certain threshold, our algorithm achieves an improved competitive guarantee of $(1, 2/3)$. Our algorithm employs an optimistic penalty-weight mechanism coupled with a dual exploration-discarding strategy to balance resource feasibility, exploration, and fairness among utilities.

CCS Concepts: • **Theory of computation → Online algorithms**.

Additional Key Words and Phrases: Resource allocation, Fairness, Online learning

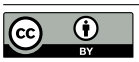
ACM Reference Format:
Qingsong Liu and Mohammad Hajiesmaili. 2025. Online Fair Allocation of Reusable Resources. *Proc. ACM Meas. Anal. Comput. Syst.* 9, 2, Article 29 (June 2025), 46 pages. https://doi.org/10.1145/3727121

1 Introduction

In a wide range of application domains, the online allocation of *reusable* resources is a pivotal design problem. Unlike non-reusable resources, whose availability decreases monotonically, the reusability of resources exhibits dynamic fluctuations in availability due to the interplay between allocation decisions and resource release events. Assigning resources reduces availability, while their return restores it. Moreover, resource availability depends on the order of decisions made, as different decisions can result in varying usage durations. Prior studies, reviewed in Section 2.1, have been proposed to address these complexities by designing algorithms that provide performance guarantees while adhering to resource constraints in both online and offline (or Bayesian) settings.

Most existing resource allocation algorithms are designed around traditional first-order metrics, such as throughput, operational cost, or energy consumption, and typically focus on optimizing a single objective function. However, as algorithmic solutions are increasingly deployed in real-world domains with significant societal and environmental impacts, it is crucial to incorporate emerging objectives, such as environmental metrics (e.g., carbon footprint, water usage, air pollution [24]) and broader societal considerations (e.g., safety and privacy [23]). These objectives, however, often conflict; e.g., achieving carbon reduction targets may necessitate higher energy consumption [25].

Authors' Contact Information: Qingsong Liu, University of Massachusetts Amherst, USA, qingsongliu@umass.edu; Mohammad Hajiesmaili, University of Massachusetts Amherst, USA, hajiesmaili@cs.umass.edu.



ACM 2476-1249/2025/6-ART29
https://doi.org/10.1145/3727121

Consequently, systems must balance trade-offs among competing performance metrics. Different notions of fairness are commonly used in prior literature to characterize these trade-offs in multi-objective design settings. The main prior studies (e.g., [8, 15, 30], see Section 2.2 for extended list), however, proposed fair allocation of non-reusable resources, with a few exceptions, e.g., [50], where fairness has been explored on reusable resource allocation. However, [50] operates within offline (or Bayesian) settings, where problem/system parameters, including utilities and usage durations, are drawn from known distributions.

In this work, our goal is to tackle the fair allocation of reusable resources in an online setting where the decision maker must operate without complete information on the underlying distributions of arrival types, utilities, or resource usage durations. In practice, many real-world applications like cloud resource management are inherently online, as the underlying distributions of system parameters are often unknown or only partially known. The online nature or model uncertainty impacts not only allocation efficiency (i.e., the uncertainty about which actions yield higher utilities per unit of resource occupation time) but also future resource availability. To make effective allocation decisions, an algorithm must balance exploring actions to learn the latent model and exploiting this knowledge to maximize cumulative utility. This trade-off is further complicated by the interaction between resource feasibility and fairness objectives, making designing online algorithms with provable performance guarantees particularly challenging.

A recent study [17] investigates a special version of online reusable resource allocation in the context of admission control (i.e., with only accept/reject decisions). Although this work marks a valuable step toward online reusable resource allocation, it does not consider fairness (focusing on only one utility type) and operates under restrictive assumptions, including only two possible actions, a single type of arrival request, and deterministic resource usage durations. Consequently, the model and algorithm in [17] fall short for applications such as LLM inference service provisioning, where user requests may vary significantly in characteristics, including query length, membership tier, and quality or latency requirements. Additionally, the service provider may employ multiple LLM models with varying performance profiles to serve these requests. In such a scenario, the service provider must fairly balance multiple objectives such as model accuracy, energy consumption, and carbon/water footprint. This fairness requirement, coupled with the stochastic nature of resource usage durations and the diversity of arriving request types, highlights the need for a generalized framework that can handle these complexities in an online setting.

To address these gaps, we study a general framework for fair allocation of reusable resources in the *online* setting. In this framework, requests arrive sequentially, each associated with a specific type capturing its features. Upon observing the request type, the controller (decision-maker) may choose an action from a finite set or reject the request (e.g., if resources are unavailable or the utility gain does not justify the associated resource consumption). Each action yields multiple utility types (reflecting different performance metrics) and incurs a stochastic resource usage duration, determining how long the resources remain occupied. Motivated by real-world observations on the independence of objective functions, e.g., for environmental metrics [24, 34], our model supports multiple general utility functions that might be arbitrarily correlated with the resource usage durations. The controller must respect a capacity constraint, ensuring total occupied resource units never exceed C at any time. Critically, the distributions of arrival types, utilities, and usage durations are unknown, and feedback is limited to the chosen action's outcome, conforming to a bandit-feedback setup. To balance multi-type utilities, we adopt the notion of max-min fairness (also referred to as egalitarian welfare in operations research [8]): we aim to maximize the minimum time-averaged utility across all utility types over the decision horizon, ensuring that no single objective is disproportionately sacrificed. We note that the max-min fairness is one of the most widely adopted fairness notions and has been extensively studied for fair allocation of non-reusable resources [8, 30].

Moreover, this notion has been used in addressing the environmental concerns of AI and computing applications. For example, [34] employs max-min fairness to simultaneously balance water usage, carbon footprint, and energy cost objectives, thereby addressing AI's environmental inequity.

Contributions. In this paper, and in Section 3, we present a general framework that unifies online reusable resource allocation and fairness optimization. Our formulation accommodates multiple actions, addresses model uncertainty (including stochastic usage times with unknown distributions), and supports diverse utility types (e.g., varied performance metrics) as well as multiple resource units. We also illustrate representative applications of this framework, such as LLM inference service provisioning and healthcare management.

► *Algorithm design.* In Section 4, we present a novel algorithm called Exploration-Discarding with Penalty Weights Update (ED-PWU). To balance the multi-type utilities and (allocation) efficiency, ED-PWU bases the decisions on penalty weights and optimistic estimates of unknown parameters. The idea of selecting actions based on penalty weights is similar to the approach taken in [15, 16, 50] for non-reusable or Bayesian settings, albeit with a distinct weight update process due to model uncertainty. Despite the benefits of these penalty-weight-based decisions, we cannot directly treat them as final due to several inherent challenges: (1) these decisions may not respect the resource capacity constraints under all sample paths of the underlying randomness. Moreover, estimation errors can lead to violations even in expectation; (2) the bandit feedback setup necessitates reserving part of the resource for exploration, so as to gradually reduce estimation errors in the latent model. To address these challenges, we design a dual exploration–discarding strategy (hence the algorithm's name) that probabilistically discards penalty-weight-based decisions (thereby preserving sufficient resource "slack" to mitigate the risk of constraint violations) while simultaneously dedicating part of the resources to forced exploration. The discarding and forced-exploration probabilities are carefully designed and adapt to the magnitude of estimation errors, striking a balance between maintaining resource feasibility (or reducing the likelihood of resource availability violations), ensuring sufficient exploration, and reducing utility loss induced by discarding and exploration.

► *Competitive analysis.* To analyze the performance of ED-PWU, we adopt a joint competitive and regret analysis approach. Denote OPT^* and ALG as the time-averaged optimum and objective value achieved by the controller. We seek a policy that satisfies an (α, β)-competitive guarantee:

$$\mathrm{ALG} \geq \alpha \cdot \mathrm{OPT}^* - O\big(T^{\beta-1}\big), \tag{1}$$

for some parameters $\alpha, \beta \in (0, 1]$. Here, α represents the classic notion of the competitive ratio of ALG. The additional error term $O\big(T^{\beta-1}\big)$ is closer to the definition of regret, i.e., if the policy ensures $\alpha = 1$, we can say it also achieves a sublinear regret guarantee of $O(T^{\beta})$. Using the above bi-criteria competitive ratio, we show that ED-PWU achieves a competitive guarantee of $((1 - \frac{1}{2C})\mathcal{L}(\epsilon^*(C)), 2/3)$ under the bandit-feedback setup, where C denotes the resource inventory (or capacity), $\varepsilon^*(C) = \arg\max_{\varepsilon>0} \mathcal{L}(\varepsilon)$, and $\mathcal{L}(\varepsilon) = \frac{1}{\varepsilon+1} \cdot (1 - (1+\varepsilon)\exp[C(\frac{\varepsilon}{1+\varepsilon} - \log(1+\varepsilon))])^+$. The competitive ratio is of the order $\alpha = 1 - O(\sqrt{\log C/C})$, which asymptotically approaches optimality, i.e., $\alpha = 1$, as capacity C increases. Notably, when the inventory exceeds a certain threshold, our model reduces to the indivisible variant of the horizon-fairness optimization problem in the online setting [12, 42]. In this scenario, our algorithm can further achieve an $(1, 2/3)$ competitive guarantee (i.e., sublinear regret of $O(T^{2/3})$). This result fills an important gap in the literature, as no prior work has achieved provable performance guarantees under the bandit feedback setup.

► *Extensions to Quasi-full-feedback setting.* Additionally, we extend our analysis (with the same algorithm) to a Quasi-full-feedback setup consistent with [17], which studies a special case of our model. In this feedback setup, the controller receives full information when the resources are available but none otherwise. This extension improves the competitive guarantee of our algorithm

to $((1-\frac{1}{2C})\mathcal{L}(\epsilon^*(C)), 1/2)$, and even to $(1, 1/2)$ when the inventory exceeds a certain threshold. In contrast, [17] achieves a $(1/2, 1/2)$-competitive guarantee only when the inventory C is limited to 1.

► *Numerical experiments.* Last, we conduct simulations in a cloud computing scenario to validate the theoretical performance of our algorithm (details provided in Appendix B).

2 Related Works

In this section, we introduce two streams of literature related to our research: reusable resource allocation and horizon fairness optimization. We do not delve into the literature on non-reusable resource allocation problems (e.g., [1, 6, 16]), as the fundamental differences between non-reusable and reusable resource cases necessitate distinct methodologies and analyses. We position our research as a bridge between these two lines of literature by comparing closely related existing works. Tables 1 and 2 summarize the key results from each of these two streams.

2.1 Reusable Resource Allocation

The problem of reusable resource allocation has been studied in various contexts, including admission control, pricing, assortment planning, and queueing systems. We review literature in offline (or Bayesian) and online settings, where the primary distinction lies in the presence of model uncertainty. In the Bayesian setting, while exact realizations of arrivals, utilities, or usage durations may be unknown at the time of decision, their underlying distributions are time-invariant and known to the decision maker. Conversely, the online setting lacks such prior knowledge, requiring real-time learning of unknown distributions. Most existing studies focus on the Bayesian setting, while the online setting remains less well understood compared to traditional (non-reusable) resource allocation problems. This is largely attributed to the complex interplay between model uncertainty and resource reusability. Depending on whether the arrival process of requests is determined adversarially or stochastically, existing works in the online setting can be further categorized into adversarial and stochastic cases. Our research falls in the category of online stochastic setting.

Table 1. Closely related literature on reusable resource allocation

Paper	fairness	stoch. or adv.	online	random usage duration	competitive ratio
[17]	✗	stochastic	✔	✗	$1/2^*$
[49]	✗	stochastic	✔	✔	$1/2$
[19]	✗	stochastic	✗	✔	$1-\min\left\{\frac{1}{2}, O\left(\sqrt{\log C/C}\right)\right\}$
[50]	✔	stochastic	✗	✔	$1-O\left(\sqrt{\xi\log(1/\xi)}\right)^\dagger$
[26, 27, 32]	✗	adversarial	✗	✗	(instance-dependent) constants
[18, 20–22]	✗	adversarial	✗	✔	(instance-dependent) constants
Ours	✔	stochastic	✔	✔	$1-O\left(\sqrt{\log C/C}\right)$ 1 (large inventory)

* This result only holds for the case of $C = 1$.

† ξ is an instance-dependent parameter that scales proportionally to $1/C$.

Online setting. In the online setting, there is relatively little research conducted on reusable resource allocation for the stochastic case. Some exceptions, e.g., [28, 29, 52], explore the problem in the context of queuing systems, where reusable resources are treated as servers and requests/customers are modeled as jobs arriving sequentially via a stationary Poisson process. Notably, in queueing systems, jobs can often wait for service, whereas our model resembles a loss system where requests are lost if no idle server is available. The work most closely related to ours is [17], which focuses on

an admission control setting with unknown utility distributions but deterministic usage durations. They propose a dynamic threshold rule inspired by an infinite-dimensional linear programming reformulation of the original problem, achieving a competitive ratio of 1/2 when the inventory is limited to a single unit. However, their work leaves several intriguing open questions, as noted in their concluding remarks. One such question is the extension to stochastic usage times with unknown distributions, potentially correlated with utilities. Another is whether an efficient algorithm (possibly threshold-based) with provable competitive guarantees exists for managing multiple identical reusable resources. Our work addresses these questions by considering a generalized model that includes multiple actions, stochastic usage durations with unknown distributions, multiple identical reusable resources, and diverse utility types. We propose an algorithm leveraging penalty weights to balance the trade-off between the utilities earned across all types and the resources consumed, which simplifies to a threshold-based policy in the context of admission control. Additionally, a recent study, [49], explores a combination of pricing and admission control, where the decision-maker determines not only acceptance but also the rental price for each arrival. Their model accommodates unknown distributions in both utility and usage times while allowing multiple resource units. The algorithm they propose employs linear function approximation based on Markov Decision Process (MDP) methodology and achieves a competitive ratio of 1/2. Although our methodologies and models differ significantly and are not directly comparable, our algorithm ensures that the competitive ratio approaches 1 as the inventory C increases.

There is another line of research focusing on the adversarial case. Specifically, recent studies in assortment planning [18, 20–22] consider scenarios where each customer type is associated with a unique choice model over the resources. These works assume that the usage duration of a resource and the price paid (can be interpreted as utility in our model) depend only on the resource type, independent of customer type. For example, [20] demonstrates that the myopic policy achieves a competitive ratio of 1/2, while [21] and [22] obtain a $1 - 1/e$ competitive ratio based on the fluid approximation guided algorithms. [18] incorporates exogenous inventory replenishment into assortment planning and designs an inventory-balancing algorithm with a constant competitive ratio. Additionally, [26, 27] consider scenarios where rewards are linearly related to usage duration and usage duration is deterministic, resulting in instance-dependent competitive ratios. Beyond assortment planning, [32] explores network pricing models with advance reservations but assumes deterministic resource usage durations. In contrast to these works, our algorithm is designed for the stochastic setting, where the competitive ratio improves as the inventory C increases. This behavior differs from the adversarial setting, where the competitive ratio remains constant regardless of the inventory size. Additionally, our model allows both the usage duration and price to depend on the customer type, and we extend our framework to address fairness considerations, which are unexplored in the aforementioned works.

Bayesian or offline setting. Within this setting, [41] investigates an assortment planning problem and proposes a policy based on affine approximations, achieving a competitive ratio of at least 1/2. Similarly, [5] proposes a policy with competitive ratios that depend on request sizes. In [19], a dynamic programming-based policy guided by linear programming is developed to achieve a competitive ratio of $1-\min\{1/2, O(\sqrt{\log C/C})\}$. However, their model assumes that usage durations are independent of request types, a feature explicitly addressed in our work. The most relevant work to ours in the Bayesian setting is [50], which also explores fairness. They attain a comparable competitive ratio under the restrictive assumption that utilities exhibit a linear dependence on resource usage durations. [10] studies the reusable resource allocation in the pricing context and prove that a well-chosen static pricing policy guarantees 78.9% of the optimum. Broadly, these works mainly rely on dynamic programming to carefully allocate their reusable resources due

Table 2. Closely related literature on online fair resource allocation

Paper[†]	fairness type	stoch. or adv.	div. or indiv.	Online	Competitive guarantee
[15, 16]	max-min fairness	stochastic	indivisible	✗	$(1, 1/2)$
[8]	Nash social welfare	stochastic	indivisible	✗	$(1, 0)$
[30]	max-min fairness	adversarial	indivisible	✔	$(1/N, 1/2)$[§]
[9]	proportional fairness	adversarial	indivisible	✔	$(1/\log N, 0)$[§]
[35]	proportional fairness	adversarial	divisible	✔	$(1, \log(\sqrt{TP_T})/\log T)$[*]
[4, 42]	general α-fairness	adversarial	divisible	✔	$(1, \log(\sqrt{TP_T})/\log T)$[*]
[45]	general α-fairness	adversarial	indivisible	✔	$(e^{-\frac{1}{e}}, \frac{1-2\alpha}{2})$ if $\alpha < \frac{1}{2}$ $(e^{-\frac{1}{e}}, 0)$ if $\alpha \geq \frac{1}{2}$
[7]	general α-fairness	stochastic	divisible	✔	$(1, 1/2)$
[12]	general α-fairness	stochastic	indivisible	✔ (Full feedback)	$(1, 1/2)$
Ours	max-min fairness	stochastic	indivisible	✔ (Bandit feedback) ✔ (Quasi-full-feedback)	$(1, 2/3)$ $(1, 1/2)$

[†] It is worth noting that none of the existing works except ours consider resource capacity constraints.
[*] P_T quantifies the accumulated variations in the environment. Notably, $T^\beta = \sqrt{TP_T}$ if $\beta = \log(\sqrt{TP_T})/\log T$.
[§] N represents the number of agents, corresponding to the number of utility types in our model.

to the full knowledge of the underlying distributions of usage durations and utilities. Another line of research (e.g., [13, 32]) focuses on admission control or pricing problems within queueing systems. Compared to these works, the model uncertainty and bandit-feedback setup in our model introduce additional challenges, requiring a careful balance between maintaining resource feasibility (necessitating discarding), exploration (i.e., information acquisition), and minimizing utility loss caused by both discarding and exploration.

2.2 Fairness in Resource Allocation

Fairness is a critical metric in resource management and has been widely studied in topics such as computing systems [11, 48] and communication systems [2]. Independent of resource reusability, our research focuses specifically on horizon fairness in resource allocation [4, 42], which aims to ensure fairness across utilities accumulated over a time horizon. This contrasts with slot fairness problems [43, 46] where fairness is addressed independently in each decision round, without considering past or future decisions. Within this framework, our study aligns with the online setting and thus we do not delve into the literature (e.g., [8, 15, 16]) on offline (Bayesian) setting where utilities are assumed to be drawn from a fixed and known distribution.

Horizon fairness has recently been explored in adversarial cases, where the arrival of items may be controlled by an adversary. For instance, [30] studies scenarios with known utility at the time of decision and designs policies under the max-min fairness criterion. Similarly, [9, 35] investigate the problem using the proportional fairness criterion, allowing policies to leverage available predictions, while [42, 45] explores a more general α-fairness objective. The policies in [9, 30, 45] achieve constant competitive ratios compared to the optimum, whereas [35, 42] attain sublinear regret relative to a static benchmark (weaker than the optimum). Notably, [35, 42] tackle the divisible version of the problem using online convex optimization techniques, where allocation variables are continuous. In contrast, our work focuses on the indivisible variant, employing penalty weights based policy to guide allocation decisions.

For stochastic cases, [7] examines regularized resource allocation problems that encompass general α-fairness, achieving an $O(T^{1/2})$ regret when resources are divisible. Instead, [12] investigates the indivisible version of the problem and attains an $O(T^{1/2})$ regret by leveraging virtual queues, which can be interpreted as penalty weights, to balance utilities across different groups. While our algorithm shares similarities with theirs, it introduces an additional trade-off between information acquisition and fairness optimization, stemming from the bandit-feedback setup. This contrasts with [12] that assumes (delayed) full-feedback setup. Notably, their concluding remarks suggest extending their approach to the bandit-feedback setting as a direction for future work—a gap that our research effectively addresses. Additionally, another line of research (e.g., [33, 39, 44, 47]) focuses on using horizon-fairness as constraints rather than as objectives, making it fundamentally different and not directly comparable to our work.

3 System Model and Motivating Applications

This section presents the framework we have studied. Before specifying our problem formulation, we introduce below some notations that will be used. In this paper, vectors are generally bolded. We denote the n-dimensional all-ones vector as $\mathbf{1}_n$. For any vector $\boldsymbol{v}$, we define $v_{\min}$ and $v_{\max}$ as $\min_i v_i$ and $\max_i v_i$, respectively. Additionally, we use the notation $(v)^+$ to represent $\max\{v, 0\}$.

3.1 Problem formulation

We consider a controller that allocates some available resources to process the arriving requests (or customers) in an online manner. Time is divided into discrete slots, and at most one request arrives at each time t. The type of the request arriving at time t, denoted as $j(t)$, captures the characteristics or features of the request. The set of all possible request types is denoted by $\mathcal{J}$. The sequence $j(1), j(2), \ldots$ consists of independently and identically distributed (i.i.d.) random variables, governed by a fixed but unknown probability distribution $\boldsymbol{p} = \{p_j\}_{j\in\mathcal{J}}$, where $p_j = \mathbf{P}(j(t) = j)$. Importantly, the controller has no prior knowledge of the distribution $\boldsymbol{p}$.

At time t, upon observing the type $j(t)$ of the arriving request, the controller selects an action $k(t)$ from a finite action set $\mathcal{K}$, which could be used to model a broad range of decisions. For example, the controller can assign the request to different servers or channels. We also allow the controller to have the option to take no action (i.e., $k(t) = k_{\text{null}}$), which is equivalent to rejecting the request. Once $k(t)$ is selected, the controller receives multiple types of utilities $\{R_{i,j(t),k(t)}(t)\}_{i\in\mathcal{I}}$, where $R_{i,j(t),k(t)}(t)$ denotes the amount of type-i utility offered by action $k(t)$ at time t. Our multi-type utility framework allows the controller to evaluate the performance across various dimensions, such as energy consumption, social welfare, total revenue earned, etc.

Executing $k(t)$ also incurs one unit of resource usage for a stochastic duration $D_{j(t),k(t)}(t) \in \{1, 2, ..., d_{\max}\}$. Specifically, the resource is occupied from time t to time $t + D_{j(t),k(t)}(t) - 1$, and becomes available again at time $t + D_{j(t),k(t)}(t)$. The controller manages a limited pool of C resource units and rejects incoming requests if all resources are occupied. The stochastic outcomes $(R_{i,j,k}(t), D_{j,k}(t))$ for any pair $(i, j, k) \in \mathcal{I} \times \mathcal{J} \times \mathcal{K}$ follow a joint distribution $\boldsymbol{O}_{i,j,k}$, i.e., $(R_{i,j,k}(t), D_{j,k}(t)) \sim \boldsymbol{O}_{i,j,k}$. We denote $r_{i,j,k} = \mathbf{E}[R_{i,j,k}(t)]$ and $d_{j,k} = \mathbf{E}[D_{j,k}(t)]$. Here we note that rejecting a request ($k(t) = k_{\text{null}}$) results in zero utility and no resource usage. Additionally, the stochastic variables $\{R_{1,j,k}(t), .., R_{\mathcal{I},j,k}(t), D_{j,k}(t)\}$ can be arbitrarily correlated. In this work, we consider the online setting where the statistics about the $\boldsymbol{O}_{i,j,k}$ are unknown to the controller.

Feedback model. We consider the bandit-feedback setup for the controller. Specifically, at each time t, the controller observes only the type $j(t)$ of the arriving request, and the stochastic outcomes/realizations $\{D_{j(t),k(t)}(t), R_{1,j(t),k(t)}(t), ..., R_{\mathcal{I},j(t),k(t)}\}$ associated with the selected action $k(t)$ and type $j(t)$. The outcomes of the other actions will not be revealed to the controller. Also, the controller cannot observe future arrivals or their corresponding utilities and resource usage

durations. Thus, the choice of $k(t)$ is based solely on (1) the observed type $j(t)$, (2) the historical observations $\mathcal{H}(t-1) = \{j(t), D_{j(t),k(t)}(t), \boldsymbol{R}_{j(t),k(t)}(t)\}_{\tau=1}^{t-1}$ from time 1 to $t-1$, and (3) the internal randomness of the controller, i.e., the developed algorithm should be a non-anticipatory policy. It is worth noting that our feedback setup is more restrictive compared to [17]. In their setting, the controller receives full information about the utility and resource usage outcomes when the resource is available, even though their focus is limited to a two-action scenario (i.e., admission control). When the resource is unavailable, requests are rejected and no feedback is provided—consistent with our feedback model. Consequently, their controller is able to collect a linear number of samples, i.e., $O(t)$ samples for each unknown parameter. In contrast, our setup lacks such property which makes our model more challenging. We relax our feedback setup in Section 6 to align with theirs, demonstrating an improved performance guarantee for our algorithm.

Objective. The objective of the controller is to enforce fairness on the time-averaged aggregate utilities across all types over an *unknown* horizon T, while adhering to the resource capacity constraint at any time:

$$\max_{\{k(t)\}_{t\in[T]}} \min_{i\in\mathcal{I}} \mathbf{E}\left[\frac{1}{T}\sum_{t=1}^{T} R_{i,j(t),k(t)}(t)\right], \qquad \sum_{\tau=1}^{t} \mathbf{I}\left\{D_{j(\tau),k(\tau)}(\tau) \geq t-\tau+1\right\} \leq C,\ \forall t \in [T].$$

Here, we adopt max–min fairness as our fairness notion, which has been extensively studied in the problems of fair allocation [8, 15, 30]. Its appeal lies in its ability to ensure a uniform minimal utility guarantee across all metrics or criteria, offering a stronger fairness assurance compared to other fairness notions like Nash social welfare and proportional fairness. Our fairness objective raises novel technical challenges and subsumes total utility maximization as a special case. In our paper, each utility type is normalized to lie within the interval $[-1, 1]$. This normalization ensures that utility types—potentially defined on different scales—become comparable under our max–min fairness framework. Actually, our algorithm and analysis naturally extend to a scalarized version of the objective: $\min_{i\in\mathcal{I}} \mathbf{E}\left[\frac{1}{T}\sum_{t=1}^{T} w_i \cdot R_{i,j(t),k(t)}(t)\right]$, where w_i represents a user-defined weight for utility type i. This scalarization accommodates varying operational priorities across different performance metrics, allowing practitioners to place more or less emphasis on any particular metric as needed. For convenience, let $Y_k(t) \in \{0, 1\}$ indicate whether action k is taken at time t. Then, the online fair allocation problem with reusable resources (OFARR) can be reformulated as the following online binary integer program,

$$\begin{aligned}
(\texttt{OFARR}) \quad & \max_{\{Y(t)\}_{t\in[T]}} \min_{i\in\mathcal{I}} \mathbf{E}\left[\frac{1}{T}\sum_{t=1}^{T}\sum_{k\in\mathcal{K}} R_{i,j(t),k}(t)\cdot Y_k(t)\right] \\
& \text{s.t.} \sum_{\tau=1}^{t}\sum_{k\in\mathcal{K}} Y_k(\tau)\cdot \mathbf{I}\{D_{j(\tau),k}(\tau) \geq t-\tau+1\} \leq C,\ \forall t\in[T], \\
& \quad \sum_{k\in\mathcal{K}} Y_k(t) \leq 1,\ \forall t\in[T]; \quad Y_k(t)\in\{0,1\},\ \forall k\in\mathcal{K} \text{ and } t\in[T].
\end{aligned}$$

Here, we let $\sum_{k\in\mathcal{K}} Y_k(t) \leq 1$ since the controller could take the null action (reject the request). We denote the optimal objective value of (OFARR) under a *non-anticipatory* policy as OPT^*, where the policy has full knowledge of the underlying distributions but no access to the actual realizations of arrivals, durations, or utilities in advance. OPT^* is also referred to as the *clairvoyant optimum*, a term commonly used in the literature [7, 19]. We remark that although we consider the stochastic setting, the term $\mathbf{I}\{D(\tau) \geq t-\tau+1\}$ induces non-stationarity into the decision-making process. Specifically, even if the controller selects the same action k at two different times $\tau_1 < \tau_2$, the corresponding resource usage-characterized by $\mathbf{I}\{D(\tau_1) \geq t-\tau_1+1\}$ and $\mathbf{I}\{D(\tau_2) \geq t-\tau_2+1)$-follows different distributions at time t. This behavior contrasts sharply with the existing literature on stochastic (non-reusable) resource allocation with knapsacks that crucially rely on model stationarity. In general,

(OFARR) is intractable due to the unknown parameter distributions, non-stationarity induced by the resource reusability, and the curse of dimensionality. Thus, our goal is to achieve near-optimality. More precisely, let ALG represent the time-averaged objective value achieved by the controller and our goal is to design a policy for the controller that satisfies the competitive guarantee defined in (1). Notably, even with full model certainty, achieving a sublinear regret guarantee (i.e., $\alpha = 1$ in (1)) is generally not possible. This was demonstrated by [17], who studied a special case of our problem with $|\mathcal{I}| = |\mathcal{J}| = 1$, $|\mathcal{K}| = 2$, and $C = 1$.

