Fairness-aware Network Revenue Management with Demand Learning

Xi Chen *
Leonard N. Stern School of Business, New York University, New York, NY 10012, USA, xc13@stern.nyu.edu

Jiameng Lyu
Yau Mathematical Sciences Center, Tsinghua University, Beijing 100084, China, lvjm21@mails.tsinghua.edu.cn

Yining Wang
Naveen Jindal School of Management, University of Texas at Dallas, Richardson, TX 75080, USA, yining.wang@utdallas.edu

Yuan Zhou
Yau Mathematical Sciences Center, Tsinghua University, Beijing 100084, China, yuan-zhou@tsinghua.edu.cn

In addition to maximizing the total revenue, decision-makers in lots of industries would like to guarantee fair consumption across different resources and avoid saturating certain resources. Motivated by these practical needs, this paper studies the price-based network revenue management problem with both demand learning and fairness concern about the consumption across different resources. We introduce the regularized revenue, i.e., the total revenue with a fairness regularization, as our objective to incorporate fairness into the revenue maximization goal. We propose a primal-dual-type online policy with the Upper-Confidence-Bound (UCB) demand learning method to maximize the regularized revenue. We adopt several innovative techniques to make our algorithm a unified and computationally efficient framework for the continuous price set and a wide class of fairness regularizers. Our algorithm achieves a worst-case regret of $O(\sqrt{NT^2 \log T})$, where $N$ denotes the number of products and $T$ denotes the number of time periods. Numerical experiments in a few NRM examples demonstrate the effectiveness of our algorithm for balancing revenue and fairness.

Key words: network revenue management, demand learning, fairness, regret analysis, linear bandit

1. Introduction

Network revenue management, as a fundamental and important model in revenue management, has been successfully applied in lots of industries, such as online retailing, airline, hotel (Talluri et al. 2004, Klein et al. 2020). Classical research on NRM aims to maximize the total revenue over $T$ time periods under resource constraints assuming the demand function is known. See, for example, the seminal works of Gallego and Van Ryzin (1997), Jasin (2014), Maglaras and Meissner (2006).

In practice, there are two main challenges in adopting NRM models. First, the demand function in NRM is usually unknown, which needs to be learned on the fly. Second, in addition to maximizing

* Author names listed in alphabetical order.
the total revenue, many retailers would like to guarantee fair consumption across different resources and avoid saturating certain resource types. As summarized in (Bateni et al. 2022), fairness is an important concern in two cases. First, many platforms (e.g., in healthcare or public utility sectors) are under the obligation to ensure fairness. Second, even those profit-seeking platforms also need to take fairness into consideration, because to maximize the long-term revenue, a platform should keep long-term cooperation with different resource suppliers. Ensuring fairness across different resource suppliers will help keep a good collaboration relationship. Let us provide two examples to illustrate the importance of fairness constraints in revenue management applications:

**Online Retailing:** In an NRM model for the online retailing industry, a retailer decides prices for products and each product consumes several types of resources. Different resources are provided by different suppliers, and it is also possible that one supplier could provide more than one type of resources. If one supplier terminates the cooperation, certain products may no longer be able to be produced. Therefore, to maximize long-term revenue and guarantee no product is out of stock, the retailer should keep good relationship with all resource suppliers. To this end, it is essential to ensure profit fairness across different suppliers. Indeed, only paying attention to the revenue maximization may cause an “unfair” scenario in which some suppliers earn a lot but the shares of others are small.

**Airline Ticketing:** For an airline company, it needs to make decisions on the prices of products (e.g., multileg airlines) subject to resource constraints (i.e., the capacity constraint of each flight), taking into consideration the fact that customers’ realized demand for the products change with prices charged accordingly. Since a multileg airline may involve several flights between different cities, if one leg of flight between two cities is saturated, all the products that include this leg will be out of stock no matter how many seats are available on the other flight legs. As a result, saturation in just a few flights not only will cause the waste of the capacity but also cannot meet the need of the passengers, which in turn leads to passenger dissatisfaction. Therefore, taking the fairness across different resources into consideration will help airline companies avoid saturating flight legs and eventually improve passenger satisfaction.

As shown in the above examples, the fairness objective is often measured in the sense of the whole selling season and cannot be decomposed as an additive objective at each time period. In this paper, we adopt several different metrics that are applied to the average resource consumption vector to measure its fairness. For example, the fairness on resource consumption could be measured by the minimum element of the average resource consumption vector, which is the famous max-min fairness metric extensively studied in economics and resource allocation literature, also included
as a special case of the weighted max-min fairness metric studied in this paper. In some cases it may also be desirable to consume the resources at similar rates, which can be incentivized by our range fairness metric. In addition to the above two examples, we propose several more practically useful fairness metrics. We further identify a wide class of fairness metrics (including all the above examples) with quite mild assumptions. All the fairness metrics in the class can be incorporated in our learning and doing framework. Please refer to Section 2.1 for more details.

The main contribution of this paper is a dynamic pricing algorithm that simultaneously learns the unknown demand function and optimizes the composite objective concerning both the NRM revenue and the fairness metric. For any fairness metric included in the class mentioned above, our algorithm achieves a regret at most $O\left(\frac{N^{5/2}}{T}\right)$ (where $N$ is the number of products and $T$ is the selling horizon, see Theorem 1 for more details). Below we discuss the various technical challenges and our contributions that overcome them in details.

1.1. Main technical challenges

There are several technical challenges to tackle the NRM problem with fairness concern and demand learning. First, the fairness metrics are all applied to the average resource consumption vector, which is a global objective calculated across all time periods. There is also a global unreplenishable inventory constraint for the resources. The global objective and constraint are usually difficult for sequential decision making problems since the decisions at each time step have to be well coordinated to jointly optimize the fairness metric and satisfy the inventory constraint. Second, the problem becomes even harder when the demand function is unknown to the retailer and the retailer has to balance the exploration vs. exploitation trade-off with only learned information or estimated demand. Here, “exploration” means that the retailer needs to explore different prices in order to learn the unknown demand function on the run, and “exploitation” means that the retailer needs to exploit the near-optimal price to simultaneously gain revenue and achieve the global fairness objective. Finally, we aim at designing a computationally efficient algorithm (i.e., that runs in polynomial time) to minimize the regret (i.e., the cost of learning and sequential decision making) about the revenue and the fairness metric.

Compared with the work of Balseiro et al. (2020), the fact that demand curves are unknown in our problem lead to several specific technical hurdles that require novel solutions. In particular, we have the following challenges:

1. In (Balseiro et al. 2020), the dual problem is unconstrained, leading to complex dual space shapes and unbounded penalty vectors. This causes two problems: first, with dual spaces having complex shapes, the dual update steps may not have closed forms or be solved efficiently
for general fairness-induced penalty functions. More importantly, with penalty vectors being unbounded the regret of the problem becomes unbounded as well because of the uncertainty exhibited in the estimated demand functions, rendering bandit learning algorithms impractical.

2. With uncertainty quantified in the estimated demand functions, solving the primal update steps becomes computationally intractable as the objective functions are not necessarily concave any more. Novel algorithms and analysis are required to solve such non-concave problems efficiently and rigorously.

1.2. Our contributions

We are able to address the above challenges with several technical innovations. At a higher level, our algorithm combines the primal-dual-type online policy proposed by Balseiro et al. (2020) with the Upper-Confidence-Bound (UCB) method, where the former solves the fairness-aware online allocation problem with the known demand function and the latter is a widely adopted principle to balance the exploration and exploitation trade-off in many online learning algorithms. However, this combination is not black-box style and we make novel technical contributions in the design and analysis of our algorithm. Moreover, we also introduce new algorithmic ingredients to make sure that our algorithm is computationally efficient. In contrast, most of the online learning algorithms for linear demand models (and even the linear bandit algorithm, LinUCB (Abbasi-Yadkori et al. 2011)) do not guarantee a polynomial time complexity. Below, we describe our main technical contributions in more details.

First, instead of using the UCB of the objective function in the usual online learning and decision making algorithms, we make decisions according to the UCB of the specially designed adjusted reward function. While our adjusted reward function involves the revenue, it does not directly include the fairness metric (as it is not obvious how to decompose such a global metric to each individual time period, as discussed previously). Instead, we include a carefully designed term in the adjusted reward function to relate it to the dual variable. This term, together with our update rule of the dual variable, helps optimize the fairness metric in a global fashion and reflects the inventory constraint at the same time.

Second, to control the estimation error of the adjusted reward during learning, we need to design a bounded domain for the dual parameter (i.e., the dual space). In contrast, Balseiro et al. (2020) adopt an unbounded dual space which is unfriendly to the analysis of our learning process, as illustrated in the first bullet point in the previous section. By adopting a bounded domain for
the dual parameter, we able able to upper bound the estimation errors as well, leading to correct regret scalings. An additional benefit of our new dual space is that due to its simpler shape, we are able to employ a closed-form dual update rule (Algorithm 2, Section 5) for any fairness metric. In contrast, Balseiro et al. (2020) may only achieve this for selected fairness metrics and the dual update in our algorithm is much simplified.

Third, the regret analysis (especially the analysis related to the dual variable) greatly relies on the magnitude of our model parameter estimations. While the natural (regularized) least-squares estimator may not provide the desired bound, we employ an additional convex program $M_t$ to compute a set of bounded model estimates. This $M_t$ program is also helpful to guarantee the concavity the estimated (adjusted) revenue function so that we may computationally efficiently find its maximizer which crucially connects to the decision we will make at each time step.

Finally, to achieve computational efficiency, we innovatively adopt an $\ell_\infty$-norm confidence radius instead of the usual $\ell_2$-norm confidence radius when computing the Upper Confidence Bounds so that we are able to maximize the UCB of the adjusted revenue function in polynomial time. Also, by reducing the $M_t$ program to a linear program with infinitely many constraints, we design a polynomial-time separation oracle and invoke the Ellipsoid method to efficiently solve the $M_t$ program. Both ingredients help our algorithm to achieve the polynomial time complexity that addresses the second bullet point of technical challenges mentioned in the previous section.

For the first 3 technical contributions, we provide more concrete explanations at the end of Section 3, after the introduction of notations and the algorithm description. For the last item, please refer to Section 4.1 and Section 4.2 for more details.

1.3. Related Works

In the section, we introduce three streams of literature related to our paper: network revenue management (NRM) with known demand function, revenue management (RM) with demand learning, and fairness in operations management. And we discuss how our paper is appropriately placed into contemporary literature by giving comparisons with closely-related existing works.

NRM with known demand function. A large body of the price-based network revenue management literature focuses on the case in which the seller knows the underlying demand function in advance. And it is known that the optimal pricing policy of this case can be computed using dynamic programming (DP). However, the well-known curse of dimensionality of DP makes the optimal pricing policy computationally intractable. As a result, many works in the literature have investigated on developing algorithms that are computationally efficient with a superior revenue
performance. The seminal work by Gallego and Van Ryzin (1994, 1997) proposed a simple but powerful heuristics. Specifically, they solve the optimal price of the fluid approximation model which is a deterministic analog of the DP and choose a static price every time. And their approach achieves an $O(\sqrt{T})$ regret. Jasin (2014) introduced an improvement to the static pricing policy by resolving the static price periodically according to the remaining inventory, and attained $O(\log T)$ regret bound. Recently, Wang and Wang (2022) proved that the resolving heuristics can achieve $O(1)$ regret as compared to the optimal policy of the DP.

**RM with demand learning.** There is a large body of literature focusing on the price-based revenue management with demand learning, which are either without inventory constraints (see, e.g., (Den Boer 2014, Den Boer and Zwart 2014, Keskin and Zeevi 2014, 2017, Bu et al. 2022) and references therein) or with inventory constraints (see, e.g., (Besbes and Zeevi 2009, Wang et al. 2014, Chen et al. 2014, Ferreira et al. 2018, Miao et al. 2021)). For dynamic pricing problems without inventory constraints, we refer the readers to (Den Boer 2015) for a detailed review. For price-based revenue management problems with inventory constraints, there are two streams of literature, either considering the nonparametric demand model (Wang et al. 2014, Chen and Shi 2019, Chen et al. 2019, Miao and Wang 2021) or the parametric demand model (see discussion below). Since our paper considers a parametric demand function, we mainly investigate the literature on the revenue management problem with inventory constraints and the parametric model. There are three main approaches for tackling the learning-while-doing challenge.

The first approach is using the Explore-Then-Commit strategy, which separates the exploration phase and exploitation phases. This simple strategy has been widely used in online learning tasks, and Besbes and Zeevi (2009, 2012) and Chen et al. (2014) applied this strategy to the NRM problem and Chen et al. (2014) achieved $O(\sqrt{T})$ regret assuming the strong concavity of the revenue function.

The second approach is using Thompson sampling to address the exploration-exploitation trade-off. Ferreira et al. (2018) introduced Thompson sampling into network revenue management and considered both discrete price model and continuous price set with the linear demand model. They obtained a Bayesian regret $\tilde{O}(\sqrt{T})$ instead of the worst-case regret. The most important step in their algorithm for the continuous price set is to solve a quadratic program, which is not guaranteed to be a convex problem and not clear how to be solved efficiently.

The third approach is incorporating the Optimism in the Face of Uncertain principle into the primal-dual optimization framework. This approach is closely related to the Bandit-with-Knapsack (BwK) model (Badanidiyuru et al. 2013), which introduces global resource constraints into the
multi-armed bandit. Agrawal and Devanur (2019) further generalized BwK to bandit with global convex constraints and concave objective. The work by Agrawal and Devanur (2016), which considered BwK in the linear bandit setting, can be applied to the NRM problem with the discrete price. However, in the continuous price setting, the regret and the running time will exponentially dependent on the number of products due to the discretization procedure. Miao et al. (2021) considered the NRM problem with continuous price and generalized linear model. To tackle the high computational complexity due to the continuous price set, they designed a UCB solver to reduce the original optimization problem to the price optimization problem of an ordinary NRM problem by randomly sampling a vector on the unit sphere and using it to linearize the $\ell_2$-norm-based UCB term. However, the price optimization problem might still be non-convex and difficult to solve despite this reduction.

Our work is closely related to (Agrawal and Devanur 2016) and (Miao et al. 2021). However, there are several significant differences. First, the primal-dual framework in (Agrawal and Devanur 2016) and (Miao et al. 2021) do not consider the fairness regularizer and their algorithms do not directly work in our setting. With the fairness concerns, our algorithm adopts very different primal and dual updates, which requires a different analysis. Second, as compared to the random sampling method in (Miao et al. 2021), we introduce the $\ell_\infty$-norm-based UCB term (Section 4.2) which is not only simpler to calculate, but also only sacrifices an $O(\sqrt{N})$ factor in the regret. In contrast, even without the fairness consideration, the regret of (Miao et al. 2021) for the NRM problem is $\tilde{O}(N^{4.5} \sqrt{T})$, about $N$ times our regret. Third, we introduce a feasibility program $\mathcal{M}_t$ (Section 4.1) to make sure the estimated revenue function is concave and computationally easy to optimize. We are able to combine the above new techniques to derive a computationally efficient low-regret learning-while-doing algorithm for the NRM problem with fairness concerns.

**Fairness in operations.** With the development of data-driven algorithms in machine learning and operations, there is a growing concern about discrimination and unfairness. As a result, the fairness issue has been well studied in the machine learning literature, e.g., fair classification (Dwork et al. 2012, Agarwal et al. 2018, Jang et al. 2022) and fair online learning (Joseph et al. 2016, Liu et al. 2017, Jabbari et al. 2017, Gupta and Kamble 2021, Kandasamy et al. 2020, Baek and Farias 2021). We refer the interested readers to the surveys of fairness in machine learning (Corbett-Davies and Goel 2018, Mehrabi et al. 2021, Hutchinson and Mitchell 2019). And we mainly focus on the fairness issue in operations problem, which has also attracted a lot of attention (Bonald et al. 2006, Chen and Wang 2018, Ma et al. 2020, Kallus and Zhou 2021, Kallus et al. 2022, Zhang et al. 2022).
Static resource allocation is one of earlier fields in operations that takes different fairness metrics into consideration (Bansal and Sviridenko 2006, Bertsimas et al. 2011, Elzayn et al. 2019, Donahue and Kleinberg 2020, Cai et al. 2021, Bateni et al. 2022). Among many fairness metrics, a widely-acknowledged fairness notion, max-min fairness (Nash Jr 1950) that maximizes the minimum resource allocation, has been widely applied (Bansal and Sviridenko 2006, Bertsimas et al. 2011). The fairness notion considered in our paper includes the max-min fairness as a special case and (please refer to Section 2.1 for more details). Another fairness metric that has been well studied is proportional fairness (Azar et al. 2010, Vlasiou et al. 2014, Bateni et al. 2022), which maximizes the overall utility of rate allocations via a logarithmic utility function.

There is a vast body of literature considering the fairness in online allocation problem with inventory constraints (with known demand models) (Elzayn et al. 2019, Ma et al. 2020, Balseiro et al. 2020, Chen et al. 2021a), where the decision-maker must take an action upon each arriving request and generates a reward and the consumption of resources. The related works by Balseiro et al. (2020), Chen et al. (2021a) considered both fairness and revenue management. The policy in the work by Chen et al. (2021a) achieves a \(O(\sqrt{T})\) bound for the cumulative unfairness and a bounded revenue regret. Balseiro et al. (2020) considered revenue maximization and fairness resource allocation simultaneously by introducing the regularized online allocation problem. Since our work is most related to (Balseiro et al. 2020), we have thoroughly discussed the technical differences in the introduction section above and we will present the comparison more concretely in Section 3. We also note that (Balseiro et al. 2020) works for the fairness-aware NRM problem in a quantity-based setting, where the decision-maker must irrevocably accept or reject each arriving request given limited resources (a special case of the online allocation problem studied in their paper). In contrast, we study the NRM problem in the price-based setting where the decision-maker has to decide the prices that influence the demand and the demand has always to be met (as long as permitted by the resource constraints). To the best of our knowledge, our work is the first to consider the fairness objective in the price-based NRM problem.

Another related line of works study the fairness in dynamic pricing (Cohen et al. 2022, 2021, Chen et al. 2021b, Li and Jain 2016, Kallus and Zhou 2021). These works focus on the price fairness, ensuring the prices are similar for different groups and stable over time.

1.4. Notations
The vectors throughout this paper are all column vectors. We denote the set \(\{1, 2, \ldots, N\}\) by \([N]\) for any \(N \in \mathbb{N}\). For vectors \(a, b \in \mathbb{R}^N\), we use \(a \preceq b\) (\(a \succeq b\) respectively) to denote \(a_i \leq b_i\) (\(a_i \geq b_i\) respectively) for all \(i \in [N]\). We use \([x_i]_i\) to denote the \(i\)-th element of the vector \(x_i\). For \(x \in \mathbb{R}^N\) and
\( \Lambda \in \mathbb{R}^{N \times N} \), we define the following norms: 
\( \|x\|_1 := \sum_{i=1}^{N} |x_i| \), 
\( \|x\|_2 := (\sum_{i=1}^{N} x_i^2)^{1/2} \), 
\( \|x\|_{\infty} := \max_{i \in [N]} |x_i| \), 
and 
\( \|x\|_{\Lambda} := \sqrt{x^T \Lambda x} \).

