Bandits with Knapsacks: Advice on Time-Varying Demands

Lixing Lyu [1] **Wang Chi Cheung** [2]

Abstract

We consider a non-stationary Bandits with Knapsack problem. The outcome distribution at each time is scaled by a non-stationary quantity that signifies changing demand volumes. Instead of studying settings with limited non-stationarity, we investigate how online predictions on the total demand volume Q allows us to improve our performance guarantees. We show that, without any prediction, any online algorithm incurs a linear-in-T regret. In contrast, with online predictions on Q, we propose an online algorithm that judiciously incorporates the predictions, and achieve regret bounds that depends on the accuracy of the predictions. These bounds are shown to be tight in settings when prediction accuracy improves across time. Our theoretical results are corroborated by our numerical findings.

1. Introduction

The multi-armed bandit problem (MAB) is a classical model on sequential decision making. The MAB problem features the trade-off between exploration and exploitation, i.e., between exploring for new information about the underlying system and exploiting the potentially optimal solution based on current information. MAB problems have been studied extensively over many decades, with diverse applications such as recommendation systems, ad allocation, resource allocation, revenue management and network routing/scheduling.

Many of these mentioned applications involve resource constraints. For example, a seller experimenting with product prices may have limited product inventory. This motivates the formulation of the *bandits with knapsack* (BwK) problem, an online knapsack problem involving model uncertainty introduced by (Badanidiyuru et al., 2013). The agent has $d \geq 1$ types of resources. At each time step t, the agent pulls an arm, which generates an array of outcomes consisting of the random reward the amounts of the d resources consumed. If some resource(s) are exhausted, then the agent stops pulling any arm. The objective is to maximize the expected total reward over a known horizon T, subject to the budget constraints on the d resources.

The BwK problem was first studied in a stationary stochastic setting, dubbed *stochastic BwK*, where the outcome of an arm follows a stationary but latent probability distribution. (Badanidiyuru et al., 2013; Agrawal & Devanur, 2014) provide online algorithms for stochastic BwK with regret sub-linear in T. The regret is the difference between the optimum and the expected cumulative reward earned by the algorithm, and a sub-linear-in-T regret implies the convergence to opimality as T grows. An alternative setting, *adversarial BwK*, is introduced by (Immorlica et al., 2019). Each arm's outcome distribution can change arbitrarily over time. Contrary to stochastic BwK, it is impossible to achieve a regret sub-linear in T, even when the outcome distribution is changed only once during the horizon (Liu et al., 2022). It begs a question: could a non-stationary BwK problem be more tractable under a less adversarial model than (Immorlica et al., 2019; Liu et al., 2022)?

Practically, while stationary models could be a strong assumption, it could be too pessimistic to assume the underlying model to be adversarily changing and completely latent. Consider the example of ad allocation. On one hand, the internet traffic is constantly changing, leading to a non-stationary model. On the other hand, forecast information is often available. Some advertisements are intrinsically more attractive than others, regardless of the internet traffic, meaning that the click probability of an ad can be regarded as stationary. How could the platform harness the probem structure? Can the allocator utilize forecast information on the internet traffic to improve his/her decisions?

Motivated by the discussions above, we consider the Non-Stationary BwK with Online Advice problem (NS-BwK-OA). An outcome involve an adversarial and a stochastic component. The mean reward and mean type-i resource consumption of pulling arm a at time t are equal to $q_t \cdot r(a)$

[1]Institute of Operations Research and Analytics, National University of Singapore, Singapore [2]Department of Industrial Systems Engineering and Management, National University of Singapore, Singapore. Correspondence to: Lixing Lyu <lyulixing@u.nus.edu>, Wang Chi Cheung <isecwc@nus.edu.sg>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

and $q_t \cdot c(a, i)$ respectively. The quantity $q_t \in \mathbb{R}_{>0}$ is a seasonal term that represents the demand volume at time t, while the stationary quantities $r(a), c(a, i)$ are the mean reward earned and resource consumed per demand unit under arm a. In a dynamic pricing setting, q_t counts the customers arrivals at time t. The quantities $r(a), c(a, i)$ respectively represent the revenue earned and type-i resource consumed per customer under arm a, which could represent a pricing scheme. Our proposed outcome model generalizes the unconstrained settings in (Tracà et al., 2021; Lykouris et al., 2020). We make the following novel contributions:

Model. We incoporate a prediction oracle in NS-BwK-OA, a non-stationary BwK problem. Such an incorporation is novel compared to existing BwK works, which always assume full model uncertainty. In NS-BwK-OA, we identify the total demand volume $Q = \sum_{t=1}^{T} q_t$ as a crucial (but latent) parameter. The oracle provides a prediction $\hat{Q}_t$ to Q at every time step t. The oracle corresponds to how a firm constantly acquires updated information about Q, and a variety of existing time series prediciton tools can be used to construct such an oracle.

Regret Lower Bounds. We derive two regret lower bounds on NS-BwK-OA. First, without the access to a prediction oracle, we show that any online algorithm suffers a linear-in-T regret, even when r, c are known and q_t is known before the arm at time t is to be chosen. Second, with the access of a prediction oracle, we establish regret lower bounds that depend on the accuracy of the estimations. When the estimates are equal to the groundtruth, the regret lower bounds reduce to that of the stochastic BwK problem.

Algorithms and Regret Upper Bounds. We design an online algorithm OA-UCB to utilize the predictions judiciously. OA-UCB is novel in its incorporation of the prediction $\hat{Q}_t$ and demand volume q_t into the estimated opportunity costs of the resources, in relation to the predicted demand volumes. We derive a regret upper bound on OA-UCB that depends on the accuracy of the predictions, even though the accuracy of each prediction is not known to the algorithm. OA-UCB is shown to achieve near optimal regret bounds when the accuracy of the predictions improves across time.

Numerical Validations. We perform numerical experiments when $\{q_t\}_{t=1}^{T}$ is governed by a time series model. The experiment highlights the benefit of predicitons. We show that an online algorithm, such as OA-UCB, that harnesses predictions judiciously can perform empirically better than existing baselines, which only has access to the bandit feedback from the latent environment.

1.1. Literature Review

The Bandits with Knapsacks (BwK) problem has been extensively studied. (Badanidiyuru et al., 2013) first introduced the stochastic BwK problem, which bears applications in dynamic pricing (Besbes & Zeevi, 2009; 2012) and ad allocation (Mehta et al., 2007). The BwK problem is generalized by (Agrawal & Devanur, 2014) to incorporate convex constraints and concave reweards. Several variants are studied, such as the settings of contextual bandits (Agrawal et al., 2016; Badanidiyuru et al., 2014), combinatorial semi-bandits (Sankararaman & Slivkins, 2018).

Non-stationary BwK problems, where the outcome distribution of each arm is changing over time, are studied recently. (Immorlica et al., 2019) achieves a $O(\log T)$ competitive ratio against the best fixed distribution benchmark in an adversarial setting. (Rangi et al., 2018) consider both stochastic and adversarial BwK problems in the single resource case. (Liu et al., 2022) design a sliding window learning algorithm with sub-linear-in-T regret, assuming the amount of non-stationarity is upper bounded and known. A sub-linear-in-T regret non-stationary BwK is only possible in restrictive settings. For example, as shown in (Immorlica et al., 2019; Liu et al., 2022) and our forthcoming Lemma 3.1, for any online algorithm, there exists a non-stationary BwK instance where the outcome distribution only changes oncee during the horizon, for which the algorihtm incurs a linear-in-T regret. Non-stationary stochastic bandits with no resource constraints are studied in (Besbes et al., 2014; Cheung et al., 2019; Zhu & Zheng, 2020), who provide sub-linear-in-T regret bounds in less restrictive non-stationary settings than (Liu et al., 2022), while the amount of non-stationariety, quantified as the variational budget or the number of change points, has to be sub-linear in T. Our work goes in an orthogonal direction. Instead of studying settings with limited non-stationariety, we seek an improved regret bound when the decision maker is endowed with information additional (in the form of prediction oracle) to the online observations.

Our work is also related to a recent stream of work on resource allocation with horizon uncertainty. (Bai et al., 2022; Aouad & Ma, 2022) consider a stochastic resource allcation setting under full model certainty. In their model, the total demand volume is a random variable, whose probability distribution is known to the decision maker, but the *realization* of the total demand volume is not known. (Balseiro et al., 2022) consider an online resource allocation setting with model uncertainty on the horizon, which is closely related to our model uncertainty setting on the total demand volume Q. A cruciall difference between the uncertainty settings in (Balseiro et al., 2022) and ours is that, the former focuses on the case when the DM is provided with *static* advices, while our work complementarily consider the case of *dynamic* advices. More precisely, (Balseiro et al., 2022) consider a model where $q_t \in \{0, 1\}$. They first consider a model when the DM knows $Q \in [Q_{lower}, Q_{upper}]$ but not the actual value of Q at the beginning, and then consider a case when the DM is additionally endowed with a static prediction $\hat{Q}$

at the beginning. In both cases, the performance guarantees are quantified as *competitive ratios* that depends on the static advices. By contrast, our study quantifies the benefit of receiving dyanmically updated advices $\hat{Q}_t$, and pinpoint conditions on $\{\hat{Q}_t\}_{t=1}^T$ that leads to a sublinear-in-T regret.

In terms of the prediction model, our work is related to an active stream of research works on online algorithm design with machine learned advice (Piotr et al., 2019; Mitzenmacher & Vassilvitskii, 2022). While traditional online algorithm research focuses on worst case performance guarantee in full model uncertainty setting, this stream of works focuses on enhancing the performance guarantee when the decision maker (DM) is provided with a machine learned advice at the start of the online dynamics. A variety of results are derived in different settings, such as online caching (Lykouris & Vassilvitskii, 2021), rent-or-buy (Purohit et al., 2018), scheduling (Mitzenmacher, 2019; Lattanzi et al., 2020), online set cover problems (Bamas et al., 2020; Almanza et al., 2021), online matching (Antoniadis et al., 2020). Our research seeks to take a further step, by investigating the case when the DM receives pregressively updated predictions across the horizon, instead of being given a fixed prediction at the beginning.

Lastly, our prediction model is also related to a line of works online optimization with predictions, which concerns improving the performance guarantee with the help of predictions. These predictions are provided to the DM at the beginning of each round sequentially. A variety of full feedback settings are studied in (Rakhlin & Sridharan, 2013a;b; Steinhardt & Liang, 2014; Jadbabaie et al., 2015), and the contextual bandit setting is studied in (Wei et al., 2020). We remark that the abovementioned works do not involve resource constraints, and they are fundamental different from ours, as shown in the forthcoming Lemma 3.1.

Notation. For a positive integer d, we denote $[d] = \{1, \ldots, d\}$. We adopt the $O(\cdot), o(\cdot), \Omega(\cdot)$ notation in (Cormen et al., 2009).

2. Problem Setup

We consider the Non-Stationary Bandit with Knapsack problem with Online Advice (NS-BwK-OA). The nature specifies an NS-BwK-OA problem instance, represented by the tuple $(\mathcal{A}, B, T, \{q_t\}_{t=1}^T, \{P_a\}_{a\in\mathcal{A}})$. We denote $\mathcal{A}$ as the set of K arms. There are d types of resources, and the decision maker (DM) is endowed with $B_i \geq 0$ units of resource i for each $i \in [d]$. The planning horizon consists of T discrete time steps. Following the convention in (Badanidiyuru et al., 2013), we assume for all $i \in [d]$ that $B_i = B = bT$, where b is the normalized budget. At time t, there are q_t units of demands arriving at the DM's platform. For example, q_t can be the number of customers visiting an online shop at time step t, and a time step can be a fifteen minute interval. We assume $q_t \in [\underline{q}, \overline{q}]$, where $0 < \underline{q} < \overline{q}$. The sequence $\{q_t\}_{t=1}^T$ is an arbitrary element of $[\underline{q}, \overline{q}]^T$ fixed by the nature before the online dynamics. The arbitrariness represents the exogenous nature of the demands.

When the DM pulls arm $a \in \mathcal{A}$, the nature samples a vector $(R(a), C(a, 1), \ldots, C(a, d)) \sim P_a$ of random outcomes. The quantity $R(a)$ is the reward earned per demand unit, and $C(a, i)$ is the amount of type i resource consumed per demand unit. The random variables $R(a), C(a, 1), \ldots, C(a, d)$ are supported in $[0, 1]$, and they can be arbitrarily correlated. We denote $r(a) = \mathbb{E}[R(a)], c(a, i) = \mathbb{E}[C(a, i)]$, and $\boldsymbol{r} = (r(a))_{a\in\mathcal{A}}$, $\boldsymbol{c} = (c(a, i))_{a\in\mathcal{A}, i\in[d]}$. To ensure feasiblity, we assume there is a null arm $a_0 \in \mathcal{A}$ such that $R(a_0) = C(a_0, i) = 0$ with certainty for all $i \in [d]$.

At each time t, the DM is provided with a *prediction oracle* $\mathcal{F}_t$. The oracle is a function $\mathcal{F}_t : [\underline{q}, \overline{q}]^{t-1} \to [\underline{q}T, \overline{q}T]$ that provides an estimation $\hat{Q}_t = \mathcal{F}_t(q_1, \ldots, q_{t-1})$ on $Q = \sum_{t=1}^T q_t$ with the past observations $\{q_s\}_{s=1}^{t-1}$. The DM knows $\mathcal{A}, B, T$, and has the access to $\mathcal{F}_t$ in a sequential manner. In contrast, the DM does not know $\{P_a\}_{a\in\mathcal{A}}, \{q_t\}_{t=1}^T$, while the upper bound $\overline{q}$ is known to the DM.

Dynamics. At each time t, three events happen. Firstly, the DM receives a prediction $\hat{Q}_t = \mathcal{F}_t(q_1, \ldots, q_{t-1})$ on Q. Secondly, based on $\hat{Q}_t$ and the history observation, the DM selects arm $A_t \in \mathcal{A}$. Thirdly, the DM observes the feedback consisting of (i) demand volume q_t, (ii) reward earned $q_t R_t$, (iii) resources consumed $\{q_t C_{t,i}\}_{i\in[d]}$. Recall that $(R_t, C_{t,1}, \ldots, C_{t,d}) \sim P_{A_t}$. Then, the DM proceeds to time $t + 1$. If some resource is depleted, *i.e.* $\exists j \in [d]$ such that $\sum_{s=1}^t q_s C_{s,j} > B_j$, then the null arm a_0 is to be pulled in the remaining horizon $t+1, \ldots, T$. We denote the stopping time here as τ. The DM aims to maximize the total reward $\mathbb{E}[\sum_{t=1}^{\tau-1} q_t R_t]$, subject to the resource constraints and model uncertainty.

On q_t. Our feedback model on q_t is more informative than (Lykouris et al., 2020), where none of $q_1, \ldots, q_T$ is observed during the horizon. In contrast, ours is less informative than (Tracà et al., 2021), where $q_1, \ldots, q_T$ are all observed at time 1. Our assumption of observing q_t at the end of time t is mild in online retail settings. For example, the number of visitors to a website within a time interval can be extracted from the electronic records when the interval ends.

While the nature sets $\{q_t\}_{t=1}^T$ to be fixed but arbitrary, the sequence is set without knowing the DM's online algorithm and prediciton oracle $\mathcal{F} = \{\mathcal{F}_t\}_{t=1}^T$. Our model is milder than the *oblivious adversary* model, where the nature sets a latent quantity (in this case $\{q_t\}_{t=1}^T$) *with* the knowledge of the DM's algorithm before the online dynamics. Our milder model allows the possibility of $\hat{Q}_t = \mathcal{F}_t(q_1, \ldots, q_{t-1})$ being a sufficiently accurate (to be quantified in our main results) estimate to Q for each t, for example when $\{q_t\}_{t=1}^T$ is govenred by a latent time series model. In contrary, an oblivious adversary can set Q to be far away from the predictions $\hat{Q}_1, \ldots, \hat{Q}_T$ in response to the information on $\mathcal{F}$.

On $\mathcal{F}$. Our prediction oracle is a general Black-Box model. We do not impose any structural or parameteric assumption on $\mathcal{F}$ or $\{q_t\}_{t=1}^T$. It is instructive to understand $\mathcal{F}$ as a side information provided to the DM by an external source. In the dynamic pricing example, $\hat{Q}_t$ could be an estimate on the customer base population provided by an external marketing research firm. A prime candidate of $\mathcal{F}$ is the cornucopia of time series prediction models proposed in decades of research works on time series (Shumway & Stoffer, 2017; Hyndman & Athanasopoulos, 2021; Lim & Zohren, 2021). These prediction models allow *one step prediction*, where for any t, the predictor $\mathcal{P}$ inputs $\{q_s\}_{s=1}^{t-1}$ and outputs an estimate $\hat{q}_t$ on q_t. The prediction $\hat{Q}_t$ can be constructed by (1) iteratively applying $\mathcal{P}$ on $\{q_s\}_{s=1}^{t-1} \cup \{\hat{q}\}_t^{t+\rho-1}$ to output $\hat{q}_{t+\rho}$, for $\rho \in \{0, \ldots, T-t\}$, (2) summing over $q_1, \ldots, q_{t-1}, \hat{q}_t, \ldots, \hat{q}_T$ and return $\hat{Q}_t$. We provide an example in the forthcoming Section 5.