3.2 Motivating examples

Before proceeding to our algorithm development, we highlight the generality of our framework by discussing some of its applications, including large language model (LLM) inference service provisioning and healthcare management.

LLM inference service provisioning. Consider a service provider offering LLM inference services [14, 51] for users' requests. These requests correspond to inference queries, which may include text-generation tasks, code-completion queries, or conversational instructions. The request type $j(t)$ can encode specific attributes such as the level of the query token length, user membership tier (e.g., free vs. paid users), or desired response quality and latency. The action $k(t) \in \mathcal{K}$ represents different inference optimization settings (e.g., quantization, model pruning) or distinct LLM models. Naturally, different LLM models and inference optimization settings yield different inference accuracy, monetary cost, and latency. After serving the request, the service provider receives multi-type utilities $\{R_{i,j(t),k(t)}(t)\}_i$, capturing performance metrics along multiple dimensions such as: (a) user satisfaction or service-level agreement (SLA) fulfillment, e.g., whether the request is answered with sufficient quality; (b) energy consumption or operational costs, e.g., using an LLM model with larger weights can provide more accurate and higher-quality responses but incurs higher energy usage; (c) revenue (giving priority to paid subscribers can increase profit). The resource capacity C reflects the maximum number of requests that can be processed in parallel (i.e., maximum service capacity [34]), often limited by GPU memory or the physical number of GPU nodes. If all resources are occupied, newly arriving requests must be rejected. Additionally, if the service provider determines that the benefit of serving the request is insufficient to justify resource usage (e.g., during peak request periods for free users), he can choose to reject the request. Each accepted request $j(t)$ under action $k(t)$ occupies the allocated resource for a stochastic duration $D_{j(t),k(t)}(t)$, indicating how long the GPU remains locked for that specific inference task. In practice, the service provider seeks to maximize a max-min fairness objective among different utility types—e.g., ensuring that user satisfaction, energy efficiency, and revenue are balanced, without letting any single metric degrade excessively. Meanwhile, the service provider must respect the service capacity constraints and thus the resulting problem fits into our framework.

Healthcare management. Our framework could be applied to other domains such as healthcare resource management. In this context, requests correspond to patients seeking medical attention. The patient type $j(t)$ captures patient characteristics, such as urgency level and medical condition. Actions $k(t) \in \mathcal{K}$ represent assigning patients to specific healthcare providers, such as nurses or doctors. Alternatively, $k(t) = k_{\text{null}}$ may indicate redirecting patients to other hospitals or outpatient departments to alleviate crowding. The utilities $\{R_{i,j,k}(t)\}_i$ could reflect patient satisfaction, recovery speed, or treatment effectiveness for different medical objectives. For example, if i denotes a reliability objective, $r_{i,j,k}$ should be higher for more urgent patients when j encodes the acuity level of the patient. The stochastic duration $D_{j,k}(t)$ reflects the expected treatment time required for the assigned healthcare provider to complete the patient's care. The model incorporates capacity constraints on healthcare resources, such as hospital beds, treatment rooms, or medical

equipment. The controller must balance resource usage and patient outcomes, and when resources are unavailable, he can mitigate overload by redirecting patients to other hospitals.

4 Algorithm

In this section, we first introduce preliminaries that facilitate our algorithm design in Section 4.

4.1 Preliminaries

We note that directly tackling (OFARR) is challenging, even when the distributions of problem parameters are known in advance. This difficulty arises from the stochastic variations in the outcomes and the unknown value of T. To mitigate this difficulty, we introduce a "steady-state" benchmark of (OFARR), denoted as (OFARR-S), which serves as the building block for our algorithm:

$$\text{(OFARR-S)} \quad \max\ \lambda \qquad \text{s.t.} \quad \begin{cases} \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} p_j \cdot r_{i,j,k} \cdot x_{j,k} \geq \lambda,\ \forall i \in \mathcal{I}, \\ \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} p_j \cdot d_j \cdot x_{j,k} \leq C, \\ \sum_{k\in\mathcal{K}} x_{j,k} \leq 1,\ \forall j \in \mathcal{J}, \quad x_{j,k} \geq 0,\ \forall k \in \mathcal{K},\ j \in \mathcal{J}. \end{cases}$$

Here $(\boldsymbol{p}, \boldsymbol{r}, \boldsymbol{d})$ represents the ground truth of arrival probabilities, utilities, and resource usage durations. The variable $x_{j,k}$ denotes the fraction of type-j requests allocated to action k, making (OFARR-S) a fluid approximation of (OFARR). Notably, [19, 50] employs a relaxed version of the original optimization problem as the foundation for their algorithms. When applied to our model, this relaxed version, (OFARR-R), assumes that the realizations of request arrivals and their resource usage durations under all possible actions align exactly with their underlying distributions (as detailed in (17) of the Appendix), enabling the careful allocation of reusable resources. However, this approach is not directly applicable to our (online) setting due to two key challenges: the unknown time horizon T, and the bandit feedback which prevents accurate estimation of the distribution of usage durations. In contrast, (OFARR-S) offers a more lightweight alternative, relying solely on the expectations of the unknown parameters and remaining independent of T, making it a suitable building block for our algorithm design. Denote the optimal objective values of (OFARR), (OFARR-R), and (OFARR-S) as OPT^*, OPT^{R}, and λ^*, respectively. The following lemma establishes that λ^* incurs only a bounded optimality gap compared to both OPT^* and OPT^{R}.

Lemma 1. *The inequalities hold surely that* $T \cdot OPT^* \leq T \cdot OPT^R \leq T \cdot \lambda^* + \mathbf{I}\{C < d_{\max}\} \cdot d_{\max} \cdot r_{\max}$.

The indicator term $\mathbf{I}\{C < d_{\max}\}$ arises from the observation that when $C \geq d_{\max}$, i.e., resources are abundant, the capacity constraints in both (OFARR-R) and (OFARR-S) can be safely ignored, as they are satisfied for all t. Consequently, we can deduce that the optimal objective values of these two problems coincide. Lemma 1 shows that λ^* can be interpreted as the target value for the optimal utility rate of each type. In our algorithm design, we estimate λ^* using past observations, which then used to adjust the penalty weights associated with the utilities.

4.2 Algorithm design

The proposed algorithm operates across multiple epochs and incorporates a multiplicative weights update (MWU) process. The MWU process iteratively generates penalty weights to effectively balance different types of utilities earned, resource units consumed as well as usage durations. More specifically, our algorithm divides the time horizon T into multiple epochs $n = 0, 1, \ldots$, where each epoch n consists of ℓ_n time steps. We initialize with $\ell_0 = d_{\max}$ and double ℓ_n for each subsequent epoch $n \geq 1$. We refer to t_n as the ending timestep of epoch n. Epoch 0 serves as a warm-up phase. For epochs $n \geq 1$, the algorithm comprises three key procedures: In procedure ①, we estimate the optimum of (OFARR-S). In procedure ②, we derive the penalty weights based on the estimated optimum of (OFARR-S) and subsequently determine the allocation decisions using these weights. In procedure ③, we employ a dual mechanism comprising forced exploration and discarding to

reduce the likelihood of capacity violations (ensure that the decision sequence generated in the procedure ② adheres to the resource capacity constraints with high probability even in the presence of stochastic outcome variations and estimation errors), while also ensuring sufficient exploration. This mechanism operates probabilistically, employing carefully designed probabilities f_n and g_n (as specified in (3)) to explore and discard the decisions computed in procedure ②. The pseudocode of our algorithm, namely Exploration-Discarding with Penalty Weights Update (ED-PWU), is shown in Algorithm 1. Next, we provide detailed explanations for these three procedures.

Procedure ①: estimating λ^*. In this procedure, we use observations collected up to epoch n, i.e., the observations during the time-steps $\{1, ..., t_{n-1}\}$, to compute an estimate $\hat{\lambda}^*(n)$ of λ^*, which is the solution to (OFARR-S). Since parameters $\boldsymbol{p}, \boldsymbol{r}$ and $\boldsymbol{d}$ are unknown in our online (learning) setting, we replace them with their estimates in (OFARR-S) and solve the following optimization problem to obtain $\hat{\lambda}^*(n)$:

$$\text{(OFARR-S)}(n) \quad \max\ \lambda \qquad \text{s.t.} \quad \begin{cases} \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} \tilde{p}_j(n)\cdot \hat{r}_{i,j,k}(n)\cdot x_{j,k} \geq \lambda,\ \forall i\in\mathcal{I}, \\ \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} \tilde{p}_j(n)\cdot \hat{d}_{j,k}(n)\cdot x_{j,k} \leq C, \\ \sum_{k\in\mathcal{K}} x_{j,k} \leq 1,\ \forall j\in\mathcal{J}; \quad x_{j,k}\geq 0,\ \forall k\in\mathcal{K},\ j\in\mathcal{J}. \end{cases}$$

Here $\tilde{\boldsymbol{p}}(n)$ is the empirical estimate of $\boldsymbol{p}$ based on the request arrivals up to epoch n. While $\boldsymbol{p}$ can be directly estimated from the observed samples, as one sample for $\boldsymbol{p}$ is obtained at each time, the estimates for $\boldsymbol{r}$ and $\boldsymbol{d}$ are constructed using their upper confidence bound (UCB) estimate, $\hat{\boldsymbol{r}}(n)$, and lower confidence bound (LCB) estimate, $\hat{\boldsymbol{d}}(n)$, respectively. The intuition behind this choice is to encourage more aggressive decisions regarding utilities while remaining conservative about resource consumption (a similar idea has also been used in related constrained online learning works [3, 36, 37]). The detailed specifications of $\hat{\boldsymbol{r}}(n)$ and $\hat{\boldsymbol{d}}(n)$ are as follows,

$$\hat{r}_{i,j,k}(n) = \min\left\{r_{\max},\ \tilde{r}_{i,j,k}(n) + r_{\max}\sqrt{2\log(|\mathcal{I}||\mathcal{J}||\mathcal{K}|/\delta_n)/N_{j,k}(t_{n-1})}\right\},\ \forall i\in\mathcal{I},\ j\in\mathcal{J},\ k\in\mathcal{K},$$

$$\hat{d}_{j,k}(n) = \max\left\{1,\ \tilde{d}_{i,j,k}(n) - d_{\max}\sqrt{2\log(|\mathcal{J}||\mathcal{K}|/\delta_n)/N_{j,k}(t_{n-1})}\right\},\ \forall j\in\mathcal{J},\ k\in\mathcal{K},$$

where δ_n is the confidence parameter; $N_{j,k}(t)$ is the number of times that the action k is chosen for arrival type j up to time t; and $\tilde{\boldsymbol{r}}(n)$ and $\tilde{\boldsymbol{d}}(n)$ are the empirical estimates for $\boldsymbol{r}$ and $\boldsymbol{d}$, respectively. By Hoeffding's inequality, we can verify that $\hat{r}_{i,j,k}(n) \geq r_{i,j,k}$, w.p. $\geq 1-\delta_n/(|\mathcal{J}||\mathcal{K}||\mathcal{I}|)$ and $\hat{d}_{j,k}(n) \leq d_{j,k}$, w.p. $\geq 1-\delta_n/(|\mathcal{J}||\mathcal{K}|)$.

Procedure ②: calculating allocation decisions via penalty weights. At this procedure in epoch n, upon observing the arriving request type $j(t)$, we sample a weight vector $\boldsymbol{\phi}(t)$ from a weight vector set $\Theta(n)$ uniformly and compute the decision $\tilde{k}(t)$ based on $\boldsymbol{\phi}(t)$ (see (2) in Algorithm 1). In (2), $\hat{\boldsymbol{r}}(n)$ and $\hat{\boldsymbol{d}}(n)$ represent the optimistic and pessimistic estimates for $\boldsymbol{r}$ and $\boldsymbol{d}$, respectively, and are identical to the estimates used in procedure ① (we let $\hat{r}_{i,j(t),k_{\text{null}}}(n) = \hat{d}_{j(t),k_{\text{null}}}(n) = 0$). The vector $\boldsymbol{\phi}(t) = (\phi_0(t), \phi_1(t), ..., \phi_{\mathcal{I}}(t))$ is a penalty weight vector, where $\phi_0(t)$ represents the penalty for resource consumption, and the weights $\{\phi_i(t)\}_{i\in\mathcal{I}}$ quantify the relative importance of different utility types to ensure fairness among them. The weight vector set $\Theta(n)$ is constructed by invoking a MWU process (see Algorithm 2) and the weight vectors in $\Theta(n)$ could be viewed as dual variables solving the online feasibility problem for (OFARR-S)(n) given $\lambda^*(n)$. Notably, in Line 6 of Algorithm 2, ϵ_n^C (as detailed in Lemma 2) represents the confidence radius for $\lambda^*(n)$, and we show in Lemma 3 that $\hat{\lambda}^*(n) - \epsilon_n^C \leq \lambda^*$ with high probability. Generally speaking, these weight vectors altogether describe the dual prices that trade-off between the utility earned of all types and the resource units consumed. They ensure that the sequence $\{\tilde{k}(t)\}_{t=t_{n-1}+1}^{t_n}$ achieves a time-averaged objective value close to λ, the target objective value. Additionally, constructing weight vectors in this way is also for a technical reason as we can ensure that $\{\tilde{k}(t)\}_{t=t_{n-1}+1}^{t_n}$ are mutually independent conditional on

the history $\mathcal{H}(t-1)$, as $j(t)$ is i.i.d. over time. Such independence reduces the difficulty of handling the complicated correlations in resource feasibility across all time steps.

Remark 1. *We remark that in admission control scenarios where only two actions "accept/reject" are available, the selection rule (2) simplifies to a threshold rule based on whether utility outweighs resource cost, i.e., $\tilde{k}(t) = k_{accept}$ only if the indicator variable $\mathbf{I}\{\sum_{i\in|\mathcal{I}|}\phi_i(n)\cdot\hat{r}_{i,j(t),k_{accept}}(n) \geq \phi_0(n)\cdot\hat{d}_{j(t),k_{accept}}(n)\}$ is true. We also note that when $C \geq d_{\max}$, resource availability is guaranteed at all time steps. Consequently, the weight of resource usage, $\phi_0(t)$, approaches zero over time, allowing our algorithm to achieve improved performance guarantees (as reported in Theorem 3).*

Algorithm 1 ED-PWU

Initialization: $t_0 = \ell_0 = d_{\max}$, $\ell_n = 2\ell_{n-1}$, and confidence parameters $\delta_n = 1/t_n^2$ for all $n = 1, 2, \ldots$
for time steps $t = 1, \ldots, \ell_0$ **do**
  If the resource is available, select an arbitrary action. Otherwise, select the null action k_{null}.
end for
for epoch $n = 1, 2, \ldots$ **do**
  Update the empirical estimates $\left(\tilde{\boldsymbol{p}}(n), \tilde{\boldsymbol{r}}(n), \tilde{\boldsymbol{d}}(n)\right)$, UCB estimate $\hat{\boldsymbol{r}}(n)$, LCB estimate $\hat{\boldsymbol{d}}(n)$, respectively.
  Solve (OFARR-S)(n) and obtain its optimal objective value $\hat{\lambda}_*(n)$. // procedure ①
  Run Algorithm 2 and obtain the weight vector set $\Theta_n = \{\boldsymbol{\phi}_n(s)\}_{s=1}^{\ell_n}$.
  $t_n = t_{n-1} + \ell_n$
  for time steps $t = t_{n-1}+1, \ldots, t_n$ **do**
    Sample a weight vector $\boldsymbol{\phi}(t)$ uniformly at random from Θ_n. // procedure ②
    Observe arrival type $j(t)$, and compute

$$\tilde{k}(t) = \arg\max_{k\in\mathcal{K}\cup k_{\text{null}}} \left\{ \sum_{i\in\mathcal{I}} \phi_i(t)\cdot\hat{r}_{i,j(t),k}(n) - \phi_0(t)\cdot\hat{d}_{j(t),k}(n) \right\}. \tag{2}$$

    Sample $\omega(t) \in [0,1]$ and set // procedure ③

$$\hat{k}(t) = \begin{cases} \tilde{k}^e(t), \text{ where } \tilde{k}^e(t) \text{ is sample uniformly at random from } \mathcal{K} & \text{if } \omega(t) \leq f_n, \quad \text{//exploration} \\ \tilde{k}(t) & \text{if } f_n < \omega(t) \leq f_n + (1-f_n)\frac{1}{1+g_n}, \\ k_{\text{null}} & \text{otherwise.} \quad \text{//discarding} \end{cases}$$

    if $\sum_{\tau=1}^{t-1}\mathbf{I}\{D_{j(\tau),k(\tau)}(\tau) \geq t-\tau+1\} < C$ **then**
      Take action $k(t) = \hat{k}(t)$
    else
      Take action $k(t) = k_{\text{null}}$
    end if
  end for
end for

Procedure ③: exploration and discarding. As noted in the introduction section, the sequence $\{\tilde{k}(t)\}_{t=t_{n-1}+1}^{t_n}$ computed in procedure ② cannot be directly adopted as the final decision sequence. This is because it is essential to carefully balance the trade-offs among maintaining resource feasibility, ensuring adequate exploration, and minimizing utility loss caused by discarding or exploration. Our algorithmic ideas to handle these trade-offs are as follows.

- *Exploration.* We allocate a probability f_n (to be defined later) for exploration. If the controller opts to explore, an action k is sampled uniformly at random from $\mathcal{K}$.
- *Discarding.* If exploration is not chosen, the base decision $\tilde{k}(t)$ is discarded independently with a carefully designed probability g_n (to be defined later) to create slack in the resource feasibility. The complement $1-g_n$ could be considered as the retaining probability.

In our analysis, we demonstrate that these two steps ensure the expected resource usage at any time is approximately $(1 - f_n - g_n)$ times the initial resource inventory C. Since this usage is represented as a sum of independent indicator random variables, it concentrates around its expectation, ensuring high-probability satisfaction of the resource capacity constraint when g_n and f_n are appropriately chosen. Furthermore, the utility loss for any type due to discarding and exploration is bounded by approximately $f_n + g_n$ times the expected utility earned by sequence $\{\tilde{k}(t)\}_t$. In order to optimize this trade-off, we choose f_n and g_n as follows.

$$f_n = \ell_n^{-\rho}, \quad g_n = \gamma_1 \cdot \epsilon_n + \gamma_2, \quad \text{where } \epsilon_n = O\left(\sqrt{\frac{\log(1/\delta_n)}{\ell_n}} + \sqrt{\frac{\log(1/\delta_n)}{\min_{j,k} N_{j,k}(n)}}\right). \tag{3}$$

Here $\rho, \gamma_1, \gamma_2 > 0$ are tuning parameters, the specifics of which are elaborated in Theorem 2. The exploration probability f_n is decaying as the epoch grows and will not cause significant overhead in the algorithm performance. We note that ϵ_n captures the estimation errors and is detailed in Lemma 3. The term $O\left(\sqrt{\frac{\log(1/\delta_n)}{\ell_n}}\right)$ in ϵ_n arises from errors in estimating $\boldsymbol{p}$, while $O\left(\sqrt{\frac{\log(1/\delta_n)}{\min_{j,k} N_{j,k}(n)}}\right)$ reflects the largest uncertainty in estimating utilities and resource usage duration, dominated by the type-action pair with the fewest samples. The coefficients γ_1 and γ_2 in g_n aim to balance the trade-off between guaranteeing resource feasibility and minimizing utility loss due to discarding. In the practical implementation of the algorithm, we use a retaining probability of $1/(1 + g_n)$ instead of $1 - g_n$ to avoid the potential negativity of $1 - g_n$ during early epochs. When n is sufficiently large, $1/(1 + g_n)$ approximates $1 - g_n$, preserving the desired behavior. The discarding and exploration complicate the analysis of point-wise resource feasibility. In Section 5.3, we develop a coupling technique to disentangle the intricate dependencies between point-wise resource feasibility (induced by the sequence $\{k(t)\}_t$) and the outcomes resulting from the sequence $\{\hat{k}(t)\}_t$.

Algorithm 2 Penalty weight vectors construction (invoked at the start of epoch n)

Input: estimates $\left(\hat{\lambda}^*(n), \hat{\boldsymbol{r}}, \hat{\boldsymbol{d}}(n), \tilde{\boldsymbol{p}}(n)\right)$

Initialize: $\eta_t = \frac{\sqrt{\log(|\mathcal{I}|+1)}}{\max\{r_{\max}, d_{\max}\}\sqrt{t}}$ for each $s = 1, 2, \ldots, \ell_n$; $\boldsymbol{\phi}(0) = \frac{1}{|\mathcal{I}|+1} \cdot \mathbf{1}_{|\mathcal{I}|+1}$.

for virtual time steps $t = 1, \ldots, \ell_n$ **do**

  Sample $j^v(t)$ from $\tilde{\boldsymbol{p}}(n)$

  Compute the *virtual* action: $\tilde{k}^v(t) = \arg\max_{k \in \mathcal{K} \cup k_{\text{null}}} \left\{\sum_{i \in \mathcal{I}} \phi_i(t) \cdot \hat{r}_{i,j^v(t),k}(n) - \phi_0(t) \cdot \hat{d}_{j^v(t),k}(n)\right\}$.

  Update the penalty vector:

$$\phi_i(t+1) = \phi_i(t) \cdot \exp\left(-\eta_t \left[\hat{r}_{i,j^v(t),\tilde{k}^v(t)}(n) - \left(\hat{\lambda}^*(n) - \epsilon_C(n)\right)\right]\right), \quad \forall i \in \{1, ..., |\mathcal{I}|\}$$

$$\phi_0(t+1) = \phi_0(t) \cdot \exp\left(-\eta_t \left[-\hat{d}_{j^v(t),\tilde{k}^v(t)}(n) + C\right]\right), \quad i = 0$$

$$\boldsymbol{\phi}(t+1) = \boldsymbol{\phi}(t+1)/\|\boldsymbol{\phi}(t)\|_1$$

end for

Return the weight vector set $\Theta_n = \{\boldsymbol{\phi}(t)\}_{t=1}^{\ell_n}$.

5 Main Theoretical Results and Analysis

In this section, we first introduce the designs of the discarding density g_n and the exploration density f_n. We then present the corresponding performance bounds of our algorithm.

5.1 The design of f_n and g_n

The following theorem is instrumental for the design of discarding density $g_n = \gamma_1 \cdot \epsilon_n + \gamma_2$ and exploration density $f_n = \ell_n^{-\rho}$.

Theorem 1. *Given any $\varepsilon > 0$, it holds that w.p. $1 - 23\delta_n$ for all $i \in \mathcal{I}$ and $n \geq 1$:*

$$\sum_{t=t_{n-1}+1}^{t_n} R_{i,j(t),k(t)}(t) \geq \underbrace{\frac{1}{1+g_n}\left(1 - \frac{1}{(1+\varepsilon)^{C-1}} \cdot \exp\left[\frac{\varepsilon(C+\epsilon_n)}{1+g_n}\right]\right)^+}_{\text{ratio term}} \ell_n \cdot \lambda^* - \underbrace{\ell_n\left(\epsilon_n^A + \epsilon_n^B + \epsilon_n^C + \epsilon_n + f_n\lambda^* + 2\varepsilon f_n (d_{max} - C)^+ \cdot \lambda^*\right) - r_{max}\sqrt{2d_{max}\ell_n \log(d_{max}/\delta_n)} - r_{max}d_{max}}_{\text{error term}}, \quad (4)$$

where $(\epsilon_A, \epsilon_B, \epsilon_C)$ and ϵ_n are quantities provided in Lemmas 2 and 3, respectively.

