Inventory Balancing with Online Learning

Wang Chi Cheung
National University of Singapore, NUS Engineering, Department of Industrial Systems Engineering and Management, Singapore, SG 117576, wangchimit@gmail.com

Will Ma
Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA 02139, willma@mit.edu

David Simchi-Levi
Institute for Data, Systems, and Society, Department of Civil and Environmental Engineering, and Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA 02139, dslevi@mit.edu

Xinshang Wang
Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge, MA 02139, xinshang@mit.edu

We study a general problem of allocating limited resources to heterogeneous customers over time, under model uncertainty. Each type of customer can be serviced using different actions, each of which stochastically consumes some combination of resources, and returns different rewards for the resources consumed. We consider a general model framework, where the resource consumption distribution associated with each (customer type, action)-combination is not known, but is consistent and can be learned over time. In addition, the sequence of customer types to arrive over time is arbitrary and completely unknown. We achieve near optimality under both model uncertainty and customer heterogeneity by judiciously synergizing two algorithmic frameworks in the literature: inventory balancing, which "reserves" a portion of each resource for high-reward customer types which could later arrive; and online learning, which shows how to "explore" the resource consumption distributions of each customer type under different actions. We define an auxiliary problem, which allows for existing competitive ratio and regret bounds to be seamlessly integrated. Furthermore, we show that the performance guarantee generated by our framework is tight, using the special case of the online bipartite matching problem with unknown match probabilities. Finally, we demonstrate the practicality and efficacy of algorithms generated by our framework using a publicly available hotel data set.

1. Introduction

Online resource allocation is a fundamental topic in many applications of operations research, such as revenue management, display advertisement allocation and appointment scheduling. In each of these settings, a manager needs to allocate limited resources to a heterogeneous pool of customers arriving in real time, while maximizing a certain notion of cumulative reward. The starting amount of each resource is exogenous, and these resources cannot be replenished during the planning horizon.

In many applications, the manager can observe a list of feature values of each arriving customer, which allows the manager to customize allocation decisions in real time. For example, a display advertising platform operator is usually provided with the internet cookie from a website visitor, upon the latter's arrival. Consequently, the operator is able to display relevant advertisements to each website visitor, in a bid to maximize the total revenue earned from clicks on these advertisements.

To achieve an optimal allocation in the presence of resource constraints, the manager's allocation decisions at any moment has to take into account the features of both the arriving customer as well as the customers who will arrive in the future. For example, in selling airline tickets, it is profitable to judiciously reserve a number of seats for business class customers, who often purchase tickets close to departure time (Talluri and van Ryzin 1998, Ball and Queyranne 2009). In healthcare applications, when making advance appointments for out-patients, it is critical to reserve certain physicians' hours for urgent patients (Feldman et al. 2014, Truong 2015). In these examples, the manager's central task is to *reserve* the right amount of each resource for future customers so as to maximize the expected reward.

While resource reservation is vital for optimizing online resource allocations, the implementation of resource reservation is often hindered by the following two challenges. First, the manager often lacks an accurate forecast model about the arrival patterns of future demand. For example, it may be difficult to model the demand spike on Black Friday as a stochastic process.

Second, the manager is often uncertain about the relationship between an arriving customer's behavior (e.g., click-through rate) and his/her observed features. For example, when selling a new product at an online platform, the manager initially has very little information about the relationship between a customer's observed feature values and his/her willingness to pay for a product.

These challenges in implementing resource reservation raise the following research question: *Can the manager perform resource reservation effectively, in the absence of any demand forecast model and under uncertain customer behavior?*

1.1. Overview of Model and Main Results

We describe an online resource allocation problem using general, neutral terminology. A central platform starts with finite and discrete amounts of inventory for multiple resources. Each unit of a resource yields a reward when consumed by a customer. Customers arrive sequentially, each of which is characterized by a context vector that describes the customer's features. Upon the arrival of each customer, the platform selects an action, which corresponds to offering a subset of resources to the customer. Then the platform accumulates the reward value for each unit of resource consumed by the customer. The objective of the central platform is to maximize the total reward collected from all the resources.

We make the following two important assumptions in our model:

1. The number of future customer arrivals and the context vectors of each one of them are unknown and picked by an adversary. As a result, the historical observation at any time step does not provide any information about future customers.

2. For each potential combination of context vector and action, there is a fixed unknown distribution over the consumption outcome. That is, two customers arriving at different time periods with identical context vectors will have the same consumption distribution. As a concrete example, in e-commerce, the context vector represents the characteristics (e.g., age, location) of an online shopper. We are assuming that the conversion rate only depends on the characteristics of the shopper and the product offered. The platform needs to learn these conversion rates in an online fashion.

Each of these two assumptions has been studied extensively, but only separately, in the literature. In models with the first assumption alone, customer behavior such as purchase probabilities is known, and the difficulty is to conduct resource reservation without any demand forecast. The conventional approach is to balance the cost of reserving each unit of a resource with the opportunity cost of allocating it. We call such techniques *inventory balancing*. Models with the second assumption alone are online learning problems (more specifically, stochastic bandit problems), in which resources have unlimited inventories, or the context vector of each customer is randomly drawn from a fixed distribution. The key trade-off in such problems is between exploring customer behavior and exploiting immediate rewards. We review research results in each of these domains in Section 1.2.

In this research, we present a unified analysis in the presence of both of these two assumptions. We make the following contributions:

- We propose a framework that integrates the inventory balancing technique with a broad class of online learning algorithms (Section 3). The framework produces online allocation algorithms with provable performance guarantees (Section 4), which can be informally expressed as

$$\mathsf{ALG} \geq \alpha \mathsf{OPT} - \mathsf{REG}, \tag{1}$$

where ALG is the performance of the algorithm produced by our framework; OPT is an upper bound on the expected revenue of an optimal algorithm which knows both the arrival sequence and the click probabilities in advance; REG represents the *regret*, i.e., the loss from exploring customer behavior; and α can be viewed as the *competitive ratio* when customer behavior is known, i.e., when $\mathsf{REG} = 0$.

- As an application of the framework, we analyze an online bipartite matching problem where edges, once assigned, are only matched with an unknown probability (Section 5). We use the framework to generate an online matching algorithm based on the Upper Confidence Bound (UCB) technique. We prove that the algorithm has performance guarantee

$$\mathsf{ALG} \geq \mathsf{OPT} - \frac{\mathsf{OPT}}{e} - \tilde{O}(\sqrt{\mathsf{OPT}}). \tag{2}$$

As a result, $\mathsf{ALG}/\mathsf{OPT}$ is bounded from below by $1 - 1/e - \tilde{O}(1/\sqrt{\mathsf{OPT}})$, which approaches the best-possible competitive ratio of $1 - 1/e$ as OPT becomes large (i.e. the regret from learning the matching probabilities becomes negligible). We also show that this is tight: we construct a setting where in (2), the loss of OPT/e is unavoidable due to not knowing the arrival sequence in advance, and the loss of $\tilde{O}(\sqrt{\mathsf{OPT}})$ is unavoidable due to not knowing the matching probabilities in advance.

- We study a dynamic assortment planning problem in which each resource can be sold at different reward rates (Section 6). We propose an online algorithm based on Thompson sampling, and test it on the hotel dataset of Bodea et al. (2009) (Section 7).

1.2. Literature Review

To our understanding, we are the first to give a unified analysis for online algorithms involving (i) resource constraints, (ii) learning customer behavior, and (iii) adversarial customer arrivals.

1.2.1. Competitive analysis. We first briefly review the literature analyzing the competitive ratio for resource allocation problems under adversarial arrivals. This technique is often called *competitive analysis*, and for a more extensive background, we refer the reader to Borodin and El-Yaniv (2005). For more on the application of competitive analysis in online matching and allocation problems, we refer to Mehta (2013). For more on the application of competitive analysis in airline revenue management problems, we refer to the discussions in Ball and Queyranne (2009).

Our work is focused on the case where competitive analysis is used to manage the consumption of resources. The prototypical problem in this domain is the Adwords problem (Mehta et al. 2007). Often, the resources are considered to have large starting capacities—this assumption is equivalently called the “small bids assumption” (Mehta et al. 2007), “large inventory assumption” (Golrezaei et al. 2014), or “fractional matching assumption” (Kalyanasundaram and Pruhs 2000). In our work, we use the best-known bound that is parametrized by the starting inventory amounts (Ma and Simchi-Levi 2017). The Adwords problem originated from the classical online matching problem (Karp et al. 1990)—see Devanur et al. (2013) for a recent unified analysis. The competitive ratio aspect of our analysis uses ideas from this analysis as well as the primal-dual analysis of Adwords (Buchbinder et al. 2007). We also refer to Devanur and Jain (2012), Kell and Panigrahi (2016), Ma and Simchi-Levi (2017) for recent generalizations of the Adwords problem.

Our model also allows for probabilistic resource consumption, resembling many recent papers in the area starting with Mehta and Panigrahi (2012). We incorporate the *assortment* framework of

Golrezaei et al. (2014), where the probabilistic consumption comes in the form of a random customer choice—see also Chen et al. (2016), Ma and Simchi-Levi (2017). However, unlike those three papers on assortment planning, which assume some *substitutability* assumption in the customer choice model, we instead allow for resources which have ran out to still be consumed, but at zero reward.

1.2.2. Online learning. The problem of learning customer behavior is conventionally studied in the field of *online learning*. For a comprehensive review on recent advances in online learning, we refer the reader to Bubeck and Cesa-Bianchi (2012), Slivkins (2017).

Our research focuses on online learning problems with resources constraints. Badanidiyuru et al. (2014), Agrawal and Devanur (2014) incorporate resource constraints into the standard multi-armed bandit problem, and propose allocation algorithms with provable upper bounds on the regret. Badanidiyuru et al. (2013), Agrawal and Devanur (2016), Agrawal et al. (2016) study extensions in which customers are associated with independently and identically distributed context vectors; the values of reward and resource consumption are determined by the customer context. Besbes and Zeevi (2009, 2012), Babaioff et al. (2015), Wang et al. (2014), Ferreira et al. (2016) study pricing strategies for revenue management problems, where a resource-constrained seller offers a price from a potential infinite price set to each arriving customer. Customers are homogeneous, in the sense that each customer has the same purchase probability under the same offered price.

Those models with resource constraints in the current literature assume that the type (if there is any) of each customer is drawn from a fixed distribution that does not change over time. As a result, there exists an underlying fixed randomized allocation strategy (typically based on an optimal linear programming solution) that converges to optimality as the number of customers becomes large. The idea of the online learning techniques involved in the above-mentioned research works is to try to converge to that fixed allocation strategy. In our model, however, there is no such fixed allocation strategy that we can discover over time. For instance, the optimal algorithm in our model may reject all the low-fare customers who arrive first and reserve all the resources for high-fare customers who arrive at the end. As a result, the optimal algorithm does not earn any reward at first, and thus cannot be identified as the best strategy by any learning technique. Our analysis is innovative as we construct learning algorithms with strong performance guarantees without trying to converge to any benchmark allocation strategy.

2. Model Formulation

Throughout this paper, we let $\mathbb{N}$ denote the set of positive integers. For any $n \in \mathbb{N}$, let $[n]$ denote the set $\{1, 2, ..., n\}$.

Consider a class of online resource allocation problems, generically modeled as follows. A central platform starts with $n \in \mathbb{N}$ resources. Each resource $i \in [n]$ has a reward $r_i > 0$ associated with it (Later in Section 6, we will allow each resource to have multiple reward values). Each resource also has an unreplenishable discrete starting inventory $b_i \in \mathbb{N}$. We denote $r_{\max} = \max_{i \in [n]} r_i$, and $b_{\min} = \min_{i \in [n]} b_i$.

There is a latent sequence of T customers who will arrive sequentially. Each customer t is associated with a context vector $x^t \in \mathcal{X}$, where $\mathcal{X}$ is a known context set. The sequence of context vectors $x^1, x^2, \ldots, x^T$ is revealed sequentially. That is, for each t, the platform must make the decision for customer t, without knowing the values of $x^{t+1}, \ldots, x^T$ nor the value of T. We will use the phrases "customer t" and "time period t" interchangeably.

When a customer arrives at the platform, the platform observes the customer context $x \in \mathcal{X}$ and takes an action $a \in \mathcal{A}$. Under context x and action a, the customer's behavior is governed by a distribution $\rho_{x,a}$ over *outcomes*. For each x, a, the distribution $\rho_{x,a}$ is *not known* to the platform. An outcome is defined by the subset of resources consumed, given by a vector $\mathbf{y}$ in $\{0,1\}^n$. If $\mathbf{y}_i = 1$ and resource i is not yet depleted, then one unit of the inventory of resource i is consumed, and a reward of r_i is earned. If $\mathbf{y}_i = 1$ but resource i is depleted, or if $\mathbf{y}_i = 0$, then no resource i is consumed and no reward is earned. For all $x \in \mathcal{X}$, $a \in \mathcal{A}$, and $\mathbf{y} \in \{0,1\}^n$, let $\rho_{x,a}(\mathbf{y})$ be the probability of outcome $\mathbf{y}$ when action a is played on context x.