Regret. To measure the performance of an algorithm, we define the regret of an algorithm as

$$\text{Regret}_T = \text{OPT} - \sum_{t=1}^{\tau-1} q_t R_t, \tag{1}$$

where OPT denotes the expected cumulative reward of an offline optimal dynamic policy given all latent information and all adversarial terms. For analytical tractabililty in our regret upper bound, we consider an alternative benchmark

$$\begin{aligned} \text{OPT}_{\text{LP}} = \max_{\boldsymbol{u} \in \Delta_K} \quad & \left(\sum_{t=1}^T q_t \right) \boldsymbol{r}^\top \boldsymbol{u} \\ \text{s.t.} \quad & \left(\sum_{t=1}^T q_t \right) \boldsymbol{c}^\top \boldsymbol{u} \le B \mathbf{1}_d, \end{aligned} \tag{2}$$

where $\Delta_K = \{\boldsymbol{w} \in [0,1]^d : \sum_{a \in \mathcal{A}} w_a = 1\}$. The benchmark (2) is justified by the following Lemma:

Lemma 2.1. *$OPT_{LP} \ge OPT$.*

The proof of Lemma 2.1 is in Appendix.

3. Regret Lower Bounds

In this section, we provide impossibility results on the NS-BwK-OA in the form of regret lower bounds. Firstly, we show that a linear-in-T regret is inevitable in the absence of the prediction oracle $\mathcal{F}$.

Lemma 3.1. *Consider a fixed but arbitrary online algorithm that knows $\{P_a\}_{a \in \mathcal{A}}$, $\{(q_s, q_s R_s, q_s C_{s,1}, \ldots q_s C_{s,d})\}_{s=1}^{t-1}$ and q_t, but does not have any access to a prediction oracle when the action A_t is to be chosen at each time t. There exists an instance such that the online algorithm suffers $Regret_T = \Omega(T)$.*

Lemma 3.1 is proved in Appednix B.1. Lemma 3.1 shows that even when all model information on time steps $1, \ldots, t$ are revealed when A_t is to be chosen, the DM still suffers $\text{Regret}_T = \Omega(T)$. Thus, NS-BwK-OA is fundamentally different from non-stationary bandits without resource constraints such as (Besbes et al., 2015), and online optimization with predictions problems such as (Rakhlin & Sridharan, 2013a). In these settings, we can achieve $\text{Regret}_T = 0$ if all model information on time steps $1, \ldots, t$ are available at the time point of choosing A_t or the action at time t. Indeed, given all model information at time t, the DM achieve the optimum by choosing an arm or an action that maximizes the reward function of time t for every $t \in [T]$.

In view of Lemma 3.1, we seek to understand if the DM can avoid $\text{Regret}_T = \Omega(T)$ when s/he is endowed with an accurate prediction on Q. Certainly, if the DM only recieves an uninformative prediction, such as a worst case prediction $\bar{q}T$, at each time step, $\text{Regret}_T = \Omega(T)$ still cannot be avoided. In contrast, if the DM received an *accurate* prediction at a time step, we demonstrate our first step for deriving a better regret bound, in the form of a more benign regret lower bound compared to Lemma 3.1. We formalize the notion of being *accurate* by the following two concepts.

For $T_0 \in [T-1]$ and $\epsilon_{T_0+1} \ge 0$, an instance $\{q_t\}_{t=1}^T$ is said to be $(T_0+1, \epsilon_{T_0+1})$-*well estimated by oracle* $\mathcal{F}$, if the prediction $\hat{Q}_{T_0+1} = \mathcal{F}_{T_0+1}(q_1, \ldots, q_{T_0})$ returned by the oracle at time T_0+1 satisfies $|Q - \hat{Q}_{T_0+1}| \in [\epsilon_{T_0+1}, 2\epsilon_{T_0+1}]$. This notion measures the power of prediction oracle $\mathcal{F}$. We say that ϵ_{T_0+1} is $(T_0+1, \{q_t\}_{t=1}^{T_0})$-*well response* by oracle $\mathcal{F}$ if ϵ_{T_0+1} satisfies $\epsilon_{T_0+1} \le \min\{\hat{Q}_{T_0+1} - \sum_{s=1}^{T_0} q_s - \underline{q}(T-T_0), \bar{q}(T-T_0) - (\hat{Q}_{T_0+1} - \sum_{t=1}^{T_0} q_t), \hat{Q}_{T_0+1}/2\}$, where $\hat{Q}_{T_0+1} = \mathcal{F}_{T_0+1}(q_1, \ldots, q_{T_0})$. This concept imposes requirements on the power of prediction by introducing a non-trivial upper bound on ϵ_{T_0+1} for the "well-estimate" notion. This can help us eliminate trivial and uninformative predictions such as $\hat{Q}_t = 0$ or $\bar{q}T$.

Theorem 3.2. *Consider the NS-BwK-OA setting, and consider a fixed but arbitrary online algorithm and prediciton oracle $\mathcal{F} = \{\mathcal{F}_t\}_{t=1}^T$. For any $T_0 \in [T-1]$ and any $\epsilon_{T_0+1} > 0$ that is $(T_0+1, \{q_t\}_{t=1}^{T_0})$-well response, there exists a $(T_0+1, \epsilon_{T_0+1})$-well estimated instance $I = \{q_t\}_{t=1}^{T_0} \cup \{q_t\}_{t=T_0+1}^T$ such that*

$$Regret_T = \Omega\left(\max\left\{ \frac{1}{Q} \sum_{t=1}^{T_0} q_t \epsilon_{T_0+1} \ , \ \Lambda \right\} \right), \tag{3}$$

where $Q = \sum_{t=1}^{T} q_t$*, and*

$$\Lambda = \min\left\{ OPT, OPT\sqrt{\frac{\overline{q}K}{B}} + \sqrt{\overline{q}K OPT} \right\}.$$

Theorem 3.2 is proved in Appendix B.2. In (3), the regret lower bound Λ is due to the uncertainty on $\{P_a\}_{a\in\mathcal{A}}$, and Λ is derived directly from (Badanidiyuru et al., 2013). The regret lower bound $\frac{1}{Q}\sum_{t=1}^{T_0} q_t \epsilon_{T_0+1}$ is due to the oracle's estimation error on $\hat{Q}_{T_0+1}$. Theorem 3.2 demonstrates a more benign regret lower bound than $\Omega(T)$, under the condition that the prediction on Q is sufficiently accurate (as formalized as $(T_0+1, \epsilon_{T_0+1})$-well estimated).

More specifically, let us consider the following *accurate prediction condition* at time T_0 by oracle $\mathcal{F}$: ϵ_{T_0+1} is $(T_0 + 1, \{q_t\}_{t=1}^{T_0})$-well response by oracle $\mathcal{F}$ and

$$\frac{\epsilon_{T_0+1}}{Q} = O(T_0^{-\alpha}) \quad \text{for some } \alpha > 0. \tag{4}$$

The condition implies that, for the prediction $\hat{Q}_{T_0+1}$ made using T_0 data points $q_1, \ldots, q_{T_0}$, it holds that $|1 - (\hat{Q}_{T_0+1}/Q)| = O(T_0^{-\alpha})$. For example, when $\{q_t\}_{t=1}^{T}$ are i.i.d. generated, the accurate prediction condition holds with $\alpha = 1/2$.

Corollary 3.3. *Consider the setting of Theorem 3.2. Suppose the accurate prediction condition (4) holds at* T_0*, then the refined regret lower bound* $Regret_T = \Omega(\max\{\underline{q}T_0^{1-\alpha}, \Lambda\})$ *holds.*

Altogether, under the accurate prediction condition, the corollary presents a strictly smaller regret lower bound than that in Lemma 3.1, which has no prediction oracle available. In complement, we design and analyze an online algorithm in the next section that reaps the benefits of predictions, and in particular nearly matches the regret lower bound in Corollary 3.3 under the accurate prediction condition. Thus, a $o(T)$-regret is possible in a non-stationary environment given accurate predictions as prescribed above, even though the amount of non-stationarity in the underlying model is not bounded in general.

4. Algorithm and Analysis

We propose the Online-Advice-UCB (OA-UCB) algorithm, displayed in Algorithm 1, for solving NS-BwK-OA. The algorithm design involve constructing confidence bounds to address the model uncertainty on $\boldsymbol{r}, \boldsymbol{c}$, as discussed in Section 4.1. In Section 4.2, we elaborate on OA-UCB, which uses Online Convex Optimization (OCO) tools to balance the trade-off among rewards and resources. Crucially, at each time t, we incorporate the prediction $\hat{Q}_t$ to scale the opportunity costs of the resources. In addition, both q_t and $\hat{Q}_t$ are judiciously integrated into the OCO tools to factor the demand volumes into the consideration of the abovemention trade-off. In Section 4.3, we provide a regret upper bound to OA-UCB, and demonstrate its near-optimality when the accurate prediction condition (4) holds and when capacity is large. In Section 4.4 we provide a sketch proof of the regret upper bound, where the complete proof is in Appendix C.

4.1. Confidence Bounds

We consider the following confidence radius function:

$$\text{rad}(v, N, \delta) = \sqrt{\frac{2v\log\left(\frac{1}{\delta}\right)}{N}} + \frac{4\log\left(\frac{1}{\delta}\right)}{N}. \tag{5}$$

The function (5) satisfies the following property:

Lemma 4.1 ((Babaioff et al., 2015; Agrawal & Devanur, 2014)). *Let random variables* $\{V_i\}_{i=1}^{N}$ *be independently distributed with support in* $[0, 1]$*. Denote* $\hat{V} = \frac{1}{N}\sum_{i=1}^{N} V_i$*, then with probability* $\geq 1 - 3\delta$*, we have*

$$\left|\hat{V} - \mathbb{E}[\hat{V}]\right| \leq rad(\hat{V}, N, \delta) \leq 3rad(\mathbb{E}[\hat{V}], N, \delta).$$

We prove Lemma 4.1 in Appendix D.1 by following the line of argument in (Babaioff et al., 2015) for the purpose of extracting the values of the coefficients in (5), which are implicit in (Babaioff et al., 2015; Agrawal & Devanur, 2014). Based on the observation $\{R_s, \{C_{s,i}\}_{i\in[d]}\}_{s\in[t-1]}$, we compute the sample means

$$\hat{R}_t(a) = \frac{1}{N_{t-1}^{+}(a)} \sum_{s=1}^{t-1} R_s \mathbf{1}_{\{A_s = a\}}, \quad \forall a \in \mathcal{A},$$

$$\hat{C}_t(a, i) = \frac{1}{N_{t-1}^{+}(a)} \sum_{s=1}^{t-1} C_{s,i} \mathbf{1}_{\{A_s = a\}}, \ \forall a \in \mathcal{A}, \ i \in [d],$$

where $N_{t-1}^{+}(a) = \max\{\sum_{s=1}^{t-1} \mathbf{1}_{\{A_s=a\}}, 1\}$. In line with the principle of Optimism in Face of Uncertatinty, we construct upper confidence bounds (UCBs) for the rewards and lower confidence bounds (LCBs) for resource consumption ammounts. For each $a \in \mathcal{A}$, we set $\text{UCB}_{r,t}(a) =$

$$\min\left\{\hat{R}_t(a) + \text{rad}(\hat{R}_t(a), N_{t-1}^{+}(a), \delta), 1\right\}. \tag{6}$$

For each $a \in \mathcal{A}, i \in [d]$, we set $\text{LCB}_{c,t}(a, i) =$

$$\max\left\{\hat{C}_t(a, i) - \text{rad}(\hat{C}_t(a, i), N_{t-1}^{+}(a), \delta), 0\right\}. \tag{7}$$

The design of the UCBs and LCBs are justified by Lemma 4.1 and the model assumption that $r(a), c(a, i) \in [0, 1]$ for all $a \in \mathcal{A}, i \in [d]$:

Lemma 4.2. *With probability* $\geq 1 - 3KTd\delta$, *we have*

$$UCB_{r,t}(a) \geq r(a), \;\; LCB_{c,t}(a,i) \leq c(a,i)$$

for all $a \in \mathcal{A}, i \in [d]$.

Lemma 4.2 is proved in Appendix D.2.

4.2. Details on OA-UCB

OA-UCB is presented in Algorithm 1. At each time step t, the algorithm first computes a composite reward term

$$\mathrm{UCB}_{r,t}(a) - \frac{\hat{Q}_t}{B} \cdot \boldsymbol{\mu}_t^\top \mathbf{LCB}_{c,t}(a), \tag{8}$$

where $\mathrm{UCB}_{r,t}(a), \hat{Q}_t$ and $\mathbf{LCB}_{c,t}(a)$ are the surrogates for the latent $r(a), Q, \mathbf{c}(a)$ respectively. The term $\frac{\hat{Q}_t}{B} \cdot \boldsymbol{\mu}_t^\top \mathbf{LCB}_{c,t}(a)$ can be interpreted as the opportunity cost of the resources. The scalarization

$$\boldsymbol{\mu}_t \in S = \{\boldsymbol{\mu} : \|\boldsymbol{\mu}\|_1 \leq 1, \boldsymbol{\mu} \geq \mathbf{0}_d\} \tag{9}$$

weighs the relative importance of the resources. The factor $\hat{Q}_t/B$ reflects that the opportunity cost increases with $\hat{Q}_t$, since with a higher total demand volume, the DM is more likely to exhaust some of the resources during the horizon, and similar reasoning holds for B. Altogether, (8) balances the trade-off between the reward of an arm and the opportunity cost of that arm's resource consumption. We select an arm that maximizes (8) at time t.

After receiving the feedback, We update the scalarization $\boldsymbol{\mu}_t$ via the Online Gradient Descent (OGD) (Agrawal & Devanur, 2014; Hazan et al., 2016) on the sequence of functions $\{f_t\}_{t=1}^T$, where

$$f_t(\boldsymbol{x}) = \frac{q_t \hat{Q}_t}{B} \left(\frac{B}{\hat{Q}_t} \mathbf{1}_d - \mathbf{LCB}_{c,t}(A_t) \right)^\top \boldsymbol{x}. \tag{10}$$

While f_t incorporates the prediction $\hat{Q}_t$ for the purpose of accounting for the estimated opportunity cost similar to (8), f_t also incoporates the actual demand q_t for accounting the actual amounts of resources consumed. In the OGD update in (11), for a resource type i, the coefficient $\mu_{t+1}(i)$ increases with $q_t \mathrm{LCB}_{c,t}(A_t, i)$, meaning that a higher amount of resource i consumption at time t leads to a higher weightage of resource i's opportunity cost at time $t+1$.

4.3. Performance Guarantees of OA-UCB

The following theorem provides a high-probability regret upper bound for Algorithm 1:

Theorem 4.3. *Consider the OA-UCB algorithm, that is provided with predictions that satisfy* $|\hat{Q}_t - Q| \leq \epsilon_t$ *for all* $t \in [T]$. *With probability* $\geq 1 - 3KTd\delta$,

$$\begin{aligned} & \mathit{OPT}_{\mathit{LP}} - \sum_{t=1}^{\tau-1} q_t R_t \\ \leq & O\left(\left(\mathit{OPT}_{\mathit{LP}} \sqrt{\frac{\bar{q}K}{B}} + \sqrt{\bar{q}K\mathit{OPT}_{\mathit{LP}}} \right) \log\left(\frac{1}{\delta}\right) \right. \quad (12) \\ & \left. + \left(\frac{1}{Q} + \frac{1}{B} \right) \sum_{t=1}^{\tau-1} q_t \epsilon_t + M\sqrt{T} \right). \quad (13) \end{aligned}$$

Algorithm 1 Online-advice-UCB (OA-UCB)

Initialize $\boldsymbol{\mu}_1 = \frac{1}{d}\mathbf{1}_d$, $M = \left(\bar{q} + \frac{\bar{q}^2}{b}\right)\sqrt{d}$, $\eta_t = \frac{\sqrt{2}}{M\sqrt{t}}$.
for $t = 1, 2, ..., T$ **do**
  Receive $\hat{Q}_t = \mathcal{F}_t(q_1, \ldots, q_{t-1})$.
  Compute $\mathrm{UCB}_{r,t}(a)$, $\mathbf{LCB}_{c,t}(a)$ for all $a \in \mathcal{A}$ by (6), (7), where $\mathbf{LCB}_{c,t}(a) = (\mathrm{LCB}_{c,t}(a,i))_{i\in[d]}$.
  Select
  $$A_t \in \operatorname*{argmax}_{a\in\mathcal{A}} \left\{ \mathrm{UCB}_t^{(r)}(a) - \tfrac{\hat{Q}_t}{B} \cdot \boldsymbol{\mu}_t^\top \mathbf{LCB}_{c,t}(a) \right\}.$$
  if $\exists j \in [d]$ such that $\sum_{s=1}^t q_s C_{t,j} > B$ **then**
    Break, and pull the null arm a_0 all the way.
  end if
  Observe q_t, receive reward $q_t R_t$, and consume $q_t C_{t,i}$ for each resource $i \in [d]$.
  Update $\boldsymbol{\mu}_{t+1}$ with OGD. $\boldsymbol{\mu}_{t+1}$ is set to be
  $$\Pi_S \left(\boldsymbol{\mu}_t - \eta_t \frac{q_t \hat{Q}_t}{B} \left(\frac{B}{\hat{Q}_t} \mathbf{1}_d - \mathbf{LCB}_{c,t}(A_t) \right) \right), \tag{11}$$
  where S is defined in (9).
end for

Theorem 4.3 is proved in the Appendix, and we provide a sketch proof in Section 4.4. The Theorem holds in the special case when we set $\epsilon_t = |\hat{Q}_t - Q|$, and ϵ_t represents an upper bound on the estimation error of $\hat{Q}_t$ on Q, for example by certain theoretical guarantees. The term (12) represents the regret due to the learning on $r(a), c(a,i)$. The first term in (13) represents the regret due to the prediction error of the prediction oracle, and the second term in (13) represents the regret due to the application of OGD.