Theorem 1 serves as the foundation for our theoretical results by providing a lower bound to the accumulated utility of type i during each epoch n. In particular, by setting the specific values of $\epsilon_n^A, \epsilon_n^B, \epsilon_n^C$ and ϵ_n into (4), we can obtain that the error term in (4) is of the order $O(\sum_{n=1}^N (f_n + \ell_n\sqrt{\log(1/\delta_n)/\min_{j,k} N_{j,k}(t_{n-1})} + \sqrt{\ell_n \log(1/\delta_n)}))$, where $N = \lceil \log(T/d_{\max}) \rceil$ is the total number of epochs. To derive a performance bound for our algorithm via Theorem 1, it suffices to establish a lower bound for $\min_{j,k} N_{j,k}(t_{n-1})$. By noticing that any policy encounters at least one instance of resource availability every $d_{\max}$ rounds, we can immediately obtain the following expected lower bound on the number of samples for all pairs (j,k):

$$\mathbf{E}\left[\min_{j,k} N_{j,k}(t_{n-1})\right] = \sum_{t=1}^{t_{n-1}} \mathbf{P}\left(k(t) = k, j(t) = j\right) \geq \sum_{n=0}^{n-1} \ell_n \cdot f_n \cdot \frac{1}{d_{\max}} \frac{p_j}{|\mathcal{K}|} = \sum_{n=0}^{n-1} \ell_n^{1-\rho} \cdot \frac{1}{d_{\max}} \frac{p_j}{|\mathcal{K}|} = O\left(2^{n(1-\rho)}\right).$$

The term $1/d_{\max}$ arises from observing that $k(t) = \hat{k}(t)$ (line 15 in Algorithm 1) occurs at least once every $d_{\max}$ rounds. Using Chernoff bounds, we can show that $\mathbf{P}\left(N_{j,k}(n) \leq (1 - \frac{\sqrt{2}}{2}) \cdot \mathbf{E}[N_{j,k}(n)]\right) \leq \exp\left(-\mathbf{E}[N_{j,k}(n)]/4\right) \leq O\left(1/4^{n(1-\rho)}\right)$. Consequently, the expectation of the error term becomes $O\left(\mathbf{E}\left[\sum_{n=1}^N \left(f_n + \ell_n\sqrt{\log(1/\delta_n)/\min_{j,k} N_{j,k}(t_{n-1})} + \sqrt{\ell_n \log(1/\delta_n)}\right)\right]\right) \leq O(T^{\frac{1+\rho}{2}}\sqrt{\log T} + T^{1-\rho})$. By setting $\rho = 1/3$, we minimize the order of this bound to $O(T^{2/3})$. Accordingly, we choose $f_n = \ell_n^{-1/3}$ in our algorithm to match this analysis, and this choice also explains the emergence of the $O(T^{2/3})$ term in our performance bound (see Theorem 2). To deal with the ratio term in (4), we set $\gamma_1 = (\varepsilon+1)/C$ and $\gamma_2 = \varepsilon$ in the definition of g_n (ε will be specified later), resulting in

$$\frac{1}{1+g_n} = \frac{1}{1+(\varepsilon+1)\epsilon_n/C + \varepsilon} = \frac{1}{1+\epsilon_n/C} \cdot \frac{1}{\varepsilon+1},$$

$$\frac{1}{(1+\varepsilon)^{C-1}} \cdot \exp\left[\frac{\varepsilon(C+\epsilon_n)}{1+g_n}\right] = (1+\varepsilon) \cdot \frac{1}{(1+\varepsilon)^C} \cdot \exp(\frac{\varepsilon}{1+\varepsilon} \cdot C) = (1+\varepsilon)\exp\left[C\left(\frac{\varepsilon}{1+\varepsilon} - \log(1+\varepsilon)\right)\right].$$

Thus, the ratio term in (4) can be rewritten as

$$\text{ratio term in (4)} = \frac{1}{1+\epsilon_n/C} \cdot \frac{1}{\varepsilon+1} \cdot \left(1 - (1+\varepsilon)\exp\left[C\left(\frac{\varepsilon}{1+\varepsilon} - \log(1+\varepsilon)\right)\right]\right)^+. \quad (5)$$

Denote $\mathcal{L}(\varepsilon)$ as

$$\mathcal{L}(\varepsilon) = \frac{1}{\varepsilon+1} \cdot \left(1 - (1+\varepsilon)\exp\left[C\left(\frac{\varepsilon}{1+\varepsilon} - \log(1+\varepsilon)\right)\right]\right)^+, \quad (6)$$

and then (5) implies that our algorithm can achieve at least $\frac{1}{1+\epsilon_n/C} \cdot \mathcal{L}(\varepsilon)$ competitive ratio against $\ell_n \cdot \lambda^*$ for any $\varepsilon > 0$ at epoch $n \geq 1$. Let $\varepsilon^*(C)$ be the optimal assignment solving Equation (6). It is not hard to verify that $\varepsilon^*(C) = O(\sqrt{\log C/C})$. Next, we show that the ratio term in (4) is of order $1 - O(\sqrt{\log C/C})$ under $\epsilon = \varepsilon^*(C)$ in the asymptotic regime of C. Note that the term

$\exp\left[C\left(\frac{\varepsilon}{1+\varepsilon}-\log(1+\varepsilon)\right)\right]$ in $\mathcal{L}(\varepsilon)$ can be bounded as follows,

$$\exp\left[C\left(\frac{\varepsilon}{1+\varepsilon}-\log(1+\varepsilon)\right)\right] \le \exp\left[C\left(\frac{\varepsilon}{1+\varepsilon}-\left(\varepsilon-\frac{\varepsilon^2}{2}\right)\right)\right] = \exp\left[C\left(-\frac{\varepsilon^2}{1+\varepsilon^2}+\frac{\varepsilon^2}{2}\right)\right]$$
$$= \exp\left[C\left(-\frac{\varepsilon^2}{2}+\varepsilon^2\frac{\varepsilon^2}{1+\varepsilon^2}\right)\right] \le \exp\left[C\left(-\frac{\varepsilon^2}{2}+\varepsilon^4\right)\right] = \exp\left[-\frac{C\varepsilon^2}{2}\right]\exp\left[C\varepsilon^4\right].$$

Using $\mathcal{L}(\varepsilon^*(C)) \ge \mathcal{L}\left(\sqrt{\log C/C}\right)$ and combining all the things together, we can find that when n is sufficiently large such that $\epsilon_n \le 0.5$:

$$\begin{aligned}
\text{ratio term in (4)} &= \frac{1}{1+\epsilon_n/C}\cdot\mathcal{L}(\varepsilon^*(C)) \ge \frac{1}{1+1/(2C)}\cdot\mathcal{L}\left(\sqrt{\log C/C}\right)\\
&\ge \frac{1}{1+1/(2C)}\cdot\frac{1}{1+\sqrt{\log C/C}}\cdot\left(1-\frac{1}{1+\sqrt{\log C/C}}\cdot\exp\left[-\frac{C\frac{\log C}{C}}{2}\right]\exp\left[C\left(\frac{\log C}{C}\right)^2\right]\right)^+\\
&\ge \left(1-\frac{1}{2C}\right)\cdot\left(1-\sqrt{\frac{\log C}{C}}\right)\cdot\left(1-\frac{1}{1+\sqrt{\log C/C}}\cdot\sqrt{\frac{\log C}{C}}\exp\left[\frac{\log^2 C}{C}\right]\right)^+\\
&= O\left(\left(1-\frac{1}{2C}\right)\cdot\left(1-\sqrt{\frac{\log C}{C}}\right)\right) = 1-O\left(\sqrt{\frac{\log C}{C}}\right).
\end{aligned}$$

Therefore, by setting $\rho = 1/3$, $\gamma_1 = (\varepsilon^*(C)+1)/C$, and $\gamma_2 = \varepsilon^*(C)$ in Algorithm 1, and combining with Lemma 1, our algorithm achieves at least $1 - O(\sqrt{\log C/C})$ competitive ratio against the optimum OPT*. We formally establish this statement in the following section.

5.2 Main competitive results

In this section, we present the theoretical guarantees of the developed algorithm. The following theorem establishes the performance bounds for our algorithm by using the discarding and exploration densities (f_n, g_n) defined in Section 5.1.

Theorem 2. *Set $\rho = 1/3, \gamma_1 = (\varepsilon^*(C)+1)/C$, and $\gamma_2 = \varepsilon^*(C)$ in (3), where $\varepsilon^*(C) = \arg\max_{\varepsilon>0}\mathcal{L}(\varepsilon) = \frac{1}{\varepsilon+1}\cdot\left(1-(1+\varepsilon)\exp\left[C\left(\frac{\varepsilon}{1+\varepsilon}-\log(1+\varepsilon)\right)\right]\right)^+$. Then our algorithm ensures that*

$$\min_{i\in\mathcal{I}}\mathbf{E}\left[\frac{1}{T}\cdot\sum_{t=1}^{T}R_{i,j(t),k(t)}(t)\right] \ge \left(1-\frac{1}{2C}\right)\cdot\mathcal{L}(\varepsilon^*(C))\cdot OPT^* - O\Big(|\mathcal{J}|d_{\max}\cdot T^{2/3}\sqrt{\log T\cdot\log(|\mathcal{J}||\mathcal{K}||\mathcal{I}|)}$$
$$+\varepsilon^*(C)\cdot(d_{max}-C)^+\cdot T^{2/3}+T^{2/3}+d_{\max}^3+d_{\max}/p_{\min}+\sum_{j\in\mathcal{J}}|\mathcal{K}|/p_j\Big)/T.$$

Here, we have $\mathcal{L}(\varepsilon^*(C)) \ge 1 - O(\sqrt{\log C/C})$ as previously proved, and its curve is depicted in Figure 1. Theorem 2 implies that the performance of our algorithm is closer to the optimum when T and C increase. A larger C means that the controller is endowed with more resource units to buffer against the model uncertainty, stochastic nature of the outcomes, and limited information feedback. Compared to [17], where the error term is of the order $O(T^{1/2}+d_{\max}^3)$, our performance bound introduces error terms that scale as $O(d_{\max}\cdot T^{2/3}+d_{\max}^3+\sum_{j\in\mathcal{J}}|\mathcal{K}|/p_j+d_{\max}/p_{\min})$. The leading term $O(d_{\max}\cdot T^{2/3})$ represents the performance loss due to forced exploration, highlighting the impact of model uncertainty and the bandit feedback structure. While our current approach adopts a fixed

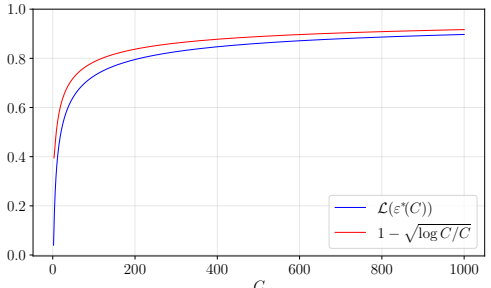


Fig. 1. The curve of $\mathcal{L}(\varepsilon^*(C))$

forced exploration schedule, a more adaptive exploration strategy may help mitigate this overhead and improve the theoretical guarantees. Therefore, an intriguing direction for future work is to design more refined exploration mechanisms that have the potential to surpass this $O(T^{2/3})$ bound. In Section 6, we demonstrate that under the Quasi-full-feedback setup, our algorithm improves this term to $O(d_{\max} \cdot T^{1/2})$, achieving the same order of dependence on the time-horizon as [17] under the same feedback setup. The coefficient $d_{\max}$ reflects the absence of specific assumptions about the distributions of $D_{j,k}(t)$, such as known variance. By Hoeffding's inequality, the accumulated estimation error for $\boldsymbol{d}$ at any given time is at least proportional to its support $d_{\max}$, and thus the coefficient $d_{\max}$ in relation to T generally cannot be improved. This contrasts with [17], which assumes deterministic resource usage durations. The coefficient $d_{\max}$ also implies that our algorithm achieves a meaningful competitive guarantee only when $d_{\max} = o(T^{1/3})$. Additionally, the additional constant term $O\left(\sum_{j\in\mathcal{J}} |\mathcal{K}|/p_j + d_{\max}/p_{\min}\right)$ arises from the uncertainty in $\boldsymbol{p}$. Intuitively, when p_j is very small—indicating that insufficient samples are available for accurate estimation under arrival type j—significant performance loss may occur if actions associated with this arrival type can yield high utility but with minimal resource usage time. In scenarios with only one request type, as considered in [17], $p_{\min} = |\mathcal{J}| = 1$, and thus $d^3_{\max}$ becomes the dominant constant term.

Finally, we would like to remark that the performance guarantee in Theorem 2 is not optimized for the case where $C \geq d_{\max}$. In fact, when resources are abundant ($C \geq d_{\max}$), resource availability is ensured at all time steps, hence, discarding becomes unnecessary. In this case, we set the discarding probability g_n in our algorithm to zero (i.e., $\gamma_1 = \gamma_2 = 0$ in (3)). As a result, our algorithm achieves an improved competitive ratio of 1. This improved performance guarantee under resource sufficiency is formally presented in the following theorem.

Theorem 3. *For the case where $C \geq d_{\max}$, set $\rho = 1/3$, $\gamma_1 = 0$, and $\gamma_2 = 0$ in (3) allows our algorithm to further guarantee that*

$$\min_{i\in\mathcal{I}} \mathbf{E}\left[\frac{1}{T}\cdot\sum_{t=1}^{T} R_{i,j(t),k(t)}(t)\right] \geq OPT^* - O\Big(|\mathcal{J}|d_{\max}\cdot T^{2/3}\sqrt{\log T\cdot\log(|\mathcal{J}||\mathcal{K}||\mathcal{I}|)} + \varepsilon^*(C)\cdot(d_{max}-C)^+\cdot T^{2/3} + T^{2/3} + d^3_{\max} + \sum_{j\in\mathcal{J}}|\mathcal{K}|/p_j\Big)/T.$$

Theorem 3 implies that our algorithm can guarantee a sublinear regret of $O(T^{2/3})$ when resources are abundant ($d \geq d_{\max}$). Notably, when $C \geq d_{\max}$, our model reduces to the indivisible variant of the horizon-fairness optimization problem in the online setting. In this context, the most closely related work is by [12] which achieves an $O(T^{1/2})$ regret bound under the full-feedback setup. In Section 6, we relax the bandit-feedback setup to a Quasi-full-feedback setup (weaker than the full feedback setup) and our algorithm guarantees the same regret order of $O(T^{1/2})$. However, it remains unclear whether the $O(T^{2/3})$ regret bound is optimal in the bandit-feedback setup, and we leave this as an open problem for future work.

We remark that the proofs of our main theoretical results (i.e., Theorems 2 and 3) are built upon Theorem 1, which involve bounding its error term and applying a union bound across epochs. Accordingly, we provide the proof of Theorem 1 in the next section. We introduce our proof strategy by breaking the proof of Theorem 1 into lemmas.

5.3 Proof of Theorem 1

We begin by introducing a coupling technique that serves as a key tool for analyzing Theorem 1. This process involves bounding the accumulated utility for every type $i \in \mathcal{I}$ in each epoch $n \in \{1, \ldots, \lceil \log_2(T/d_{\max})\rceil\}$, and thereby leads to the proof of Theorem 1. Note that analyzing the utility collected within each epoch is not straightforward due to the intricate dependence of

resource availability on both the previously allocated resources and their usage durations. We disentangle this intricate dependence via the following coupling argument. For notational convenience, let $(\tilde{\boldsymbol{R}}(t), \tilde{A}(t), \tilde{D}(t))$ represent the stochastic outcomes (i.e., the yield utilities, the amount of consumed resource, and the resource usage duration) under action $\tilde{k}(t)$. By our model setup, we have $\tilde{\boldsymbol{R}}(t) = (R_{1,j(t),\tilde{k}(t)}(t), \ldots, R_{\mathcal{I},j(t),\tilde{k}(t)}(t))$, $\tilde{A}(t) = 1$, and $\tilde{D}(t) = D_{j(t),\tilde{k}(t)}(t)$. Similarly, we denote $(\tilde{\boldsymbol{R}}^e(t), \tilde{A}^e(t), \tilde{D}^e(t))$ and $(\hat{\boldsymbol{R}}(t), \hat{A}(t), \hat{D}(t))$ as the stochastic outcomes under the actions $\tilde{k}^e(t)$ and $\hat{k}(t)$, respectively. We also define two indicator variables $I_1(t) = \mathbf{I}\{\omega(t) \le f_n\}$ and $I_2(t) = \mathbf{I}\{f_n < \omega(t) \le f_n + {}^{(1-f_n)}/_{(1+g_n)}\}$, which are mutually exclusive and are Bernoulli variables with mean f_n and ${}^{(1-f_n)}/_{(1+g_n)}$, respectively. Note that the random variables $\{(I_1(t), I_2(t))\}_{t=t_{n-1}+1}^{t_n}$ are jointly independent, and they are independent of $\{(\tilde{\boldsymbol{R}}(t), \tilde{A}(t), \tilde{D}(t))\}_{t=t_{n-1}+1}^{t_n}$ as well. Then by the definition of $\hat{k}(t)$, the stochastic outcomes $(\hat{\boldsymbol{R}}(t), \hat{A}(t), \hat{D}(t))$ can be decomposed as

$$\begin{aligned} \hat{R}_i(t) &= I_2(t) \cdot \tilde{R}_i(t) + I_1(t) \cdot \tilde{R}_i^e(t), \text{ for all } i \in \mathcal{I}; \\ \hat{A}(t) = I_2(t) \cdot \tilde{A}(t) + I_1(t) \cdot \tilde{A}^e(t); \quad \hat{D}(t) &= I_2(t) \cdot \tilde{D}(t) + I_1(t) \cdot \tilde{D}^e(t). \end{aligned} \tag{7}$$

With a slight abuse of notation, let $(\boldsymbol{R}(t), A(t), D(t))$ be the actual outcomes under the final decision $k(t)$, i.e., $R_i(t) = R_{i,j(t),k(t)}(t), D(t) = D_{j(t),k(t)}(t)$ and $A(t) = \mathbf{I}\{k(t) \ne k_{\text{null}}\}$. Then

$$R_i(t) = \hat{R}_i(t) \cdot \mathbf{I}\left\{\sum_{\tau=1}^{t-1} A(\tau)\mathbf{I}\{D(\tau) \ge t - \tau + 1\} \le C - 1\right\}, \text{ for all } i \in \mathcal{I}; \tag{8}$$

$$A(t) = \hat{A}(t) \cdot \mathbf{I}\left\{\sum_{\tau=1}^{t-1} A(\tau)\mathbf{I}\{D(\tau) \ge t - \tau + 1\} \le C - 1\right\}; \; D(t) = \hat{D}(t) \cdot \mathbf{I}\left\{\sum_{\tau=1}^{t-1} A(\tau)\mathbf{I}\{D(\tau) \ge t - \tau + 1\} \le C - 1\right\}.$$

The equations in (8), elaborate on the intricate dependency between the outcomes $\{\boldsymbol{R}(t), A(t), D(t)\}$ at time t and those from preceding time steps. This dependency is captured by the indicator random variable which determines whether the controller can make an allocation at time t. Based on (8), we handle the accumulated utility of type i within epoch n below:

$$\begin{aligned} \sum_{t=t_{n-1}+1}^{t_n} R_i(t) &= \sum_{t=t_{n-1}+1}^{t_n} \hat{R}_i(t) \cdot \mathbf{I}\left\{\sum_{\tau=1}^{t-1} A_i(\tau)\mathbf{I}\{D_i(\tau) \ge t - \tau + 1\} \le C - 1\right\} \\ &\ge \sum_{t=t_{n-1}+1}^{t_n} \hat{R}_i(t) \cdot \mathbf{I}\left\{\sum_{\tau=1}^{t-1} \hat{A}(\tau)\mathbf{I}\{\hat{D}(\tau) \ge t - \tau + 1\} \le C - 1\right\} \qquad (9) \\ &= \sum_{t=t_{n-1}+1}^{t_n} \hat{R}_i(t) \cdot \mathbf{I}\left\{\sum_{\tau=\max\{t-d_{\max},1\}}^{t-1} \hat{A}(\tau)\mathbf{I}\{\hat{D}(\tau) \ge t - \tau + 1\} \le C - 1\right\} \qquad (10) \\ &\ge \sum_{t=\min\{t_{n-1}+1+d_{\max},t_n\}}^{t_n} \hat{R}_i(t) \cdot \mathbf{I}\left\{\sum_{\tau=\max\{t-d_{\max},1\}}^{t-1} \hat{A}(\tau)\mathbf{I}\{\hat{D}(\tau) \ge t - \tau + 1\} \le C - 1\right\}, \qquad (11) \end{aligned}$$

where (9) is because the coupling (8) ensures $A(t) \le \hat{A}(t)$ and $D(t) \le \hat{D}(t)$ almost surely; (10) holds since $\hat{D}(t) \le d_{\max}$. Here we would like to remark that the inequality (11) facilitates our analysis, because conditioned on the parameters $\{\Theta_n, \hat{\boldsymbol{r}}(n), \hat{\boldsymbol{d}}(n), f_n, g_n\}$, which is $\sigma(\mathcal{H}(t_{n-1}))$-measurable, the random outcomes $\{(\hat{\boldsymbol{R}}(t), \hat{A}(t), \hat{D}(t))\}_{t=t_{n-1}+1}^{t_n}$ defined in (7) are i.i.d. Consequently, the random outcomes $\{(\hat{\boldsymbol{R}}(t), \hat{A}(t), \hat{D}(t))\}_{t=t_{n-1}+1}^{t_n}$ are much easier to analyse than the actual random outcomes $\{(\boldsymbol{R}(t), A(t), D(t))\}_{t=t_{n-1}+1}^{t_n}$.

Next, we establish a lower bound for the sum on the r.h.s of (11), which plays an important role in the proof of Theorem 1. Denote $V(t) = \mathbf{I}\big\{\sum_{\tau=\max\{t-d_{\max},1\}}^{t-1} \hat{A}(\tau)\mathbf{I}\{\hat{D}(\tau) \ge t - \tau + 1\} \le C - 1\big\}$,

which represents the resource availability at time t induced by the sequence $\{\hat{k}(t)\}_t$. Consequently, the sum on the r.h.s. of (11) can be rewritten as follows:

$$(11) = \sum_{t=t_{n-1}+1+d_{\max}}^{t_n} \hat{R}_i(t) \cdot V(t).$$

The analysis on the r.h.s of (11) involves lower bounding the probability of resource availability $V(t)$ and the utilities earned $\hat{\boldsymbol{R}}(t)$ induced by the sequence $\{\hat{k}(t)\}_t$ at all time steps within epoch n. Notably, $\hat{k}(t)$ is closely tied to $\tilde{k}(t)$, which is determined by the penalty weight vectors in Θ_n. These weight vectors could be viewed as dual variables solving the online feasibility problem for (OFARR-S)(n) given $\lambda^*(n)$. To this end, the following lemma bounds the estimation error of $\hat{\lambda}^*(n)$.

Lemma 2. *Define* $\epsilon_n^A = 2\sqrt{\frac{d_{\max}\log(1/\delta_n)}{C\cdot t_{n-1}}} + \frac{2d_{\max}\log(1/\delta_n)}{C\cdot t_{n-1}}$, $\epsilon_n^B = r_{\max}\sqrt{\frac{2\log(|\mathcal{I}|/\delta_n)}{t_{n-1}}}$, *and* $\epsilon_n^C = 2r_{\max}\sqrt{\frac{\log(1/\delta_n)}{t_{n-1}}} + 2r_{\max}\frac{\log(1/\delta_n)}{t_{n-1}} + 2|\mathcal{J}|\,(r_{\max}+d_{\max})\sqrt{\frac{2\log(|\mathcal{J}||\mathcal{K}||\mathcal{I}|/\delta_n)}{\min_{j\in\mathcal{J},k\in\mathcal{K}} N_{j,k}(t_{n-1})}}$. *For any* $n \geq 1$, *we can guarantee that*

$$\textbf{(a)}\quad \hat{\lambda}^*(n) \geq \lambda^* - \epsilon_n^B - \epsilon_n^A, \quad \text{w.p. } 1-4\delta_n, \qquad \textbf{(b)}\quad \hat{\lambda}^*(n) \leq \lambda^* + \epsilon_n^C, \quad \text{w.p. } 1-3\delta_n.$$

Subsequently, the following lemma establishes a connection between (OFARR-S)(n) and the time-averaged expected resource consumption and utilities induced by the sequence $\{\tilde{k}(t)\}_{t=t_{n-1}+1}^{t_n}$.

Lemma 3. *For notational convenience, define oracle* $\kappa_n(\boldsymbol{\phi}, j) = \arg\max_{k\in\mathcal{K}\cup k_{null}}\{\sum_{i\in\mathcal{I}}\phi_i\cdot\hat{r}_{i,j,k}(n) - \phi_0\cdot\hat{d}_{j,k}(n)\}$. *Denote* $\epsilon_n = 6d_{\max}\sqrt{\frac{2}{t_{n-1}}\log\frac{2}{\delta_n}} + d_{\max}\sqrt{\frac{8\log(|\mathcal{I}|+1)}{\ell_n}} + d_{\max}\sqrt{\frac{2\log(|\mathcal{I}||\mathcal{J}||\mathcal{K}|)/\delta_n}{\min_{j,k}N_{j,k}(t_{n-1})}}$, *then we have that*

$$\frac{1}{\ell_n}\sum_{s=1}^{\ell_n}\sum_{j\in\mathcal{J}} p_j\cdot r_{i,j,\kappa_n(\boldsymbol{\phi}_n(s),j)} \geq \lambda^* - \epsilon_n^A - \epsilon_n^B - \epsilon_n^C - \epsilon_n, \quad \text{w.p. } 1-11\delta_n,$$

$$\frac{1}{\ell_n}\sum_{s=1}^{\ell_n}\sum_{j\in\mathcal{J}} p_j\cdot d_{j,\kappa_n(\boldsymbol{\phi}_n(s),j)} \leq C + \epsilon_n, \quad \text{w.p. } 1-11\delta_n.$$

The lemma 3 builds upon lemma 2 and the properties of the MWU process. Leveraging lemma 3, we derive the following lower bounds for $V(t)$ and $\hat{\boldsymbol{R}}(t)$ under our algorithm:

Lemma 4. *Our algorithm ensures that for any* $\varepsilon > 0$, *with probability* $1-11\delta_n$ *the inequality*

$$\begin{aligned}\mathbf{E}[V(t)\mid\mathcal{H}(t_{n-1})] &\geq \left(1 - \frac{1}{(1+\varepsilon)^{C-1}}\cdot\exp\left[\frac{\varepsilon(1-f_n)}{1+g_n}\cdot\frac{1}{\ell_n}\sum_{s=1}^{\ell_n}\sum_{j\in\mathcal{J}} p_j\cdot d_{j,f_n(\boldsymbol{\phi}_n(s),j)} + \varepsilon\cdot f_n\cdot d_{\max}\right]\right)^+ \\ &\geq \left(1 - \frac{1}{(1+\varepsilon)^{C-1}}\cdot\exp\left[\frac{\varepsilon(C+\epsilon_n)}{1+g_n}\right]\cdot\exp\left[\varepsilon\cdot f_n\cdot(d_{max}-C)^+\right]\right)^+\end{aligned}$$

holds simultaneously for all $t \in \{t_{n-1}+1+d_{\max},\ldots,t_n\}$. *Also, for any* $i\in\mathcal{I}$, *with probability* $1-11\delta_n$ *we have that simultaneously for all* $t\in\{t_{n-1}+1+d_{\max},\ldots,t_n\}$:

$$\mathbf{E}[\hat{R}_i(t)\mid\mathcal{H}(t_{n-1})] \geq \frac{1-f_n}{1+g_n}\cdot\frac{1}{\ell_n}\sum_{s=1}^{\ell_n}\sum_{j\in\mathcal{J}} p_j\cdot r_{i,j,\kappa_n(\boldsymbol{\phi}_n(s),j)} \geq \frac{1}{1+g_n}\lambda^* - \epsilon_n^A - \epsilon_n^B - \epsilon_n^C - \epsilon_n - f_n\cdot\lambda^*.$$

We note that, in the special case where $C \geq d_{\max}$, it is straightforward to see that $\mathbf{E}[V(t)\mid\mathcal{H}(t_{n-1})] = 1$ as resource availability is always ensured. This enhanced technical result forms the basis of Theorem 3. Generally, Lemma 4 establishes that given any $\varepsilon > 0$ and $i\in\mathcal{I}$, the following inequality

holds simultaneously for all $t \in \{t_{n-1}+1+d_{\max}, \ldots, t_n\}$

$$\mathbf{E}\left[\hat{R}_i(t)\cdot V(t) \mid \mathcal{H}(t_{n-1})\right] \geq \frac{1}{1+g_n}\left(1-\frac{1}{(1+\varepsilon)^{C-1}}\exp\left[\frac{\varepsilon(C+\epsilon_n)}{1+g_n}\right]\exp\left[\varepsilon\cdot f_n(d_{max}-C)^+\right]\right)^+\cdot\lambda^* - \left(\epsilon_n^A+\epsilon_n^B+\epsilon_n^C+\epsilon_n+f_n\lambda^*\right), \quad \text{w.p. } 1-22\delta_n. \tag{12}$$

When n is sufficient large such that $\varepsilon\cdot f_n\cdot(d_{max}-C)^+ \leq 1$, e.g., $n \geq 3\log\varepsilon + 2\log d_{\max}$, the inequality $e^x \leq 1+2x$ for any $x\in[0,1]$ ensures that

$$\mathbf{E}\left[\hat{R}_i(t)\cdot V(t) \mid \mathcal{H}(t_{n-1})\right] \geq \frac{1}{1+g_n}\left(1-\frac{1}{(1+\varepsilon)^{C-1}}\cdot\exp\left[\frac{\varepsilon(C+\epsilon_n)}{1+g_n}\right]\right)^+\cdot\lambda^* - \left(\epsilon_n^A+\epsilon_n^B+\epsilon_n^C+\epsilon_n+f_n\lambda^*+2\varepsilon\cdot f_n\cdot(d_{max}-C)^+\cdot\lambda^*\right), \quad \text{w.p. } 1-22\delta_n. \tag{13}$$

Inequality (13) provides a conditional expectation lower bound for the r.h.s of (11). To complete the the proof of Theorem 1, we still need a conditional Chernoff inequality, stated below:

Lemma 5 (Conditional Chernoff Inequality [40]). *Let the random variables $\{X_t\}_{t=1}^N$ satisfy the following conditions: (I) The variables $X_1,\ldots,X_N \in [0,B]^N$ are jointly independent given a σ-algebra $\mathcal{F}$; (II) There exist real numbers $\epsilon\in[0,1]$ and $\nu>0$ such that $\mathbf{P}\left(\mathbf{E}\left[\sum_{t=1}^N X_t\middle|\mathcal{F}\right]<\nu\right)\leq\epsilon$. Then for any $\delta\in(0,1)$, the following inequality holds: $\mathbf{P}\left(\sum_{t=1}^N X_t < \nu-\sqrt{2B\nu\log(1/\delta)}\right)\leq\delta+\epsilon$.*

However, Theorem 1 cannot be directly derived from Lemma 5 applied to (13). This is because the random variables $\{\hat{R}_i(t)V(t)\}_{t=t_{n-1}+d_{\max}+1}^{t_n}$ remain correlated even when conditioned on $\mathcal{H}(t_{n-1})$. To address this, we employ Lemma 5 on carefully selected subsets of $\{\hat{R}_i(t)V(t)\}_{t=t_{n-1}+d_{\max}+1}^{t_n}$, partitioning the entire set into disjoint groups. For this purpose, let $\Gamma = \lceil \ell_n - d_{\max}/d_{\max}\rceil$. For $q\in\{1,\ldots,d_{\max}\}$ and $p\in\{1,\ldots,\Gamma\}$, we define the time index $t(p;q)=(t_{n-1}+d_{\max})+q+(p-1)d_{\max}$. We argue that for any fixed $q\in\{1,\ldots,d_{\max}\}$, the random variables in the following set are i.i.d. conditioned on $\mathcal{H}(t_{n-1})$:

$$\Psi(q)=\left\{V\left(t(p;q)\right)\cdot\hat{R}_i\left(t(p;q)\right)\right\}_{p=1}^{\Gamma}.$$

To establish the conditional independence, note that $\hat{R}_i(t)V(t)$ is $\sigma(\{(\hat{\boldsymbol{R}}(\tau),\hat{A}(\tau),\hat{D}(\tau))\}_{\tau=t-d_{\max}}^{t})$-measurable for any time t. More specifically, there exists a deterministic function z_i such that $\hat{R}_i(t)V(t)=z_i(\{(\hat{\boldsymbol{R}}(\tau),\hat{A}(\tau),\hat{D}(\tau))\}_{\tau=t-d_{\max}}^{t})$, where z_i depends only on i and remains invariant with respect to t. Since the time indexes within $\Psi(q)$ are separated by at least $d_{\max}$ time steps, we know that for any $p,p'\in\{1,\ldots,\Gamma\}$ with $p\neq p'$, the index sets $\{t(p;q)-d_{\max},\ldots,t(p,q)\}$ and $\{t(p';q)-d_{\max},\ldots,t(p',q)\}$ do not overlap. By observing that $\{(\hat{\boldsymbol{R}}(\tau),\hat{A}(\tau),\hat{D}(\tau))\}_{\tau=t_{n-1}+1}^{t_n}$ are independent conditioned on $\mathcal{H}(t_{n-1})$, we conclude that the random variables within $\Psi(q)$ are independent conditioned on $\mathcal{H}(t_{n-1})$.