We use \( I_N \) to denote the identity matrix of order \( N \). For \( A \in \mathbb{R}^{M \times N} \), we define the following matrix norms: 
\( \|A\|_{F} := \sum_{i,j} A_{ij}^2 \), 
\( \|A\|_2 := \sup_{\|x\|_2=1} \|Ax\|_2 \), 
\( \|A\|_{\infty} := \sup_{\|x\|_\infty=1} \|Ax\|_\infty \) and it is easy to obtain 
\( \|A\|_{\infty} = \max_{i \in [M]} \sum_{j=1}^{N} |A_{ij}| \). For square matrix \( A \in \mathbb{R}^{N \times N} \), we use \( \lambda_{\max}(A) \) to denote the largest eigenvalue of \( A \). For a symmetric matrix \( A \in \mathbb{R}^{N \times N} \), we use \( A \leq 0 \) to represent that \( A \) is negative semi-definite. We use \( [B|\alpha] \) to denote the augmented matrix by adding \( \alpha \in \mathbb{R}^N \) to the right of the matrix \( B \) as a new column.

We use the big-O notation \( f(T) = \mathcal{O}(g(T)) \) to denote that \( \limsup_{T \to \infty} f(T)/g(T) \leq +\infty \). We use \( \tilde{\mathcal{O}}(\cdot) \) to further omit the logarithmic dependency on \( N, M, \) and \( T \).

### 1.5. Organization

The remainder of this paper is organized as follows. In Section 2, we formulate our problem by introducing the model assumptions and performance measure; we also give plenty of examples of the fairness regularizers to illustrate the potential guidance our paper might bring to the practical scenarios. In Section 3, we present our algorithm and discuss the high-level ideas of the algorithm design. Then we discuss reward and demand estimation (Section 4) and the design of the mirror descent solver (Section 5) in details, which are two key building blocks of our algorithm. In Section 6, we present and prove the main theorem that upper bounds the regret of our algorithm. To demonstrate the empirical performance of our policy, we conduct several numerical experiments and present the results in Section 7. In the end, we give a summary of our paper in Section 8. The proofs of most technical lemmas and the additional experimental results are included in the supplementary materials.

### 2. Model Description and Assumptions

In an NRM model with \( N \) types of products and \( M \) types of resources, a retailer sells \( N \) types of products during a selling season with \( T \) time periods. Each product is defined as a combination of \( M \) types of unreplenishable resources by the consumption matrix \( A \in \mathbb{R}^{M \times N} \), where \( A_{ij} \) means that selling one unit of the type-\( i \) product consumes \( A_{ij} \) unit of the type-\( j \) resource. At each time period \( t \), the retailer must determine the prices for the \( N \) products, i.e., the price vector \( p_t \in \mathbb{R}_+^N \). The retailer then observes the consumer’s demand vector \( d_t \in \mathbb{R}_+^N \) which is realized from an unknown underlying demand function \( D(p_t) \), and finally consumes the resources according to the consumption matrix \( A \). The retailer needs to choose the prices during the selling season to accomplish the following 3 goals:
1. to gradually learn the underline demand function $D(p_t)$ from the observed demands,

2. to maximize the total revenue based on the learned information and given the un replenishable resource inventory,

3. to balance the consumption of the different types of resources via maximizing the fairness regularizer $\phi(\cdot)$, which will be defined soon.

More specifically, the initial inventory levels of the $M$ resources are $I_0 = (I_{0,1}, \ldots, I_{0,M})^\top \in \mathbb{R}_+^M$. At the end of time $t$, the inventory levels become $I_t = I_{t-1} - Ad_t$ for $t = 1, 2, 3, \ldots$ For convenience we also define the normalized inventory level $\gamma = (\gamma_1, \ldots, \gamma_M)^\top := I_0/T$, which is the average amount of resources that can be used at a time period.

For simplicity, we assume that the price range for each product is $[\underline{p}_i, \bar{p}_i]$ and the retailer has to choose $p_t$ in the price set $\mathcal{P}$ at each time $t$. The price set $\mathcal{P}$ can either be $[\underline{p}_i, \bar{p}_i]^N$ or a discrete subset in $[\underline{p}_i, \bar{p}_i]^N$. For brevity, we focus on the case $\mathcal{P} = [\underline{p}_i, \bar{p}_i]^N$, which is much more challenging. One can easily adapt our algorithm and analysis to the discrete price set.

The realized demand $d_t$ is a random variable centered at $D(p_t)$, i.e.,

$$d_t = D(p_t) + \varepsilon_t$$

where $\varepsilon_t$ is a zero-mean noise variable (see Assumption 2 for the more precise statement). We consider the linear demand function (which is the most commonly analyzed demand model in literature, e.g., (Keskin and Zeevi 2014) and (Ferreira et al. 2018))

$$D(p_t) = \alpha + Bp_t,$$

where $\alpha \in \mathbb{R}^N$ and $B \in \mathbb{R}^{N \times N}$ are the model parameters unknown to the retailer. For convenience, we also denote these unknown parameters by $\theta = (\alpha, B) \in \Theta \subseteq \mathbb{R}^{N^2 + N}$, where $\Theta$ is the parameter space.

The revenue collected by the retailer at time $t$ is $r_t = \langle d_t, p_t \rangle$. We also denote the corresponding expected revenue by

$$r(p_t) := \mathbb{E}[r_t|p_t] = \langle p_t, D(p_t) \rangle.$$

The objective of the retailer is to design a policy $\pi = (\pi_1, \ldots, \pi_T)$ with $\pi_t : \mathcal{H}_t \mapsto p_t$ (where $\mathcal{H}_t = \{p_s, d_s\}_{s<t}$ is the historical prices and demands before time $t$) to satisfy the inventory constraint
$I_t \geq 0$ for all $t \in \{1, 2, \ldots, T\}$ and maximize the following expected total revenue plus the fairness regularizer on resource consumption:

$$
\mathbb{E} \left[ \sum_{t=1}^{T} r_t(p_t) + T \phi \left( \frac{1}{T} \sum_{t=1}^{T} A_t \right) \right].
$$

Below in Section 2.1 we will discuss more about the fairness regularizer $\phi(\cdot)$; in Section 2.2 we introduce some standard assumptions on the linear demand model and define the regret that our online policy aims to minimize.

### 2.1. Fairness Regularizer $\phi(\cdot)$: Assumptions and Examples

The fairness on resource consumption is measured by the regularizer $\phi \left( \frac{1}{T} \sum_{t=1}^{T} A_t \right)$ (i.e., the regularizer function $\phi(\cdot)$ applied to the average resource consumption vector).

**Assumption 1.** Throughout this paper we impose the following assumptions on $\phi(\cdot)$.

1. $\phi(s)$ is $L$-Lipschitz continuous with respect to the $\| \cdot \|_\infty$-norm on its effective domain, i.e.,
   $$|\phi(s_1) - \phi(s_2)| \leq L \| s_1 - s_2 \|_\infty \text{ for any } s_1, s_2 \leq \gamma.$$

2. There exists $\overline{\phi}$ such that $0 \leq \phi(s) \leq \overline{\phi}$ for all $0 \leq s \leq \gamma$.

3. $\phi(s)$ is concave.

In the following, we present several regularizers satisfying the above assumptions as examples. We will use $s = \frac{1}{T} \sum_{t=1}^{T} A_t$ and $s_i$ refers to the average consumption of the type-$i$ resource.

**Example 1: Weighted Max-min Fairness Regularizer.** The first example is rooted in the famous max-min fairness guarantee, which has been well studied in the literature on static resources allocation (Bansal and Sviridenko 2006, Bertsimas et al. 2011). The idea behind the max-min fairness guarantee is to promote fairness by maximizing the minimum resource allocation. In our paper, we consider the following weighted max-min fairness regularizer to promote fairness in resource consumption. It is worthy to note that the max-min regularizer in (Balseiro et al. 2020) can be seen as a special case of our weighted max-min regularizer by setting the parameters correspondingly.

Formally, we define the weighted max-min fairness regularizer as $\phi(s) := \lambda \min_i (w_i s_i)$, where $\lambda$ is the parameter to balance between the total revenue goal and the fairness objective, and in the online retailing setting the parameter $w_i$ could be selected as the revenue of the resource supplier due to the consumption of one unit type $i$ resource.
Example 2: Group Max-min Fairness Regularizer. We may divide the different types of resources into groups and only focus on promoting the minimum consumption of each resource group. In practice, each supplier may provide several types of resources (which naturally forms a group) and the group max-min fairness would be useful if we wish to guarantee fairness among the suppliers.

Formally, we define the group max-min fairness regularizer as $\phi(s) := \lambda \min_i((U\tilde{s})_i)$, where $\tilde{s} = (w_1s_1, \cdots, w_ms_m)^T$, $w_i$ is similarly defined as in Example 1, and $U \in \mathbb{R}^{K \times M}$ is a 0-1 matrix describing the grouping scheme. In particular, we require that in each column there is exactly 1 non-zero element and in each row there is at least 1 non-zero element, where a simple example of $U$ is as follows,

$$U = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}.$$  

In this example, there are two resource suppliers. The first supplier provides the type-1 and the type-3 resources and the second supplier provides the type-2 and the type-4 resources.

Example 3: Range Fairness Regularizer. Range is a fundamental statistical quantity that measures the difference between the highest and the lowest value of a population. The Range fairness regularizer provides incentive to minimize the range among the entries of the weighted average consumption vector $\tilde{s} = (w_1s_1, \cdots, w_ms_m)^T$.

Formally, we define the range fairness regularizer as $\phi(s) := \lambda (\min_i(w_is_i) - \max_i(w_is_i) + \max_i(w_i\gamma_i))$, where $-\left[\min_i(w_is_i) - \max_i(w_is_i)\right]$ is the range of $\tilde{s}$ and $\max_i(w_i\gamma_i)$ is introduced to guarantee the positiveness of the regularizer. When $w_i$ is chosen to be per-unit revenue of the type-i resource supplier, the range fairness regularizer can be applied to promote the revenue fairness across different suppliers; when $w_i = 1/\gamma_i$, this regularizer can evaluate the evenness of resource availability and help to avoid the pre-mature saturation of a few resource types.

Example 4: Load Balancing Regularizer. We finally present the load-balancing regularizer proposed in Balseiro et al. (2020). The regularizer is defined as $\phi(s) := \lambda (\min_i((\gamma_i - s_i)/\gamma_i))$, which measures the minimum relative resource availability, and also helps to make sure that no resource is too demanded.

1In general, we may combine the grouping operation with any fairness regularizer satisfying the Assumption 1 to obtain a group version of the fairness regularizer, but for simplicity, we only present the group version of the weighted max-min regularizer here.
2.2. Model Assumptions and Performance Measure

**Assumption 2.** Throughout this paper we impose the following assumptions on the demand model \( d_t = \alpha + B p_t + \varepsilon_t \):

1. The noise \( \{\varepsilon_t\}_{t=1}^T \) is a martingale difference sequence adapted to the filtration \( \{\mathcal{F}_t\}_{t=1}^T \) where \( \mathcal{F}_t = \{p_1,d_1,\cdots,p_t,d_t,p_{t+1}\} \), i.e., \( \mathbb{E}[\varepsilon_t|\mathcal{F}_{t-1}] = 0 \).

2. There exists \( \bar{d} \) such that \( d_t \leq \bar{d} \) almost surely for all \( t \in \{1,2,\ldots,T\} \).

3. The underlying true parameter \( B \) in the linear demand model is negative definite;
   \footnote{\( B \) is negative definite (not necessarily symmetric) if for any \( z \in \mathbb{R}^n \), it holds that \( z^T B z < 0 \).} there exists \( L_B \geq 1 \) such that \( \sqrt{\alpha_t^2 + \|B^T e_i\|_2^2} \leq L_B \) for every \( i \in [N] \), where \( e_i \) is the \( i \)-th unit vector and \( B^T e_i \) is the \( i \)-th row of \( B \).

All items in Assumption 2 are quite standard in literature. The third item is usually seen in papers focusing on the linear demand model (see, e.g., Keskin and Zeevi 2014, Ferreira et al. 2018, Bu et al. 2022)). By the definition \( r(p_t) = \langle d_t,p_t \rangle \) and according to Assumption 2, we may upper bound \( r(p_t) \) by \( \bar{\tau} := N\bar{p}\bar{d} \).

We now discuss the performance measure for the retailer’s policy. We would like to compare the objective value achieved by the retailer with the optimal offline policy, i.e., the one that knows all the model parameters \( \theta \):

\[
J_{\text{opt}} := \max_{\pi = (\pi_1,\ldots,\pi_T)} \mathbb{E} \left[ \sum_{t=1}^T r(p_t) + T \phi \left( \frac{1}{T} \sum_{t=1}^T A d_t \right) \right] \quad \text{s.t.} \quad \sum_{t=1}^T A d_t \leq T\gamma \quad \text{a.s.} \quad (1)
\]

\( J_{\text{opt}} \) upper bounds the objective value achieved by any online policy (i.e., the one without access to \( \theta \)). In light of this, we define the regret of a policy \( \pi \) up to time horizon \( T \) as

\[
\mathcal{R}(T) := J_{\text{opt}} - \mathbb{E} \left[ \sum_{t=1}^T r(p_t) + T \phi \left( \frac{1}{T} \sum_{t=1}^T A d_t \right) \right]. \quad (2)
\]

Note that solving the exact value of \( J_{\text{opt}} \) is quite complicated due to the stochastic nature and adaptivity available to choose \( \pi_1,\pi_2,\ldots,\pi_T \) in sequence. We now introduce the following fluid model which is the deterministic and non-adaptive analogue of \( J_{\text{opt}} \) and easier to analyze.

\[
J_D := \max_{p_1,\ldots,p_T \in [p,p]} \left[ \sum_{t=1}^T r(p_t) + T \phi \left( \frac{1}{T} \sum_{t=1}^T A D(p_t) \right) \right] \quad \text{s.t.} \quad \sum_{t=1}^T A D(p_t) \leq T\gamma. \quad (3)
\]
We assert that there exists an optimal solution \( \{p^*_1, p^*_2, \ldots, p^*_T\} \) to \( J_D \) such that \( p^*_1 = p^*_2 = \cdots = p^*_T = p^* \), since otherwise, we can set \( p_t = p' = \frac{1}{T} \sum_{t=1}^{T} p^*_t \) for every \( t \), and the objective value of \( \{p'_1, p'_2, \ldots, p'_T\} \) becomes no smaller due to the concavity of \( r(p) \) (since \( r(p) = \langle p, \alpha + Bp \rangle \) and \( B \) is negative definite by Assumption 2) and Jenson’s inequality. Therefore we have the following equivalent definition of \( J_D \).

\[
J_D = \max_{p \in [p, \bar{p}]^N} \left[ Tr(p) + T \phi(AD(p)) \right] \quad \text{s.t. } AD(p) \leq \gamma. \tag{4}
\]

We denote \( p^* \) by the optimal solution to Eq. (4).

The following proposition shows that the fluid model \( J_D \) is an upper bound of \( J_{opt} \), and its proof is deferred to the supplementary materials.

**PROPOSITION 1.** \( J_{opt} \leq J_D \).

By Proposition 1, we upper bound the regret of any policy \( \pi \) as follows, which will serve as the starting point of the analysis of our proposed policy.

\[
R(T) \leq T [r(p^*) + \phi(AD(p^*))] - \mathbb{E} \left[ \sum_{t=1}^{T} r(p_t) + T \phi \left( \frac{1}{T} \sum_{t=1}^{T} Ad_t \right) \right]. \tag{5}
\]

### 3. Primal-dual Type Algorithm with Demand Learning

The pseudo-code of our main algorithm is given in Algorithm 1. To better introduce our algorithm design, we first imagine that the demand function were known and only explain the primal-dual framework. We then add the learning component for the demand function and address the additional challenges raised due to the unknown demand.

Given the demand function \( D(\cdot) \), thanks to Proposition 1, we may use the fluid model \( J_D \), which upper bounds \( J_{opt} \), to calculate an upper estimation of the regret of any online policy.\(^3\) We now focus on the primal formulation of the fluid model, and rewrite it with an auxiliary variable \( s \).

\[
p^* := \max_{p \in [p, \bar{p}]^N} \left\{ r(p) + \phi(AD(p)) \quad \text{s.t. } AD(p) \leq \gamma \right\} \tag{11}
\]

\[
= \max_{p \in [p, \bar{p}]^N, -\gamma \leq s \leq \gamma} \left\{ r(p) + \phi(AD(p)) \quad \text{s.t. } AD(p) = s \right\}.
\]

Note that we deliberately impose a lower bound constraint \( s \geq -\gamma \) in the new formulation. This does not change the optimal value of the program since \( AD(p) \) is entry-wise non-negative for non-negative \( A \) and \( D(p) \).

\(^3\) Indeed, this relaxation would not ruin our aimed \( \tilde{O}(\sqrt{T}) \) regret, as the difference between \( J_D \) and \( J_{opt} \) is also \( \tilde{O}(\sqrt{T}) \) for the network revenue management problem either without (Gallego and Van Ryzin 1997) or with the fairness concern (as we will see later in this paper).
Algorithm 1 Primal-dual + UCB for fairness-aware NRM with demand learning

1: Initialize the dual variable $\mu_1 = (0, \ldots, 0) \in \mathbb{R}^M$.

2: **for** $t = 1, 2, \ldots, T$ **do**

3: Compute the regularized least-squares estimator

\[
(\hat{\alpha}_t, \hat{\beta}_t) \in \text{arg min}_{(\alpha, \beta)} \left\{ (N + 1)(\|\alpha\|_2^2 + \|\beta\|_F^2) + \sum_{s \leq t} \|d_s - (\alpha + Bp_s)\|_2^2 \right\}. \tag{6}
\]

4: Find $(\bar{\alpha}_t, \bar{\beta}_t)$ around $(\hat{\alpha}_t, \hat{\beta}_t)$ such that $\bar{\beta}_t$ is negative semi-definite by solving the program

\[
(\bar{\alpha}_t, \bar{\beta}_t) \in \mathcal{M}_t := \left\{ (\alpha, \beta) : \| (\bar{\beta} - \bar{\beta}_t) \|_{\Lambda_t} \leq \kappa, \| \bar{\beta} \|_2 \leq 2L_B \forall i \in [N] \text{ and } \bar{\beta} + \bar{\beta}_t^\top \leq 0 \right\}, \tag{7}
\]

where we define the block matrices $\bar{\beta} := \hat{\beta}|_{[\hat{\alpha}]}$ and $\hat{\beta} := \hat{\beta}_t|_{[\hat{\alpha}_t]}$ ($e_i$ is the $i$-th canonical basis vector), $\Lambda_t$ and $\kappa$ are defined in Eqs. (20, 21) respectively. In the rare case when Eq. (7) is infeasible, we arbitrarily choose $(\bar{\alpha}_t, \bar{\beta}_t)$ as long as $\bar{\beta}_t$ is negative semi-definite and $\|\bar{\beta}_t e_i\|_2 \leq 2L_B \forall i \in [N]$ (e.g., set $\bar{\beta}_t = 0$).