Comparison between regret lower and upper bounds. The regret term (12) matches the lower bound term Λ in Theorem 3.2 within a logarithmic factor. Next, we compare the regret upper bound term $(\frac{1}{Q} + \frac{1}{B}) \sum_{t=1}^{\tau-1} q_t \epsilon_t$ and the lower bound term $\frac{1}{Q} \sum_{t=1}^{T_0} q_t \epsilon_{T_0+1}$ in Theorem 3.2. We first assure that the lower and upper bound results are consistent, in the sense that our regret upper bound is indeed in $\Omega(\frac{1}{Q} \sum_{t=1}^{T_0} q_t \epsilon_{T_0+1})$ on the lower bounding instances constructed for the proof of Theorem 3.2. In those instances, T_0 is set in a way that the resource is not fully exhausted at time T_0 under any policy, thus the stopping time τ of OA-UCB satisfies $\tau > T_0$ with certainty. More details are provided in Appendix B.3.

Next, we highlight that the regret upper and lower bounds are nearly matching (modulo multiplicative factors of $\log(1/\delta)$ and $\overline{q}/\underline{q}$, as well as the additive $O(M\sqrt{T})$ term), under the high capacity condition $B = \Theta(Q)$ and the accurate prediction condition (4) for all $T_0 \geq 1$. The first condition is similar to the large capacity assumption in the literature (Besbes & Zeevi, 2009; 2012; Liu et al., 2022), while the second condition is a natural condition that signfies a non-trivial estimation by the prediciton oracle, as discussed in Section 3. On one hand, By setting $T_0 = \Theta(T)$ for the highest possible lower bound in Corollary 3.3, we yield the regret lower bound $\Omega(\max\{\underline{q}T_0^{1-\alpha}, \Lambda\}) = \Omega(\max\{\underline{q}T^{1-\alpha}, \Lambda\})$. On the other hand, the second term in (13) is upper bounded as

$$\left(\frac{1}{Q}+\frac{1}{B}\right)\sum_{t=1}^{\tau-1} q_t\epsilon_t = O\left(\sum_{t=1}^{\tau-1} q_t\frac{\epsilon_t}{Q}\right)$$
$$= O\left(\sum_{t=1}^{\tau-1} q_t t^{-\alpha}\right) = O(\overline{q}T^{1-\alpha}).$$

Altogether, our claim on the nearly matching bounds is established.

4.4. Proof Sketch of Theorem 4.3

We provide an overview on the proof of Theorem 4.3, which is fully proved in Appendix C. We first provide bounds on the regret induced by the estimation errors of the UCBs and LCBs. Now, with probability $\geq 1-3KTd\delta$, the inequalities

$$\left|\sum_{t=1}^{\tau-1} q_t \mathrm{UCB}_{r,t}(A_t) - \sum_{t=1}^{\tau-1} q_t R_t\right|$$
$$\leq O\left(\log\left(\frac{1}{\delta}\right)\left(\sqrt{\overline{q}K\sum_{t=1}^{\tau-1} q_t R_t} + \overline{q}K\log\left(\frac{T}{K}\right)\right)\right), \tag{14}$$

$$\left|\sum_{t=1}^{\tau-1} q_t \mathrm{LCB}_{c,t}(A_t, i) - \sum_{t=1}^{\tau-1} q_t C_{t,i}\right|$$
$$\leq O\left(\log\left(\frac{1}{\delta}\right)\left(\sqrt{\overline{q}KB} + \overline{q}K\log\left(\frac{T}{K}\right)\right)\right) \ \forall i \in [d] \tag{15}$$

hold. Inequalities (14, 15) are proved in Appendix D.3. Next, by the optimality of A_t in Line 1 in Algorithm 1,

$$\mathrm{UCB}_{r,t}(A_t) - \frac{\hat{Q}_t}{B}\cdot \boldsymbol{\mu}_t^\top \mathbf{LCB}_{c,t}(A_t)$$
$$\geq \mathbf{UCB}_{r,t}^\top \boldsymbol{u}^* - \frac{\hat{Q}_t}{B}\cdot \boldsymbol{\mu}_t^\top \mathbf{LCB}_{c,t}\boldsymbol{u}^*,$$

which is equivalent to

$$\mathbf{UCB}_{r,t}^\top \boldsymbol{u}^* - \mathrm{UCB}_{r,t}(A_t) + \frac{\hat{Q}_t}{B}\boldsymbol{\mu}_t^\top\left(\frac{B}{\hat{Q}_t}\mathbf{1}_d - \mathbf{LCB}_{c,t}^\top \boldsymbol{u}^*\right)$$
$$\leq \frac{\hat{Q}_t}{B}\cdot \boldsymbol{\mu}_t^\top\left(\frac{B}{\hat{Q}_t}\mathbf{1}_d - \mathbf{LCB}_{c,t}(A_t)\right).$$

Multiply q_t on both side, and sum over t from 1 to $\tau-1$. By applying the OGD performance guarantee in (Hazan et al., 2016) with $\{f_t\}_{t=1}^T, S$ defined in (10, 9) respectively, we argue that, for all $\boldsymbol{\mu} \in S$,

$$\sum_{t=1}^{\tau-1} q_t \mathbf{UCB}_{r,t}^\top \boldsymbol{u}^* - \sum_{t=1}^{\tau-1} q_t \mathrm{UCB}_{r,t}(A_t)$$
$$+\sum_{t=1}^{\tau-1} q_t\frac{\hat{Q}_t}{B}\cdot \boldsymbol{\mu}_t^\top\left(\frac{B}{\hat{Q}_t}\mathbf{1}_d - \mathbf{LCB}_{c,t}^\top \boldsymbol{u}^*\right)$$
$$\leq \sum_{t=1}^{\tau-1} q_t\frac{\hat{Q}_t}{B}\cdot\left(\frac{B}{\hat{Q}_t}\mathbf{1}_d - \mathbf{LCB}_{c,t}(A_t)\right)^\top \boldsymbol{\mu} + O\left(M\sqrt{T}\right).$$

If $\tau \leq T$, then there exists $j_0 \in [d]$ such that $\sum_{t=1}^{\tau} q_t C_{t,j_0} > B$. Take $\boldsymbol{\mu} = \frac{\mathrm{OPT_{LP}}}{Q}\boldsymbol{e}_{j_0} \in S$ (This is because $\mathrm{OPT_{LP}} = Q\boldsymbol{r}^\top \boldsymbol{u}^* \leq Q$). Analysis yields

$$\mathrm{OPT_{LP}} - \sum_{t=1}^{\tau-1} q_t \mathrm{UCB}_{r,t}(A_t) \leq O\left(\log\left(\frac{1}{\delta}\right)\mathrm{OPT_{LP}}\sqrt{\frac{\overline{q}K}{B}}\right.$$
$$\left.+\left(\frac{1}{Q}+\frac{1}{B}\right)\sum_{t=1}^{\tau-1} q_t\epsilon_t + M\sqrt{T}\right). \tag{16}$$

If $\tau > T$, it is the case that $\tau - 1 = T$, and no resource is exhausted at the end of the horizon. Take $\boldsymbol{\mu} = \mathbf{0}$. Similar analaysis to the previous case shows that $\mathrm{OPT_{LP}} - \sum_{t=1}^{\tau-1} q_t \mathrm{UCB}_{r,t}(A_t) \leq$

$$O\left(\frac{1}{Q}\sum_{t=1}^{\tau} q_t\epsilon_t + M\sqrt{T}\right). \tag{17}$$

Combine (14), (16), (17) and the fact that $\mathrm{OPT_{LP}} \geq \sum_{t=1}^{\tau-1} q_t R_t$, the theorem holds.

5. Numerical Experiments

We present numerical results, and compare our algorithm with several existing algorithms on BwK.

Demand sequence $\{q_t\}_{t=1}^T$: We apply an AR(1) model to generate $\{q_t\}$:

$$q_t = \alpha + \beta q_{t-1} + \varepsilon_t,$$

where $\varepsilon_1, \dots \varepsilon_T \sim N(0, \sigma^2)$ are independent.

Estimations $\{\hat{Q}_t\}_{t=1}^T$: At round t, given history observation $\{q_s\}_{s=1}^{t-1}$, there are many time series prediction tools in Python, MatLab or R that perform predictions to yield $\{\hat{q}_s\}_{s=t}^T$, where $\hat{q}_s$ is a estimate on q_s. We define the estimation $\hat{Q}_t$ as $\sum_{s=1}^{t-1} q_s + \sum_{s=t}^T \hat{q}_s$. To achieve time-efficiency, we consider a "power-of-two" policy for updating the $\hat{Q}_t$ on Q, as shown in Algorithm 2. That is, we only recompute $\hat{Q}_t$ when $t = 2^k$ for some $k \in \mathbb{N}^+$. The estimation error of Algorithm 2 in terms of additive gap is plotted in Figure 1.

Algorithm 2 Estimation Generation Policy

Input: Time step t, history observation $\{q_s\}_{s=1}^{t-1}$, previous estimation $\hat{Q}_{t-1}$.

1: **if** $t = 2^k$ for some $k \in \mathbb{N}^+$ **then**
2: Compute predictions $\{\hat{q}_s\}_{s=t}^T$, and update $\hat{Q}_t = \sum_{s=1}^{t-1} q_s + \sum_{s=t}^T \hat{q}_s$.
3: **else**
4: Set $\hat{Q}_t = \hat{Q}_{t-1}$.
5: **end if**
6: **return** $\hat{Q}_t$.

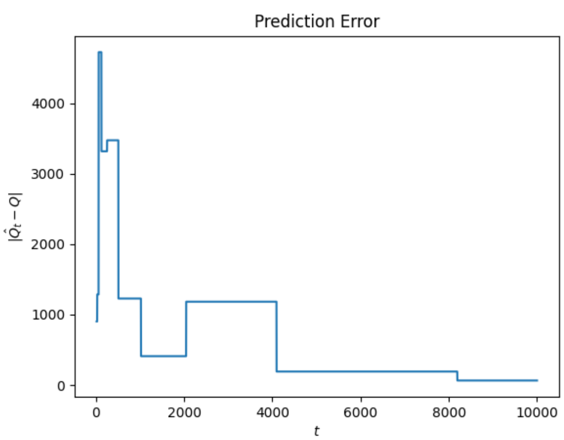

Figure 1. Estimation error

Benchmarks: We compare OA-UCB with three existing online algorithms for BwK. The first algorithm is the PrimalDualBwK algorithm in (Badanidiyuru et al., 2013), which we call "PDB" in the following. The second algorithm is the UCB algorithm presented in (Agrawal & Devanur, 2014), which we call "AD-UCB". The third algorithm is the Sliding-Window UCB in (Liu et al., 2022), which we call "SW-UCB". In implementing SW-UCB, we set the sliding window size according to the suggestion in (Liu et al., 2022), and we input the required non-stationarity measures by computing them from the ground truth $\{q_t\}_{t=1}^{15000}$.

In the experiment, we simulate our algorithm and the benchmarks on a family of instances, with $K = 10$, $d = 3$, $b = 3$, $\alpha = 2$, $\beta = 0.5$, $\sigma = 0.5$, and T varies from 5000 to 15000. Each arms's per-unit-demand outcome $(R(a), \{C_i(a)\}_{i=1}^d)$ follows the standard Gaussian distribution truncated in $[0, 1]^{d+1}$, which has mean denoted as $(r(a), \boldsymbol{c}(a))$. We perform two groups of the experiment. In each group, we first generate a sample of $(\boldsymbol{r}, \boldsymbol{c}, \{q_t\}_{t=1}^{15000})$. Then, for each fixed T, we simulate each algorithm ten times with demand volume sequence $\{q_t\}_{t=1}^T$, and compute the regret based on the sample average.

Figures 2 and 3 plots the regret of each algorithm on different horizon lengths, in each of the two groups. The superiority in numerical performance for OA-UCB does not mean that our algorithm is strictly superior to the baselines. Indeed, our algorithm OA-UCB receives online advice by Algorithm 2, while the benchmarks do not. The numerical results instead indicate the benefit of prediciting the underlying non-stationary demand sequence, and showcase how a suitably designed algoirhtm such as OA-UCB could reap the benefit of predictions.

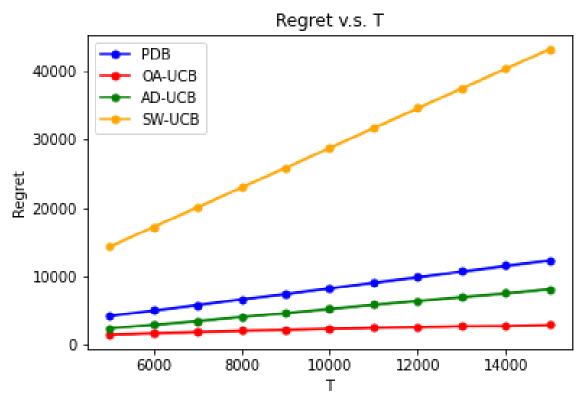

Figure 2. Regret on Experiment Group 1

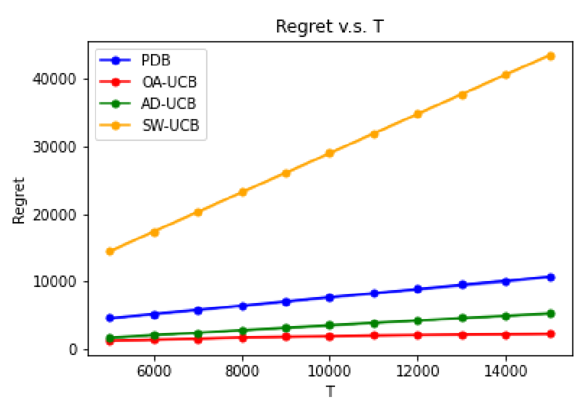

Figure 3. Regret on Experiment Group 2

Figure 4 depict the trend of the accumulated rewards as time progresses, with horizon $T = 10000$. The black dotted lines indicate the stopping times of each algorithm respectively. Compared with our algorithm OA-UCB, PDB and SW-UCB could appear conservative, meaning that they focus too much on not exhausting resources. In contrast, AD-UCB seems a little aggressive, meaning that it prefers to choose the arm with high reward and resource consumption.

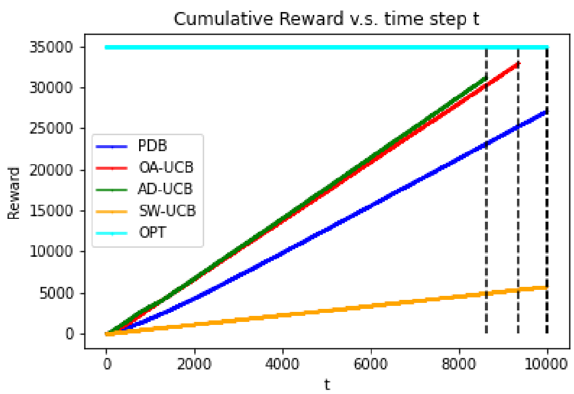

Figure 4. Cumulative Reward growing.

6. Conclusion

We study a non-stationary bandit with knapsack problem, in the presense of a prediction oracle on the latent total demand volume Q. Our oracle is novel compared to existing models in online optimization with machine-learned advice (such as (Lykouris & Vassilvitskii, 2021)), in that ours returns a (possibly refined) prediction on Q every time step. There are many interesting future directions, such as investigating the models (Bamas et al., 2020; Lykouris & Vassilvitskii, 2021; Purohit et al., 2018; Mitzenmacher, 2019) in the presense of sequential prediction oracles similar to ours. It is also interesting to investigate other forms of predictions, such as predictions with distributional information (Bertsimas et al., 2019; Diakonikolas et al., 2021). Customizing prediction oracles for NS-BwK-OA is also an interesting direction (Anand et al., 2020).

Acknowledgements

The authors would like to thank the reviewing team for the constructive suggestions that improve the positioning of the paper and the clarity in the exposition. This research work is funded by Institute of Operations Research and Analytics, National University of Singapore, a Singapore Ministry of Education AcRF Tier 2 Grant (Grant number: T2EP20121-0035) and a Singapore Ministry of Education AcRF Tier 3 Grant (Grant number: MOE-2019-T3-1-010).

References

Agrawal, S. and Devanur, N. R. Bandits with concave rewards and convex knapsacks. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pp. 989–1006, 2014.

Agrawal, S., Devanur, N. R., and Li, L. An efficient algorithm for contextual bandits with knapsacks, and an extension to concave objectives. In *Conference on Learning Theory*, pp. 4–18. PMLR, 2016.

Almanza, M., Chierichetti, F., Lattanzi, S., Panconesi, A., and Re, G. Online facility location with multiple advice. *Advances in Neural Information Processing Systems*, 34: 4661–4673, 2021.

Anand, K., Ge, R., and Panigrahi, D. Customizing ML predictions for online algorithms. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 303–313. PMLR, 13–18 Jul 2020.

Antoniadis, A., Gouleakis, T., Kleer, P., and Kolev, P. Secretary and online matching problems with machine learned advice. *Advances in Neural Information Processing Systems*, 33:7933–7944, 2020.

