Online Stochastic Optimization with Wasserstein-Based Non-stationarity

Jiashuo Jiang[†]   Xiaocheng Li[‡]   Jiawei Zhang[†]

† Department of Technology, Operations & Statistics, Stern School of Business, New York University
‡ Department of Management Science & Engineering, Stanford University

Abstract: We consider a general online stochastic optimization problem with multiple budget constraints over a horizon of finite time periods. In each time period, a reward function and multiple cost functions are revealed, and the decision maker needs to specify an action from a convex and compact action set to collect the reward and consume the budgets. Each cost function corresponds to the consumption of one budget constraint. The reward function and the cost functions of each time period are drawn from an unknown distribution, which is non-stationary across time. The objective of the decision maker is to maximize the cumulative reward subject to the budget constraints. This formulation captures a wide range of applications including online linear programming and network revenue management, among others. In this paper, we consider two settings: (i) a data-driven setting where the true distribution is unknown but a prior estimate (possibly inaccurate) is available; (ii) an uninformative setting where the true distribution is completely unknown. We propose a unified Wasserstein-distance based measure to quantify the inaccuracy of the prior estimate in Setting (i) and the non-stationarity of the environment in Setting (ii). We show that the proposed measure leads to a necessary and sufficient condition for the attainability of a sublinear regret in both settings. For Setting (i), we propose an *informative gradient descent* algorithm. The algorithm takes a primal-dual perspective and it integrates the prior information of the underlying distributions into an online gradient descent procedure in the dual space. The algorithm also naturally extends to the uninformative setting (ii). Under both settings, we show the corresponding algorithm achieves a regret of optimal order. We illustrate the algorithm performance through numerical experiments.

1. Introduction

In this paper, we study a general online stochastic optimization problem with m budget constraints, each with an initial capacity, over a finite horizon of discrete time periods. At each time t, a reward function $f_t : \mathcal{X} \to \mathbb{R}$ and a cost function $\boldsymbol{g}_t : \mathcal{X} \to \mathbb{R}^m$ are drawn independently from a distribution. Upon the observation of the functions, the decision maker specifies a decision $\boldsymbol{x}_t \in \mathcal{X}$, where $\mathcal{X}$ is assumed to be a convex and compact set. Accordingly, a reward $f(\boldsymbol{x}_t)$ is collected, and each budget $i \in \{1, 2, \dots, m\}$ is consumed by an amount of $g_{it}(\boldsymbol{x}_t)$, where $\boldsymbol{g}_t(\boldsymbol{x}) = (g_{1t}(\boldsymbol{x}), \dots, g_{mt}(\boldsymbol{x}))^\top$. The decision maker's objective is to maximize the totally collected reward subject to the budget capacity constraints.

Our formulation generalizes several existing problems studied in the literature. When f_t and $\boldsymbol{g}_t$ are linear functions for each t, our formulation reduces to the online linear programming (OLP)

problem (Buchbinder and Naor, 2009). Our formulation also covers the network revenue management (NRM) problem (Talluri and Van Ryzin, 2006), including the quantity-based model, the price-based model and the choice-based model (Talluri and Van Ryzin, 2004). Note that for the OLP literature, the reward function and cost functions are assumed to be drawn from an unknown distribution which is stationary over time (or from a random permutation), while the NRM literature always assumes a precise knowledge of the true distribution, though it can be nonidentical across time. In this paper, our focus is an unknown non-stationary setting where the functions f_t and $\boldsymbol{g}_t$ are drawn from a distribution $\mathcal{P}_t$ that is nonidentical over time and unknown to the decision maker. Specifically, we mainly consider two settings: (i) a data-driven setting where there exists an available prior estimate $\hat{\mathcal{P}}_t$ (possibly inaccurate) for the true distribution $\mathcal{P}_t$ of each time period and (ii) an uninformative setting where the true distribution $\mathcal{P}_t$ is completely unknown. When the prior estimates are identical to the true distributions, the data-driven setting reduces to the known non-stationary setting considered in the NRM literature. When the distribution is identical over time, the uninformative setting reduces to the unknown stationary setting considered in the OLP literature.

For both settings, we assume that the true distribution falls into an uncertainty set, which controls the inaccuracy of the prior estimate in the data-driven setting and the non-stationarity of the distributions in the uninformative setting. Our goal is to derive near-optimal policies for both settings, which perform well over the entire uncertainty set. We compare the performances of our algorithms/policies to the "offline" optimization problem which maximizes the total reward with full information/knowledge of all the f_t's and $\boldsymbol{g}_t$'s. We use *regret* as the performance metric, which is defined as worst-case additive optimality gap (over the uncertainty set) between the total reward generated by an algorithm/policy and the offline optimal objective value.

1.1. Main Results and Contributions

For the data-driven setting (in Section 4), the true distribution $\mathcal{P}_t$ is unknown, but we assume the availability of a prior estimate $\hat{\mathcal{P}}_t$. The prior estimate can be based on some history data. We propose a Wasserstein-based deviation measure, *Wasserstein-based deviation budget* (WBDB), to quantify the deviation of the prior estimate from the true distribution. Based on WBDB, we introduce an uncertainty set driven by the notion of WBDB and a parameter W_T (called *deviation budget*), and the uncertainty set encapsulates all the distributions $\mathcal{P}_t$'s that have WBDB no greater than W_T. We illustrate the sharpness of WBDB by showing that if the variation budget W_T is linear in T, sublinear regret could not be achieved by any admissible policy. Next, we develop a new *Informative Gradient Descent* with prior estimate (IGDP) algorithm, which adaptively combines the prior distribution knowledge into an update in the dual space. Our algorithm is motivated

by the traditional online gradient descent (OGD) algorithm (Hazan, 2016). The OGD algorithm applies a linear update rule according to the gradient information at the current period and has been shown to work well in the stationary setting, even when the distribution is unknown (Lu et al., 2020; Sun et al., 2020; Li et al., 2020). The OGD type algorithm has also been developed under stationary setting with unknown distributions in Agrawal and Devanur (2014b) by assuming further stronger conditions on the dual optimal solution, and with known distributions for service level problems (Jiang et al., 2019). However, the update in OGD for each time period only involves information gathered up to the current time period, but for the non-stationary setting, we also need to take advantage of the prior estimates of the future time periods. Specifically, based on a primal-dual convex relaxation of the underlying offline problem, we obtain a prescribed allocation of the budgets over the entire horizon based on the prior estimates. Then, the IGDP algorithm uses this allocation to adjust the gradient descent direction. This idea is new for the related literature in that the IGDP descent direction at each period does not simply come from the historical observations, but it is also informed by the distribution knowledge of the entire horizon. We show that the IGDP algorithm achieves the *first* optimal regret upper bound $O(\max\{\sqrt{T}, W_T\})$.

A few recent works also study similar online decision making problems in a non-stationary environment where $\mathcal{P}_t$ may vary over time. Devanur et al. (2019) study the case of known distribution ($W_T = 0$) and obtain a $1 - O(1/\sqrt{c})$ competitive ratio, where c denotes the minimal capacity of the budget constraints. It remains unclear how to generalize the method in (Devanur et al., 2019) to the setting of $W_T > 0$ where the distribution knowledge is inaccurate or absent. A line of works (Vera and Banerjee, 2020; Banerjee and Freund, 2020a,b) study the known distribution setting ($W_T = 0$) with an additional assumption that the underlying distribution takes a finite support. These works develop algorithms that achieve bounded regret that bears no dependency on T. Compared to this stream of literature, the main results of our paper do not assume the finite supportedness. In addition, when the underlying distribution is finite, we extend the previous algorithm and analysis for the case of $W_T > 0$. Another recent work (Cheung et al., 2020) studies the non-stationary problem and proposes dual-based algorithms that utilize the trajectories sampled from the prior estimate distribution, which can be classified as a static policy (see Section 6.2). We show a major disadvantage of the static policy is that when directly applied to the data-driven setting with estimation erros ($W_T > 0$), it could incur linear regret even when W_T is sublinear; and also the Wasserstein distance is in general tighter than the total variation distance used therein.

For the uninformative setting, we assume no prior knowledge on the true distribution. This setting is consistent with the unknown distribution setting in the literature of OLP problem (Molinaro and Ravi, 2013; Agrawal et al., 2014; Gupta and Molinaro, 2014) and the setting of blind

NRM (Besbes and Zeevi, 2012; Jasin, 2015). We modify the WBDB by replacing the prior estimate of each distribution with their uniform mixture distribution to propose a new measure called *Wasserstein-based non-stationarity measure* (WBNB). By its definition, the WBNB captures the cumulative deviation for all the distribution $\mathcal{P}_t$ from their centric distribution, and it thus reflects the intensity of the non-stationarity associated with $\mathcal{P}_t$'s. In this sense, the WBNB concerns the global change of the distributions, whereas the previous non-stationarity measures (Besbes et al., 2014, 2015; Cheung et al., 2019) in an unconstrained setting characterize the local and temporal change of the distributions over time. In Section 5.1, we illustrate by a simple example that such temporal change measures actually fail in a constrained setting. Thus it addresses the necessity of such a global measure and reveals the interaction between the constraints and the non-stationary environment. Note that a simultaneous and independent work (Balseiro et al., 2020) also uses global change of the distributions to derive a measure of non-stationarity. However, their measure is based on the total variation metric between distributions. With the same example, we illustrate the advantage of using Wasserstein distance instead of total variation distance or KL-divergence. Specifically, the Wasserstein distance compares both the support and densities between two distributions, while total variation distance or KL-divergence compares only the densities. Therefore the Wasserstein-based measure is sharper and more proper for the general online stochastic optimization problem. We formulate the uncertainty set accordingly with the WBNB and propose *Uninformative Gradient Descent* Algorithm (UGD) as a natural reduction of the IGD algorithm in the uninformative setting. We prove that UGD algorithm achieves a regret bound of optimal order.

As a probability distance metric, the Wasserstein distance has been widely used as a measure of the deviation between estimate and true distribution in the distributionally robust optimization literature (e.g. Esfahani and Kuhn (2018)) to represent confidence set and it has demonstrated good performance both theoretically and empirically. To the best of our knowledge, we are the first to use the Wasserstein distance in an online optimization/learning context. From a modeling perspective, the two proposed measures WBDB and WBNB contribute to the study of non-stationary environment for online optimization/learning problem. Specifically, the data-driven setting relaxes the common assumption adopted in the NRM literature that the true distributions are known to the decision maker by allowing the prior estimates to deviate from the true distributions. This deviation can be interpreted as an estimation or model misspecification error, and WBDB establishes a connection between the deviation and algorithmic performance. The uninformative setting generalizes a stream of online learning literature (e.g. Besbes et al. (2015)), which mainly concerned with the unconstrained settings and includes bandits problem (Garivier and Moulines, 2008; Besbes et al., 2014) and reinforcement learning problem (Cheung et al., 2019; Lecarpentier and Rachelson,

2019) as special cases. WBDB adds to the current dictionary of non-stationarity definitions and it specializes for a characterization of the constrained setting.

Finally, we conclude our discussion with several extensions. For the majority of this paper, we analyze algorithm performance under the above two settings without imposing additional structures on the random functions f_t and $\boldsymbol{g}_t$'s. When these functions have a stronger structure such as taking a finite support, we show that a better regret bound can be obtained through a re-solving design. The result complements to the existing works on re-solving algorithms (Jasin and Kumar, 2012; Bumpensanti and Wang, 2020; Vera and Banerjee, 2020) in that it provides a robustness analysis of the re-solving algorithm when the underlying distribution is misspecified. Also, throughout the paper, we provide a number of lower bound results. These lower bound results illustrate the following points: (i) our proposed algorithm has an optimal order of regret in the worst-case sense; (ii) the Wasserstein's distance is sharp; (iii) a dynamic algorithm is necessary to achieve sublinear regret in a non-stationary environment.

1.2. Other Related Literature

As mentioned above, our formulation of the online stochastic optimization problem roots in two major applications: the online linear programming problem and the network revenue management problem. The online linear programming (OLP) problem (Molinaro and Ravi, 2013; Agrawal et al., 2014; Gupta and Molinaro, 2014) covers a wide range of applications through different ways of specifying the underlying LP, including secretary problem (Ferguson et al., 1989), online knapsack problem (Arlotto and Xie, 2020; Jiang and Zhang, 2020), resource allocation problem (Vanderbei et al., 2015; Asadpour et al., 2020), quantity-based network revenue management problem (Jasin and Kumar, 2012; Jasin, 2015), network routing problem (Buchbinder and Naor, 2009), matching problem (Mehta et al., 2005), etc. Notably, the problem has been studied under either (i) the stochastic input model where the coefficient in the objective function, together with the corresponding column in the constraint matrix is drawn from an unknown distribution $\mathcal{P}$, or (ii) the random permutation model where they arrive in a random order. As noted in the paper (Li et al., 2020), the random permutation model exhibits similar concentration behavior as the stochastic input model. The non-stationary setting of our paper relaxes the i.i.d. structure and it can be viewed as a third paradigm for analyzing the OLP problem.

The network revenue management (NRM) problem has been extensively studied in the literature and a main focus is to propose near-optimal policies with strong theoretical guarantees. One popular way is to construct a deterministic linear program as an upper bound of the optimal revenue and use its optimal solution to derive heuristic policies. Specifically, Talluri and Van Ryzin (1998) propose a static bid-price policy based on the dual variable of the linear programming upper bound and proves

that the revenue loss is $O(\sqrt{k})$ when each period is repeated k times and the capacities are scaled by k. Subsequently, Reiman and Wang (2008) show that by re-solving the linear programming upper bound once, one can obtain an $o(\sqrt{k})$ upper bound on the revenue loss. Then, Jasin and Kumar (2012) show that under a non-degeneracy condition for the underlying LP, a policy which re-solves the linear programming upper bound at each time period will lead to an $O(1)$ revenue loss, which is independent of the scaling factor k. The relationship between the performances of the control policies and the number of times of re-solving the linear programming upper bound is further discussed in their later paper (Jasin and Kumar, 2013). Recently, Bumpensanti and Wang (2020) propose an infrequent re-solving policy and show that their policy achieves an $O(1)$ upper bound of the revenue loss even without the "non-degeneracy" assumption. With a different approach, Vera and Banerjee (2020) prove the same $O(1)$ upper bound for the NRM problem and their approach is further generalized in (Vera et al., 2019; Banerjee and Freund, 2020a,b) for other online decision making problems, including online stochastic knapsack, online probing, bin packing, and dynamic pricing. Note that all the approaches mentioned above are mainly developed for the stochastic/stationary setting. When the arrival process of customers is non-stationary over time, Adelman (2007) develops a strong heuristic based on a novel approximate dynamic programming (DP) approach. This approach is further investigated under various settings in the literature (for example (Zhang and Adelman, 2009; Kunnumkal and Talluri, 2016)). Remarkably, although the approximate DP heuristic is one of the strongest heuristics in practice, it does not feature for a theoretical bound. Finally, by using non-linear basis functions to approximate the value of the DP, Ma et al. (2020) develop a novel approximate DP policy and derive a constant competitiveness ratio dependent on the problem parameters. Compared to this line of works, our contribution is two-fold. First, the two main settings that we consider generalize the existing framework of network revenue management in two aspects: (i) we do not assume the knowledge of the underlying distribution; (ii) we do not require the distribution to take a finite support. Second, with the finite-support condition on the underlying distribution, we characterize the relationship between the algorithm performance and the misspecification error of the underlying distribution.

Besides, our problem is also related to the literature of Online Convex Optimization (OCO) and model predictive control. We will discuss the model predictive control literature here and leave the discussion on OCO after presenting our formulation in the next secton. Specifically, our problem can also be formulated as a nonlinear model predictive control (MPC) problem (Rawlings, 2000). In MPC, the decision maker executes the online decisions based on some benchmarks. In our paper, we adopt the offline optimum as the benchmark and develop online algorithms based on the offline optimum. Specifically, for the data-driven setting, we assume that the prior estimates deviate from the true distribution, which corresponds to the broad literature on *robust* MPC (Bemporad and

Morari, 1999), where there are model uncertainties and noises. Compared t this line of works, our paper is the first to propose the use of Wasserstein distance to measure the model uncertainty, and we derive tight regret bound that relates the best achievable algorithm performance with the model uncertainty. Our algorithm also extends to the uninformative setting where there are no prior estimates, and the information about the underlying benchmark has to be learned on-the-fly. In this way, our algorithm for the uninformative setting establishes a connection between online learning and MPC for a time-varying (non-stationary) environment. We refer interested readers to Wagener et al. (2019) for a more detailed discussion over the connections between online learning and MPC.