In each period t, events occur in the following sequence. First, the customer context x^t is revealed. Second, the platform plays an action $a^t \in \mathcal{A}$ on x^t. The action a^t is determined by an *online algorithm* that sees x^t and all the information prior to period t. Third, the platform observes the outcome $\mathbf{y}^t \in \{0,1\}^n$ in period t, drawn according to ρ_{x^t,a^t}, and collects rewards.

The sequence of context vectors $(x^1, x^2, ..., x^T)$ and the mapping ρ can be interpreted as being chosen by an *oblivious adversary*, who cannot see any information related to $(a^1, a^2, \ldots, a^T, \mathbf{y}^1, \mathbf{y}^2, \ldots, \mathbf{y}^T)$. As a result, we will treat all of the adversarially-chosen parameters T, $(x^1, x^2, ..., x^T)$, and ρ as being deterministic and chosen ahead-of-time. An algorithm is evaluated on the expected reward it earns, for any fixed set of $T, (x^t)_{t=1}^T, \rho$. In Section 4, we will bound the expected reward earned by our algorithms from Section 3, in comparison to that earned by an optimal algorithm which knows all of $T, (x^t)_{t=1}^T, \rho$ in advance, with the bound holding for any values of $T, (x^t)_{t=1}^T, \rho$.

3. Online Allocation Algorithm

In this section, we present a framework which generates online allocation algorithms by integrating a broad class of online learning techniques, such as Upper Confidence Bounds (UCBs) and Thompson

Sampling, with the inventory-balancing technique that hedges against an unknown sequence of customer contexts.

The framework first creates an auxiliary problem, which exclusively focuses on the exploration-exploitation trade-off, by removing all the inventory constraints from the original problem. In other words, there is no need to conduct resource reservation in the auxiliary problem. As a result, we can apply existing online learning techniques on the auxiliary problem, and achieve regrets sub-linear in T. Next, given any online learning algorithm for the auxiliary problem, the framework converts it into another algorithm that performs both learning and resource reservation for the original model.

3.1. Auxiliary Problem

The auxiliary problem is a contextual stochastic bandit problem, in which we define the context set $\mathcal{X}$, action set $\mathcal{A}$, and distributions $\{\rho_{x,a}\}_{x\in\mathcal{X},a\in\mathcal{A}}$ in the same way as in the original problem. The distributions $\{\rho_{x,a}\}_{x\in\mathcal{X},a\in\mathcal{A}}$ are still unknown from the beginning and needs to be learned.

The auxiliary problem differs from the original problem in two ways.

First, we define all of the resources to have unlimited inventory in the auxiliary problem. As a result, algorithms for the auxiliary problem are not concerned with any global inventory constraints, i.e. if the distributions $\{\rho_{x,a}\}_{x\in\mathcal{X},a\in\mathcal{A}}$ were known, then the optimal algorithm would simply maximize the immediate reward for each period.

Second, the reward of resource $i\in[n]$ in period $t\in[T]$ is now defined as r_i^t, which depends on t. In each period $t\in[T]$, the online algorithm is given $r_1^t,\dots,r_n^t$ before having to make decision a^t; however it does not know the reward values for future periods. Thus, we can view $(r_1^t,r_2^t,\dots,r_n^t)$ as additional contextual information that is observed by online algorithms in the beginning of period t. We assume that r_i^t is chosen by an *adaptive adversary*, so that r_i^t may depend on the actions played and outcomes realized in periods $1,...,t-1$ (whereas the sequence $(x^1,x^2,\dots,x^T)$ of context vectors are still fixed a priori in both the original problem and the auxiliary problem). We restrict the adversary so that all of the chosen rewards r_i^t are bounded from above by $r_{\max}$.

Let U denote a random variable that encapsulates all the external information used by an online algorithm. For example, U can represent the random seed used by a sampling algorithm. The adversary cannot see the realization of U. Without loss of generality, let

$$\mathcal{F}_t=(x^1,x^2,...,x^t,\mathbf{y}^1,\mathbf{y}^2,...,\mathbf{y}^{t-1},r_1^t,r_2^t,...,r_n^t,U)$$

denote all the information that an online algorithm uses to make decision a^t. If U is a constant, the algorithm is a deterministic algorithm; otherwise, the algorithm is randomly drawn from a family of deterministic algorithms, according to the distribution of U. Then, an online algorithm for the auxiliary problem can be represented by a list of oracles $\{\mathcal{O}^t(\cdot)\}_{t\geq 1}$ such that the algorithm makes decision $a^t = \mathcal{O}^t(\mathcal{F}_t)$ in period $t \in [T]$.

Given an online algorithm $\{\mathcal{O}^t(\cdot)\}_{t\geq 1}$, we define its regret for any realized sample path $\mathcal{F}_T$ as

$$\mathsf{REG}(\mathcal{F}_T) = \sum_{t\in[T]} \left[R^t(a_*^t) - R^t(\mathcal{O}^t(\mathcal{F}_t))\right], \tag{3}$$

where

$$R^t(a) = \sum_{\mathbf{y}\in\{0,1\}^n} \rho_{x^t,a}(\mathbf{y}) \sum_{i\in[n]} \mathbf{y}_i r_i^t$$

is the expected reward of taking action $a \in \mathcal{A}$ in period t in the auxiliary problem and

$$a_*^t = \arg\max_{a\in\mathcal{A}} R^t(a)$$

is the optimal action in period t.

The goal in the auxiliary problem is to minimize the expected value of the regret (3). Typically, we expect $\mathbb{E}[\mathsf{REG}(\mathcal{F}_T)] = \tilde{O}(\sqrt{T})$, where the notation $\tilde{O}(\cdot)$ omits logarithmic factors. Depending on the specific problem setting, the constants hidden in $\tilde{O}(\cdot)$ may depend on parameters that are specific to the structure of ρ. The value of $\mathbb{E}[\mathsf{REG}(\mathcal{F}_T)]$ will be used in our performance guarantee in Section 4.

3.2. Integrated Algorithm

Suppose we are given an online learning algorithm $\{\mathcal{O}^t(\cdot)\}_{t\geq 1}$ for the auxiliary problem. Our integrated algorithm defines the time-dependent rewards for the auxiliary problem based on the resources that have been consumed by that time period.

For each time $t \in [T]$ and resource $i \in [n]$, let N_i^t denote the number of units of resource i that have been consumed by the *end* of time t. N_i^0 is understood to equal b_i for all i.

Then, we define the function $\Psi(x) = \frac{e^x - 1}{e-1}$, which is commonly used in inventory-constrained online problems to hedge against adversarial arrivals (see Buchbinder et al. (2007)). This is a convex function which increases from 0 to 1 over $[0,1]$.

We are now ready to define our integrated algorithm. For each time period $t \in [T]$:

1. For each resource i, define its *discounted reward* for time t to be

$$r_i^t = r_i\left(1 - \Psi\left(\frac{N_i^{t-1}}{b_i}\right)\right), \tag{4}$$

where N_i^{t-1} is the amount of resource i that has been consumed at the start of time t;

2. Play action $a^t = \mathcal{O}^t(\mathcal{F}_t)$, where the input

$$\mathcal{F}_t = \left(x^1, x^2, \ldots, x^t, \mathbf{y}^1, \mathbf{y}^2, \ldots, \mathbf{y}^{t-1}, r_1^t, r_2^t, \ldots, r_n^t, U\right)$$

for the oracle is constructed based on the discounted rewards generated in the previous step.

Our integrated algorithm has a very specific rule (4) of choosing the value of r_i^t, which depends on the random outcomes in previous periods. In the auxiliary problem, however, we more generally allow r_i^t to take any value generated in an adaptive way. Such a relaxation does not restrict the scope of online learning techniques that we can apply. This is because for models with adaptive contextual information (recall that we view r_i^t as part of the contextual information), most online learning algorithms $\{\mathcal{O}^t(\cdot)\}_{t\geq 1}$ achieve near optimality under context vectors generated by an adaptive adversary.

4. Analysis of Online Algorithm

In this section, we prove a performance guarantee for algorithms generated by our framework. Later in Section 5.3, we will prove that this performance guarantee is tight for a special case of our model.

4.1. LP Upper Bound on Optimum

The expected reward of an algorithm which knows in advance both the distributions $\{\rho_{x,a}\}_{x\in\mathcal{X},a\in\mathcal{A}}$ and the arrival sequence $x^1, x^2, ..., x^T$ can be upper-bounded by the following LP, which is a standard result in the revenue management literature.

Primal:

$$\mathsf{OPT} = \max \sum_{i\in[n]} z_i r_i \tag{5}$$

$$\begin{aligned} z_i &\leq \sum_{a\in\mathcal{A}} \sum_{t\in[T]} s_{a,t} \sum_{\mathbf{y}\in\{0,1\}^n} \rho_{x^t,a}(\mathbf{y})\mathbf{y}_i & i\in[n] \\ z_i &\leq b_i & i\in[n] \\ \sum_{a\in\mathcal{A}} s_{a,t} &\leq 1 & t\in[T] \\ s_{a,t} &\geq 0 & a\in\mathcal{A}, t\in[T] \end{aligned}$$

In the LP, the variable $s_{a,t}$ encapsulates the unconditional probability of an algorithm taking action a in period t. Given a fixed underlying problem instance, we set OPT to be the optimal objective of the LP. Its dual can be written as follows.

Dual:

$$\min \sum_{i\in[n]} b_i\lambda_i + \sum_{t\in[T]} \gamma_t \tag{6}$$

$$\gamma_t \geq \sum_{\mathbf{y}\in\{0,1\}^n} \rho_{x^t,a}(\mathbf{y}) \sum_{i\in[n]} \mathbf{y}_i(r_i - \lambda_i) \qquad a\in\mathcal{A}, t\in[T] \tag{7}$$

$$\begin{aligned} \lambda_i &\leq r_i & i\in[n] \\ \lambda_i, \gamma_t &\geq 0 & i\in[n], t\in[T] \end{aligned}$$

4.2. Proof of Performance Guarantee

We prove the performance guarantee using a primal-dual approach. More precisely, we construct a primal solution to the LP (5), whose objective value equals the total reward of our algorithm plus the cumulative regret due to not knowing ρ in advance. We also construct a dual solution to (6), which bounds OPT from above. The performance guarantee is obtained by finding a relationship in the form of (1) between the objective values of the constructed primal and dual solutions.

We will let ALG be the random variable representing the reward earned by the algorithm under consideration. Recall that a_*^t denotes the optimal action during period t in the auxiliary problem, while N_i^t denotes number of units of resource i consumed by the end of time t; these values will also be treated as random in the analysis.

Definition 1 (Random Dual Variables). Define the following dual variables for all i, t, which are random variables:

$$\Lambda_i = r_i \cdot \Psi\left(\frac{N_i^T}{b_i}\right)$$

$$\Gamma_t = \max_{a\in\mathcal{A}} \left\{ \sum_{\mathbf{y}\in\{0,1\}^n} \rho_{x^t,a}(\mathbf{y}) \sum_{i\in[n]} \mathbf{y}_i r_i \left[1-\Psi\left(\frac{N_i^{t-1}}{b_i}\right)\right]\right\} = \sum_{\mathbf{y}\in\{0,1\}^n} \rho_{x^t,a_*^t}(\mathbf{y}) \sum_{i\in[n]} \mathbf{y}_i r_i \left[1-\Psi\left(\frac{N_i^{t-1}}{b_i}\right)\right].$$

Lemma 1 **(Feasibility)**. *The dual variables defined in Definition 1 are feasible on every sample path. Therefore, if we set $\lambda_i = \mathbb{E}[\Lambda_i]$ for all i and $\gamma_t = \mathbb{E}[\Gamma_t]$ for all t, then this provides a feasible solution to the Dual LP.*

The following theorem gives our main result, which states that the optimality gap of our online allocation algorithm is at most a constant fraction of OPT plus the regret in the auxiliary problem.

Theorem 1. *The total reward ALG of the algorithm generated by our framework satisfies*

$$\mathsf{OPT} \le \frac{(1+b_{min})(1-e^{-1/b_{min}})}{1-1/e} \cdot \mathbb{E}[\mathsf{ALG}] + \mathbb{E}[\mathsf{REG}(\mathcal{F}_T)]. \tag{8}$$

When $b_{min} \to \infty$, the above expression can be written as

$$\mathsf{OPT} - \mathbb{E}[\mathsf{ALG}] \le \frac{1}{e}\mathsf{OPT} + \left(1-\frac{1}{e}\right)\mathbb{E}[\mathsf{REG}(\mathcal{F}_T)]. \tag{9}$$

Recall that $b_{\min}$ denotes the smallest starting inventory among the resources. The expression $(1+b_{\min})(1-e^{-1/b_{\min}})$ in bound (8) represents the best-known dependence on $b_{\min}$ in the competitive ratio (Ma and Simchi-Levi 2017). The expression decreases to 1 as $b_{\min} \to \infty$.