Aouad, A. and Ma, W. A nonparametric framework for online stochastic matching with correlated arrivals. *arXiv preprint arXiv:2208.02229*, 2022.

Audibert, J.-Y., Munos, R., and Szepesvári, C. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19): 1876–1902, 2009.

Babaioff, M., Dughmi, S., Kleinberg, R., and Slivkins, A. Dynamic pricing with limited supply, 2015.

Badanidiyuru, A., Kleinberg, R., and Slivkins, A. Bandits with knapsacks. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pp. 207–216. IEEE, 2013.

Badanidiyuru, A., Langford, J., and Slivkins, A. Resourceful contextual bandits. In *Conference on Learning Theory*, pp. 1109–1134. PMLR, 2014.

Bai, Y., El Housni, O., Jin, B., Rusmevichientong, P., Topaloglu, H., and Williamson, D. Fluid approximations for revenue management under high-variance demand. *SSRN*, 2022. URL https://ssrn.com/abstract=4136445.

Balseiro, S., Kroer, C., and Kumar, R. Online resource allocation under horizon uncertainty. *arXiv preprint arXiv:2206.13606*, 2022.

Bamas, É., Maggiori, A., and Svensson, O. The primal-dual method for learning augmented algorithms. *Advances in Neural Information Processing Systems*, 33:20083–20094, 2020.

Bertsimas, D., Sim, M., and Zhang, M. Adaptive distributionally robust optimization. *Management Science*, 65(2): 604–618, 2019.

Besbes, O. and Zeevi, A. Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Operations Research*, 57(6):1407–1420, 2009.

Besbes, O. and Zeevi, A. Blind network revenue management. *Operations research*, 60(6):1537–1550, 2012.

Besbes, O., Gur, Y., and Zeevi, A. Stochastic multi-armed-bandit problem with non-stationary rewards. *Advances in neural information processing systems*, 27, 2014.

Besbes, O., Gur, Y., and Zeevi, A. Non-stationary stochastic optimization. *Operations research*, 63(5):1227–1244, 2015.

Cheung, W. C., Simchi-Levi, D., and Zhu, R. Learning to optimize under non-stationarity. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1079–1087. PMLR, 2019.

Chung, F. and Lu, L. Concentration inequalities and martingale inequalities: a survey. *Internet mathematics*, 3(1): 79–127, 2006.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. *Introduction to Algorithms, 3rd Edition*. MIT Press, 2009. ISBN 978-0-262-03384-8.

Diakonikolas, I., Kontonis, V., Tzamos, C., Vakilian, A., and Zarifis, N. Learning online algorithms with distributional advice. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2687–2696. PMLR, 18–24 Jul 2021.

Dubhashi, D. P. and Panconesi, A. *Concentration of measure for the analysis of randomized algorithms*. Cambridge University Press, 2009.

Freedman, D. A. On tail probabilities for martingales. *the Annals of Probability*, pp. 100–118, 1975.

Hazan, E. et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

Hyndman, R. and Athanasopoulos, G. *Forecasting: principles and practice*. OTexts, 2021. URL https://otexts.com/fpp3/.

Immorlica, N., Sankararaman, K. A., Schapire, R., and Slivkins, A. Adversarial bandits with knapsacks. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 202–219. IEEE, 2019.

Jadbabaie, A., Rakhlin, A., Shahrampour, S., and Sridharan, K. Online optimization: Competing with dynamic comparators. In *Artificial Intelligence and Statistics*, pp. 398–406. PMLR, 2015.

Lattanzi, S., Lavastida, T., Moseley, B., and Vassilvitskii, S. Online scheduling via learned weights. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1859–1877. SIAM, 2020.

Lim, B. and Zohren, S. Time-series forecasting with deep learning: a survey. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379(2194), 2021.

Liu, S., Jiang, J., and Li, X. Non-stationary bandits with knapsacks. *arXiv preprint arXiv:2205.12427*, 2022.

Lykouris, T. and Vassilvitskii, S. Competitive caching with machine learned advice. *Journal of the ACM (JACM)*, 68 (4):1–25, 2021.

Lykouris, T., Mirrokni, V., and Leme, R. P. Bandits with adversarial scaling. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6511–6521, 2020.

Mehta, A., Saberi, A., Vazirani, U., and Vazirani, V. Adwords and generalized online matching. *Journal of the ACM (JACM)*, 54(5):22–es, 2007.

Mitzenmacher, M. Scheduling with predictions and the price of misprediction. *arXiv preprint arXiv:1902.00732*, 2019.

Mitzenmacher, M. and Vassilvitskii, S. Algorithms with predictions. *Commun. ACM*, 65(7):33–35, jun 2022.

Piotr, I., Yaron, S., Ali, V., and Sergei, V. Summer workshop on learning-based algorithms. 2019. URL http://www.mit.edu/~vakilian/ttic-workshop.html.

Purohit, M., Svitkina, Z., and Kumar, R. Improving online algorithms via ml predictions. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Rakhlin, A. and Sridharan, K. Online learning with predictable sequences. In *Conference on Learning Theory*, pp. 993–1019. PMLR, 2013a.

Rakhlin, S. and Sridharan, K. Optimization, learning, and games with predictable sequences. *Advances in Neural Information Processing Systems*, 26, 2013b.

Rangi, A., Franceschetti, M., and Tran-Thanh, L. Unifying the stochastic and the adversarial bandits with knapsack. *arXiv preprint arXiv:1811.12253*, 2018.

Sankararaman, K. A. and Slivkins, A. Combinatorial semi-bandits with knapsacks. In *International Conference on Artificial Intelligence and Statistics*, pp. 1760–1770. PMLR, 2018.

Shumway, R. and Stoffer, D. *Time Series Analysis and Its Applications: With R Examples*. Springer texts in statistics. Springer, 2017. URL https://github.com/nickpoison/tsa4/blob/master/textRcode.md.

Steinhardt, J. and Liang, P. Adaptivity and optimism: An improved exponentiated gradient algorithm. In *International Conference on Machine Learning*, pp. 1593–1601. PMLR, 2014.

Tracà, S., Rudin, C., and Yan, W. Regulating greed over time in multi-armed bandits. *J. Mach. Learn. Res.*, 22: 3–1, 2021.

Wei, C.-Y., Luo, H., and Agarwal, A. Taking a hint: How to leverage loss predictors in contextual bandits? In *Conference on Learning Theory*, pp. 3583–3634. PMLR, 2020.

Zhu, F. and Zheng, Z. When demands evolve larger and noisier: Learning and earning in a growing environment. In *International Conference on Machine Learning*, pp. 11629–11638. PMLR, 2020.

A. Proof for Section 2

A.1. Proof for Lemma 2.1

Let's first consider

$$\mathrm{OPT}'_{\mathrm{LP}} = \max_{\boldsymbol{x}_t \in \Delta_{|\mathcal{A}|},\ \forall t \in [T]} \sum_{t=1}^{T} q_t \boldsymbol{r}^\top \boldsymbol{x}_t \qquad \text{s.t.} \quad \sum_{t=1}^{T} q_t \boldsymbol{c}^\top \boldsymbol{x}_t \le B\boldsymbol{1}_d, \tag{18}$$

where $\Delta_{|\mathcal{A}|} = \{\boldsymbol{w} : \sum_{a\in\mathcal{A}} w_a = 1\}$ is the probability simplex across all arms. It is evident that $\mathrm{OPT}'_{\mathrm{LP}} \ge \mathrm{OPT}$, since for a fixed policy π that achieves OPT, the solution $\bar{\boldsymbol{x}} = \{\bar{x}_{t,a}\}_{t\in[T],a\in\mathcal{A}}$ defined as

$$\bar{x}_{t,a} = \mathbb{E}[\mathbf{1}(\text{action } a \text{ is chosen at } t \text{ under } \pi)]$$

is feasible to OPT', and the objective value of $\bar{\boldsymbol{x}}$ in $\mathrm{OPT}'_{\mathrm{LP}}$ is equal to the expected revenue earned in the online process.

Next, we claim that $\mathrm{OPT}_{\mathrm{LP}} = \mathrm{OPT}'_{\mathrm{LP}}$. Indeed, for each feasible solution $(\boldsymbol{x}_t)_{t\in[T]}$ to $\mathrm{OPT}'_{\mathrm{LP}}$, the solution

$$\boldsymbol{u} = \frac{\sum_{t=1}^{T} q_t \boldsymbol{x}_t}{\sum_{t=1}^{T} q_t},$$

is feasible to $\mathrm{LP}_{\mathrm{OPT}}$ and has the same objective value as $(\boldsymbol{x}_t)_{t\in[T]}$. Altogether, the Lemma is proved.

B. Proofs for Section 3, and Consistency Remarks

In this section, we provide proofs to the lower bound results. In both proofs, we consider an arbitrary but fixed deterministic online algorithm, that is, conditioned on the realization of the history in $1, \ldots, t-1$ and q_t, $\hat{Q}_t$, the chosen arm A_t is deterministic. This is without loss of generality, since the case of random online algorithm can be similarly handled by replace the chosen arm A_t with a probability distribution over the arms, but we focus on deterministic case to ease the exposition. Lastly, in Section B.3 we demonstrate that our regret upper and lower bounds are consistent on the lower bounding instances we constructed in Section B.2.

B.1. Proof for Lemma 3.1

Our lower bound example involve two instances $I^{(1)}, I^{(2)}$ with determinstic rewards and deterministic consumption amounts. Both instances involve two non-dummy arms $1, 2$ in addition to the null arm a_0, and there is $d = 1$ resource type. Instances $I^{(1)}, I^{(2)}$ differ in their respective seqeunces of demand volumes $\{q_t^{(1)}\}_{t=1}^T, \{q_t^{(2)}\}_{t=1}^T$, but for other parameters are the same in the two instances.

In both $I^{(1)}, I^{(2)}$, arm 1 is associated with (deterministc) reward $r(1) = 1$ and (deterministic) consumption amount $c(1,1) = 1$, while arm 2 is associated with (deterministc) reward $r(2) = 3/4$ and (deterministic) consumption amount $c(2,1) = 1/2$. Both instances share the same horizon T, a positive even integer, and the same capacity $B = T/2$. The sequences of demand volumes $\{q_t^{(1)}\}_{t=1}^T, \{q_t^{(2)}\}_{t=1}^T$ of instances $I^{(1)}, I^{(2)}$ are respectively defined as

$$q_t^{(1)} = \begin{cases} 1 & \text{if } t \in \{1, \ldots, T/2\}, \\ 1/16 & \text{if } t \in \{T/2+1, \ldots, T\}, \end{cases}$$
$$q_t^{(2)} = 1, \quad \text{for all } t \in \{1, \ldots, T\}.$$

Then the optimal reward for $I^{(1)}$ is at least $\frac{T}{2}$ (always select the arm 1 until the resource is fully consumed), and the optimal reward for $I^{(2)}$ is $\frac{3T}{4}$ (always select arm 2 until the resource is fully consumed).

Consider the first $T/2$ rounds, and consider an arbitrary online algorithm that knows $\{P_a\}_{a\in\mathcal{A}}$, $\{(q_s, q_s R_s, q_s C_{s,1}, \ldots q_s C_{s,d})\}_{s=1}^{t-1}$ and q_t when the action A_t is to be chosen at each time t. Under this setting, the DM recieves the same set of observations in the first $T/2$ time steps in each of instances $I^{(1)}, I^{(2)}$. Consequently, the sequence of arm pulls in the first $T/2$ time steps are the same. Now, we denote $N_a = \sum_{t=1}^{T/2} \mathbf{1}(A_t = a)$ for $a \in \{1, 2\}$. By the previous remark, N_a is the number of times arm a is pulled during time steps $1, \ldots, T/2$ in each of the two instances.

Observe that $N_1 + N_2 \le \frac{T}{2}$, which implies $N_1 \le \frac{T}{4}$ or $N_2 \le \frac{T}{4}$. We denote $\text{Reward}_T(I^{(i)}), \text{Regret}_T(I^{(i)})$ as the expected reward and the expected regret of the policy in instance $I^{(i)}$. In what follows, we demonstrate that

$$\max_{i \in \{1,2\}} \text{Regret}_T(I^{(i)}) \ge \frac{T}{32}, \tag{19}$$

which proves the Lemma.

Case 1: $N_1 \le \frac{T}{4}$**.** We consider the algorithm on $I^{(1)}$, which earns

$$\text{Reward}_T(I^{(1)}) \le \frac{T}{4} \cdot 1 + \frac{T}{4} \cdot \frac{3}{4} + \frac{T}{2} \frac{1}{16} = \frac{15}{32} T.$$

Hence,

$$\text{Regret}_T(I^{(1)}) \ge \frac{T}{2} - \text{Reward}_T(I^{(1)}) \ge \frac{1}{32} T.$$

Case 2: $N_2 \le \frac{T}{4}$**.** We consider the algorithm on $I^{(2)}$, which earns

$$\text{Reward}_T(I^{(2)}) = \left(\frac{T}{2} - N_2\right) \cdot 1 + N_2 \cdot \frac{3}{4} + \left(\frac{T}{2} - \left(\frac{T}{2} - N_2\right) \cdot 1 - N_2 \cdot \frac{1}{2}\right) \cdot \frac{\frac{3}{4}}{\frac{1}{2}} = \frac{T}{2} + \frac{N_2}{4} \le \frac{9}{16} T.$$

Hence,

$$\text{Regret}_T(I^{(2)}) \ge \frac{3}{4} T - \text{Reward}_T(I^{(2)}) \ge \frac{3}{16} T.$$

Altogether, the inequality (19) is shown.

B.2. Proof for Theorem 3.2

By the Theorem's assumption that $\epsilon_{T_0+1} > 0$ is $(T_0 + 1, \{q_t\}_{t=1}^{T_0})$-*well response* by $\mathcal{F} = \{\mathcal{F}_t\}$, we know that

$$0 < \epsilon_{T_0+1} \le \min \left\{ \hat{Q}_{T_0+1} - \sum_{t=1}^{T_0} q_t - \underline{q}(T - T_0), \overline{q}(T - T_0) - \hat{Q}_{T_0+1} - \sum_{t=1}^{T_0} q_t, \frac{\hat{Q}_{T_0+1}}{2} \right\}, \tag{20}$$

where $\hat{Q}_{T_0+1} = \mathcal{F}_{T_0+1}(q_1, \ldots, q_{T_0})$. To proceed, we take a demand volume seqeunce $\{q_t\}_{t=1}^{T_0} \in [\underline{q}, \overline{q}]^{T_0}$ that satisfies the assumption. In what follows, we first construct two deterministic instances $I^{(1)}, I^{(2)}$ which only differ in their respective seqeunces of demand volumes $\{q_t^{(1)}\}_{t=T_0+1}^{T}, \{q_t^{(2)}\}_{t=T_0+1}^{T}$, but the two instances are the same on other parameters, and that $q_t^{(1)} = q_t^{(2)} = q_t$ for $t \in \{1, \ldots, T_0\}$. Both $I^{(1)}, I^{(2)}$ only involve one resource constraint. We estbalish the Theorem by showing three claims:

1. Both $I^{(1)}, I^{(2)}$ are $(T_0 + 1, \epsilon_{T_0+1})$-well-estimated by $\mathcal{F}$, and the underlying online algorithm and prediction oracle (which are assumed to be fixed but arbitrary in the Theorem statement) suffer

$$\text{Regret}_T(I^{(i)}) \ge \frac{\sum_{t=1}^{T_0} q_t \epsilon_{T_0+1}}{6Q^{(i)}} \quad \text{for some } i \in \{1, 2\}. \tag{21}$$

   In (21), we define $\text{Regret}_T(I^{(i)})$ as the regret of the algorithm on instance $I^{(i)}$, and $Q^{(i)} = \sum_{t=1}^{T} q_t^{(i)}$.

2. Among the set of instances $\{J_c^{(i)}\}_{i \in [K]}$ (see **Instances** $\{J_c^{(i)}\}_{i \in [K]}$), the online algorithm suffers

$$\text{Regret}_T(J_c^{(i)}) \ge \frac{1}{128} \min \left\{ 1, \sqrt{\frac{K\overline{q}}{B}} \right\} \text{opt}(J_c^{(i)}) \quad \text{for some } i \in [K], \tag{22}$$

   where $\text{opt}(I)$ denote the optimum of instance I, even when the DM has complete knowledge on $q_1, \ldots, q_T$, and $\hat{Q}_t$ is equal to the ground truth Q in each of the instances in $\{J_c^{(i)}\}_{i \in [K]}$.

3. Among the set of instances $\{J_r^{(i)}\}_{i\in[K]}$ (see **Instances** $\{J_r^{(i)}\}_{i\in[K]}$), the online algorithm suffers

$$\text{Regret}_T(J_r^{(i)}) \geq \frac{1}{20}\sqrt{\bar{q}K\text{opt}(J_r^{(i)})} \quad \text{for some } i\in[K], \tag{23}$$

even when the DM has complete knowledge on $q_1,\ldots,q_T$, and $\hat{Q}_t$ is equal to the ground truth Q in each of the instances in $\{J_r^{(i)}\}_{i\in[K]}$.

Once we establish inequalities (21, 22, 23), the Theorem is shown. We remark that (22, 23) are direct consequences of (Badanidiyuru et al., 2013). We first extract the instances $\{J_c^{(i)}\}_{i\in[K]}$, $\{J_r^{(i)}\}_{i\in[K]}$ that are constructed in (Badanidiyuru et al., 2013), then we construct the instances $I^{(1)}, I^{(2)}$. After that, we prove (21), which establish the Theorem.