5: Obtain $\tilde{D}_t(p) = \bar{\alpha}_t + \bar{B}_t p$ and $\tilde{r}_t(p) = \langle p, \tilde{D}_t(p) \rangle$ to estimate $D(p)$ and $r(p)$ respectively.

6: Update the primal variables

\[
p_t \in \text{arg max}_{p \in [\underline{p}, \overline{p}]} \left\{ \tilde{r}_t(p) - \mu_t^\top A \tilde{D}_t(p) + 2\Delta_t'(p) \right\}, \quad s_t \in \text{arg max}_{-\gamma \leq s \leq \gamma} \left\{ \phi(s) + \mu_t^\top s \right\}, \tag{8}
\]

where $2\Delta_t'(p)$ is the confidence radius of the adjusted reward estimation $[\tilde{r}_t(p) - \mu_t^\top A \tilde{D}_t(p)]$ which will be constructed later in Eq. (23).

7: Charge the price $p_t$, observe the demand $d_t$, consume resources $Ad_t$ and update the inventory level $I_t = I_{t-1} - Ad_t$ (the algorithm stops whenever any resource is depleted).

8: Obtain an estimated subgradient of dual function $q(\cdot)$ at $\mu_t$: $\tilde{g}_t = -A \tilde{D}_t(p_t) + s_t$.

9: Update the dual variable by invoking the mirror descent solver $\zeta^D$ (Definition 1):

\[
\mu_{t+1} = \zeta^D(\mu_t, \tilde{g}_t; \mathcal{D}, \eta), \tag{9}
\]

where we set $\eta = \sqrt{\frac{C_1}{C_2}}$ ($C_1$ and $C_2$ are also defined in Definition 1) and

\[
\mathcal{D} := \{ \mu \in \mathbb{R}^M \mid \| \mu \|_1 \leq C \}, \quad C := L + ((\tau + \phi)/\gamma), \quad \gamma := \min_{i \in [M]} \gamma_i. \tag{10}
\]

10: **end for**

We then transform primal formulation to an unconstrained optimization problem using the Lagrangian dual method. By the well-known weak duality, we have that

\[
p^* = \max_{p \in [\underline{p}, \overline{p}], \gamma \leq s \leq \gamma} \min_{\mu \in \mathbb{R}^M} \left\{ r(p) + \phi(s) - \mu^\top AD(p) + \mu^\top s \right\}
\]
\[
\min \mu \mathbb{P} \max_{p \in [\mathbb{P}]^N, -\gamma \leq s \leq \gamma} \{ r(p) + \phi(s) - \mu^\top AD(p) + \mu^\top s \}. \tag{12}
\]

In light of this, we define the dual function

\[
q(\mu) := \max_{p \in [\mathbb{P}]^N, -\gamma \leq s \leq \gamma} \{ r(p) + \phi(s) - \mu^\top AD(p) + \mu^\top s \} = r^\sharp(A^\top \mu) + (-\phi)^*(\mu), \tag{13}
\]

where for every \(\mu \in \mathbb{R}^M\) we define

\[
r^\sharp(A^\top \mu) := \max_{p \in [\mathbb{P}]^N} \{ r(p) - \mu^\top AD(p) \}. \tag{14}
\]

and the convex conjugate (following the convention, e.g., Chapter 3 in (Boyd et al. 2004))

\[
(-\phi)^*(\mu) := \max_{-\gamma \leq s \leq \gamma} \{ \phi(s) + \mu^\top s \}. \tag{15}
\]

By Eq. (12), for every \(\mu \in \mathbb{R}^M\), we have that

\[
p^* \leq q(\mu). \tag{16}
\]

In the primal-dual framework, we iteratively optimize the primal variables \(p, s\) (Line 6) and the dual variable \(\mu\) (Line 9). When updating the dual variable at Line 9, we use the mirror descent method with the help of a mirror descent solver \(\zeta^D\) defined as follows.

**Definition 1 (Mirror Descent Solver).** A mirror descent solver \(\zeta^D(\mu_t, \tilde{g}_t; D, \eta)\) takes \(\mu_t \in D\) and \(\tilde{g}_t\) as input and returns the updated dual variable \(\mu_{t+1} \in D\) at each time \(t\). For a sequence of input \{\(\tilde{g}_t\)\} and the initial dual variable \(\mu_1\), if we repeatedly apply \(\zeta^D\) and produce a sequence of dual variables \{\(\mu_t\)\}. The solver makes sure that for all \(\mu \in D\),

\[
\sum_{t=1}^T \langle \mu_t, \tilde{g}_t \rangle \leq \sum_{t=1}^T \langle \mu, \tilde{g}_t \rangle + \frac{C_1}{\eta} + C_2\eta T, \tag{17}
\]

where \(C_1\) and \(C_2\) are constants that only depend on \(D\).

In other words, the mirror descent solver should generate a sequence \{\(\mu_t\)\} to minimize \(\sum_{t=1}^T \langle \cdot, \tilde{g}_t \rangle\) against any stationary benchmark with the regret at most \(C_1/\eta + C_2\eta T\).

The above-described primal-dual framework is similar to (and inspired by) the algorithm proposed in (Balseiro et al. 2020). However, the key differences are two folds explained as follows.

**Demand Learning.** In contrast to the known demand function in (Balseiro et al. 2020), the demand function is not known to the decision-maker beforehand in our setting. Our algorithm learns

\footnote{When optimizing \(p\), we need to deal with the upper confidence bounds of the estimated quantities, which will be explained soon.}
the parameterized demand function from historical data via the regularized least-squares estimate (Line 3). We then solve another convex program (Line 4) to make sure the estimated parameters \((\tilde{\alpha}, \tilde{B})\) are bounded and \(\tilde{B}\) is negative semi-definite. Finally, we use the Upper Confidence Bound of the \textit{adjusted reward function} \([\tilde{r}_t(p) - \mu_t^\top A\tilde{D}_t(p)]\) (Line 6) to compute the primal update. We will explain this in more details in Section 4.

An additional feature of our demand estimator is that the reward Upper Confidence Bound is defined based on the \(\ell_2\)-norm of \(\Lambda^{-1/2}_t \tilde{p}_t\) where the \(\ell_2\)-norm is usually adopted in the linear bandit literature. Together with the convex program solved in Line 4, our definition of the reward Upper Confidence Bound renders the primal update a combination of a few convex optimization problems (Eq. (8)) that can be efficiently solved, which will be further explained in Section 4.2.\(^5\)

**A New Dual Space.** The dual space \(\mathcal{D}\) is a crucial component in the design of the mirror descent solver \(\varsigma^D\) and affects the regret analysis. Balseiro et al. (2020) adopt a dual space \(\mathcal{D}^{\text{Bal}} = \{ \mu \in \mathbb{R}^M \mid \sup_{a \in \gamma} \phi(a) + \mu^\top a \}\) which might have different shapes for difference fairness regularizer \(\phi(\cdot)\). The fundamental reason that we cannot directly adopt \(\mathcal{D}^{\text{Bal}}\) in our problem, however, is the unboundedness of \(\mathcal{D}^{\text{Bal}}\) which would lead to an unbounded regret due to the unbounded estimation error of the adjusted reward \([\tilde{r}_t(p) - \mu_t^\top A\tilde{D}_t(p)]\) during the learning process (see Eq. (25) for more details). To deal with this issue, we construct a novel, simply and uniformly shaped, and bounded dual space (Eq. (10)). We prove that our dual space encompasses all potential stationary benchmark dual variables \(\overline{p}\) which is necessary for the desired \(O(\sqrt{T})\) regret.

Thanks to the newly constructed dual space, an extra benefit enjoyed by our algorithm is that, together with a carefully chosen variant of the exponentiated gradient descent (EG\(\pm\)) algorithm as the mirror descent solver, we are able to obtain a \textit{uniform} and \textit{closed-form} update of the dual variables for \textit{all} fairness regularizers. In contrast, Balseiro et al. (2020) have to design the dual update step on a case-by-case basis for the fairness regularizers. Also, the closed-form update improves the computational efficiency of the algorithm and is a desired feature in (Balseiro et al. 2020) that is partially achieved for a few selected fairness regularizers.

### 4. Demand and Reward Estimation

The regularized least-squares estimator (Line 3 of Algorithm 1) for demand parameters is frequently used in the linear bandit literature (see, e.g., (Dani et al. 2008, Rusmevichientong and Tsitsiklis 2010, Abbasi-Yadkori et al. 2011)). However, as mentioned before, we need to work with the upper

\(^5\)On the downside, we sacrifice an \(O(\sqrt{N})\) factor in the regret bound. However, we view this degradation relatively small compared to the existing \(O(N^2)\) factor which seems necessary in the regret due to the \(N^2\) parameters in \(B\) to learn.
confidence bound of a specially defined \textit{adjusted reward function}. Also, we employ an additional step (Line 4) to make sure the estimated parameters ($\tilde{\alpha}, \tilde{B}$) are bounded, which is crucial to the regret analysis (more specifically, the analysis of the mirror descent solver). Line 4 also guarantees the negative semi-definiteness of $\tilde{B}$; when computing the Upper Confidence Bound for the estimation, we innovatively use an $\ell_8$-norm confidence radius instead of the usual $\ell_2$-norm confidence radius – both ingredients help the algorithm to compute the upper confidence bound \textit{in polynomial time}. We will show how to computationally efficiently find $(\tilde{\alpha}, \tilde{B}) \in \mathcal{M}_t$ (Line 4) and implement the UCB-type primal update (Line 6) in Section 4.1 and Section 4.2 respectively.

To explain our Upper Confidence Bound method in more details, we first introduce some notations. For convenience, we define the stopping time

$$\tau = \max\{t : \min_i I_{t,i} > 0, t \leq T\},$$

which is the last time period when the inventory levels of all resources remain positive. Most of our analysis will be done only for time periods up to $\tau$. We define

$$f_t(p) := r(p) - \mu^T AD(p)$$

(18)

to be the \textit{adjusted reward function} at price $p$ and with respect to $\mu_t$. Note that this corresponds to the optimization objective in Eq. (14) when $\mu = \mu_t$. Our estimation for $f_t(p)$ is

$$\tilde{f}_t(p) := \tilde{r}_t(p) - \mu^T A\tilde{D}_t(p),$$

(19)

which corresponds to the first part in the optimization objective of $p_t$ in Eq. (8). We also define estimators with regard to $(\tilde{\alpha}_t, \tilde{B}_t)$ as

$$\hat{D}_t(p) = \hat{\alpha}_t + \hat{B}_t p, \quad \tilde{\alpha}_t(p) = \left\langle p_t, \hat{D}_t(p) \right\rangle, \quad \tilde{f}_t(p) := \tilde{r}_t(p) - \mu^T A\tilde{D}_t(p).$$

\textbf{Bounding the estimation errors.} We now derive the estimation errors of $\hat{D}_t, \tilde{r}_t, \tilde{f}_t$ and $\hat{\alpha}_t, \tilde{\alpha}_t, \tilde{f}_t$, as well as their corresponding upper confidence bounds. For any price vector $p$, we let $\hat{p} := (p, 1)$, and then introduce the \textit{regularized information matrix} at time $t$ to be

$$\Lambda_t := (N + 1) \cdot I_{N+1} + \sum_{s \leq t} \tilde{p}_s \tilde{p}_s^T.$$  

(20)

Let

$$\kappa := 2\sqrt{2\tilde{d}^2 (N + 1) \ln (NT(1 + \tilde{p}^2 T)) + 2(N + 1) L_B^2}, \quad \text{and}$$

$$\Delta_t^p(p) := \sqrt{N + 1} \kappa \Lambda_t^{-1/2} \tilde{p}_t \infty \quad \text{and} \quad \Delta_t^r(p) := \sqrt{N + 1} N \tilde{p}_t \kappa \Lambda_t^{-1/2} \tilde{p}_t \infty.$$  

(21)
to be the confidence radii of \( \hat{\tau}_t(p) \) and \( \hat{D}_t(p) \) respectively. Note that here we use \( \| \Lambda_t^{-1/2} \tilde{w}_t \|_x \) instead of the \( \ell_2 \)-norm confidence radius \( \| \Lambda_t^{-1/2} \tilde{w}_t \|_2 \) commonly seen in literature. We finally define

\[
\Delta_t^f(p) := \Delta_t^r(p) + \| \mu_t \|_1 \cdot \| A \|_x \Delta_t^P(p)
\]  

(23)

to be the confidence radius for the adjusted reward estimator \( \hat{f}_t \). We utilize the famous Confidence Ellipsoid Lemma in (Abbasi-Yadkori et al. 2011) to analyze our \( \ell_x \)-type confidence region, and have the following lemma. Our lemma states that the confidence radii defined above (Eq. (22) and Eq. (23)) hold with overwhelming probability.

**Lemma 1.** With probability at least \((1 - O(T^{-1}))\), for all \( t \leq \tau \) and all \( p \in [\bar{p}, \overline{p}] \), we have

\[
\| \hat{D}_t(p) - D(p) \|_x \leq \Delta_t^D(p), \quad |\hat{\tau}_t(p) - \tau(p)| \leq \Delta_t^r(p), \quad \text{and} \quad |\hat{f}_t(p) - f_t(p)| \leq \Delta_t^f(p).
\]

As we will later show in Lemma 5, with probability at least \((1 - O(T^{-1}))\), we are able to find a feasible \((\tilde{\alpha}_t, \tilde{B}_t) \in \mathcal{M}_t \) in Line 4. Combining the definition of \( \mathcal{M}_t \) and Lemmas 1 and 5, we have the following corollary.

**Corollary 1.** With probability at least \((1 - O(T^{-1}))\), for all \( t \leq \tau \) and all \( p \in [\bar{p}, \overline{p}] \), we have

\[
\| \hat{D}_t(p) - D(p) \|_x \leq 2\Delta_t^D(p), \quad |\hat{\tau}_t(p) - \tau(p)| \leq 2\Delta_t^r(p), \quad \text{and} \quad |\hat{f}_t(p) - f_t(p)| \leq 2\Delta_t^f(p).
\]

**The program \( \mathcal{M}_t \).** Note that by Lemma 1, \( \hat{D}_t, \hat{\tau}_t, \hat{f}_t \) already serve as good estimators. However, in the rest part of the algorithm (as well as the analysis), we will mainly work with \( \tilde{D}_t, \tilde{\tau}_t, \tilde{f}_t \), which are defined based on \((\tilde{\alpha}_t, \tilde{B}_t) \) derived by solving the program \( \mathcal{M}_t \) in Line 4 of Algorithm 1. This is due to the following two requirements.

1. The analysis of the mirror descent solver requires an upper bound on the estimated gradient \( \| \tilde{g}_t \|_x \) which relies on the bound of \( \max_i \| \tilde{B}_t e_i \|_2 \) (Eqs. (31,32)).

2. The primal update (Eq. (8)) involves maximizing \( \tilde{r}_t \), a quadratic form of \( \tilde{B}_t \), which can be efficiently optimized only when \( \tilde{B}_t \) is negative semi-definite so that \( \tilde{r}_t \) is concave.

By solving the program \( \mathcal{M}_t \), we find \((\tilde{\alpha}_t, \tilde{B}_t) \) that simultaneously satisfies the above two requirements and stays close to \((\hat{\alpha}_t, \hat{B}_t) \) (in terms of the \( \| \cdot \|_{\Lambda_t} \) norm). In this way, we facilitate both the regret analysis and the efficient computation of the algorithm. Please refer to Sections 4.1 and 4.2 for the efficient implementations of solving \( \mathcal{M}_t \) and the primal update respectively.

**UCB of the adjusted reward function.** When the desired event in Corollary 1 happens, we define
\[ \overline{f}_t(p) := \overline{f}_t(p) + 2\Delta_t^f(p) \] (24)

and have that \( f_t(p) \leq \overline{f}_t(p) \) for all \( t \leq \tau \) and \( p \in [p_\min, p_\max] \). Note that \( \overline{f}_t(\cdot) \) is exactly the optimization objective of \( p_t \) (at Line 6 of Algorithm 1), which is indeed an Upper Confidence Bound (UCB) of the maximization objective in \( r^*(\cdot) \) (Eq. (14), namely \( f_t(\cdot) \)).

Since \( \mu_t \in D \) for all \( t \in [T] \) and \( D = \{ \mu \in \mathbb{R}^M \mid \|\mu\|_1 \leq C \} \), we further upper bound \( \Delta_t^f(p) \) by

\[ \Delta_t^f(p) \leq \Delta_t^\tau(p) + C\|A\|_x \Delta_t^D(p). \] (25)

Therefore, we set

\[ \Delta_t(p) := \Delta_t^\tau(p) + \max\{C\|A\|_x, 1\} \Delta_t^D(p) \] (26)

to upper bound all the confidence radii \( \Delta_t(p), \Delta_t^D(p) \) and \( \Delta_t^f(p) \).

**Bounding the total estimation error.** In our regret analysis, we will relate the regret incurred at time \( t \) to the confidence radii at price \( p_t \) at the time (which aligns with the general Upper Confidence Bound principle – bounding the regret by the confidence radii of the selected actions). And thus we will be interested in the summation of the estimation errors. The following lemma adapts the celebrated Elliptical Potential Lemma (see, e.g., Theorem 11.7 in (Cesa-Bianchi and Lugosi 2006) and Lemma 9 in (Dani et al. 2008)) to upper bound the total estimation error.

**Lemma 2.** With probability \( 1 \), we have the following upper bound for the total estimation error:

\[ \sum_{t=1}^{\tau} \Delta_t(p_t) \leq O \left( \sqrt{N + 1} \max\{p_\max, 1\} (Np + \max\{C\|A\|_x, 1\}) \right) \times \sqrt{N T \log(N + 1 + p_\max^2 T)}, \]

where only a universal constant is hidden in the \( O(\cdot) \) notation.

### 4.1. Solving \( \mathcal{M}_t \) via Ellipsoid Method

In this subsection, we describe how to implement Line 4 and find a feasible solution to \( \mathcal{M}_t \) in polynomial time via the Ellipsoid method. The main lemma of this subsection is Lemma 5.