To establish the identical distribution, note that $\{(\hat{\boldsymbol{R}}(t),\hat{A}(t),\hat{D}(t))\}_{t=t_{n-1}+1}^{t_n}$ are identically distributed conditioned on $\mathcal{H}(t_{n-1})$. Given that $\hat{R}_i(t)V(t)=z_i(\{(\hat{R}(\tau),\hat{A}(\tau),\hat{D}(\tau))\}_{\tau=t-d_{\max}}^{t})$ and z_i is independent of t, we conclude that the random variables in $\Psi(q)$ are also identically distributed conditioned on $\mathcal{H}(t_{n-1})$.

With the conditional i.i.d. nature of the variables in $\Psi(q)$ for any q, we can apply the conditional Chernoff inequality (Lemma 5) to these variables to establish Theorem 1 using $\mathcal{F}=\mathcal{H}(t_{n-1})$ and $B=r_{\max}$. Specifically, by summing $\mathbf{E}\left[\hat{R}_i(t)\cdot V(t)\mid\mathcal{H}(t_{n-1})\right]$ over $t\in\{t(p;q)\}_{q=1}^{\Gamma}$, inequality (13) implies that, for any $\varepsilon>0$, with probability at least $1-22\delta_n$, the following inequality holds:

$$\sum_{p=1}^{\Gamma}\mathbf{E}\left[\hat{R}_i\left(t(p;q)\right)\cdot V\left(t\left(p;q\right)\right)\mid\mathcal{H}(t_{n-1})\right]=\nu(q)\geq\left(1-\frac{1}{(1+\varepsilon)^{C-1}}\cdot\exp\left[\frac{\varepsilon(C+\epsilon_n)}{1+g_n}\right]\right)^+\cdot\Gamma\cdot\lambda^* - \Gamma\cdot(\epsilon_n^A+\epsilon_n^B+\epsilon_n^C+\epsilon_n+f_n\lambda^*+2\varepsilon\cdot(d_{max}-C)^+\lambda^*)$$

for all $q \in \{1, \ldots, d_{\max}\}$. Note that $\nu(q) \leq r_{\max} \cdot \Gamma$ for all $q \in \{1, \ldots, d_{\max}\}$ with certainty. Using Lemma 5, we further deduce that, with probability at least $1 - \delta_n/d_{\max}$:

$$\sum_{p=1}^{\Gamma} \hat{R}_i(t(p;q)) \cdot V(t(p;q)) \geq \nu(q) - r_{\max}\sqrt{2\Gamma \log d_{\max}/\delta_n}, \text{ for any } q \in \{1, \ldots, d_{\max}\}. \tag{14}$$

Summing (14) over q and applying the union bound yield

$$\sum_{t=t_{n-1}+d_{\max}+1}^{t_n} \hat{R}_i(t) \cdot V(t) \geq \left(1 - \frac{1}{(1+\varepsilon)^{C-1}} \exp\left[\frac{\varepsilon(C+\epsilon_n)}{1+g_n}\right]\right)^+ (\ell_n - d_{\max})\lambda^* - r_{\max} d_{\max}\sqrt{2\Gamma \log d_{\max}/\delta_n}$$
$$- r_{\max} d_{\max} - (\ell_n - d_{\max}) \cdot (\epsilon_n^A + \epsilon_n^B + \epsilon_n^C + \epsilon_n + f_n\lambda^* + 2\varepsilon \cdot f_n \cdot (d_{max} - C)^+\lambda^*), \text{ w.p. } 1 - 23\delta_n. \tag{15}$$

Finally, by leveraging (11) that $\sum_{t_{n-1}+1}^{t_n} R_i(t) \geq \sum_{t_{n-1}+d_{\max}+1}^{t_n} \hat{R}_i(t) \cdot V(t)$, we establish Theorem 1. Theorems 2 and 3 are applications of Theorem 1, and we provide their proofs in the Appendix.

6 Extending to Quasi-full-feedback Setup

In this section, we extend our analysis to a Quasi-full-feedback setup where the controller can observe the outcomes for all possible actions if the resource is available (i.e., the resource is not blocked). This feedback setup, aligned with the setup in [17], is more informative than the bandit feedback setup but less comprehensive than the full-feedback setup. Notably, in this case, the controller can collect sufficient samples for all pairs $(j,k) \in \mathcal{J} \times \mathcal{K}$, thereby eliminating the need for forced exploration. As a result, we set $f_n = 0$ (i.e., $\rho = \infty$) in our algorithm, simplifying the exploration-exploitation trade-off. The following theorem demonstrates that, leveraging this richer feedback structure, our algorithm achieves a tighter performance guarantee.

Theorem 4. *Under the Quasi-full-feedback setup, our algorithm with $\rho = \infty$ ($f_n = 0$), $\gamma_1 = (\varepsilon^*(C) + 1)/C$, and $\gamma_2 = \varepsilon^*(C)$ ensures that*

$$\min_{i \in \mathcal{I}} \mathbf{E}\left[\frac{1}{T} \cdot \sum_{t=1}^{T} R_i(t)\right] \geq \left(1 - \frac{1}{2C}\right) \cdot \mathcal{L}(\varepsilon^*(C)) \cdot OPT^*$$
$$- O\Big(|\mathcal{J}| d_{\max}\sqrt{T \log T \cdot \log(|\mathcal{J}||\mathcal{K}||\mathcal{I}|)} + d_{\max}^3 + \sum_{j \in \mathcal{J}} \frac{1}{p_j} \frac{|\mathcal{K}|}{(1 - 0.5/C)\,\mathcal{L}(\epsilon^*(C))}\Big)/T.$$

Moreover, for the special case of $C \geq d_{\max}$, we have that (regret guarantee)

$$\min_{i \in \mathcal{I}} \mathbf{E}\left[\frac{1}{T} \cdot \sum_{t=1}^{T} R_i(t)\right] \geq OPT^* - O\Big(|\mathcal{J}| d_{\max}\sqrt{T \log T \cdot \log(|\mathcal{J}||\mathcal{K}||\mathcal{I}|)} + d_{\max}^3 + \sum_{j \in \mathcal{J}} \frac{|\mathcal{K}|}{p_j}\Big)/T.$$

The proof of Theorem 4 exploits that our system could receive a linear number of samples. Specifically, after t time steps (for sufficiently large t), the system obtains approximately $\frac{p_j}{|\mathcal{K}|}\left(1 - \frac{1}{2C}\right)\mathcal{L}(\epsilon^*(C)) \cdot t = O(t)$ samples for all pairs (j,k), as established in Lemma 4. Notably, when $C \geq d_{\max}$, this Quasi-full-feedback setup aligns with the full-feedback setup and our formulation reduces to the indivisible variant of the horizon-fairness optimization problem. This corresponds to the framework presented by [12], and we achieve the same regret order as in their work.

7 Conclusion

This paper introduces a general framework that unifies online reusable resource allocation with fairness optimization, addressing the challenges of balancing multiple performance metrics under model uncertainty and a bandit feedback setup. We develop an algorithm that achieves a competitive guarantee of $((1 - \frac{1}{2C})\mathcal{L}(\epsilon^*(C)), 2/3)$, in which the competitive ratio approaches optimality as the resource capacity increases. Furthermore, we show that under a Quasi-full-feedback setup, our

algorithm can achieve an improved competitive guarantee of $((1 - \frac{1}{2C})\mathcal{L}(\epsilon^*(C)), 1/2)$. This work lays the theoretical groundwork for developing more equitable computing systems.

Future research could explore more complex scenarios, such as situations where different requests consume varying amounts of resources. In addition, investigating tight lower bounds—particularly in the bandit feedback setting—remains an open problem. Establishing a rigorous, matching lower bound would not only complement the current upper-bound analysis but also provide deeper insights into the fundamental performance limits of online algorithms.

Acknowledgments

This work is supported by NSF CNS-2325956, CAREER-2045641, CPS-2136199, CNS-2102963, and CNS-2106299. We thank the anonymous reviewers and our shepherd Sean Sinclair for their valuable feedback.

References

[1] Shipra Agrawal and Nikhil R Devanur. Fast algorithms for online stochastic convex programming. In *Proceedings of the twenty-sixth annual ACM-SIAM symposium on Discrete algorithms*, pages 1405–1424. SIAM, 2014.

[2] Eitan Altman, Konstantin Avrachenkov, and Andrey Garnaev. Generalized α-fair resource allocation in wireless networks. In *2008 47th IEEE Conference on Decision and Control*, pages 2414–2419. IEEE, 2008.

[3] Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Linear stochastic bandits under safety constraints. *Advances in Neural Information Processing Systems*, 32, 2019.

[4] Fatih Aslan, George Iosifidis, Jose A Ayala-Romero, Andres Garcia-Saavedra, and Xavier Costa-Perez. Fair resource allocation in virtualized o-ran platforms. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 8(1):1–34, 2024.

[5] Jackie Baek and Will Ma. Bifurcating constraints to improve approximation ratios for network revenue management with reusable resources. *Operations Research*, 2022.

[6] Santiago Balseiro, Haihao Lu, and Vahab Mirrokni. Dual mirror descent for online allocation problems. In *International Conference on Machine Learning*, pages 613–628. PMLR, 2020.

[7] Santiago Balseiro, Haihao Lu, and Vahab Mirrokni. Regularized online allocation problems: Fairness and beyond. In *International Conference on Machine Learning*, pages 630–639. PMLR, 2021.

[8] Santiago R Balseiro and Shangzhou Xia. Uniformly bounded regret in dynamic fair allocation. *arXiv preprint arXiv:2205.12447*, 2022.

[9] Siddhartha Banerjee, Vasilis Gkatzelis, Artur Gorokh, and Billy Jin. Online nash social welfare maximization with predictions. In *Proceedings of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1–19. SIAM, 2022.

[10] Omar Besbes, Adam N Elmachtoub, and Yunjie Sun. Static pricing: Universal guarantees for reusable resources. *Operations Research*, 70(2):1143–1152, 2022.

[11] Thomas Bonald and James Roberts. Multi-resource fairness: Objectives, algorithms and performance. In *Proceedings of the 2015 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, pages 31–42, 2015.

[12] Semih Cayci, Swati Gupta, and Atilla Eryilmaz. Group-fair online allocation in continuous time. *Advances in Neural Information Processing Systems*, 33:13750–13761, 2020.

[13] Yiwei Chen, Retsef Levi, and Cong Shi. Revenue management of reusable resources with advanced reservations. *Production and Operations Management*, 26(5):836–859, 2017.

[14] Xiangxiang Dai, Jin Li, Xutong Liu, Anqi Yu, and John Lui. Cost-effective online multi-llm selection with versatile reward models. *arXiv preprint arXiv:2405.16587*, 2024.

[15] Nikhil R Devanur, Kamal Jain, Balasubramanian Sivan, and Christopher A Wilkens. Near optimal online algorithms and fast approximation algorithms for resource allocation problems. In *Proceedings of the 12th ACM conference on Electronic commerce*, pages 29–38, 2011.

[16] Nikhil R Devanur, Kamal Jain, Balasubramanian Sivan, and Christopher A Wilkens. Near optimal online algorithms and fast approximation algorithms for resource allocation problems. *Journal of the ACM (JACM)*, 66(1):1–41, 2019.

[17] Matthew Faw, Orestis Papadigenopoulos, Constantine Caramanis, and Sanjay Shakkottai. Learning to maximize welfare with a reusable resource. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 6(2):1–30, 2022.

[18] Yiding Feng, Rad Niazadeh, and Amin Saberi. Online assortment of reusable resources with exogenous replenishment. *Available at SSRN 3795056*, 2021.
[19] Yiding Feng, Rad Niazadeh, and Amin Saberi. Near-optimal bayesian online assortment of reusable resources. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pages 964–965, 2022.
[20] Xiao-Yue Gong, Vineet Goyal, Garud N Iyengar, David Simchi-Levi, Rajan Udwani, and Shuangyu Wang. Online assortment optimization with reusable resources. *Management Science*, 68(7):4772–4785, 2022.
[21] Vineet Goyal, Garud Iyengar, and Rajan Udwani. Asymptotically optimal competitive ratio for online allocation of reusable resources. *arXiv preprint arXiv:2002.02430*, 2020.
[22] Vineet Goyal, Garud Iyengar, and Rajan Udwani. Online allocation of reusable resources via algorithms guided by fluid approximations. *arXiv preprint arXiv:2010.03983*, 2020.
[23] Mohammad H Hajiesmaili. Trade-off analysis in learning-augmented algorithms with societal design criteria. *ACM SIGMETRICS Performance Evaluation Review*, 51(2):53–58, 2023.
[24] Yuelin Han, Zhifeng Wu, Pengfei Li, Adam Wierman, and Shaolei Ren. The unpaid toll: Quantifying the public health impact of ai. *arXiv preprint arXiv:2412.06288*, 2024.
[25] Walid Hanafy, Roozbeh Bostanedoost, Noman Bashir, David Irwin, Mohammad Hajiesmaili, and Prashant Shenoy. The war of the efficiencies: Understanding the tension between carbon and energy optimization. In *ACM Workshop on Hot Topics in Sustainable Computing*, 2023.
[26] Tianming Huo and Wang Chi Cheung. Online reusable resource allocations with multi-class arrivals. *Available at SSRN 4320423*, 2022.
[27] Tianming Huo and Wang Chi Cheung. Online reusable resource assortment planning with customer-dependent usage durations. *Available at SSRN 4504713*, 2023.
[28] Huiwen Jia, Cong Shi, and Siqian Shen. Online learning and pricing for network revenue management with reusable resources. *Advances in Neural Information Processing Systems*, 35:4830–4842, 2022.
[29] Huiwen Jia, Cong Shi, and Siqian Shen. Online learning and pricing for service systems with reusable resources. *Operations Research*, 72(3):1203–1241, 2024.
[30] Yasushi Kawase and Hanna Sumita. Online max-min fair allocation. In *International Symposium on Algorithmic Game Theory*, pages 526–543. Springer, 2022.
[31] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.
[32] Yanzhe Lei and Stefanus Jasin. Real-time dynamic pricing for revenue management with reusable resources, advance reservation, and deterministic service time requirements. *Operations Research*, 68(3):676–685, 2020.
[33] Fengjiao Li, Jia Liu, and Bo Ji. Combinatorial sleeping bandits with fairness constraints. *IEEE Transactions on Network Science and Engineering*, 7(3):1799–1813, 2019.
[34] Pengfei Li, Jianyi Yang, Adam Wierman, and Shaolei Ren. Towards environmentally equitable AI via geographical load balancing. In *Proceedings of the 15th ACM International Conference on Future and Sustainable Energy Systems*, pages 291–307, 2024.
[35] Luofeng Liao, Yuan Gao, and Christian Kroer. Nonstationary dual averaging and online fair allocation. *Advances in Neural Information Processing Systems*, 35:37159–37172, 2022.
[36] Qingsong Liu and Zhixuan Fang. Learning the optimal control for evolving systems with converging dynamics. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 8(2):1–39, 2024.
[37] Xin Liu, Bin Li, Pengyi Shi, and Lei Ying. An efficient pessimistic-optimistic algorithm for stochastic linear bandits with general constraints. *Advances in Neural Information Processing Systems*, 34:24075–24086, 2021.
[38] Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.
[39] Vishakha Patil, Ganesh Ghalme, Vineet Nair, and Yadati Narahari. Achieving fairness in the stochastic multi-armed bandit problem. *Journal of Machine Learning Research*, 22(174):1–31, 2021.
[40] Odile Pons. *Inequalities in analysis and probability*. World Scientific, 2021.
[41] Paat Rusmevichientong, Mika Sumida, and Huseyin Topaloglu. Dynamic assortment optimization for reusable products with random usage durations. *Management Science*, 66(7):2820–2844, 2020.
[42] Tareq Si Salem, Georgios Iosifidis, and Giovanni Neglia. Enabling long-term fairness in dynamic resource allocation. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 6(3):1–36, 2023.
[43] Sean R Sinclair, Siddhartha Banerjee, and Christina Lee Yu. Sequential fair allocation: Achieving the optimal envy-efficiency tradeoff curve. *ACM SIGMETRICS Performance Evaluation Review*, 50(1):95–96, 2022.
[44] Abhishek Sinha. Banditq: Fair bandits with guaranteed rewards. In *The 40th Conference on Uncertainty in Artificial Intelligence*, 2024.
[45] Abhishek Sinha, Ativ Joshi, Rajarshi Bhattacharjee, Cameron Musco, and Mohammad Hajiesmaili. No-regret algorithms for fair resource allocation. *Advances in Neural Information Processing Systems*, 36:48083–48109, 2023.
[46] Mohammad Sadegh Talebi and Alexandre Proutiere. Learning proportionally fair allocations with low regret. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 2(2):1–31, 2018.

[47] Shufan Wang, Guojun Xiong, and Jian Li. Online restless multi-armed bandits with long-term fairness constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 15616–15624, 2024.

[48] Wei Wang, Baochun Li, and Ben Liang. Dominant resource fairness in cloud computing systems with heterogeneous servers. In *IEEE INFOCOM 2014-IEEE Conference on Computer Communications*, pages 583–591. IEEE, 2014.

[49] Ziwei Wang, Jingtong Zhao, Yixuan Liu, and Jie Song. Reinforcement learning algorithm for reusable resource allocation with time-varying reward. *Available at SSRN 4680309*, 2023.

[50] Xilin Zhang and Wang Chi Cheung. Online resource allocation for reusable resources. *arXiv preprint arXiv:2212.02855*, 2022.

[51] Xuechen Zhang, Zijian Huang, Ege Onur Taga, Carlee Joe-Wong, Samet Oymak, and Jiasi Chen. Efficient contextual llm cascades through budget-constrained policy learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

[52] Yueyang Zhong, John R Birge, and Amy Ward. *Learning the scheduling policy in time-varying multiclass many server queues with abandonment*, volume 6. SSRN, 2022.

A Proofs

This section provides the proofs for all listed lemmas and theorems, with the exception of Theorem 1. Before delving into the proofs, we first introduce several supporting lemmas.

Lemma 6 (Azuma-Hoeffding Inequality [31]). *Let N be a positive integer and B be a positive real number. Suppose the random variables $X_1, \ldots, X_N$ constitute a martingale difference sequence with respect to the filtration $\{\mathcal{F}_n\}_{n=0}^N$, i.e. $\mathbf{E}[X_n|\mathcal{F}_{n-1}] = 0$ almost surely for every $n \in \{1, \ldots, N\}$. In addition, suppose $|X_n| \leq B$ almost surely for every $n \in \{1, \ldots, N\}$. For any $\delta \in (0, 1)$, it holds that*

$$\mathbf{P}\left(\left|\frac{1}{N}\sum_{n=1}^{N} X_n\right| \leq B\sqrt{\frac{2\log(2/\delta)}{N}}\right) \geq 1 - \delta.$$

Lemma 7 (Multiplicative Chernoff Inequality [31]). *Suppose random variables $\{X_t\}_{t=1}^N$ are independent, and that $X_t \in [0, B]$ for all $t \in \{1, \ldots, N\}$. Denote $\mu = \mathbf{E}[\sum_{t=1}^N X_t]$. The following concentration inequalities hold for any fixed but arbitrary $\delta \in (0, 1)$:*

$$\mathbf{P}\left(\sum_{t=1}^{N} X_t > \mu + 2\sqrt{B\mu\log\frac{1}{\delta}} + 2B\log\frac{1}{\delta}\right) \leq \delta, \quad \mathbf{P}\left(\sum_{t=1}^{N} X_t < \mu - \sqrt{2B\mu\log\frac{1}{\delta}}\right) \leq \delta.$$

Lemma 8 (Multiplicative Weight Update). *Let $\{\ell(s)\}_{s=1}^{\tau}$ be an arbitrary sequence of vectors, where $\ell(s) = (\ell_i(s))_{i\in\{0,1,\ldots,|\mathcal{I}|\}} \in [-B, B]^{|\mathcal{I}|+1}$ for each $s \in \{1, \ldots, \tau\}$. Consider the sequence of vectors $\vartheta(1), \ldots, \vartheta(\tau)$, where $\vartheta(s) = (\vartheta_i(s))_{i\in\{0,1,\ldots,|\mathcal{I}|\}} \in \Delta^{|\mathcal{I}|+1}$ is defined as*

$$\vartheta_i(s) = \frac{\exp\left[-\eta(s)\sum_{n=1}^{s-1}\ell_i(n)\right]}{\sum_{\iota\in\{0,1,\ldots,|\mathcal{I}|\}}\exp\left[-\eta(s)\sum_{n=1}^{s-1}\ell_\iota(n)\right]}, \text{ and } \eta(s) = \frac{\sqrt{\log(|\mathcal{I}|+1)}}{B\sqrt{s}} \tag{16}$$

for each $s \in \{1, \ldots \tau\}, i \in \{0, 1, ..., |\mathcal{I}|\}$. Then for any $i \in \{0, 1, ..., |\mathcal{I}|\}$, it holds that

$$\frac{1}{\tau}\sum_{s=1}^{\tau}\ell_i(s) \geq \frac{1}{\tau}\sum_{s=1}^{\tau}\sum_{\iota\in\{0,1,\ldots,|\mathcal{I}|\}}\vartheta_\iota(s)\ell_\iota(s) - 2B\sqrt{\frac{\log(|\mathcal{I}|+1)}{\tau}}.$$

The proof of this lemma can be found in Chapter 7.5 of [38].

A.1 Proof of Lemma 1

We begin by adopting a similar analysis from [19] to establish $\text{OPT}^* \leq \text{OPT}^{\text{R}}$ by showing that any optimal solution to (OFARR) is feasible to (OFARR-R). We then derive $\text{OPT}^{\text{R}} \leq \lambda_* + d_{\max} \cdot r_{\max}/T$ by constructing a feasible solution for the dual of (OFARR-S) using an optimal dual solution to (OFARR-R).

Recall that OPT^{R} is the optimal objective value to (OFARR-R), which is defined as follows,

$$\begin{aligned}
\text{(OFARR-R)} \quad \max\ & \lambda \\
\text{s.t.}\ & \sum_{t=1}^{T}\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} p_j \cdot r_{i,j,k} \cdot x_{j,k}(t) \geq T\cdot\lambda,\ \forall i \in \mathcal{I}, \\
& \sum_{\tau=1}^{t}\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} p_j \cdot \mathbf{P}(D_{j,k}(\tau) \geq t-\tau+1)\cdot x_{j,k}(t) \leq C,\ \forall t \in [T], \\
& \sum_{k\in\mathcal{K}} x_{j,k}(t) \leq 1,\ \forall j \in \mathcal{J},\ t \in [T]; \quad x_{j,k}(t) \geq 0,\ \forall k \in \mathcal{K},\ j \in \mathcal{J}.
\end{aligned} \tag{17}$$

To establish $\text{OPT}^{\text{R}} \geq \text{OPT}^*$, consider a policy π that achieves the optimum OPT^* for (OFARR), i.e., $\min_i \left\{\mathbf{E}\left[\frac{1}{T}\sum_{t=1}^{T}\sum_{k\in\mathcal{K}} R_{i,j(t),k}(t) Y_k^{\pi}(t)\right]\right\} = \text{OPT}^*$. Define $x_{j,k}(t) = \mathbf{P}(Y_k^{\pi}(t) = 1 | j(t) = j)$, and we claim that $\{x_{j,k}(t)\}_{j,k,t}$ is feasible to (OFARR-R) and the objective value of (OFARR-R) under $\{x_{j,k}(t)\}_{j,k,t}$ equals $\min_i \left\{\mathbf{E}\left[\frac{1}{T}\sum_{t=1}^{T}\sum_{k\in\mathcal{K}} R_{i,j(t),k}(t) Y_k^{\pi}(t)\right]\right\}$. Thus, verifying both the feasibility and equivalence of the objective value can establish $\text{OPT}^{\text{R}} \geq \text{OPT}^*$.