2. Problem Formulation

Consider the following convex optimization problem

$$
\begin{aligned}
\max \quad & \sum_{t=1}^{T} f_t(\boldsymbol{x}_t) && \text{(CP)} \\
\text{s.t.} \quad & \sum_{t=1}^{T} g_{it}(\boldsymbol{x}_t) \le c_i, \quad i = 1, ..., m, \\
& \boldsymbol{x}_t \in \mathcal{X}, \quad t = 1, ..., T,
\end{aligned}
$$

where the decision variables are $\boldsymbol{x}_t \in \mathcal{X}$ for $t = 1, ..., T$ and $\mathcal{X}$ is a compact convex set in $\mathbb{R}^k$. The function f_t's are functions in the space $\mathcal{F} = \mathcal{F}(\mathcal{X})$ of concave continuous functions and g_{it}'s are functions in the space $\mathcal{G} = \mathcal{G}(\mathcal{X})$ of convex continuous functions, both of which are supported on $\mathcal{X}$. Compactly, we define the vector-value function $\boldsymbol{g}_t(\boldsymbol{x}) = (g_{1t}(\boldsymbol{x}), ..., g_{mt}(\boldsymbol{x}))^\top : \mathbb{R}^k \to \mathbb{R}^m$. Throughout the paper, we use i to index the constraint and t (or sometimes j) to index the decision variables, and we use bold symbols to denote vectors/matrices and normal symbols to denote scalars.

In this paper, we study the *online stochastic optimization* problem where the functions in the optimization problem (CP) are revealed in an online fashion and one needs to determine the value of decision variables sequentially. Specifically, at each time t, the functions $(f_t, \boldsymbol{g}_t)$ are first revealed, and then the decision maker needs to decide the value of $\boldsymbol{x}_t$. Different from the offline setting, at each time t, we do not have the information of the future part of the optimization problem (from time $t+1$ to T). Given the history $\mathcal{H}_{t-1} = \{f_j, \boldsymbol{g}_j, \boldsymbol{x}_j\}_{j=1}^{t-1}$, the decision of $\boldsymbol{x}_t$ can be expressed as a policy function of the history and the observation at the current time period. That is,

$$
\boldsymbol{x}_t = \pi_t(f_t, \boldsymbol{g}_t, \mathcal{H}_{t-1}) \tag{1}
$$

where the policy function π_t can be time-dependent. We denote the policy $\boldsymbol{\pi} = (\pi_1, ..., \pi_T)$. The decision variables $\boldsymbol{x}_t$'s must conform to the constraints in (CP) throughout the procedure, and the objective is aligned with the maximization objective of the offline problem (CP).

2.1. Parameterized Form, Probability Space, and Assumptions

Consider a parametric form of the underlying problem (CP) where the functions $(f_t, \boldsymbol{g}_t)$ are parameterized by a (random) vector $\boldsymbol{\theta}_t \in \Theta \subset \mathbb{R}^l$. Specifically,

$$f_t(\boldsymbol{x}_t) \coloneqq f(\boldsymbol{x}_t; \boldsymbol{\theta}_t), \quad g_{it}(\boldsymbol{x}_t; \boldsymbol{\theta}_t) \coloneqq g_i(\boldsymbol{x}_t; \boldsymbol{\theta}_t)$$

for each $i = 1, ..., m$ and $t = 1, ..., T$. We denote the vector-valued constraint function by $\boldsymbol{g}(\boldsymbol{x}; \boldsymbol{\theta}) = (g_1(\boldsymbol{x}; \boldsymbol{\theta}), ..., g_m(\boldsymbol{x}; \boldsymbol{\theta}))^\top : \mathcal{X} \to \mathbb{R}^m$. Then the problem (CP) can be rewritten as the following parameterized convex program

$$\max \quad \sum_{t=1}^{T} f(\boldsymbol{x}_t; \boldsymbol{\theta}_t) \tag{PCP}$$
$$\text{s.t.} \quad \sum_{t=1}^{T} g_i(\boldsymbol{x}_t; \boldsymbol{\theta}_t) \le c_i, \quad i = 1, ..., m,$$
$$\boldsymbol{x}_t \in \mathcal{X}, \quad t = 1, ..., T,$$

where the decision variables are $(\boldsymbol{x}_1, ..., \boldsymbol{x}_T)$. We note that this parametric form (PCP) avoids the complication of dealing with probability measure in function space. It is introduced mainly for presentation purpose, and it will change the nature of the problem. Moreover, we assume the knowledge of f and $\boldsymbol{g}$ a priori. Here and hereafter, we will use (PCP) as the underlying form of the online stochastic optimization problem.

The problem of online stochastic optimization, as its name refers, involves stochasticity on the functions for the underlying optimization problem. The parametric form (PCP) reduces the randomness from the function to the parameters $\boldsymbol{\theta}_t$'s, and therefore the probability measure can be defined in the parameter space of Θ. First, we consider the following distance function between two parameters $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$,

$$\rho(\boldsymbol{\theta}, \boldsymbol{\theta}') \coloneqq \sup_{\boldsymbol{x} \in \mathcal{X}} \|(f(\boldsymbol{x}; \boldsymbol{\theta}), \boldsymbol{g}(\boldsymbol{x}; \boldsymbol{\theta})) - (f(\boldsymbol{x}; \boldsymbol{\theta}'), \boldsymbol{g}(\boldsymbol{x}; \boldsymbol{\theta}'))\|_\infty \tag{2}$$

where $\| \cdot \|_\infty$ is the supremum norm in $\mathbb{R}^{m+1}$. Without loss of generality, let Θ be a set of class representatives, that is, for any $\boldsymbol{\theta} \neq \boldsymbol{\theta}' \in \Theta$, $\rho(\boldsymbol{\theta}, \boldsymbol{\theta}') > 0$. In this way, the parameter space Θ can be viewed as a metric space equipped with metric $\rho(\cdot, \cdot)$. We choose supremum norm because in our model, one *single* action $\boldsymbol{x}_t$ is made at each period t and it can take any arbitrary value in $\mathcal{X}$. As a result, the comparison between $(f(\cdot; \boldsymbol{\theta}), \boldsymbol{g}(\cdot; \boldsymbol{\theta}))$ and $(f(\cdot; \boldsymbol{\theta}'), \boldsymbol{g}(\cdot; \boldsymbol{\theta}'))$ should be made at each point $\boldsymbol{x}$ and thus the supremum norm is a natural choice for defining $\rho(\boldsymbol{\theta}, \boldsymbol{\theta}')$. In this way, the definition of ρ is based on the vector-valued function $(f, \boldsymbol{g}) : \mathcal{X} \to \mathbb{R}^{m+1}$. Thus it captures the effect of different parameters on the function value rather than the original Euclidean difference in the parameter

space. Let $\mathcal{B}_{\Theta}$ be the smallest σ-algebra in Θ that contains all open subsets (under metric ρ) of Θ. We denote the distribution of $\boldsymbol{\theta}_t$ as $\mathcal{P}_t$ which can be viewed as a probability measure on $(\Theta, \mathcal{B}_{\Theta})$.

Throughout the paper, we make the following assumptions. Assumption 1 (a) and (b) impose boundedness on function f and g_i's. Assumption 1 (c) states the ratio between f and g_i is uniformly bounded by q for all $\boldsymbol{x}$ and $\boldsymbol{\theta}$. Intuitively, it tells that for each unit consumption of resource, the maximum amount of revenue earned is upper bounded by q. In this paper, this condition will mainly be used to give an upper bound on the dual optimal solution. In Assumption 1 (d), we assume $\mathcal{P}_t$'s are independent of each other but we do not assume the exact knowledge of them. Also, there can be dependence between components in the vector-value functions $(f, \boldsymbol{g})$. In Assumption 1 (e), we require some convexity structure for the underlying functions. In the rest of the paper, this assumption will only be used to ensure that the Lagrangian problem $\max_{\boldsymbol{x} \in \mathcal{X}} \{f(\boldsymbol{x}; \boldsymbol{\theta}) - \boldsymbol{p}^{\top} g(\boldsymbol{x}; \boldsymbol{\theta})\}$ can be efficiently solved for any fixed $\boldsymbol{p} \geq \boldsymbol{0}$.

Assumption 1 (Boundedness and Independence) *We assume*

(a) $|f(\boldsymbol{x}; \boldsymbol{\theta})| \leq 1$ *for all* $\boldsymbol{x} \in \mathcal{X}, \boldsymbol{\theta} \in \Theta$.

(b) $g_i(\boldsymbol{x}; \boldsymbol{\theta}) \in [0, 1]$ *for all* $\boldsymbol{x} \in \mathcal{X}, \boldsymbol{\theta} \in \Theta$ *and* $i = 1, ..., m$. *In particular,* $\boldsymbol{0} \in \mathcal{X}$ *and* $g_i(\boldsymbol{0}; \boldsymbol{\theta}) = 0$ *for all* $\boldsymbol{\theta} \in \Theta$.

(c) *There exists a positive constant* q *such that for any* $\boldsymbol{\theta} \in \Theta$ *and each* i, *we have that* $f(\boldsymbol{x}; \boldsymbol{\theta}) \leq q \cdot g_i(\boldsymbol{x}; \boldsymbol{\theta})$ *holds for any* $\boldsymbol{x} \in \mathcal{X}$ *when* $g_i(\boldsymbol{x}; \boldsymbol{\theta}) > 0$.

(d) $\boldsymbol{\theta}_t \sim \mathcal{P}_t$ *and* $\mathcal{P}_t$'s *are independent with each others.*

(e) *The function* $f(\boldsymbol{x}; \boldsymbol{\theta})$ *is concave over* $\boldsymbol{x}$ *and the function* $g_i(\boldsymbol{x}; \boldsymbol{\theta})$ *is convex over* $\boldsymbol{x}$ *for any* $\boldsymbol{\theta} \in \Theta$ *and* $i = 1, \ldots, m$.

In the following, we illustrate the online formulation through two main application contexts: online linear programming and online network revenue management. We choose the more general convex formulation (PCP) to uncover the key mathematical structures for this online optimization problem, but we will occasionally return to these two examples to generate intuitions throughout the paper.

2.2. Examples

Online linear programming (LP): The online LP problem (Molinaro and Ravi, 2013; Agrawal et al., 2014; Gupta and Molinaro, 2014) can be viewed as an example of the online stochastic optimization formulation of (PCP). Specifically, the decision variable $x_t \in \mathcal{X} = [0, 1]$, the functions f and g_i are linear functions, and the parameter $\boldsymbol{\theta}_t = (r_t, \boldsymbol{a}_t)$ where $\boldsymbol{a}_t = (a_{1t}, ..., a_{mt})$. That is, $f(x_t; \boldsymbol{\theta}_t) = r_t x_t$ and $g_i(x_t; \boldsymbol{\theta}_t) = a_{it} x_t$. At each time $t = 1, ..., T$, the coefficient in the objective r_t together with the corresponding column $\boldsymbol{a}_t$ in the constraint matrix is revealed, and then one needs

to determine the value of x_t immediately. As mentioned earlier, the online LP problem covers a wide range of applications, through different specifications of $\boldsymbol{\theta}_t$, including secretary problem, online knapsack problem, resource allocation problem, generalized assignment problem, network routing problem, and matching problem.

Price-based network revenue management (NRM): In the price-based NRM problem (Gallego and Van Ryzin, 1994), a seller is selling a given stock of products over a finite time horizon by posting a price at each time. The demand is price-sensitive and the firm's objective is to maximize the total collected revenue. This problem could be cast in the formulation (PCP) as follows. The parameter θ_t refers to the type of the t-th arriving customer. There is usually a finite number of customer types, so the parameter set Θ is finite. The type of each customer arrival can be based on the side information such as demographic features and purchasing history. Each different customer type specifies a different demand function between the posted price and the realized demand/revenue. Specifically, the decision variable $\boldsymbol{x}_t$ represents to the price posted by the decision maker at time t. The constraint function $\boldsymbol{g}(\boldsymbol{x}_t; \theta_t)$ denotes the resource consumption (demand) under the price $\boldsymbol{x}_t$ and $f(\boldsymbol{x}_t; \theta_t)$ denotes the collected revenue. In such context, there is usually an extra layer of randomness for the functions $f(\boldsymbol{x}_t; \theta_t)$ and $\boldsymbol{g}(\boldsymbol{x}_t; \theta_t)$ given the parameter θ_t, different from our main setting where f and $\boldsymbol{g}$ are deterministic and known. We will discuss this extension in Section 6.4.

Choice-based network revenue management: In the choice-based NRM problem (Talluri and Van Ryzin, 2004), the seller offers an assortment of the products to the customer arriving in each time period, and the customer chooses a product from the assortment to purchase according to a given choice model. Now we discuss how our formulation (PCP) covers the choice-based NRM problem as a special case. As in the price-based NRM model, the parameter θ_t represents the customer type at time t, and the parameter set Θ is finite. Then we denote by $\mathcal{S}$ the set of all possible assortments the seller can offer to the customers. At each time t, the decision variable encodes the assortment decision, i.e., $\boldsymbol{x}_t = (x_{t,1}, \ldots, x_{t,|\mathcal{S}|}) \in [0,1]^{|\mathcal{S}|}$ and $\sum_{s \in \mathcal{S}} x_{t,s} = 1$. Here, $x_{t,s}$ denotes the probability that an assortment $s \in \mathcal{S}$ is offered at period t. The parameter θ_t specifies the choice model of the t-th customer arrival, and, together with the assortment decision $\boldsymbol{x}_t$, determines the revenue through the function $f(\boldsymbol{x}_t; \theta_t)$ and the resource consumption (demand) through the function $g(\boldsymbol{x}_t; \theta_t)$. As in the price-based NRM problem, there is an extra layer of randomness for the functions $f(\boldsymbol{x}_t; \boldsymbol{\theta}_t)$ and $\boldsymbol{g}(\boldsymbol{x}_t; \boldsymbol{\theta}_t)$ given the parameter θ_t and the assortment $\boldsymbol{x}_t$. We will also elaborate more in Section 6.4.

2.3. Performance Measure

We denote the offline optimal solution of optimization problem (CP) as $\boldsymbol{x}^* = (\boldsymbol{x}_1^*, ..., \boldsymbol{x}_T^*)$, and the offline (online) objective value as R_T^* (R_T). Specifically,

$$R_T^* := \sum_{t=1}^{T} f_t(\boldsymbol{x}_t^*)$$

$$R_T(\boldsymbol{\pi}) := \sum_{t=1}^{T} f_t(\boldsymbol{x}_t).$$

where the online objective value depends on the policy $\boldsymbol{\pi}$. Aligned with general online learning/optimization problems, we focus on minimizing the gap between the online and offline objective values. Specifically, the *optimality gap* is defined as follows:

$$\mathrm{Reg}_T(\mathcal{H}, \boldsymbol{\pi}) := R_T^* - R_T(\boldsymbol{\pi})$$

where the problem profile $\mathcal{H}$ encapsulates the realization of the random parameters, i.e., $\mathcal{H} := (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_T)$. Note that R_T^*, $R_T(\pi)$, $\boldsymbol{x}_t^*$ and $\boldsymbol{x}_t$ are all dependent on the problem profile $\mathcal{H}$, but we omit $\mathcal{H}$ in these terms for notation simplicity when there is no ambiguity. We define the performance measure of the online stochastic optimization problem formally as *regret*

$$\mathrm{Reg}_T(\boldsymbol{\pi}) := \max_{\mathcal{P} \in \Xi} \ \mathbb{E}_{\mathcal{H} \sim \mathcal{P}}[\mathrm{Reg}_T(\mathcal{H}, \boldsymbol{\pi})] \tag{3}$$

where $\mathcal{P} = (\mathcal{P}_1, ..., \mathcal{P}_T)$ denotes the probability measure of all time periods and the expectation is taken with respect to the parameter $\boldsymbol{\theta}_t \sim \mathcal{P}_t$; compactly, we write the problem profile $\mathcal{H} \sim \mathcal{P}$. We consider the worst-case regret for all the distribution $\mathcal{P}$ in a certain set Ξ where the set Ξ will be specified in later sections.