Instances $\{J_c^{(i)}\}_{i\in[K]}$**.** These instances are single resource instances, with determinsitic rewards but stochastic consumption. According to (Badanidiyuru et al., 2013), we first set parameters

$$\eta = \frac{1}{32}\min\left\{1, \sqrt{\frac{K}{B}}\right\}, \quad T = \frac{16B}{\eta(1/2-\eta)},$$

and set $q_t = \bar{q}$ for all $t\in[T]$. The instances $J_c^{(1)},\ldots,J_c^{(K)}$ share the same $B, T, \{q_t\}_{t=1}^T$, and the instances share the same (deterministic) reward function:

$$R(a) = r(a) = \begin{cases} 1 & \text{if } a\in[K]\setminus\{a_0\} \\ 0 & \text{if } a = a_0 \end{cases}.$$

In contrast, instances $J_c^{(1)},\ldots,J_c^{(K)}$ differ in the resource consumption model. We denote $C^{(i)}(a)$ as the random consumption of arm a in instance $J_c^{(i)}$. The probability distribution of $C^{(i)}(a)$ for each $a, i\in[K]$ is defined as follow:

$$C^{(i)}(a) \sim \begin{cases} \text{Bern}(1/2) & \text{if } a\in[K]\setminus\{a_0, i\} \\ \text{Bern}(1/2-\eta) & \text{if } a = i \\ \text{Bern}(0) & \text{if } a = a_0 \end{cases},$$

where Bern(p) denotes the Bernoulli distribution with mean d. The regret lower bound (22) is a direct consequence of Lemma 6.10 in (Badanidiyuru et al., 2013), by incorporating the scaling factor $\bar{q}$ into the rewards earned by the DM and the optimal reward.

Instances $\{J_r^{(i)}\}_{i\in[K]}$**.** These instances are single resource instances, with random rewards but deterministic consumption. These instances share the same $B, T > K$ (set arbitrarily), the same demand volume seqeunce, which is $q_t = \bar{q}$ for all $t\in[T]$, and the same resource consumption model, in that $c(a) = 0$ for all $a\in\mathcal{A}$. These instances only differ in the reward distributions. We denote $R^{(i)}(a)$ as the random reward of arm a in instance $J_r^{(i)}$. The probability distribution of $R^{(i)}(a)$ for each $a, i\in[K]$ is defined as follow:

$$R^{(i)}(a) \sim \begin{cases} \text{Bern}\left(\frac{1}{2}-\frac{1}{4}\sqrt{\frac{K}{T}}\right) & \text{if } a\in[K]\setminus\{a_0, i\} \\ \text{Bern}(1/2) & \text{if } a = i \\ \text{Bern}(0) & \text{if } a = a_0 \end{cases}.$$

The regret lower bound (23) is a direct consequence of Claim 6.2a in (Badanidiyuru et al., 2013), by incorporating the scaling factor $\bar{q}$ into the rewards earned by the DM and the optimal reward.

Construct $I^{(1)}, I^{(2)}$**.** We first describe $\{q_t^{(1)}\}_{t=1}^T, \{q_t^{(2)}\}_{t=1}^T$. As previously mentioned, for $t\in\{1,\ldots,T_0\}$, we have $q_t^{(1)} = q_t^{(2)} = q_t$. To define $q_t^{(1)}, q_t^{(2)}$ for $t\in\{T_0+1,\ldots,T\}$, first recall that $|\hat{Q}_{T_0+1} - Q| \geq \epsilon_{T_0+1}$, where ϵ_{T_0+1} satisfies (20). By (20), we know that

$$\underline{q}(T-T_0) \leq \hat{Q}_{T_0+1} - \sum_{t=1}^{T_0} q_t - \epsilon_{T_0+1} < \hat{Q}_{T_0+1} - \sum_{t=1}^{T_0} q_t + \epsilon_{T_0+1} \leq \bar{q}(T-T_0)$$

We set $q^{(1)}_{T_0+1} = \ldots = q^{(1)}_T \in [\underline{q}, \overline{q}]$ and $q^{(2)}_{T_0+1} = \ldots = q^{(2)}_T \in [\underline{q}, \overline{q}]$ such that based on current instance $\{q_t\}_{t=1}^{T_0}$ we have ever received, we construct the following two subsequent instances $I^{(1)} = \{q^{(1)}_t\}_{t=T_0+1}^T$, $I^{(2)} = \{q^{(2)}_t\}_{t=T_0+1}^T$, such that

$$Q^{(1)} = \sum_{t=1}^T q^{(1)}_t = \hat{Q}_{T_0+1} - \epsilon_{T_0+1}, \quad Q^{(2)} = \sum_{t=1}^T q^{(2)}_t = \hat{Q}_{T_0+1} + \epsilon_{T_0+1},$$

which is valid by the stated inequalities.

Next, we define the parameters $\{r(a)\}_{a\in\mathcal{A}}, \{c(a,1)\}_{a\in\mathcal{A}}, B$, and recall that $d = 1$. Similar to the proof for Lemma 3.1, we only consider deterministic instances, so it is sufficient to define the mean rewards and consumption amounts. To facilitate our discussion, we specify $\mathcal{A} = [K] = \{1, 2, \ldots, K\}$, with $K \geq 3$ and arm K being the null arm. The parameters $\{r(a)\}_{a\in\mathcal{A}}, \{c(a,1)\}_a, B$ shared between instances $I^{(1)}, I^{(2)}$ are defined as follows:

$$r(a) = \begin{cases} 1 & \text{if } a = 1, \\ (1+c)/2 & \text{if } a = 2, \\ 0 & \text{if } a \in \{3, \ldots, K\}, \end{cases}$$

and

$$c(a,1) = \begin{cases} 1 & \text{if } a = 1, \\ c & \text{if } a = 2, \\ 0 & \text{if } a \in \{3, \ldots, K\}, \end{cases}$$

where

$$c = \frac{\hat{Q}_{T_0+1} - \epsilon_{T_0+1}}{\hat{Q}_{T_0+1} + \epsilon_{T_0+1}}.$$

Finally, we set

$$B = \hat{Q}_{T_0+1} - \epsilon_{T_0+1}.$$

Inequality (20) ensures that $c, B > 0$.

Proving (21). To evaluate the regrets in the two instances, we start with the optimal reawrds. The optimal reward in $I^{(1)}$ is $\hat{Q}_{T_0+1} - \epsilon_{T_0+1}$, which is achieved by pulling arm 1 until the resource is exhasuted. The optimal reward for $I^{(2)}$ is $\hat{Q}_{T_0+1}$, which is achieved by pulling arm 2 until the resource is exhasuted.

Consider the execution of the fixed but arbitrary online algorithm during time steps $1, \ldots, T_0$ in each of the instances. The prediction oracle returns the same prediction $\hat{Q}_t$ for $t \in \{1, \ldots, T_0\}$ in both instances, since both instances share the same r, c, B, T and $q^{(1)}_t = q^{(2)}_t$ for $t \in \{1, \ldots, T_0\}$. Consequently, the fixed but arbitrary online algorithm has the same sequence of arm pulls $A_1, \ldots, A_{T_0}$ during time steps $1, \ldots, T_0$ in both instances $I^{(1)}, I^{(2)}$. Now, for each arm $i \in \{1, 2\}$, we define $N_i = \{t \in \{1, \ldots, T_0\} : A_t = i\}$, which has the same realization in instances $I^{(1)}, I^{(2)}$. Since $N_1 \cup N_2 \subseteq [T_0]$, at least one of the cases $\sum_{t\in N_1} q_t \leq \frac{1}{2}\sum_{s=1}^{T_0} q_s$ or $\sum_{t\in N_2} q_t \leq \frac{1}{2}\sum_{s=1}^{T_0} q_s$ holds.

We denote $\text{Reward}_T(I^{(i)})$ as the expected reward of the online algorithm in instance $I^{(i)}$. We proceed with the following case consideration:

Case 1: $\sum_{t\in N_1} q_t \leq \frac{1}{2}\sum_{s=1}^{T_0} q_s$. We consider the online algorithm's execution on $I^{(1)}$, which yields

$$\begin{aligned} \text{Reward}_T(I^{(1)}) &\leq \frac{\sum_{s=1}^{T_0} q_s}{2} \cdot 1 + \frac{\sum_{s=1}^{T_0} q_s}{2} \cdot \frac{1}{2}(1+c) + \left(\hat{Q}_{T_0+1} - \epsilon_{T_0+1} - \sum_{s=1}^{T_0} q_s\right) \cdot 1 \\ &= \left(\sum_{s=1}^{T_0} q_s\right)\left(-\frac{1}{4} + \frac{1}{4}c\right) + \hat{Q}_{T_0+1} - \epsilon_{T_0+1}. \end{aligned}$$

Hence,

$$\text{Regret}_T(I^{(1)}) \geq \left(\sum_{s=1}^{T_0} q_s\right) \cdot \frac{1}{4}(1-c) = \frac{\sum_{s=1}^{T_0} q_s \epsilon_{T_0+1}}{2(\hat{Q}_{T_0+1} + \epsilon_{T_0+1})} \geq \frac{\sum_{s=1}^{T_0} q_s \epsilon_{T_0+1}}{6(\hat{Q}_{T_0+1} - \epsilon_{T_0+1})} = \frac{\sum_{s=1}^{T_0} q_s \epsilon_{T_0+1}}{6Q^{(1)}},$$

where the last inequality is by the *well response* condition that guarantees that $2\epsilon_{T_0+1} \leq \hat{Q}_{T_0+1}$. For the last equality, recall $\hat{Q}_{T_0+1} - \epsilon_{T_0+1} = \sum_{t=1}^{T} q_t^{(1)}$.

Case 2: $\sum_{t\in N_2} q_t \leq \frac{1}{2}\sum_{s=1}^{T_0} q_s$. We consider the online algorithm's execution on $I^{(2)}$, which yields

$$\begin{aligned}
\text{Reward}_T(I^{(2)}) &\leq \left(\sum_{s=1}^{T_0} q_s - \sum_{t\in N_2} q_t\right)\cdot 1 + \sum_{t\in N_2} q_t \cdot \frac{1}{2}(1+c) + \left(B - \left(\sum_{s=1}^{T_0} q_s - \sum_{t\in N_2} q_t\right) - \sum_{t\in N_2} q_t \cdot c\right)\cdot \frac{\frac{1}{2}(1+c)}{c} \\
&= \sum_{s=1}^{T_0} q_s\left(1 - \frac{1+c}{2c}\right) + \left(\sum_{s\in N_2} q_s\right)\left[-1 + \frac{1+c}{2} + \frac{1+c}{2c} - \frac{1+c}{2}\right] + B\cdot\frac{1+c}{2c} \\
&= -\frac{\sum_{s=1}^{T_0} q_s \epsilon_{T_0+1}}{\hat{Q}_{T_0+1} - \epsilon_{T_0+1}} + \left(\sum_{s\in N_2} q_s\right)\cdot\frac{\epsilon_{T_0+1}}{\hat{Q}_{T_0+1} - \epsilon_{T_0+1}} + \hat{Q}_{T_0} \\
&\leq -\frac{\sum_{s=1}^{T_0} q_s \epsilon_{T_0+1}}{2(\hat{Q}_{T_0+1} - \epsilon_{T_0+1})} + \hat{Q}_{T_0+1}.
\end{aligned}$$

Hence,

$$\text{Regret}_T(I^{(2)}) \geq \frac{\sum_{s=1}^{T_0} q_s \epsilon_{T_0+1}}{2(\hat{Q}_{T_0+1} - \epsilon_{T_0+1})} \geq \frac{\sum_{s=1}^{T_0} q_s \epsilon_{T_0+1}}{2(\hat{Q}_{T_0+1} + \epsilon_{T_0+1})} = \frac{\sum_{s=1}^{T_0} q_s \epsilon_{T_0+1}}{2Q^{(2)}}.$$

Altogether, the Theorem is proved.

B.3. Consistency Between Regret Upper and Lower Bounds

Recall that in the proof of Theorem 3.2, we constructed two instances $I^{(1)}, I^{(2)}$ such that (see (21):

$$\text{Regret}_T(I^{(i)}) \geq \frac{\sum_{t=1}^{T_0} q_t \epsilon_{T_0+1}}{6Q^{(i)}} \quad \text{for some } i \in \{1,2\}, \tag{24}$$

where $\text{Regret}_T(I^{(i)})$ is the regret of an arbitrary but fixed online algorithm on $I^{(i)}$, with its prediction oracle satisfying that

$$|Q^{(i)} - \hat{Q}_t| \leq \epsilon_{T_0+1} \qquad \text{for each } i \in \{1,2\}. \tag{25}$$

In the lower bound analysis on $I^{(1)}, I^{(2)}$, we establish the regret lower bound (24) solely hinging on the model uncertainty on $Q^{(1)}, Q^{(2)}$, and the bound (24) still holds when the DM knows $\{P_a\}_{a\in\mathcal{A}}$.

In particular, we can set the online algorihtm to be OA-UCB, with an oracle that satisfies the property (25) above. Now, also recall in our construction that $q_t^{(1)} = q_t^{(2)} = q_t$ for all $t \in [T_0]$, thus the predictions $\hat{Q}_t$ for $t \in [T_0]$ are the same in the two instances, whereas $Q^{(1)} = \hat{Q}_{T_0+1} - \epsilon_{T_0+1}$ but $Q^{(2)} = \hat{Q}_{T_0+1} + \epsilon_{T_0+1}$, while we still have $Q^{(2)} \leq 3Q^{(1)}$, so that $Q^{(1)} = \Theta(Q^{(2)})$. Therefore, (24) is equivalent to

$$\max_{i\in\{1,2\}}\{\text{Regret}_T(I^{(i)})\} \geq \Omega\left(\frac{\sum_{t=1}^{T_0} q_t \epsilon_{T_0+1}}{Q^{(1)}}\right). \tag{26}$$

To demonstrate the consistency, it suffices to show

$$\max_{i\in\{1,2\}}\left\{\frac{1}{Q^{(1)}}\sum_{t=1}^{\tau-1} q_t \epsilon_t^{(i)}\right\} = \Omega\left(\frac{\sum_{t=1}^{T_0} q_t \epsilon_{T_0+1}}{Q^{(1)}}\right). \tag{27}$$

where $\epsilon_t^{(i)} = |\hat{Q}_t - Q^{(i)}|$ is the prediction error on instance $I^{(i)}$ at time t. Indeed, to be consistent, we should have Theorem

4.3 holds true for both instances, while (26) still holds true. We establish (27) as follows:

$$\begin{aligned}\max_{i\in\{1,2\}}\left\{\sum_{t=1}^{\tau-1} q_t\epsilon_t^{(i)}\right\} &\geq \sum_{t=1}^{\tau-1} q_t\frac{\epsilon_t^{(1)}+\epsilon_t^{(2)}}{2}\\ &= \sum_{t=1}^{\tau-1} q_t\frac{|\hat{Q}_t-\hat{Q}_{T_0+1}+\epsilon_{T_0+1}|+|\hat{Q}_t-\hat{Q}_{T_0+1}-\epsilon_{T_0+1}|}{2}\\ &\geq \sum_{t=1}^{\tau-1} q_t\frac{2\epsilon_{T_0+1}}{2} \qquad (28)\\ &\geq \sum_{t=1}^{T_0} q_t\epsilon_{T_0+1}. \qquad (29)\end{aligned}$$

Step (28) is by the triangle inequality, and step (29) is by the fact that for any algorithm that fully exhausts the resource, its stopping time $\tau > T_0$ (In the case when OA-UCB does not fully consume all the resource at the end of time T, by definition we have $\tau = T+1 > T_0$). By construction, the common budget B in both instances is strictly larger than $\sum_{t=1}^{T_0} q_t$, thus the resource is always not exhausted at T_0, since at time $t\in[T_0]$ the DM consumes at most q_t units of resource. Altogether, (27) is shown and consistency is verified.

C. Proof of Theorem 4.3

Before we embark on the proof, we first state a well known result on online gradient descent:

Lemma C.1 (Theorem 3.1 in (Hazan et al., 2016))**.** *Suppose $\{f_t\}$ are convex functions, then Online Gradient Descent presented in Algorithm 3 applied on $\{f_t\}$ with step sizes $\{\eta_t = \frac{D}{G\sqrt{t}}\}$ guarantees the following for all $T\geq 1$:*

$$\sum_{t=1}^{T} f_t(\boldsymbol{x}_t) - \min_{\boldsymbol{x}^*\in S}\sum_{t=1}^{T} f_t(\boldsymbol{x}^*) \leq \frac{3}{2}GD\sqrt{T},$$

where $D = diam(S)$ and $G = \max_t \|\nabla f_t\|$.