We first introduce the definition of a separation oracle for a convex set \( K \), which is closely related to the Ellipsoid method.

**Definition 2 (Separation Oracle).** For a closed convex set \( K \subseteq \mathbb{R}^n \), a *separation oracle* for \( K \), namely \( \text{SEP}_K \), is an algorithm that takes a point \( x \in \mathbb{R}^n \) as input and correctly decides whether \( x \in K \). In the case that \( x \notin K \), the separation oracle also returns a hyperplane that separates \( x \) from \( K \). The hyperplane may be characterized by its norm vector \( c \in \mathbb{R}^n \) such that \( c^\top x > c^\top y \) for all \( y \in K \).
The ellipsoid method reduces a convex program feasibility problem to the construction of an efficient separation oracle for the corresponding convex body. The following lemma characterizes such a reduction. The lemma is a simplification of the Theorem 3.2.1 in (Grötschel et al. 2012) modulo the numerical error due to the arithmetic operations on real numbers.\footnote{The numerical error analysis is often tedious but straightforward, which is also the case in this subsection. Therefore, we choose to omit this part and emphasize the main algorithmic idea more clearly.}

**Lemma 3.** Suppose we could perform exact arithmetic operations on real numbers. Let $\text{Ball}(x,r)$ denote the closed ball with radius $r$ and centered at $x \in \mathbb{R}^n$. Given a closed convex set $K \subseteq \mathbb{R}^n$, suppose that there exist $R, r > 0$ such that $K \subseteq \text{Ball}(x_0, R)$ and $\text{Ball}(x_1, r) \subseteq K$ for some $x_0, x_1 \in \mathbb{R}^n$. Given $R, r, x_0$, and a separation oracle for $K$, namely $\text{SEP}_K$, the Ellipsoid method will return a point in $K$ using $O(n^2 \log(R/r))$ calls to the separation oracle and $\text{poly}(n, \log(R/r))$ arithmetic operations.

It is easy to verify that our $M_t$ is a closed convex set in $\mathbb{R}^{N \times (N+1)}$. To apply Lemma 3 to $M_t$, we first upper and lower bound the shape of $M_t$ as follows.

**Lemma 4.** Given the desired event described in Lemma 1, we have that

$$M_t \subseteq \text{Ball}(\hat{B}_t, \kappa \sqrt{N}) \quad \text{and} \quad \text{Ball}([B - T^{-2} \cdot I_N | \alpha], T^{-4}) \subseteq M_t,$$

where we treat the matrices $\hat{B}_t$ and $[B - T^{-2} \cdot I_N | \alpha]$ as $N \times (N + 1)$-dimensional vectors.

**The separation oracle.** It remains to design the separation oracle $\text{SEP}_{M_t}$. Given $\tilde{B} = [\tilde{B} | \tilde{\alpha}]$, we need to verify the following two types of constraints specified in the definition of $M_t$ (Eq. (7)).

- $\| (\tilde{B} - \tilde{B}_t)^\top e_i \|_{M_t} \leq \kappa, \forall i \in [N]$. This condition can be verified for each $i \in [N]$ by straightforward computation. When the condition is not met for some $i \in [N]$, we have that $\kappa < \| (\tilde{B} - \tilde{B}_t)^\top e_i \|_{M_t} = \| \Lambda_t^{1/2} (\tilde{B} - \tilde{B}_t)^\top e_i \|_2$, and there exists $c = \frac{\Lambda_t^{1/2} (\tilde{B} - \tilde{B}_t)^\top e_i}{\| \Lambda_t^{1/2} (\tilde{B} - \tilde{B}_t)^\top e_i \|_2}$ such that

$$c^{\top} \Lambda_t^{1/2} (\tilde{B} - \tilde{B}_t)^\top e_i > \kappa \geq c^{\top} \Lambda_t^{1/2} (B' - \tilde{B}_t)^\top e_i$$

for every $B' = [B' | \alpha']$ where $(\alpha', B') \in M_t$, which defines the separation hyperplane.

- $\| \tilde{B}^\top e_i \|_2 \leq 2L_B, \forall i \in [N]$. This condition can also be verified for each $i \in [N]$ by straightforward computation. When the condition is not met for some $i \in [N]$, we have that $2L_B < \| \tilde{B}^\top e_i \|_2$, and there exists $c = \frac{\tilde{B}^\top e_i}{\| \tilde{B}^\top e_i \|_2}$ such that

$$c^{\top} \tilde{B}^\top e_i > 2L_B \geq c^{\top} B' e_i$$

for every $B' = [B' | \alpha']$ where $(\alpha', B') \in M_t$, which defines the separation hyperplane.
\[ \begin{align*}
\bullet \; & \mathbf{B} + \tilde{\mathbf{B}}^\top \leq 0. \; \text{This condition is equivalent to } \lambda_{\text{max}}(\mathbf{B} + \tilde{\mathbf{B}}^\top) \leq 0 \text{ which can be efficiently verified.}
\end{align*} \]

If this condition is not satisfied, we can efficiently find a vector \( \mathbf{c} \in \mathbb{R}^N \) such that
\[ \langle \tilde{\mathbf{B}} + \tilde{\mathbf{B}}^\top, \mathbf{c}\mathbf{c}^\top \rangle > 0 \geq \langle \mathbf{B}' + (\mathbf{B}')^\top, \mathbf{c}\mathbf{c}^\top \rangle \]
for every \((\alpha', \mathbf{B}') \in \mathcal{M}_t\), which defines the separation hyperplane.

Combining Lemma 4, and the separation oracle constructed above, we may invoke Lemma 3 with \( R = \kappa\sqrt{N} \) and \( r = T^{-4} \), and conclude this subsection with the following lemma.

**Lemma 5.** With probability at least \((1 - \mathcal{O}(T^{-1}))\), for all \( t \leq \tau \), \( \mathcal{M}_t \) is feasible, and we can find \((\tilde{x}_t, \tilde{B}_t) \in \mathcal{M}_t \) via the Ellipsoid method using only \( \text{poly}(N, \log T, \log \overline{d}, \log \overline{p}, \log L_B) \) arithmetic operations on real numbers.

### 4.2. Efficient Primal Update

We now show that, thanks to the new \( \ell_\infty \)-norm-based confidence region, we may efficiently implement the primal update (Line 6) by solving \( \mathcal{O}(N) \) convex optimization problems. We focus on the optimization problem for \( p_t \) as the one for \( s_t \) is already convex. Note that
\[ \begin{align*}
\max_{\mathbf{p} \in [\mathbf{0}, \mathbf{p}]^N} & \left\{ \tilde{\mathbf{J}}_t(p) + 2\Delta_t^f(p) \right\} \\
= & \max_{\mathbf{p} \in [\mathbf{0}, \mathbf{p}]^N} \left\{ \langle \mathbf{p} - \mathbf{A}^\top \mathbf{\mu}_t, \tilde{\mathbf{D}}_t(p) \rangle + 2\sqrt{N + 1} \kappa (N\overline{p} + \|\mathbf{\mu}_t\|_1 \cdot \|\mathbf{A}\|_\infty) \lambda_t^{-1/2} \tilde{\mathbf{p}} \rangle \|_{\infty} \right\} \\
= & \max_{\mathbf{p} \in [\mathbf{0}, \mathbf{p}]^N} \left\{ \langle \mathbf{p} - \mathbf{A}^\top \mathbf{\mu}_t, \tilde{\mathbf{D}}_t(p) \rangle + 2\sqrt{N + 1} \kappa (N\overline{p} + \|\mathbf{\mu}_t\|_1 \cdot \|\mathbf{A}\|_\infty) \lambda_t^{-1/2} \tilde{\mathbf{p}} \rangle \right\} \\
= & \max_{\lambda \in \{\pm e_1, \ldots, \pm e_{N+1}\}} \max_{\mathbf{p} \in [\mathbf{0}, \mathbf{p}]^N} \left\{ \langle \mathbf{p} - \mathbf{A}^\top \mathbf{\mu}_t, \tilde{\mathbf{D}}_t(p) \rangle + 2\sqrt{N + 1} \kappa (N\overline{p} + \|\mathbf{\mu}_t\|_1 \cdot \|\mathbf{A}\|_\infty) \lambda \lambda_t^{-1/2} \tilde{\mathbf{p}} \rangle \right\},
\end{align*} \]

where \( e_i \) (\( i \in \{1, 2, \ldots, N + 1\} \)) is the \( i \)-th canonical basis vector in \( \mathbb{R}^{N+1} \). For any \( \lambda \in \{\pm e_1, \ldots, \pm e_{N+1}\} \), define the convex program (which is convex due to the negative semi-definiteness of \( \tilde{\mathbf{B}} \) guaranteed in Line 4)
\[ P_t^{(\lambda)} := \arg\max_{\mathbf{p} \in [\mathbf{0}, \mathbf{p}]^N} \left\{ \langle \mathbf{p} - \mathbf{A}^\top \mathbf{\mu}_t, \tilde{\mathbf{D}}_t(p) \rangle + 2\sqrt{N + 1} \kappa (N\overline{p} + \|\mathbf{\mu}_t\|_1 \cdot \|\mathbf{A}\|_\infty) \lambda \lambda_t^{-1/2} \tilde{\mathbf{p}} \rangle \right\}. \]

It is easy to verify that \( \arg\max_{\mathbf{p} \in [\mathbf{0}, \mathbf{p}]^N} \left\{ \tilde{\mathbf{J}}_t(p) + 2\Delta_t^f(p) \right\} \subseteq \cup_{\lambda \in \{\pm e_1, \ldots, \pm e_{N+1}\}} P_t^{(\lambda)} \). Therefore, to compute the primal update for \( p_t \), we only need to first solve \( 2(N + 1) \) convex programs to identify \( p_t^{(\lambda)} \in P_t^{(\lambda)} \) for every \( \lambda \in \{\pm e_1, \ldots, \pm e_{N+1}\} \), and then select
\[ p_t = \arg\max_{\mathbf{p} \in \{p_t^{(\lambda)} : \lambda \in \{\pm e_1, \ldots, \pm e_{N+1}\}\}} \left\{ \langle \mathbf{p} - \mathbf{A}^\top \mathbf{\mu}_t, \tilde{\mathbf{D}}_t(p) \rangle + 2\sqrt{N + 1} \kappa (N\overline{p} + \|\mathbf{\mu}_t\|_1 \cdot \|\mathbf{A}\|_\infty) \lambda \lambda_t^{-1/2} \tilde{\mathbf{p}} \rangle \right\}. \quad (27) \]
5. Mirror Descent Solver $\zeta^D$ and its Closed-form Dual Update

In this section, we design the mirror descent solver $\zeta^D$ to satisfy Definition 1. Given the dual space $\mathcal{D}$, for any reference function $h$ that is $\sigma$-strongly convex with respect to $\| \cdot \|_1$ over $\mathcal{D}$, the online mirror descent (OMD) algorithm operates in the following way to update the dual variable:

$$\mu_{t+1} \in \arg \min_{\mu \in \mathcal{D}} \left\{ \langle \mu, \tilde{g}_t \rangle + \frac{1}{\eta} D_h(\mu, \mu_t) \right\},$$

where $D_h(x, y) = h(x) - h(y) - \nabla h(y)^T (x - y)$ is the Bregman divergence. It is well-known (see, e.g., Hazan et al. (2016)) that if $\| \tilde{g}_t \|_x \leq \mathcal{G}$ for all $t$, then if we start with any given $\mu_1 \in \mathcal{D}$, the $\{\mu_t\}$ sequence produced by Eq. (28) guarantees that for any stationary benchmark $\mu \in \mathcal{D}$,

$$\sum_{t=1}^{T} \langle \mu_t, \tilde{g}_t \rangle \leq \sum_{t=1}^{T} \langle \mu, \tilde{g}_t \rangle + \frac{\sup_{\mu \in \mathcal{D}} D_h(\mu, \mu_1)}{\eta} + \frac{\eta \mathcal{G}^2}{2\sigma} T,$$

which matches the requirement of Definition 1.

The popular choices of the reference functions are the negative entropy function $h(x) = \sum_{i=1}^{n} x_i \ln x_i$ (so that $D_h(x, y) = \sum_{i=1}^{n} x_i (\ln x_i - \ln y_i)$, and the OMD algorithm becomes the exponentiated gradient algorithm) and the Euclidean norm $h(x) = \frac{1}{2} \|x\|_2^2$ (so that $D_h(x, y) = \frac{1}{2} \|x - y\|_2^2$ and the OMD algorithm becomes the projected gradient descent algorithm). However, based on the different shapes of the dual space $\mathcal{D}$, one has to carefully choose $h$ to guarantee its strong convexity and proper definition (e.g., the negative entropy function is not properly defined when any of the coordinate becomes negative). Due to this reason, Balseiro et al. (2020) have to design the reference function on a case-by-case basis for various fairness regularizers $\phi$ which shape their dual space $\mathcal{D}$. When designing $h$, Balseiro et al. (2020) also aim to simplify the update rule (Eq. (28)) with the hope of a closed-form update, so as to reduce the computational cost. However, they are only able to achieve this goal for selected fairness regularizers.

In our work, thanks to the simplicity of newly designed dual space $\mathcal{D} = \{\mu \in \mathbb{R}^M \mid \|\mu\|_1 \leq C\}$ (Eq. (10)), we are able to use a uniform mirror descent solver $\zeta^D$ that enjoys the closed-form update for all fairness regularizers. Our $\zeta^D$ is similar to the OMD algorithm with the negative entropy function. The only issue, however, is that the negative entropy function does not apply to negative coordinates covered by our dual space $\mathcal{D}$. To this end, we employ the special variant of the algorithm that separately deals with the positive weights and negative weights in $\mu_t$. The algorithm was proposed by Kivinen and Warmuth (1997) and called $\text{EG}^\pm$ (Exponentiated Gradient Algorithm with Positive and Negative Weights).

The $\text{EG}^\pm$ algorithms is formally described in Algorithm 2. Note that instead of a single vector $\mu_t$, the algorithm keeps two vectors $\mu_{t+1}^+$ and $\mu_{t+1}^-$, and the update of the two vectors are in simple
closed forms. While both vectors are in $\mathbb{R}_+^M$, they respectively represent (the absolute values of) the positive and negative weights in $\mu_t$ (see Eq. (30)). Due to this technical reason, to use $\text{EG}^\pm$ as our mirror descent solver, we need to slightly modify the description of our main Algorithm 1. First, we initialize the two vectors as

$$\mu_1^+ = \mu_1^- = (C/M, \ldots, C/M),$$

which replaces the initialization (Line 1) of Algorithm 1. We also replace the dual update (Eq. (9)) of Algorithm 1 by

$$(\mu_{t+1}^+, \mu_{t+1}^-) = \text{EG}^\pm(\mu_t^+, \mu_t^-, \tilde{g}_t; \mathcal{D}, \eta), \quad \mu_{t+1} = \mu_{t+1}^+ - \mu_{t+1}^-.$$  

(30)

**Algorithm 2** \text{EG}^\pm(\mu_t^+, \mu_t^-, \tilde{g}_t; \mathcal{D}, \eta)

1: Compute the $\mu_{t+1}^+$ and $\mu_{t+1}^-$ vectors as follows:

2: for $i = 1, 2, \ldots, M$ do

3: \[ [\mu_{t+1}^+]_i = \frac{C[\mu_t^+], \exp(-\eta C[\tilde{g}_t],_i)}{\sum_{i=1}^M ([\mu_t^+], \exp(-\eta C[\tilde{g}_t],_i) + [\mu_t^-], \exp(\eta C[\tilde{g}_t],_i))}, \]

4: \[ [\mu_{t+1}^-]_i = \frac{C[\mu_t^-], \exp(\eta C[\tilde{g}_t],_i)}{\sum_{i=1}^M ([\mu_t^+], \exp(-\eta C[\tilde{g}_t],_i) + [\mu_t^-], \exp(\eta C[\tilde{g}_t],_i))}. \]

4: end for

5: return $(\mu_{t+1}^+, \mu_{t+1}^-)$.

It remains to choose $G$ as the upper bound of $\|\tilde{g}_t\|_\infty$. To this end, we set

$$G := 2 \max\{\overline{p}, 1\}(N + 1)L_B\|A\|_\infty + \overline{\gamma}.$$ 

Since for all $i \in [N]$ we have $\|\tilde{B}_i e_i\| \leq (N + 1)\|\tilde{B}_i e_i\|_2$ and $\|\tilde{B}_i e_i\|_2 \leq 2L_B$, which is guaranteed in Line 4, it is easy to obtain $\|\tilde{B}_i\|_\infty = \max_i \|\tilde{B}_i e_i\|_1 \leq 2(N + 1)L_B$. And thus we could have the following upper bound of $\|\tilde{D}_t(p_i)\|_\infty$

$$\|\tilde{D}_t(p_i)\|_\infty = \|\tilde{B}_i \hat{p}_t\|_\infty \leq \|\tilde{B}_i\|_\infty \max\{\overline{p}, 1\} \leq 2(N + 1)L_B \max\{\overline{p}, 1\}. \quad (31)$$

Therefore, we may upper bound $\|\tilde{g}_t\|_\infty$ by $G$:

$$\|\tilde{g}_t\|_\infty = \|A \tilde{D}_t(p_i) - s_t\|_\infty \leq \|A\|_\infty \|\tilde{D}_t(p_i)\|_\infty + \overline{\gamma} \leq 2 \max\{\overline{p}, 1\}(N + 1)L_B\|A\|_\infty + \overline{\gamma} = G. \quad (32)$$

By directly applying Theorem 2 in Hoeven et al. (2018), we have the following lemma showing that $\text{EG}^\pm$ satisfies our requirement of the mirror descent solver.

**Lemma 6.** By adopting $\text{EG}^\pm$ as our mirror descent solver $\varsigma^D$, Definition 1 is satisfied with

$$C_1 = \ln(2M) \quad \text{and} \quad C_2 = C^2G^2/2.$$  

(33)
6. Regret Analysis

With the main technical tools ready in hand, we now prove the following main theorem which upper bounds the regret of our Algorithm 1.

**Theorem 1.** When combining Algorithm 1 with our EG± mirror descent solver (Algorithm 2), we may upper bound the regret of the algorithm by

\[
\mathcal{R}(T) \leq (\|A\|_\infty \bar{d}/\bar{\gamma} + \mathcal{O}(1)) \left[ r(p^*) + \phi(AD(p^*)) \right] + \mathcal{O}(\|A\|_\infty \bar{d}\sqrt{T} \log(MT)) \\
+ 2\sqrt{C_1 C_2 T} + \mathcal{O} \left( \sqrt{N + \frac{1}{\gamma} \max\{p, 1\}} (Np + \max\{C\|A\|_\infty, 1\}) \right) \times \sqrt{NT \log(N + 1 + \bar{p}^2 T)},
\]

where we may choose values for \(C_1\) and \(C_2\) according to Eq. (33) and only universal constants are hidden in the \(\mathcal{O}(\cdot)\) notations.