The specification of the set Ξ imposes more structure on the distributions of $(\mathcal{P}_1, ..., \mathcal{P}_T)$ and this is one of the main themes of our paper. In the canonical setting of online stochastic learning problem, all the distributions $\mathcal{P}_t$'s are the same, i.e., $\mathcal{P}_t = \mathcal{P}_0$ for $t = 1, ..., T$. Meanwhile, the adversarial setting of online learning problem refers to the case when $\mathcal{P}_t$'s are adversarially chosen. Our work aims to bridge these two ends of the spectrum with a novel notion of non-stationarity, and to relate the algorithm performance with certain structural property of $\mathcal{P} = (\mathcal{P}_1, ..., \mathcal{P}_T)$. In the same spirit, the work on non-stationary stochastic optimization (Besbes et al., 2015) proposes an elegant notion of non-stationarity called variation budget. Subsequent works consider similar notions in the settings of bandits (Besbes et al., 2014; Russac et al., 2019) and reinforcement learning (Cheung et al., 2019). To the best of our knowledge, all the previous works along this line consider unconstrained setting and thus our work contributes to this line of work in illustrating how the constraints interact with the non-stationarity.

Discussion on the (constrained) online learning/optimization literature.

Now we discuss the positioning of our work against the vast literature on the (constrained) online learning/optimization problem. Generally speaking, the formulations on this topic fall into two categories: (i) first-observe-then-decide and (ii) first-decide-then-observe. Our formulation belongs to the first category in that at each time t, the decision maker first observes the parameter $\boldsymbol{\theta}_t$ (and hence the functions $(f(\boldsymbol{x};\boldsymbol{\theta}_t), \boldsymbol{g}(\boldsymbol{x};\boldsymbol{\theta}_t))$), and then determines the value of $\boldsymbol{x}_t$. This formulation covers many applications in operations research and operations management, such as online LP and NRM. In these application contexts, the observations represent customers/orders arriving sequentially to the system, and the decision variables capture accordingly the acceptance/rejection/pricing decisions of the customers. Technically, the nature of having the observation before making the decision enables the stronger dynamic benchmark that allows a different $\boldsymbol{x}_t$ across different time periods as the definition of R_T^* in above.

One representative problem for the second category is the online convex optimization (OCO) problem. The OCO problem can be viewed as a first-decide-then-observe problem in that at each time t, the decision maker first chooses the decision variable $\boldsymbol{x}_t$ and then observes the function f_t (incurring a loss of $f_t(\boldsymbol{x}_t)$). The OCO problem is mainly motivated from machine learning applications such as online linear regression or online support vector machine (Hazan, 2016). From an information perspective, the OCO problem can be viewed as a partial information setting, whereas our online stochastic optimization can be viewed as a full information setting (Lattimore and Szepesvári, 2020). Accordingly, the standard OCO problem generally adopts the (weaker) static benchmark, i.e., $\boldsymbol{x}_t^*$'s need to be the same when defining R_T^*. This discrepancy in regret benchmark prevents direct applications of OCO algorithms and analyses to our context. There are results that consider a dynamic or partially dynamic regret benchmark for OCO in a non-stationary environment (Besbes et al., 2015) or an adversarial environment (Hall and Willett, 2013; Jadbabaie et al., 2015), but all these works consider the unconstrained problem. A line of works study the problem of online convex optimization with constraints (OCOwC) under a static or stochastic generation of the constraint functions $\boldsymbol{g}_t$'s. Specifically, Jenatton et al. (2016); Yuan and Lamperski (2018); Yi et al. (2021) all consider the OCOwC problem with static constraint ($\boldsymbol{g}_t = \boldsymbol{g}$ for some $\boldsymbol{g}$), while Neely and Yu (2017) mainly study a setting where $\boldsymbol{g}_t$ is i.i.d. generated from some distribution. To the best of our knowledge, no existing work along this line of literature allows non-stationarily generated constraint functions.

Another representative first-decide-then-observe problem is the bandits with knapsacks (BwK) problem which can also be viewed as a constrained online learning problem. The existing BwK results consider either a stochastic setting (Badanidiyuru et al., 2013; Agrawal and Devanur, 2014a) or an adversarial setting (Rangi et al., 2018; Immorlica et al., 2019). The algorithms along this

line of literature are mainly based on the underlying primal and dual LPs. Specifically, Agrawal and Devanur (2014b) develop a fast dual-based algorithm for the online stochastic optimization problem (first-observe-then-decide) and analyzes the algorithm performance under further stronger conditions on the dual optimal solution. Agrawal and Devanur (2014a) then extend to the first-decide-then-observe problem of BwK. The idea has been recently further applied to online learning in revenue management problems in (Miao et al., 2021). But these learning models and algorithms are developed in the stationary (stochastic) environment, which cannot be applied to the non-stationary setting.

We remark that both the first-decide-then-observe and the first-observe-then-decide frameworks can be useful in modeling some application context. For example, an Adwords problem under pay for conversions can be modeled by OCOwO or BwK problems, while an Adwords problem under pay for impressions is usually modeled by our online stochastic optimization framework or OLP problem (Mehta et al., 2005).

3. Known Distribution and Informative Gradient Descent

We begin our discussion with the case when the distributions $\mathcal{P}_t$'s are all known a priori. We use this case to motivate and present our prototypical algorithm – *informative gradient descent* which incorporates the prior information of $\mathcal{P}_t$'s with the online gradient descent algorithm. In the following sections, we will discuss the case when the distributions $\mathcal{P}_t$'s are not known precisely and analyze the algorithm performance accordingly.

3.1. Deterministic Upper Bound and Dual Problem

We first introduce the standard deterministic upper bound for the regret benchmark – the "offline" optimum $\mathbb{E}[R_T^*]$. We define the following expectation for a function $u(\boldsymbol{x}; \boldsymbol{\theta}) : \mathcal{X} \to \mathbb{R}$ and a probability measure $\mathcal{P}$ in the parameter space Θ,

$$\mathcal{P}u(\boldsymbol{x}(\boldsymbol{\theta}); \boldsymbol{\theta}) \coloneqq \int_{\boldsymbol{\theta}' \in \Theta} u(\boldsymbol{x}(\boldsymbol{\theta}'); \boldsymbol{\theta}') d\mathcal{P}(\boldsymbol{\theta}')$$

where $\boldsymbol{x}(\boldsymbol{\theta}) : \Theta \to \mathcal{X}$ is a measurable function. Thus $\mathcal{P}u(\cdot)$ can be viewed as a deterministic functional that maps function $\boldsymbol{x}(\boldsymbol{\theta})$ to a real value and it is obtained by taking expectation with respect to the parameter $\boldsymbol{\theta} \sim \mathcal{P}$.

Then, consider the following optimization problem

$$
\begin{aligned}
R_T^{\text{UB}} = \max \quad & \sum_{t=1}^{T} \mathcal{P}_t f(\boldsymbol{x}_t(\boldsymbol{\theta}_t); \boldsymbol{\theta}_t) \\
\text{s.t.} \quad & \sum_{t=1}^{T} \mathcal{P}_t g_i(\boldsymbol{x}_t(\boldsymbol{\theta}_t); \boldsymbol{\theta}_t) \leq c_i, \quad i = 1, ..., m, \\
& \boldsymbol{x}_t(\boldsymbol{\theta}_t) : \Theta \to \mathcal{X} \text{ is a measurable function for } t = 1, ..., T.
\end{aligned}
\tag{4}
$$

where $\boldsymbol{\theta}_t$ follows the distribution $\mathcal{P}_t$. The optimization problem (4) can be viewed as a convex relaxation of (PCP) where the objective and constraints are all replaced with their expected counterparts. Here $\boldsymbol{x}_{1:T} = (\boldsymbol{x}_1(\boldsymbol{\theta}_1), ..., \boldsymbol{x}_T(\boldsymbol{\theta}_T))$ encapsulates all the primal decision variables. The primal variables are expressed in a function form of $\boldsymbol{\theta}_t$ in that for each different $\boldsymbol{\theta}_t$, we allow a different choice of the primal variables. This is aligned with the "first-observe-then-decide" setting where the decision maker first observes the realization of $\boldsymbol{\theta}_t \sim \mathcal{P}_t$ and then decide the value of $\boldsymbol{x}_t$. As a standard result in literature (Gallego and Van Ryzin, 1994), Lemma 1 establishes the optimal objective value R_T^{UB} as an upper bound for $\mathbb{E}[R_T^*]$.

Lemma 1 *It holds that $R_T^{UB} \geq \mathbb{E}[R_T^*]$.*

The deterministic upper bound and the optimization problem (4) are commonly used to design algorithms in the literature. To proceed, we introduce the Lagrangian of (4),

$$L(\boldsymbol{p}, \boldsymbol{x}_{1:T}) := \boldsymbol{c}^\top \boldsymbol{p} + \sum_{t=1}^{T} \mathcal{P}_t \left(f(\boldsymbol{x}_t(\boldsymbol{\theta}_t); \boldsymbol{\theta}_t) - \boldsymbol{p}^\top \boldsymbol{g}(\boldsymbol{x}_t(\boldsymbol{\theta}_t); \boldsymbol{\theta}_t) \right)$$

where $\boldsymbol{\theta}_t$ follows the distribution $\mathcal{P}_t$. The (Lagrange multipliers) vector $\boldsymbol{p} = (p_1, ..., p_m)^\top$ conveys a meaning of shadow price for each budget, and $p_i \geq 0$ is the multiplier/dual variable associated with the i-th constraint. Furthermore, we define the following function based on a point-wise optimization for the primal variables,

$$h(\boldsymbol{p}; \boldsymbol{\theta}) := \max_{\boldsymbol{x}(\boldsymbol{\theta}) \in \mathcal{X}} \left\{ f(\boldsymbol{x}(\boldsymbol{\theta}); \boldsymbol{\theta}) - \boldsymbol{p}^\top \boldsymbol{g}(\boldsymbol{x}(\boldsymbol{\theta}); \boldsymbol{\theta}) \right\}.$$

Here the point-wise optimization emphasizes that the primal variables can be dependent on (and as a measurable function of) the parameter $\boldsymbol{\theta}_t$. For example, the pricing and assortment decisions can be made upon the observation of the customer type. Then the dual problem of (4) becomes

$$\min_{\boldsymbol{p} \geq \boldsymbol{0}} L(\boldsymbol{p}) := \boldsymbol{c}^\top \boldsymbol{p} + \sum_{t=1}^{T} \mathcal{P}_t h(\boldsymbol{p}; \boldsymbol{\theta}_t)$$

where $\boldsymbol{\theta}_t$ follows the distribution $\mathcal{P}_t$ as before.

Let $\boldsymbol{p}^*$ denote an optimal dual solution, i.e.,

$$\boldsymbol{p}^* \in \mathrm{argmin}_{\boldsymbol{p} \geq \boldsymbol{0}} L(\boldsymbol{p}) \tag{5}$$

and for each t,

$$\boldsymbol{\gamma}_t := \mathcal{P}_t \boldsymbol{g}(\boldsymbol{x}^*(\boldsymbol{\theta}); \boldsymbol{\theta}) \quad \text{where} \quad \boldsymbol{x}^*(\boldsymbol{\theta}) = \mathrm{argmax}_{\boldsymbol{x} \in \mathcal{X}} \{ f(\boldsymbol{x}; \boldsymbol{\theta}) - (\boldsymbol{p}^*)^\top \cdot \boldsymbol{g}(\boldsymbol{x}; \boldsymbol{\theta}) \}. \tag{6}$$

Here, $\boldsymbol{x}^*(\boldsymbol{\theta})$ is the associated primal (optimal) solution under the dual optimal solution $\boldsymbol{p}^*$, and $\boldsymbol{\gamma}_t$ can be interpreted as the expected budget consumption in the t-th time period under the optimal primal-dual pair $(\boldsymbol{x}^*(\boldsymbol{\theta}), \boldsymbol{p}^*)$. Accordingly, for each t, we define

$$L_t(\boldsymbol{p}) := \boldsymbol{\gamma}_t^\top \boldsymbol{p} + \mathcal{P}_t h(\boldsymbol{p}; \boldsymbol{\theta}).$$

The following proposition states the relation between $L(\cdot)$ and $L_t(\cdot)$ and establishes an upper bound for the benchmark using the dual problem.

Proposition 1 *For each $t = 1, ..., T$, it holds that*

$$\boldsymbol{p}^* \in \operatorname{argmin}_{\boldsymbol{p} \geq 0} L_t(\boldsymbol{p})$$

where $\boldsymbol{p}^*$ is defined in (5) *as one dual optimal solution. Moreover, we have*

$$\mathbb{E}[R_T^*] \leq \min_{\boldsymbol{p} \geq 0} \sum_{t=1}^T L_t(\boldsymbol{p}) = \sum_{t=1}^T \min_{\boldsymbol{p}_t \geq 0} L_t(\boldsymbol{p}_t).$$

The first part of the proposition says that all the L_t's share the same minimizer of $\boldsymbol{p}^*$ as the dual problem. In fact, this is the reason why we introduce $\boldsymbol{\gamma}_t$'s to define L_t's and the key how the prior knowledge of non-stationarity can be used for algorithm design. This property of a shared minimizer can be useful for online optimization in a non-stationary environment. First, we note that the primal optimal solution can be largely determined by the dual optimal solution $\boldsymbol{p}^*$. At each time t, the decision maker only has observations of the (random) realizations of L_s for $s = 1, ..., t$. Intuitively, the proposition implies that these past observations can be effectively used to estimate the dual optimal solution $\boldsymbol{p}^*$ as all the L_t's share the same optimal solution $\boldsymbol{p}^*$. Another advantage of this dual-based representation is that, through the adjustment of $\boldsymbol{\gamma}_t$'s, the dual optimal solution (of $L_t(\cdot)$) for each time period is the same, whereas the primal optimal solution $\boldsymbol{x}_t$ to (PCP) may be different from each t. As mentioned earlier, there are two types of regret benchmarks in the literature of online optimization: static oracle and dynamic oracle. The static oracle refers to the case where we compare to an offline decision maker adopting a common optimal solution throughout the entire horizon, and the dynamic oracle allows the offline decision maker to take an individual optimal solution for each time period. Proposition 1 states that in our setting, the static oracle and the dynamic oracle connect with each other in the dual space through a careful construction of $L_t(\cdot)$: the primal optimal solution is dynamic (different over time) while the dual optimal solution is static after the adjustment of $\boldsymbol{\gamma}_t$'s. This connection makes it possible to apply the gradient descent-based algorithms from OCO literature for the dual space in the non-stationary setting. Besbes et al. (2015) derive a similar argument to connect the dynamic oracle and static oracle for the unconstrained setting as a backbone for the algorithm design and regret analysis therein.

3.2. Main Algorithm and Regret Analysis

Now we present our main algorithm – *Informative Gradient Descent* – fully described as Algorithm 1. The algorithm is described as a meta algorithm with an input $\boldsymbol{\gamma}$. In the following sections, we will discuss how to apply the algorithm to different settings with different specifications of $\boldsymbol{\gamma}$. When the distributions $\mathcal{P}_t$'s are known, the algorithm is motivated from the dual-based representation in Proposition 1 and the input $\boldsymbol{\gamma}$ is accordingly defined by (6). Specifically, the algorithm maintains a dual vector/price $\boldsymbol{p}_t$, and at each time t, it performs a stochastic gradient descent update for $\boldsymbol{p}_t$ with respect to the function $L_t(\cdot)$. To see that the expectation of the dual gradient update (7) is the gradient with respect to the function $L_t(\cdot)$ evaluated at $\boldsymbol{p}_t$,

$$
\begin{aligned}
\mathbb{E}\left[\boldsymbol{g}(\tilde{\boldsymbol{x}}_t;\boldsymbol{\theta}_t)-\boldsymbol{\gamma}_t\right] &= -\boldsymbol{\gamma}_t + \mathcal{P}_t\boldsymbol{g}(\tilde{\boldsymbol{x}}_t;\boldsymbol{\theta}) \\
&= -\frac{\partial}{\partial\boldsymbol{p}}\left(\boldsymbol{\gamma}_t^\top\boldsymbol{p}+\mathcal{P}_t h(\boldsymbol{p};\boldsymbol{\theta})\right)\bigg|_{\boldsymbol{p}=\boldsymbol{p}_t}
\end{aligned}
$$

where the first line comes from taking expectation with respect to $\boldsymbol{\theta}_t$ and the second line comes from the definition of $\tilde{\boldsymbol{x}}_t$ in the algorithm. At each time t, the primal decision variable $\boldsymbol{x}_t$ is determined based on the dual price $\boldsymbol{p}_t$ and the observation $\boldsymbol{\theta}_t$ jointly, in the same manner as the definition of the function $h(\boldsymbol{p};\boldsymbol{\theta})$. Assumption 1 (e) ensures the optimization problem that defines $h(\boldsymbol{p};\boldsymbol{\theta})$ can be solved efficiently (See further discussion in Section A3).