Proof. $\mathsf{OPT} \le \sum_{i=1}^{n} b_i \mathbb{E}[\Lambda_i] + \sum_{t=1}^{T} \mathbb{E}[\Gamma_t]$, by Lemma 1 and weak duality. Using the definitions of Γ_t, Λ_i, and $R^t(\cdot)$, we obtain

$$\begin{aligned}\mathsf{OPT} \le & \mathbb{E}\left\{\sum_{i\in[n]} b_i r_i \cdot \Psi\left(\frac{N_i^T}{b_i}\right) + \sum_{t\in[T]} \sum_{\mathbf{y}\in\{0,1\}^n} \rho_{x^t,a_*^t}(\mathbf{y}) \sum_{i\in[n]} \mathbf{y}_i r_i \left[1 - \Psi\left(\frac{N_i^{t-1}}{b_i}\right)\right]\right\} \\ = & \sum_{i\in[n]} b_i r_i \cdot \sum_{t\in[T]} \mathbb{E}\left[\left(\Psi(\frac{N_i^t}{b_i}\right) - \Psi\left(\frac{N_i^{t-1}}{b_i}\right)\right] + \mathbb{E}\left[\sum_{t\in[T]} R^t(a_*^t)\right].\end{aligned}$$

Recall that $\mathcal{F}_t$ is the input to the auxiliary problem at the start of time t, which determines the values of N_i^{t-1}. Conditioned on $\mathcal{F}_t$, the algorithm's action $a^t = \mathcal{O}^t(\mathcal{F}_t)$ is determined. Thus, for any resource i,

$$\mathbb{E}\left[\Psi\left(\frac{N_i^t}{b_i}\right)\middle|\mathcal{F}_t\right] - \Psi\left(\frac{N_i^{t-1}}{b_i}\right) = \sum_{\mathbf{y}\in\{0,1\}^n} \rho_{x^t,a^t}(\mathbf{y})\mathbf{y}_i \left[\Psi\left(\frac{\min\{b_i, N_i^{t-1}+1\}}{b_i}\right) - \Psi\left(\frac{N_i^{t-1}}{b_i}\right)\right]. \quad (10)$$

We explain equation (10). Conditional on any $\mathcal{F}_t$, the vector of outcomes $\mathbf{y}$ is distributed according to ρ_{x^t,a^t}. If $\mathbf{y}_i = 1$ and resource i is not yet depleted, i.e. $N_i^{t-1} < b_i$, then a unit of resource i is consumed, leading to $N_i^t = N_i^{t-1} + 1 \le b_i$. If $\mathbf{y}_i = 1$ but resource i is depleted, i.e. $N_i^{t-1} = b_i$, then resource i cannot be consumed further, leading to $N_i^t = b_i$.

Using the tower property of conditional expectation over the randomness in $\mathcal{F}_t$, and substituting in equation (10), we obtain

$$
\begin{aligned}
\mathsf{OPT} \le& \mathbb{E}\left[\sum_{i\in[n]} b_i r_i \cdot \sum_{t\in[T]} \left(\Psi\left(\frac{N_i^t}{b_i}\right) - \Psi\left(\frac{N_i^{t-1}}{b_i}\right)\right)\right] + \mathbb{E}\left[\sum_{t\in[T]} R^t(a_*^t)\right] \\
=& \sum_{t\in[T]} \mathbb{E}\left[\mathbb{E}\left[\sum_{i\in[n]} b_i r_i \cdot \left(\Psi\left(\frac{N_i^t}{b_i}\right) - \Psi\left(\frac{N_i^{t-1}}{b_i}\right)\right) \middle| \mathcal{F}_t\right]\right] + \mathbb{E}\left[\sum_{t\in[T]} R^t(a_*^t)\right] \\
=& \sum_{t\in[T]} \mathbb{E}\left[\sum_{\mathbf{y}\in\{0,1\}^n} \rho_{x^t,a^t}(\mathbf{y}) \sum_{i\in[n]} b_i r_i \cdot \mathbf{y}_i \left(\Psi\left(\frac{\min\{b_i, N_i^{t-1}+1\}}{b_i}\right) - \Psi\left(\frac{N_i^{t-1}}{b_i}\right)\right)\right] + \mathbb{E}\left[\sum_{t\in[T]} R^t(a_*^t)\right] \\
=& \sum_{t\in[T]} \mathbb{E}\left[\sum_{\mathbf{y}\in\{0,1\}^n} \rho_{x^t,a^t}(\mathbf{y}) \sum_{i\in[n]} r_i \cdot \mathbf{y}_i \frac{\Psi\left(\frac{\min\{b_i, N_i^{t-1}+1\}}{b_i}\right) - \Psi\left(\frac{N_i^{t-1}}{b_i}\right)}{1/b_i}\right] + \mathbb{E}\left[\sum_{t\in[T]} R^t(a_*^t)\right] \\
=& \sum_{t\in[T]} \mathbb{E}\left[\sum_{\mathbf{y}\in\{0,1\}^n} \rho_{x^t,a^t}(\mathbf{y}) \sum_{i\in[n]} r_i \cdot \mathbf{y}_i \left(\frac{\Psi\left(\frac{\min\{b_i, N_i^{t-1}+1\}}{b_i}\right) - \Psi\left(\frac{N_i^{t-1}}{b_i}\right)}{1/b_i} + 1 - \Psi\left(\frac{N_i^{t-1}}{b_i}\right)\right)\right] \\
&+ \mathbb{E}\left[\sum_{t\in[T]} (R^t(a_*^t) - R^t(a^t))\right] \\
=& \sum_{t\in[T]} \mathbb{E}\left[\sum_{\mathbf{y}\in\{0,1\}^n} \rho_{x^t,a^t}(\mathbf{y}) \sum_{i\in[n]} r_i \cdot \mathbf{y}_i \left(\frac{\Psi\left(\frac{\min\{b_i, N_i^{t-1}+1\}}{b_i}\right) - \Psi\left(\frac{N_i^{t-1}}{b_i}\right)}{1/b_i} + 1 - \Psi\left(\frac{N_i^{t-1}}{b_i}\right)\right)\right] \\
&+ \mathbb{E}\left[\mathsf{REG}(\mathcal{F}_T)\right].
\end{aligned}
\tag{11}
$$

Since $\Psi(1) = 1$, we must have

$$
\begin{aligned}
&\frac{\Psi\left(\frac{\min\{b_i, N_i^{t-1}+1\}}{b_i}\right) - \Psi\left(\frac{N_i^{t-1}}{b_i}\right)}{1/b_i} + 1 - \Psi\left(\frac{N_i^{t-1}}{b_i}\right) \\
=& \mathbb{1}(N_i^{t-1} < b_i) \cdot \left[\frac{\Psi\left(\frac{N_i^{t-1}+1}{b_i}\right) - \Psi\left(\frac{N_i^{t-1}}{b_i}\right)}{1/b_i} + 1 - \Psi\left(\frac{N_i^{t-1}}{b_i}\right)\right] \\
\le& \mathbb{1}(N_i^{t-1} < b_i) \cdot \frac{(1+b_{\min})(1-e^{-1/b_{\min}})}{1-1/e}.
\end{aligned}
$$

Substituting back into (11), we obtain

$$
\begin{aligned}
\mathsf{OPT} \le& \sum_{t\in[T]} \mathbb{E}\left[\sum_{\mathbf{y}\in\{0,1\}^n} \rho_{x^t,a^t}(\mathbf{y}) \sum_{i\in[n]} r_i \cdot \mathbf{y}_i \mathbb{1}(N_i^{t-1} < b_i) \cdot \frac{(1+b_{\min})(1-e^{-1/b_{\min}})}{1-1/e}\right] + \mathbb{E}[\mathsf{REG}(\mathcal{F}_T)] \\
=& \frac{(1+b_{\min})(1-e^{-1/b_{\min}})}{1-1/e} \cdot \mathbb{E}[\mathsf{ALG}] + \mathbb{E}[\mathsf{REG}(\mathcal{F}_T)],
\end{aligned}
$$

which completes the proof. □

5. Online Matching

In this section, we present a specific application of the framework on the online matching problem. In this problem, each resource i corresponds to an advertiser who is willing to spend at most $b_i \cdot r_i$ dollars for receiving clicks on the advertiser's advertisements. The context set is $\mathcal{X} = \{0,1\}^n$, where we recall that n is the number of resources/advertisers. For each $x \in \mathcal{X}$, x_i indicates whether a customer with context x can be matched to advertiser i.

Each advertiser has $K \in \mathbb{N}$ different advertisements, e.g., K videos/banners. Upon the arrival of a customer $x \in \mathcal{X}$, the platform needs to pick an advertiser $i \in [n]$ such that $x_i = 1$, and display an advertisement $k \in [K]$ that belongs to advertiser i.

The action set can be written as $\mathcal{A} = \{(i,k) : i \in [n], k \in [K]\}$. When action (i,k) is played on customer context x, the customer will click on the displayed advertisement with probability $p_{(i,k)}\mathbb{1}(x_i = 1)$. The platform earns reward r_i from each click on any advertisement belonging to advertiser i.

The values of $p_{(i,k)}$ are unknown from the beginning. For each x and (i,k), the distribution $\rho_{x,(i,k)}$ can be written as

$$\begin{aligned} \rho_{x,(i,k)}(\mathbf{e}_i) &= \mathbb{1}(x_i = 1) \cdot p_{(i,k)}, \\ \rho_{x,(i,k)}(\mathbf{0}) &= 1 - \rho_{x,(i,k)}(\mathbf{e}_i), \\ \rho_{x,(i,k)}(\mathbf{y}) &= 0 \text{ for all other outcomes } \mathbf{y} \text{ in } \{0,1\}^n. \end{aligned}$$

5.1. Algorithm for the auxiliary problem

The auxiliary problem is a variant of the classic stochastic multi-armed bandit problem, in which the expected reward of each arm (i,k) in each period t is scaled by an observable factor $r_i^t \mathbb{1}(x_i^t = 1)$. The following algorithm for the auxiliary problem is based on the UCB technique.

Let $D_{(i,k)}^t \subseteq \{1,2,...,t\}$ denote the set of periods s such that action (i,k) is taken in period s and $x_i^s = 1$. Define function $\mathsf{rad}(\alpha,\mu,N) := \sqrt{\frac{\alpha\mu}{N}} + \frac{\alpha+1}{N}$. Recall that $\mathbf{y}^t$ is the random outcome in period t. Let

$$\bar{p}_{(i,k)}^t = \frac{\sum_{s \in D_{(i,k)}^{t-1}} \mathbf{y}_i^s}{|D_{(i,k)}^{t-1}| + 1}$$

be an estimate for $p_{(i,k)}$, and

$$U_{(i,k)}^t = \mathsf{rad}\left(72\log(2nKt^2), \bar{p}_{(i,k)}^t, |D_{(i,k)}^{t-1}| + 1\right)$$

be the size of its confidence interval.

Upon the arrival of customer x^t, the algorithm takes action (i,k) that maximizes the upper confidence bound $r_i^t \cdot (\bar{p}_{(i,k)}^t + U_{(i,k)}^t(\bar{p}_{(i,k)}^t)) \cdot \mathbb{1}(x_i^t = 1)$.

Lemma 2. *(Kleinberg et al. 2008) Consider n independently and identically distributed random variables $X_1,\dots,X_n$ in $[0,1]$. Let $\mu=\mathbb{E}[X_1]$, and $X=\sum_{i=1}^{n}X_i/n$. Then, for any $\delta>0$, we have with probability at least $1-\delta$,*

$$|X-\mu|<\sqrt{\frac{72X\log\frac{2}{\delta}}{n}}+\frac{72\log\frac{2}{\delta}}{n}<3\left[\sqrt{\frac{72\mu\log\frac{2}{\delta}}{n}}+\frac{72\log\frac{2}{\delta}}{n}\right].$$

Proposition 1. *In any fixed period t, with probability at least $1-t^{-2}$ we have*

$$\left|p_{(i,k)}-\bar{p}^{t}_{(i,k)}\right|\le U^{t}_{(i,k)}\le 3\mathsf{rad}\left(72\log(2nKt^2),p_{(i,k)},|D^{t-1}_{(i,k)}|+1\right)$$

for all $(i,k)\in\mathcal{A}$.