To verify the feasibility of $\{x_{j,k}(t)\}_{j,k,t}$ for (OFARR-R), note that the policy π satisfies the resource capacity constraints at any time. Specifically, the inequality $\sum_{\tau=1}^{t}\sum_{k\in\mathcal{K}} \mathbf{I}\{D_{j(\tau),k}(\tau) \geq t-\tau+1\} Y_k^{\pi}(\tau) \leq C$ holds for all $t \in [T]$. Taking the expectation of the left-hand side over $Y_k^{\pi}(\tau)$, $D_{j,k}(\tau)$, and $j(\tau)$ for $\tau = 1, \ldots, t$ yields:

$$\begin{aligned}
&\mathbf{E}\left[\sum_{\tau=1}^{t}\sum_{k\in\mathcal{K}} \mathbf{I}\left\{D_{j(\tau),k}(\tau) \geq t-\tau+1\right\} \cdot Y_k^{\pi}(\tau)\right] \\
=&\sum_{\tau=1}^{t}\sum_{k\in\mathcal{K}}\sum_{j\in\mathcal{J}} p_j \mathbf{E}\left[\mathbf{I}\left\{D_{j,k}(\tau) \geq t-\tau+1\right\}\right] \cdot x_{j,k}(\tau) \qquad (18)\\
=&\sum_{\tau=1}^{t}\sum_{k\in\mathcal{K}}\sum_{j\in\mathcal{J}} p_j \mathbf{P}\left(D_{j,k}(\tau) \geq t-\tau+1\right) \cdot x_{j,k}(\tau) \leq C.
\end{aligned}$$

Similarly, taking the expectation over the accumulated utility of each type $i \in \mathcal{I}$ yields:

$$\mathbf{E}\left[\sum_{t=1}^{T}\sum_{k\in\mathcal{K}} R_{i,j(t),k}(t) \cdot Y_k^{\pi}(t)\right] = \sum_{t=1}^{T}\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} p_j \cdot r_{i,j,k} \cdot x_{j,k}(t).$$

Hence, the claim regarding the equivalence of the objective value is verified.

To complete the proof of Lemma 1, we next demonstrate that $T\cdot\text{OPT}^{\text{R}} \leq T\cdot\lambda^* + r_{\max} d_{\max}$ holds with certainty. Since $D_{j,k}(t) \in [1, d_{\max}]$ for all j, k, t, the resource capacity constraints in (OFARR-R) can be equivalently written as $\sum_{\tau=\max\{t-d_{\max},1\}}^{t}\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} p_j \mathbf{P}\left(D_{j,k}(\tau) \geq t-\tau+1\right) x_{j,k}(\tau) \leq C,\ \forall t \in [T]$. Similarly, the resource capacity constraints in (OFARR-S) can be written as $\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}}\sum_{\tau=1}^{d_{\max}} p_j \cdot \mathbf{P}(D_{j,k}(\tau) \geq \tau) \cdot x_{j,k} \leq C$. Therefore, the dual of (OFARR-R) can be expressed as:

$$\begin{aligned}
&\text{(OFARR-R-D)}: \min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\rho}} \sum_{t=1}^{T}\left(\sum_{j\in\mathcal{J}} \beta_{j,t} + C\cdot\alpha_t\right)\\
&\text{s.t. } \beta_{j,t} + p_j \sum_{\tau=t}^{\min\{t+d_{\max}-1,T\}} \mathbf{P}\left(D_{j,k}(\tau) \geq \tau-t+1\right)\cdot\alpha_\tau - p_j\sum_{i\in\mathcal{I}} r_{i,j,k}\cdot\rho_i \geq 0,\ \forall j\in\mathcal{J},\ k\in\mathcal{K},\ t\in[T],\\
&\quad T\sum_{i\in\mathcal{I}}\rho_i \geq 1; \quad \rho_i \geq 0,\ \forall i\in\mathcal{I}; \quad \alpha_t \geq 0,\ \forall t\in[T]; \quad \beta_{j,t}\geq 0,\ \forall j\in\mathcal{J},\ t\in[T].
\end{aligned}$$

which is equivalent to

$$\begin{aligned}
&\text{(OFARR-R-D)}: \min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\rho}} \sum_{t=1}^{T}\left(\sum_{j\in\mathcal{J}} p_j\cdot\beta_{j,t} + C\cdot\alpha_t\right)\\
&\text{s.t. } \beta_{j,t} + \sum_{\tau=t}^{\min\{t+d_{\max}-1,T\}} \mathbf{P}\left(D_{j,k}(\tau) \geq \tau-t+1\right)\cdot\alpha_\tau - \sum_{i} r_{i,j,k}\cdot\rho_i \geq 0,\ \forall j\in\mathcal{J},\ k\in\mathcal{K},\ t\in[T],\\
&\quad T\sum_{i\in\mathcal{I}}\rho_i \geq 1; \quad \rho_i \geq 0,\ \forall i\in\mathcal{I}; \quad \alpha_t \geq 0,\ \forall t\in[T]; \quad \beta_{j,t}\geq 0,\ \forall j\in\mathcal{J},\ t\in[T].
\end{aligned}$$

Similarly, the dual of (OFARR-S) is given by:

$$\begin{aligned}(\text{OFARR-S-D}): \min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\rho}} \ & \sum_{j\in\mathcal{J}} p_j\cdot\beta_j + C\cdot\alpha \\ \text{s.t. } & \beta_j + \alpha\cdot d_{j,k} - \sum_{i\in\mathcal{I}} r_{i,j,k}\cdot\rho_i \geq 0, \quad \forall j\in\mathcal{J},\ k\in\mathcal{K}, \\ & \sum_{i\in\mathcal{I}}\rho_i \geq 1; \quad \rho_i\geq 0,\ \forall i\in\mathcal{I}; \quad \alpha\geq 0; \quad \beta_j\geq 0,\ \forall j\in\mathcal{J}.\end{aligned}$$

Let $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \boldsymbol{\rho}^*)$ be an optimal solution to (OFARR-S-D). Define the solution $(\bar{\boldsymbol{\alpha}} = (\bar{\alpha}_t)_t, \bar{\boldsymbol{\beta}} = (\bar{\beta}_{j,t})_{j,t}, \bar{\boldsymbol{\rho}} = (\bar{\rho}_i)_i)$ as follows,

$$\begin{aligned}\bar{\alpha}_t &= \alpha^*/T && \text{for all } t\in\{1,\dots,T\},\\ \bar{\beta}_{j,t} &= \begin{cases}\beta_j^*/T & \text{for all } j\in\mathcal{J}, t\in\{1,\dots,T-d_{\max}\},\\ r_{\max}/T & \text{for all } j\in\mathcal{J}, t\in\{T-d_{\max}+1,\dots,T\},\end{cases}\\ \bar{\rho}_i &= \rho_i^*/T && \text{for all } i\in\mathcal{I}.\end{aligned}$$

We claim that the solution $(\bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{\beta}}, \bar{\boldsymbol{\rho}})$ is feasible to (OFARR-R-D). Note that the objective value of (OFARR-R-D) under $(\bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{\beta}}, \bar{\boldsymbol{\rho}})$ can be bounded as

$$\frac{r_{\max}d_{\max}}{T} + \sum_{j\in\mathcal{J}}\left(1-\frac{d_{\max}}{T}\right)p_j\beta_j^* + C\cdot\alpha^* \leq \frac{r_{\max}d_{\max}}{T} + \lambda^*.$$

Thus, to establish the bound $T\cdot\mathrm{OPT}^{\mathrm{R}} \leq T\cdot\lambda^* + d_{\max}r_{\max}$, it suffices to show the feasibility of $(\bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{\beta}}, \bar{\boldsymbol{\rho}})$ to (OFARR-R-D). Firstly, by the optimality of $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \boldsymbol{\rho}^*)$ to (OFARR-S-D), we have $\sum_{i\in\mathcal{I}}\rho_i^* = 1$. Therefore, it is sufficient to verify the feasibility of the first set of constraints in (OFARR-R-D). For the case where $t\in\{T-d_{\max}+1,\dots,T\}$, the constraints are satisfied since $\bar{\beta}_{j,t} = r_{\max}/T$ while $\sum_{i\in\mathcal{I}} r_{i,j,k}\cdot\rho_i \leq r_{\max}/T$. For the another case where $t\in\{1,\dots,T-d_{\max}\}$, the constraints remain feasible because

$$\begin{aligned}&\bar{\beta}_{j,t} + \sum_{\tau=t}^{\min\{t+d_{\max},T\}} \mathbf{P}\left(D_{j,k}(\tau)\geq\tau-t+1\right)\bar{\alpha}_\tau - \sum_{i\in\mathcal{I}} r_{i,j,k}\cdot\bar{\rho}_i\\ =&\frac{1}{T}\left[\beta_j^* + \sum_{\tau=t}^{t+d_{\max}}\mathbf{P}\left(D_{j,k}(\tau)\geq\tau-t+1\right)\alpha_\tau^* - \sum_{i\in\mathcal{I}} r_{i,j,k}\cdot\rho_i^*\right]\\ =&\frac{1}{T}\left[\beta_j^* + \sum_{\tau=1}^{d_{\max}}\mathbf{P}\left(D_{j,k}(\tau)\geq\tau\right)\alpha_\tau^* - \sum_{i\in\mathcal{I}} r_{i,j,k}\cdot\rho_i^*\right]\\ =&\frac{1}{T}\left[\beta_j^* + d_{j,k}\cdot\alpha_\tau^* - \sum_{i\in\mathcal{I}} r_{i,j,k}\cdot\rho_i^*\right]\geq 0.\end{aligned}$$

For the special case where $C\geq d_{\max}$, the resource capacity constraints in both (OFARR-R) and (OFARR-S) can be safely ignored, as they are guaranteed to hold for all $t\in[T]$. Consequently,

(OFARR-R) and (OFARR-S) simplify to the following forms, respectively:

$$\text{(OFARR-R)} \quad \max\ \lambda \qquad \text{s.t.} \sum_{t=1}^{T}\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} p_j\cdot r_{i,j,k}\cdot x_{j,k}(t)\geq T\cdot\lambda, \quad \forall i\in\mathcal{I},$$
$$\sum_{k\in\mathcal{K}} x_{j,k}(t)\leq 1, \quad \forall j\in\mathcal{J},\ t\in[T]; \qquad x_{j,k}(t)\geq 0, \quad \forall k\in\mathcal{K},\ j\in\mathcal{J},\ t\in[T].$$
$$\text{(OFARR-S)} \quad \max\ \lambda \qquad \text{s.t.} \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} p_j\cdot r_{i,j,k}\cdot y_{j,k}\geq \lambda, \quad \forall i\in\mathcal{I},$$
$$\sum_{k\in\mathcal{K}} y_{j,k}\leq 1, \quad \forall j\in\mathcal{J}; \qquad y_{j,k}\geq 0, \quad \forall k\in\mathcal{K},\ j\in\mathcal{J}.$$

Denote $\{x_{j,k}(t)\}_{j,k,t}$ and $\boldsymbol{y}^*$ as the optimal solutions to (OFARR-R) and (OFARR-S), respectively, when $C\geq d_{\max}$. It follows that $\lambda^*\leq \mathrm{OPT}^{\mathrm{R}}$, as the solution $\boldsymbol{x}$, defined by $x_{j,k}(t)=y^*_{j,k}$, is feasible to (OFARR-R). The corresponding objective value of $\boldsymbol{x}$ under (OFARR-R) is exactly λ^*. Conversely, $\lambda^*\geq \mathrm{OPT}^{\mathrm{R}}$ holds because the objective value of (OFARR-S) under $\boldsymbol{y}$, where $y_{j,k}=1/T\cdot\sum_{t=1}^{T}x^*_{j,k}(t)$, equals $\mathrm{OPT}^{\mathrm{R}}$, and $\boldsymbol{y}$ is feasible to (OFARR-S). Therefore, we conclude that $\lambda^*=\mathrm{OPT}^{\mathrm{R}}$ when $C\geq d_{\max}$. Put all the things together completes the proof of Lemma 1.

A.2 Proof of Lemma 2

To establish that $\lambda^*\leq\hat{\lambda}^*(n)+\epsilon_n^A+\epsilon_n^B$ holds w.p. $1-4\delta_n$, we construct the following solution $\hat{\boldsymbol{x}}(n)$ for (OFARR-S)(n),

$$\hat{x}_{j,k}(n)=\frac{1}{1+\epsilon_n^A}\cdot x^*_{j,k},\ \forall j\in\mathcal{J},\ k\in\mathcal{K},$$

where $\boldsymbol{x}^*$ is an optimal solution to (OFARR-S). We claim that with probability at least $1-2\delta_n$, $\hat{\boldsymbol{x}}(n)$ is feasible to (OFARR-S)(n). Furthermore, it holds that with probability at least $1-2\delta_n$:

$$\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}}\tilde{p}_j(n)\cdot\hat{r}_{i,j,k}(n)\cdot\hat{x}_{j,k}(n)\geq\lambda^*-\epsilon_n^B-\epsilon_n^A. \tag{20}$$

These two properties clearly imply that inequality $\lambda^*-\epsilon_n^A-\epsilon_n^B\leq\hat{\lambda}^*(n)$ holds with probability at least $1-4\delta_n$. Thus, our focus shifts to establishing the feasibility of $\hat{\boldsymbol{x}}(n)$ and verifying (20).

To verify the feasibility, we apply the multiplicative Chernoff inequality (Lemma 7) to $X_t=\sum_{k\in\mathcal{K}}d_{j(t),k}\cdot x^*_{j(t),k}$ for $t\in\{1,\ldots,t_{n-1}\}$. Observe that $\frac{1}{t_{n-1}}\sum_{s=1}^{t_{n-1}}X_s=\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}}\tilde{p}_j(n)\cdot d_{j,k}\cdot x^*_{j,k}$ and $\mathbf{E}[\frac{1}{t_{n-1}}\sum_{s=1}^{t_{n-1}}X_s]=\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}}p_j\cdot d_{j,k}\cdot x^*_{j,k}$. Thus, with probability at least $1-\delta_n$, it holds that

$$\begin{aligned}
&\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}}\tilde{p}_j(n)\cdot d_{j,k}\cdot x^*_{j,k}\\
\leq&\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}}p_j\cdot d_{j,k}\cdot x^*_{j,k}+\sqrt{\frac{4d_{\max}}{t_{n-1}}\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}}p_j\cdot d_{j,k}\cdot x^*_{j,k}\cdot\log\frac{1}{\delta_n}}+\frac{2d_{\max}}{t_{n-1}}\log\frac{1}{\delta_n}\\
\leq&C+\sqrt{\frac{4d_{\max}C}{t_{n-1}}\log\frac{1}{\delta_n}}+\frac{2d_{\max}}{t_{n-1}}\log\frac{1}{\delta_n}\\
=&C\left(1+2\sqrt{\frac{d_{\max}}{C\cdot t_{n-1}}\log\frac{1}{\delta_n}}+\frac{2d_{\max}}{C\cdot t_{n-1}}\log\frac{1}{\delta_n}\right)=C\left(1+\epsilon_n^A\right).
\end{aligned}$$

Additionally, the concentration inequality for the LCB estimate $\hat{\boldsymbol{d}}(n)$ implies that

$$\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}}\tilde{p}_j(n)\cdot\hat{d}_{j,k}(n)\cdot x^*_{j,k}\le\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}}\tilde{p}_j(n)\cdot d_{j,k}\cdot x^*_{j,k},\quad\text{w.p. } 1-\delta_n.$$

Then according to the definition of $\hat{\boldsymbol{x}}(n)$ we can derive that $\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}}\tilde{p}_j(n)\cdot\hat{d}_{j,k}(n)\cdot\hat{x}_{j,k}(n)=\frac{1}{1+\epsilon_n^A}\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}}\tilde{p}_j(n)\cdot\hat{d}_{j,k}(n)\cdot x^*_{j,k}\le C$ holds with probability at least $1-2\delta_n$. This indicates that $\hat{\boldsymbol{x}}(n)$ is feasible to (OFARR-S)(n) with probability at least $1-2\delta_n$.

To establish (20), we apply the multiplicative Chernoff inequality again to $X_t=\sum_{k\in\mathcal{K}}r_{i,j(t),k}\cdot x^*_{j(t),k}$ for $t\in\{1,\dots,t_{n-1}\}$, which yields:

$$\begin{aligned}
&\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}}\tilde{p}_j(n)\cdot r_{i,j,k}\cdot x^*_{j,k}\\
&\ge\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}}p_j\cdot r_{i,j,k}\cdot x^*_{j,k}-\sqrt{\frac{2r_{\max}\cdot\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}}p_j\cdot r_{i,j,k}\cdot x^*_{j,k}\cdot\log(|\mathcal{I}|/\delta_n)}{t_{n-1}}}\quad\text{w.p. } 1-\delta_n/|\mathcal{I}|\\
&\ge\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}}p_j\cdot r_{i,j,k}\cdot x^*_{j,k}-r_{\max}\sqrt{\frac{2\log(|\mathcal{I}|/\delta_n)}{t_{n-1}}}\quad\text{w.p. } 1-\delta_n/|\mathcal{I}|\\
&\ge\lambda^*-r_{\max}\sqrt{\frac{2\log(|\mathcal{I}|/\delta_n)}{t_{n-1}}}=\lambda^*-\epsilon_n^B,\quad\text{w.p. } 1-\delta_n/|\mathcal{I}|,
\end{aligned}$$

The concentration bound for $\hat{\boldsymbol{r}}(n)$ further implies that for any $i\in\mathcal{I}$:

$$\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}}\tilde{p}_j(n)\cdot\hat{r}_{i,j,k}(n)\cdot x^*_{j,k}\ge\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}}\tilde{p}_j(n)\cdot r_{i,j,k}\cdot x^*_{j,k}\quad\text{w.p. } 1-\delta_n/|\mathcal{I}|.$$

Combining these results, we establish that for all $i\in\mathcal{I}$

$$\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}}\tilde{p}_j(n)\cdot\hat{r}_{i,j,k}(n)\cdot y^*_{j,k}\ge\lambda^*-\epsilon_n^B,\quad\text{w.p. } 1-2\delta_n.$$

Finally, the inequality (20) can be derived using the definition of $\hat{\boldsymbol{x}}(n)$:

$$\begin{aligned}
\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}}\tilde{p}_j(n)\cdot\hat{r}_{i,j,k}(n)\cdot\hat{x}_{j,k}(n)&=\frac{1}{1+\epsilon_n^A}\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}}\tilde{p}_j(n)\cdot\hat{r}_{i,j,k}(n)\cdot x^*_{j,k}\\
&\ge\frac{\lambda^*-\epsilon_n^B}{1+\epsilon_n^A}\ge\lambda^*-\epsilon_n^B-\epsilon_n^A,
\end{aligned}$$

where the last inequality holds since $\lambda^*\le r_{\max}\le 1+\epsilon_n^A+\epsilon_n^B$.

Thus, the remaining thing is to prove $\mathbf{P}(\hat{\lambda}^*(n)-\epsilon_n^C\le\lambda^*)\ge 1-3\delta_n$. Recall that the dual of (OFARR-S) is given by

$$\begin{aligned}
\text{(OFARR-S-D)}\ :\ \min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\rho}}\ &\sum_{j\in\mathcal{J}}p_j\beta_j+C\alpha\\
\text{s.t. }&\beta_j+d_{j,k}\cdot\alpha-\sum_{i\in\mathcal{I}}r_{i,j,k}\cdot\rho_i\ge 0,\ \forall j\in\mathcal{J},\ k\in\mathcal{K}\\
&\sum_{i\in\mathcal{I}}\rho_i\ge 1;\quad\rho_i\ge 0,\ \forall i\in\mathcal{I};\quad\alpha\ge 0,\ \beta_j\ge 0,\ \forall j\in\mathcal{J}.
\end{aligned}$$

Similarly, the dual of (OFARR-S)(n) is as follows:

$$\begin{aligned}(\text{OFARR-S-D})(n): \min_{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\rho}}} & \sum_{j\in\mathcal{J}} \tilde{p}_j(n)\cdot\hat{\beta}_j + C\cdot\hat{\alpha}\\ \text{s.t. } & \hat{\beta}_j + \hat{d}_{j,k}(n)\cdot\hat{\alpha} - \sum_{i\in\mathcal{I}} \hat{r}_{i,j,k}(n)\cdot\hat{\rho}_i \geq 0,\ \forall j\in\mathcal{J},\ k\in\mathcal{K}\\ & \sum_{i\in\mathcal{I}}\hat{\rho}_i \geq 1;\quad \hat{\rho}_i\geq 0,\ \forall i\in\mathcal{I};\quad \hat{\alpha}\geq 0,\ \hat{\beta}_j\geq 0,\ \forall j\in\mathcal{J}.\end{aligned}$$

Since the feasible domains of (OFARR-S-D) and (OFARR-S-D)(n) are identical, any optimal solution to (OFARR-S-D) must also be feasible to (OFARR-S-D)(n). Assume $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \boldsymbol{\rho}^*)$ is an optimal solution to (OFARR-S-D) and then it holds that

$$\lambda^* = \sum_{j\in\mathcal{J}} p_j\beta_j^* + C\alpha^*,\quad \sum_{i\in\mathcal{I}}\rho_i^* = 1,\quad \beta_j^* = \sum_{i\in\mathcal{I}} r_{i,j,k}\cdot\rho_i^* - d_{j,k}\cdot\alpha^*,\ \forall j\in\mathcal{J},\ k\in\mathcal{K}.$$

Now we claim that the constructed solution $(\hat{\boldsymbol{\alpha}}^*, \hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\rho}}^*)$, defined as

$$\begin{aligned}&\hat{\alpha}^* = \alpha^*,\\ &\hat{\beta}_j^* = \beta_j^* + 2(r_{\max}+d_{\max})\sqrt{\frac{2\log(|\mathcal{J}||\mathcal{K}||\mathcal{I}|/\delta_n)}{\min_{k\in\mathcal{K}} N_{j,k}(t_{n-1})}},\quad \forall j\in\mathcal{J},\\ &\hat{\rho}_i^* = \rho_i^*,\quad \forall i\in\mathcal{I}.\end{aligned}$$

is feasible to (OFARR-S-D)(n) with probability as least $1-2\delta_n$. To prove this claim, note that $\sum_{i\in\mathcal{I}}\rho_i^* = 1$ and $\beta_j^* = \sum_{i\in\mathcal{I}} r_{i,j,k}\cdot\rho_i^* - d_{j,k}\cdot\alpha^* \geq 0, \forall j\in\mathcal{J},\ k\in\mathcal{K}$, which together imply that $\alpha^*\leq r_{\max}$. Hence, the following inequality holds that for any pair (j,k):

$$\begin{aligned}&\hat{\beta}_j^* + \hat{d}_{j,k}(n)\cdot\hat{\alpha}^* - \sum_{i\in\mathcal{I}}\hat{r}_{i,j,k}(n)\cdot\hat{\rho}_i^*\\ &= \beta_j + (2r_{\max}+2d_{\max})\sqrt{\frac{2\log(|\mathcal{J}||\mathcal{K}||\mathcal{I}|/\delta_n)}{\min_{k\in\mathcal{K}} N_{j,k}(t_{n-1})}} + \hat{d}_{j,k}(n)\cdot\alpha^* - \sum_{i\in\mathcal{I}}\hat{r}_{i,j,k}(n)\cdot\rho_i^*\\ &\overset{(a)}{\geq} \beta_j^* + (2r_{\max}+2d_{\max})\sqrt{\frac{2\log(|\mathcal{J}||\mathcal{K}||\mathcal{I}|/\delta_n)}{\min_{k\in\mathcal{K}} N_{j,k}(t_{n-1})}}\\ &\quad + \left(d_{j,k} - 2d_{\max}\sqrt{\frac{2\log(|\mathcal{J}||\mathcal{K}|/\delta_n)}{N_{j,k}(t_{n-1})}}\right)\cdot\alpha^* - \sum_{i\in\mathcal{I}}\hat{r}_{i,j,k}(n)\cdot\rho_i^*\quad \text{w.p. } 1-\delta_n/(|\mathcal{J}||\mathcal{K}|)\\ &\overset{(b)}{\geq} \beta_j^* + (2r_{\max}+2d_{\max})\sqrt{\frac{2\log(|\mathcal{J}||\mathcal{K}||\mathcal{I}|/\delta_n)}{\min_{k\in\mathcal{K}} N_{j,k}(t_{n-1})}} + \left(d_{j,k} - 2d_{\max}\sqrt{\frac{2\log(|\mathcal{J}||\mathcal{K}|/\delta_n)}{N_{j,k}(t_{n-1})}}\right)\cdot\alpha^*\\ &\quad - \sum_{i\in\mathcal{I}}\left(r_{i,j,k} + 2r_{\max}\sqrt{\frac{2\log(|\mathcal{J}||\mathcal{K}||\mathcal{I}|/\delta_n)}{N_{j,k}(t_{n-1})}}\right)\cdot\rho_i^*\quad \text{w.p. } 1-2\delta_n/(|\mathcal{J}||\mathcal{K}|)\\ &\geq \beta_j^* + (2r_{\max}+2d_{\max})\sqrt{\frac{2\log(|\mathcal{J}||\mathcal{K}||\mathcal{I}|/\delta_n)}{\min_{k\in\mathcal{K}} N_{j,k}(t_{n-1})}} + d_{j,k}\cdot\alpha^* - \sum_{i\in\mathcal{I}} r_{i,j,k}\cdot\rho_i^*\\ &\quad - 2d_{\max}\sqrt{\frac{2\log(|\mathcal{J}||\mathcal{K}||\mathcal{I}|/\delta_n)}{N_{j,k}(t_{n-1})}}\cdot\alpha^* - 2r_{\max}\sqrt{\frac{2\log(|\mathcal{J}||\mathcal{K}|/\delta_n)}{N_{j,k}(t_{n-1})}}\sum_{i\in\mathcal{I}}\rho_i^*\quad \text{w.p. } 1-2\delta_n/(|\mathcal{J}||\mathcal{K}|)\end{aligned}$$

where (a) and (b) hold due to the concentration bounds for $\hat{\boldsymbol{d}}(n)$ and $\hat{\boldsymbol{r}}(n)$, respectively. Further leveraging the facts that $\sum_i \rho_i^* = 1$ and $\alpha^* \le r_{\max}$, we have

$$
\begin{aligned}
&\hat{\beta}_j^* + \hat{d}_{j,k}(n)\cdot\hat{\alpha}^* - \sum_{i\in\mathcal{I}} \hat{r}_{i,j,k}(n)\cdot\hat{\rho}_i^* \\
\ge\ & \beta_j^* + (2r_{\max}+2d_{\max})\sqrt{\frac{2\log(|\mathcal{J}||\mathcal{K}||\mathcal{I}|/\delta_n)}{\min_{k\in\mathcal{K}} N_{j,k}(t_{n-1})}} + d_{j,k}\cdot\alpha^* - \sum_{i\in\mathcal{I}} r_{i,j,k}\cdot\rho_i^* \\
&- 2d_{\max}\sqrt{\frac{2\log(|\mathcal{J}||\mathcal{K}||\mathcal{I}|/\delta_n)}{N_{j,k}(t_{n-1})}} - 2r_{\max}\sqrt{\frac{2\log(|\mathcal{J}||\mathcal{K}|/\delta_n)}{N_{j,k}(t_{n-1})}} \quad \text{w.p. } 1-2\delta_n/(|\mathcal{J}||\mathcal{K}|) \\
\ge\ & \beta_j^* + d_{j,k}\cdot\alpha^* - \sum_{i\in\mathcal{I}} r_{i,j,k}\cdot\rho_i^* \ge 0, \quad \text{w.p. } 1-2\delta_n/(|\mathcal{J}||\mathcal{K}|).
\end{aligned}
$$

This completes the proof that $(\hat{\boldsymbol{\alpha}}^*, \hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\rho}}^*)$ is feasible to (OFARR-S-D)(n) with probability as least $1-2\delta_n$. Next, we show that the objective value of (OFARR-S-D)(n) achieved by $(\hat{\boldsymbol{\alpha}}^*, \hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\rho}}^*)$ has a gap of at most ϵ_n^C compared to λ^*, with high-probability.