Algorithm 3 Online Gradient Descent

1: Initialize convex set S, $\boldsymbol{x}_1\in\mathcal{K}$, step sizes $\{\eta_t\}_{t=1}^T$.
2: **for** $t = 1,2,...,T$ **do**
3: Play $\boldsymbol{x}_t$ and observe cost $f_t(\boldsymbol{x}_t)$.
4: Update

$$\boldsymbol{x}_{t+1} = \Pi_S\left(\boldsymbol{x}_t - \eta_t\nabla f_t(\boldsymbol{x}_t)\right).$$

5: **end for**

Now we begin the proof of Theorem 4.3. Denote $\mathbf{UCB}_{r,t} = (\mathrm{UCB}_{r,t}(a))_{a\in\mathcal{A}}$, $\mathbf{LCB}_{c,t} = (\mathbf{LCB}_{c,t}(a,i))_{a\in\mathcal{A},i\in[d]}$. We first claim that, at a time step $t\leq\tau$,

$$\boldsymbol{e}_{A_t} \in \arg\max_{\boldsymbol{u}\in\Delta_{|\mathcal{A}|}} \mathbf{UCB}_{c,t}^\top\boldsymbol{u} - \frac{\hat{Q}_t}{B}\cdot\boldsymbol{\mu}_t^\top\mathbf{LCB}_{c,t}^\top\boldsymbol{u}. \qquad (30)$$

In fact, the following linear optimization problem

$$\begin{aligned}\max \quad & \mathbf{UCB}_{r,t}^\top\boldsymbol{u} - \frac{\hat{Q}_t}{B}\cdot\boldsymbol{\mu}_t^\top\mathbf{LCB}_{c,t}^\top\boldsymbol{u}\\ \text{s.t.} \quad & \boldsymbol{u}\in\Delta_{|\mathcal{A}|}\end{aligned}$$

has an extreme point solution such that $\boldsymbol{u}^* = \boldsymbol{e}_a$ for some $a\in\mathcal{A}$. According to the definition of A_t, we know that $\boldsymbol{u}^* = \boldsymbol{e}_{A_t}$. Then the claim holds. Suppose $\boldsymbol{u}^*$ is an optimal solution of (2), then we have $\mathrm{OPT}_{\mathrm{LP}} = Q\boldsymbol{r}^\top\boldsymbol{u}^*$, $Q\boldsymbol{c}^\top\boldsymbol{u}^*\leq B\mathbf{1}$ and $\boldsymbol{u}^*\in\Delta_{|\mathcal{A}|}$. By the optimality of (30), we have

$$\mathrm{UCB}_{r,t}(A_t) - \frac{\hat{Q}_t}{B}\cdot\boldsymbol{\mu}_t^\top\mathbf{LCB}_{c,t}(A_t) = \mathbf{UCB}_{r,t}^\top\boldsymbol{e}_{A_t} - \frac{\hat{Q}_t}{B}\cdot\boldsymbol{\mu}_t^\top\mathbf{LCB}_{c,t}^\top\boldsymbol{e}_{A_t} \geq \mathbf{UCB}_{r,t}^\top\boldsymbol{u}^* - \frac{\hat{Q}_t}{B}\cdot\boldsymbol{\mu}_t^\top\mathbf{LCB}_{c,t}\boldsymbol{u}^*,$$

which is equivalent to

$$\mathbf{UCB}_{r,t}^\top \boldsymbol{u}^* - \text{UCB}_{r,t}(A_t) + \frac{\hat{Q}_t}{B} \cdot \boldsymbol{\mu}_t^\top \left(\frac{B}{\hat{Q}_t} \mathbf{1}_d - \mathbf{LCB}_{c,t}^\top \boldsymbol{u}^* \right) \leq \frac{\hat{Q}_t}{B} \cdot \boldsymbol{\mu}_t^\top \left(\frac{B}{\hat{Q}_t} \mathbf{1}_d - \mathbf{LCB}_{c,t}(A_t) \right).$$

Times q_t on both side and sum all t from 1 to $\tau - 1$ and apply Lemma C.1 with $f_t(\boldsymbol{x}) = \frac{q_t \hat{Q}_t}{B} \left(\frac{B}{\hat{Q}_t} \mathbf{1}_d - \mathbf{LCB}_{c,t}(A_t) \right)^\top \boldsymbol{x}$, $S = \{\boldsymbol{\mu} : \|\boldsymbol{\mu}\|_1 \leq 1, \boldsymbol{\mu} \geq \mathbf{0}_d\}$, $D = \text{diam}(S) = \sqrt{2}$, $G = \max_t \|\nabla f_t\| = M$, then we obtain

$$\begin{aligned}
&\sum_{t=1}^{\tau-1} q_t \mathbf{UCB}_{r,t}^\top \boldsymbol{u}^* - \sum_{t=1}^{\tau-1} q_t \text{UCB}_{r,t}(A_t) + \sum_{t=1}^{\tau-1} q_t \frac{\hat{Q}_t}{B} \cdot \boldsymbol{\mu}_t^\top \left(\frac{B}{\hat{Q}_t} \mathbf{1}_d - \mathbf{LCB}_{c,t}^\top \boldsymbol{u}^* \right) \\
\leq &\sum_{t=1}^{\tau-1} q_t \frac{\hat{Q}_t}{B} \cdot \boldsymbol{\mu}_t^\top \left(\frac{B}{\hat{Q}_t} \mathbf{1}_d - \mathbf{LCB}_{c,t}(A_t) \right) \\
\leq &\min_{\boldsymbol{\mu}^* \in S} \sum_{t=1}^{\tau-1} q_t \frac{\hat{Q}_t}{B} \cdot \left(\frac{B}{\hat{Q}_t} \mathbf{1}_d - \mathbf{LCB}_{c,t}(A_t) \right)^\top \boldsymbol{\mu}^* + O\left(M\sqrt{T}\right) \\
\leq &\sum_{t=1}^{\tau-1} q_t \frac{\hat{Q}_t}{B} \cdot \left(\frac{B}{\hat{Q}_t} \mathbf{1}_d - \mathbf{LCB}_{c,t}(A_t) \right)^\top \boldsymbol{\mu} + O\left(M\sqrt{T}\right), \quad \forall \boldsymbol{\mu} \in S.
\end{aligned} \tag{31}$$

Recap by lemma 4.2 that with probability $\geq 1 - 3KTd\delta$, we have

$$\mathbf{LCB}_{c,t} \leq \boldsymbol{c}.$$

Hence, with probability $\geq 1 - 3KTd\delta$,

$$\sum_{t=1}^{\tau-1} q_t \frac{\hat{Q}_t}{B} \cdot \boldsymbol{\mu}_t^\top \left(\frac{B}{\hat{Q}_t} \mathbf{1}_d - \mathbf{LCB}_{c,t}^\top \boldsymbol{u}^* \right) \geq \sum_{t=1}^{\tau-1} q_t \frac{\hat{Q}_t}{B} \cdot \boldsymbol{\mu}_t^\top \left(\frac{B}{\hat{Q}_t} \mathbf{1}_d - \boldsymbol{c}^\top \boldsymbol{u}^* \right) \tag{32a}$$

$$= \sum_{t=1}^{\tau-1} q_t \frac{\hat{Q}_t}{B} \cdot \boldsymbol{\mu}_t^\top \left(\frac{B}{\hat{Q}_t} \mathbf{1}_d - \frac{B}{Q} \mathbf{1}_d \right) + \sum_{t=1}^{\tau-1} q_t \frac{\hat{Q}_t}{B} \cdot \boldsymbol{\mu}_t^\top \left(\frac{B}{Q} \mathbf{1}_d - \boldsymbol{c}^\top \boldsymbol{u}^* \right) \tag{32b}$$

$$\geq \sum_{t=1}^{\tau-1} q_t \frac{\hat{Q}_t}{B} \cdot \boldsymbol{\mu}_t^\top \left(\frac{B}{\hat{Q}_t} \mathbf{1}_d - \frac{B}{Q} \mathbf{1}_d \right) \tag{32c}$$

$$= \sum_{t=1}^{\tau-1} q_t \frac{\hat{Q}_t}{B} \left(\frac{B}{\hat{Q}_t} - \frac{B}{Q} \right) \|\boldsymbol{\mu}_t\|_1, \tag{32d}$$

where (32a) comes from Lemma 4.2, (32b) comes from rearranging the sum, and (32c) comes from the fact the definition of $\boldsymbol{u}^*$. We first consider the case $\tau \leq T$, which implies that there exists $j_0 \in [d]$ such that

$$\sum_{t=1}^{\tau} q_t C_{t,j_0} > B \quad \Rightarrow \quad \sum_{t=1}^{\tau-1} q_t C_{t,j_0} > B - \overline{q}. \tag{33}$$

Take $\boldsymbol{\mu} = \lambda \boldsymbol{e}_{j_0}$, where $\lambda \in [0, 1]$ is a constant that we tune later. In this case, with probability $\geq 1 - 3KT\delta$,

$$\sum_{t=1}^{\tau-1} q_t \mathbf{UCB}_{r,t}^\top \boldsymbol{u}^* \geq \sum_{t=1}^{\tau-1} q_t \boldsymbol{r}_t^\top \boldsymbol{u}^* = \text{OPT}_{\text{LP}} \frac{Q_{\tau-1}}{Q}, \tag{34}$$

and

$$\lambda \sum_{t=1}^{\tau-1} q_t \frac{\hat{Q}_t}{B} \left(\frac{B}{\hat{Q}_t} \mathbf{1}_d - \mathbf{LCB}_{c,t}(A_t) \right)^\top \boldsymbol{e}_{j_0} = \lambda \sum_{t=1}^{\tau-1} q_t \frac{\hat{Q}_t}{B} \left(\frac{B}{\hat{Q}_t} - \mathrm{LCB}_{c,t}(A_t, j_0) \right)$$
$$= \lambda \sum_{t=1}^{\tau-1} q_t \frac{\hat{Q}_t}{B} \left(\frac{B}{\hat{Q}_t} - \frac{B}{Q} \right) + \lambda \sum_{t=1}^{\tau-1} q_t \frac{\hat{Q}_t}{B} \left(\frac{B}{Q} - C_{t,j_0} \right) \tag{35}$$
$$+ \lambda \sum_{t=1}^{\tau-1} q_t \frac{\hat{Q}_t}{B} \left(C_{t,j_0} - \mathrm{LCB}_{c,t}(A_t, j_0) \right).$$

Then we deal with each term respectively:

$$\sum_{t=1}^{\tau-1} q_t \frac{\hat{Q}_t}{B} \left(\frac{B}{Q} - C_{t,j_0} \right) = \sum_{t=1}^{\tau-1} q_t \frac{Q}{B} \left(\frac{B}{Q} - C_{t,j_0} \right) + \sum_{t=1}^{\tau-1} q_t \frac{\hat{Q}_t - Q}{B} \left(\frac{B}{Q} - C_{t,j_0} \right) \tag{36a}$$
$$\leq Q_{\tau-1} - \frac{Q}{B} \sum_{t=1}^{\tau-1} q_t C_{t,j_0} + \frac{1}{B} \sum_{t=1}^{\tau-1} q_t \epsilon_t \left| \frac{B}{Q} - C_{t,j_0} \right| \tag{36b}$$
$$< Q_{\tau-1} - Q + \frac{Q}{B} \overline{q} + \frac{1}{B} \sum_{t=1}^{\tau-1} q_t \epsilon_t \frac{B}{Q} + \frac{1}{B} \sum_{t=1}^{\tau-1} q_t \epsilon_t C_{t,j_0} \tag{36c}$$
$$\leq Q_{\tau-1} - Q + \frac{Q}{B} \overline{q} + \left(\frac{1}{Q} + \frac{1}{B} \right) \sum_{t=1}^{\tau-1} q_t \epsilon_t, \tag{36d}$$

where (36a) comes from rearranging the sum, (36c) comes from the (33), and (36d) comes from the assumption that C_{t,j_0} is supported in $[0, 1]$. Similarly,

$$\begin{aligned} \sum_{t=1}^{\tau-1} q_t \frac{\hat{Q}_t}{B} \left(C_{t,j_0} - \mathrm{LCB}_{c,t}(A_t, j_0) \right) &= \sum_{t=1}^{\tau-1} q_t \frac{Q}{B} \left(C_{t,j_0} - \mathrm{LCB}_{c,t}(A_t, j_0) \right) + \sum_{t=1}^{\tau-1} q_t \frac{\hat{Q}_t - Q}{B} \left(C_{t,j_0} - \mathrm{LCB}_{c,t}(A_t, j_0) \right) \\ &\leq \frac{Q}{B} \left| \sum_{t=1}^{\tau-1} q_t \left(C_{t,j_0} - \mathrm{LCB}_{c,t}(A_t, j_0) \right) \right| + \frac{1}{B} \sum_{t=1}^{\tau-1} q_t \epsilon_t, \end{aligned} \tag{37}$$

where the equality comes from rearranging the sum, and the inequality comes from the assuption that $|\hat{Q}_t - Q| \leq \epsilon_t$, $0 \leq \mathrm{LCB}_{c,t}(A_t, j_0), C_{t,j_0} \leq 1$. Combine (35), (36) and (37), we obtain

$$\begin{aligned} \lambda \sum_{t=1}^{\tau-1} q_t \frac{\hat{Q}_t}{B} \left(\frac{B}{\hat{Q}_t} \mathbf{1}_d - \mathbf{LCB}_{c,t}(A_t) \right)^\top \boldsymbol{e}_{j_0} &\leq \lambda \sum_{t=1}^{\tau-1} q_t \frac{\hat{Q}_t}{B} \left(\frac{B}{\hat{Q}_t} - \frac{B}{Q} \right) + \lambda \left(Q_{\tau-1} - Q + \frac{Q}{B} \overline{q} + \left(\frac{1}{Q} + \frac{1}{B} \right) \sum_{t=1}^{\tau-1} q_t \epsilon_t \right) \\ &+ \lambda \left(\frac{Q}{B} \left| \sum_{t=1}^{\tau-1} q_t \left(C_{t,j_0} - \mathrm{LCB}_{c,t}(A_t, j_0) \right) \right| + \frac{1}{B} \sum_{t=1}^{\tau-1} q_t \epsilon_t \right) \\ &\leq \lambda \left(Q_{\tau-1} - Q + \frac{Q}{B} \overline{q} + \frac{Q}{B} \left| \sum_{t=1}^{\tau-1} q_t \left(C_{t,j_0} - \mathrm{LCB}_{c,t}(A_t, j_0) \right) \right| \right) \\ &+ \sum_{t=1}^{\tau-1} q_t \frac{\hat{Q}_t}{B} \left(\frac{B}{\hat{Q}_t} - \frac{B}{Q} \right) + O \left(\left(\frac{1}{Q} + \frac{1}{B} \right) \sum_{t=1}^{\tau-1} q_t \epsilon_t \right), \end{aligned} \tag{38}$$

where the second inequality comes from the assumption that $\lambda \in [0, 1]$. Finally, combine (31), (32), (34), (38), we obtain

$$\mathrm{OPT}_{\mathrm{LP}} \frac{Q_{\tau-1}}{Q} - \sum_{t=1}^{\tau-1} q_t \mathrm{UCB}_{r,t}(A_t) + \sum_{t=1}^{\tau-1} q_t \frac{\hat{Q}_t}{B} \left(\frac{B}{\hat{Q}_t} - \frac{B}{Q} \right) \|\boldsymbol{\mu}_t\|_1$$
$$\leq \lambda \left(Q_{\tau-1} - Q + \frac{Q}{B} \overline{q} + \frac{Q}{B} \left| \sum_{t=1}^{\tau-1} q_t \left(C_{t,j_0} - \mathrm{LCB}_{c,t}(A_t, j_0) \right) \right| \right) + \sum_{t=1}^{\tau-1} q_t \frac{\hat{Q}_t}{B} \left(\frac{B}{\hat{Q}_t} - \frac{B}{Q} \right) + O \left(\left(\frac{1}{Q} + \frac{1}{B} \right) \sum_{t=1}^{\tau-1} q_t \epsilon_t \right),$$

which is equivalent to

$$\mathrm{OPT}_{\mathrm{LP}} - \sum_{t=1}^{\tau-1} q_t \mathrm{UCB}_{r,t}(A_t) \leq \mathrm{OPT}_{\mathrm{LP}} \left(1 - \frac{Q_{\tau-1}}{Q}\right) + \lambda \left(Q_{\tau-1} - Q + \frac{Q}{B}\bar{q} + \frac{Q}{B} \left| \sum_{t=1}^{\tau-1} q_t \left(C_{t,j_0} - \mathrm{LCB}_{c,t}(A_t, j_0)\right) \right| \right)$$
$$+ \sum_{t=1}^{\tau-1} q_t \frac{\hat{Q}_t}{B} \left(\frac{B}{\hat{Q}_t} - \frac{B}{Q} \right) (1 - \|\boldsymbol{\mu}_t\|_1) + O\left(\left(\frac{1}{Q} + \frac{1}{B} \right) \sum_{t=1}^{\tau-1} q_t \epsilon_t \right) + O\left(M\sqrt{T} \right).$$