**Remark 1.** Recall that \(C = L + ((\tau + \bar{\phi})/\bar{\gamma})\), \(\tau = N\bar{p}\bar{d}\), and \(\kappa\) is defined in Eq. (21). Assuming the problem parameters \(\bar{d}, \bar{p}, \bar{\phi}, L, L_B, \|A\|_\infty \leq \mathcal{O}(1)\) and \(\gamma \geq \Omega(1)\), we have that \(C_1 \leq \tilde{\mathcal{O}}(1), C, C_2 \leq \mathcal{O}(N), \kappa \leq \tilde{\mathcal{O}}(N)\), and \(\mathcal{R}(T) \leq \tilde{\mathcal{O}}(N^{5/2} \sqrt{T})\).

The rest of this section is devoted to the proof of Theorem 1. Recall that \(\tau = \max\{t : \min_{i \in [M]} C_{t,i} > 0, t \leq T\}\) is the stopping time till when the inventory levels of all resources remain positive. For convenience, for \(t \leq \tau\), we define

\[
g_t := -AD(p_t) + s_t.
\]

For \(t > \tau\), we set all relevant quantities to zeros:

\[
g_t = \hat{g}_t = s_t = 0, \quad p_t = 0, \quad D(p_t) = d_t = 0, \quad \forall t > \tau.
\]

By Eq. (5), we have that

\[
\mathcal{R}(T) \leq T[r(p^*) + \phi(AD(p^*))] - \mathbb{E} \left[ \sum_{t=1}^{T} r(p_t) + T \phi \left( \frac{1}{T} \sum_{t=1}^{T} AD(p_t) \right) \right]. \tag{35}
\]

The proof of our main theorem follows the general framework of the primal-dual analysis of online optimization problems (e.g., (Beck and Teboulle 2003, Hazan et al. 2016, Balseiro et al. 2022, 2020)), and will be detailed in 5 steps. The main differences from (Balseiro et al. 2020) is that in Step II, we need to deal with the estimation error in the dual expression that relates to the fairness regularizer (note the \(\bar{g}_t\) term in \(\sum_{t=1}^{T} \langle \mu_t, \bar{g}_t \rangle + \sum_{t=1}^{T} \phi(s_t) - T \phi \left( \frac{1}{T} \sum_{t=1}^{T} AD(p_t) \right)\)). We bound this part in Steps III and IV. This error can be upper bounded by estimation error of \(\bar{g}_t\) multiplied by the \(\ell_1\)-norm of the dual variables. Our definition of the dual space \(\mathcal{D} = \{\mu \in \mathbb{R}^M \mid \|\mu\|_1 \leq C\}\) again kicks in to help upper bound the error.

\(^7\) We will treat \(D(p_t)\) as a symbol rather than a function of \(p_t\) for \(t > \tau\).
6.1. Step I: Replacing $d_t$ with $D(p_t)$ and Introducing $\tilde{R}$

The first step is to replace the real demand $d_t$ on the Right-Hand-Side of Eq. (35) with the expected demand $D(p_t)$ so that the resulting expression $\tilde{R}$ is easier to deal with. By applying the Azuma-Hoeffding inequality and a union bound, we have the following lemma.

**Lemma 7.** With probability at least $(1 - 1/T)$, it holds that

$$\forall i \in [M], \quad \left| \sum_{t=1}^{T} [Ad_t]_i - \sum_{t=1}^{T} [AD(p_t)]_i \right| \leq \mathcal{O}(\|A\|_x \bar{d} \sqrt{T \log(MT)}). \quad (36)$$

Lemma 7 implies that

$$\left| \sum_{t=1}^{T} Ad_t - \sum_{t=1}^{T} AD(p_t) \right| \leq \mathcal{O}(\|A\|_x \bar{d} \sqrt{T \log(MT)}). \quad (37)$$

Together with the Lipschitz continuity of $\phi(\cdot)$, Eq. (37) implies that

$$\left| \phi \left( \frac{1}{T} \sum_{t=1}^{T} Ad_t \right) - \phi \left( \frac{1}{T} \sum_{t=1}^{T} AD(p_t) \right) \right| \leq \frac{L}{T} \left| \sum_{t=1}^{T} Ad_t - \sum_{t=1}^{T} AD(p_t) \right| \leq \mathcal{O} \left( L \|A\|_x \bar{d} \sqrt{\frac{\log(MT)}{T}} \right). \quad (38)$$

Now we define the random variable

$$\tilde{R} := T[r(p^*) + \phi(AD(p^*))] - \left[ \sum_{t=1}^{\tau} r(p_t) + T \phi \left( \frac{1}{T} \sum_{t=1}^{T} AD(p_t) \right) \right]. \quad (39)$$

Eq. (38) (which holds with probability at least $(1 - 1/T)$) implies that

$$\mathcal{R}(T) \leq \mathbb{E}[\tilde{R}] + \mathcal{O}(L \|A\|_x \bar{d} \sqrt{T \log(MT)}), \quad (40)$$

and we may turn to upper bound $\mathbb{E}[\tilde{R}]$ instead. In the following steps, we will upper bound the value of $\tilde{R}$ conditioned on that the desired events of Lemma 7 and Lemma 1 hold. Note that this scenario happens with probability at least $(1 - \mathcal{O}(T^{-1}))$, and $\tilde{R}$ is at most $T[r(p^*) + \phi(AD(p^*))]$ in the rare opposite case.

6.2. Step II: Bounding the Fluid Optimum by the UCB of the Dual

The goal of the second step is, given the desired event of Corollary 1, to establish for each $t \leq \tau$ that

$$r(p^*) + \phi(AD(p^*)) \leq r(p_t) + \phi(s_t) + \left< \mu_t, -A\hat{D}_t(p_t) + s_t \right> + 2(\Delta_t^r(p_t) + \Delta_t^f(p_t)), \quad (41)$$

where $p^*$ is the optimal solution of the fluid model (Eq. (11)), and recall the definitions of the primal variables $p_t$ and $s_t$ in Eq. (8). The Right-Hand-Side of Eq. (41) can be viewed as the Upper-Confidence-Bound of the dual function $q(\mu_t)$.
To prove Eq. (41), we first introduce \( p^*_t := \arg \max_{p \in [E]^N} \{ r(p) - \langle \mu_t, AD(p) \rangle \} \) which can be viewed as the desired choice for \( p_t \) (without the estimation errors of \( r(\cdot) \) and \( D(\cdot) \)). In the following claim we upper bound the fluid optimum by the exact dual function.

**Claim 1.** \( r(p^*) + \phi(AD(p^*)) \leq r(p^*_t) + \phi(s_t) + \langle \mu_t, -AD(p^*_t) + s_t \rangle . \)

Claim 1 is essentially a restatement of the weak duality and its proof is deferred to the supplementary materials. Comparing Claim 1 and our goal (Eq. (41)), we only need to upper bound the estimation errors. In particular, it suffices to have that

\[
|r(p_t) - \bar{r}_t(p_t)| \leq 2\Delta^*_t(p_t) \quad \text{and} \quad r(p^*_t) + \langle \mu_t, -AD(p^*_t) \rangle \leq \bar{r}_t(p_t) + \langle \mu_t, -A\bar{D}_t(p_t) \rangle + 2\Delta^t_t(p_t),
\]

where the first inequality is exactly guaranteed by the first item of Corollary 1, for the second inequality, we have that

\[
r(p^*_t) - \langle \mu_t, AD(p^*_t) \rangle = f_t(p^*_t) \leq \bar{f}_t(p^*_t) \leq \bar{r}_t(p_t) = \bar{r}_t(p_t) + \langle \mu_t, -A\bar{D}_t(p_t) \rangle + 2\Delta^t_t(p_t),
\]

where the first inequality is by the third item of Corollary 1 and the second inequality is due to our Upper-Confidence-Bound-style primal update (Eq. (8)).

Now we have established Eq. (41). Together with the definition of \( \hat{R} \) (Eq. (39)) and that \( \hat{y}_t = -A\bar{D}_t(p_t) + s_t \) (Line 8 of Algorithm 1), we obtain that

\[
\hat{R} \leq (T - \tau)[r(p^*) + \phi(AD(p^*))] + \sum_{t=1}^{\tau} 2(\Delta^*_t(p_t) + \Delta^t_t(p_t)) + \sum_{t=1}^{\tau} \phi(s_t) - T\phi \left( \frac{1}{T} \sum_{t=1}^{T} AD(p_t) \right).
\] (42)

Observe that in Eq. (42) we have \( \sum_{t=1}^{\tau} \phi(s_t) - T\phi \left( \frac{1}{T} \sum_{t=1}^{T} AD(p_t) \right) \) and \( \sum_{t=1}^{\tau} \langle \mu_t, \hat{y}_t \rangle \) In the next two steps, we will bound them separately.

### 6.3. Step III: Upper Bounding

\[
\sum_{t=1}^{\tau} \phi(s_t) - T\phi \left( \frac{1}{T} \sum_{t=1}^{T} AD(p_t) \right)
\]

In this step we upper bound the term \( \sum_{t=1}^{\tau} \phi(s_t) - T\phi \left( \frac{1}{T} \sum_{t=1}^{T} AD(p_t) \right) \) in Eq. (42) by the dot products between a carefully selected dual variable \( \overline{\mu} \) and \( \{g_t\} \) (Eq. (34)). It is also important to guarantee that \( \overline{\mu} \) stays in the range of our novel dual space \( D \), which is used by our later analysis. Formally, we prove the following lemma.

**Lemma 8.** Let

\[
\overline{\mu} = \arg \max_{\mu \in \mathbb{R}^M} \{ -(-\phi)^*(\mu) + \langle \mu, \overline{s} \rangle \}.
\] (43)
It holds that
\[
\sum_{t=1}^{T} \phi(s_t) - T \phi \left( \frac{1}{T} \sum_{t=1}^{T} AD(p_t) \right) \leq \sum_{t=1}^{T} \langle \boldsymbol{\mu}, g_t \rangle = - \sum_{t=1}^{\tau} \langle \boldsymbol{\mu}, g_t \rangle. \tag{44}
\]

Moreover, we have that \( \boldsymbol{\mu} \in \mathcal{D} \).

Noting that when \( t > \tau \) we have \( s_t = 0 \) and thus \( \phi(s_t) \geq 0 \) (by Assumption 1). Together with Eq. (42) and Lemma 8 we further upper bound \( \tilde{R} \) by
\[
\tilde{R} \leq (T - \tau) \left[ r(p^*) + \phi(AD(p^*)) \right] + \sum_{t=1}^{\tau} 2 \left( \Delta_t(p_t) + \Delta_t^f(p_t) \right) + \sum_{t=1}^{\tau} \langle \mu_t, \tilde{g}_t \rangle - \sum_{t=1}^{\tau} \langle \boldsymbol{\mu}, g_t \rangle. \tag{45}\]

6.4. Step IV: Upper Bounding \( \sum_{t=1}^{\tau} \langle \mu_t, \tilde{g}_t \rangle \)

In this step, we upper bound the term \( \sum_{t=1}^{\tau} \langle \mu_t, \tilde{g}_t \rangle \) in Eq. (45) (as well as Eq. (42)) by combining the properties of the mirror descent solver and our Upper-Confidence-Bound-type estimator. Intuitively, we would like to replace \( \tilde{g}_t \) by \( g_t \) so that the term could be compared with the other term \( \sum_{t=1}^{\tau} \langle \boldsymbol{\mu}, g_t \rangle \) in Eq. (45). Formally, we will establish Eq. (48).

For any \( \mu \in \mathcal{D} \), by applying the definition of the mirror descent solver in Definition 1, we have that
\[
\sum_{t=1}^{\tau} \langle \mu_t, \tilde{g}_t \rangle \leq \sum_{t=1}^{\tau} \langle \mu_t, g_t \rangle + \sum_{t=1}^{\tau} \langle \mu_t, \Delta_t(p_t) \rangle + \frac{C_1}{\eta} + C_2 \eta T \tag{46}
\]
where \( C_1 \) and \( C_2 \) are the constant parameters in Definition 1. By Hölder’s Inequality, (for any \( \mu \in \mathcal{D} \)) we have that
\[
\left| \langle \mu, A\tilde{D}_t(p_t) - AD(p_t) \rangle \right| \leq \| \mu \|_1 \cdot \| A\tilde{D}_t(p_t) - AD(p_t) \|_\infty \leq 2C \| A \|_\infty \Delta_t^D(p_t). \tag{47}
\]

Note that here we crucially rely on our definition of the dual space \( \mathcal{D} = \{ \mu \in \mathbb{R}^M | \| \mu \|_1 \leq C \} \). Combining Eq. (46) and Eq. (47), for any \( \mu \in \mathcal{D} \), we establish that
\[
\sum_{t=1}^{\tau} \langle \mu_t, \tilde{g}_t \rangle \leq \sum_{t=1}^{\tau} \langle \mu_t, g_t \rangle + \sum_{t=1}^{\tau} \langle \delta_t, g_t \rangle + \frac{C_1}{\eta} + C_2 \eta T + \sum_{t=1}^{\tau} 2C \| A \|_\infty \Delta_t^D(p_t), \tag{48}
\]

Recall the definition of \( \boldsymbol{\mu} \) in Eq. (43). Let \( \mu = \boldsymbol{\mu} + \delta \) where \( \delta \in \mathbb{R}^M_+ \) satisfying \( \boldsymbol{\mu} + \delta \in \mathcal{D} \) will be determined later. Plugging our choice of \( \mu \) into Eq. (48), we have that
\[
\sum_{t=1}^{\tau} \langle \mu_t, \tilde{g}_t \rangle \leq \sum_{t=1}^{\tau} \langle \boldsymbol{\mu}, g_t \rangle + \sum_{t=1}^{\tau} \langle \delta_t, g_t \rangle + \frac{C_1}{\eta} + C_2 \eta T + \sum_{t=1}^{\tau} 2C \| A \|_\infty \Delta_t^D(p_t). \tag{49}
\]
Recalling $\Delta_t(p) := \Delta_t(p) + \max\{C\|A\|_2, 1\}\Delta_t^D(p)$ and combining Eq. (45) and Eq. (49), we obtain that

$$\bar{R} \leq (T - \tau)[r(p^*) + \phi(AD(p^*))] + \sum_{t=1}^{\tau} \langle \delta, g_t \rangle + \frac{C_1}{\eta} + C_2\tau T + \frac{\tau}{t=1} 4\Delta_t(p_t).$$

### 6.5. Step V: Choosing Parameters and Putting Things Together

We finally choose the proper parameters to upper bound $\bar{R}$ and conclude the proof. For the choice of $\delta$, we discuss the following two cases.

**Case 1: $\tau = T$.** If none of the resources depletes before time horizon $T$, i.e., $\tau = T$, we set $\delta = 0$. Now, Eq. (50) implies that

$$\bar{R} \leq \frac{C_1}{\eta} + C_2\eta T + \sum_{t=1}^{\tau} 4\Delta_t(p_t).$$

**Case 2: $\tau < T$.** If $\tau < T$, then there exists a resource $i \in [M]$ such that

$$\sum_{t=1}^{\tau} [Ad_t]_i + \|A\|_x \bar{d} \geq T\gamma_i.$$  (51)

We now set $\delta = ([r(p^*) + \phi(AD(p^*))]/\gamma_i) e_i$, where $e_i$ is the $i$-th unit vector, it is easy to verify that $\mu = \bar{p} + \delta \in \mathcal{D}$. Thus, combining Eq. (36) and Eq. (51), we have that

$$\sum_{t=1}^{\tau} \langle \delta, g_t \rangle \leq \frac{([r(p^*) + \phi(AD(p^*))]/\gamma_i) \sum_{t=1}^{\tau} (s_t - [AD(p_t)]),}{\gamma_i}$$

$$\leq ([r(p^*) + \phi(AD(p^*))]/\gamma_i) \sum_{t=1}^{\tau} (s_t - [Ad_t]) + \mathcal{O}(\|A\|_x \bar{d} \sqrt{T\log(MT)}$$

$$\leq ([r(p^*) + \phi(AD(p^*))]/\gamma_i) ((T - \tau)\gamma_i + \|A\|_x \bar{d}) + \mathcal{O}(\|A\|_x \bar{d} \sqrt{T\log(MT)}).$$

Plugging Eq. (52) back into Eq. (50), we obtain that

$$\bar{R} \leq \left(\|A\|_x \bar{d}/\gamma_i\right) [r(p^*) + \phi(AD(p^*))] + \mathcal{O}(\|A\|_x \bar{d} \sqrt{T\log(MT)}) + \frac{C_1}{\eta} + C_2\eta T + \frac{\tau}{t=1} 4\Delta_t(p_t).$$

Combining the above two cases and setting $\eta = \sqrt{\frac{C_1}{C_2}T}$, together with Lemma 2, we get that

$$\bar{R} \leq \left(\|A\|_x \bar{d}/\gamma_i\right) [r(p^*) + \phi(AD(p^*))] + \mathcal{O}(\|A\|_x \bar{d} \sqrt{T\log(MT)})$$

$$+ 2\sqrt{C_1C_2T} + \mathcal{O} \left( \sqrt{N + 1} \max\{\bar{p}, 1\}(N\bar{p} + \max\{C\|A\|_x, 1\}) \right) \times \sqrt{NT\log(N + 1 + \bar{p}^2T)}.$$  

Together with Eq. (40) and the discussion about the rare case when either of the desired events of Corollary 1 and Lemma 7 fails, we conclude that

$$R(T) \leq \left(\|A\|_x \bar{d}/\gamma_i + \mathcal{O}(1)\right) [r(p^*) + \phi(AD(p^*))] + \mathcal{O}(\|A\|_x \bar{d} \sqrt{T\log(MT)})$$

$$+ 2\sqrt{C_1C_2T} + \mathcal{O} \left( \sqrt{N + 1} \max\{\bar{p}, 1\}(N\bar{p} + \max\{C\|A\|_x, 1\}) \right) \times \sqrt{NT\log(N + 1 + \bar{p}^2T)}.$$
7. Numerical Experiments

In this section, we present the numerical experiments on the synthetic data sets to illustrate the effectiveness of our algorithm. For consistency, we use the NRM example presented in (Besbes and Zeevi 2012, Ferreira et al. 2018). In this example, the retailer sells two products \((N = 2)\) using three resources \((M = 3)\), and the resource consumption matrix is defined as

\[
U = \begin{bmatrix}
1 & 1 \\
3 & 1 \\
0 & 5 \\
\end{bmatrix}.
\]

The underlying linear demand function is defined as

\[
D(p) = \begin{bmatrix}
8 \\
9 \\
\end{bmatrix} + \begin{bmatrix}
-1.5 & 0 \\
0 & -3 \\
\end{bmatrix} p.
\]

In contrast to (Besbes and Zeevi 2012, Ferreira et al. 2018) which use a discrete price set in their experiments, we use the continuous price set to test the effectiveness of our algorithm for handling large price sets. We assume that the price range for each product is \([r_1, s_1]\).