Proof. For any fixed i,k,t, applying Lemma 2 to the sequence of random variables $\{\mathbf{y}_i^s\}_{s\in D^{t-1}_{(i,k)}}$, we can obtain with probability at least $1-\delta$,

$$\left|\sum_{s\in D^{t-1}_{(i,k)}}(\mathbf{y}_i^s-p_{(i,k)})\right|<\sqrt{72\log\frac{2}{\delta}\cdot\sum_{s\in D^{t-1}_{(i,k)}}\mathbf{y}_i^s}+72\log\frac{2}{\delta}<3\left[\sqrt{72\log\frac{2}{\delta}\cdot|D^{t-1}_{(i,k)}|p_{(i,k)}}+72\log\frac{2}{\delta}\right].$$

Since

$$\begin{aligned}&|\bar{p}^{t}_{(i,k)}-p_{(i,k)}|\\=&\left|\frac{\sum_{s\in D^{t-1}_{(i,k)}}(\mathbf{y}_i^s-p_{(i,k)})}{|D^{t-1}_{(i,k)}|+1}-\frac{p_{(i,k)}}{|D^{t-1}_{(i,k)}|+1}\right|\\\le&\left|\frac{\sum_{s\in D^{t-1}_{(i,k)}}(\mathbf{y}_i^s-p_{(i,k)})}{|D^{t-1}_{(i,k)}|+1}\right|+\frac{1}{|D^{t-1}_{(i,k)}|+1},\end{aligned}$$

we have with probability at least $1-\delta$,

$$\begin{aligned}&|\bar{p}^{t}_{(i,k)}-p_{(i,k)}|\\<&\frac{\sqrt{72\log\frac{2}{\delta}\cdot\sum_{s\in D^{t-1}_{(i,k)}}\mathbf{y}_i^s}}{|D^{t-1}_{(i,k)}|+1}+\frac{72\log\frac{2}{\delta}+1}{|D^{t-1}_{(i,k)}|+1}<3\left[\frac{\sqrt{72\log\frac{2}{\delta}\cdot|D^{t-1}_{(i,k)}|p_{(i,k)}}}{|D^{t-1}_{(i,k)}|+1}+\frac{72\log\frac{2}{\delta}+1}{|D^{t-1}_{(i,k)}|+1}\right]\end{aligned}$$

$$\Longrightarrow|\bar{p}^{t}_{(i,k)}-p_{(i,k)}|<\sqrt{\frac{72\log\frac{2}{\delta}\cdot\bar{p}^{t}_{(i,k)}}{|D^{t-1}_{(i,k)}|+1}}+\frac{72\log\frac{2}{\delta}+1}{|D^{t-1}_{(i,k)}|+1}<3\left[\sqrt{\frac{72\log\frac{2}{\delta}\cdot p_{(i,k)}}{|D^{t-1}_{(i,k)}|+1}}+\frac{72\log\frac{2}{\delta}+1}{|D^{t-1}_{(i,k)}|+1}\right].$$

The proposition is proved by choosing $\delta=\frac{1}{nKt^2}$. □

5.2. Analysis

The following proposition gives a regret upper bound on the UCB algorithm on the auxiliary problem. The bound is sub-linear in the expected optimal value $\mathbb{E}[\mathsf{OPT}]$ of the Primal linear program, and the bound is in particular sub-linear in T.

PROPOSITION 2. *In the auxiliary problem, the total regret of the UCB algorithm is*

$$\mathbb{E}[\mathsf{REG}(\mathcal{F}_T)] = \tilde{O}\left(\sqrt{nK\mathbb{E}[\mathsf{OPT}]}\right).$$

Proof. Conditioned on the high probability event that $|p_{(i,k)} - \bar{p}^t_{(i,k)}| \le U^t_{(i,k)}$ for all t, (i,k), the total regret can be bounded from above as follows:

$$\begin{aligned}
\mathsf{REG}(\mathcal{F}_T) &\le \sum_{i\in[n],k\in[K]} \sum_{t=1}^{T} r_i^t \cdot 2U^t_{a^t} \mathbb{1}(a^t = (i,k)) \\
&\le 6 \sum_{i\in[n],k\in[K]} \sum_{t=1}^{T} r_i^t \mathbb{1}(a^t = (i,k)) \mathsf{rad}\left(72\log(2nKt^2), p_{(i,k)}, |D^{t-1}_{(i,k)}| + 1\right).
\end{aligned}$$

Recall that $r_i^t = r_i(1 - \Psi(N_i^{t-1}/b_i))$, so we must have $r_i^t \le r_i \mathbb{1}(N_i^{t-1} < b_i)$. Let $d_{(i,k)} = |\{s \in D^T_{(i,k)} : N_i^{s-1} < b_i\}|$ denote the number of periods in which action (i,k) is taken and resource i still has positive remaining inventory. For any fixed $(i,k) \in \mathcal{A}$, we have

$$\begin{aligned}
&\sum_{t=1}^{T} r_i^t \mathbb{1}(a^t = (i,k)) \mathsf{rad}\left(72\log(2nKt^2), p_{(i,k)}, |D^{t-1}_{(i,k)}| + 1\right) \\
=&\sum_{t=1}^{T} r_i^t \mathbb{1}(a^t = (i,k)) \left[\sqrt{\frac{72\log(2nKt^2)\cdot p_{(i,k)}}{|D^{t-1}_{(i,k)}|+1}} + \frac{72\log(2nKt^2)+1}{|D^{t-1}_{(i,k)}|+1}\right] \\
\le&\sum_{t=1}^{T} r_i \mathbb{1}(N_i^{t-1} < b_i) \mathbb{1}(a^t = (i,k)) \left[\sqrt{\frac{72\log(2nKT^2)\cdot p_{(i,k)}}{|D^{t-1}_{(i,k)}|+1}} + \frac{72\log(2nKT^2)+1}{|D^{t-1}_{(i,k)}|+1}\right] \\
=&\sum_{s=1}^{d_{(i,k)}} r_i \left[\sqrt{\frac{72\log(2nKT^2)\cdot p_{(i,k)}}{s}} + \frac{72\log(2nKT^2)+1}{s}\right] \\
\le& r_i \cdot 2\sqrt{72\log(2nKT^2)\cdot p_{(i,k)} d_{(i,k)}} + O(r_{\max}\log^2(nKT^2)) \qquad (12) \\
\le& 12\sqrt{2r_{\max}\log(2nKT^2)}\sqrt{r_i p_{(i,k)} d_{(i,k)}} + O(r_{\max}\log^2(nKT^2)).
\end{aligned}$$

Inequality (12) follows from the fact that $\sum_{i=1}^{d} 1/\sqrt{i} \le 2\sqrt{d}$. Substituting this back into the upper bound on REG, we obtain

$$\begin{aligned}
\mathsf{REG}(\mathcal{F}_T) &\le 6 \sum_{i\in[n],k\in[K]} \sum_{t=1}^{T} r_i^t \mathbb{1}(a^t = (i,k)) \mathsf{rad}\left(72\log(2nKt^2), p_{(i,k)}, |D^{t-1}_{(i,k)}| + 1\right) \\
&\le 72\sqrt{2r_{\max}\log(2nKT^2)} \sum_{i\in[n],k\in[K]} \sqrt{r_i p_{(i,k)} d_{(i,k)}} + O(r_{\max} nK \log^2(nKT^2)) \\
&\le 72\sqrt{2r_{\max}\log(2nKT^2)} \sqrt{nK \sum_{i\in[n],k\in[K]} r_i p_{(i,k)} d_{(i,k)}} + O(r_{\max} nK \log^2(nKT^2)), \qquad (13)
\end{aligned}$$

where inequality (13) is by the Cauchy-Schwartz inequality.

For a period t, if the event $|p_{(i,k)} - \bar{p}^t_{(i,k)}| \le U^t_{(i,k)}$ fails to hold for some (i,k), then there is an additional regret at most $r_{\max}$. In expectation, the total amount of such additional regret is at most

$$\sum_{t=1}^{T} \frac{r_{\max}}{t^2} = O(1),$$

which is negligible.

Altogether, by Jensen's inequality, the total expected regret of the UCB algorithm is

$$\begin{aligned}
\mathbb{E}[\mathsf{REG}(\mathcal{F}_T)] &\le 72\sqrt{2r_{\max}\log(2nKT^2)}\,\mathbb{E}\left[\sqrt{nK\sum_{i\in[n],k\in[K]} r_i p_{(i,k)} d_{(i,k)}}\right] + O(r_{\max} nK \log^2(nKT^2)) \\
&\le 72\sqrt{2r_{\max}\log(2nKT^2)}\sqrt{nK\mathbb{E}[\sum_{i\in[n],k\in[K]} r_i p_{(i,k)} d_{(i,k)}]} + O(r_{\max} nK \log^2(nKT^2)) \\
&= 72\sqrt{2r_{\max}\log(2nKT^2)}\sqrt{nK\mathbb{E}[\mathsf{ALG}]} + O(r_{\max} nK \log^2(nKT^2)) \\
&\le 72\sqrt{2r_{\max}\log(2nKT^2)}\sqrt{nK\mathbb{E}[\mathsf{OPT}]} + O(r_{\max} nK \log^2(nKT^2)).
\end{aligned}$$

□

Altogether, using Theorem 1, we can obtain the following performance guarantee when $b_{\min} \to \infty$.

COROLLARY 1. *In the online matching problem with unknown matching probabilities, the total reward* ALG *of our algorithm satisfies*

$$\mathsf{OPT} - \mathbb{E}[\mathsf{ALG}] \le \frac{1}{e}\mathsf{OPT} + \left(1 - \frac{1}{e}\right)\mathbb{E}[\mathsf{REG}(\mathcal{F}_T)] = \frac{1}{e}\mathsf{OPT} + \tilde{O}(\sqrt{nK\mathbb{E}[\mathsf{OPT}]}). \tag{14}$$

5.3. Lower Bound

In this section, we establish a lower bound on the regret of any online algorithm for the online matching problem. Specifically, we prove that the regret bound (14) is tight in the sense that both of the loss terms OPT/e, $\tilde{O}(\sqrt{\mathbb{E}[\mathsf{OPT}]})$ are unavoidable due to the uncertainty on the click probabilities $p_{(i,k)}$ and the uncertainty on the sequence of customer contexts.

We construct a randomized worst-case instance as follows. The capacity values are the same $b_i = b$ for all $i \in [n]$. Let π be a random permutation of $[n]$. There are $T = 2bn$ customers, split into n "groups" of $2b$ customers each. The customers in each group $j \in [n]$ all have the same context (feature) vector $x^{(j)}$, where

$$x^{(j)}_i = 1 \text{ if and only if } \pi(i) \ge j.$$

In other words, if we view $\pi(i)$ as a random score of resource i, then the customers become increasingly selective as customers in group j are only interested in resources i with scores higher than j.

Let $\boldsymbol{\ell} = (\ell_1, \ldots, \ell_n) \in [K]^n$ be a random vector of "secret arms". The distribution $\rho_{x,(i,k)}$ is given by

$$\begin{aligned}\rho_{x,(i,k)}(\mathbf{e}_i) &= \mathbb{1}(x_i = 1)\left(\frac{1-\varepsilon}{2} + \mathbb{1}(k = \ell_i)\cdot\varepsilon\right)\\ \rho_{x,(i,k)}(\mathbf{0}) &= 1 - \rho_{x,(i,k)}(\mathbf{e}_i)\\ \rho_{x,(i,k)}(\mathbf{y}) &= 0 \text{ for all other outcomes } \mathbf{y} \text{ in } \{0,1\}^n\end{aligned}$$

Here, $\varepsilon \in (0, 1/2]$ will be defined in our analysis. We choose $\varepsilon \leq 1/2$ just for technical convenience.

This problem instance is a randomized one because we draw both π and $\boldsymbol{\ell}$ uniformly at random. Note that for all realization of π and $\boldsymbol{\ell}$, OPT will be bn.

A *deterministic policy* is a mapping, for any $t \in \mathbb{N}$, from any history of observed contexts and outcomes, $(x^1, \mathbf{y}^1, \ldots, x^t)$ in $\mathcal{X}^t \times \{0,1\}^{n\times(t-1)}$, to an action to play on context x^t, in $\mathcal{A}$. Our proof strategy is to upper-bound the performance of any deterministic policy on this randomized instance (it suffices to consider deterministic policies because when given the randomized instance, there always exists an optimal policy which is deterministic).

Theorem 2 **(Lower Bound)**. *Let n, b, K be any positive integers satisfying $b \geq K \geq 3$. Then there exists a randomized instance (with a random arrival sequence and a random mapping from contexts to outcomes) such that for any deterministic or randomized algorithm,*

$$\mathsf{OPT} - \mathbb{E}[\mathsf{ALG}] \geq \frac{\mathsf{OPT}}{e} + \Theta(\sqrt{K\mathsf{OPT}}).$$

We prove this theorem through Lemmas 3, 4, 5, and Proposition 3. The proof is based on an information-theoretic analysis.

Let $\mathcal{T}_j = \{2b(j-1)+1, \ldots, 2bj\}$ denote the indices of the customers in group j, for all $j \in [n]$. Let $\mathcal{A}_i = \{(i,k) : k \in [K]\}$ denote the set of actions that correspond to resource i, for all $i \in [n]$. Let Y_t be the indicator random variable for whether customer t accepted her offer, for all $t \in [T]$.