Firstly, by the definition of $\hat{\lambda}^*(n)$, it is upper bounded by the objective value achieved by $(\hat{\boldsymbol{\alpha}}^*, \hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\rho}}^*)$ in (OFARR-S-D)(n). Hence, we have that

$$
\begin{aligned}
\hat{\lambda}^*(n) \le& \sum_{j\in\mathcal{J}} \tilde{p}_j(n)\cdot\hat{\beta}_j^* + C\cdot\hat{\alpha}^* \\
\le& \sum_{j\in\mathcal{J}} \tilde{p}_j(n)\cdot\beta_j^* + C\cdot\alpha^* + \sum_{j\in\mathcal{J}} 2\tilde{p}_j(n)(r_{\max}+d_{\max})\sqrt{\frac{2\log(|\mathcal{J}||\mathcal{K}||\mathcal{I}|/\delta_n)}{\min_{k\in\mathcal{K}} N_{j,k}(t_{n-1})}} \\
=& \sum_{j\in\mathcal{J}} \tilde{p}_j(n)\cdot\beta_j^* + C\cdot\alpha^* + \sum_{j\in\mathcal{J}} 2(r_{\max}+d_{\max})\sqrt{\frac{2\log(|\mathcal{J}||\mathcal{K}||\mathcal{I}|/\delta_n)}{\min_{k\in\mathcal{K}} N_{j,k}(t_{n-1})}} \\
\overset{(a)}{\le}& \sum_{j\in\mathcal{J}} p_j\beta_j^* + 2\sqrt{\frac{r_{\max}(\sum_{j\in\mathcal{J}} p_j\beta_j^*)\log(1/\delta_n)}{t_{n-1}}} + 2\frac{r_{\max}\log(1/\delta_n)}{t_{n-1}} \\
&+ C\cdot\alpha^* + \sum_{j\in\mathcal{J}} 2(r_{\max}+d_{\max})\sqrt{\frac{2\log(|\mathcal{J}||\mathcal{K}||\mathcal{I}|/\delta_n)}{\min_{k\in\mathcal{K}} N_{j,k}(t_{n-1})}} \quad \text{w.p. } 1-\delta_n \\
=& \lambda^* + 2\sqrt{\frac{r_{\max}(\sum_{j\in\mathcal{J}} p_j\beta_j^*)\log(1/\delta_n)}{t_{n-1}}} + 2\frac{r_{\max}\log(1/\delta_n)}{t_{n-1}} \\
&+ \sum_{j\in\mathcal{J}} 2(r_{\max}+d_{\max})\sqrt{\frac{2\log(|\mathcal{J}||\mathcal{K}||\mathcal{I}|/\delta_n)}{\min_{k\in\mathcal{K}} N_{j,k}(t_{n-1})}} \\
\le& \lambda^* + 2r_{\max}\left[\sqrt{\frac{\log(1/\delta_n)}{t_{n-1}}} + \frac{\log(1/\delta_n)}{t_{n-1}}\right] + \sum_{j\in\mathcal{J}} 2(r_{\max}+d_{\max})\sqrt{\frac{2\log(|\mathcal{J}||\mathcal{K}||\mathcal{I}|/\delta_n)}{\min_{k\in\mathcal{K}} N_{j,k}(t_{n-1})}} \\
\le& \lambda^* + 2r_{\max}\left[\sqrt{\frac{\log(1/\delta_n)}{t_{n-1}}} + \frac{\log(1/\delta_n)}{t_{n-1}}\right] + 2|\mathcal{J}|(r_{\max}+d_{\max})\sqrt{\frac{2\log(|\mathcal{J}||\mathcal{K}||\mathcal{I}|/\delta_n)}{\min_{j\in\mathcal{J},k\in\mathcal{K}} N_{j,k}(t_{n-1})}} \\
=& \lambda_* + \epsilon_n^C.
\end{aligned}
$$

Step (a) applies the multiplicative Chernoff inequality, leveraging the key fact that $\beta_j^* \le r_{\max}$. This holds because the condition $\sum_{i\in\mathcal{I}} \rho_i^* = 1$ is satisfied due to the optimality of the solution $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \boldsymbol{\rho}^*)$ to (OFARR-S-D). Combining these arguments establishes the following bound

$$\hat{\lambda}^*(n) \le \epsilon_n^C + \lambda^*, \quad \text{w.p. } 1 - 3\delta_n,$$

which completes the proof of Lemma 2.

A.3 Proof of Lemma 3

The proof critically relies on a pivotal application of Lemma 8, which serves as the foundation for constructing Algorithm 2. Specifically, for each $s \in \{1, \ldots, \ell_n\}$, we define

$$f_i(s) = \begin{cases} \hat{r}_{i,j^v(s),k^v(s)}(n) - (\hat{\lambda}^*(n) - \epsilon_n^C) & \text{if } i = 1, 2, ..., |\mathcal{I}|, \\ -\hat{d}_{j^v(s),k^v(s)}(n) + C & \text{if } i = 0. \end{cases} \tag{21}$$

Since $\hat{r}_{i,j,k}(n) \le r_{\max}$ and $\hat{d}_{j,k}(n) \le d_{\max}$, it follows that $|f_i(s)| \le \max\{r_{\max}, d_{\max}\}$ for all $i = 0, 1, ..., |\mathcal{I}|$ and $s \in \{1, \ldots, \ell_n\}$. Moreover, under the specification of $\{\boldsymbol{f}(s)\}_{s=1}^{\ell_n}$ in (21), it can be directly verified that the MWU weigh vector $\vartheta(s)$ in lemma 8 is equal to $\boldsymbol{\phi}_n(s)$ for each $s \in \{1, \ldots, \ell_n\}$, where $\{\boldsymbol{\phi}_n(s)\}_{s=1}^{\ell_n}$ are the weight vectors in $\Theta(n)$. Applying Lemma 8 under these conditions yields the following inequalities (which simultaneously hold with certainty):

$$\left[\frac{1}{\ell_n}\sum_{s=1}^{\ell_n} \hat{r}_{i,j^v(s),k^v(s)}(n)\right] - \left(\hat{\lambda}^*(n) - \epsilon_n^C\right) \ge \Phi(n) - r_{\max}\sqrt{\frac{8\log(|\mathcal{I}|+1)}{\ell_n}} \text{ for each } i \in |\mathcal{I}|, \tag{22}$$

$$-\left[\frac{1}{\ell_n}\sum_{s=1}^{\ell_n} \hat{d}_{j^v(s),k^v(s)}(n)\right] + C \ge \Phi(n) - d_{\max}\sqrt{\frac{8\log(|\mathcal{I}|+1)}{\ell_n}}, \tag{23}$$

where

$$\Phi(n) = \frac{1}{\ell_n}\sum_{s=1}^{\ell_n}\left[\sum_{\iota\in\{1,\ldots,\mathcal{I}\}} \phi_{n,\iota}(s)\left(\hat{r}_{\iota,j^v(s),k^v(s)}(n) - (\hat{\lambda}^*(n) - \epsilon_n^C)\right) + \phi_{n,0}(s)\left(-\hat{d}_{j^v(s),k^v(s)}(n) + C\right)\right]. \tag{24}$$

Next, we proceed to prove the following three inequalities:

$$\mathbf{P}\left(\Phi(n) \ge -r_{\max}\sqrt{\frac{2}{\ell_n}\log\frac{2}{\delta_n}} - d_{\max}\sqrt{\frac{2}{\ell_n}\log\frac{2}{\delta_n}} - r_{\max}\sqrt{\frac{2}{t_n}\log\frac{2}{\delta_n}} - d_{\max}\sqrt{\frac{2}{t_n}\log\frac{2}{\delta_n}}\right) \ge 1 - 9\delta_n, \tag{25}$$

$$\mathbf{P}\left(\sum_{s=1}^{\ell_n} r_{i,j^v(s),k^v(s)}(n) \le \sum_{s=1}^{\ell_n}\sum_{j\in\mathcal{J}} p_j \cdot r_{i,j,\kappa_n(\boldsymbol{\phi}_n(s),j)}(n) + 2r_{\max}\sqrt{2\ell_n\log\frac{|\mathcal{I}|}{\delta_n}}\right) \ge 1 - 2\frac{\delta_n}{|\mathcal{I}|}, \tag{26}$$

$$\mathbf{P}\left(\sum_{s=1}^{\ell_n} d_{j^v(s),k^v(s)}(n) \ge \sum_{s=1}^{\ell_n}\sum_{j\in\mathcal{J}} p_j \cdot d_{j,\kappa_n(\boldsymbol{\phi}_n(s),j)}(n) - 2d_{\max}\sqrt{2\ell_n\log\frac{1}{\delta_n}}\right) \ge 1 - 2\delta_n. \tag{27}$$

To establish (25), we assume that $\boldsymbol{x}^* = \{x^*_{j,k}\}_{j\in\mathcal{J},k\in\mathcal{K}}$ is an optimal solution to (OFARR-S). According to the selection rule for $k^v(t)$ in Algorithm 2, it holds that

$$\Phi(n) \geq \frac{1}{\ell_n}\sum_{s=1}^{\ell_n}\left[\sum_{\iota\in\{1,\ldots,|\mathcal{I}|\}}\phi_{n,\iota}(s)\left(\left\{\sum_{k\in\mathcal{K}}\hat{r}_{\iota,j^v(s),k}(n)\cdot x^*_{j^v(s),k}\right\} - (\hat{\lambda}^*(n)-\epsilon_n^C)\right)\right.$$
$$\left. + \phi_{n,0}(s)\left(-\left\{\sum_{k\in\mathcal{K}}\hat{d}_{j^v(s),k}(n)\cdot x^*_{j^v(s),k}\right\}+C\right)\right]. \tag{28}$$

The concentration bounds for $\hat{\boldsymbol{r}}(n)$ and $\hat{\boldsymbol{d}}(n)$ imply that

$$\begin{aligned}
\Phi(n) \geq& \frac{1}{\ell_n}\sum_{s=1}^{\ell_n}\left[\sum_{\iota\in\{1,\ldots,|\mathcal{I}|\}}\phi_{n,\iota}(s)\left(\left\{\sum_{k\in\mathcal{K}} r_{\iota,j^v(s),k}\cdot x^*_{j^v(s),k}\right\} - (\hat{\lambda}^*(n)-\epsilon_n^C)\right)\right.\\
&\left. + \phi_{n,0}(s)\left(-\left\{\sum_{k\in\mathcal{K}} d_{j^v(s),k}\cdot x^*_{j^v(s),k}\right\}+C\right)\right] \quad \text{w.p.} \geq 1-2\delta_n\\
\geq& \frac{1}{\ell_n}\sum_{s=1}^{\ell_n}\left[\sum_{\iota\in\{1,\ldots,\mathcal{I}\}}\phi_{n,\iota}(s)\left(\left\{\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}}\tilde{p}_j(n)\cdot r_{\iota,j,k}\cdot x^*_{j,k}\right\} - (\hat{\lambda}^*(n)-\epsilon_n^C)\right)\right.\\
&\left. + \phi_{n,0}(s)\left(-\left\{\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}}\tilde{p}_j(n)\cdot d_{j,k}\cdot x^*_{j,k}\right\}+C\right)\right]\\
&- r_{\max}\sqrt{\frac{2}{\ell_n}\log\frac{2}{\delta_n}} - d_{\max}\sqrt{\frac{2}{\ell_n}\log\frac{2}{\delta_n}} \quad \text{w.p.} \geq 1-2\delta_n \qquad (29)\\
\geq& \frac{1}{\ell_n}\sum_{s=1}^{\ell_n}\left[\sum_{\iota\in\{1,\ldots,\mathcal{I}\}}\phi_{n,\iota}(s)\left(\left\{\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} p_j\cdot r_{\iota,j,k}\cdot x^*_{j,k}\right\} - (\hat{\lambda}^*(n)-\epsilon_n^C)\right)\right.\\
&\left. + \phi_{n,0}(s)\left(-\left\{\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} p_j\cdot d_{j,k}\cdot x^*_{j,k}\right\}+C\right)\right] - r_{\max}\sqrt{\frac{2}{t_{n-1}}\log\frac{2}{\delta_n}} - d_{\max}\sqrt{\frac{2}{t_{n-1}}\log\frac{2}{\delta_n}}\\
&- r_{\max}\sqrt{\frac{2}{\ell_n}\log\frac{2}{\delta_n}} - d_{\max}\sqrt{\frac{2}{\ell_n}\log\frac{2}{\delta_n}} \quad \text{w.p.} \geq 1-2\delta_n \qquad (30)\\
\geq& \frac{1}{\ell_n}\sum_{s=1}^{\ell_n}\left[\sum_{\iota\in\{1,\ldots,|\mathcal{I}|\}}\phi_{n,\iota}(s)\left(\left\{\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} p_j\cdot r_{\iota,j,k}\cdot x^*_{j,k}\right\} - \lambda^*\right)\right.\\
&\left. + \phi_{n,0}(s)\left(-\left\{\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} p_j\cdot d_{j,k}\cdot x^*_{j,k}\right\}+C\right)\right] - r_{\max}\sqrt{\frac{2}{t_{n-1}}\log\frac{2}{\delta_n}} - d_{\max}\sqrt{\frac{2}{t_{n-1}}\log\frac{2}{\delta_n}}\\
&- r_{\max}\sqrt{\frac{2}{\ell_n}\log\frac{2}{\delta_n}} - d_{\max}\sqrt{\frac{2}{\ell_n}\log\frac{2}{\delta_n}} \quad \text{w.p.} \geq 1-3\delta_n \qquad (31)\\
\geq& -(r_{\max}+d_{\max})\sqrt{\frac{2}{t_{n-1}}\log\frac{2}{\delta_n}} - (r_{\max}+d_{\max})\sqrt{\frac{2}{\ell_n}\log\frac{2}{\delta_n}}. \qquad (32)
\end{aligned}$$

Step (29) leverages the fact that $j(s)$ is sampled from the distribution $\hat{\boldsymbol{p}}(n)$ and applies the Azuma Hoeffding inequality (see Lemma 6) with the filtration $\{\mathcal{F}(s)\}_{s=1}^{\ell_n}$ defined as $\mathcal{F}(s) = \sigma(\{\hat{\boldsymbol{r}}(n), \hat{\boldsymbol{d}}(n), \tilde{\boldsymbol{p}}(n)\} \cup \{j^v(\tau)\}_{\tau=1}^{s})$. This leads to the conclusion that the inequality in (29) holds with probability $\geq 1 - 2\delta_n$. Step (30) applies the Hoeffding inequality for the estimates $\tilde{\boldsymbol{p}}(n)$, which are computed based on the observations up to time t_{n-1}. Step (31) follows from the Lemma 2, which establishes that $\hat{\lambda}^*(n) \leq \lambda^* + \epsilon_n^C$ holds with probability at least $1 - 3\delta_n$. Step (32) is justified by the feasibility of $\boldsymbol{y}^*$ for (OFARR-S) and the fact that $\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} p_j \cdot d_{j,k} \cdot y^*_{j,k} \leq C$. Notably, the inequality in (32) holds with certainty. Put these things together completes the proof of (25).

Finally, we prove inequalities (26) and (27) both using the Azuma-Hoeffding inequality. Inequality (27)) can be shown by considering $\{r_{i,j^v(s),k^v(s)} - \sum_{j\in\mathcal{J}} \tilde{p}_j(n) \cdot r_{i,j,\kappa_n(\boldsymbol{\phi}_n(s),j)}\}_{s=1}^{\ell_n}$, which forms a martingale difference sequence with respect to the filtration $\{\mathcal{F}(s)\}_{s=1}^{\ell_n}$ defined as $\mathcal{F}(s) = \sigma(\{\hat{\boldsymbol{r}}(n), \hat{\boldsymbol{d}}(n), \tilde{\boldsymbol{p}}(n)\} \cup \{j^v(\tau)\}_{\tau=1}^{s})$. Crucially, in the conditional expectation $\mathbf{E}[r_{i,j^v(s),k^v(s)}|\mathcal{F}(s-1)]$, the randomness lies solely in $j^v(s)$, as $k^v(s)$ is deterministic conditioned on $j^v(s) \cup \mathcal{F}(s-1)$. Since the weight vector $\boldsymbol{\phi}_n(s)$ is $\mathcal{F}(s-1)$-measurable, applying the Azuma-Hoeffding inequality yields

$$\sum_{s=1}^{\ell_n} r_{i,j^v(s),k^v(s)} \leq \sum_{s=1}^{\ell_{n-1}} \sum_{j\in\mathcal{J}} \tilde{p}_j(n) \cdot r_{i,j,\kappa_n(\boldsymbol{\phi}_n(s),j)} + r_{\max}\sqrt{2\ell_n \log\frac{|\mathcal{I}|}{\delta_n}}, \quad \text{w.p. } \geq 1 - \delta_n/|\mathcal{I}|.$$

Then applying the Hoeffding inequality into $\hat{\boldsymbol{p}}(n)$ gives

$$\sum_{s=1}^{\ell_n} \sum_{j\in\mathcal{J}} \tilde{p}_j(n) \cdot r_{i,j,\kappa_n(\boldsymbol{\phi}_n(s),j)} \leq \sum_{s=1}^{\ell_n} \sum_{j\in\mathcal{J}} p_j \cdot r_{i,j,\kappa_n(\boldsymbol{\phi}_n(s),j)} + r_{\max}\sqrt{2\ell_n \log\frac{|\mathcal{I}|}{\delta_n}}, \quad \text{w.p. } \geq 1 - \delta_n/|\mathcal{I}|.$$

Thus, the inequality (26) follows.

Similarly, inequality (27) can be shown by considering $\{d_{j^v(s),k^v(s)} - \sum_{j\in\mathcal{J}} \tilde{p}_j(n) \cdot d_{j,\kappa_n(\boldsymbol{\phi}_n(s),j)}\}_{s=1}^{\ell_n}$, which also forms a martingale difference sequence with respect to the filtration $\{\mathcal{F}(s)\}_{s=1}^{\ell_n}$. By applying the Azuma-Hoeffding inequality, we obtain

$$\sum_{s=1}^{\ell_n} d_{j^v(s),k^v(s)} \leq \sum_{s=1}^{\ell_n} \sum_{j\in\mathcal{J}} \tilde{p}_j(n) \cdot d_{j,\kappa_n(\boldsymbol{\phi}_n(s),j)} + d_{\max}\sqrt{2\ell_n \log\frac{1}{\delta_n}}, \quad \text{w.p. } \geq 1 - \delta_n.$$

Applying the Hoeffding inequality again to $\hat{\boldsymbol{p}}(n)$ yields

$$\sum_{s=1}^{\ell_n} \sum_{j\in\mathcal{J}} \tilde{p}_j(n) \cdot d_{j,\kappa_n(\boldsymbol{\phi}_n(s),j)} \leq \sum_{s=1}^{\ell_n} \sum_{j\in\mathcal{J}} p_j \cdot d_{j,\kappa_n(\boldsymbol{\phi}_n(s),j)} + d_{\max}\sqrt{2\ell_n \log\frac{1}{\delta_n}}, \quad \text{w.p. } \geq 1 - \delta_n.$$

Combining these results completes the proof of inequality (27). Therefore, based on inequalities (25), (26), and (27), for all $i \in \mathcal{I}$ we have that

$$\begin{aligned}
&\frac{1}{\ell_n}\sum_{s=1}^{\ell_n}\sum_{j\in\mathcal{J}} p_j \cdot r_{i,j,\kappa_n(\boldsymbol{\phi}_n(s),j)} \geq \frac{1}{\ell_n}\sum_{s=1}^{\ell_n} r_{i,j^v(s),k^v(s)} - 2r_{\max}\sqrt{\frac{2}{\ell_n}\log\frac{1}{\delta_n}} \quad \text{w.p. } 1-2\delta_n \\
&= \frac{1}{\ell_n}\sum_{s=1}^{\ell_n} \hat{r}_{i,j^v(s),k^v(s)}(n) - 2r_{\max}\sqrt{\frac{2}{\ell_n}\log\frac{1}{\delta_n}} - \frac{1}{\ell_n}\sum_{s=1}^{\ell_n} r_{\max}\sqrt{\frac{2\log(|\mathcal{I}||\mathcal{J}||\mathcal{K}|/\delta_n)}{N_{j^v(s),k^v(s)}(t_{n-1})}}
\end{aligned}$$

$$\geq \frac{1}{\ell_n}\sum_{s=1}^{\ell_n}\hat{r}_{i,j^v(s),k^v(s)}(n) - 2r_{\max}\sqrt{\frac{2}{\ell_n}\log\frac{1}{\delta_n}} - r_{\max}\sqrt{\frac{2\log(|\mathcal{I}||\mathcal{J}||\mathcal{K}|/\delta_n)}{\min_{j,k}N_{j,k}(t_{n-1})}}$$

$$\geq \left(\hat{\lambda}^*(n) - \epsilon_n^C\right) + \Phi(n) - r_{\max}\sqrt{\frac{8\log(|\mathcal{I}|+1)}{\ell_n}} - 2r_{\max}\sqrt{\frac{2}{\ell_n}\log\frac{1}{\delta_n}} - r_{\max}\sqrt{\frac{2\log(|\mathcal{I}||\mathcal{J}||\mathcal{K}|/\delta_n)}{\min_{j,k}N_{j,k}(t_n)}}$$

$$\geq \lambda^* - \epsilon_n^A - \epsilon_n^B - \epsilon_n^C + \Phi(n) - r_{\max}\sqrt{\frac{8\log(|\mathcal{I}|+1)}{\ell_n}} - 2r_{\max}\sqrt{\frac{2}{\ell_n}\log\frac{1}{\delta_n}} - r_{\max}\sqrt{\frac{2\log(|\mathcal{I}||\mathcal{J}||\mathcal{K}|/\delta_n)}{\min_{j,k}N_{j,k}(t_{n-1})}}$$

$$\geq \left(\lambda^* - \epsilon_n^A - \epsilon_n^B - \epsilon_n^C\right) - (r_{\max}+d_{\max})\sqrt{\frac{2}{t_{n-1}}\log\frac{2}{\delta_n}} - (r_{\max}+d_{\max})\sqrt{\frac{2}{\ell_n}\log\frac{2}{\delta_n}}$$
$$- r_{\max}\sqrt{\frac{8\log(|\mathcal{I}|+1)}{\ell_n}} - 2r_{\max}\sqrt{\frac{2}{\ell_n}\log\frac{1}{\delta_n}} - r_{\max}\sqrt{\frac{2\log(|\mathcal{I}||\mathcal{J}||\mathcal{K}|/\delta_n)}{\min_{j,k}N_{j,k}(t_{n-1})}} \quad \text{w.p. } 1-9\delta_n$$

$$\geq \lambda^* - \epsilon_n^A - \epsilon_n^B - \epsilon_n^C - (4r_{\max}+2d_{\max})\sqrt{\frac{2}{t_{n-1}}\log\frac{2}{\delta_n}} - r_{\max}\sqrt{\frac{8\log(|\mathcal{I}|+1)}{\ell_n}}$$
$$- r_{\max}\sqrt{\frac{2\log(|\mathcal{I}||\mathcal{J}||\mathcal{K}|/\delta_n)}{\min_{j,k}N_{j,k}(t_{n-1})}} \geq \lambda^* - \epsilon_n^A - \epsilon_n^B - \epsilon_n^C - \epsilon_n,$$

which establishes that the inequality $\frac{1}{\ell_n}\sum_{s=1}^{\ell_n}\sum_{j\in\mathcal{J}}p_j\cdot r_{i,j,f_n(\boldsymbol{\phi}_n(s),j)} \geq \lambda^* - \epsilon_n^A - \epsilon_n^B - \epsilon_n^C - \epsilon_n$ hold simultaneously for all $i\in\mathcal{I}$ with probability at least $1-11\delta_n$. Similarly, it holds that

$$\frac{1}{\ell_n}\sum_{s=1}^{\ell_n}\sum_{j\in\mathcal{J}}p_j\cdot d_{j,f_n(\boldsymbol{\phi}_n(s),j)} \leq \frac{1}{\ell_n}\sum_{s=1}^{\ell_n}d_{j^v(s),k^v(s)} + 2d_{\max}\sqrt{\frac{2}{\ell_n}\log\frac{1}{\delta_n}} \quad \text{w.p. } 1-2\delta_n$$

$$= \frac{1}{\ell_n}\sum_{s=1}^{\ell_n}\hat{d}_{j^v(s),k^v(s)}(n) + 2d_{\max}\sqrt{\frac{2}{\ell_n}\log\frac{1}{\delta_n}} + \frac{1}{\ell_n}\sum_{s=1}^{\ell_n}d_{\max}\sqrt{\frac{2\log(|\mathcal{J}||\mathcal{K}|)/\delta_n}{N_{j^s(s),k^s(s)}(t_{n-1})}}$$

$$\leq \frac{1}{\ell_n}\sum_{s=1}^{\ell_n}\hat{d}_{j^v(s),k^v(s)}(n) + 2d_{\max}\sqrt{\frac{2}{\ell_n}\log\frac{1}{\delta_n}} + d_{\max}\sqrt{\frac{2\log(|\mathcal{J}||\mathcal{K}|)/\delta_n}{\min_{j,k}N_{j,k}(t_{n-1})}}$$

$$\leq C - \Phi(n) + d_{\max}\sqrt{\frac{8\log(|\mathcal{I}|+1)}{\ell_n}} + 2d_{\max}\sqrt{\frac{2}{\ell_n}\log\frac{1}{\delta_n}} + d_{\max}\sqrt{\frac{2\log(|\mathcal{J}||\mathcal{K}|)/\delta_n}{\min_{j,k}N_{j,k}(t_{n-1})}}$$

$$\leq C + (r_{\max}+d_{\max})\sqrt{\frac{2}{\ell_n}\log\frac{2}{\delta_n}} + (r_{\max}+d_{\max})\sqrt{\frac{2}{t_{n-1}}\log\frac{2}{\delta_n}}$$
$$+ d_{\max}\sqrt{\frac{8\log(|\mathcal{I}|+1)}{\ell_n}} + 2d_{\max}\sqrt{\frac{2}{\ell_n}\log\frac{1}{\delta_n}} + d_{\max}\sqrt{\frac{2\log(|\mathcal{J}||\mathcal{K}|)/\delta_n}{\min_{j,k}N_{j,k}(t_{n-1})}} \quad \text{w.p. } 1-9\delta_n$$

$$\leq C + (2r_{\max}+4d_{\max})\sqrt{\frac{2}{t_{n-1}}\log\frac{2}{\delta_n}} + d_{\max}\sqrt{\frac{8\log(|\mathcal{I}|+1)}{\ell_n}} + d_{\max}\sqrt{\frac{2\log(|\mathcal{J}||\mathcal{K}|)/\delta_n}{\min_{j,k}N_{j,k}(t_{n-1})}}$$

$$\leq C + \epsilon_n.$$

This completes the proof that $\frac{1}{\ell_n}\sum_{s=1}^{\ell_n}\sum_{j\in\mathcal{J}}p_j\cdot d_{j,f_n(\boldsymbol{\phi}_n(s),j)} \leq C+\epsilon_n$ holds with probability at least $1-11\delta_n$. Combining all things together, the Lemma 3 follows.