We now provide an alternative perspective to interpret Algorithm 1 for the case when the distributions are known. Note that by definition (6), $\boldsymbol{\gamma}_t$'s represent the "optimal" way to allocate the resource budget over time according to the dual optimal solution $\boldsymbol{p}^*$. Specifically, a larger (resp. smaller) value of $\gamma_{i,t}$, where $\gamma_{i,t}$ denotes the i-th component of $\boldsymbol{\gamma}_t$, indicates that more (resp. less) budget should be allocated to time period t for constraint i. In Algorithm 1, from the update rule (7) of the dual variable at time period t, we know that if the budget consumption of constraint i is larger (resp. smaller) than $\gamma_{i,t}$, i.e., $g_i(\tilde{\boldsymbol{x}}_t;\boldsymbol{\theta}_t) > \gamma_{i,t}$ (resp. $g_i(\tilde{\boldsymbol{x}}_t;\boldsymbol{\theta}_t) < \gamma_{i,t}$), then we have that $p_{i,t+1} \le p_{i,t}$ (resp. $p_{i,t+1} > p_{i,t}$), where $p_{i,t}$ denotes the i-th component of $\boldsymbol{p}_t$. That is, if more (resp. less) budget is consumed in the earlier periods, then the dual price will be more likely to increase (resp. decrease), and consequently, less (resp. more) budget will be consumed in the future periods. In this way, the dual variable $\boldsymbol{p}_t$ dynamically balances the budget consumption: it ensures that for each t, the cumulative budget consumption of Algorithm 1 during the first t time periods always stay "close" to the optimal scheme of $\sum_{j=1}^t \boldsymbol{\gamma}_j$. Later in Section 6.2, we will show that a dynamic policy that incorporates the resource consumption process into the online decisions is necessary in a non-stationary environment, and any static policy can incur a linear regret even when the underlying non-stationarity intensity is sublinear.

Algorithm 1 Informative Gradient Descent Algorithm (IGD($\boldsymbol{\gamma}$))

1: Input: parameters $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_T)$.

2: Initialize the initial dual price $\boldsymbol{p}_1 = \boldsymbol{0}$ and initial constraint capacity $\boldsymbol{c}_1 = \boldsymbol{c}$

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

4:    Observe $\boldsymbol{\theta}_t$ and solve

$$\tilde{\boldsymbol{x}}_t = \operatorname{argmax}_{\boldsymbol{x} \in \mathcal{X}} \{ f(\boldsymbol{x}; \boldsymbol{\theta}_t) - \boldsymbol{p}_t^\top \cdot \boldsymbol{g}(\boldsymbol{x}; \boldsymbol{\theta}_t) \}$$

   where $\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{\theta}_t) = (g_1(\boldsymbol{x}, \boldsymbol{\theta}_t), \ldots, g_m(\boldsymbol{x}, \boldsymbol{\theta}_t))^\top$

5:    Set

$$\boldsymbol{x}_t = \begin{cases} \tilde{\boldsymbol{x}}_t, & \text{if } \boldsymbol{c}_t \text{ permits a consumption of } \boldsymbol{g}(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) \\ \boldsymbol{0}, & \text{otherwise} \end{cases}$$

6:    Update the dual price

$$\boldsymbol{p}_{t+1} = \left(\boldsymbol{p}_t + \frac{1}{\sqrt{T}} \left(\boldsymbol{g}(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) - \boldsymbol{\gamma}_t \right) \right) \vee \boldsymbol{0} \tag{7}$$

   where the element-wise maximum operator $u \vee v = \max\{v, u\}$

7:    Update the remaining capacity

$$\boldsymbol{c}_{t+1} = \boldsymbol{c}_t - \boldsymbol{g}(\boldsymbol{x}_t; \boldsymbol{\theta}_t)$$

8: **end for**

9: Output: $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$

Now we analyze the performance of Algorithm 1 for the known distribution setting. The following lemma says that the dual price vector $\boldsymbol{p}_t$ remains bounded throughout the procedure. Its proof largely relies on Assumption 1 (c), and also, the main usage of Assumption 1 (c) throughout our analysis is to ensure the boundedness of the dual vector.

Lemma 2 *Under Assumption 1, the dual price vector satisfies $\|\boldsymbol{p}_t\|_\infty \leq q + 1$ for $t = 1, 2, \ldots, T$. Here $\boldsymbol{p}_t$ is computed by (7) of IGD($\boldsymbol{\gamma}$) in Algorithm 1 with $\boldsymbol{\gamma}$ specified by (6), and the constant q is defined in Assumption 1 (c).*

The following theorem builds upon Proposition 1 and Lemma 2 and it states that the regret of Algorithm 1 is upper bounded by $O(\sqrt{T})$.

Theorem 1 *Under Assumption 1, if we consider the set $\Xi = \{ \mathcal{P} : \mathcal{P} = (\mathcal{P}_1, \ldots, \mathcal{P}_T), \forall \mathcal{P}_1, \ldots, \forall \mathcal{P}_T \}$, then the regret of IGD($\boldsymbol{\gamma}$) in Algorithm 1, with $\boldsymbol{\gamma}$ defined by (6), has the following upper bound*

$$Reg_T(\pi_{IGD}) \leq O(\sqrt{T})$$

where π_{IGD} stands for the policy specified by IGD(γ) in Algorithm 1.

The $O(\sqrt{T})$ regret upper bound of Algorithm 1 under a known non-stationary environment complements to several existing results in the literature. For the NRM problem, Talluri and Van Ryzin (1998) derives a $O(\sqrt{k})$ regret bound when the system size is scaled by k times, i.e., each period in the original problem is split into k statistically independent and identical periods and the capacities are scaled up k times. Following works subsequently improves this bound to $O(1)$ (e.g. (Reiman and Wang, 2008; Jasin and Kumar, 2012; Bumpensanti and Wang, 2020; Vera and Banerjee, 2020)). However, these methods are developed for a stationary setting under the finite support assumption over the distributions and do not account for the achievability of a sublinear regret under a non-stationary setting where the distribution at each period could be arbitrarily different from each other. Along this line, a recent work (Banerjee and Freund, 2020a) considers a setting that the length horizon T is random, but the paper still requires the underlying distribution to be finite-support and a "pseudo-stationary" environment where a concentration over time holds. The most similar result to Theorem 1 is from (Devanur et al., 2019) where the authors derives a $1 - 1/\sqrt{c_{\min}}$ competitive ratio under the non-stationary setting, where $c_{\min} = \min\{c_1, c_2, \ldots, c_m\}$ is the minimal budget and they assume that $f_t(\cdot)$ and $\boldsymbol{g}_t(\cdot)$ are all linear functions for each t. However, the competitive ratio result cannot be translated into a regret bound in our setting since we do not assume any relationship between the horizon length T and the initial budget $\boldsymbol{c}$. Specifically, the method therein is based on showing that the arrivals possess a concentration property and applying Chernoff-type inequality to derive high probability bounds on the event that all the constraints are not violated. In contrast, our method is based on using the dual variable $\boldsymbol{p}_t$ to balance the budget consumption on every sample path and showing that the dual variable is bounded over time according to our update rule (7). A recent work (Cheung et al., 2020) also derives an $O(\sqrt{T})$ regret bound for the setting of non-stationary environment with known distribution. The analysis therein builds upon a static dual-based algorithm; we defer more discussions to Section 6.2 and Section 7. In the following sections, we continue our pursuit and investigate how the IGD algorithm rolls out in a non-stationary environment when the true distribution is unknown.

Proposition 2 *Under Assumption 1, there is no algorithm that can achieve a regret better than $\tilde{\Omega}(\sqrt{T})$ where $\tilde{\Omega}$ hides a logarithmic term.*

Proposition 2 constructs a problem instance showing that even for a stationary setting where $\mathcal{P}_t$ for each t is identical to each other and known a priori, the lower bound of any online policy is $\tilde{\Omega}(\sqrt{T})$. For the problem instance, the underlying distribution $\mathcal{P}_t$ takes an infinite support. We note that when the parameter distribution has a finite support, an $O(1)$ regret bound can be derived following

the approach in Vera and Banerjee (2020); Banerjee and Freund (2020a,b), which implies that the gap between $O(\sqrt{T})$ and $O(1)$ is caused by whether the support of the parameter distribution is infinite or finite. In Section 6.1, we will further exploit the finite support structure and achieve better regret bound. The proof of Proposition 2 builds upon the analysis of Lemma 1 in Arlotto and Gurvich (2019). Different from the existing lower bound examples (Arlotto and Gurvich, 2019; Bumpensanti and Wang, 2020; Balseiro et al., 2020), the distribution of our problem instance bears no dependence on the horizon T; that is, the same static problem instance establishes the lower bound for all T. Theorem 1 and Proposition 2 altogether state that Algorithm 1 is optimal in a worst-case sense when no additional structure is imposed on $\mathcal{P}_t$'s.

4. Non-stationary Environment with Prior Estimate: Wasserstein Based Ambiguity and Analysis

In this section, we consider a "data-driven" setting where the true distribution is unknown, but a prior estimate of the true distribution is available. The setting relaxes the assumption on the exact knowledge of the true distribution in the last section. In practice, the availability of the prior estimate may characterize the predictable patterns of the non-stationarity in various application contexts. For example, the decision maker may not be able to foresee the future demand (distribution), but (s)he can construct some estimate based on history data or domain expertise of demand seasonality, the day-of-week effect, and demand surge due to pre-scheduled promotions or shopping festivals. When such prior estimate is accurate (the same as the true distribution), the setting of prior estimate in this section reduces to the discussion in the last section. However, when the prior estimate deviates from the true distribution, as is often the case in reality, then two natural questions are: (i) how to properly measure the inaccuracy of the prior estimate from the true distribution, (ii) how to design and analyze algorithm with such prior estimate. We answer these two questions in this section.

4.1. Wasserstein-Based Measure of Deviation

Consider the decision maker has a prior estimate/prediction $\hat{\mathcal{P}}_t$ for the true distribution $\mathcal{P}_t$ for each t, and all the predictions $\{\hat{\mathcal{P}}_t\}_{t=1}^T$ are made available at the very beginning of the procedure. We use the Wasserstein distance between $\hat{\mathcal{P}}_t$ and $\mathcal{P}_t$ to measure the deviation of the prior estimate from the true distribution. In following, we first formalize the definition and then discuss the suitability of the proposed Wasserstein-based measure.

The Wasserstein distance, also known as Kantorovich-Rubinstein metric or optimal transport distance (Villani, 2008; Galichon, 2018), is a distance function defined between probability distributions on a metric space. Its notion has a long history dating back over decades ago and

gains increasingly popularity in recent years with a wide range of applications including generative modeling (Arjovsky et al., 2017), robust optimization (Esfahani and Kuhn, 2018), statistical estimation (Blanchet et al., 2019), etc. In our context, the Wasserstein distance for two probability distributions $\mathcal{Q}_1$ and $\mathcal{Q}_2$ on the metric parameter space $(\Theta, \mathcal{B}_\Theta)$ is defined as follows,

$$\mathcal{W}(\mathcal{Q}_1, \mathcal{Q}_2) := \inf_{\mathcal{Q}_{1,2} \in \mathcal{J}(\mathcal{Q}_1, \mathcal{Q}_2)} \int \rho(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) d\mathcal{Q}_{1,2}(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) \tag{8}$$

where $\mathcal{J}(\mathcal{Q}_1, \mathcal{Q}_2)$ denotes all the joint distributions $\mathcal{Q}_{1,2}$ for $(\boldsymbol{\theta_1}, \boldsymbol{\theta_2})$ that have marginals $\mathcal{Q}_1$ and $\mathcal{Q}_2$. The distance function $\rho(\cdot, \cdot)$ is defined earlier in (2).

We define the following Wasserstein-based deviation budget (WBDB) to measure the cumulative deviation of the prior estimate,

$$\mathcal{W}_T(\mathcal{P}, \hat{\mathcal{P}}) := \sum_{t=1}^{T} \mathcal{W}(\mathcal{P}_t, \hat{\mathcal{P}}_t)$$

where $\mathcal{P} = (\mathcal{P}_1, ..., \mathcal{P}_T)$ denotes the true distribution and $\hat{\mathcal{P}} = (\hat{\mathcal{P}}_1, ..., \hat{\mathcal{P}}_T)$ denotes the prior estimate.

Based on the notion of WBDB, we define a set of distributions

$$\Xi_P(W_T) := \{\mathcal{P} : \mathcal{W}_T(\mathcal{P}, \hat{\mathcal{P}}) \le W_T, \mathcal{P} = (\mathcal{P}_1, ..., \mathcal{P}_T)\}$$

for a non-negative constant W_T, which we call as *deviation budget*. In this section, we consider a regret based on the set Ξ_P as defined in (3). In this way, the regret characterizes a worst-case performance of a certain algorithm for all the distributions $\mathcal{P} = (\mathcal{P}_1, ..., \mathcal{P}_T)$ within the set Ξ_P prescribed by some W_T. Specifically, the deviation budget W_T defines the set Ξ_P by inducing an upper bound for the deviation of prior estimate. Our next theorem provides an intuitive result that W_T is an inevitable loss (in terms of the algorithm regret) as a result of the inaccuracy of the prior estimate.

Theorem 2 *Under Assumption 1, if we consider the set $\Xi_P(W_T) := \{\mathcal{P} : \mathcal{W}_T(\mathcal{P}, \hat{\mathcal{P}}) \le W_T, \mathcal{P} = (\mathcal{P}_1, ..., \mathcal{P}_T)\}$, there is no algorithm that can achieve a regret better than $\Omega(\max\{\sqrt{T}, W_T\})$.*

Theorem 2 states that the lower bound of the regret is $\Omega(\max\{\sqrt{T}, W_T\})$. The theorem characterizes the best achievable algorithm performance under an inaccurate prior estimate, and precisely, the lower bound is linear in respect with the deviation of the prior estimate from the true distribution. The $\Omega(\sqrt{T})$ part inherits the result in Proposition 2 and it captures the intrinsic uncertainty of the underlying stochastic process over a time horizon T. The $\Omega(W_T)$ part captures the uncertainty arising from the inaccurate prior estimate.

4.2. Informative Gradient Descent Algorithm with Prior Estimate

Now we apply our informative gradient descent algorithm to the setting of prior estimate. A natural idea is to pretend that the prior estimate $\hat{\mathcal{P}}_t$ is indeed the true distribution $\mathcal{P}_t$. To implement the idea, we define

$$\hat{L}(\boldsymbol{p}) = \boldsymbol{c}^\top \boldsymbol{p} + \sum_{t=1}^{T} \hat{\mathcal{P}}_t h(\boldsymbol{p}; \boldsymbol{\theta})$$

where the true distribution $\mathcal{P}_t$ is replaced by its estimate $\hat{\mathcal{P}}_t$ for each component in function $L(\cdot)$. Thus it can be viewed as an approximation for the true dual function $L(\cdot)$ based on prior estimate. Let $\hat{\boldsymbol{p}}^*$ denote an optimal solution to $\hat{L}(\cdot)$,

$$\hat{\boldsymbol{p}}^* \in \operatorname{argmin}_{\boldsymbol{p} \geq 0} \hat{L}(\boldsymbol{p}) \tag{9}$$

and for each t, define

$$\hat{\boldsymbol{\gamma}}_t := \hat{\mathcal{P}}_t \boldsymbol{g}(\hat{\boldsymbol{x}}(\boldsymbol{\theta}); \boldsymbol{\theta}) \quad \text{where } \hat{\boldsymbol{x}}(\boldsymbol{\theta}) = \operatorname{argmax}_{\boldsymbol{x} \in \mathcal{X}} \{ f(\boldsymbol{x}; \boldsymbol{\theta}) - (\hat{\boldsymbol{p}}^*)^\top \cdot \boldsymbol{g}(\boldsymbol{x}; \boldsymbol{\theta}) \}. \tag{10}$$

Here, $\hat{\boldsymbol{\gamma}}_t$ denotes the "optimal" expected budget consumption in the t-th time under the prior estimate. In the setting of prior estimate, we do not have the exact knowledge of the true distributions $\mathcal{P}_t$'s and therefore $\boldsymbol{\gamma}_t$'s, so we alternatively use $\hat{\boldsymbol{\gamma}}_t$ as a substitute. Thus, the algorithm for the prior estimate setting, denoted by IGD($\hat{\boldsymbol{\gamma}}$), implements Algorithm 1 with the input $\hat{\boldsymbol{\gamma}}$ defined by (10). Specifically, the dual update step will become

$$\boldsymbol{p}_{t+1} = \left(\boldsymbol{p}_t + \frac{1}{\sqrt{T}} \left(\boldsymbol{g}(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) - \hat{\boldsymbol{\gamma}}_t \right) \right) \vee \boldsymbol{0} \tag{11}$$

in Algorithm 1.