We can write ALG, the random variable for the total reward earned by the deterministic policy, as

$$\mathsf{ALG} = \sum_{i=1}^{n} \min\Big\{\sum_{j=1}^{i}\sum_{t\in\mathcal{T}_j} \mathbb{1}(Y_t = 1 \cap a^t \in \mathcal{A}_{\pi^{-1}(i)}), b\Big\}. \tag{15}$$

To upper-bound $\mathbb{E}[\mathsf{ALG}]$, we need to upper-bound $\mathbb{E}[\sum_{j=1}^{i}\sum_{t\in\mathcal{T}_j}\mathbb{1}(Y_t = 1 \cap a^t \in \mathcal{A}_{\pi^{-1}(i)})]$. Thus, we will focus on analyzing $\Pr[Y_t = 1 \cap a^t \in \mathcal{A}_{\pi^{-1}(i)}]$ for an arbitrary $i \in [n]$, $j \leq i$, and $t \in \mathcal{T}_j$.

$$\begin{aligned}&\Pr[Y_t = 1 \cap a^t \in \mathcal{A}_{\pi^{-1}(i)}]\\ =&\Pr[Y_t = 1 | a^t = (\pi^{-1}(i), \ell_{\pi^{-1}(i)})]\cdot\Pr[a^t = (\pi^{-1}(i), \ell_{\pi^{-1}(i)})]\\ &+\Pr[Y_t = 1 | a^t \in \mathcal{A}_{\pi^{-1}(i)} \cap a^t \neq (\pi^{-1}(i), \ell_{\pi^{-1}(i)})]\cdot\Pr[a^t \in \mathcal{A}_{\pi^{-1}(i)} \cap a^t \neq (\pi^{-1}(i), \ell_{\pi^{-1}(i)})]\\ =&\frac{1+\varepsilon}{2}\Pr[a^t = (\pi^{-1}(i), \ell_{\pi^{-1}(i)})] + \frac{1-\varepsilon}{2}\Pr[a^t \in \mathcal{A}_{\pi^{-1}(i)} \cap a^t \neq (\pi^{-1}(i), \ell_{\pi^{-1}(i)})]\\ =&\frac{1-\varepsilon}{2}\Pr[a^t \in \mathcal{A}_{\pi^{-1}(i)}] + \varepsilon\cdot\Pr[a^t = (\pi^{-1}(i), \ell_{\pi^{-1}(i)})]\end{aligned} \tag{16}$$

The difficult term to analyze is $\Pr[a^t = (\pi^{-1}(i), \ell_{\pi^{-1}(i)})]$. Note that the distribution of a^t is affected by the entire realized vector of secret arms $\boldsymbol{\ell}$, as well as the realized values of $\pi^{-1}(1), \ldots, \pi^{-1}(j-1)$.

Now, consider an alternate universe where for each resource $m \in [n]$, all of the actions $(m,1), \ldots, (m,K)$ result in the customer accepting with probability $\frac{1-\varepsilon}{2}$, regardless of the value of ℓ_m. We can also consider the execution of the fixed, deterministic policy in this alternate universe, where we will use random variables $\overline{a}^t, \overline{Y}_t$ to refer to its execution.

LEMMA 3 **(Using information theory to get an initial bound)**. *Let $j \in [n]$ be any customer group and let t be any customer from $\mathcal{T}_j$. Let $S \subseteq [n]$ be any set of resources. Condition on any sequence of $j-1$ resources with lowest scores*

$$\pi^{-1}([j-1]) := (\pi^{-1}(1), \ldots, \pi^{-1}(j-1))$$

and vector of secret arms $\boldsymbol{\ell}$. Then

$$\begin{aligned}\sum_{m \in S} \Pr[a^t = (m, \ell_m) | \pi^{-1}([j-1]), \boldsymbol{\ell}] \leq & \sum_{m \in S} \Pr[\overline{a}^t = (m, \ell_m) | \pi^{-1}([j-1]), \boldsymbol{\ell}] \\ & + \varepsilon \sqrt{\sum_{s=1}^{t-1} \sum_{m \notin \pi^{-1}([s-1])} \Pr[\overline{a}^s = (m, \ell_m) | \pi^{-1}([j-1]), \boldsymbol{\ell}]}. \end{aligned} \tag{17}$$

Proof. For brevity, we will omit the conditioning on $\pi^{-1}(1), \ldots, \pi^{-1}(j-1)$ and $\boldsymbol{\ell}$ throughout the proof. We will also use $\mathbf{Z}^s$ to denote the vector of random variables $(Y_1, \ldots, Y_s)$ and $\mathbf{z}^s$ to denote a vector in $\{0,1\}^s$, for any $s \in [t-1]$.

First, note that a^t is the rule of the deterministic policy for choosing the action at time t, dependent on sequence of observations $\mathbf{Z}^{t-1}$ and the sequence of contexts $x^1, \ldots, x^t$ (which is captured by $\pi^{-1}(1), \ldots, \pi^{-1}(j-1)$).

$$\begin{aligned}\sum_{m \in S} \Pr[a^t = (m, \ell_m)] &= \sum_{\mathbf{z}^{t-1} \in \{0,1\}^{t-1}} \Pr[\mathbf{Z}^{t-1} = \mathbf{z}^{t-1}] \sum_{m \in S} \Pr[a^t = (m, \ell_m) | \mathbf{Z}^{t-1} = \mathbf{z}^{t-1}] \\ &\leq \sum_{\mathbf{z}^{t-1} \in \{0,1\}^{t-1}} \Pr[\overline{\mathbf{Z}}^{t-1} = \mathbf{z}^{t-1}] \sum_{m \in S} \Pr[\overline{a}^t = (m, \ell_m) | \overline{\mathbf{Z}}^{t-1} = \mathbf{z}^{t-1}] \\ &\quad + \delta(\overline{\mathbf{Z}}^{t-1}, \mathbf{Z}^{t-1}) \\ &\leq \sum_{m \in S} \Pr[\overline{a}^t = (m, \ell_m)] + \sqrt{\frac{1}{2} \mathsf{KL}(\overline{\mathbf{Z}}^{t-1} \| \mathbf{Z}^{t-1})}, \end{aligned} \tag{18}$$

where the first inequality is from the definition that

$$\delta(\overline{\mathbf{Z}}^{t-1}, \mathbf{Z}^{t-1}) = \sum_{\mathbf{z}^{t-1} \in \{0,1\}^{t-1}} |\Pr[\overline{\mathbf{Z}}^{t-1} = \mathbf{z}^{t-1}] - \Pr[\mathbf{Z}^{t-1} = \mathbf{z}^{t-1}]|,$$

and the second inequality is due to Pinsker's inequality.

$$
\begin{aligned}
&\mathsf{KL}(\overline{\mathbf{Z}}^{t-1} \| \mathbf{Z}^{t-1}) \\
=&\sum_{\mathbf{z}^{t-1}\in\{0,1\}^{t-1}} \Pr[\overline{\mathbf{Z}}^{t-1}=\mathbf{z}^{t-1}]\cdot\ln\frac{\Pr[\overline{\mathbf{Z}}^{t-1}=\mathbf{z}^{t-1}]}{\Pr[\mathbf{Z}^{t-1}=\mathbf{z}^{t-1}]} \\
=&\sum_{s=1}^{t-1}\sum_{\mathbf{z}^{s-1}\in\{0,1\}^{t-1}} \Pr[\overline{\mathbf{Z}}^{s-1}=\mathbf{z}^{s-1}]\left(\sum_{y_s\in\{0,1\}}\Pr[\overline{Y}_s=y_s|\overline{\mathbf{Z}}^{s-1}=\mathbf{z}^{s-1}]\cdot\ln\frac{\Pr[\overline{Y}_s=y_s|\overline{\mathbf{Z}}^{s-1}=\mathbf{z}^{s-1}]}{\Pr[Y_s=y_s|\mathbf{Z}^{s-1}=\mathbf{z}^{s-1}]}\right),
\end{aligned}
$$

where the second equality comes from the Chain Rule for KL-divergences. Now, consider the term inside the parentheses. Conditioned on $\mathbf{z}^{s-1}$ (and $\pi^{-1}([j-1])$, which have been omitted in the notation), actions $\overline{a}^s$ and a^s are deterministic and equal. If this action is (m,ℓ_m) for some $m\in[n]$ and $m\notin\pi^{-1}([s-1])$, then $\overline{Y}_s$ is 1 w.p. $\frac{1-\varepsilon}{2}$ while Y_s is 1 w.p. $\frac{1+\varepsilon}{2}$, and the term inside the parentheses is the KL-divergence of $\mathsf{Ber}(\frac{1+\varepsilon}{2})$ from $\mathsf{Ber}(\frac{1-\varepsilon}{2})$, equal to $\varepsilon\cdot\ln\frac{1+\varepsilon}{1-\varepsilon}$. Otherwise, $\overline{Y}_s$ and Y_s are identically distributed, and the term inside the parentheses is zero.

Therefore,

$$
\begin{aligned}
&\mathsf{KL}(\overline{\mathbf{Z}}^{t-1} \| \mathbf{Z}^{t-1}) \\
=&\sum_{s=1}^{t-1}\sum_{m\notin\pi^{-1}([s-1])}\Pr[\overline{a}^s=(m,\ell_m)]\left(\varepsilon\cdot\ln\frac{1+\varepsilon}{1-\varepsilon}\right) \\
\le&\sum_{s=1}^{t-1}\sum_{m\notin\pi^{-1}([s-1])}\Pr[\overline{a}^s=(m,\ell_m)]\left(2\varepsilon^2\right)
\end{aligned}
$$

(the inequality is because $\varepsilon\le 1/2$) and substituting into (18) completes the proof of the lemma.

□

Definition 2. Define the following random variables for all $i,j\in[n]$:

- $Q_{i,j}=\sum_{t\in\mathcal{T}_j}\mathbb{1}(a^t\in\mathcal{A}_{\pi^{-1}(i)})$ is the total number of group-j customers on whom an action corresponding to resource $\pi^{-1}(i)$ is played;
- $Q^*_{i,j}=\sum_{t\in\mathcal{T}_j}\mathbb{1}(a^t=(\pi^{-1}(i),\ell_{\pi^{-1}(i)}))$ is the total number of group-j customers on whom action $(\pi^{-1}(i),\ell_{\pi^{-1}(i)})$ is played.

Let $q_{i,j},q^*_{i,j}$ denote the expected values of $Q_{i,j},Q^*_{i,j}$, respectively. We will use $\overline{Q}_{i,j},\overline{Q}^*_{i,j},\overline{q}_{i,j},\overline{q}^*_{i,j}$ to refer to the respective quantities under the alternate universe.

Lemma 4 **(Removing dependence on t, π, and $\boldsymbol{\ell}$)**. *Let $D\subseteq[n]$ be any set of scores, and $\pi^{-1}(D)$ be the corresponding set of resources with scores D. For any group $j\in[n]$,*

$$\sum_{i\in D}\mathbb{E}[Q^*_{i,j}]\le\frac{1}{K}\sum_{i\in D}\mathbb{E}[\overline{Q}_{i,j}]+2b\varepsilon\sqrt{\frac{2bj}{K}}.$$

Proof. Consider the probability

$$\Pr[\overline{a}^s = (m, \ell_m) | \pi^{-1}([j-1]), \boldsymbol{\ell}]$$

from the RHS of inequality (17). Since $\overline{a}^s$, which refers to the alternate universe, is unaffected by the value of ℓ_m, the probability is identical after removing the conditioning on ℓ_m. We can do this for all $s = 1, \ldots, t$.

Let $\boldsymbol{\ell}_{-m}$ denote the fixed vector of secret arms for resources other than m. We take an average over the randomness in ℓ_m (drawn uniformly from $[K]$) and apply the law of total probability to obtain:

$$\begin{aligned}
&\mathbb{E}[\Pr[\overline{a}^s = (m, \ell_m) | \pi^{-1}([j-1]), \boldsymbol{\ell}]] \\
=&\mathbb{E}[\Pr[\overline{a}^s = (m, \ell_m) | \pi^{-1}([j-1]), \boldsymbol{\ell}_{-m}]] \\
\le&\mathbb{E}[\frac{1}{K}\Pr[\overline{a}^s \in \mathcal{A}_m | \pi^{-1}([j-1]), \boldsymbol{\ell}_{-m}]] \\
=&\frac{1}{K}\Pr[\overline{a}^s \in \mathcal{A}_m],
\end{aligned}$$

where the inequality is because the probability that $\overline{a}^s$ turns out to be the "secret arm" ℓ_m of resource m is $1/K$ if $\overline{a}^s \in \mathcal{A}_m$, and 0 otherwise.