Let $\lambda = \frac{\mathrm{OPT}_{\mathrm{LP}}}{Q} \leq 1$ (This is because $\mathrm{OPT}_{\mathrm{LP}} = Q\boldsymbol{r}^\top \boldsymbol{u}^* \leq Q$), then we can further derive with probability $\geq 1 - 3KTd\delta$,

$$\begin{aligned}
\mathrm{OPT}_{\mathrm{LP}} - \sum_{t=1}^{\tau-1} q_t \mathrm{UCB}_{r,t}(A_t) &\leq \mathrm{OPT}_{\mathrm{LP}} \left(1 - \frac{Q_{\tau-1}}{Q}\right) + \frac{\mathrm{OPT}_{\mathrm{LP}}}{Q} \left(Q_{\tau-1} - Q + \frac{Q}{B}\bar{q} + \frac{Q}{B} \left| \sum_{t=1}^{\tau-1} q_t \left(C_{t,j_0} - \mathrm{LCB}_{c,t}(A_t, j_0)\right) \right| \right) \\
&+ \sum_{t=1}^{\tau-1} q_t \frac{\hat{Q}_t}{B} \left(\frac{B}{\hat{Q}_t} - \frac{B}{Q} \right) (1 - \|\boldsymbol{\mu}_t\|_1) + O\left(\left(\frac{1}{Q} + \frac{1}{B} \right) \sum_{t=1}^{\tau-1} q_t \epsilon_t \right) + O\left(M\sqrt{T} \right) \\
&= \frac{\mathrm{OPT}_{\mathrm{LP}}}{B}\bar{q} + \frac{\mathrm{OPT}_{\mathrm{LP}}}{B} \left| \sum_{t=1}^{\tau-1} q_t \left(C_{t,j_0} - \mathrm{LCB}_{c,t}(A_t, j_0)\right) \right| + \sum_{t=1}^{\tau-1} q_t \frac{\hat{Q}_t}{B} \left(\frac{B}{\hat{Q}_t} - \frac{B}{Q} \right) (1 - \|\boldsymbol{\mu}_t\|_1) \\
&+ O\left(\left(\frac{1}{Q} + \frac{1}{B} \right) \sum_{t=1}^{\tau-1} q_t \epsilon_t \right) + O\left(M\sqrt{T} \right) \\
&\leq O\left(\log\left(\frac{1}{\delta}\right) \mathrm{OPT}_{\mathrm{LP}} \left(\sqrt{\frac{\bar{q}K}{B}} + \frac{\bar{q}K}{B} \log\left(\frac{T}{K}\right) \right) + \left(\frac{1}{Q} + \frac{1}{B} \right) \sum_{t=1}^{\tau-1} q_t \epsilon_t + M\sqrt{T} \right) \\
&= O\left(\log\left(\frac{1}{\delta}\right) \mathrm{OPT}_{\mathrm{LP}} \sqrt{\frac{\bar{q}K}{B}} + \left(\frac{1}{Q} + \frac{1}{B} \right) \sum_{t=1}^{\tau-1} q_t \epsilon_t + M\sqrt{T} \right),
\end{aligned} \tag{39}$$

where the second inequality comes from Lemma D.10 and the following

$$\sum_{t=1}^{\tau-1} q_t \frac{\hat{Q}_t}{B} \left(\frac{B}{\hat{Q}_t} - \frac{B}{Q} \right) (1 - \|\boldsymbol{\mu}_t\|_1) \leq \sum_{t=1}^{\tau} q_t \frac{\hat{Q}_t}{B} \left| \frac{B}{Q} - \frac{B}{\hat{Q}_t} \right| = \frac{1}{Q} \sum_{t=1}^{\tau} q_t \left| \hat{Q}_t - Q \right| \leq \frac{1}{Q} \sum_{t=1}^{\tau} q_t \epsilon_t.$$

The above concludes our arguments for the case $\tau \leq T$. In complement, we then consider the case $\tau > T$, which means that $\tau = T + 1$, and no resource is fully exhausted during the horizon. With probability $\geq 1 - 3KT\delta$, we have

$$\sum_{t=1}^{T} q_t \mathbf{UCB}_{r,t}^\top \boldsymbol{u}^* \geq \sum_{t=1}^{T} q_t \boldsymbol{r}_t^\top \boldsymbol{u}^* = \mathrm{OPT}_{\mathrm{LP}}. \tag{40}$$

Take $\boldsymbol{\mu} = 0$ and combine (31), (32), (40), with probablity $\geq 1 - 3KT\delta$, we have

$$\begin{aligned}
\mathrm{OPT}_{\mathrm{LP}} - \sum_{t=1}^{\tau-1} q_t \mathrm{UCB}_{r,t}(A_t) &\leq - \sum_{t=1}^{\tau-1} q_t \frac{\hat{Q}_t}{B} \left(\frac{B}{\hat{Q}_t} - \frac{B}{Q} \right) \|\boldsymbol{\mu}_t\|_1 + O\left(M\sqrt{T} \right) \\
&\leq O\left(\frac{1}{Q} \sum_{t=1}^{\tau-1} q_t \epsilon_t + M\sqrt{T} \right).
\end{aligned} \tag{41}$$

Combine (41) and (39), for any stopping time τ, with probability $\geq 1 - 3KTd\delta$, we have

$$\mathrm{OPT}_{\mathrm{LP}} - \sum_{t=1}^{\tau-1} q_t \mathrm{UCB}_{r,t}(A_t) \leq O\left(\log\left(\frac{1}{\delta}\right) \mathrm{OPT}_{\mathrm{LP}} \sqrt{\frac{\bar{q}K}{B}} + \left(\frac{1}{Q} + \frac{1}{B} \right) \sum_{t=1}^{\tau-1} q_t \epsilon_t + M\sqrt{T} \right).$$

By Lemma D.9, we can further derive it to the high probability bound, that with probability $\geq 1 - 3KTd\delta$,

$$\begin{aligned}
\mathrm{OPT}_{\mathrm{LP}} - \sum_{t=1}^{\tau-1} q_t R_t &\leq O\left(\log\left(\frac{1}{\delta}\right) \left(\mathrm{OPT}_{\mathrm{LP}} \sqrt{\frac{\bar{q}K}{B}} + \sqrt{\bar{q}K \sum_{t=1}^{\tau-1} q_t R_t} + \bar{q}K \log\left(\frac{T}{K}\right) \right) + \left(\frac{1}{Q} + \frac{1}{B} \right) \sum_{t=1}^{\tau-1} q_t \epsilon_t + M\sqrt{T} \right) \\
&\leq O\left(\log\left(\frac{1}{\delta}\right) \left(\mathrm{OPT}_{\mathrm{LP}} \sqrt{\frac{\bar{q}K}{B}} + \sqrt{\bar{q}K \mathrm{OPT}_{\mathrm{LP}}} \right) + \left(\frac{1}{Q} + \frac{1}{B} \right) \sum_{t=1}^{\tau-1} q_t \epsilon_t + M\sqrt{T} \right),
\end{aligned}$$

where the second inequality comes from the fact that $\mathrm{OPT}_{\mathrm{LP}} \geq \sum_{t=1}^{\tau-1} q_t R_t$. Now we finish the proof of Theorem 4.3.

D. Proofs for Confidence Radii

This section contains proofs for the confidence radius results, which largely follow the literature, but we provide complete proofs since we are in a non-stationary setting. Section D.1 provides the proof for Lemma 4.1, which allows us to extract the implicit constants in existing proofs in (Babaioff et al., 2015; Agrawal & Devanur, 2014). Section D.2 provides the proof for Lemma 4.2. Finally, section D.3, we prove inequalities (14, 15).

D.1. Proof for Lemma 4.1, due to (Babaioff et al., 2015; Agrawal & Devanur, 2014)

In this subsection, we prove Lemma 4.1 by following the line of arguments in (Babaioff et al., 2015). We emphasize that a version of the Lemma has been proved in (Babaioff et al., 2015). We dervie the Lemma for the purpose of extracting the values of the constant coefficients. We first extract some relevant concentration inequalities in the following two Lemmas.

Lemma D.1 (Theorem 8 in (Chung & Lu, 2006))**.** *Suppose $\{U_i\}_{i=1}^n$ are independent random variables satisfying $U_i \leq M$, for $1 \leq i \leq n$ almost surely. Let $U = \sum_{i=1}^n U_i$, $\|U\|^2 = \sum_{i=1}^n \mathbb{E}[U_i^2]$. With probability $\geq 1 - e^{-x}$, we have*

$$U - \mathbb{E}[U] \leq \sqrt{2\|U\|^2 x} + \frac{2x}{3}\max\{M, 0\}.$$

Lemma D.2 (Theorem 6 in (Chung & Lu, 2006))**.** *Suppose U_i are independent random variables satisfying $U_i - \mathbb{E}[U_i] \leq M$, $M > 0$, for $1 \leq i \leq n$. Let $U = \sum_{i=1}^n U_i$, Var$(U) = \sum_{i=1}^n$ Var(U_i), then with probability $\geq 1 - e^{-x}$, we have*

$$U - \mathbb{E}[U] \leq \sqrt{2\textit{Var}(U)x} + \frac{2Mx}{3}.$$

Using Lemma D.2, we first derive the following Lemma that bounds the empirical mean:

Lemma D.3. *Let $\{X_i\}_{i=1}^n$ be independent random variables supported in $[0, 1]$. Let $X = \sum_{i=1}^n X_i$ and Var$(X) = \sum_{i=1}^n$ Var(X_i). For any fixed $x > 0$, With probability $\geq 1 - 2e^{-x}$, we have*

$$|X - \mathbb{E}[X]| \leq \sqrt{2\textit{Var}(X)x} + \frac{2x}{3}.$$

Proof of Lemma D.3. Apply Lemma D.2 with $U_i = X_i$, $U_i = -X_i$, respectively, and $M = 1$, then with probability $\geq 1 - 2e^{-x}$, we have

$$|X - \mathbb{E}[X]| \leq \sqrt{2\text{Var}(X)x} + \frac{2x}{3}.$$

□

Next, we bound the difference between the ground truth variance and its empirical counterpart using Lemma D.1:

Lemma D.4. *Suppose X_i are independent random variables supported in $[0, 1]$. Let $X = \sum_{i=1}^n X_i$, Var$(X) = \sum_{i=1}^n$ Var(X_i), $V_n = \sum_{i=1}^n (X_i - \mathbb{E}[X_i])^2$then with probability $\geq 1 - 3e^{-x}$, we have*

$$\sqrt{\textit{Var}(X)} \leq \sqrt{V_n} + 2\sqrt{x}.$$

Proof of Lemma D.4. The proof follows the line of argument in (Audibert et al., 2009). First, we apply Lemma D.1 with $U_i = -(X_i - \mathbb{E}[X_i])^2$ and $M = 0$. With probability $\geq 1 - e^{-x}$, we have

$$\text{Var}(X) \leq V_n + \sqrt{2\left(\sum_{i=1}^n \mathbb{E}\left[(X_i - \mathbb{E}[X_i])^4\right]\right)x} \leq V_n + \sqrt{2\left(\sum_{i=1}^n \mathbb{E}\left[(X_i - \mathbb{E}[X_i])^2\right]\right)x} = V_n + \sqrt{2\text{Var}(X)x}. \quad (42)$$

Since $X_i \in [0,1]$ almost surely for all $i \in [n]$, we have

$$\mathrm{Var}(X_i) = \mathbb{E}[X_i^2] - \mathbb{E}[X_i]^2 \leq \mathbb{E}[X_i] - \mathbb{E}[X_i]^2 \leq \frac{1}{4}.$$

Now, observe that

$$\mathrm{Var}(X) = \sum_{i=1} \mathrm{Var}(X_i) \leq \sum_{i=1}^{n} \frac{1}{4} = \frac{n}{4} \quad \Rightarrow \sqrt{\mathrm{Var}(X)} \leq \frac{\sqrt{n}}{2}.$$

If $2\sqrt{x} \geq \frac{\sqrt{n}}{2}$, then the Lemma evidently holds. Otherwise, we assume $2\sqrt{x} \leq \frac{\sqrt{n}}{2}$, which is equivalent to $x \leq \frac{n}{16}$. Combining Lemma D.3 and (42), with probability $\geq 1 - 3e^{-x}$, we have,

$$\begin{aligned}
\mathrm{Var}(X) &\leq V_n + \sqrt{2\mathrm{Var}(X)x} + \frac{(X - \mathbb{E}[X])^2}{n} \\
&\leq V_n + \sqrt{2\mathrm{Var}(X)x} + \frac{1}{n}\left(2\mathrm{Var}(X)x + \frac{4}{3}x\sqrt{2\mathrm{Var}(X)x} + \frac{4x^2}{9}\right) \\
&\leq V_n + \sqrt{2\mathrm{Var}(X)x} + \frac{1}{n}\left(2\sqrt{\mathrm{Var}(X)x} \cdot \frac{\sqrt{n}}{2}\frac{\sqrt{n}}{4} + \frac{4}{3}\sqrt{2\mathrm{Var}(X)x} \cdot \frac{n}{16} + \frac{4x}{9} \cdot \frac{n}{16}\right) \\
&= \sqrt{\mathrm{Var}(X)x}\left(\frac{13}{12}\sqrt{2} + \frac{1}{4}\right) + \left(V_n + \frac{x}{36}\right).
\end{aligned}$$

Consequently, we can derive an upper bound for $\sqrt{\mathrm{Var}(X)}$:

$$\sqrt{\mathrm{Var}(X)} \leq \frac{\sqrt{x}}{2}\left(\frac{13}{12}\sqrt{2} + \frac{1}{4}\right) + \frac{1}{2}\sqrt{x\left(\frac{13}{12}\sqrt{2} + \frac{1}{4}\right)^2 + 4\left(V_n + \frac{x}{36}\right)} \leq \sqrt{V_n} + 2\sqrt{x},$$

which proves the Lemma. □

Lemma D.5. *Suppose X_i are independent random variables supported in $[0,1]$. Let $X = \sum_{i=1}^n X_i$, then with probability $\geq 1 - 3e^{-x}$, we have*

$$|X - \mathbb{E}[X]| \leq \sqrt{2Xx} + 4x.$$

Proof of Lemma D.5. Apply Lemma D.3 and Lemma D.4, we directly derive that with probability $\geq 1 - 3e^{-x}$

$$|X - \mathbb{E}[X]| \leq \sqrt{2\mathrm{Var}(X)x} + \frac{2x}{3} \leq \sqrt{2V_n x} + \left(2\sqrt{2} + \frac{2}{3}\right)x < \sqrt{2V_n x} + 4x \leq \sqrt{2Xx} + 4x,$$

where the last inequality comes from the fact that for random variable whose support is $[0,1]$, then its variance is always smaller than its mean. □

Lemma D.6 (Theorem 1.1 in (Dubhashi & Panconesi, 2009))**.** *Suppose X_i are independent random variables supported in $[0,1]$. Let $X = \sum_{i=1}^n X_i$, then for any $R > 2e\mathbb{E}[X]$, we have*

$$\mathbb{P}(X > R) \leq 2^{-R}.$$

Now we turn back to the proof of Lemma 4.1. Denote $\delta = e^{-x}$. Apply Lemma D.5 then with probability $\geq 1 - 3\delta$, we have,

$$N\left|\hat{V} - \mathbb{E}\left[\hat{V}\right]\right| \leq \sqrt{2N\hat{V}\log\left(\frac{1}{\delta}\right)} + 4\log\left(\frac{1}{\delta}\right),$$

which is equivalent to ,

$$\left|\hat{V} - \mathbb{E}\left[\hat{V}\right]\right| \leq \mathrm{rad}\left(\hat{V}, N, \delta\right). \tag{43}$$

Besides,

$$\begin{aligned}\mathbb{P}\left(\text{rad}\left(\hat{V},N,\delta\right)>3\text{rad}\left(\mathbb{E}\left[\hat{V}\right],N,\delta\right)\right)&\le\mathbb{P}\left(\hat{V}>9\mathbb{E}\left[\hat{V}\right]+32\log\left(\frac{1}{\delta}\right)\right)\\&\le 2^{-9\mathbb{E}\left[\hat{V}\right]-32\log\left(\frac{1}{\delta}\right)}\\&\le\delta.\end{aligned}\tag{44}$$

Therefore, combining (43) and (44), the lemma holds.

D.2. Proof for Lemma 4.2

By Lemma 4.1, with probability $\ge 1-3KT\delta$, we have

$$|r(a)-\hat{R}_t(a)|\le\text{rad}(\hat{R}_t(a),N^+_{t-1}(a),\delta).$$

Hence with probability $\ge 1-3KT\delta$,

$$\begin{cases}r(a)\le\hat{R}_t(a)+\text{rad}(\hat{R}_t(a),N^+_{t-1}(a),\delta)\\ r(a)\le 1\end{cases}\quad\Rightarrow\quad\begin{aligned}r(a)&\le\min\left\{\hat{R}_t(a)+\text{rad}(\hat{R}_t(a),N^+_{t-1}(a),\delta),1\right\}\\&=\text{UCB}_{r,t}(a).\end{aligned}$$

Similarly, with probability $\ge 1-3KTd\delta$,

$$\textbf{LCB}_{c,t}(a)\preceq\boldsymbol{c}(a).$$

D.3. Proof for Inequalities 14, 15

We first provide the two lemmas:

Lemma D.7 (Theorem 1.6 in (Freedman, 1975))**.** *Suppose $\{U_i\}_{i=1}^n$ is a martingale difference sequence supported in $[0,1]$ with respect to the filtration $\{\mathcal{F}_i\}_{i=1}^n$. Let $U=\sum_{i=1}^n U_i$, and $V=\sum_{i=1}^n \text{Var}(U_i|\mathcal{F}_{i-1})$. Then for any $a>0$, $b>0$, we have*

$$\mathbb{P}\left(|U|\ge a,V\le b\right)\le 2e^{-\frac{a^2}{2(a+b)}}.$$

Lemma D.8. *Suppose $\{X_i\}_{i=1}^n$ are random variables supported in $[0,1]$, where X_i is $\mathcal{F}_i$-measurable and $\{\mathcal{F}_i\}_{i=1}^n$ is a filtration. Let $M_i=\mathbb{E}[X_i|\mathcal{F}_{i-1}]$ for each $i\in\{1,\dots,n\}$, and $M=\sum_{i=1}^n M_i$. Then with probability $\ge 1-2n\delta$, we have*

$$\left|\sum_{i=1}^n(X_i-M_i)\right|\le O\left(\sqrt{M\log\left(\frac{1}{\delta}\right)}+\log\left(\frac{1}{\delta}\right)\right).$$

Proof of Lemma D.8. The proof follows the line of Theorem 4.10 in (Babaioff et al., 2015). Let $U_i=X_i-M_i$ for each $i\in\{1,\dots,n\}$. Clearly, $\{U_i\}_{i=1}^n$ is a martingale difference sequence with respect to the filtration $\{\mathcal{F}_i\}_{i=1}^n$. Since

$$\text{Var}(U_i|\mathcal{F}_{i-1})=\text{Var}(X_i|\mathcal{F}_{i-1})=\mathbb{E}[X_i^2|\mathcal{F}_{i-1}]-\mathbb{E}[X_i|\mathcal{F}_{i-1}]^2\le\mathbb{E}[X_i|\mathcal{F}_{i-1}]=M_i\text{ almost surely},$$

we have $V=\sum_{i=1}^n\text{Var}(U_i|\mathcal{F}_{i-1})\le\sum_{i=1}^n M_i=M$ almost surely. Apply Lemma D.7 with $a=\sqrt{2b\log\left(\frac{1}{\delta}\right)}+2\log\left(\frac{1}{\delta}\right)$ for any $b\ge 1$, it follows that with probability $\le 2\delta$,

$$|U|=\left|\sum_{i=1}^n U_i\right|\ge O\left(\sqrt{b\log\left(\frac{1}{\delta}\right)}+\log\left(\frac{1}{\delta}\right)\right)\quad\&\quad V\le b,$$

Take the union bound over all integer b from 1 to n, noting that $V\le M$ and $b-1\le M\le b$ for some $b\in\{1,\dots,n\}$ almost surely, with probability $\ge 1-2n\delta$ we have

$$\left|\sum_{i=1}^n(X_i-M_i)\right|\le O\left(\sqrt{M\log\left(\frac{1}{\delta}\right)}+\log\left(\frac{1}{\delta}\right)\right).$$

Altogether, the lemma holds.