4.3. Regret Analysis

The analysis of IGD($\hat{\boldsymbol{\gamma}}$) is slightly more complicated than that of IGD($\boldsymbol{\gamma}$) in theorem 1 because the algorithm is built upon the function $\hat{L}(\cdot)$ defined by the prior estimate instead of the true distribution. So we first study how to bound the difference between the function $\hat{L}(\cdot)$ and $L(\cdot)$ using the deviation between prior estimate and true distribution. For a probability measure $\mathcal{Q}$ over the metric parameter space $(\Theta, \mathcal{B}_\Theta)$, we denote

$$L_{\mathcal{Q}}(\boldsymbol{p}) := \mathcal{Q} h(\boldsymbol{p}; \boldsymbol{\theta}) = \int_{\boldsymbol{\theta}' \in \Theta} h(\boldsymbol{p}; \boldsymbol{\theta}') d\mathcal{Q}(\boldsymbol{\theta}').$$

Then the function $\hat{L}(\cdot)$ and $L(\cdot)$ can be expressed as

$$\hat{L}(\boldsymbol{p}) = \boldsymbol{p}^\top \boldsymbol{c} + \sum_{t=1}^{T} L_{\hat{\mathcal{P}}_t}(\boldsymbol{p}) \quad \text{and} \quad L(\boldsymbol{p}) = \boldsymbol{p}^\top \boldsymbol{c} + \sum_{t=1}^{T} L_{\mathcal{P}_t}(\boldsymbol{p})$$

Lemma 3 states that the function $L_{\mathcal{Q}}(\boldsymbol{p})$ has certain "Lipschitz continuity" in regard with the underlying distribution $\mathcal{Q}$. Specifically, the supremum norm between two functions $L_{\mathcal{Q}_1}(\boldsymbol{p})$ and $L_{\mathcal{Q}_2}(\boldsymbol{p})$ is bounded by the Wasserstein distance between two distributions $\mathcal{Q}_1$ and $\mathcal{Q}_2$ up to a constant dependent on the dimension and the boundedness of the function's argument.

Lemma 3 *For two probability measures $\mathcal{Q}_1$ and $\mathcal{Q}_2$ over the metric parameter space $(\Theta, \mathcal{B}_\Theta)$, we have that*

$$\sup_{\boldsymbol{p} \in \Omega_{\bar{p}}} |L_{\mathcal{Q}_1}(\boldsymbol{p}) - L_{\mathcal{Q}_2}(\boldsymbol{p})| \leq \max\{1, \bar{p}\} \cdot (m+1) \mathcal{W}(\mathcal{Q}_1, \mathcal{Q}_2) \tag{12}$$

where $\Omega_{\bar{p}} = [0, \bar{p}]^m$ and $\bar{p}$ is an arbitrary positive constant.

Note that the Lipschitz constant in Lemma 3 involves an upper bound of the function argument $\boldsymbol{p}$. The following lemma provides such an upper bound for the dual price $\boldsymbol{p}_t$ in IGD($\hat{\boldsymbol{\gamma}}$). The derivation is essentially the same as Lemma 2.

Lemma 4 *For each $t = 1, 2, \ldots, T$, we have that $\|\boldsymbol{p}_t\|_\infty \leq q + 1$ with probability 1, where $\boldsymbol{p}_t$ is specified by* (11) *in IGD($\hat{\boldsymbol{\gamma}}$).*

The rest of the analysis for IGD($\hat{\boldsymbol{\gamma}}$) is similar to that of IGD($\boldsymbol{\gamma}$) in Theorem 1. The regret of IGD($\hat{\boldsymbol{\gamma}}$) is formally stated in Theorem 3. Notably, the algorithm's regret matches the lower bound of Theorem 2 and thus it establishes the optimality of the algorithm.

Theorem 3 *Under Assumption 1, suppose a prior estimate $\hat{\mathcal{P}}$ is available and the regret is defined based on the set $\Xi_P(W_T)$, then the regret of IGD($\hat{\boldsymbol{\gamma}}$) has the following upper bound*

$$Reg_T(\pi_{IGDP}) \leq O(\max\{\sqrt{T}, W_T\})$$

where π_{IGDP} stands for the policy specified by IGD($\hat{\boldsymbol{\gamma}}$).

We remark the algorithm IGD($\hat{\boldsymbol{\gamma}}$) does not depend on or utilize the knowledge of the quantity W_T. On the upside, this avoids the assumption on the prior knowledge of W_T (as the knowledge of variation budget V_T (Besbes et al., 2014, 2015; Cheung et al., 2019)). On the downside, there is nothing the algorithm can do even when it knows a priori W_T is small or large. Technically, it means for our algorithm IGD($\hat{\boldsymbol{\gamma}}$), the WBNB contributes nothing in the dimension of algorithm design, and it will only influence the algorithm analysis. In particular, if we compare Theorem 3 with Theorem 1, the extra term W_T captures how the deviation of the prior estimate from the true distribution will deteriorate the performance of the gradient-based algorithm. When W_T is small, the $O(\sqrt{T})$ will be dominant and we do not need to worry about the deviation because its effect on the regret is secondary. In this light, the regret bound illuminates the effect of model misspecification/estimation error on the algorithm's performance in a non-stationary environment.

5. Non-stationary Environment Without Prior Estimate

In this section, we consider an uninformative setting where the true distribution is completely unknown to the decision maker. To one end, the discussion in this section can be viewed as a reduction of the results in the last section to a setting with an "uninformative" prior estimate. To the other, the uninformative setting draws an interesting comparison with the literature on (unconstrained) online learning/optimization in non-stationary environment (Besbes et al., 2014, 2015; Cheung et al., 2019) and its analysis exemplifies the interaction between the constraints and the non-stationarity.

5.1. Wasserstein-based Non-stationarity

We first illustrate how the non-stationarity over $\{\mathcal{P}_t\}_{t=1}^T$ interplays with the constraints through the following example adapted from (Golrezaei et al., 2014). The original usage of the example in their paper is to stress the importance of balancing resource consumption in an online setting. Specifically, consider the following two linear programs as the underlying problem (PCP) for two online stochastic optimization problems,

$$
\begin{aligned}
\max \quad & x_1 + ... + x_c + (1+\kappa)x_{c+1} + ... + (1+\kappa)x_T && (13) \\
\text{s.t.} \quad & x_1 + ... + x_c + x_{c+1} + ... + x_T \le c \\
& 0 \le x_t \le 1 \ \text{ for } t = 1, ..., T. \\
\max \quad & x_1 + ... + x_c + (1-\kappa)x_{c+1} + ... + (1-\kappa)x_T && (14) \\
\text{s.t.} \quad & x_1 + ... + x_c + x_{c+1} + ... + x_T \le c \\
& 0 \le x_t \le 1 \ \text{ for } t = 1, ..., T.
\end{aligned}
$$

where $\kappa \in (0,1)$, $c = \frac{T}{2}$ and the true distributions for both scenarios are point-mass distributions. Without loss of generality, we assume c is an integer. For the first LP (13), the optimal solution is to wait and accept the later half of the orders, while for the second LP (14), the optimal solution is to accept the first half of the orders and deplete the resource at half time. The contrast between these two LPs (two scenarios of whether the first half or the second half is more profitable) creates difficulty for the online decision making. Without knowledge of the future orders, there is no way we can obtain a sub-linear regret in both scenarios simultaneously. Because if we exhaust too much resource in the first half of the time, then for the first scenario (13), we do not have enough capacity to accept all the relatively profitable orders in the second half. On the contrary, if we have too much remaining resource at the half way, then for the second scenario (14), those relatively profitable orders that we miss in the first half are irrevocable.

An equivalent view of these two examples is to consider the existence of an adversary: the adversary is aware of the policy of the decision maker at the very beginning and then chooses the distribution $\mathcal{P}_t$ in an adversarial manner. Specifically, the adversary acts against us and aims to maximize the optimality gap between the offline optimal objective value and the online objective value. For example, in (13) and (14), the adversary can make a decision of which scenario for us to enter for the second half of the time based on our remaining inventory at the half way. The adversary view augments our previous interpretation of W_T as the maximal derivation (of the prior estimate from the true distribution): the regret definition based on Ξ_P in the last section can be viewed as a partially adversarial setting where the adversary chooses the true distribution against our will in a sequential manner but the choice of the distributions $\mathcal{P}_t$'s is subject to the set Ξ_P. Then the parameter W_T that defines the set Ξ_P serves as a measure of both the estimation error and the intensity of adversity.

Proposition 3 *The worst-case regret of online constrained stochastic optimization in an adversarial setting is $\Omega(T)$.*

Proposition 3 states that a fully adversarial setting where $\mathcal{P}_t$ can change arbitrarily over t does not permit a sub-linear regret. The same observation is also made in the literature (Besbes et al., 2014, 2015; Cheung et al., 2019) for unconstrained online learning problems where there is no function $\boldsymbol{g}(\cdot)$ and the decision $\boldsymbol{x}_t$ is made before the revealing of $f(\cdot)$. Specifically, Besbes et al. (2015) propose a novel measure of non-stationarity as follows (in the language of our paper),

$$V_T := \sum_{t=1}^{T-1} TV(\mathcal{P}_t, \mathcal{P}_{t+1})$$

where $TV(\cdot, \cdot)$ denotes the total variation distance between two distributions. The quantity V_T represents the cumulative *temporal change* of the distributions by comparing $\mathcal{P}_t$ and $\mathcal{P}_{t+1}$. Unfortunately, such temporal measure fails in the constrained setting. Note that for both (13) and (14), there is only one change point throughout the whole procedure thus the non-stationarity; their *temporal change* measure is $O(1)$ but a sub-linear regret is still unattainable.

Now, we propose the definition of the Wasserstein-based non-stationarity budget (WBNB) as

$$\mathcal{W}_T(\mathcal{P}) := \sum_{t=1}^{T} \mathcal{W}(\mathcal{P}_t, \bar{\mathcal{P}}_T)$$

where $\mathcal{P} = (\mathcal{P}_1, ..., \mathcal{P}_T)$ and $\bar{\mathcal{P}}_T$ is defined to be the uniform mixture distribution of $\{\mathcal{P}_t\}_{t=1}^{T}$, i.e., $\bar{\mathcal{P}}_T := \frac{1}{T} \sum_{t=1}^{T} \mathcal{P}_t$. The non-stationarity measure WBNB can be viewed as a degeneration of our previous deviation measure WBDB in that WBNB replaces all the prior estimates $\hat{\mathcal{P}}_t$'s with the uniform mixture $\bar{\mathcal{P}}_T$. The caveat is that in the uninformative setting, no distribution knowledge

is assumed, so the decision maker does not have access to $\bar{\mathcal{P}}_T$ unlike the prior estimate in the last section. As we will see shortly, the knowledge of $\bar{\mathcal{P}}_T$ does not affect anything in terms of the algorithm design and analysis.

Based on the notion of WBNB, we define a set of distributions

$$\Xi_U(W_T) = \{\mathcal{P} : \mathcal{W}_T(\mathcal{P}) \leq W_T, \mathcal{P} = (\mathcal{P}_1, ..., \mathcal{P}_T)\}$$

for a non-negative constant W_T, which we call as variation budget. Throughout this section, we consider a regret based on the set Ξ_U as defined in (3) in aim to characterize a "worst-case" performance of certain policy/algorithm for all the distributions in the set Ξ_U.

The variation budget W_T defines the uncertainty set Ξ_U by providing an upper bound on the non-stationarity of the distributions. Our next theorem states that it is impossible to get rid of W_T in the regret bound of any algorithm, which illustrates the sharpness of our definition of WBNB. Intuitively, it means that apart from the intrinsic stochasticity term $O(\sqrt{T})$, the (unknown) non-stationarity of the underlying distributions defined by WBNB appears to be a second bottleneck for algorithm performance. The proof of the theorem follows the same argument as Theorem 2.

Theorem 4 *Under Assumption 1, if we consider the set $\Xi_U(W_T) = \{\mathcal{P} : \mathcal{W}_T(\mathcal{P}) \leq W_T, \mathcal{P} = (\mathcal{P}_1, ..., \mathcal{P}_T)\}$, there is no algorithm that can achieve a regret better than $\Omega(\max\{\sqrt{T}, W_T\})$.*

5.2. Algorithm and Regret Analysis

One pillar of designing IGD(γ) and IGD($\hat{\gamma}$) is the budget allocation plan γ_t's (or $\hat{\gamma}_t$) prescribed by either the true distribution or the prior estimate. In the uninformative setting, the most straight-forward (and probably optimal) plan is to allocate the budget evenly over the entire horizon. Algorithm 2 – *uninformative gradient descent* (UGD) – implements the intuition by evenly distributing the budget without referring to any information. Thus the UGD algorithm can be viewed as IGD$\left(\frac{c}{T}\right)$. Returning to the point mentioned earlier on the knowledge of the centric distribution $\bar{\mathcal{P}}_T$, it does not matter we know it or not; because as long as all the prior estimate distributions $\hat{\mathcal{P}}_t$ are the same over time, we always have the same budget allocation plan. Furthermore, when all the $\mathcal{P}_t$'s are the same, which means the variation budget $W_T = 0$, Algorithm 2 and its analysis collapse into several recent studies on the gradient-based online algorithm under a stationary environment (Lu et al., 2020; Li et al., 2020).

Theorem 5 *Under Assumption 1, if we consider the set $\Xi_U(W_T) = \{\mathcal{P} : \mathcal{W}_T(\mathcal{P}) \leq W_T, \mathcal{P} = (\mathcal{P}_1, ..., \mathcal{P}_T)\}$, then the regret of Algorithm 2 has the following upper bound*

$$Reg_T(\pi_{UGD}) \leq O(\max\{\sqrt{T}, W_T\})$$

where π_{UGD} stands for the policy specified by Algorithm 2.

Algorithm 2 Uninformative Gradient Descent Algorithm (UGD)

1: Initialize the initial dual price $\boldsymbol{p}_1 = \boldsymbol{0}$ and initial constraint capacity $\boldsymbol{c}_1 = \boldsymbol{c}$.

2: **for** $t = 1, ..., T$ **do**

3:     Observe $\boldsymbol{\theta}_t$ and solve

$$\tilde{\boldsymbol{x}}_t = \operatorname{argmax}_{\boldsymbol{x} \in \mathcal{X}} \{ f(\boldsymbol{x}; \boldsymbol{\theta}_t) - \boldsymbol{p}_t^\top \cdot \boldsymbol{g}(\boldsymbol{x}; \boldsymbol{\theta}_t) \}$$

   where $\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{\theta}_t) = (g_1(\boldsymbol{x}, \boldsymbol{\theta}_t), ..., g_m(\boldsymbol{x}, \boldsymbol{\theta}_t))^\top$

4:     Set

$$\boldsymbol{x}_t = \begin{cases} \tilde{\boldsymbol{x}}_t, & \text{if } \boldsymbol{c}_t \text{ permits a consumption of } \boldsymbol{g}(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) \\ \boldsymbol{0}, & \text{otherwise} \end{cases}$$

5:     Update the dual price

$$\boldsymbol{p}_{t+1} = \left(\boldsymbol{p}_t + \frac{1}{\sqrt{T}} \left(\boldsymbol{g}(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) - \frac{\boldsymbol{c}}{T} \right) \right) \vee \boldsymbol{0} \tag{15}$$

   where the element-wise maximum operator $u \vee v = \max\{v, u\}$

6:     Update the remaining capacity

$$\boldsymbol{c}_{t+1} = \boldsymbol{c}_t - \boldsymbol{g}(\boldsymbol{x}_t; \boldsymbol{\theta}_t)$$

7: **end for**

8: Output: $\boldsymbol{x} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_T)$

Theorem 5 states the upper bound of Algorithm 2, which matches the regret lower bound in Theorem 4. Remarkably, the factors on T and W_T are additive in the regret upper bound of Algorithm 2. In comparison, the factor on T and the variation budget V_T are usually multiplicative in the regret upper bounds in the line of works that adopts the temporal change variation budget as nonstationary measure (Besbes et al., 2014, 2015; Cheung et al., 2019). The price of such an advantage for WBNB is that the WBNB is a more restrictive notion than the variation budget; recall that in (13) and (14), the temporal change variational budget is $O(1)$, but the WBNB is $O(\kappa T)$. Again, as the setting with prior estimate, the knowledge of the quantity W_T does not affect the algorithm design. By putting together Theorem 4 and Theorem 5, we argue that the knowledge of W_T does not help to further improve the algorithm performance.