Then, for any set $S \subseteq [n]$ of resources, we apply inequality (17) to obtain:

$$\begin{aligned}
&\sum_{m \in S} \Pr[a^t = (m, \ell_m)] \\
&= \sum_{m \in S} \mathbb{E}[\Pr[a^t = (m, \ell_m) | \pi^{-1}([j-1]), \boldsymbol{\ell}]] \\
&\le \sum_{m \in S} \mathbb{E}[\Pr[\overline{a}^t = (m, \ell_m) | \pi^{-1}([j-1]), \boldsymbol{\ell}]] + \varepsilon \cdot \mathbb{E}\left[\sqrt{\sum_{s=1}^{t-1} \sum_{m \notin \pi^{-1}([s-1])} \Pr[\overline{a}^s = (m, \ell_m) | \pi^{-1}([j-1]), \boldsymbol{\ell}]}\right] \\
&\le \sum_{m \in S} \mathbb{E}[\Pr[\overline{a}^t = (m, \ell_m) | \pi^{-1}([j-1]), \boldsymbol{\ell}]] + \varepsilon \cdot \sqrt{\sum_{s=1}^{t-1} \mathbb{E}\left[\sum_{m \notin \pi^{-1}([s-1])} \Pr[\overline{a}^s = (m, \ell_m) | \pi^{-1}([j-1]), \boldsymbol{\ell}]\right]} \\
&\le \sum_{m \in S} \mathbb{E}[\Pr[\overline{a}^t = (m, \ell_m) | \pi^{-1}([j-1]), \boldsymbol{\ell}]] + \varepsilon \cdot \sqrt{\sum_{s=1}^{t-1} \mathbb{E}\left[\sum_{m \in [n]} \Pr[\overline{a}^s = (m, \ell_m) | \pi^{-1}([j-1]), \boldsymbol{\ell}]\right]} \\
&\le \frac{1}{K} \sum_{m \in S} \Pr[\overline{a}^t \in \mathcal{A}_m] + \varepsilon \sqrt{\frac{1}{K} \sum_{s=1}^{t-1} \sum_{m \in [n]} \Pr[\overline{a}^s \in \mathcal{A}_m]} \\
&\le \frac{1}{K} \sum_{m \in S} \Pr[\overline{a}^t \in \mathcal{A}_m] + \varepsilon \sqrt{\frac{t}{K}}.
\end{aligned}$$

The second inequality is Jensen's inequality (the square root function is concave).

By the definition of $Q_{i,j}$ and $Q^*_{i,j}$, we sum over the $2b$ values of t in $\mathcal{T}_j$ to obtain

$$\begin{aligned}
&\sum_{i\in D}\mathbb{E}[Q^*_{i,j}]\\
=&\sum_{t\in\mathcal{T}_j}\mathbb{E}\left[\sum_{m\in\pi^{-1}(D)}\Pr[a^t=(m,\ell_m)]\right]\\
=&\sum_{t\in\mathcal{T}_j}\mathbb{E}\left[\mathbb{E}\left[\sum_{m\in S}\Pr[a^t=(m,\ell_m)]\middle|\pi^{-1}(D)=S\right]\right]\\
\le&\sum_{t\in\mathcal{T}_j}\mathbb{E}\left[\mathbb{E}\left[\frac{1}{K}\sum_{m\in S}\Pr[\overline{a}^t\in\mathcal{A}_m]+\varepsilon\sqrt{\frac{t}{K}}\middle|\pi^{-1}(D)=S\right]\right]\\
=&\frac{1}{K}\sum_{t\in\mathcal{T}_j}\mathbb{E}\left[\sum_{m\in\pi^{-1}(D)}\Pr[\overline{a}^t\in\mathcal{A}_m]\right]+\sum_{t\in\mathcal{T}_j}\varepsilon\sqrt{\frac{t}{K}}\\
=&\frac{1}{K}\sum_{i\in D}\mathbb{E}[\overline{Q}_{i,j}]+\sum_{t\in\mathcal{T}_j}\varepsilon\sqrt{\frac{t}{K}}\\
\le&\frac{1}{K}\sum_{i\in D}\mathbb{E}[\overline{Q}_{i,j}]+\varepsilon 2b\sqrt{\frac{2bj}{K}}.
\end{aligned}$$

The last inequality uses the fact that $t\le 2bj$ for all $t\in\mathcal{T}_j$.

□

Lemma 5 **(Argument for randomized permutation)**. *For any customer group $j\in[n]$ and compatible resource with score $i\ge j$, both $\mathbb{E}[Q_{i,j}]$ and $\mathbb{E}[\overline{Q}_{i,j}]$ are upper-bounded by $2b/(n-j+1)$.*

Proof. We prove the result for $\mathbb{E}[Q_{i,j}]$ (the proof for $\mathbb{E}[\overline{Q}_{i,j}]$ is identical):

$$\begin{aligned}
\mathbb{E}[Q_{i,j}]&=\sum_{t\in\mathcal{T}_j}\Pr[a^t\in\mathcal{A}_{\pi^{-1}(i)}]\\
&=\sum_{t\in\mathcal{T}_j}\sum_{\pi^{-1}([j-1])}\Pr[\pi^{-1}([j-1])]\cdot\Pr[a^t\in\mathcal{A}_{\pi^{-1}(i)}|\pi^{-1}([j-1])]\\
&=\sum_{t\in\mathcal{T}_j}\sum_{\pi^{-1}([j-1])}\Pr[\pi^{-1}([j-1])]\sum_{m\notin\pi^{-1}([j-1])}\Pr[\pi(m)=i|\pi^{-1}(i)]\cdot\Pr[a^t\in\mathcal{A}_m|\pi^{-1}([j-1]),\pi(m)=i]\\
&=\sum_{t\in\mathcal{T}_j}\sum_{\pi^{-1}([j-1])}\Pr[\pi^{-1}([j-1])]\sum_{m\notin\pi^{-1}([j-1])}\frac{1}{n-j+1}\Pr[a^t\in\mathcal{A}_m|\pi^{-1}([j-1])]\\
&\le\sum_{t\in\mathcal{T}_j}\sum_{\pi^{-1}([j-1])}\Pr[\pi^{-1}([j-1])]\cdot\frac{1}{n-j+1}(1)\\
&=\frac{2b}{n-j+1}.
\end{aligned}$$

The first equality is by definition and the linearity of expectation; the second and third equalities are by the law of total probability; and the fourth equality is by the fact that a^t is independent of $\pi^{-1}(i)$, which completes the proof of the lemma.

□

Now, combining (15), (16), and definitions, we get that

$$\mathbb{E}[\mathsf{ALG}] \leq \sum_{i=1}^{n} \min\Big\{ \sum_{j=1}^{i} \big(\frac{1-\varepsilon}{2}\mathbb{E}[Q_{i,j}] + \varepsilon \cdot \mathbb{E}[Q^*_{i,j}]\big), b \Big\}, \tag{19}$$

where we have also used the fact that $\min\{\cdot, b\}$ is concave. For all $i \in [n]$, let

$$H^n_{n-i} := (1 + \frac{1}{2} + \ldots + \frac{1}{n}) - (1 + \frac{1}{2} + \ldots + \frac{1}{n-i}) = \sum_{j=1}^{i} \frac{1}{n-j+1}. \tag{20}$$

Now, let $n' \in [n]$ be the largest value such that $H^n_{n-n'} \leq 1$.

$$\begin{aligned}
&\sum_{i=1}^{n'}\sum_{j=1}^{i} \big(\frac{1-\varepsilon}{2}\mathbb{E}[Q_{i,j}] + \varepsilon \cdot \mathbb{E}[Q^*_{i,j}]\big) \\
\leq &\sum_{i=1}^{n'} (1-\varepsilon) b \cdot H^n_{n-i} + \varepsilon \cdot \sum_{j=1}^{n'}\sum_{i=j}^{n'} \mathbb{E}[Q^*_{i,j}] \\
&\quad \text{(by Lemma 5)} \\
\leq &\sum_{i=1}^{n'} (1-\varepsilon) b \cdot H^n_{n-i} + \varepsilon \cdot \sum_{j=1}^{n'} \left(\frac{1}{K} \sum_{i=j}^{n'} \mathbb{E}[\overline{Q}_{i,j}] + \varepsilon 2b \sqrt{\frac{2bj}{K}} \right) \\
&\quad \text{(by Lemma 4)} \\
= &\sum_{i=1}^{n'} (1-\varepsilon) b \cdot H^n_{n-i} + \frac{\varepsilon}{K} \cdot \sum_{i=1}^{n'}\sum_{j=1}^{i} \mathbb{E}[\overline{Q}_{i,j}] + \varepsilon^2 2b \cdot \sum_{j=1}^{n'} \sqrt{\frac{2bj}{K}} \\
\leq &\sum_{i=1}^{n'} (1-\varepsilon) b \cdot H^n_{n-i} + \frac{\varepsilon 2b}{K} \cdot \sum_{i=1}^{n'} H^n_{n-i} + \varepsilon^2 2b \cdot \sum_{j=1}^{n'} \sqrt{\frac{2bj}{K}} \\
&\quad \text{(by Lemma 5)} \\
= &b \cdot \sum_{i=1}^{n'} H^n_{n-i} \cdot \left[1 - \varepsilon \left(1 - \frac{2}{K} \right) \right] + \varepsilon^2 2b \cdot \sum_{j=1}^{n'} \sqrt{\frac{2bj}{K}} \\
\leq &b \cdot \sum_{i=1}^{n'} H^n_{n-i} \cdot \left[1 - \varepsilon \left(1 - \frac{2}{K} \right) \right] + \varepsilon^2 2b \cdot n \sqrt{\frac{2bn}{K}}.
\end{aligned}$$

Since $\min\{x, y\} \leq x$ and $\min\{x, y\} \leq y$, we can obtain

$$\mathbb{E}[\mathsf{ALG}]$$

$$
\begin{aligned}
&\leq \sum_{i=1}^{n} \min\Big\{ \sum_{j=1}^{i} \big(\frac{1-\varepsilon}{2}\mathbb{E}[Q_{i,j}] + \varepsilon\cdot\mathbb{E}[Q^*_{i,j}]\big), b\Big\} \\
&\leq \sum_{i=1}^{n'}\sum_{j=1}^{i} \big(\frac{1-\varepsilon}{2}\mathbb{E}[Q_{i,j}] + \varepsilon\cdot\mathbb{E}[Q^*_{i,j}]\big) + \sum_{i=n'+1}^{n} b \\
&\leq b\cdot\sum_{i=1}^{n'} H^n_{n-i}\cdot\left[1-\varepsilon\left(1-\frac{2}{K}\right)\right] + \varepsilon^2 2b\cdot n\sqrt{\frac{2bn}{K}} + \sum_{i=n'+1}^{n} b \\
&= b\cdot\sum_{i=1}^{n}\min(H^n_{n-i},1) - b\cdot\sum_{i=1}^{n'} H^n_{n-i}\cdot\varepsilon\left(1-\frac{2}{K}\right) + \varepsilon^2 2b\cdot n\sqrt{\frac{2bn}{K}}. \qquad (21)
\end{aligned}
$$

The last equality is because $H^n_{n-i}\leq 1$ for all $i\leq n'$.

Make the technical assumptions $b\geq K\geq 3$, and set

$$\varepsilon := \frac{1}{34}\sqrt{\frac{K}{bn}}$$

which satisfies the condition that $\varepsilon\leq 1/2$.

Substituting back into (21), we obtain

$$
\begin{aligned}
&\mathbb{E}[\mathsf{ALG}] \\
&\leq b\cdot\sum_{i=1}^{n}\min(H^n_{n-i},1) - b\cdot\sum_{i=1}^{n'} H^n_{n-i}\cdot\varepsilon\left(1-\frac{2}{K}\right) + \varepsilon^2 2b\cdot n\sqrt{\frac{2bn}{K}} \\
&= b\cdot\sum_{i=1}^{n}\min(H^n_{n-i},1) - \sqrt{\frac{Kb}{n}}\left[\frac{1}{34}\left(1-\frac{2}{K}\right)\sum_{i=1}^{n'} H^n_{n-i} - \frac{\sqrt{2}}{578}n\right] \\
&\leq b\cdot\sum_{i=1}^{n}\min(H^n_{n-i},1) - \sqrt{\frac{Kb}{n}}\left[\frac{1}{34}\left(1-\frac{2}{3}\right)\sum_{i=1}^{n'} H^n_{n-i} - \frac{\sqrt{2}}{578}n\right] \\
&= b\cdot\sum_{i=1}^{n}\min(H^n_{n-i},1) - \sqrt{\frac{Kb}{n}}\left[\frac{1}{102}\sum_{i=1}^{n'} H^n_{n-i} - \frac{\sqrt{2}}{578}n\right]. \qquad (22)
\end{aligned}
$$

To complete the analysis, we need elementary facts about the harmonic sums H^n_{n-i} defined in (20):

Proposition 3.

$$\sum_{i=1}^{n'} H^n_{n-i} \leq n - 2n/e + 2; \qquad (23)$$

$$\sum_{i=n'+1}^{n}\min(H^n_{n-i},1) \leq n/e + 1. \qquad (24)$$

Proof. Since n' was defined to be the largest value such that $H^n_{n-n}\leq 1$, it can be checked that $n' = \lfloor n(1-1/e)\rfloor$. For all $i = 1,\ldots,n'$, $\min(H^n_{n-1},1) = H^n_{n-1}$, while for all $i = n'+1,\ldots,n$, $\min(H^n_{n-1},1) = 1$.