In addition, we choose the weighted min-max fairness regularizer

\[
\phi(s) := \lambda \min_i (w_i s_i)
\]

with \(w_i = 1\) for all \(i\). And we test two initial inventory levels \((\gamma = (15, 12, 30)\) and \(\gamma = (10, 8, 20)\)) and four fairness regularization level \((\lambda \in \{0, 0.5, 1.0, 1.5\})\). For brevity, we present the numerical results of \(\gamma = (15, 12, 30)\) in this section and leave the numerical results of initial inventory level \(\gamma = (10, 8, 20)\) to Section EC.5 in the supplementary materials.

**Implementation Details.** We implement Algorithm 1 with \(C = 5\), \(\eta = 0.01/T\) and time horizon \(T \in \{100, 500, 1000, 2000, 3000, 4000, 6000, 8000, 10000\}\). We generate the demand noise from the truncated Gaussian distribution

\[
\text{clip}\left(\mathcal{N}(0,1), 1\right), \quad \text{where} \quad \text{clip}(x, 1) = \begin{cases}
-1 & x < -1; \\
\text{sign}(x) & |x| \leq 1; \\
1 & x > 1;
\end{cases}
\]

We set the coefficient of \(\Lambda_i^{-1/2}\) in Eq. (27) as \(20\sqrt{\ln T}\). Although the regression regularization parameter in Eq. (6) was set as \(N + 1\) to obtain better dependency on \(N\) of the theoretical regret bound, we fix this parameter to be 0.001 in our experiments for the better empirical performance. We conduct 10 trials independently for each case, and plot the average result of these trials in Figure 1 and Figure 2. We also use the shaded region around each curve to indicate the 95% confidence interval across the 10 trials.
Figure 1  The performance of Algorithm 1 with $\gamma = (15, 12, 30)$ and $\lambda \in \{0, 0.5, 1.0, 1.5\}$. Here the $x$-axis of the left figure is the square root of the total time periods $T$ and the $y$-axis is the cumulative regret defined in Eq. (5). The $x$-axis of the right figure is the total time periods $T$ and the $y$-axis is the relative regret defined in Eq. (53).

Results. In the left of Figure 1 is the plot of the regret of Algorithm 1 with $\gamma = (15, 12, 30)$ and $\lambda \in \{0, 0.5, 1.0, 1.5\}$ versus the square root of the total time periods $T$. This figure clearly demonstrates the regret of our algorithm grows at rate $O(\sqrt{T})$ for all regularization levels $\lambda \in \{0, 0.5, 1.0, 1.5\}$, which is consistent with the theoretical guarantee of Theorem 1. In the right of Figure 1 we plot the relative regret of Algorithm 1 versus the total time periods $T$, where the relative regret is defined as

$$
\frac{T[r(p^*) + \phi(AD(p^*))] - E \left[ \sum_{t=1}^{T} r(p_t) + T \phi \left( \frac{1}{T} \sum_{t=1}^{T} Ad_t \right) \right]}{T[r(p^*) + \phi(AD(p^*))]}.
$$

(53)

Note that the narrow 95% confidence intervals indicate the stability and robustness of our algorithm.
We further empirically study the impact of the fairness regularization level $\lambda$ to the utilization of the resources. In the left of Figure 2 is the plot of the max-min fairness versus the total time periods $T$ with $\gamma = (15, 12, 30)$ and $\lambda \in \{0, 0.5, 1.0, 1.5\}$, where the max-min fairness is defined as the minimum element of the average resource consumption vector $\min_i \left( \frac{1}{T} \sum_{t=1}^{T} [A_{dt}]_i \right)$. In the right of Figure 2 we plot the average reward versus the total time periods $T$, where the average reward is defined as $\frac{1}{T} \sum_{t=1}^{T} r(p_t)$. These figures show the max-min fairness increases and the average reward decreases as $\lambda$ grows, indicating the natural trade-off between fairness and the average reward.

We also find that the max-min fairness could be enhanced greatly with a small sacrifice of the average reward reduction. In addition, the variance of max-min fairness is large when there is no fairness regularizer ($\lambda = 0$), and the variance becomes very small after introducing the fairness regularizer. This interesting phenomenon shows that the fairness regularizer could not only enhance the max-min fairness, but also provide good control of its variance.

8. Conclusion

This paper studies the price-based network revenue management with both fairness concern and demand learning, which is motivated by the practical needs of industries such as online retailing and airline applications. To tackle the challenges of this task, we make several innovative technical contributions, which have the potential to be applied to other operations management problems. We propose a primal-dual-type online policy with Upper-Confidence-Bound (UCB) learning method to simultaneously learn the unknown demand function and optimize the composite objective concerning both the NRM revenue and the fairness metric. Both theoretical analysis and numerical results show the effectiveness and the ability to balance the trade-off between revenue and fairness of the developed policy.

For future directions, one can consider adapting the framework in this paper to other revenue management applications with both fairness concern and demand learning. One could also study the demand learning of non-parametric demand functions and consider adapting our framework to other global ancillary objectives beyond fair consumption across resources.

References

Abbasi-Yadkori Y, Pál D, Szepesvári C (2011) Improved algorithms for linear stochastic bandits. Advances in Neural Information Processing Systems 24:2312–2320.

Agarwal A, Beygelzimer A, Dudík M, Langford J, Wallach H (2018) A reductions approach to fair classification. International Conference on Machine Learning, 60–69.

Agrawal S, Devanur N (2016) Linear contextual bandits with knapsacks. Advances in Neural Information Processing Systems 29:3458–3467.
Agrawal S, Devanur NR (2019) Bandits with global convex constraints and objective. *Operations Research* 67(5):1486–1502.

Azar Y, Buchbinder N, Jain K (2010) How to allocate goods in an online market? *European Symposium on Algorithms*, 51–62.

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

Baek J, Farias V (2021) Fair exploration via axiomatic bargaining. *Advances in Neural Information Processing Systems* 34:22034–22045.

Balseiro S, Lu H, Mirrokni V (2020) Regularized online allocation problems: Fairness and beyond. *arXiv preprint arXiv:2007.00514*.

Balseiro SR, Lu H, Mirrokni V (2022) The best of many worlds: Dual mirror descent for online allocation problems. *Operations Research Forthcoming*.

Bansal N, Sviridenko M (2006) The santa claus problem. *Proceedings of the 38th annual ACM symposium on Theory of computing*, 31–40.

Bateni M, Chen Y, Ciocan DF, Mirrokni V (2022) Fair resource allocation in a volatile marketplace. *Operations Research* 70(1):288–308.

Beck A, Teboulle M (2003) Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters* 31(3):167–175.

Bertsimas D, Farias VF, Trichakis N (2011) The price of fairness. *Operations research* 59(1):17–31.

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

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

Bonald T, Massoulié L, Proutiere A, Virtamo J (2006) A queueing analysis of max-min fairness, proportional fairness and balanced fairness. *Queueing systems* 53(1):65–84.

Boyd S, Boyd SP, Vandenberghe L (2004) *Convex optimization* (Cambridge university press).

Bu J, Simchi-Levi D, Xu Y (2022) Online pricing with offline data: Phase transition and inverse square law. *Management Science* Forthcoming.

Cai W, Gaebler J, Garg N, Goel S (2021) Fair allocation through selective information acquisition. *AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society*, 22–28.

Cesa-Bianchi N, Lugosi G (2006) *Prediction, learning, and games* (Cambridge university press).

Chen G, Li X, Ye Y (2021a) Fairer LP-based online allocation via analytic center. *arXiv preprint arXiv:2110.14621*.

Chen Q, Jasin S, Duenyas I (2014) Adaptive parametric and nonparametric multi-product pricing via self-adjusting controls. *Ross School of Business Paper* (1258).
Chen Q, Jasin S, Duenyas I (2019) Nonparametric self-adjusting control for joint learning and optimization of multiproduct pricing with finite resource capacity. *Mathematics of Operations Research* 44(2):601–631.

Chen X, Zhang X, Zhou Y (2021b) Fairness-aware online price discrimination with nonparametric demand models. *arXiv preprint arXiv:2111.08221*.

Chen Y, Shi C (2019) Network revenue management with online inverse batch gradient descent method. *Available at SSRN 3331939*.

Chen Y, Wang H (2018) Why are fairness concerns so important? lessons from a shared last-mile transportation system. *Available at SSRN 3168324*.

Cohen MC, Elmachtoub AN, Lei X (2022) Price discrimination with fairness constraints. *Management Science* Forthcoming.

Cohen MC, Miao S, Wang Y (2021) Dynamic pricing with fairness constraints. *Available at SSRN 3930622*.

Corbett-Davies S, Goel S (2018) The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*.

Dani V, Hayes TP, Kakade SM (2008) Stochastic linear optimization under bandit feedback. *Conference on Learning Theory* 355–366.

Den Boer AV (2014) Dynamic pricing with multiple products and partially specified demand distribution. *Mathematics of operations research* 39(3):863–888.

Den Boer AV (2015) Dynamic pricing and learning: historical origins, current research, and new directions. *Surveys in operations research and management science* 20(1):1–18.

Den Boer AV, Zwart B (2014) Simultaneously learning and optimizing using controlled variance pricing. *Management science* 60(3):770–783.

Donahue K, Kleinberg J (2020) Fairness and utilization in allocating resources with uncertain demand. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 658–668.

Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R (2012) Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226.

Elzayn H, Jabbari S, Jung C, Kearns M, Neel S, Roth A, Schutzman Z (2019) Fair algorithms for learning in allocation problems. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 170–179.

Ferreira KJ, Simchi-Levi D, Wang H (2018) Online network revenue management using thompson sampling. *Operations research* 66(6):1586–1602.

Gallego G, Van Ryzin G (1994) Optimal dynamic pricing of inventories with stochastic demand over finite horizons. *Management science* 40(8):999–1020.
Gallego G, Van Ryzin G (1997) A multiproduct dynamic pricing problem and its applications to network yield management. *Operations Research* 45(1):24–41.

Grötschel M, Lovász L, Schrijver A (2012) *Geometric algorithms and combinatorial optimization*, volume 2 (Springer Science & Business Media).

Gupta S, Kamble V (2021) Individual fairness in hindsight. *Journal of Machine Learning Research* 22(144):1–35.

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

Hoeven D, Erven T, Kotłowski W (2018) The many faces of exponential weights in online learning. *Conference On Learning Theory*, 2067–2092.

Hutchinson B, Mitchell M (2019) 50 years of test (un)fairness: Lessons for machine learning. *Proceedings of the conference on fairness, accountability, and transparency*, 49–58.

Jabbari S, Joseph M, Kearns M, Morgenstern J, Roth A (2017) Fairness in reinforcement learning. *International conference on machine learning*, 1617–1626.

Jang T, Shi P, Wang X (2022) Group-aware threshold adaptation for fair classification. *Proceedings of the AAAI Conference on Artificial Intelligence* 36(6):6988–6995.

Jasin S (2014) Reoptimization and self-adjusting price control for network revenue management. *Operations Research* 62(5):1168–1178.

Joseph M, Kearns M, Morgenstern JH, Roth A (2016) Fairness in learning: Classic and contextual bandits.

Kallus N, Mao X, Zhou A (2022) Assessing algorithmic fairness with unobserved protected class using data combination. *Management Science* 68(3):1959–1981.

Kallus N, Zhou A (2021) Fairness, welfare, and equity in personalized pricing. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 296–314.

Kandasamy K, Sela GE, Gonzalez JE, Jordan MI, Stoica I (2020) Online learning demands in max-min fairness. *arXiv preprint arXiv:2012.08648v1*.

Keskin NB, Zeevi A (2014) Dynamic pricing with an unknown demand model: Asymptotically optimal semi-myopic policies. *Operations research* 62(5):1142–1167.

Keskin NB, Zeevi A (2017) Chasing demand: Learning and earning in a changing environment. *Mathematics of Operations Research* 42(2):277–307.

Kivinen J, Warmuth MK (1997) Exponentiated gradient versus gradient descent for linear predictors. *Information and computation* 132(1):1–63.

Klein R, Koch S, Steinhardt C, Strauss AK (2020) A review of revenue management: Recent generalizations and advances in industry applications. *European journal of operational research* 284(2):397–412.
Li KJ, Jain S (2016) Behavior-based pricing: An analysis of the impact of peer-induced fairness. *Management Science* 62(9):2705–2721.

Liu Y, Radanovic G, Dimitrakakis C, Mandal D, Parkes DC (2017) Calibrated fairness in bandits. *arXiv preprint arXiv:1707.01875*.

Ma W, Xu P, Xu Y (2020) Group-level fairness maximization in online bipartite matching. *arXiv preprint arXiv:2011.13908*.

Maglaras C, Meissner J (2006) Dynamic pricing strategies for multiproduct revenue management problems. *Manufacturing & Service Operations Management* 8(2):136–148.

Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A (2021) A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54(6):1–35.

Miao S, Wang Y (2021) Network revenue management with nonparametric demand learning: $\sqrt{T}$-regret and polynomial dimension dependency. *Available at SSRN 3948140*.

Miao S, Wang Y, Zhang J (2021) A general framework for resource constrained revenue management with demand learning and large action space. *Available at SSRN 3841273*.

Nash Jr JF (1950) The bargaining problem. *Econometrica: Journal of the econometric society* 155–162.

Rusmevichientong P, Tsitsiklis JN (2010) Linearly parameterized bandits. *Mathematics of Operations Research* 35(2):395–411.

Talluri KT, Van Ryzin G, Van Ryzin G (2004) *The theory and practice of revenue management* (Springer).

Vlasiou M, Zhang J, Zwart B (2014) Insensitivity of proportional fairness in critically loaded bandwidth sharing networks. *arXiv preprint arXiv:1411.4841*.

Wang Y, Wang H (2022) Constant regret resolving heuristics for price-based revenue management. *Operations Research* Forthcoming.

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

Zhang Z, Shi P, Ward AR (2022) Routing for fairness and efficiency in a queueing model with reentry and continuous customer classes. *Proceedings of the American Control Conference (ACC)*.
Supplementary materials

EC.1. Some useful technical lemmas

**Lemma EC.1 (Azuma-Hoeffding Inequality).** Let \( \{(D_k, \mathcal{F}_k)\}_{k=1}^{\infty} \) be a martingale difference sequence for which there are constants \( \{(a_k, b_k)\}_{k=1}^{n} \) such that \( D_k \in [a_k, b_k] \) almost surely for all \( k = 1, \ldots, n \). Then, for all \( t \geq 0 \),

\[
\Pr \left[ \left| \sum_{k=1}^{n} D_k \right| \geq t \right] \leq 2 \exp \left( -\frac{2t^2}{\sum_{k=1}^{n} (b_k - a_k)^2} \right).
\]

Recall that \( \mathcal{B} = [B|\alpha] \in \mathbb{R}^{N \times (N+1)} \), \( \hat{\mathcal{B}} = [\hat{B}|\hat{\alpha}] \in \mathbb{R}^{N \times (N+1)} \), and \( \vec{p} := (p, 1) \). Thus we have \( \mathcal{B}^\top e_i = (\mathcal{B}^\top e_i)^\top \vec{p}_i + \varepsilon_i \). Noting \( \|\mathcal{B}^\top e_i\|_2 \leq L_B, \varepsilon_i \leq \tilde{d} \) and \( \|\vec{p}_i\|_2 \leq \vec{p}\sqrt{N+1} \) and applying the Theorem 2 in Abbasi-Yadkori et al. (2011), we have the following confidence ellipsoid lemma.

**Lemma EC.2.** Recall that \( \Lambda_i = (N+1) \cdot \mathbf{1}_{N+1} + \sum_{s \leq t} \vec{p}_s \vec{p}_s^\top \), for any \( \delta > 0 \), with probability at least \( 1 - \delta \), for all \( t \) we have

\[
\Pr \left[ \|\mathcal{B} - \hat{\mathcal{B}}\| e_i \mathbf{1}_{\Lambda_i} \leq \tilde{d} \sqrt{(N+1) \ln \left( \frac{1 + \|p\|^2}{\delta} \right)} + \sqrt{N+1}L_B \right] \geq 1 - \delta.
\]  

(EC.1)

EC.2. Proofs Omitted in Section 2

EC.2.1. Proof of Proposition 1

**Proof of Proposition 1.** By the constraint of Eq. (1) , it is easy to obtain \( \mathbb{E} \left[ \sum_{t=1}^{T} A_{d_t} \right] \leq T \gamma \). Therefore,

\[
J_{\text{opt}} \leq \left\{ \begin{array}{c}
\max_{\pi} \mathbb{E} \left[ \sum_{t=1}^{T} r(p_t) + T \phi \left( \frac{1}{T} \sum_{t=1}^{T} A_{d_t} \right) \right] \\
\text{s.t.} \quad \mathbb{E} \left[ \sum_{t=1}^{T} AD(p_t) \right] \leq T \gamma
\end{array} \right\}.
\]

With the concavity of \( r(p_t) \) by Jenson’s inequality we obtain

\[
\sum_{t=1}^{T} r(p_t) \leq T r \left( \frac{1}{T} \sum_{t=1}^{T} p_t \right).
\]

Therefore, we have

\[
\left\{ \begin{array}{c}
\max_{\pi} \mathbb{E} \left[ \sum_{t=1}^{T} r(p_t) + T \phi \left( \frac{1}{T} \sum_{t=1}^{T} A_{d_t} \right) \right] \\
\text{s.t.} \quad \mathbb{E} \left[ \sum_{t=1}^{T} AD(p_t) \right] \leq T \gamma
\end{array} \right\} \leq \left\{ \begin{array}{c}
\max_{\pi} \mathbb{E} \left[ T r \left( \frac{1}{T} \sum_{t=1}^{T} p_t \right) + T \phi \left( \frac{1}{T} \sum_{t=1}^{T} A_{d_t} \right) \right] \\
\text{s.t.} \quad \mathbb{E} \left[ \sum_{t=1}^{T} AD(p_t) \right] \leq T \gamma
\end{array} \right\}.
\]
Example 1: Weighted Max-min Fairness Regularizer: \( \phi(s) := \lambda \min_i (w_i s_i) \).