6. Extensions and Discussions

6.1. Improved Regret Bound under Finite Support Assumption

In previous sections, we impose no additional structure on the underlying distributions $\mathcal{P}_t$'s other than the nonstationarity variation W_T. A natural question is whether a better regret bound is

achievable when the underlying distribution has more structures. In this subsection, we provide a positive answer to the question under an online LP formulation. Specifically, for the online linear programming problem, $\boldsymbol{\theta}_t = (r_t, \boldsymbol{a}_t)$ and $x_t \in [0, 1]$. The functions $f(x_t, \boldsymbol{\theta}_t) = r_t x_t$ and $g_i(x_t, \boldsymbol{\theta}_t) = a_{it} x_t$ are linear. We make the following assumption on the randomness of $\boldsymbol{\theta}_t$.

Assumption 2 *For each t, the distribution of $\boldsymbol{\theta}_t$ has a finite support, denoted by $\{\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(n)}\}$. Moreover, there exists $p_{\min} > 0$ such that for each $t = 1, \ldots, T$ and each $j = 1, \ldots, n$, it holds that $P(\boldsymbol{\theta}_t = \boldsymbol{\theta}^{(j)}) \geq p_{\min}$.*

Under this finite support assumption, the online LP problem is also known as the quantity based network revenue management problem and has been studied extensively in the literature. Here we consider a setting where the true distribution $\mathcal{P}_t$ that governs $\boldsymbol{\theta}_t$ is unknown but a prior estimate $\hat{\mathcal{P}}_t$ is available. Jasin and Kumar (2012); Vera et al. (2019) develop re-solving algorithms and derive bounded regret results when the underlying distribution is known, i.e., $\hat{P}_t$ is accurate. Compared to these existing works, our setting of prior estimate captures the potential misspecification or estimation error of the underlying distribution. Thus our analysis examines the robustness of the online algorithm against such error.

Next, we introduce an algorithm that extends the Fluid Bayes Selector algorithm in (Vera et al., 2019). Specifically, our algorithm uses the prior estimate instead of the true distribution for prescribing decisions. To describe the algorithm, we denote by $\boldsymbol{c}_t$ the remaining capacity at the beginning of each period t, and denote by $\mathcal{H}(t) = \{\boldsymbol{\theta}_t, \ldots, \boldsymbol{\theta}_T\}$ the (future) trajectory from period t to period T. Let $\mathcal{H}_j(t)$ be the number of times that $\boldsymbol{\theta}_\tau$ is realized as $\boldsymbol{\theta}^{(j)}$ for $\tau = t, \ldots, T$. Then, the trajectory $\mathcal{H}(t)$ can be equivalently represented by $\mathcal{H}(t) = (\mathcal{H}_1(t), \ldots, \mathcal{H}_n(t))$. The offline hindsight problem (PCP) with the remaining capacity $\boldsymbol{c}_t$ from period t to period T can be written as

$$R_t^*(\mathbf{c}_t, \mathcal{H}(t)) = \max \quad \sum_{j=1}^{n} r_j z_j,$$

$$\text{s.t.} \quad \sum_{j=1}^{n} a_{ji} z_j \leq c_{ti}, \quad \forall i = 1, \ldots, m,$$

$$0 \leq x_j \leq \mathcal{H}_j(t), \quad \forall j = 1, \ldots, n$$

where the decision variable z_j can be interpreted as the number of accepted orders of the j-th type, and each unit acceptance is associated with an reward r_j and a resource consumption vector of $(a_{j1}, ..., a_{jm})$. By taking expectation of the arrival counts $\mathcal{H}_j(t)$, an upper bound on the expected optimal objective value of the hindsight problem can be obtained as follows

$$R_t^{\mathrm{UB}}(\mathbf{c}_t, \mathcal{P}) = \max \quad \sum_{j=1}^{n} r_j z_j, \tag{16}$$

$$\text{s.t.} \quad \sum_{j=1}^{n} a_{ji} z_j \leq c_{ti}, \quad \forall i = 1, \ldots, m,$$

$$0 \leq z_j \leq \mathbb{E}_{\mathcal{H} \sim \mathcal{P}}[\mathcal{H}_j(t)], \quad \forall j = 1, \ldots, n,$$

where $\mathcal{P} = (\mathcal{P}_t, \ldots, \mathcal{P}_T)$ encapsulates the underlying distributions. The Fluid Bayes Selector in (Vera et al., 2019) solves the LP (16) and uses its optimal solution to guide the online decisions. As the resource capacity c_{it} changes over time and the LP (16) is solved at each time period, this type of algorithms is known as a re-solving algorithm.

When the true distribution $\mathcal{P}$ is unknown but some prior estimate $\hat{\mathcal{P}}$ is available, a natural idea is to calculate the right-hand-side of constraints (the expectations) in (16) using the prior estimate distributions. Specifically, in Algorithm 3, the decision maker first observes the parameter type j_t and refer to the optimal solution of (16) (using $\hat{\mathcal{P}}$ instead of $\mathcal{P}$ as its input) to choose the value of x_t.

Algorithm 3 Re-solving with Prior Estimate

1: Initialize constraint capacity $\boldsymbol{c}_1 = \boldsymbol{c}$.

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

3:      Solve $R_t^{\mathrm{UB}}(\boldsymbol{c}_t, \hat{\mathcal{P}})$ in (16) and obtain $\{\hat{z}_j(\boldsymbol{c}_t)\}_{j=1}^{n}$ as one optimal solution.

4:      Observe the realization of $\boldsymbol{\theta}_t$ as $\boldsymbol{\theta}^{(j_t)}$.

5:      **if** $\hat{z}_{j_t}(\boldsymbol{c}_t) \geq \frac{1}{2} \cdot \mathbb{E}_{\mathcal{H} \sim \hat{\mathcal{P}}}[\mathcal{H}_{j_t}(t)]$, **then** set $x_t = 1$.

6:      **else** set $x_t = 0$.

7:      **end if**

8: **end for**

9: Output: $\boldsymbol{x} = (x_1, \ldots, x_T)$

Theorem 6 *Suppose a prior estimate $\hat{\mathcal{P}}$ is available and the regret is defined based on the set $\Xi_P(W_T)$. Then, under Assumption 1 and Assumption 2, the regret of Algorithm 3 has the following upper bound*

$$Reg_T(\pi_{Resolve}) \leq O\left(\frac{1}{p_{\min}}(1 + W_T)\right)$$

where $\pi_{Resolve}$ stands for the policy specified by Algorithm 3.

Theorem 6 states the regret bound of Algorithm 3. We provide two ways to interpret the regret bound. First, when the prior estimate is accurate ($W_T = 0$), the algorithm and its analysis reduce to the results in (Vera and Banerjee, 2020). Compared to the $O(\sqrt{T})$ result in Section 3, the bounded regret here relies on the finite-supportedness and the existence of $p_{\min}$ in Assumption

2. Specifically, the assumption essentially requires the distributions $\mathcal{P}_t$ to be almost stationary over time (some concentration property holds) and thus guarantees that the budget consumption stays close to its expectation. Banerjee and Freund (2020a,b) further relax the assumption into a concentration condition and obtain more general results for problems such as online bin packing. When Assumption 2 or the concentration condition does not hold, the analyses in (Vera and Banerjee, 2020; Banerjee and Freund, 2020a,b) no longer work, but our regret bound of $O(\sqrt{T})$ in Section 3 is still valid as it does not impose any condition on $\mathcal{P}_t$'s. Second, when the prior estimate is inaccurate ($W_T > 0$), the regret bound in Theorem 6 captures the algorithm performance deterioration caused by the inaccurate prior estimation. Importantly, this term related to W_T is not reducible even with stronger condition of the functions f and g_i such as Assumption 2. In this sense, the result extends the existing ones which assumes the knowledge of true distributions (Vera and Banerjee, 2020; Banerjee and Freund, 2020a,b) to a setting with model (distribution) misspecification.

6.2. Sub-optimality of Static Policies

Both of our main algorithms – Algorithm 1 and Algorithm 2 are gradient based. Compared to Algorithm 3 and other existing re-solving algorithms (Jasin and Kumar, 2012; Bumpensanti and Wang, 2020; Vera and Banerjee, 2020), the gradient-based algorithms feature for simplicity and computational efficiency. In addition, the gradient-based algorithms have an adaptive and dynamic design that is crucial in stabilizing the resource consumption (i.e., not to exhaust the resource too early or have too much resource left-over). As discussed in Section 3.2, this is achieved inherently by using the realized resource consumption at each time period in the gradient update. We argue that such a dynamic design that relates the online decisions with the realized resource consumption process is necessary in achieving an optimal order of regret. In contrast, for a *static policy*, the online decisions can be dependent on the realized parameters $\boldsymbol{\theta}_t$'s but will not be affected by the dynamic of the resource consumption process. We remark that a static policy can utilize the prior estimate and be time-dependent; by "static", it means the policy remains the same regardless the realization of the resource consumption process. Examples of static policies include the bid-price policy (Talluri and Van Ryzin, 1998) and the offline-to-online policy (Cheung et al., 2020).

Definition 1 *A static policy π is described by a set of functions $\{h_t^\pi\}_{t=1}^T$, where $h_t^\pi : \Theta \to \mathcal{X}$ is allowed to be a random function. At each period t, given the type $\boldsymbol{\theta}_t$, the policy π will take the action $h_t^\pi(\boldsymbol{\theta}_t)$ if the budget constraints are not violated.*

Next, we illustrate the sub-optimality of any static policy's for both of the two settings. First, for the uninformative setting in Section 5, a static policy clearly cannot work since there is no

prior information that can be used to design the policy. Second, for the data-driven setting in Section 4, it is not as obvious whether a static policy can achieve the same order of regret optimality as the gradient-based algorithms. For example, what if we simply use the "offline" optimal dual solution $\boldsymbol{p}^*$ or $\hat{\boldsymbol{p}}^*$ to form a static decision rule throughout the procedure? This implements bid-price policy (Talluri and Van Ryzin, 1998) for the network revenue management problem under the known distribution setting. In Section D2.2, we show that this bid-price policy can incur a linear regret under an environment with slight non-stationarity (arbitrarily small W_T). The following proposition provides a more general statement on the sub-optimality of any static policy under a non-stationarity environment.

Proposition 4 *Suppose π is a static policy such that $\mathbb{E}_{\mathcal{H}\sim\hat{\mathcal{P}}}[R_T^{UB} - R_T(\pi)] \leq C_1 \cdot \sqrt{T}$ for some constant $C_1 > 0$, where $\hat{\mathcal{P}} = \{\hat{\mathcal{P}}_1, \ldots, \hat{\mathcal{P}}_T\}$ denotes the set of prior estimates. There exists a true distribution $\mathcal{P} = \{\mathcal{P}_1, \ldots, \mathcal{P}_T\}$ such that $\mathcal{W}_T(\mathcal{P}, \hat{\mathcal{P}}) \leq W_T = 4\sqrt{C_1} \cdot T^{3/4}$ and $\mathbb{E}_{\mathcal{H}\sim\mathcal{P}}[R_T^{UB} - R_T(\pi)] \geq C_2 \cdot T$ for some constant $C_2 > 0$.*

In the proposition, π denotes an arbitrary static policy that achieves an $O(\sqrt{T})$ regret when the prior estimate $\hat{\mathcal{P}}$ is accurate. But when there is a difference between the prior estimate $\hat{\mathcal{P}}$ and $\mathcal{P}$ and even if the deviation budget W_T is sublinear in T, the static policy π may still incur a linear regret for some problem instances. The implication is that for a static policy that works well under a known distribution setting, its performance can drastically deteriorate when there exists an estimation error or non-stationarity. Thus the dynamic design of gradient update or re-solving is both effective and necessary in overcoming the estimation error and the environment non-stationarity.

6.3. Advantage of Wasserstein distance

In the previous sections, we use Wasserstein distance to define both the deviation budget WBDN and the non-stationarity budget WBNB. We note that the analyses and regret bounds still hold if we change the underlying distance to total variation distance or KL-divergence. We choose the Wasserstein distance because it is a tighter measure than the total variation distance or the KL-divergence for defining WBNB. If we revisit the examples (13) and (14), a smaller value of $\kappa \in (0, 1)$ should indicate a smaller variation/non-stationarity between the first half and the second half of observations in both examples. However, the total variation distance fails to characterize this subtlety in that for any non-zero value of κ, the total variation distance between $\mathcal{P}_t$ and $\mathcal{P}_{t'}$ for $t \leq \frac{T}{2} < t'$ is always 1 (since $\mathcal{P}_t$ and $\mathcal{P}_{t'}$ have different supports). In other words, if we replace the Wasserstein distance with the total variation distance in our definition of WBNB, then the

quantity will always be $\frac{T}{2}$ for all $\kappa \in (0,1)$. Formally, for any two distributions $\mathcal{P}_1$ and $\mathcal{P}_2$, the following inequality holds (Gibbs and Su, 2002),

$$\mathcal{W}(\mathcal{P}_1, \mathcal{P}_2) \leq \mathrm{TV}(\mathcal{P}_1, \mathcal{P}_2).$$

Interestingly, this coincides with the intuitions in the literature of generative adversarial network (GAN) where Arjovsky et al. (2017) replace the KL-divergence with the Wasserstein distance in training GANs. Simultaneously and independently, Balseiro et al. (2020) analyze the dual mirror descent algorithm under a similar setting as our results in this section. The paper only discusses our uninformative setting, but not the known true distribution setting and the prior estimate setting. For the uninformative setting, their definition of non-stationarity is parallel to WBNB; both can be viewed as a reduction from the more general WBDB. The key difference is that Balseiro et al. (2020) from WBNB consider the total variation distance, which inherits the definition of variation functional from (Besbes et al., 2015).

A recent paper Cheung et al. (2020) proposes to measure the deviation budget as follows:

$$\Psi(\mathcal{P}, \hat{\mathcal{P}}) = \frac{1}{m+1} \cdot \sup_{\boldsymbol{x}(\boldsymbol{\theta}):\Theta \to \mathcal{X}} \left\| \sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{\theta}_t \sim \mathcal{P}_t} \left[(f(\boldsymbol{x}(\boldsymbol{\theta}_t), \boldsymbol{\theta}_t), \boldsymbol{g}(\boldsymbol{x}(\boldsymbol{\theta}_t), \boldsymbol{\theta}_t)) \right] - \mathbb{E}_{\boldsymbol{\theta}_t \sim \hat{\mathcal{P}}_t} \left[(f(\boldsymbol{x}(\boldsymbol{\theta}_t), \boldsymbol{\theta}_t), \boldsymbol{g}(\boldsymbol{x}(\boldsymbol{\theta}_t), \boldsymbol{\theta}_t)) \right] \right\|_2$$

where m is the number of resource constraints. The measure is used to analyze the proposed offline-to-online policy, and it captures the supremum distance between the true environment and the estimated environment under the same decision function $\boldsymbol{x}(\boldsymbol{\theta})$. Apart from the static policy's sub-optimality discussed earlier, the measure is also looser than the Wasserstein-based measure. Specifically, as the previous examples (13) and (14), if we choose the function $\boldsymbol{x}(\boldsymbol{\theta})$ such that $\boldsymbol{x}(\boldsymbol{\theta}) = 0$ when $\boldsymbol{\theta} = 1$ and $\boldsymbol{\theta} = 1 - \kappa$ (the reward r_t is 1 or $1 - \kappa$) and $\boldsymbol{x}(\boldsymbol{\theta}) = 1$ when $\boldsymbol{\theta} = 1 + \kappa$ (the reward r_t is $1 + \kappa$), then $\Psi(\mathcal{P}, \hat{\mathcal{P}})$ is always lower bounded by $\frac{T}{2(m+1)}$ regardless of the value of κ. In contrast, the Wasserstein distance is more sensitive and can capture the intensity of the parameter κ.