□

Now, we paraphrase inequalities 14, 15 as Lemmas D.9, D.10, and provide their proofs.

Lemma D.9. *With probability $\geq 1 - 3KT\delta$, we have*

$$\left|\sum_{t=1}^{\tau-1} q_t UCB_{r,t}(A_t) - \sum_{t=1}^{\tau-1} q_t R_t\right| \leq O\left(\log\left(\frac{1}{\delta}\right)\left(\sqrt{\overline{q}K\sum_{t=1}^{\tau-1} q_t R_t} + \overline{q}K\log\left(\frac{T}{K}\right)\right)\right).$$

Lemma D.10. *With probability $\geq 1 - 3KTd\delta$, we have*

$$\left|\sum_{t=1}^{\tau-1} q_t LCB_{c,t}(A_t, i) - \sum_{t=1}^{\tau-1} q_t C_{t,i}\right| \leq O\left(\log\left(\frac{1}{\delta}\right)\left(\sqrt{\overline{q}KB} + \overline{q}K\log\left(\frac{T}{K}\right)\right)\right), \quad \forall i \in [d].$$

Proof of Lemma D.9. First with probability $\geq 1 - 2T\delta$, we have

$$\left|\sum_{t=1}^{\tau-1} q_t r(A_t) - \sum_{t=1}^{\tau-1} q_t R_t\right| = \overline{q}\left|\sum_{t=1}^{\tau-1} \frac{q_t}{\overline{q}}(r(A_t) - R_t)\right| \tag{45a}$$

$$\leq O\left(\sqrt{\overline{q}\log\left(\frac{1}{\delta}\right)\sum_{t=1}^{\tau-1} q_t r(A_t)} + \overline{q}\log\left(\frac{1}{\delta}\right)\right) \tag{45b}$$

$$\leq O\left(\sqrt{\overline{q}\log\left(\frac{1}{\delta}\right)\sum_{t=1}^{\tau-1} q_t \mathrm{UCB}_{r,t}(A_t)} + \overline{q}\log\left(\frac{1}{\delta}\right)\right), \tag{45c}$$

where (45c) comes from Lemma 4.2. Inequality (45b) comes from Lemma D.8, where we apply $X_t = \frac{q_t R_t}{\overline{q}}$ and $\mathcal{F}_{t-1} = \sigma(\{A_s, q_s, R_s, \{C_{s,i}\}_{i=1}^d, \hat{Q}_s\}_{s=1}^{t-1} \cup \{q_t\})$. Then with probability $\geq 1 - 3KT\delta$, we also have

$$\left|\sum_{t=1}^{\tau-1} q_t \mathrm{UCB}_{r,t}(A_t) - \sum_{t=1}^{\tau-1} q_t r(A_t)\right| \leq 6\sum_{t=1}^{\tau-1} q_t \mathrm{rad}(r(A_t), N_{t-1}^{+}(A_t), \delta) \tag{46a}$$

$$\leq 6\sum_{a\in\mathcal{A}:N_{\tau-1}(a)>0}\ \sum_{n=1}^{N_{\tau-1}(a)} q_n(a)\mathrm{rad}\left(r(a), n, \delta\right) \tag{46b}$$

$$= 6\overline{q}\sum_{a\in\mathcal{A}:N_{\tau-1}(a)>0}\ \sum_{n=1}^{N_{\tau-1}(a)} \frac{q_n(a)}{\overline{q}}\left(\sqrt{\frac{2r(a)\log\left(\frac{1}{\delta}\right)}{n}} + \frac{4}{n}\log\left(\frac{1}{\delta}\right)\right) \tag{46c}$$

$$\leq 6\overline{q}\sum_{a\in\mathcal{A}:N_{\tau-1}(a)>0}\left(2\sqrt{2r(a)\frac{Q_{\tau-1}(a)}{\overline{q}}\log\left(\frac{1}{\delta}\right)} + 4\left(1 + \log(N_{\tau-1}(a))\right)\log\left(\frac{1}{\delta}\right)\right) \tag{46d}$$

$$\leq 12\left(\sqrt{2\overline{q}K\log\left(\frac{1}{\delta}\right)\sum_{a\in\mathcal{A}} r(a)Q_{\tau-1}(a)} + 2\overline{q}K\log\left(\frac{T}{K}\right)\log\left(\frac{1}{\delta}\right) + 2\overline{q}K\log\left(\frac{1}{\delta}\right)\right) \tag{46e}$$

$$= 12\left(\sqrt{2\overline{q}K\log\left(\frac{1}{\delta}\right)\sum_{t=1}^{\tau-1} q_t r(A_t)} + 2\overline{q}K\log\left(\frac{T}{K}\right)\log\left(\frac{1}{\delta}\right) + 2\overline{q}K\log\left(\frac{1}{\delta}\right)\right) \tag{46f}$$

$$\leq 12\left(\sqrt{2\overline{q}K\log\left(\frac{1}{\delta}\right)\sum_{t=1}^{\tau-1} q_t \mathrm{UCB}_{r,t}(A_t)} + 2\overline{q}K\log\left(\frac{T}{K}\right)\log\left(\frac{1}{\delta}\right) + 2\overline{q}K\log\left(\frac{1}{\delta}\right)\right), \tag{46g}$$

where

- (46a) comes from the following, with probability $\geq 1 - 3KT\delta$,

$$\begin{aligned}|\text{UCB}_{r,t}(A_t) - r(A_t)| &\leq \left|\hat{R}_{t-1}(A_t) - r(A_t)\right| + \text{rad}(\hat{R}_{t-1}(A_t), N^+_{t-1}(A_t), \delta)\\ &\leq 2\text{rad}(\hat{R}_{t-1}(A_t), N_{t-1}(A_t), \delta)\\ &\leq 6\text{rad}(r(A_t), N_{t-1}(A_t), \delta).\end{aligned}$$

- (46b) comes from rearranging the sum. $q_n(a)$ means the n-th adversarial term that the algorithm selects a.

- (46c) comes from the definition of $\text{rad}(\cdot, \cdot, \cdot)$.

- (46d) comes from the following

$$\sum_{i=1}^{n} \frac{w_i}{\sqrt{i}} = \sum_{i=1}^{n} \frac{2w_i}{2\sqrt{i}} \leq \sum_{i=1}^{n} \frac{2w_i}{\sqrt{\sum_{j=1}^{i} w_j} + \sqrt{\sum_{j=1}^{i-1} w_j}} = \sum_{i=1}^{n} 2\left(\sqrt{\sum_{j=1}^{i} w_j} - \sqrt{\sum_{j=1}^{i-1} w_j}\right) = 2\sqrt{\sum_{i=1}^{n} w_i},$$

  and

$$\sum_{i=1}^{n} \frac{w_i}{i} \leq \sum_{i=1}^{n} \frac{1}{i} \leq (1 + \log(n)).$$

  where $w_i \in (0, 1]$.

- In (46d) and (46e) $Q_t(a) = \sum_{s\in[t], A_s = a} q_s$.

- (46e) comes from Jansen inequality.

Combine (45) and (46), we have

$$\sum_{t=1}^{\tau-1} q_t \text{UCB}_{r,t}(A_t) \leq \sum_{t=1}^{\tau-1} q_t r_t + O\left(\sqrt{\bar{q}K \log\left(\frac{1}{\delta}\right) \sum_{t=1}^{\tau-1} q_t \text{UCB}_{r,t}(A_t)} + \bar{q}K \log\left(\frac{T}{K}\right) \log\left(\frac{1}{\delta}\right) + \bar{q}K \log\left(\frac{1}{\delta}\right)\right),$$

which is equivalent to

$$\left(\sqrt{\sum_{t=1}^{\tau-1} q_t \text{UCB}_{r,t}(A_t)} - O\left(\sqrt{\bar{q}K \log\left(\frac{1}{\delta}\right)}\right)\right)^2 \leq \sum_{t=1}^{\tau-1} q_t r_t + O\left(\bar{q}K \log\left(\frac{T}{K}\right) \log\left(\frac{1}{\delta}\right) + \bar{q}K \log\left(\frac{1}{\delta}\right)\right),$$

Hence,

$$\begin{aligned}\sqrt{\sum_{t=1}^{\tau-1} q_t \text{UCB}_{r,t}(A_t)} &\leq O\left(\sqrt{\bar{q}K \log\left(\frac{1}{\delta}\right)}\right) + \sqrt{\sum_{t=1}^{\tau-1} q_t r_t + O\left(\bar{q}K \log\left(\frac{T}{K}\right) \log\left(\frac{1}{\delta}\right) + \bar{q}K \log\left(\frac{1}{\delta}\right)\right)}\\ &\leq \sqrt{\sum_{t=1}^{\tau-1} q_t r_t} + O\left(\sqrt{\bar{q}K \log\left(\frac{T}{K}\right) \log\left(\frac{1}{\delta}\right)} + \sqrt{\bar{q}K \log\left(\frac{1}{\delta}\right)}\right).\end{aligned} \tag{47}$$

Combine (45) and (46), (47), we finish the proof. □

Proof of Lemma D.10. The proof is quite similar to Lemma D.9, so we omit the descriptive details. Similarly, with probability $\geq 1 - 2Td\delta$, we have

$$\begin{aligned}\left|\sum_{t=1}^{\tau-1} q_t c(A_t, i) - \sum_{t=1}^{\tau-1} q_t C_{t,i}\right| &= \overline{q}\left|\sum_{t=1}^{\tau-1} \frac{q_t}{\overline{q}}(c(A_t) - C_{t,i})\right| \\ &\leq O\left(\sqrt{\overline{q}\log\left(\frac{1}{\delta}\right)\sum_{t=1}^{\tau-1} q_t c(A_t, i)} + \overline{q}\log\left(\frac{1}{\delta}\right)\right) \\ &\leq O\left(\sqrt{\overline{q}\log\left(\frac{1}{\delta}\right)\sum_{t=1}^{\tau-1} q_t \text{UCB}_{c,t}(A_t, i)} + \overline{q}\log\left(\frac{1}{\delta}\right)\right),\end{aligned} \tag{48}$$

Then with probability $\geq 1 - 3KTd\delta$, we also have

$$\begin{aligned}\left|\sum_{t=1}^{\tau-1} q_t \text{LCB}_{c,t}(A_t, i) - \sum_{t=1}^{\tau-1} q_t c(A_t, i)\right| &\leq 6\sum_{t=1}^{\tau-1} q_t \text{rad}(c(A_t, i), N_{t-1}^+(A_t), \delta) \\ &\leq 6 \sum_{a\in\mathcal{A}:N_{\tau-1}(a)>0} \sum_{n=1}^{N_{\tau-1}(a)} q_n(a)\text{rad}\left(c(a,i), n, \delta\right) \\ &= 6\overline{q} \sum_{a\in\mathcal{A}:N_{\tau-1}(a)>0} \sum_{n=1}^{N_{\tau-1}(a)} \frac{q_n(a)}{\overline{q}}\left(\sqrt{\frac{2c(a,i)\log\left(\frac{1}{\delta}\right)}{n}} + \frac{4}{n}\log\left(\frac{1}{\delta}\right)\right) \\ &\leq 6\overline{q} \sum_{a\in\mathcal{A}:N_{\tau-1}(a)>0} \left(2\sqrt{2c(a,i)\frac{Q_{\tau-1}(a)}{\overline{q}}\log\left(\frac{1}{\delta}\right)} + 4\left(1 + \log(N_{\tau-1}(a))\right)\log\left(\frac{1}{\delta}\right)\right) \\ &\leq 12\left(\sqrt{2\overline{q}K\log\left(\frac{1}{\delta}\right)\sum_{a\in\mathcal{A}} c(a,i)Q_{\tau-1}(a)} + 2\overline{q}K\log\left(\frac{T}{K}\right)\log\left(\frac{1}{\delta}\right) + 2\overline{q}K\log\left(\frac{1}{\delta}\right)\right) \\ &\leq 12\left(\sqrt{2\overline{q}K\log\left(\frac{1}{\delta}\right)\sum_{t=1}^{\tau-1} q_t\text{UCB}_{c,t}(A_t, i)} + 2\overline{q}K\log\left(\frac{T}{K}\right)\log\left(\frac{1}{\delta}\right) + 2\overline{q}K\log\left(\frac{1}{\delta}\right)\right).\end{aligned} \tag{49}$$

Similarly,

$$\begin{aligned}\left|\sum_{t=1}^{\tau-1} q_t \text{UCB}_{c,t}(A_t, i) - \sum_{t=1}^{\tau-1} q_t c(A_t, i)\right| &\leq 6\sum_{t=1}^{\tau-1} q_t \text{rad}(c(A_t, i), N_{t-1}^+(A_t), \delta) \\ &\leq O\left(\sqrt{\overline{q}K\log\left(\frac{1}{\delta}\right)\sum_{t=1}^{\tau-1} q_t\text{UCB}_{c,t}(A_t, i)} + \overline{q}K\log\left(\frac{T}{K}\right)\log\left(\frac{1}{\delta}\right) + \overline{q}K\log\left(\frac{1}{\delta}\right)\right).\end{aligned} \tag{50}$$

Combine (48) and (50), we have

$$\sum_{t=1}^{\tau-1} q_t\text{UCB}_{c,t}(A_t, i) \leq \sum_{t=1}^{\tau-1} q_t C_{t,i} + O\left(\sqrt{\overline{q}K\log\left(\tfrac{1}{\delta}\right)\sum_{t=1}^{\tau-1} q_t\text{UCB}_{c,t}(A_t, i)} + \overline{q}K\log\left(\tfrac{T}{K}\right)\log\left(\tfrac{1}{\delta}\right) + \overline{q}K\log\left(\tfrac{1}{\delta}\right)\right),$$

which is equivalent to

$$\left(\sqrt{\sum_{t=1}^{\tau-1} q_t\text{UCB}_{c,t}(A_t, i)} - O\left(\sqrt{\overline{q}K\log\left(\frac{1}{\delta}\right)}\right)\right)^2 \leq \sum_{t=1}^{\tau-1} q_t C_{t,i} + O\left(\overline{q}K\log\left(\frac{T}{K}\right)\log\left(\frac{1}{\delta}\right) + \overline{q}K\log\left(\frac{1}{\delta}\right)\right),$$

Hence,

$$\sqrt{\sum_{t=1}^{\tau-1} q_t \mathrm{UCB}_{c,t}(A_t, i)} \leq O\left(\sqrt{\overline{q}K\log\left(\frac{1}{\delta}\right)}\right) + \sqrt{\sum_{t=1}^{\tau-1} q_t C_{t,i} + O\left(\overline{q}K\log\left(\frac{T}{K}\right)\log\left(\frac{1}{\delta}\right) + \overline{q}K\log\left(\frac{1}{\delta}\right)\right)}$$
$$\leq \sqrt{\sum_{t=1}^{\tau-1} q_t C_{t,i}} + O\left(\sqrt{\overline{q}K\log\left(\frac{T}{K}\right)\log\left(\frac{1}{\delta}\right)} + \sqrt{\overline{q}K\log\left(\frac{1}{\delta}\right)}\right) \tag{51}$$
$$\leq \sqrt{B} + O\left(\sqrt{\overline{q}K\log\left(\frac{T}{K}\right)\log\left(\frac{1}{\delta}\right)} + \sqrt{\overline{q}K\log\left(\frac{1}{\delta}\right)}\right),$$

where the last inequality comes from the definition of the stopping time τ. Combine (48) and (49), (51), we finish the proof. ☐