A.4 Proof of Lemma 4

We begin by noting that:

$$\mathbf{E}[V(t) \mid \mathcal{H}(t_{n-1})] \geq 1 - \underbrace{\mathbf{E}\left[\mathbf{I}\left\{\sum_{\tau=\max\{t-d_{\max},1\}}^{t-1} \hat{A}(\tau)\mathbf{I}\{\hat{D}(\tau) \geq t-\tau+1\} > C-1\right\} \,\middle|\, \mathcal{H}(t_{n-1})\right]}_{=M_n(t)}.$$

The proof of the lemma involves two key steps. Firstly, we demonstrate that for any $t \in \{t_{n-1}+1+d_{\max}, \ldots, t_n\}$ and any fixed $\varepsilon > 0$, the following inequality holds with certainty:

$$M_n(t) \leq \frac{1}{(1+\varepsilon)^{C-1}} \cdot \exp\left[\frac{\varepsilon(1-f_n)}{1+g_n} \cdot \frac{1}{\ell_n}\sum_{s=1}^{\ell_n} p_j d_{j,\kappa_n(\boldsymbol{\phi}_n(s),j)} + \varepsilon \cdot f_n \cdot d_{\max}\right]. \tag{33}$$

Importantly, the right-hand side of (33) is independent of t and only relies on the epoch index n. This independence arises because the weight vector construction ensures that $\tilde{k}(t)$ is i.i.d. for all $t \in \{t_{n-1}+1, ..., t_n\}$, conditional on the history $\mathcal{H}(t_{n-1})$. Applying Lemma 3 yields:

$$\exp\left[\frac{\varepsilon(1-f_n)}{1+g_n} \cdot \frac{1}{\ell_n}\sum_{s=1}^{\ell_n} p_j \cdot d_{j,\kappa_n(\boldsymbol{\phi}_n(s),j)} + \varepsilon \cdot f_n \cdot d_{\max}\right] \leq \exp\left[\frac{\varepsilon(C+\epsilon_n)}{1+g_n}\right] \cdot \exp\left[\varepsilon \cdot (d_{max}-C)^+\right], \tag{34}$$

which holds with probability at least $1-11\delta$. Combining (33) and (34), we establish the first part of Lemma 4. Secondly, we demonstrate that the following inequality holds:

$$\mathbf{E}[\hat{R}_i(t) \mid \mathcal{H}(t_{n-1})] \geq \frac{1-f_n}{1+g_n} \cdot \frac{1}{\ell_n}\sum_{s=1}^{\ell_n}\sum_{j\in\mathcal{J}} p_j \cdot r_{i,j,\kappa_n(\boldsymbol{\phi}_n(s),j)}. \tag{35}$$

Applying Lemma 3 again, we derive that:

$$\begin{aligned}\frac{1}{1+g_n} \cdot \frac{1}{\ell_n}\sum_{s=1}^{\ell_n}\sum_{j\in\mathcal{J}} p \cdot r_{i,j,\kappa_n(\boldsymbol{\phi}_n(s),j)} \geq& \frac{1-f_n}{1+g_n} \cdot \left(\lambda^* - \epsilon_n^A - \epsilon_n^B - \epsilon_n^B - \epsilon_n\right) \\ \geq& \frac{1}{1+g_n}\lambda^* - \epsilon_n^A - \epsilon_n^B - \epsilon_n^C - \frac{\epsilon_n}{1+g_n} - f_n \cdot \lambda^*,\end{aligned}$$

which holds for all $i \in \mathcal{I}$ with probability at least $1-11\delta$. Thus, the second part of Lemma 4 is established.

Thus, what remains is to establish (33) and (35). Inequality (33) is demonstrated through the following sequence of calculations, where all equalities and inequalities hold certainty:

$$\begin{aligned}M_n(t) =& \mathbf{E}\left[\mathbf{I}\left\{\sum_{\tau=\max\{t-d_{\max},1\}}^{t-1} \hat{A}(\tau)\mathbf{I}\{\hat{D}(\tau) \geq t-\tau+1\} > C-1\right\} \,\middle|\, \mathcal{H}(t_{n-1})\right] \\ =& \mathbf{E}\left[\mathbf{I}\left\{(1+\varepsilon)^{\sum_{\tau=\max\{t-d_{\max}+1,1\}}^{t-1} \hat{A}(\tau)\mathbf{I}\{\hat{D}(\tau)\geq t-\tau+1\}} > (1+\varepsilon)^{C-1}\right\} \,\middle|\, \mathcal{H}(t_{n-1})\right] \\ \leq& \frac{1}{(1+\varepsilon)^{C-1}} \cdot \mathbf{E}\left[(1+\varepsilon)^{\sum_{\tau=\max\{t-d_{\max},1\}}^{t-1} \hat{A}(\tau)\mathbf{I}\{\hat{D}(\tau)\geq t-\tau+1\}} \,\middle|\, \mathcal{H}(t_{n-1})\right]\end{aligned} \tag{36}$$

$$= \frac{1}{(1+\varepsilon)^{C-1}} \cdot \prod_{\tau=\max\{t-d_{\max},1\}}^{t-1} \mathbf{E}\left[(1+\varepsilon)^{\hat{A}(\tau)\mathbf{I}\{\hat{D}(\tau)\geq t-\tau+1\}} \,\middle|\, \mathcal{H}(t_{n-1})\right] \tag{37}$$

$$\leq \frac{1}{(1+\varepsilon)^{C-1}} \cdot \prod_{\tau=\max\{t-d_{\max},1\}}^{t-1} \left(1+\varepsilon\cdot\mathbf{E}\left[\hat{A}(\tau)\mathbf{I}\{\hat{D}(\tau)\geq t-\tau+1\} \,\middle|\, \mathcal{H}(t_{n-1})\right]\right) \tag{38}$$

$$\leq \frac{1}{(1+\varepsilon)^{C-1}} \cdot \prod_{\tau=\max\{t-d_{\max},1\}}^{t-1} \Big(1+\varepsilon\cdot\mathbf{P}\left(I_1(\tau)\right)\mathbf{E}\left[\tilde{A}(\tau)\mathbf{I}\{\tilde{D}(\tau)\geq t-\tau+1\} \,\middle|\, \mathcal{H}(t_{n-1})\right] \tag{39}$$

$$+\varepsilon\cdot\mathbf{P}\left(I_2(\tau)\right)\mathbf{E}\left[\tilde{A}^e(\tau)\mathbf{I}\{\tilde{D}^e(\tau)\geq t-\tau+1\} \,\middle|\, \mathcal{H}(t_{n-1})\right]\Big)$$

Step (36) is derived using the Markov inequality. Step (37) follows from the joint independence of the sequence $\{\hat{A}(\tau)\mathbf{I}\{\hat{D}(\tau)\geq t-\tau+1\}\}_{\tau=\max\{t-d_{\max},1\}}^{t-1}$ conditioned on $\mathcal{H}(t_{n-1})$, as established by the coupling argument. Step (38) utilizes the inequality $(1+\varepsilon)^a \leq 1+\varepsilon\cdot a$ for all $a\in[0,1], \varepsilon>0$. Step (39) follows from the facts that $(I_1(t), I_2(t))$ is independent of the history $\mathcal{H}(t_{n-1})$ for $t\in\{t_{n-1}+1,\ldots,t_n\}$ and

$$(\hat{A}(\tau),\hat{D}(\tau)) = \begin{cases} (\tilde{A}(\tau),\tilde{D}(\tau)) & \text{if } I_1(t) \text{ holds,} \\ (\tilde{A}^e(\tau),\tilde{D}^e(\tau)) & \text{if } I_2(t) \text{ holds,} \end{cases}$$

Recall that $\tilde{A}(t)=\tilde{A}^e(t)=1,\ \forall t\in[T]$, thus

$$\begin{aligned} M_n(t) \leq & \frac{1}{(1+\varepsilon)^{C-1}} \cdot \prod_{\tau=\max\{t-d_{\max},1\}}^{t-1} \Big(1+\varepsilon\cdot\mathbf{P}\left(I_1(\tau)\right)\mathbf{E}\left[\mathbf{I}\{\tilde{D}(\tau)\geq t-\tau+1\} \,\middle|\, \mathcal{H}(t_{n-1})\right] \\ & +\varepsilon\cdot\mathbf{P}\left(I_2(\tau)\right)\mathbf{E}\left[\mathbf{I}\{\tilde{D}^e(\tau)\geq t-\tau+1\} \,\middle|\, \mathcal{H}(t_{n-1})\right]\Big) \\ \leq & \frac{1}{(1+\varepsilon)^{C-1}} \cdot \prod_{\tau=\max\{t-d_{\max},1\}}^{t-1} \exp\Big(\varepsilon\cdot\mathbf{P}(I_1(\tau))\mathbf{E}\left[\mathbf{I}\{\tilde{D}(\tau)\geq t-\tau+1\} \,\middle|\, \mathcal{H}(t_{n-1})\right] \\ & +\varepsilon\cdot\mathbf{P}(I_2(\tau))\mathbf{E}\left[\mathbf{I}\{\tilde{D}^e(\tau)\geq t-\tau+1\} \,\middle|\, \mathcal{H}(t_{n-1})\right]\Big) \end{aligned} \tag{40}$$

$$\frac{1}{(1+\varepsilon)^{C-1}} \cdot \exp\Bigg(\frac{\varepsilon(1-f_n)}{1+g_n} \cdot \sum_{\tau=\max\{t-d_{\max},1\}}^{t-1} \mathbf{E}\left[\mathbf{I}\{\tilde{D}(\tau)\geq t-\tau+1\} \,\middle|\, \mathcal{H}(t_{n-1})\right]$$

$$+\varepsilon\cdot f_n \cdot \sum_{\tau=\max\{t-d_{\max},1\}}^{t-1} \mathbf{E}\left[\mathbf{I}\{\tilde{D}^e(\tau)\geq t-\tau+1\} \,\middle|\, \mathcal{H}(t_{n-1})\right]\Bigg) \tag{41}$$

Step (40) follows from the inequality $1+\varepsilon\leq e^{\epsilon},\ \forall\varepsilon>0$. To further deal with (41), note that

$$\begin{aligned} & \sum_{\tau=\max\{t-d_{\max},1\}}^{t-1} \mathbf{E}\left[\mathbf{I}\{\tilde{D}(\tau)\geq t-\tau+1\} \,\middle|\, \mathcal{H}(t_{n-1})\right] \\ \leq & \sum_{\tau=t-d_{\max}+1}^{t} \sum_{s=t-\tau+1}^{d_{\max}} \mathbf{E}\left[\mathbf{I}\{\tilde{D}(\tau)=s\} \,\middle|\, \mathcal{H}(t_{n-1})\right] \\ = & \sum_{s=1}^{d_{\max}} \sum_{\tau=t-s+1}^{t} \mathbf{E}\left[\mathbf{I}\{\tilde{D}(\tau)=s\} \,\middle|\, \mathcal{H}(t_{n-1})\right] \end{aligned}$$

$$= \sum_{s=1}^{d_{\max}} \sum_{\tau=t-s+1}^{t} \mathbf{E}\left[\mathbf{I}\{\tilde{D}(t)=s\} \,\middle|\, \mathcal{H}(t_{n-1})\right] \tag{42}$$
$$= \sum_{s=1}^{d_{\max}} \mathbf{E}\left[s\cdot\mathbf{I}\{\tilde{D}(t)=s\} \,\middle|\, \mathcal{H}(t_{n-1})\right]$$
$$= \sum_{s=1}^{d_{\max}} \mathbf{E}\left[s\cdot\mathbf{I}\{\tilde{D}(t)=s\} \,\middle|\, \mathcal{H}(t_{n-1})\right] \le \mathbf{E}\left[\tilde{D}(t) \,\middle|\, \mathcal{H}(t_{n-1})\right]$$

In (41), it is crucial to note that the range of summation, namely $\{\max\{t-d_{\max},1\},\dots,t-1\}$, lies entirely within the time interval of n, as ensured by our assumption $t \in \{t_{n-1}+1+d_{\max},\dots,t_n\}$. Recall the construction of $\tilde{k}(t)$ and the definition of $(\tilde{R}(t),\tilde{A}(t),\tilde{D}(t)) \sim \mathcal{O}_{j(t),\tilde{k}(t)}$ in our coupling argument. These two facts imply that $\{(\tilde{A}_i(\tau),\tilde{D}_i(\tau))\}_{\tau=\max\{t-d_{\max},1\}}^{t}$ are i.i.d conditioned on $\mathcal{H}(t_{n-1})$, which leads to (42). Similarly,

$$\sum_{\tau=\max\{t-d_{\max},1\}}^{t-1} \mathbf{E}\left[\mathbf{I}\{\tilde{D}^e(\tau)\ge t-\tau+1\} \,\middle|\, \mathcal{H}(t_{n-1})\right] \le \mathbf{E}\left[\tilde{D}^e(t) \,\middle|\, \mathcal{H}(t_{n-1})\right].$$

Therefore, $M_n(t)$ can be further upper-bounded as

$$\begin{aligned}
M_n(t) \le& \frac{1}{(1+\varepsilon)^{C-1}}\cdot\exp\left(\frac{\varepsilon(1-f_n)}{1+g_n}\cdot\sum_{\tau=t-d_{\max}+1}^{t}\sum_{s=t-\tau+1}^{d_{\max}}\mathbf{E}\left[\mathbf{I}\{\tilde{D}(\tau)=s\} \,\middle|\, \mathcal{H}(t_{n-1})\right]\right.\\
&\left.+\varepsilon\cdot f_n\cdot\sum_{\tau=t-d_{\max}+1}^{t}\sum_{s=t-\tau+1}^{d_{\max}}\mathbf{E}\left[\mathbf{I}\{\tilde{D}^e(\tau)=s\} \,\middle|\, \mathcal{H}(t_{n-1})\right]\right)\\
\le& \frac{1}{(1+\varepsilon)^{C-1}}\cdot\exp\left(\frac{\varepsilon(1-f_n)}{1+g_n}\cdot\mathbf{E}\left[\tilde{D}(t) \,\middle|\, \mathcal{H}(t_{n-1})\right]\right.\\
&\left.+\varepsilon\cdot f_n\cdot\cdot\mathbf{E}\left[\tilde{D}^e(t) \,\middle|\, \mathcal{H}(t_{n-1})\right]\right)\\
=& \frac{1}{(1+\varepsilon)^{C-1}}\cdot\exp\left[\frac{\varepsilon(1-f_n)}{1+g_n}\cdot\frac{1}{\ell_n}\sum_{s=1}^{\ell_n}p_j\cdot d_{j,f_n(\phi_n(s),j)}+\varepsilon\cdot f_n\cdot\mathbf{E}[\tilde{D}^e(t) \,\middle|\, \mathcal{H}(t_{n-1})]\right] \qquad (43)\\
\le& \frac{1}{(1+\varepsilon)^{C-1}}\cdot\exp\left[\frac{\varepsilon(1-f_n)}{1+g_n}\cdot\frac{1}{\ell_n}\sum_{s=1}^{\ell_n}p_j\cdot d_{j,f_n(\phi_n(s),j)}+\varepsilon\cdot f_n\cdot d_{\max}\right]. \qquad (44)
\end{aligned}$$

Step (43) follows from taking the conditional expectation and recalling the construction of $\tilde{k}(t)$ in Algorithm 1. Step (44) holds because $\tilde{D}^e(t)\le d_{\max},\ \forall t\in[T]$. This completes the proof of (33).

To establish (35), note that the coupling argument ensures that for every $t\in\{t_{n-1}+1,\dots,t_n\}$ and every $i\in\mathcal{I}$, the following inquality holds with certainty:

$$\begin{aligned}
\mathbf{E}[\hat{R}_i(t)\mid\mathcal{H}(t_{n-1})] &= \mathbf{E}[I_2(t)\cdot\tilde{R}_i(t)+I_2(t)\cdot\tilde{R}_i^e(t)\mid\mathcal{H}(t_{n-1})]\\
&\ge \mathbf{E}[I_2(t)\cdot\tilde{R}_i(t)\mid\mathcal{H}(t_{n-1})]\\
&= \frac{1-f_n}{1+g_n}\cdot\frac{1}{\ell_n}\sum_{s=1}^{\ell_n}\sum_{j\in\mathcal{J}}p_j\cdot r_{i,j,\kappa_n(\phi_n(s),j)}
\end{aligned}$$

Thus, the inequality (35) follows, and we complete the proof of Lemma 4.

A.5 Proof of Theorem 2

Firstly, by unpacking the definitions of f_n and g_n in Theorem 1 and setting $\varepsilon = \varepsilon^*(C)$, we have

$$\begin{aligned}\sum_{t=t_{n-1}+1}^{t_n} R_i(t) \geq \frac{1}{1+\epsilon_n/C} \cdot \mathcal{L}(\varepsilon^*(C)) \cdot \ell_n \cdot \lambda^* - r_{\max}\sqrt{2d_{\max}\ell_n \log(d_{\max}/\delta_n)} - 2r_{\max}d_{\max} \\ - \ell_n \cdot \left(\epsilon_n^A + \epsilon_n^B + \epsilon_n^C + \epsilon_n + f_n\lambda^* + 2\varepsilon^*(C) \cdot f_n \cdot (d_{max} - C)^+ \cdot \lambda^*\right), \quad \text{w.p. } 1 - 23\delta_n,\end{aligned} \tag{45}$$

when $n \geq 3\log \varepsilon^*(C) + 2\log d_{\max}$, where $\epsilon_n^A, \epsilon_n^B, \epsilon_n^C$, and ϵ_n are defined as follows

$$\epsilon_n^A = 2\sqrt{\frac{d_{\max}\log(1/\delta_n)}{C \cdot t_{n-1}}} + \frac{2d_{\max}\log(1/\delta_n)}{C \cdot t_{n-1}}, \epsilon_n^B = r_{\max}\sqrt{\frac{2\log(|\mathcal{I}|/\delta_n)}{t_{n-1}}},$$

$$\epsilon_n^C = 2r_{\max}\left[\sqrt{\frac{\log(1/\delta_n)}{t_{n-1}}} + \frac{\log(1/\delta_n)}{t_{n-1}}\right] + 2|\mathcal{J}|\,(r_{\max} + d_{\max})\sqrt{\frac{2\log(|\mathcal{J}||\mathcal{K}||\mathcal{I}|/\delta_n)}{\min_{j\in\mathcal{J},k\in\mathcal{K}} N_{j,k}(t_{n-1})}},$$

$$\epsilon_n = 6d_{\max}\sqrt{\frac{2}{t_{n-1}}\log\frac{2}{\delta_n}} + d_{\max}\sqrt{\frac{8\log(|\mathcal{I}|+1)}{\ell_n}} + d_{\max}\sqrt{\frac{2\log(|\mathcal{I}||\mathcal{J}||\mathcal{K}|/\delta_n)}{\min_{j,k} N_{j,k}(t_{n-1})}},$$

To proceed with analyzing (45), we next derive a lower bound for $\min_{j,k} N_{j,k}(t_{n-1})$. Recall that $N_{j,k}(t)$ denotes the number of times that the action k is chosen for arrival type j up to time t. By definition, at epoch n, the probability of selecting action $k \in \mathcal{K}$ (i.e., $k \neq k_{\text{null}}$) at round t when the arrival type is j satisfies:

$$\mathbf{P}\left(k(t) = k, j(t) = j\right) \geq \frac{1}{\ell_n^{1/3}} \cdot \frac{1}{d_{\max}} \cdot \frac{p_j}{|\mathcal{K}|}.$$

The term $1/d_{\max}$ arises from the observation that any policy experiences at least one instance of resource availability every $d_{\max}$ rounds. Note that the actual number of selections of action k for arrival type j may be significantly higher in practice, as this also accounts for exploratory selections of each action. At the beginning of epoch n, the expected value for $N_{j,k}(t_{n-1})$ is given by:

$$\mathbf{E}[N_{j,k}(t_{n-1})] = 1 + \sum_{t=1}^{t_{n-1}} \mathbf{P}\left(k(t) = k, j(t) = j\right) \geq 1 + \sum_{i=0}^{n-1} \ell_i \cdot \ell_i^{-1/3} \cdot \frac{1}{d_{\max}} \cdot \frac{p_j}{|\mathcal{K}|} \geq \frac{p_j}{d_{\max}|\mathcal{K}|} \cdot \ell_n^{2/3}$$

The first equality follows from initializing $N_{j,k}(0) = 1$. Using the Chernoff bound, we can further show that for any $\delta \in (0,1)$:

$$\mathbf{P}\left(N_{j,k}(n) \leq (1-\delta)\mathbf{E}[N_{j,k}(n)]\right) \leq \exp\left(-\frac{\delta^2\mathbf{E}[N_{j,k}(n)]}{2}\right) \leq \exp\left(-\frac{p_j\delta^2 2^{2n/3}}{2|\mathcal{K}|d_{\max}^{1/3}}\right).$$

Setting $\delta = \frac{\sqrt{2}}{2}$ leads to

$$\mathbf{P}\left(N_{j,k}(n) \leq \left(1 - \frac{\sqrt{2}}{2}\right) \cdot \mathbf{E}[N_{j,k}(n)]\right) \leq \exp\left(-\frac{p_j 2^{2n/3}}{4|\mathcal{K}|d_{\max}^{1/3}}\right)$$

Building on this and applying the union bound, the following inequalities hold simultaneously when $n \geq O\left(2\log d_{\max} + \log(1/p_{\min}) + \log\left(\log\left(|\mathcal{I}||\mathcal{J}||\mathcal{K}|\right)\right)\right)$,

$$\begin{aligned}
&\epsilon_n \leq 0.5,\\
&\ell_n \cdot \left(\epsilon_n^A + \epsilon_n^B + \epsilon_n^C + \epsilon_n + f_n\lambda^* + 2\varepsilon^*(C)\cdot f_n \cdot (d_{max} - C)^+ \cdot \lambda^*\right)\\
&\leq O\left(|\mathcal{J}| d_{\max}\ell_n^{2/3}\sqrt{\log(\frac{|\mathcal{J}||\mathcal{K}||\mathcal{I}|}{\delta_n})} + \frac{d_{\max}}{C}\log(\frac{1}{\delta_n}) + r_{\max}\log(\frac{1}{\delta_n}) + \varepsilon^*(C)(d_{max}-C)^+\ell_n^{2/3} + \ell_n^{2/3}\right)\\
&= O\left(|\mathcal{J}| d_{\max}\ell_n^{2/3}\sqrt{n\cdot\log(|\mathcal{J}||\mathcal{K}||\mathcal{I}|)} + \frac{d_{\max}}{C}\cdot n + r_{\max}\cdot n + \varepsilon^*(C)\cdot(d_{max}-C)^+\cdot\ell_n^{2/3} + \ell_n^{2/3}\right)\\
&\qquad \text{w.p. } 1 - \sum_{j\in\mathcal{J}}\exp\left(-p_j 2^{2n/3}/(4|\mathcal{K}|d_{\max}^{1/3})\right).
\end{aligned}$$

Thus, with probability at least $1 - 23\delta_n - \sum_{j\in\mathcal{J}}\exp\left(-p_j 2^{2n/3}/(4|\mathcal{K}|d_{\max}^{1/3})\right)$, we have that:

$$\begin{aligned}
\sum_{t=t_{n-1}+1}^{t_n} R_i(t) \geq \left(1-\frac{1}{2C}\right)\cdot\mathcal{L}(\varepsilon^*(C))\cdot\ell_n\cdot\lambda^* - O\Bigg(&|\mathcal{J}| d_{\max}\cdot\ell_n^{2/3}\sqrt{\log(|\mathcal{J}||\mathcal{K}||\mathcal{I}|/\delta_n)} + \frac{d_{\max}}{C}n\\
&+r_{\max}\cdot n + \varepsilon^*(C)\cdot(d_{max}-C)^+\cdot\ell_n^{2/3} + \ell_n^{2/3} + r_{\max}\sqrt{d_{\max}\ell_n\log(d_{\max}/\delta_n)} + r_{\max}d_{\max}\Bigg),
\end{aligned} \tag{46}$$

when $n \geq n_0 = O\left(2\log d_{\max} + 3\log\epsilon^*(C) + \log(1/p_{\min}) + \log\left(\log\left(|\mathcal{I}||\mathcal{J}||\mathcal{K}|\right)\right)\right)$. Summing over n from 0 to $\lceil\log(T/d_{\max})\rceil$ and applying a union bound, we can obtain that with probability at least $1 - \sum_{n=1}^{T}(23\delta_n + \sum_{j\in\mathcal{J}}\exp(-p_j 2^{2n/3}/(4|\mathcal{K}|d_{\max}^{1/3})))$:

$$\begin{aligned}
\sum_{t=1}^{T} R_i(t) \geq &\left(1-\frac{1}{2C}\right)\cdot\mathcal{L}(\varepsilon^*(C))\cdot T\cdot\lambda^* - O\Bigg(\sum_{n=1}^{\log(T/d_{\max})}\Bigg[|\mathcal{J}| d_{\max}\ell_n^{2/3}\sqrt{\log(|\mathcal{J}||\mathcal{K}||\mathcal{I}|/\delta_n)} + \frac{d_{\max}}{C}n\\
&+r_{\max}\cdot n + \varepsilon^*(C)\cdot(d_{max}-C)^+\cdot\ell_n^{2/3} + \ell_n^{2/3} + r_{\max}\sqrt{d_{\max}\ell_n\log(d_{\max}/\delta_n)} + r_{\max}d_{\max}\Bigg]\Bigg) - O\left(\sum_{n=1}^{n_0}\ell_n\right)\\
\geq &\left(1-\frac{1}{2C}\right)\cdot\mathcal{L}(\varepsilon^*(C))\cdot T\cdot\lambda^*\\
&- O\left(|\mathcal{J}|\cdot d_{\max}T^{2/3}\sqrt{\log T\cdot\log(|\mathcal{J}||\mathcal{K}||\mathcal{I}|)} + \varepsilon^*(C)\cdot(d_{max}-C)^+\cdot T^{2/3} + T^{12/23} + d^2 3_{\max} + \frac{d_{\max}}{p_{\min}}\right).
\end{aligned}$$

Taking the expectation yields

$$\begin{aligned}
\mathrm{E}\left[\sum_{t=1}^{T} R_i(t)\right] \geq &\left(1-\frac{1}{2C}\right)\cdot\mathcal{L}(\varepsilon^*(C))\cdot T\cdot\lambda^* - O\left(\sum_{n=1}^{T}\left(23\delta_n + \sum_{j\in\mathcal{J}}\exp\left(-p_j 2^n/4|\mathcal{K}|\right)\right)\right)\\
&- O\left(|\mathcal{J}| d_{\max}T^{2/3}\sqrt{\log T\cdot\log(|\mathcal{J}||\mathcal{K}||\mathcal{I}|)} + \varepsilon^*(C)\cdot(d_{max}-C)^+\cdot T^{2/3} + T^{2/3} + d_{\max}^3 + \frac{d_{\max}}{p_{\min}}\right)\\
= &\left(1-\frac{1}{2C}\right)\cdot\mathcal{L}(\varepsilon^*(C))\cdot T\cdot\lambda^*\\
&- O\left(|\mathcal{J}| d_{\max}T^{\frac{2}{3}}\sqrt{\log T\cdot\log(|\mathcal{J}||\mathcal{K}||\mathcal{I}|)} + \varepsilon^*(C)\cdot(d_{max}-C)^+\cdot T^{\frac{2}{3}} + T^{\frac{2}{3}} + d_{\max}^3 + \frac{d_{\max}}{p_{\min}} + \sum_{j\in\mathcal{J}}\frac{|\mathcal{K}|}{p_j}\right).
\end{aligned}$$

Using Lemma 1 which states that $T \cdot \lambda^* \geq T \cdot \mathrm{OPT}^* - d_{\max} \cdot r_{\max}$, then the Theorem 2 follows.