An additional practical benefit of Wasserstein distance is that the measure features for natural data-driven bounds. For example, the prior estimate $\hat{\mathcal{P}}_t$ can be constructed by the empirical distribution over N_t history samples over $\mathcal{P}_t$. Suppose that there exists a constant c_1 such that $\rho(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \leq c_1 \cdot \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2$. Then the following bound can be obtained for $\mathcal{W}(\mathcal{P}_t, \hat{\mathcal{P}}_t)$ (Theorem 1 of Fournier and Guillin (2015)), which explicitly relates the deviation (from $\hat{\mathcal{P}}_t$ to $\mathcal{P}_t$) with the number of history samples:

$$P\left(\mathcal{W}(\mathcal{P}_t, \hat{\mathcal{P}}_t) \leq \epsilon \right) \geq 1 - c_2 \cdot \exp(-c_3 \cdot N_t \epsilon^{\max\{K, 2\}})$$

where K is the dimension of the parameter $\boldsymbol{\theta}_t$, and c_2, c_3 are two positive constants. With $\epsilon = \frac{1}{\sqrt{T}}$ and $N_t = \frac{1}{c_3} \cdot T^{\frac{1}{2} \cdot \max\{K,2\}} \cdot \log(T/\delta)$, we have the following inequality holds

$$P\left(\mathcal{W}(\mathcal{P}_t, \hat{\mathcal{P}}_t) \leq \frac{1}{\sqrt{T}}\right) \geq 1 - \frac{c_2 \cdot \delta}{T}$$

holds for each t. Therefore, from union bound over $t = 1, \ldots, T$, we know that when $N_t = \frac{1}{c_3} \cdot T^{\frac{1}{2} \cdot \max\{K,2\}} \cdot \log(T/\delta)$, we have

$$W_T = \sum_{t=1}^{T} \mathcal{W}(\mathcal{P}_t, \hat{\mathcal{P}}_t) \leq \sqrt{T}$$

holds with probability at least $1 - c_2 \cdot \delta$.

6.4. Algorithm with Random Reward and Consumption

In this subsection, we consider another natural extension of our model where the reward and the budget consumption can be stochastic after the parameter $\boldsymbol{\theta}_t$ is revealed and the action $\boldsymbol{x}_t$ is made. To be specific, given each $\boldsymbol{x}$ and $\boldsymbol{\theta}$, the reward $f(\boldsymbol{x}; \boldsymbol{\theta})$ and the budget consumption $\boldsymbol{g}(\boldsymbol{x}; \boldsymbol{\theta})$ are all random variables. Such an extension allows us to cover the price-based NRM problem and the choice-based NRM problem first described in Section 2.2.

We note that the algorithms and results in the previous sections can be directly extended to this setting. We first illustrate the setting of known distribution described in Section 3 where the distributions $\mathcal{P}_t$'s are known a priori. For each $\boldsymbol{x}$ and $\boldsymbol{\theta}$, we introduce

$$\hat{f}(\boldsymbol{x}; \boldsymbol{\theta}) := \mathbb{E}[f(\boldsymbol{x}; \boldsymbol{\theta})] \text{ and } \hat{\boldsymbol{g}}(\boldsymbol{x}; \boldsymbol{\theta}) := \mathbb{E}[\boldsymbol{g}(\boldsymbol{x}; \boldsymbol{\theta})]$$

where the expectation is taken with respect the functions (f and $\boldsymbol{g}$) and conditional on $\boldsymbol{x}$ and $\boldsymbol{\theta}$. Then, we use $\hat{f}, \hat{\boldsymbol{g}}$ to revise the definition of the function h by

$$h(\boldsymbol{p}; \boldsymbol{\theta}) := \max_{\boldsymbol{x} \in \mathcal{X}} \{\hat{f}(\boldsymbol{x}; \boldsymbol{\theta}) - \boldsymbol{p}^T \hat{\boldsymbol{g}}(\boldsymbol{x}; \boldsymbol{\theta})\}.$$

Accordingly, the function L can be defined by the new function h

$$L(\boldsymbol{p}) := \boldsymbol{c}^T \boldsymbol{p} + \sum_{t=1}^{T} \mathcal{P}_t h(\boldsymbol{p}; \boldsymbol{\theta}) \tag{17}$$

and then the definition of $\boldsymbol{\gamma}$ is given by

$$\boldsymbol{\gamma}_t := \mathcal{P}_t \hat{\boldsymbol{g}}(\boldsymbol{x}^*(\boldsymbol{\theta}); \boldsymbol{\theta}) \text{ where } \boldsymbol{x}^*(\boldsymbol{\theta}) = \operatorname{argmax}_{\boldsymbol{x} \in \mathcal{X}} \{\hat{f}(\boldsymbol{x}; \boldsymbol{\theta}) - (\boldsymbol{p}^*)^\top \cdot \hat{\boldsymbol{g}}(\boldsymbol{x}; \boldsymbol{\theta})\} \tag{18}$$

where $\boldsymbol{p}^* \in \operatorname{argmin}_{\boldsymbol{p} \geq 0} L(\boldsymbol{p})$. Also, the benchmark is set as R_T^{UB} in (4) with the functions $f, \boldsymbol{g}$ replaced by their expectations $\hat{f}, \hat{\boldsymbol{g}}$. While implementing the algorithm IGD($\boldsymbol{\gamma}$) in Algorithm 1, the decision $\tilde{\boldsymbol{x}}_t$ is then set as

$$\tilde{\boldsymbol{x}}_t = \operatorname{argmax}_{\boldsymbol{x} \in \mathcal{X}} \{\hat{f}(\boldsymbol{x}; \boldsymbol{\theta}_t) - \boldsymbol{p}_t^T \cdot \hat{\boldsymbol{g}}(\boldsymbol{x}; \boldsymbol{\theta}_t)\}. \tag{19}$$

Note that (19) can be solved efficiently (see discussions on function h in Section A3). We have the following regret upper bound regarding the algorithm $IGD(\boldsymbol{\gamma})$ for this new setting with random reward and resource consumption. The result can be viewed as a generalization of Theorem 1.

Theorem 7 *Under Assumption 1, if we consider the set $\Xi = \{\mathcal{P} : \mathcal{P} = (\mathcal{P}_1, ..., \mathcal{P}_T), \forall \mathcal{P}_1, \ldots, \forall \mathcal{P}_T\}$, then the regret of $IGD(\boldsymbol{\gamma})$ has the following upper bound*

$$Reg_T(\pi_{IGD}) \leq O(\sqrt{T})$$

where π_{IGD} stands for the policy specified by $IGD(\boldsymbol{\gamma})$ with $\boldsymbol{\gamma}$ computed from (18).

For the data-driven setting with prior estimate, the distance between two parameters $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$ is adjusted by

$$\rho(\boldsymbol{\theta}, \boldsymbol{\theta}') := \sup_{\boldsymbol{x} \in \mathcal{X}} \|(\hat{f}(\boldsymbol{x}; \boldsymbol{\theta}), \hat{\boldsymbol{g}}(\boldsymbol{x}; \boldsymbol{\theta})) - (\hat{f}(\boldsymbol{x}; \boldsymbol{\theta}'), \hat{\boldsymbol{g}}(\boldsymbol{x}; \boldsymbol{\theta}'))\|_\infty, \tag{20}$$

and the definition of $\mathcal{W}(\mathcal{P}_t, \hat{\mathcal{P}}_t)$ follows. Then, let the parameters $\hat{\boldsymbol{\gamma}}$ be computed from (10) with respect to $\hat{f}, \hat{\boldsymbol{g}}$. As a corollary of Theorem 7, we have the following regret bound under the data-driven setting.

Corollary 1 *Under Assumption 1, suppose a prior estimate $\hat{\mathcal{P}}$ is available and the regret is defined based on the set $\Xi_P(W_T) = \{\mathcal{P} : \sum_{t=1}^T \mathcal{W}(\mathcal{P}_t, \hat{\mathcal{P}}_t) \leq W_t\}$, where the distance $\mathcal{W}(\mathcal{P}_t, \hat{\mathcal{P}}_t)$ is defined following the distance in (20), then the regret of $IGD(\hat{\boldsymbol{\gamma}})$ has the following upper bound*

$$Reg_T(\pi_{IGDP}) \leq O(\max\{\sqrt{T}, W_T\})$$

where π_{IGDP} stands for the policy specified by $IGD(\hat{\boldsymbol{\gamma}})$.

For the uninformative setting where the prior estimates are not available, it is direct to see that the regret bound in Corollary 1 continues to hold as long as WBNB is defined following the distance in (20).

6.5. Algorithm with Infrequent Update

In this subsection, we study the performance of our algorithm with an infrequent update scheme. Specifically, the infrequent update refers to that the dual update step (7) for Algorithm 1 is not done at each period, instead, the dual variable vector $\boldsymbol{p}$ is updated only after K periods. In this way, our algorithm can be viewed as a batched algorithm, where the arrivals from period t to period $t + K$, for some t, are viewed as a batch of samples to update the dual variable. Note that the dual vector determines the online decision rule of the primal variable $\boldsymbol{x}_t$. An infrequent scheme has the practical benefits of inducing more consistency in the decision rule over time.

Algorithm 4 Batched Informative Gradient Descent Algorithm (B-IGD($\boldsymbol{\gamma}, K, \alpha_{T,K}$))

1: Input: parameters $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_T)$, the batch size K, and the step size $\alpha_{T,K}$.

2: Initialize the initial dual price $\boldsymbol{p}_1 = \boldsymbol{0}$ and initial constraint capacity $\boldsymbol{c}_1 = \boldsymbol{c}$

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

4:     Specify the constant $l = \lfloor t/K \rfloor + 1$, where $\lfloor \cdot \rfloor$ denotes the floor function.

5:     Observe $\boldsymbol{\theta}_t$ and solve

$$\tilde{\boldsymbol{x}}_t = \mathrm{argmax}_{\boldsymbol{x} \in \mathcal{X}} \{ f(\boldsymbol{x}; \boldsymbol{\theta}_t) - \boldsymbol{p}_l^\top \cdot \boldsymbol{g}(\boldsymbol{x}; \boldsymbol{\theta}_t) \}$$

   where $\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{\theta}_t) = (g_1(\boldsymbol{x}, \boldsymbol{\theta}_t), \ldots, g_m(\boldsymbol{x}, \boldsymbol{\theta}_t))^\top$

6:     Set

$$\boldsymbol{x}_t = \begin{cases} \tilde{\boldsymbol{x}}_t, & \text{if } \boldsymbol{c}_t \text{ permits a consumption of } \boldsymbol{g}(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) \\ \boldsymbol{0}, & \text{otherwise} \end{cases}$$

7:     If $t = l \cdot K$, then we update the dual price

$$\boldsymbol{p}_{l+1} = \left(\boldsymbol{p}_l + \alpha_{T,K} \cdot \sum_{\tau = (l-1)K+1}^{lK} (\boldsymbol{g}(\tilde{\boldsymbol{x}}_\tau; \boldsymbol{\theta}_\tau) - \boldsymbol{\gamma}_\tau) \right) \vee \boldsymbol{0} \qquad (21)$$

   where the element-wise maximum operator $u \vee v = \max\{v, u\}$

8:     Update the remaining capacity

$$\boldsymbol{c}_{t+1} = \boldsymbol{c}_t - \boldsymbol{g}(\boldsymbol{x}_t; \boldsymbol{\theta}_t)$$

9: **end for**

10: Output: $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$

The algorithm is formally described in Algorithm 4. It takes the same structure as Algorithm 1 and Algorithm 2 but only updates the dual variables every K time periods. We have the following regret bound for Algorithm 4 for the known distribution setting.

Theorem 8 *Under Assumption 1, if we consider the set $\Xi = \{\mathcal{P} : \mathcal{P} = (\mathcal{P}_1, \ldots, \mathcal{P}_T), \forall \mathcal{P}_1, \ldots, \forall \mathcal{P}_T\}$, then the regret of B-IGD($\boldsymbol{\gamma}, K, \alpha_{T,K}$), with $\boldsymbol{\gamma}$ computed from (6), has the following upper bound*

$$Reg_T(\pi_{B\text{-}IGD}) \leq O\left(TK\alpha_{T,K} + \frac{1}{\alpha_{T,K}} + K \right)$$

where $\pi_{B\text{-}IGD}$ stands for the policy specified by IGD($\boldsymbol{\gamma}$) in Algorithm 4.

With a choice of $\alpha_{T,K} = \frac{1}{\sqrt{TK}}$, we obtain an regret upper bound of $O(\sqrt{TK})$ for the known distribution setting. Following the same spirit, for the data-driven setting, we obtain the following regret bound for Algorithm 4.

Corollary 2 *Suppose a prior estimate $\hat{\mathcal{P}}$ is available and the regret is defined based on the set $\Xi_P(W_T)$. Then, under Assumption 1, the regret of B-IGD($\hat{\gamma}, K, \alpha_{T,K}$), with $\hat{\gamma}$ computed from (10), has the following upper bound*

$$Reg_T(\pi_{B\text{-}IGD}) \leq O(TK\alpha_{T,K} + \frac{1}{\alpha_{T,K}} + K + W_T)$$

where $\pi_{B\text{-}IGD}$ stands for the policy specified by B-IGD($\hat{\gamma}, K, \alpha_{T,K}$).

With a choice of $\alpha_{T,K} = \frac{1}{\sqrt{TK}}$, we obtain the regret bound $O(\max\{\sqrt{TK}, W_T\})$ in Corollary 2 under the data-driven setting. For the uninformative setting, it is clear to see that the regret bound in Corollary 2 still holds when W_T refers to the WBNB. For both regret bounds in Theorem 8 and Corollary 2, the additional factor of $\sqrt{K}$ is the price paid for the infrequent update.

7. Numerical Experiments

7.1. Experiment I: Online Linear Programming

We first present some synthetic experiments for the setting of online LP where both the reward and cost function are linear, i.e., $f_t(x) = r_t x$ and $g_{it}(x) = a_{it} x$ for $i = 1, ..., m$ and $t = 1, ..., T$. Suppose that both the true distribution and the prior estimate distribution of a_{it} follow Uniform[0.1, 1.1] throughout the entire horizon. We consider three different settings for the true distribution and the prior estimate distribution of r_t (summarized in Table 1). The parameters α and β are to be specified: α reflects the intensity of the non-stationarity over time and β represents the error of the prior estimate. Specifically, the deviation budget WBDB (in Section 4) grows linearly with β, while the non-stationarity budget WBNB (in Section 5) grows linearly with α.

The first setting considers a uniform distribution for r_t: the true distribution of r_t follows Uniform[0, 1] for the first half of the time and Uniform[0, α] for the second half, while the prior estimate distribution follows Uniform[0, 1 + β] for the first half of the time and Uniform[0, α + β] for the second half.

The second setting considers a truncated normal distribution for r_t: the true distribution of r_t follows Normal(1, 1) for the first half of the time and Normal(α, 1) for the second half, while the prior estimate distribution follows Normal(1 + β, 1) for the first half of the time and Normal(α + β, 1) for the second half. All the normal distributions are made non-negative by truncating at 0.