Therefore, the LHS of inequality (24) equals $(n - \lfloor n(1-1/e) \rfloor) \cdot 1$, which is at most $n - (n(1 - 1/e) - 1) = n/e + 1$, which equals the RHS of inequality (24).

For inequality (23), note that its LHS is at most $\sum_{i=1}^{n'} \ln(n/(n-i))$. In turn,

$$\begin{aligned}\sum_{i=1}^{n'} \ln \frac{1}{1-i/n} &\le \int_1^{n'+1} \ln \frac{1}{1-x/n} dx \\ &\le \int_0^{n(1-1/e)+1} \ln \frac{1}{1-x/n} dx \\ &= n \int_0^{1-1/e+1/n} \ln \frac{1}{1-y} dy\end{aligned}$$

where the first inequality uses the fact that the function $\ln \frac{1}{1-x/n}$ is increasing over $x \in [1, n'+1]$. The final integral can be evaluated to equal

$$1 - 1/e + 1/n + (1/e - 1/n) \ln(1/e - 1/n)$$

which is at most $1 - 1/e + 1/n + (1/e - 1/n)(-1) = 1 - 2/e + 2/n$ as long as $n \ge 3$. This completes the proof of inequality (23).

□

Applying Proposition 3 to expression (22) and using the fact that $b - \sqrt{Kb/n}/102 > 0$, we bound expression (22) from above by

$$\begin{aligned}&bn\left(1 - \frac{1}{e} + \frac{3}{n}\right) - \sqrt{\frac{Kb}{n}}\left[\frac{1}{102}\left(1 - \frac{2}{e} + \frac{2}{n}\right)n - \frac{\sqrt{2}}{578}n\right] \\ \le& bn\left(1 - \frac{1}{e}\right) + 3b - \frac{\sqrt{nKb}}{C}.\end{aligned}$$

$C > 1$ is an absolute constant. As long as $b \le n$ and K is sufficiently large, the inequality $3b < \sqrt{nKb}/C$ holds. Since $\mathsf{OPT} = bn$ for all realization of π and $\boldsymbol{\ell}$, this completes the proof of Theorem 2.

6. Extension to Multiple Reward Rates

In this section, we consider the generalization to the setting where each resource i could be depleted (sold) at varying rates (prices), instead of a single rate r_i, following Ma and Simchi-Levi (2017). We assume that for each resource i, its set of reward rates $\mathcal{P}_i$ is known in advance. This introduces an aspect of "admission control" to the problem, where sometimes it is desirable to completely reject a customer, who is only willing to purchase a resource at a low price, to reserve resources for higher-paying customers..

We impose additional structure on the mapping from contexts and actions to distributions over outcomes. We assume that each $\mathcal{P}_i$ is finite and that the action set $\mathcal{A}$ is a non-empty downward-closed set of *combinations* (i, P) of resources i and prices $P \in \mathcal{P}_i$. $\mathcal{A}$ can be thought of as the feasible assortments of (resource, price)-combinations that the firm can offer. For example, actions $a \in \mathcal{A}$ can be constrained so that $|\{(j,P) \in a : j = i\}| \leq 1$ for all i, which says that the firm can set at most one price for each resource, or alternatively constrained only in total cardinality, so that the firm can offer the same resource at multiple prices (where presumably additional benefits are attached with the higher price).

We only allow the firm to offer combinations (i,P)s for which resource i has not ran out. Note that this is in contrast to the model described in Section 2, where actions can be arbitrarily chosen and resources which have ran out are not consumed. Since $\mathcal{A}$ is downward-closed, it always contains the empty assortment $\emptyset$, which the firm can offer if it has ran out of all resources. When the firm offers an assortment a, the outcome is described by a vector $\mathbf{y} \in \{0,1\}^{|\mathcal{P}_1|+\cdots+|\mathcal{P}_n|}$ describing which combinations (i,P) were consumed. Only combinations $(i,P) \in a$ could be consumed, and for each resource i, at most one combination corresponding to i could be consumed.

Assumption 1 **(Substitutability)**. *Consider any context $x \in \mathcal{X}$ and any two actions $a, a' \in \mathcal{A}$ with $a \subseteq a'$. Then for any combination $(i,P) \in a$, we have $\sum_{\mathbf{y}:\mathbf{y}_{(i,P)}=1} \rho_{x,a}(\mathbf{y}) \geq \sum_{\mathbf{y}:\mathbf{y}_{(i,P)}=1} \rho_{x,a'}(\mathbf{y})$.*

Colloquially, Assumption 1 reads that augmenting an assortment (from a to a') can only decrease the chances of selling the combinations already in the assortment. It is a very mild assumption which holds under any *rational* choice model. Assumption 1 is made Golrezaei et al. (2014) and Ma and Simchi-Levi (2017).

We still define OPT as the optimal objective value of the LP relaxation:

Primal:

$$\max \sum_{a \in \mathcal{A}} \sum_{t \in [T]} s_{a,t} \sum_{(i,P) \in a} \sum_{\mathbf{y}:\mathbf{y}_{(i,P)}=1} \rho_{x^t,a}(\mathbf{y}) P \tag{25}$$

$$\sum_{a \in \mathcal{A}} \sum_{t \in [T]} s_{a,t} \sum_{(i,P) \in a} \sum_{\mathbf{y}:\mathbf{y}_{(i,P)}=1} \rho_{x^t,a}(\mathbf{y}) \leq b_i \qquad i \in [n]$$

$$\sum_{a \in \mathcal{A}} s_{a,t} \leq 1 \qquad t \in [T]$$

$$s_{a,t} \geq 0 \qquad a \in \mathcal{A}, t \in [T]$$

Dual:

$$\min \sum_{i \in [n]} b_i \lambda_i + \sum_{t \in [T]} \gamma_t \tag{26}$$

$$\gamma_t \geq \sum_{(i,P) \in a} \sum_{\mathbf{y}:\mathbf{y}_{(i,P)}=1} \rho_{x^t,a}(\mathbf{y})(P - \lambda_i) \qquad a \in \mathcal{A}, t \in [T] \tag{27}$$

$$\lambda_i, \gamma_t \geq 0 \qquad i \in [n], t \in [T]$$

We modify our online resource allocation algorithm from Section 3 for the current setting with multiple reward rates. The only change is in the definition of rewards in the auxiliary online learning problem.

In Section 3, at each point in time t, we defined a virtual reward r_i^t for each resource i, based on the fraction N_i^{t-1}/b_i of that resource depleted at that time. r_i^t was defined by multiplying r_i by a *penalty factor* $(1-\Psi(N_i^{t-1}/b_i))$, where $\Psi(\cdot)$ increased from 0 to 1 as the fraction depleted increased from 0 to 1. Now that resource i has multiple reward rates in $\mathcal{P}_i$, the change from Ma and Simchi-Levi (2017) is that we instead subtract a *virtual cost*. Specifically, for each combination (i,P), its virtual reward at time t is defined to be

$$r_{(i,P)}^t = P - \Phi_{\mathcal{P}_i}\left(\frac{N_i^{t-1}}{b_i}\right), \tag{28}$$

where $\Phi_{\mathcal{P}_i}(\cdot)$ increases from 0 to $\max \mathcal{P}_i$ as the fraction of resource i depleted increases from 0 to 1. Note that it is possible for the virtual reward $r_{(i,P)}^t$ to be negative.

THEOREM 3. *The total reward* ALG *earned by the algorithm that uses virtual costs (28) satisfies*

$$\mathsf{OPT} \le \frac{(1+b_{min})(1-e^{-1/b_{min}})}{1-\exp(-\min_i \alpha_i^{(1)})}\cdot \mathbb{E}[\mathsf{ALG}] + \mathbb{E}[\mathsf{REG}(\mathcal{F}_T)]. \tag{29}$$

Proof. Redefine

$$R^t(a) = \sum_{(i,P)\in a}\sum_{\mathbf{y}:\mathbf{y}_{(i,P)}=1} \rho_{x^t,a}(\mathbf{y})\left[P - \Phi_{\mathcal{P}_i}\left(\frac{N_i^{t-1}}{b_i}\right)\right]$$

as the expected reward of action a in the auxiliary problem. Define the dual variables to LP (26) as

$$\Lambda_i = \Phi_{\mathcal{P}_i}\left(\frac{N_i^T}{b_i}\right), \quad \Gamma_t = R^t(a_*^t).$$

These dual variables can be readily verified to be feasible for LP (26). Based on the strong duality for linear program, we know that

$$\begin{aligned}
\mathsf{OPT} &\le \mathbb{E}\left[\sum_{i\in[n]} b_i\Lambda_i + \sum_{t\in[T]}\Gamma_t\right]\\
&= \mathbb{E}\left[\sum_{i\in[n]} b_i\Phi_{\mathcal{P}_i}\left(\frac{N_i^T}{b_i}\right) + \sum_{t\in[T]} R^t(a^t) + \sum_{t\in[T]}(R^t(a_*^t) - R^t(a^t))\right]\\
&= \mathbb{E}\left[\sum_{i\in[n]} b_i\Phi_{\mathcal{P}_i}\left(\frac{N_i^T}{b_i}\right) + \sum_{t\in[T]} R^t(a^t)] + \mathbb{E}[\mathsf{REG}(\mathcal{F}_T)\right].
\end{aligned}$$

Based on the theory in Ma and Simchi-Levi (2017), it is then a simple exercise to adapt the analysis in Theorem 1 to obtain

$$\mathbb{E}\left[\sum_{i\in[n]} b_i\Phi_{\mathcal{P}_i}\left(\frac{N_i^T}{b_i}\right) + \sum_{t\in[T]} R^t(a^t)\right] \le \frac{(1+b_{\min})(1-e^{-1/b_{\min}})}{1-\exp(-\min_i \alpha_i^{(1)})}\cdot \mathbb{E}[\mathsf{ALG}].$$

□

Table 1 Prices of 8 products from the dataset.

Category	$P_{i,1}$	$P_{i,2}$
King	307	361
Queen	304	361
Suites	384	496
Two-double	306	342

Compared to Theorem 1, the only change in inequality (29) in Theorem 3 is in the denominator, where the denominator $1-e^{-1}$ in Theorem 1 has been replaced by denominator $\min_i(1-e^{-\alpha_i^{(1)}})$ in Theorem 3. For each resource i, $1-e^{-\alpha_i^{(1)}}$ is the *competitive ratio associated with* price set $\mathcal{P}_i$, and the competitive ratio is equal to $1-1/e$ when $\mathcal{P}_i$ is a singleton.

7. Numerical Study

We conduct numerical experiments using dataset Hotel 1 of Bodea et al. (2009). Our numerical setting is a dynamic assortment planning problem, similar to that in Ma and Simchi-Levi (2017), but we consider their extension in which customer purchase probabilities are not observable.

7.1. Simulation Model

In the numerical experiments, we focus on a dynamic assortment planning problem with multiple reward rates (see Section 6). We consider a hotel with $n=4$ room categories: King rooms, Queen rooms, Suites, and Two-double rooms. Each room category is a resource, indexed by $i=1,2,3,4$. The inventory level of each of these resources is the number of available rooms in the corresponding category.

Rooms of each category i can be offered at two prices $\mathcal{P}_i=\{P_{i,1},P_{i,2}\}$, for $i=1,2,3,4$. Each of the $m=8$ combinations, indexed by $j=1,2,...,8$, of room category and price is a product. Table 1 summarizes the prices of all the $m=8$ products from the data set. In the experiments, we double the higher price $P_{i,2}$ of each room category i in order to differentiate the performance of different algorithms.

Each customer has a feature (context) vector $x\in\mathcal{X}\subseteq\mathbb{R}^9$. $x_1=1$ is a constant feature. Features $x_2,...,x_9$ represent the customer's personal information, such as the party size and the VIP level. (See Ma and Simchi-Levi (2017) for a more detailed discussion on feature selection.) Each product $j\in\{1,2,...,8\}$ has a latent vector $\beta_j^*\in\mathbb{R}^9$. We assume that customers follow the MNL choice model. For each customer $x\in\mathcal{X}$, the personalized attraction value of product j is $e^{x^\top\beta_j^*}$. The action set $\mathcal{A}$ consists of all the possible assortments formed by the 8 products. When assortment $a\subseteq\{1,2,...,8\}$ is offered to customer $x\in\mathcal{X}$, the customer will purchase product $j\in a$ with probability

$$\frac{e^{x^\top\beta_j^*}}{v_0+\sum_{j'\in a}e^{x^\top\beta_{j'}^*}},$$

where v_0 is the attraction value for the no-purchase option. We vary v_0 in the experiments.

We consider a Bayesian environment. The prior distribution for each β_j^*, $j \in \{1,2,...,8\}$, is generated as follows. First, calculate the maximum likelihood estimator $\bar{\beta}_j$ for β_j^* from all the transactions in the dataset. Then, we assume that each element $\beta_{j,k}^*$, for $k = 1,2,...,9$, of β_j^* is an independent uniform random variable over $[\bar{\beta}_{j,k} - \epsilon, \bar{\beta}_{j,k} + \epsilon]$. We vary the uncertainty level ϵ in the tests. $\epsilon = 0$ corresponds to the model of Ma and Simchi-Levi (2017), in which the algorithms know the true values of β_j^*.