We first show \( \phi(\cdot) \) is \( L \)-Lipschitz continuous with respect to the \( \| \cdot \|_\infty \)-norm in the following way,
\[
\phi(s) - \phi(t) = \lambda \min_i (w_i s_i) - \min_i (w_i t_i)) \\
\leq \lambda \max_i (|w_i s_i - w_i t_i|) \\
\leq (\lambda \max_i w_i) \| s - t \|_\infty.
\]

Next we will show the concavity of \( \phi(\cdot) \). For any \( s, t \) and \( \alpha \in [0, 1] \), we have
\[
\phi(\alpha s + (1 - \alpha)t) = \lambda \min_i (\alpha w_i s_i + (1 - \alpha) w_i t_i) \\
\geq \lambda (\alpha \min_i (w_i s_i) + (1 - \alpha) \min_i (w_i t_i)) \\
\geq \alpha \phi(s) + (1 - \alpha) \phi(t).
\]

Example 2: Group Max-min Fairness Regularizer: \( \phi(s) := \lambda \min_i (U \tilde{s}_i) \), where \( \tilde{s} = (w_1 s_1, \ldots, w_m s_m)^T \).

We first show \( \phi(\cdot) \) is \( L \)-Lipschitz continuous with respect to the \( \| \cdot \|_\infty \)-norm in the following way,
\[
\phi(s) - \phi(t) = \lambda \min_i ((U \tilde{s})_i) - \min_i ((U \tilde{t})_i)) \\
\leq \lambda \| U (\tilde{s} - \tilde{t}) \|_\infty \\
\leq \lambda \| U \|_\infty \| \tilde{s} - \tilde{t} \|_\infty \\
\leq (\lambda \| U \|_\infty \max_i w_i) \| s - t \|_\infty.
\]

Next we will show the concavity of \( \phi(\cdot) \). For any \( s, t \) and \( \alpha \in [0, 1] \), we have
\[
\phi(\alpha s + (1 - \alpha)t) = \lambda \min_i ((\alpha U \tilde{s})_i) + (1 - \alpha) ((U \tilde{t})_i) \\
\geq \lambda (\alpha \min_i ((U \tilde{s})_i) + (1 - \alpha) \min_i ((U \tilde{t})_i)) \\
\geq \alpha \phi(s) + (1 - \alpha) \phi(t).
\]
Example 3: Range Fairness Regularizer: \( \phi(s) := \lambda(\min_i(w_i s_i) - \max_i(w_i s_i) + \max_i(w_i \gamma_i)) \).

Example 4: Load Balancing Regularizer: \( \phi(s) := \lambda(\min_i((\gamma_i - s_i)/\gamma_i)) \).

We have shown \( \min_i(w_i s_i) \) is \( L \)-Lipschitz continuous with respect to the \( \| \cdot \|_x \)-norm and concave. By this fact, it is easy to note that Example 3 and Example 4 are also satisfy Assumption 1.

**EC.3. Proofs Omitted in Section 3**

**EC.3.1. Proof of Lemma 1**

*Proof of Lemma 1.* Let \( \delta \) in Lemma EC.2 be \( 1/(NT) \), and thus with probability \( 1 - 1/T \) for any \( i \in [N] \) and \( t \leq \tau \) it holds that

\[
\| (B - \widehat{B}_t)^T e_i \|_{\Lambda_t}^2 = \| A_i^{1/2} ((B - \widehat{B}_t)^T e_i) \|_2^2 \\
\leq 2d^2 (N + 1) \ln (NT(1 + p^2 T)) + 2(N + 1) L_B^2, \tag{EC.2}
\]

where the inequality is due to \( (a + b)^2 \leq 2a^2 + 2b^2 \).

By Eq. (EC.2) and Cauchy-Schwarz inequality, we have

\[
\| (\widehat{B}_t - B) \tilde{p} \|_\infty \leq \max_{i \in [N]} \| ((B - \widehat{B}_t)^T e_i) \| \tilde{p} \\
\leq \max_{i \in [N]} \| ((B - \widehat{B}_t)^T e_i) \| A_i^{1/2} \Lambda_t^{-1/2} \tilde{p} \\
\leq \max_{i \in [N]} \| A_i^{1/2} ((B - \widehat{B}_t)^T e_i) \|_2 \Lambda_t^{-1/2} \tilde{p}_2 \\
\leq \sqrt{2d^2 (N + 1) \ln (NT(1 + p^2 T)) + 2(N + 1) L_B^2 \sqrt{p^\top \Lambda_t^{-1} \tilde{p}}}. \tag{EC.3}
\]

Noting that \( \| (\widehat{B}_t - B) \tilde{p} \|_2 \leq \sqrt{N} \| (\widehat{B}_t - B) \tilde{p} \|_\infty \), by Eq. (EC.3) we have

\[
\| (\widehat{B}_t - B) \tilde{p} \|_2 \leq \sqrt{N} \| (\widehat{B}_t - B) \tilde{p} \|_\infty \\
\leq \sqrt{2d^2 N(N + 1) \ln (NT(1 + p^2 T)) + 2N(N + 1) L_B^2 \sqrt{p^\top \Lambda_t^{-1} \tilde{p}}}. \tag{EC.4}
\]

*Proof of \( \widehat{r}(p) - r(p) \leq \Delta (p) \).

By Eq. (EC.4), we have

\[
\langle p, \tilde{B} \tilde{p} \rangle - \langle p, B \tilde{p} \rangle \leq \| p \|_2 \| (\widehat{B}_t - B) \tilde{p} \|_2 \\
\leq \sqrt{N} p \sqrt{2d^2 N(N + 1) \ln (NT(1 + p^2 T)) + 2N(N + 1) L_B^2 \sqrt{p^\top \Lambda_t^{-1} \tilde{p}}}, \tag{EC.5}
\]

where the first inequality is due to Cauchy-Schwarz inequality.

Since for any vector \( x \in \mathbb{R}^d \) it holds that

\[
\| x \|_2 \leq \sqrt{d} \| x \|_\infty \leq \sqrt{d} \| x \|_2.
\]
Therefore, we have
\[
\sqrt{p^T \Lambda^{-1} p} \leq \sqrt{N+1} \|\Lambda^{-1/2} \tilde{p}\|_\infty \leq \sqrt{N+1} \sqrt{p^T \Lambda^{-1} p}. 
\] (EC.6)

Combining Eq. (EC.5) and Eq. (EC.6), we have
\[
|\hat{r}_t(p) - r(p)| = \left| \langle p, \hat{B} \tilde{p} \rangle - \langle p, B \tilde{p} \rangle \right|
\leq \sqrt{N \tilde{p}^2} \sqrt{2 \tilde{d}^2 (N+1) \ln \left( NT(1 + \tilde{p}^2 T) \right)} + 2(N+1)L_B \sqrt{\tilde{p}^T \Lambda^{-1} \tilde{p}}
\leq N(N+1) \tilde{p} \sqrt{2 \tilde{d}^2 \ln \left( NT(1 + \tilde{p}^2 T) \right)} + 2L_B^2 \|\Lambda^{-1/2} \tilde{p}\|_\infty 
\leq \Delta_t^F(p). 
\] (EC.7)

**Proof of** \[
\|\hat{D}_t(p) - D(p)\|_\infty \leq \Delta_t^D(p). 
\]
Combining Eq. (EC.3) and Eq. (EC.6), it is easy to obtain that
\[
\|\hat{D}_t(p) - D(p)\|_\infty = \|(\hat{B}_t - B) \tilde{p}\|_\infty 
\leq \sqrt{2 \tilde{d}^2 (N+1) \ln \left( NT(1 + \tilde{p}^2 T) \right)} + 2(N+1)L_B \sqrt{\tilde{p}^T \Lambda^{-1} \tilde{p}}
\leq (N+1) \sqrt{2 \tilde{d}^2 \ln \left( NT(1 + \tilde{p}^2 T) \right)} + 2L_B^2 \|\Lambda^{-1/2} \tilde{p}\|_\infty 
\leq \Delta_t^D(p). 
\] (EC.8)

**Proof of** \[
|\hat{f}_t(p) - f_t(p)| \leq \Delta_t^f(p). 
\]
By Eq. (EC.8), with probability 1 - 1/T we have
\[
\left| \langle \mu_t, A \hat{D}_t(p) - AD(p) \rangle \right| \leq \|\mu_t\|_1 \cdot \|A(\hat{D}_t(p) - D(p))\|_\infty 
\leq \|\mu_t\|_1 \cdot \|A\|_\infty \Delta_t^D(p). 
\] (EC.9)

Since \[
f_t(p) = r(p) - \langle \mu_t, AD(p) \rangle, \quad \hat{f}_t(p) = \hat{r}_t(p) - \langle \mu_t, A \hat{D}_t(p) \rangle \quad \text{and} \quad \Delta_t^f(p) = \Delta_t^r(p) + \|\mu_t\|_1 \cdot \|A\|_\infty \Delta_t^D(p). 
\]
Combining Eq. (EC.7) and Eq. (EC.9), we have
\[
|\hat{f}_t(p) - f_t(p)| \leq |\hat{r}_t(p) - r_t(p)| + \left| \langle \mu_t, A \hat{D}_t(p) - AD(p) \rangle \right|
\leq \Delta_t^r(p) + \|\mu_t\|_1 \cdot \|A\|_\infty \Delta_t^D(p)
= \Delta_t^f(p).
\]

Therefore, we complete the proof of Lemma 1. \(\Box\)
EC.3.2. Proof of Corollary 1

Proof of Corollary 1. The proof will be conditioned on when we find a feasible \((\tilde{\alpha}_t, \tilde{B}_t) \in \mathcal{M}_t\) for all \(t \leq \tau\) (which happens with probability \((1 - \mathcal{O}(T^{-1}))\) by Lemma 5) and the desired event of Lemma 1.

For each \(t \leq \tau\) and \(p \in [\underline{p}, \overline{p}]\), we first upper bound \(\|\tilde{D}_t(p) - D(p)\|_\infty\). Note that

\[
\|\tilde{D}_t(p) - D(p)\|_\infty \leq \|\tilde{D}_t(p) - \hat{D}_t(p)\|_\infty + \|\hat{D}_t(p) - D(p)\|_\infty \leq \|\tilde{D}_t(p) - \hat{D}_t(p)\|_\infty + \Delta_t^D(p).
\]

Therefore, we only need to show that \(\|\tilde{D}_t(p) - \hat{D}_t(p)\|_\infty \leq \Delta_t^D(p)\) to prove that \(\|\tilde{D}_t(p) - D(p)\|_\infty \leq 2\Delta_t^D(p)\). For every \(i \in [N]\), let \(\tilde{B}_i = [\tilde{\alpha}_i, \tilde{B}_i]\), and we verify that

\[
|e_i^\top (\tilde{D}_t(p) - \hat{D}_t(p))| = |e_i^\top (\tilde{B}_i - \hat{B}_i)\overline{p}| = |e_i^\top (\tilde{B}_i - \hat{B}_i)\Lambda_t^{-1/2} \Lambda_t^{-1/2} \overline{p}|
\leq \|\Lambda_t^{-1/2} (\tilde{B}_i - \hat{B}_i) \Lambda_t^{-1/2} \overline{p}\| \leq \kappa \|\Lambda_t^{-1/2} \overline{p}\| \leq \sqrt{N + 1} \kappa \|\Lambda_t^{-1/2} \overline{p}\|_\infty = \Delta_t^D(p).
\]

Here, the first inequality is due to Cauchy-Schwarz, the second inequality is due to that \((\tilde{\alpha}_t, \tilde{B}_t) \in \mathcal{M}_t\).

We now upper bound \(|\tilde{r}_i(p) - r(p)|\) as follows.

\[
|\tilde{r}_i(p) - r(p)| \leq |\tilde{r}_i(p) - \hat{r}(p)| + |\hat{r}_i(p) - r(p)| \leq N\overline{p}\|\tilde{D}_t(p) - \hat{D}_t(p)\|_\infty + |\hat{r}_i(p) - r(p)| \leq \Delta_i^r(p) + \Delta_i^r(p),
\]

where in the inequality, we upper bound \(\|\tilde{D}_t(p) - \hat{D}_t(p)\|_\infty\) by \(\Delta_t^D(p)\) due to the paragraph above.

We finally upper bound \(\|\tilde{f}_t(p) - f_t(p)\|\). Note that

\[
|\tilde{f}_t(p) - f_t(p)| \leq |\tilde{f}_t(p) - \hat{f}_t(p)| + |\hat{f}_t(p) - f_t(p)|
\leq |\tilde{r}_i(p) - \hat{r}(p)| + \|\mu_t\| \cdot \|A\| \|\tilde{D}_t(p) - \hat{D}_t(p)\|_\infty + |\hat{f}_t(p) - f_t(p)|
\leq \Delta_i^r(p) + \|\mu_t\| \cdot \|A\| \|\Delta_t^D(p)\| + \Delta_i^f(p) = 2\Delta_i^f(p),
\]

where the second inequality is by the definitions of \(\hat{f}_t\) and \(\tilde{f}_t\), and the third inequality uses the upper bounds for \(|\tilde{r}_i(p) - \hat{r}(p)|\) and \(\|\tilde{D}_t(p) - \hat{D}_t(p)\|_\infty\) derived in the previous parts of this proof.

\(\Box\)

EC.3.3. Proof of Lemma 2

Proof of Lemma 2. Since \(\Lambda_t = (N + 1) \cdot I_{N+1} + \sum_{s \leq t} \tilde{p}_s \tilde{p}_s^\top\), for every \(t \geq 1\), \(\tilde{p}_t^\top \Lambda_t^{-1} \tilde{p}_t \leq \overline{p}^2\). By this fact, we have

\[
\sqrt{\tilde{p}_t^\top \Lambda_t^{-1} \tilde{p}_t} = \min\{\overline{p}, \sqrt{\tilde{p}_t^\top \Lambda_t^{-1} \tilde{p}_t}\}
\leq \max\{\overline{p}, 1\} \min\{1, \sqrt{\tilde{p}_t^\top \Lambda_t^{-1} \tilde{p}_t}\}.
\]
Note \( \| \Lambda_t^{-1/2} \tilde{p}_t \|_\infty \leq \sqrt{\tilde{p}_t^T \Lambda_t^{-1} \tilde{p}_t} \) and recall the definition of \( \Delta_t'(p) \) and \( \Delta_t^D(p) \) we have

\[
\Delta_t'(p) \leq 2(N + 1)p \sqrt{2d^2 \ln \left( NT(1 + \bar{p}^2T) \right)} + 2L_B^2 \sqrt{\tilde{p}_t^T \Lambda_t^{-1} \tilde{p}_t}, \tag{EC.11}
\]

\[
\Delta_t^D(p) \leq 2(N + 1)\sqrt{2d^2 \ln \left( NT(1 + \bar{p}^2T) \right)} + 2L_B^2 \sqrt{\tilde{p}_t^T \Lambda_t^{-1} \tilde{p}_t}. \tag{EC.12}
\]

Recalling the definition \( \kappa = 2\sqrt{2d^2 (N + 1) \ln \left( NT(1 + \bar{p}^2T) \right)} + 2(N + 1)L_B^2 \) and \( \Delta_t(p) := \Delta_t'(p) + \max\{C\|A\|_x, 1\} \Delta_t^D(p) \), by Eqs. (EC.10, EC.11, EC.12), we obtain

\[
\Delta_t(p) \leq \sqrt{N + 1}\kappa (N\bar{p} + \max\{C\|A\|_x, 1\}) \sqrt{\tilde{p}_t^T \Lambda_t^{-1} \tilde{p}_t} \\
\leq \sqrt{N + 1}\kappa \max\{\bar{p}, 1\} (N\bar{p} + \max\{C\|A\|_x, 1\}) \min \left\{ 1, \sqrt{\tilde{p}_t^T \Lambda_t^{-1} \tilde{p}_t} \right\}. \tag{EC.13}
\]

Since \( \Lambda_t = \Lambda_{t-1} + \tilde{p}_t\tilde{p}_t^T = \Lambda_{t-1}^{1/2}(I_{N+1} + \Lambda_{t-1}^{-1/2}\tilde{p}_t\tilde{p}_t^T \Lambda_{t-1}^{-1/2})\Lambda_{t-1}^{1/2} \), we have

\[
\det(\Lambda_t) = \det(\Lambda_{t-1})(1 + \tilde{p}_t^T \Lambda_{t-1}^{-1} \tilde{p}_t) \\
\geq \det(\Lambda_{t-1}) \exp \left( \frac{1}{2} \min\{1, \tilde{p}_t^T \Lambda_{t-1}^{-1} \tilde{p}_t\} \right), \tag{EC.14}
\]

where the first equality is due to \( \det(I + xx^T) = 1 + \|x\|_2^2 \) and the last inequality is due to \( \exp(x/2) \leq 1 + x \) when \( x \in [0, 1] \).

Therefore, with Eq. (EC.14) we have

\[
\sum_{t=1}^{\tau} \min \left\{ 1, \sqrt{\tilde{p}_t^T \Lambda_t^{-1} \tilde{p}_t} \right\} \leq \sqrt{\tau} \sqrt{\sum_{t=1}^{\tau} \min\{1, \tilde{p}_t^T \Lambda_t^{-1} \tilde{p}_t\}} \\
\leq \sqrt{\tau} \sqrt{2\ln\det(\Lambda_\tau) - 2\ln\det(\Lambda_0)} \\
\leq \sqrt{\tau} \sqrt{2\ln \left( \frac{\text{trace}(\Lambda_\tau)}{N + 1} \right)^{N+1}} \\
\leq \sqrt{\tau} \mathcal{O}\left( \sqrt{N \log(N + 1 + \bar{p}^2T)} \right), \tag{EC.15}
\]

where the first inequality is due to Cauchy-Schwarz inequality, the third inequality is due to the AM-GM inequality, and the last inequality is due to \( \text{trace}(\Lambda_t) = \text{trace}(\left((N + 1) \cdot I_{N+1}\right) + \text{trace}(\sum_{s \neq t} \tilde{p}_s \tilde{p}_s^T) \leq (N + 1)(N + 1 + T\bar{p}^2) \).

Combining Eq. (EC.13) and Eq. (EC.15), we have

\[
\sum_{t=1}^{\tau} \Delta_t(p_t) \leq \mathcal{O} \left( \sqrt{N + 1}\kappa \max\{\bar{p}, 1\} (N\bar{p} + \max\{C\|A\|_x, 1\}) \right) \times \sqrt{NT \log(N + 1 + \bar{p}^2T)},
\]

where we complete the proof. \( \Box \)
EC.3.4. Proof of Lemma 4

Lemma EC.3. With probability at least \(1 - O(T^{-1})\), for all \(t \leq \tau\), we have \((\alpha, B - T^{-2} \cdot I_N) \in \mathcal{M}_t\), where \((\alpha, B)\) is the underlying true parameter and \(\mathcal{M}_t = \{(\hat{\alpha}, \hat{B}) : \|\hat{B} - \hat{B}_t\|_{\Lambda_t} \leq \kappa, \forall i \in [N] \text{ and } \hat{B} + \hat{B}^\top \leq 0\}\).