A.6 Proof of Theorem 3

Clearly, when $C \geq d_{\max}$, it follows that $g_n = 0$ and $\mathbf{E}\left[V(t) \mid \mathcal{H}(t_{n-1})\right] = 1$, i.e., resource availability is always ensured. Consequently, for any $i \in \mathcal{I}$, the following inequality holds simultaneously for all $t \in \{t_{n-1}+1+d_{\max}, \ldots, t_n\}$:

$$\mathbf{E}\left[\hat{R}_i(t) \cdot V(t) \mid \mathcal{H}(t_{n-1})\right] \geq \lambda^* - \left(\epsilon_n^A + \epsilon_n^B + \epsilon_n^C + \epsilon_n + f_n\lambda^*\right), \quad \text{w.p. } 1 - 22\delta_n, \tag{47}$$

when $n \geq 3\log\epsilon^*(C) + 2\log d_{\max}$. Following the same analysis as in the proof of Theorem 2, it holds that for all $n \geq 1$:

$$\ell_n \cdot \left(\epsilon_n^A + \epsilon_n^B + \epsilon_n^C + \epsilon_n + f_n\lambda^*\right) \leq O\left(|\mathcal{J}| d_{\max} \cdot \ell_n^{2/3}\sqrt{n \cdot \log(|\mathcal{J}||\mathcal{K}||\mathcal{I}|)} + \frac{d_{\max}}{C} \cdot n + r_{\max} \cdot n + \ell_n^{2/3}\right)$$
$$\text{w.p. } 1 - \sum_{j\in\mathcal{J}} \exp\left(-p_j 2^{2n/3}/(4|\mathcal{K}| d_{\max}^{1/3})\right).$$

Following a similar analysis as in (15), we obtain that with probability at least $1 - 23\delta_n$:

$$\sum_{t=t_{n-1}+d_{\max}+1}^{t_n} \hat{R}_i(t) \cdot V(t) \geq (\ell_n - d_{\max}) \cdot \lambda^* - r_{\max}\sqrt{2 d_{\max} \ell_n \log d_{\max}/\delta_n}$$
$$- r_{\max} d_{\max} - (\ell_n - d_{\max}) \cdot (\epsilon_n^A + \epsilon_n^B + \epsilon_n^C + \epsilon_n + f_n\lambda^*).$$

By leveraging (11), we conclude that with probability at least $1 - \sum_{n=1}^{T}\left(23\delta_n + \sum_{j\in\mathcal{J}} \exp\left(-\frac{p_j 2^{\frac{2n}{3}}}{4|\mathcal{K}| d_{\max}^{1/3}}\right)\right)$:

$$\sum_{t=1}^{T} R_i(t) \geq \lambda^* - O\left(|\mathcal{J}| \cdot d_{\max} T^{2/3}\sqrt{\log T \cdot \log(|\mathcal{J}||\mathcal{K}||\mathcal{I}|)} + T^{2/3} + d_{\max}^3\right).$$

Taking the expectation and applying Lemma 1 yield the desired bound.

A.7 Proof of Theorem 4

Unravelling the definitions of f_n (note that $f_n = 0$) and g_n in Theorem 1 and setting $\varepsilon = \varepsilon^*(C)$ yield the following result:

$$\sum_{t=t_{n-1}+1}^{t_n} R_i(t) \geq \frac{1}{1+\epsilon_n/C} \cdot \mathcal{L}(\varepsilon^*(C)) \cdot \ell_n \cdot \lambda^* - r_{\max}\sqrt{2 d_{\max} \ell_n \log(d_{\max}/\delta_n)} - 2 r_{\max} d_{\max}$$
$$- \ell_n \cdot \left(\epsilon_n^A + \epsilon_n^B + \epsilon_n^C + \epsilon_n\right), \quad \text{w.p. } 1 - 23\delta_n,$$

when $n \geq 3\log\varepsilon^*(C) + 2\log d_{\max}$. Similarly, applying the same unraveling process in Lemma 4, the following inequality holds with probability $1 - 11\delta_n$

$$\mathbf{E}[V(t) \mid \mathcal{H}(t_{n-1})] \geq \left(1 - \frac{1}{2C}\mathcal{L}(\epsilon^*(C))\right)$$

simultaneously for all $t \in \{t_{n-1}+1+d_{\max}, \ldots, t_n\}$. Taking expectation with respect to $\mathcal{H}(t_{n-1})$ and telescope summation across times within epoch $n \geq 1$ gives

$$\sum_{t_{n-1}+1}^{t_n} \mathbf{E}[V(t)] \geq \left(1 - \frac{1}{2C}\right)\mathcal{L}(\epsilon^*(C)) \cdot (\ell_n - d_{\max}).$$

Consequently, we can derive the following lower bound on quantity $N_{j,k}(t_{n-1})$ for any pair (j,k):

$$\begin{aligned}\mathbf{E}[N_{j,k}(t_{n-1})] &= \frac{p_j}{|\mathcal{K}|}\left(d_{\max} + \sum_{i=1}^{n-1}\sum_{t=t_{i-1}+1}^{t_i}\mathbf{E}[V(t)]\right)\\ &\geq \frac{p_j}{|\mathcal{K}|}d_{\max} + \frac{p_j}{|\mathcal{K}|}\left(1-\frac{1}{2C}\right)\mathcal{L}(\epsilon^*(C))\cdot(t_{n-1}-d_{\max}-(n-1)d_{\max})\\ &\geq \frac{p_j}{|\mathcal{K}|}\left(1-\frac{1}{2C}\right)\mathcal{L}(\epsilon^*(C))\cdot(t_{n-1}-(n-1)d_{\max})\,.\end{aligned}$$

Using Chernoff bounds, it holds that

$$\begin{aligned}\mathbf{P}\left(N_{j,k}(n) \leq \left(1-\frac{\sqrt{2}}{2}\right)\cdot\mathbf{E}[N_{j,k}(n)]\right) &\leq \exp\left(-\frac{\mathbf{E}[N_{j,k}(n)]}{4}\right)\\ &\leq \frac{4|\mathcal{K}|}{p_j\left(1-\frac{1}{2C}\right)\mathcal{L}(\epsilon^*(C))\cdot(t_{n-1}-(n-1)d_{\max})}.\end{aligned}$$

Building on this and applying the union bound, the following inequalities hold simultaneously when $n \geq O\left(2\log d_{\max} + \log(1/p_{\min}) + \log\left(\log\left(|\mathcal{I}||\mathcal{J}||\mathcal{K}|\right)\right)\right)$,

$$\begin{aligned}&\epsilon_n \leq 0.5,\\ &\ell_n\cdot\left(\epsilon_n^A+\epsilon_n^B+\epsilon_n^C+\epsilon_n\right)\\ &\leq O\left(|\mathcal{J}|d_{\max}\cdot\ell_n^{1/2}\sqrt{\log(|\mathcal{J}||\mathcal{K}||\mathcal{I}|/\delta_n)}+\frac{d_{\max}}{C}\log(1/\delta_n)+r_{\max}\log(1/\delta_n)\right)\\ &= O\left(|\mathcal{J}|d_{\max}\cdot\ell_n^{1/2}\sqrt{n\cdot\log(|\mathcal{J}||\mathcal{K}||\mathcal{I}|)}+\frac{d_{\max}}{C}\cdot n + r_{\max}\cdot n\right)\\ &\qquad \text{w.p. } 1-\sum_{j\in\mathcal{J}}\frac{4|\mathcal{K}|}{p_j\left(1-\frac{1}{2C}\right)\mathcal{L}(\epsilon^*(C))\cdot(t_{n-1}-(n-1)d_{\max})}.\end{aligned}$$

Thus, with probability at least $1-23\delta_n-\sum_{j\in\mathcal{J}}\frac{4|\mathcal{K}|}{p_j\left(1-\frac{1}{2C}\right)\mathcal{L}(\epsilon^*(C))\cdot(t_{n-1}-(n-1)d_{\max})}$ we have that:

$$\begin{aligned}\sum_{t=t_{n-1}+1}^{t_n}R_i(t) \geq \left(1-\frac{1}{2C}\right)\cdot\mathcal{L}(\varepsilon^*(C))\cdot\ell_n\cdot\lambda^* - O\Big(&|\mathcal{J}|d_{\max}\cdot\ell_n^{1/2}\sqrt{\log(|\mathcal{J}||\mathcal{K}||\mathcal{I}|/\delta_n)}\\ &+\frac{d_{\max}}{C}n + r_{\max}\cdot n + r_{\max}\sqrt{d_{\max}\ell_n\log(d_{\max}/\delta_n)} + r_{\max}d_{\max}\Big),\end{aligned} \tag{48}$$

when $n \geq n_0 = O\left(2\log d_{\max} + 3\log\epsilon^*(C) + \log(1/p_{\min}) + \log\left(\log\left(|\mathcal{I}||\mathcal{J}||\mathcal{K}|\right)\right)\right)$. Summing over n from 0 to $\lceil\log(T/d_{\max})\rceil$ and applying a union bound, we can obtain that with probability at least $1-\sum_{n=1}^{T}(23\delta_n+\sum_{j\in\mathcal{J}}\frac{4|\mathcal{K}|}{p_j(1-1/2C)\cdot\mathcal{L}(\epsilon^*(C))\cdot(t_{n-1}-(n-1)d_{\max})})$:

$$\begin{aligned}\sum_{t=1}^{T}R_i(t) \geq \left(1-\frac{1}{2C}\right)\cdot\mathcal{L}(\varepsilon^*(C))\cdot T\cdot\lambda^* - O\Bigg(\sum_{n=1}^{\log(T/d_{\max})}\Big[&|\mathcal{J}|d_{\max}\cdot\ell_n^{1/2}\sqrt{\log(|\mathcal{J}||\mathcal{K}||\mathcal{I}|/\delta_n)}\\ &+\frac{d_{\max}}{C}n + r_{\max}\cdot n + r_{\max}\sqrt{d_{\max}\ell_n\log(d_{\max}/\delta_n)} + r_{\max}d_{\max}\Big]\Bigg) - O\left(\sum_{n=1}^{n_0}\ell_n\right)\end{aligned}$$

$$\geq \left(1-\frac{1}{2C}\right)\cdot \mathcal{L}(\varepsilon^*(C))\cdot T\cdot \lambda^* - O\left(|\mathcal{J}|\cdot d_{\max}T^{1/2}\sqrt{\log T\cdot \log(|\mathcal{J}||\mathcal{K}||\mathcal{I}|)}+T^{1/2}+d_{\max}^3\right).$$

Taking the expectation yields

$$\begin{aligned}
&\mathbf{E}\left[\sum_{t=1}^{T} R_i(t)\right]\\
&\geq \left(1-\frac{1}{2C}\right)\mathcal{L}(\varepsilon^*(C))\cdot T\lambda^* - O\left(\sum_{n=1}^{T}\left(23\delta_n+\sum_{j\in\mathcal{J}}\frac{4|\mathcal{K}|}{p_j\left(1-\frac{1}{2C}\right)\mathcal{L}(\epsilon^*(C))\left(t_{n-1}-(n-1)d_{\max}\right)}\right)\right)\\
&\quad - O\left(|\mathcal{J}|d_{\max}\cdot T^{1/2}\sqrt{\log T\cdot\log(|\mathcal{J}||\mathcal{K}||\mathcal{I}|)}+d_{\max}^3\right)\\
&= \left(1-\frac{1}{2C}\right)\mathcal{L}(\varepsilon^*(C))\cdot T\lambda^*\\
&\quad - O\left(|\mathcal{J}|d_{\max}\cdot T^{1/2}\sqrt{\log T\cdot\log(|\mathcal{J}||\mathcal{K}||\mathcal{I}|)}+d_{\max}^3+\sum_{j\in\mathcal{J}}\frac{1}{p_j}\cdot\frac{|\mathcal{K}|}{\left(1-\frac{1}{2C}\right)\mathcal{L}(\epsilon^*(C))}\right).
\end{aligned}$$

Applying Lemma 1, which states that $T\cdot\lambda^*\geq T\cdot \mathrm{OPT}^* - d_{\max}\cdot r_{\max}$, completes the proof of the first result in Theorem 4.

For the second result in Theorem 4, note that when $C\geq d_{\max}$, it follows that $g_n=0$ and $\mathbf{E}\left[V(t)\mid\mathcal{H}(t_{n-1})\right]=1$, i.e., resource availability is always ensured. Consequently, the following inequality holds simultaneously for all $t\in\{t_{n-1}+1+d_{\max},\ldots,t_n\}$ and $i\in\mathcal{I}$:

$$\mathbf{E}\left[\hat{R}_i(t)\cdot V(t)\mid\mathcal{H}(t_{n-1})\right]\geq\lambda^*-\left(\epsilon_n^A+\epsilon_n^B+\epsilon_n^C+\epsilon_n\right),\quad \text{w.p. } 1-22\delta_n, \tag{49}$$

when $n\geq 3\log\epsilon^*(C)+2\log d_{\max}$. Additionally, the ensured resource availability implies that:

$$\mathbf{E}[N_{j,k}(t_{n-1})]\geq\frac{p_j}{|\mathcal{K}|}\cdot t_{n-1},$$

Using Chernoff bounds, we obtain:

$$\mathbf{P}\left(N_{j,k}(n)\leq\left(1-\frac{\sqrt{2}}{2}\right)\cdot\mathbf{E}[N_{j,k}(n)]\right)\leq\exp\left(-\frac{\mathbf{E}[N_{j,k}(n)]}{4}\right)\leq\frac{4|\mathcal{K}|}{p_j\cdot t_{n-1}}.$$

Thus, we have that for all $n\geq 1$:

$$\ell_n\cdot\left(\epsilon_n^A+\epsilon_n^B+\epsilon_n^C+\epsilon_n\right)\leq O\left(|\mathcal{J}|d_{\max}\cdot\ell_n^{1/2}\sqrt{n\cdot\log(|\mathcal{J}||\mathcal{K}||\mathcal{I}|)}+\frac{d_{\max}}{C}\cdot n+r_{\max}\cdot n+\ell_n^{1/2}\right)$$
$$\text{w.p. } 1-\sum_{j\in\mathcal{J}}\frac{4|\mathcal{K}|}{p_j\cdot t_{n-1}}.$$

Following a similar analysis as in (15), it follows that with probability at least $1-23\delta_n$:

$$\begin{aligned}
\sum_{t=t_{n-1}+d_{\max}+1}^{t_n}\hat{R}_i(t)\cdot V(t)&\geq(\ell_n-d_{\max})\cdot\lambda^*-r_{\max}\sqrt{2d_{\max}\ell_n\log d_{\max}/\delta_n}\\
&\quad - r_{\max}d_{\max}-(\ell_n-d_{\max})\cdot(\epsilon_n^A+\epsilon_n^B+\epsilon_n^C+\epsilon_n).
\end{aligned}$$

By leveraging (11), we conclude that with probability at least $1 - \sum_{n=1}^{T} \left(23\delta_n + \sum_{j\in\mathcal{J}} \frac{4|\mathcal{K}|}{p_j \cdot t_{n-1}}\right)$:

$$\sum_{t=1}^{T} R_i(t) \geq \lambda^* - O\left(|\mathcal{J}| \cdot d_{\max}\sqrt{T\log T \cdot \log(|\mathcal{J}||\mathcal{K}||\mathcal{I}|)} + d_{\max}^3\right).$$

Taking the expectation and applying Lemma 1 yield the second result in Theorem 4.

B Simulations

In this section, we present a series of simulation experiments conducted in a cloud computing scenario to evaluate the effectiveness of the proposed algorithm. We begin by describing the experimental setting and then summarize and analyze the numerical results.

B.1 Experimental Setup

In our experiments, each incoming request corresponds to a computing task, and the different task types reflect varying workloads or job sizes. In practice, workloads often exhibit heavy-tailed characteristics (e.g., Pareto distributions), where small tasks arrive more frequently but large tasks still make up a substantial portion of the distribution. To capture a simplified version of this phenomenon, we define four task types $|\mathcal{J}| = 4$, with type-1 representing the smallest workload and type-4 representing the largest. We assign the following discrete arrival probabilities for them:

$$\boldsymbol{p} = (p_1, p_2, p_3, p_4) = (0.4, 0.3, 0.2, 0.1).$$

The action space $\mathcal{K}$ includes several different worker nodes (or virtual machines, VMs) to which a task can be assigned, along with the null action for rejection. Specifically, we have four distinct worker nodes $\mathcal{K} = \{1, 2, 3, 4\}$. Each worker k has a different processing speed s_k, which translates into different average execution times for tasks. Higher speed implies shorter average running times (and thus potentially higher utility rates). For simplicity, we set

$$s_1 = 1.0, \quad s_2 = 1.5, \quad s_3 = 2.0, \quad s_4 = 3.0.$$

Let w_j denote the base workload (job size) of task type j. Intuitively, larger j indicates a bigger workload. Suppose

$$w_1 = 3, \quad w_2 = 6, \quad w_3 = 12, \quad w_4 = 18.$$

When a task of type-j is processed by worker k, the *average* running time $d_{j,k}$ is given by

$$d_{j,k} = \frac{w_j}{s_k}.$$

As expected, tasks with larger workload w_j take longer to complete, but this is mitigated by assigning them to faster workers (larger s_k). Table 3 illustrates the specific values of these average durations.

Table 3. Illustrative average running times $d_{j,k} = \frac{w_j}{s_k}$.

Type j	**Worker** $k = 1$	**Worker** $k = 2$	**Worker** $k = 3$	**Worker** $k = 4$
1 (s_k is small)	3	2	1.5	1
2 (s_k is medium)	6	4	3	2
3 (s_k is large)	12	8	6	4
4 (s_k is very large)	18	12	9	6

We focus on two types of utilities $\mathcal{I} = \{1, 2\}$: (i) profit (revenue) and (ii) energy consumption (modeled as a cost or negative utility).

- **Profit** $\left(r_{1,j,k}\right)$: Larger workloads generally yield higher profit, and faster workers typically deliver higher profit rates. We parameterize the profit via a base profit b_j for each task type j and a worker-specific multiplier α_k:

$$r_{1,j,k} \;=\; \alpha_k \, b_j,$$

where

$$b_1 = 3, \quad b_2 = 6, \quad b_3 = 9, \quad b_4 = 12, \quad \text{and} \quad \alpha_1 = 1.0, \quad \alpha_2 = 1.2, \quad \alpha_3 = 1.5, \quad \alpha_4 = 1.8.$$

- **Energy consumption** $\left(r_{2,j,k}\right)$: Energy cost is proportional to the task's duration and the worker's energy consumption rate. Let μ_k denote the per-unit-time energy usage rate of worker k. Then

$$r_{2,j,k} \;=\; \mu_k \cdot d_{j,k},$$

where $d_{j,k} = w_j/s_k$ is the expected running time. In our experimental setup, we set

$$\mu_1 = 1.0, \quad \mu_2 = 1.3, \quad \mu_3 = 1.5, \quad \mu_4 = 2.0.$$

Hence, faster workers have higher power (energy) consumption rates, reflecting the real-world scenarios where there is a trade-off between speed and energy efficiency.

In our simulations, we normalize $\boldsymbol{r}$ as $\boldsymbol{r} = \boldsymbol{r}/r_{\max}$. Note that the actual outcomes $(R_{1,j,k}(t), R_{2,j,k}(t))$ in practice are random and fluctuate around their respective means. To model this randomness, we represent these outcomes as Bernoulli random variables with means $(r_{1,j,k}, r_{2,j,k})$. Similarly, to reflect stochastic variations in job execution, we require $D_{j,k}(t)$ to be stochastic. Since we operate in discrete time, we also require the durations $D_{j,k}(t)$ to be integer values. To achieve this while ensuring $\mathbf{E}[D_{j,k}(t)] = d_{j,k}$, we adopt a simple two-point rounding approach when $d_{j,k}$ is not an integer. Specifically, for each pair (j,k), we define $a = \lfloor d_{j,k} \rfloor$ and $\delta = d_{j,k} - a$. We then draw $D_{j,k}(t)$ as follows:

$$D_{j,k}(t) \;=\; \begin{cases} a+1, & \text{with probability } \delta, \\ a, & \text{with probability } 1-\delta. \end{cases}$$

When $d_{j,k}$ is an integer, $D_{j,k}(t)$ is uniformly distributed over $\{d_{j,k}-1, d_{j,k}, d_{j,k}+1\}$. This approach ensures that $D_{j,k}(t)$ is integer-valued while maintaining the desired expectation $\mathbf{E}[D_{j,k}(t)] = d_{j,k}$.

We assume that the system can process a maximum number of tasks in parallel, limited by the number of GPUs, C, i.e., the system has C reusable resource units. When all C units are occupied, any newly arriving task must be rejected. In our simulations, we vary C from small to large values to observe how the algorithm's performance changes accordingly.

Compared baselines. We implement and compare two baseline algorithms: the offline static algorithm and the hybrid algorithm proposed by [50]. Notably, both baselines are tailored for the offline setting. In particular, the second baseline relies on a linear relationship between utilities and resource usage durations. It is applicable only in the experimental setup where we disregard the first type of utility (i.e., profit) and focus solely on the second type of utility (energy consumption), which is linearly dependent on resource usage durations.

- *Offline static algorithm.* Denote $\{y^*_{j,k}\}_{j,k}$ as an optimal solution for (OFARR). The offline static algorithm selects action k for a type-j arrival with probability $x^*_{j,k}$.
- *Hybrid Algorithm [50].* This algorithm involves solving (OFARR-R), adjusted dynamically based on the current resource occupation to compute a distribution $\boldsymbol{x}(t)$. Allocations are then determined using an adaptive weighting process informed by $\boldsymbol{x}(t)$.

Performance metrics and parameter settings. We run our algorithm for $n = 15$ epochs. We evaluate the empirical performance of the compared algorithms by showing how the empirical competitive ratio

$$\text{Empirical competitive ratio}(n) = \frac{\min_i \sum_{t=1}^{t_n} R_i(t)}{t_n \cdot \lambda^*}$$

evolves as the epoch index n increases under different values of C. Here t_n is the ending time slot of epoch n and λ^* denotes the optimal value of (OFARR), and the length of epoch n is $d_{\max} \cdot 2^n$. Note that $d_{\max} = 19$ in our experimental setup.

B.2 Empirical results

Figure 2 shows the empirical performance of our algorithm for different values of C. We can see from this figure that as the number of epochs increases, the empirical competitive ratios eventually stabilize for all C. Additionally, larger values of C lead to higher competitive ratios. Last, as depicted in the right most plot in Figure 3, when C exceeds $d_{\max}$, our algorithm achieves a competitive ratio of 1, aligning well with our theoretical results.

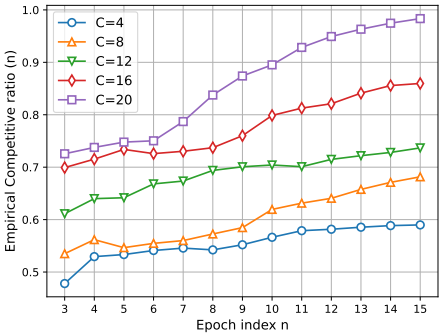


Fig. 2. Empirical performance of our algorithm under various values of C ($|\mathcal{I}| = 2$)

Figure 3 compares the empirical performance of our algorithm with that of the offline static policy. Although the offline static policy achieves fairly consistent utility performance (note that a higher empirical competitive ratio indicates higher utility performance) across various values of C, our algorithm delivers superior utility performance as C increases. This is because, although the offline static policy, by definition, targets a utility rate of λ^* for each type, random fluctuations in resource usage durations lead to temporary resource unavailability, preventing it from consistently meeting the desired utility rate.

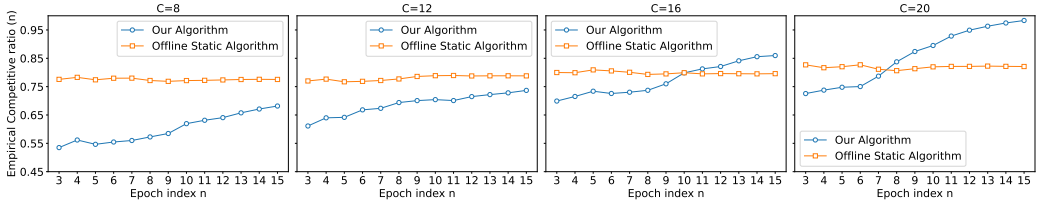


Fig. 3. Empirical performance of our algorithm and the offline static policy ($|\mathcal{I}| = 2$)

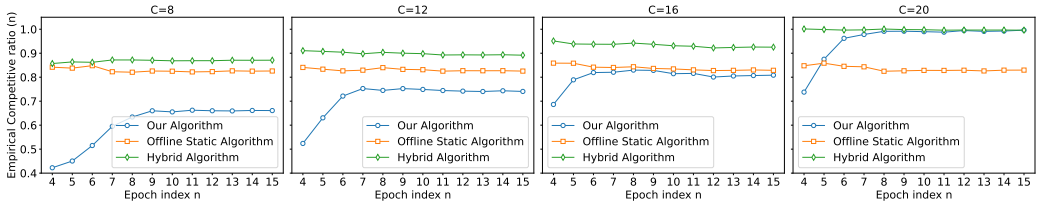


Fig. 4. Empirical performance of all compared algorithms ($|\mathcal{I}| = 1$)

To incorporate more baselines into the comparison, we disregard the first type of utility (i.e., profit) and focus exclusively on the second type of utility (i.e., energy consumption). As a result, the fairness objective simplifies to maximizing the second type of utility. In this scenario, utility depends linearly on resource usage durations, making the hybrid baseline algorithm applicable. Figure 5 presents the empirical performance of our algorithm across different values of C, while Figure 4 compares the empirical performance of our algorithm with that of the baselines. Although the empirical performance of our algorithm is initially worse than the hybrid baseline, which has full knowledge of the problem parameters, the performance gap decreases as the capacity C increases.

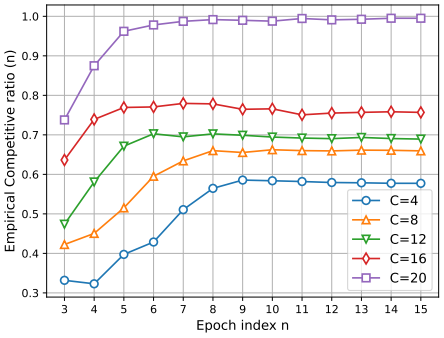


Fig. 5. Empirical performance of our algorithm under various values of C ($|\mathcal{I}| = 1$)

Received January 2025; revised April 2025; accepted April 2025