This numerical setting essentially follows Cheung and Simchi-Levi (2017) except that we impose inventory constraints here. The Thompson sampling algorithm in Cheung and Simchi-Levi (2017) solves the auxiliary problem of this setting.

PROPOSITION 4 **(Cheung and Simchi-Levi (2017))**. *Suppose that* $\beta = (\beta_1, \beta_2, ..., \beta_8)$ *is drawn from a known prior distribution* π_0. *For the auxiliary problem, there is a Thompson sampling algorithm with Bayesian regret*

$$\mathbb{E}_{\beta\sim\pi_0}[\mathsf{REG}(\mathcal{F}_T)] = \mathbb{E}_{\beta\sim\pi_0}[\mathbb{E}[\mathsf{REG}(\mathcal{F}_T)|\beta]] = \tilde{O}(Dm\sqrt{BT}).$$

In our numerical model, $D = 9$ is the length of feature vectors, $m = 8$ is the number of products, $B = 8$ is the maximum size of any assortment, and $T \approx 200$ is the number of customers.

Applying this Thompson sampling algorithm to our framework, and letting $b_{\min} \to \infty$, we can obtain the following performance guarantee by Theorem 3

$$\mathbb{E}_{\beta\sim\pi_0}[\mathsf{OPT}] \le \frac{1}{1 - \exp(-\min_{i\in[n]} \alpha_i^{(1)})} \cdot \mathbb{E}_{\beta\sim\pi_0}[\mathsf{ALG}] + \tilde{O}(Dm\sqrt{BT}).$$

Based on the prices in Table 1, we can easily calculate $1 - \exp(-\min_{i\in[n]} \alpha_i^{(1)}) \approx 0.58$. For details of the calculation, we refer to Ma and Simchi-Levi (2017).

7.2. Numerical Results

For each test case, we simulate 500 replicates and report the average performance of each algorithm. For each replicate, we uniformly draw a sample path of customer arrivals, i.e., a sequence of feature vectors, from 31 different instances constructed in Ma and Simchi-Levi (2017). Each sample path contains about 200 customers. For each replicate, we also randomly draw the latent vectors β_j^* for all products $j \in \{1,2,...,8\}$ from their prior distributions.

We compare the following algorithms

- IB-TS: the inventory-balancing algorithm generated by our framework using the Thompson sampling algorithm in Cheung and Simchi-Levi (2017) as the oracles.

Table 2 Performance of algorithms relative to OPT. $v_0 = 5$, $\epsilon = 1$.

Inventory scale	IB-TS	Gdy-TS	Conserv-TS
0.1	93.6%	90.3%	99.3%
0.15	95.5%	90.7%	98.2%
0.2	95.7%	90.8%	98.0%
0.25	96.1%	91.4%	97.0%
0.3	95.8%	92.1%	96.1%
0.35	95.1%	92.6%	96.0%
0.4	94.2%	92.5%	95.1%
0.45	93.5%	92.8%	94.5%
0.5	93.2%	93.2%	94.2%
0.55	91.5%	92.9%	92.9%
0.6	91.0%	92.9%	93.2%

Table 3 Performance of algorithms relative to OPT. $v_0 = 40$, $\epsilon = 1$.

Inventory scale	IB-TS	Gdy-TS	Conserv-TS
0.1	87.7%	84.1%	92.8%
0.15	90.4%	85.5%	91.4%
0.2	91.8%	87.2%	89.3%
0.25	91.5%	87.5%	88.3%
0.3	92.0%	88.4%	87.8%
0.35	91.6%	89.2%	86.5%
0.4	91.3%	89.0%	86.4%
0.45	91.4%	89.8%	86.1%
0.5	92.8%	90.8%	86.5%
0.55	91.8%	90.1%	86.1%
0.6	92.2%	90.7%	86.7%

- Gdy-TS: same as IB-TS but the framework uses the original reward values, instead of the virtual rewards, as the input for the oracles.
- Conserv-TS: same as IB-TS but the algorithm assumes that there are only 4 higher-price products, i.e., products with prices $P_{\cdot,2}$.

Tables 2 to 6 report the performance of these algorithms under different test parameters. In particular, the first column of each table is a parameter that scales the initial inventory levels of all the four resources. In general, Gdy-TS performs better when inventory is more abundant. This is because the greedy algorithm is the optimal algorithm when there is no need to reserve resources. On the other hand, Conserv-TS has better performance when inventory is more scarce. This is because there is no need to sell resources at lower prices when we can sell all of them. Overall, our IB-TS algorithm performs much better when total inventory is close to total demand.

References

Agrawal, Shipra, Nikhil R. Devanur. 2014. Bandits with concave rewards and convex knapsacks. *Proceedings of the fifteenth ACM conference on Economics and computation - EC '14* 989–1006.

Table 4 Performance of algorithms relative to OPT. $v_0 = 100$, $\epsilon = 1$.

Inventory scale	IB-TS	Gdy-TS	Conserv-TS
0.1	87.1%	86.2%	87.7%
0.15	89.9%	87.6%	86.4%
0.2	90.5%	88.0%	86.2%
0.25	91.9%	90.2%	85.6%
0.3	91.7%	90.2%	84.3%
0.35	91.5%	91.1%	84.3%
0.4	92.1%	90.8%	83.5%
0.45	92.4%	91.5%	84.7%
0.5	93.3%	91.4%	85.1%
0.55	93.2%	92.2%	84.7%
0.6	92.4%	92.6%	84.2%

Table 5 Performance of algorithms relative to OPT. $v_0 = 40$, $\epsilon = 0.01$.

Inventory scale	IB-TS	Gdy-TS	Conserv-TS
0.1	93.3%	91.9%	99.2%
0.15	93.5%	89.9%	97.4%
0.2	92.7%	88.5%	95.3%
0.25	93.2%	89.6%	93.5%
0.3	92.9%	91.1%	92.7%
0.35	94.9%	95.0%	92.4%
0.4	96.4%	95.7%	93.1%
0.45	96.8%	97.4%	93.7%
0.5	98.4%	98.4%	95.2%
0.55	98.3%	99.6%	95.2%
0.6	97.9%	99.0%	95.0%

Table 6 Performance of algorithms relative to OPT. $v_0 = 40$, $\epsilon = 5$.

Inventory scale	IB-TS	Gdy-TS	Conserv-TS
0.1	84.8%	82.0%	91.4%
0.15	87.5%	84.2%	89.9%
0.2	88.5%	83.7%	89.6%
0.25	88.9%	84.3%	88.5%
0.3	89.3%	84.7%	87.7%
0.35	89.4%	86.1%	86.3%
0.4	89.9%	86.2%	86.0%
0.45	89.1%	85.3%	85.4%
0.5	88.9%	85.8%	84.6%
0.55	88.6%	85.3%	84.5%
0.6	88.6%	85.6%	84.7%

Agrawal, Shipra, Nikhil R. Devanur. 2016. *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*. 3450–3458.

Agrawal, Shipra, Nikhil R. Devanur, Lihong Li. 2016. An efficient algorithm for contextual bandits with knapsacks, and an extension to concave objectives. *Proceedings of the 29th Conference on Learning*

Theory, COLT 2016, New York, USA, June 23-26, 2016 4–18.

Babaioff, Moshe, Shaddin Dughmi, Robert Kleinberg, Aleksandrs Slivkins. 2015. Dynamic Pricing with Limited Supply. *ACM Trans. Economics and Comput.* **3**(1) 4:1–4:26.

Badanidiyuru, Ashwinkumar, Robert Kleinberg, Aleksandrs Slivkins. 2013. Bandits with knapsacks. *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*. IEEE, 207–216.

Badanidiyuru, Ashwinkumar, John Langford, Aleksandrs Slivkins. 2014. Resourceful contextual bandits. Maria Florina Balcan, Vitaly Feldman, Csaba Szepesvri, eds., *Proceedings of The 27th Conference on Learning Theory*, *Proceedings of Machine Learning Research*, vol. 35. PMLR, Barcelona, Spain, 1109–1134. URL `http://proceedings.mlr.press/v35/badanidiyuru14.html`.

Ball, Michael O, Maurice Queyranne. 2009. Toward robust revenue management: Competitive analysis of online booking. *Operations Research* **57**(4) 950–963.

Besbes, Omar, Assaf Zeevi. 2009. Dynamic Pricing Without Knowing the Demand Function: Risk Bounds and Near-Optimal Algorithms. *Operations Research* **57**(6) 1407–1420. doi:10.1287/opre.1080.0640. URL `http://pubsonline.informs.org/doi/abs/10.1287/opre.1080.0640`.

Besbes, Omar, Assaf Zeevi. 2012. Blind Network Revenue Management. *Operations Research* **60**(6) 1537–1550. doi:10.1287/opre.1120.1103. URL `http://pubsonline.informs.org/doi/abs/10.1287/opre.1120.1103`.

Bodea, Tudor, Mark Ferguson, Laurie Garrow. 2009. Data setchoice-based revenue management: Data from a major hotel chain. *Manufacturing & Service Operations Management* **11**(2) 356–361.

Borodin, Allan, Ran El-Yaniv. 2005. *Online computation and competitive analysis*. cambridge university press.

Bubeck, Sébastien, Nicolò Cesa-Bianchi. 2012. Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. *Foundations and Trends in Machine Learning* **5**(1) 1–122.

Buchbinder, Niv, Kamal Jain, Joseph Seffi Naor. 2007. Online primal-dual algorithms for maximizing ad-auctions revenue. *European Symposium on Algorithms*. Springer, 253–264.

Chen, Xi, Will Ma, David Simchi-Levi, Linwei Xin. 2016. Dynamic recommendation at checkout under inventory constraint. *manuscript on SSRN* .

Cheung, Wang Chi, David Simchi-Levi. 2017. Thompson sampling for online personalized assortment optimization problems with multinomial logit choice models. *Manuscript* URL `https://ssrn.com/abstract=3075658`.

Devanur, Nikhil R, Kamal Jain. 2012. Online matching with concave returns. *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*. ACM, 137–144.

Devanur, Nikhil R, Kamal Jain, Robert D Kleinberg. 2013. Randomized primal-dual analysis of ranking for online bipartite matching. *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 101–107.

Feldman, Jacob, Nan Liu, Huseyin Topaloglu, Serhan Ziya. 2014. Appointment scheduling under patient preference and no-show behavior. *Operations Research* **62**(4) 794–811.

Ferreira, Kris Johnson, David Simchi-Levi, He Wang. 2016. Online network revenue management using thompson sampling. *Accepted by Operations Research* .

Golrezaei, Negin, Hamid Nazerzadeh, Paat Rusmevichientong. 2014. Real-time optimization of personalized assortments. *Management Science* **60**(6) 1532–1551.

Kalyanasundaram, Bala, Kirk R Pruhs. 2000. An optimal deterministic algorithm for online b-matching. *Theoretical Computer Science* **233**(1) 319–325.

Karp, Richard M, Umesh V Vazirani, Vijay V Vazirani. 1990. An optimal algorithm for on-line bipartite matching. *Proceedings of the twenty-second annual ACM symposium on Theory of computing*. ACM, 352–358.

Kell, Nathaniel, Debmalya Panigrahi. 2016. Online budgeted allocation with general budgets. *Proceedings of the 2016 ACM Conference on Economics and Computation*. ACM, 419–436.

Kleinberg, Robert, Aleksandrs Slivkins, Eli Upfal. 2008. Multi-armed Bandits in Metric Spaces. *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing*. STOC '08, ACM, New York, NY, USA, 681–690. doi:10.1145/1374376.1374475. URL `http://doi.acm.org/10.1145/1374376.1374475`.

Ma, Will, David Simchi-Levi. 2017. Tight weight-dependent competitive ratios for online edge-weighted matching, with application to revenue management .

Mehta, Aranyak. 2013. Online matching and ad allocation. *Foundations and Trends® in Theoretical Computer Science* **8**(4) 265–368.

Mehta, Aranyak, Debmalya Panigrahi. 2012. Online matching with stochastic rewards. *Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on*. IEEE, 728–737.

Mehta, Aranyak, Amin Saberi, Umesh Vazirani, Vijay Vazirani. 2007. Adwords and generalized online matching. *Journal of the ACM (JACM)* **54**(5) 22.

Slivkins, Aleksandrs. 2017. *Introduction to Multi-Armed Bandits*. September. URL `http://slivkins.com/work/MAB-book.pdf`.

Talluri, Kalyan, Garrett van Ryzin. 1998. An analysis of bid-price controls for network revenue management. *Management Science* **44**(11-part-1) 1577–1593.

Truong, Van-Anh. 2015. Optimal advance scheduling. *Management Science* **61**(7) 1584–1597.

Wang, Zizhuo, Shiming Deng, Yinyu Ye. 2014. Close the gaps: A learning-while-doing algorithm for single-product revenue management problems. *Operations Research* **62**(2) 318–331.