Proof of Lemma EC.3. Note that by the triangle inequality, for all \(t \leq \tau\) and \(i \in [N]\) we have

\[
\|[(B - T^{-2} \cdot I_N|\alpha) - \hat{B}_t]e_i\|_{\Lambda_t} \leq \|(B - \hat{B}_t)e_i\|_{\Lambda_t} + \|[T - 2 \cdot I_N]e_i\|_{\Lambda_t}.
\]

EC.16

For \(\|[(B - T^{-2} \cdot I_N|\alpha) - \hat{B}_t]e_i\|_{\Lambda_t}\), by Eq. (EC.2), with probability \(1 - 1/T\) for any \(i \in [N]\) and \(t \leq \tau\) it holds that

\[
\|(B - \hat{B}_t)e_i\|_{\Lambda_t} \leq \sqrt{2d\ln (NT(1 + p^2T)) + 2(N + 1)L_B^2}.
\]

EC.17

Let \(\operatorname{diag}(\Lambda_t)_i\) be the \(i\)-th element of the diagonal of \(\Lambda_t\). For \(\|[T - 2 \cdot I_N]e_i\|_{\Lambda_t}\), we have

\[
\|[T - 2 \cdot I_N]e_i\|_{\Lambda_t}^2 \leq \operatorname{diag}(\Lambda_t)_i \cdot T^{-4} \leq (N + 1 + p^2T)/T^4,
\]

EC.18

where the last inequality is due to \(\operatorname{diag}(\Lambda_t)_i = N + 1 + \operatorname{diag}(\sum_{s < t} \tilde{p}_s \tilde{p}_s^\top)\) \(\leq N + 1 + p^2T\).

Invoking Eq. (EC.17) and Eq. (EC.18) into Eq. (EC.16), we have

\[
\|[(B - T^{-2} \cdot I_N|\alpha) - \hat{B}_t]e_i\|_{\Lambda_t} \leq \sqrt{2d\ln (NT(1 + p^2T)) + 2(N + 1)L_B^2 + \sqrt{(N + 1 + p^2T)/T^4}}
\]

EC.19

\[
\leq 2\sqrt{2d\ln (NT(1 + p^2T)) + 2(N + 1)L_B^2} = \kappa.
\]

By assumption \(\|B\|_2 \leq L_B\) and \(L_B \geq 1\), we have

\[
\|[B - T^{-2} \cdot I_N]e_i\|_2 \leq \|B\|_2 + \|[T - 2 \cdot I_N]e_i\|_2 \leq L_B + 1/T^2
\]

EC.20

\[
\leq 2L_B.
\]

And it is easy to note that \((B - T^{-2} \cdot I_N) + (B - T^{-2} \cdot I_N)^\top \leq 0\) with the assumption \(B + B^\top \leq 0\). Therefore, we complete the proof of this lemma. \(\square\)

Lemma EC.4. When the desired event in Lemma EC.3 happens, \([B - T^{-2} \cdot I_N|\alpha] + X \in \mathcal{M}_t\) for any \(X \in \{X \in \mathbb{R}^{N \times (N + 1)} : \|X\|_F \leq 1/T^4\}\).
Proof of Lemma EC.4  
Note that by the triangle inequality, for all $t \leq \tau$ and $i \in [N]$ we have
\[
\|([B - T^{-2} \cdot I_N]\alpha + X - \hat{B}_t)^\top e_i\|_{\Lambda_t} \leq \|([B - T^{-2} \cdot I_N]\alpha - \hat{B}_t)^\top e_i\|_{\Lambda_t} + \|X^\top e_i\|_{\Lambda_t}. \tag{EC.21}
\]
When the desired event in Lemma EC.3 happens, by Eq. (EC.19), for any $i \in [N]$ and $t \leq \tau$ it holds that
\[
\|([B - T^{-2} \cdot I_N]\alpha - \hat{B}_t)^\top e_i\|_{\Lambda_t} \leq \sqrt{2d^2 (N+1) \ln(NT(1+p^2T)) + 2(N+1)L_B^2} + \sqrt{(N+1)(1+p^2T)/T^4}. \tag{EC.22}
\]
And for any $X \in \{X \in \mathbb{R}^{N \times (N+1)} \mid \|X\|_F \leq 1/T^4\}$, we can upper bound $\|X^\top e_i\|_{\Lambda_t}$ as follows
\[
\|X^\top e_i\|_{\Lambda_t}^2 \leq \lambda_{\max}(\Lambda_t) \|X^\top e_i\|_2^2 \\
\leq \lambda_{\max}(\Lambda_t) \cdot T^{-8} \\
\leq (N+1)(1+p^2T)/T^8, \tag{EC.23}
\]
where the first inequality is due to $x^\top \Lambda x \leq \lambda_{\max}(\Lambda) \|x\|_2^2$ for any symmetric matrix $\Lambda$, the second inequality is due to $\|X^\top e_i\|_2 \leq \|X\|_F \leq 1/T^4$, and the last inequality is due to $\lambda_{\max}(\Lambda_t) = N+1 + \lambda_{\max}(\sum_{s \leq t} \hat{p}_s \hat{p}_s^\top) \leq N+1 + \text{trace}(\sum_{s \leq t} \hat{p}_s \hat{p}_s^\top) \leq (N+1)(1+p^2T)$.

Invoking Eq. (EC.22) and Eq. (EC.23) into Eq. (EC.21), we have
\[
\|([B - T^{-2} \cdot I_N]\alpha + X - \hat{B}_t)^\top e_i\|_{\Lambda_t} \\
\leq \sqrt{2d^2 (N+1) \ln(NT(1+p^2T)) + 2(N+1)L_B^2} + \sqrt{(N+1)(1+p^2T)/T^4} + \sqrt{(N+1)(1+p^2T)/T^8} \\
\leq 2\sqrt{2d^2 (N+1) \ln(NT(1+p^2T)) + 2(N+1)L_B^2} = \kappa. \tag{EC.24}
\]
And by Eq. (EC.20), we have
\[
\|([B - T^{-2} \cdot I_N]\alpha + X)^\top e_i\|_2 \leq \|([B - T^{-2} \cdot I_N]\alpha)^\top e_i\|_2 + \|X^\top e_i\|_2 \\
\leq L_B + 1/T^2 + 1/T^4 \\
\leq 2L_B. \tag{EC.25}
\]

Now we need to prove the third constraint is satisfied by $[B - T^{-2} \cdot I_N]\alpha + X$. To facilitate our discussion let $\tilde{X} \in \mathbb{R}^{N \times N}$ be the square matrix after deleting the last column of $X$.

Since we have the assumption that $B + B^\top \leq 0$, we only need to show $(\tilde{X} - T^{-2} \cdot I_N) + (\tilde{X} - T^{-2} \cdot I_N)^\top \leq 0$ to prove that $(B - T^{-2} \cdot I_N + \tilde{X}) + (B - T^{-2} \cdot I_N + \tilde{X}) \leq 0$.

By the fact that $\lambda_{\max}(\Lambda) = \|\Lambda\|_2$ and $\|\Lambda\|_2 \leq \|\Lambda\|_F$ if $\Lambda$ is symmetric, we have
\[
\lambda_{\max}(\tilde{X} + \tilde{X}^\top - 2T^{-2} \cdot I_N) = \lambda_{\max}(\tilde{X} + \tilde{X}^\top) - 2/T^2 \\
\leq \|(\tilde{X} + \tilde{X}^\top)\|_F - 2/T^2 \\
\leq 2\|\tilde{X}\|_F - 2/T^2 \\
\leq 0,
\]

where the last inequality is due to \( \| X \|_F \leq 1/T^4 \).

Therefore, combining \((B - T^{-2} \cdot I_N + \hat{X}) + (B - T^{-2} \cdot I_N + \hat{X})^\top \leq 0\) with Eq. (EC.24) and Eq. (EC.25), we complete the proof of this lemma.

\( \square \)

**Proof of Lemma 4.** Note that \( \Lambda_t = (N + 1) \cdot I_{N+1} + \sum_{s < t} \tilde{p}_s \tilde{p}_s^\top \), For all \( i \in [N] \), we have

\[
\|(\tilde{B} - \tilde{B}_i)^\top e_i\|_2 \leq \|(\tilde{B} - \tilde{B}_i)^\top e_i\|_{\Lambda_t}.
\]

And thus, for \( (\tilde{\alpha}, \tilde{B}) \in \mathcal{M}_t \), it holds that

\[
\|(\tilde{B} - \tilde{B}_i)^\top e_i\|_2 \leq \kappa \quad \forall i \in [N].
\]

Therefore, we have \( \| \tilde{B} - \tilde{B}_i \|_F \leq \kappa \sqrt{N} \), i.e., \( \mathcal{M}_t \subseteq \text{Ball}(\tilde{B}_t, \kappa \sqrt{N}) \) when treating the matrix \( \tilde{B}_t \) as an \( N \times (N+1) \)-dimensional vector.

Combining Lemma EC.3 and Lemma EC.4, we will have the following conclusion to complete the proof of the second part of this lemma.

Given the desired event Eq. (EC.2) in Lemma 1, for all \( t \leq \tau \), it holds that \([B - T^{-2} \cdot I_N | \alpha] + X \in \mathcal{M}_t \) for any \( X \in \mathbb{R}^{N \times (N+1)} \| X \|_F \leq 1/T^4 \}, \) i.e., \( \text{Ball}([B - T^{-2} \cdot I_N | \alpha], T^{-4}) \subseteq \mathcal{M}_t \), when treating the matrix \([B - T^{-2} \cdot I_N | \alpha] \) as an \( N \times (N+1) \)-dimensional vector. \( \square \)

**EC.4. Proofs Omitted in Section 6**

**EC.4.1. Proof of Lemma 7**

**Proof of Lemma 7.** Combining the definition of \( \| A \|_\infty \) and the boundedness of \( d_t \), we have

\[
\| A \varepsilon_t \|_\infty \leq \| A \|_\infty \| \varepsilon_t \|_\infty \leq \| A \|_\infty \tilde{d}.
\]

By the assumption on the demand noise, \( \{ \varepsilon_t \}_{t=1}^T \) is martingale difference sequence, so as \( \{ [A \varepsilon_t] \}_{t=1}^T \). Therefore, applying Azuma-Hoeffding’s inequality (Lemma EC.1), for all \( i \in [M] \) with probability \( 1 - 1/(MT) \) we have that

\[
\sum_{i=1}^T [Ad_i]_i - \sum_{i=1}^T [AD(p_t)]_i = O(\| A \|_\infty d \sqrt{T \log(MT)}).
\] (EC.26)

By a union bound, we have that with probability at least \((1 - 1/T)\), Eq. (EC.26) holds for every \( i \in [M] \), proving the lemma. \( \square \)
EC.4.2. Proof of Claim 1

Proof of Claim 1. Recall that the goal is to prove
\[ r(p^*) + \phi(AD(p^*)) \leq r(p_t^*) + \phi(s_t) + \langle \mu_t, -AD(p_t^*) + s_t \rangle. \]  
(EC.27)

By Eq. (14) and the definitions of \( p_t^* \) and \( s_t \) (Eq. (8)), it holds that
\[
\begin{align*}
    r(p_t^*) &= \max_{p \in [p, \bar{p}]} \{ r(p) - \langle A^\top \mu_t, D(p) \rangle \} = r^+(A^\top \mu_t) + \langle \mu_t, AD(p_t^*) \rangle, \\
    \phi(s_t) + \langle \mu_t, s_t \rangle &= \max_{-\gamma \leq s \leq \gamma} \{ \phi(s) + \langle \mu_t, s \rangle \} = (-\phi)^*(\mu_t),
\end{align*}
\]
which leads to
\[
\begin{align*}
    r(p_t^*) + \phi(s_t) &= r^+(A^\top \mu_t) + (-\phi)^*(\mu_t) + \langle \mu_t, AD(p_t^*) \rangle - \langle \mu_t, s_t \rangle \\
    &= q(\mu_t) + \langle \mu_t, AD(p_t^*) - s_t \rangle \geq p^* + \langle \mu_t, AD(p_t^*) - s_t \rangle, \quad (EC.28)
\end{align*}
\]
where the second equality is by the definition of \( q \) (Eq. (13)) and the last inequality is due to the weak duality (Eq. (16)). Combining Eq. (EC.28) and the definition of \( p^* \) (Eq. (11)), we prove Eq. (EC.27). \( \square \)

EC.4.3. Proof of Lemma 8

Proof of Lemma 8. We start by proving Eq. (44). By the definition that \( \pi = \frac{1}{T} \sum_{t=1}^{T} AD(p_t) \) and \( \bar{\mu} = \arg \max_{\mu \in \mathbb{R}^M} \{ -(-\phi)^*(\mu) + \langle \mu, \pi \rangle \} \), we have that
\[
    -(-\phi)^*(\bar{\mu}) + \langle \bar{\mu}, \pi \rangle = \max_{\mu \in \mathbb{R}^M} \{ -(-\phi)^*(\mu) + \langle \mu, \pi \rangle \}.
\]

By Assumption 1, we have that \(-\phi(s)\) is convex and closed with the closed domain \( \{ s : -\gamma \leq s \leq \gamma \} \) (since \( \phi(\cdot) \) is continuous), which implies that
\[
    (-\phi)^*(s) = -\phi(s) \quad \forall s : -\gamma \leq s \leq \gamma.
\]
Thus, for \( \bar{s} \in \{ s : -\gamma \leq s \leq \gamma \} \), it holds that
\[
    -\phi(\bar{s}) = \max_{\mu \in \mathbb{R}^M} \{ -(-\phi)^*(\mu) + \langle \mu, \bar{s} \rangle \} = \max_{\mu \in \mathbb{R}^M} \{ -(\phi)^*(-\mu) + \langle \mu, \bar{s} \rangle \} = -(\phi)^*(\bar{\mu}) + \langle \bar{\mu}, \bar{s} \rangle \quad (EC.29)
\]
Let \( \bar{s} = \frac{1}{T} \sum_{t=1}^{T} \bar{s}_t \). Recall the definition of \((\phi)^*(\cdot)\) (Eq. (15)), we obtain that
\[
    (\phi)^*(\bar{\mu}) = \max_{-\gamma \leq s \leq \gamma} \{ \phi(s) + \langle \bar{\mu}, s \rangle \} \geq \frac{1}{T} \sum_{t=1}^{T} \phi(s_t) + \langle \bar{\mu}, s_t \rangle = \frac{1}{T} \sum_{t=1}^{T} \phi(s_t) + \langle \bar{\mu}, \bar{s} \rangle.
\]
Combining the two equations above, we have that
\[
    \phi(\bar{s}) \geq \frac{1}{T} \sum_{t=1}^{T} \phi(s_t) + T \langle \bar{\mu}, \bar{s} - \bar{s} \rangle. \quad (EC.30)
\]
Finally, by the definition of $g_t$ (Eq. (34)), we have that $T \langle \tilde{\mu}, \tilde{s} - \tilde{x} \rangle = \sum_{t=1}^{T} \langle \tilde{\mu}, g_t \rangle$. Together with Eq. (EC.30), we prove Eq. (44).

We now turn to show that $\tilde{\mu} \in \mathcal{D}$. By the definition of $\phi^*(\cdot)$ (Eq. (15)), for all $s : -\gamma \leq s \leq \gamma$, we have that

$$(-\phi^*(\tilde{\mu})) \geq \phi(s) + \langle \tilde{\mu}, s \rangle.$$  

Together with Eq. (EC.29), for all $s : -\gamma \leq s \leq \gamma$, we have that

$$\phi(s) \geq \phi(s) + \langle \tilde{\mu}, s - s \rangle.$$  \hspace{1cm} (EC.31)

By the definition of the dual norm, it holds that

$$\|\tilde{\mu}\|_1 = \max_{v : \|v\|_{\infty} = 1} \langle \tilde{\mu}, v \rangle.$$  

Let $v^* = \arg \max_{v : \|v\|_{\infty} = 1} \langle \tilde{\mu}, v \rangle$. Since $s$ is an interior point of $[-\gamma, \gamma]$, there exists a small real number $\alpha$ such that $s + \alpha v^* \in [-\gamma, \gamma]$. Plug $s = s + \alpha v^*$ into Eq. (EC.31), we obtain that

$$\phi(s) - \phi(s + \alpha v^*) \geq \langle \tilde{\mu}, \alpha v^* \rangle = \alpha \|\tilde{\mu}\|_1.$$  

By Assumption 1, we have that $\phi(\cdot)$ is $L$-Lipschitz continuous with respect to the $\|\cdot\|_{\infty}$-norm. Therefore, we have $\|\tilde{\mu}\|_1 \leq L$, thus $\tilde{\mu} \in \mathcal{D} = \{\mu \in \mathbb{R}^M : \|\mu\|_1 \leq C\}$ (since $C \geq L$). \hspace{1cm} $\square$

**EC.5. Additional Experimental Results**

In this section, we present the numerical results of initial inventory level $\gamma = (10, 8, 20)$.

In the left of Figure EC.1 is the plot of the regret of Algorithm 1 with $\gamma = (10, 8, 20)$ and $\lambda \in \{0, 0.5, 1.0, 1.5\}$ versus the square root of the total time periods $T$. In the right of Figure EC.1 we plot the relative regret of Algorithm 1 versus the total time periods $T$.

In the left of Figure EC.2 is the plot of the max-min fairness versus the total time periods $T$ with $\gamma = (10, 8, 20)$ and $\lambda \in \{0, 0.5, 1.0, 1.5\}$, where the max-min fairness is defines as the minimum element of the average resource consumption vector $\min_i \left( \frac{1}{T} \sum_{t=1}^{T} [Ad_i] \right)$. In the right of Figure EC.2 we plot the average reward versus the total time periods $T$, where the average reward is defines as $\frac{1}{T} \sum_{t=1}^{T} r(p_t)$.

It is easy to find that the numerical results of initial inventory level $\gamma = (10, 8, 20)$ are almost the same as those in the case $\gamma = (15, 12, 30)$ presented in Section 7, which justifies the effectiveness of our algorithm for different initial inventory levels.
Figure EC.1 The performance of Algorithm 1 with $\gamma = (10, 8, 20)$ and $\lambda \in \{0, 0.5, 1.0, 1.5\}$. Here the x-axis of the left figure is the square root of the total time periods $T$ and the y-axis is the cumulative regret defined in Eq. (5). The x-axis of the right figure is the total time periods $T$ and the y-axis is the relative regret defined in Eq. (53).

Figure EC.2 The max-min fairness $\min_i \left( \frac{1}{T} \sum_{t=1}^{T} [Ad_t]_i \right)$ and the average reward $\frac{1}{T} \sum_{t=1}^{T} r(p_t)$ at regularization levels $\lambda \in \{0, 0.5, 1.0, 1.5\}$ under the initial inventory level $\gamma = (10, 8, 20)$. 