The third setting considers a mixture of the previous two distributions of uniform and truncated normal: the true distribution of r_t follows a uniform mixture of Uniform[0, 1] and Normal(1, 1) for the first half of the time and a uniform mixture of Uniform[0, α] and Normal(α, 1) for the second half, while the prior estimate distribution follows a uniform mixture of Uniform[0, 1 + β] and Normal(1 + β, 1) for the first half of the time and a uniform mixture of Uniform[0, α + β] and

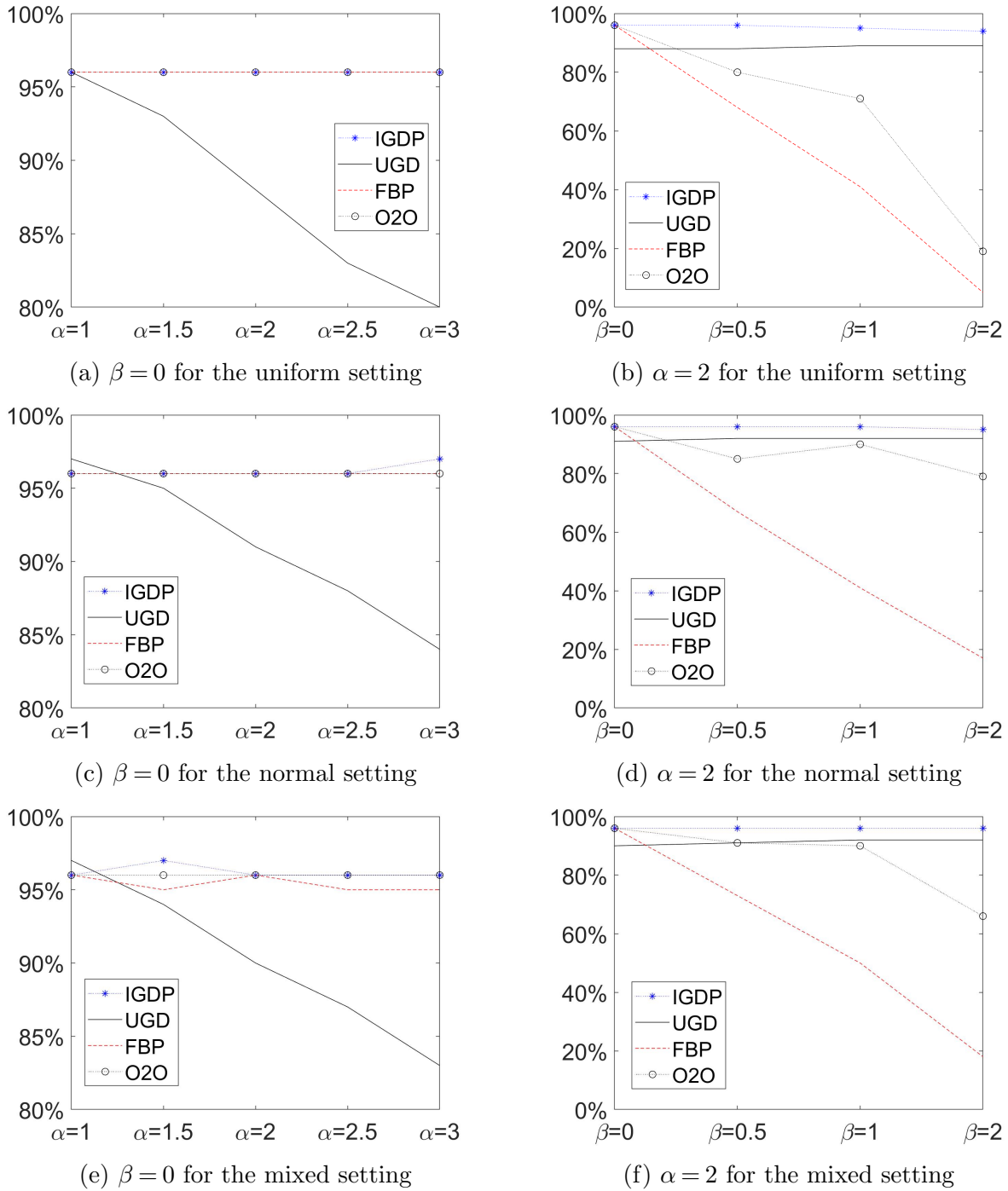|		True Distribution		Prior Estimate	
	Periods	r_t	a_{it}	$\hat{r}_t$	$\hat{a}_{it}$
The uniform setting	$t = 1, \ldots, \lfloor \frac{T}{2} \rfloor$	$\text{Unif}[0,1]$	$\text{Unif}[0.1, 1.1]$	$\text{Unif}[0, 1+\beta]$	$\text{Unif}[0.1, 1.1]$
	$t = \lfloor \frac{T}{2} \rfloor + 1, \ldots, T$	$\text{Unif}[0, \alpha]$	$\text{Unif}[0.1, 1.1]$	$\text{Unif}[0, \alpha + \beta]$	$\text{Unif}[0.1, 1.1]$
The normal setting	$t = 1, \ldots, \lfloor \frac{T}{2} \rfloor$	$\text{Norm}(1,1)$	$\text{Unif}[0.1, 1.1]$	$\text{Norm}(1+\beta, 1)$	$\text{Unif}[0.1, 1.1]$
	$t = \lfloor \frac{T}{2} \rfloor + 1, \ldots, T$	$\text{Norm}(\alpha, 1)$	$\text{Unif}[0.1, 1.1]$	$\text{Norm}(\alpha + \beta, 1)$	$\text{Unif}[0.1, 1.1]$
The mixed setting	$t = 1, \ldots, \lfloor \frac{T}{2} \rfloor$	$\text{Mix}(1)$	$\text{Unif}[0.1, 1.1]$	$\text{Mix}(\beta)$	$\text{Unif}[0.1, 1.1]$
	$t = \lfloor \frac{T}{2} \rfloor + 1, \ldots, T$	$\text{Mix}(\alpha)$	$\text{Unif}[0.1, 1.1]$	$\text{Mix}(\alpha + \beta)$	$\text{Unif}[0.1, 1.1]$

Table 1   Input of Experiment I: $\text{Unif}[a, b]$ denotes the uniform distribution over the interval $[a, b]$. $\text{Norm}(a, b)$ denotes a normal distribution with mean a and standard deviation b, truncated at 0. $\text{Mix}(a)$ denotes a uniform mixture of $\text{Unif}[0, a]$ and $\text{Norm}(a, 1)$.

$\text{Normal}(\alpha + \beta, 1)$ for the second half. As the second setting, all the normal distributions are made non-negative by truncating at 0.

For the numerical experiments, we implement Algorithm 1 – IGD($\hat{\boldsymbol{\gamma}}$) and Algorithm 2 (UGD) under different choices of α and β to study the algorithm performance under both settings in our paper. Besides, we implement the classic fixed bid price control heuristics (FBP) proposed in Talluri and Van Ryzin (1998). The FBP method uses $\hat{\boldsymbol{p}}^*$ computed from the prior estimate (9) as the bid price. It then accepts the order (setting $x_t = 1$) when $r_t \geq \boldsymbol{a}_t^\top \hat{\boldsymbol{p}}^*$ and there is enough resource; otherwise, it will reject the order (setting $x_t = 0$). We also compare with the offline-to-online (O2O) algorithm proposed in Cheung et al. (2020). The O2O algorithm is a dual-based online algorithm for a non-stationary setting. It first runs an offline procedure to construct a set of dual variables using prior estimates $\{\hat{\mathcal{P}}_t\}_{t=1}^T$. Then, at each period t, after observing the realized parameter $\boldsymbol{\theta}_t$, the O2O algorithm draws a dual variable $\boldsymbol{p}$ uniformly randomly from the set, and then implements $\tilde{\boldsymbol{x}}_t = \text{argmax}_{\boldsymbol{x} \in \mathcal{X}} \{f(\boldsymbol{x}; \boldsymbol{\theta}_t) - \boldsymbol{p}^\top \cdot \boldsymbol{g}(\boldsymbol{x}; \boldsymbol{\theta}_t)\}$. Note that the O2O algorithm is a static policy discussed in Section 6.2. In our implementation, we select different step sizes to generate multiple sets of dual variables. Then, we apply each set of dual variables to the online problem and we select the best reward to report the performance of the O2O algorithm.

Figure 1 and Tables 3-5 report the performances of the four algorithms with the horizon $T = 1000$, number of constraints $m = 10$, and initial resource capacity $c_i = 200$ for each $i = 1, \ldots, m$.

In Figure 1, the case of fixed α and varying β examines the robustness of an algorithm to prior estimation error (different β) under a nonstationary environment ($\alpha > 0$). In comparison, Algorithm UGD and IGDP have significant better performance than FBP and O2O. This shows the effectiveness of the gradient-based dynamic update and distinguishes our algorithms of UGD and IGDP from static policies such as FBP and O2O. Specifically, the performances of FBP and O2O deteriorate quickly as the estimation error β increases, while both gradient-based algorithms remain stable. Note that FBP computes a fixed dual price vector and O2O constructs a set of fixed

(a) $\beta = 0$ for the uniform setting

(b) $\alpha = 2$ for the uniform setting

(c) $\beta = 0$ for the normal setting

(d) $\alpha = 2$ for the normal setting

(e) $\beta = 0$ for the mixed setting

(f) $\alpha = 2$ for the mixed setting

Figure 1 The percentage of the total reward collected by IGDP, UGD, O2O and FBP over the upper bound for different α and different β under different setting.

dual vectors completely based on the prior estimates. However, as the estimation errors increase, the gap between the "true" dual vector and the pre-computed dual vector(s) in FBP or O2O will also increase, and the static nature of the two algorithms prevents a dynamic correction of the

dual vector. As a result, the performances of FBP and O2O become worse as the estimation error increases. In contrast, the dynamic dual update of UGD and IGDP naturally combines the demand realization and the prior estimates (through parameters $\hat{\boldsymbol{\gamma}}$). Therefore, the performance of IGDP remains stable as the estimation errors increase. These numerical evidence reinforces our previous theoretical arguments on the sub-optimality of static policies in Section 6.2.

In Figure 1, the case of varying α and $\beta = 0$ corresponds to the setting of Section 5 where the value of α indicates the intensity of the non-stationarity for the underlying distribution $\mathcal{P}_t$'s. The performance of UGD deteriorates as α increases, and thus it validates the role of WBNB in characterizing the algorithm performance. The algorithms IGDP, O2O and FBP provide unsurprisingly better and more stable performance because all of them utilize the prior estimate (which is exactly the true distribution when $\beta = 0$). This comparison highlights the effectiveness of WBNB in characterizing the learnability of a non-stationary environment when there is no prior knowledge available (as for UGD), and also it underscores the usefulness of prior knowledge.

Based on these experiments, we make the following remarks: First, both of FBP and O2O algorithms can deal with the non-stationarity to some extent (as in Figure 1(a)). However, their performances highly depends on the accuracy of the prior estimates. If the deviation of the prior estimates from the true distributions is large, both performances can become very poor. Second, the performance of UGD is robust to the deviation of prior estimate since it does not utilize any prior estimate. Meanwhile, the downside is that UGD is more sensitive to the intensity of non-stationarity compared to the other two algorithms. Third, the performance of IGDP is relatively more robust to both the non-stationarity and the deviation of the prior estimates. Specifically, IGDP obtains parameters $\boldsymbol{\gamma}_t$ for each t to handle the non-stationarity and also utilizes the gradient descent updates in a dynamic way to hedge against the deviation. In this light, IGDP combines both the advantages of FBP, UGD and O2O.

7.2. Experiment II: Network Revenue Management with Resolving Heuristics

Our second numerical experiment is adapted from the network revenue management experiment in Jasin (2015). We consider a hub-and-spoke model with 8 cities, 14 connecting flights, and 41 itineraries. The detailed itinerary structure and the normalized capacities in our experiment are referred to Table 1 and Table 2 in Jasin (2015), respectively. To model the non-stationarity, we add random noises to the arrival probabilities of the itineraries. Specifically, we denote $\mathcal{P}_0 \in \mathbb{R}^{41}$ as the arrival probabilities in Jasin (2015) and we set $\alpha \in [0, 1]$ as a parameter, which will be specified later. For each t, the arrival probabilities are set as

$$\mathcal{P}_t = (\mathcal{P}_0 + \alpha \cdot \mathrm{Unif}(41, 1))/\mathrm{sum}(\mathcal{P}_0 + \alpha \cdot \mathrm{Unif}(41, 1))$$

where $\mathrm{Unif}(41,1)$ denotes a 41-dimensional vector of i.i.d. uniformly distributed random variables over $[0,1]$ and $\mathrm{sum}(\cdot)$ denotes the summation of all the components of a vector. For each time t, an independent random noise vector will be generated. In our experiment, the probability vector $\mathcal{P}_t$ is only generated once at the very beginning, and it will remain the same in all the simulation trials. Intuitively, the value of α represents the magnitude of adjustment compared to the original probability $\mathcal{P}_0$, and thus it reflects the intensity of non-stationarity of the underlying distributions. Similarly, we introduce another parameter β to reflect the deviation of the prior estimates $\hat{\mathcal{P}}_t$ from the true distribution $\mathcal{P}_t$. Let

$$\hat{\mathcal{P}}_t = (\mathcal{P}_t + \beta \cdot \mathrm{Unif}(41,1))/\mathrm{sum}(\mathcal{P}_t + \beta \cdot \mathrm{Unif}(41,1)).$$

We report the algorithm performance for different (α, β) in Table 2.

In the table, we also implement a re-solving version of $\mathrm{IGD}(\hat{\gamma})$ which periodically invokes a re-solving procedure based on the remaining resource and the prior estimate to the dual updates in $\mathrm{IGD}(\hat{\gamma})$. Specifically, the re-solving heuristic re-solves the upper bound function $\hat{L}(\cdot)$ based on the current time period t and the remaining budget c_t to obtain an updated p_t on a regular basis. For each time $t \in \mathcal{T} \subset \{1, ..., T\}$, we solve the following problem

$$\hat{p}_t^* = \mathrm{argmin}_{p \geq 0} c_t^\top p + \sum_{j=t}^{T} \hat{\mathcal{P}}_j h(p, \theta)$$

and use its optimal solution as the dual price for the t-th time period. Also, we update $\hat{\gamma}_t$'s accordingly. Here the set $\mathcal{T}$ contains the time periods that we will re-solve the problem. For each time $t \in \mathcal{T}$, the preceding dual price p_{t-1} will be discarded and after time period t, we will continue to implement the gradient-based update as in $\mathrm{IGD}(\hat{\gamma})$ until the next re-solving time. The re-solving method has gain great popularity in both theory and application (Gallego et al., 2019), and in this experiment, we investigate how the technique can be used to further boost the performance of $\mathrm{IGD}(\hat{\gamma})$.

Table 2 reports the performance of $\mathrm{IGD}(\hat{\gamma})$ and its re-solving heuristics with frequency $k = 1, 50, 100, 200$. First, we note that the original version of $\mathrm{IGD}(\hat{\gamma})$ exhibits stably well performance across different combinations of α (the non-stationary intensity) and β (the estimation error). Second, the re-solving heuristic further boosts the performance of $\mathrm{IGD}(\hat{\gamma})$. We observe an improved performance even when the frequency $k = 200$ (which means only re-solving for 5 times throughout the horizon), but the performance improvement becomes marginal as we further increase the re-solving frequency. From a computational perspective, the re-solving heuristic for $\mathrm{IGD}(\hat{\gamma})$ provides a good trade-off between computational efficiency and algorithm performance, especially for large-scale system. It naturally blends the computational advantage of $\mathrm{IGD}(\hat{\gamma})$ and the algorithmic adaptivity of the re-solving technique. In recent years, the literature has developed a good

understanding of both the advantages and drawbacks of the re-solving technique in a stationary environment (See (Cooper, 2002; Reiman and Wang, 2008; Jasin and Kumar, 2012; Jasin, 2015; Bumpensanti and Wang, 2020) among others). In parallel, the experiment here raises an interesting but challenging future direction of understanding the re-solving technique in a non-stationary environment.

References

Adelman, Daniel. 2007. Dynamic bid prices in revenue management. *Operations Research* **55**(4) 647–661.

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

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

Agrawal, Shipra, Zizhuo Wang, Yinyu Ye. 2014. A dynamic near-optimal algorithm for online linear programming. *Operations Research* **62**(4) 876–890.

Arjovsky, Martin, Soumith Chintala, Léon Bottou. 2017. Wasserstein generative adversarial networks. *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. 214–223.

Arlotto, Alessandro, Itai Gurvich. 2019. Uniformly bounded regret in the multisecretary problem. *Stochastic Systems* .

Arlotto, Alessandro, Xinchang Xie. 2020. Logarithmic regret in the dynamic and stochastic knapsack problem with equal rewards. *Stochastic Systems* .

Asadpour, Arash, Xuan Wang, Jiawei Zhang. 2020. Online resource allocation with limited flexibility. *Management Science* **66**(2) 642–666.

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

Balseiro, Santiago, Haihao Lu, Vahab Mirrokni. 2020. The best of many worlds: Dual mirror descent for online allocation problems. *arXiv preprint arXiv:2011.10124* .

Banerjee, Siddhartha, Daniel Freund. 2020a. Good prophets know when the end is near. *Available at SSRN 3479189* .

Banerjee, Siddhartha, Daniel Freund. 2020b. Uniform loss algorithms for online stochastic decision-making with applications to bin packing. *SIGMETRICS* .

Bemporad, Alberto, Manfred Morari. 1999. Robust model predictive control: A survey. *Robustness in identification and control*. Springer, 207–226.

Berbeglia, Gerardo, Gwenaël Joret. 2020. Assortment optimisation under a general discrete choice model: A tight analysis of revenue-ordered assortments. *Algorithmica* **82**(4) 681–720.

Besbes, Omar, Yonatan Gur, Assaf Zeevi. 2014. Stochastic multi-armed-bandit problem with non-stationary rewards. *Advances in neural information processing systems*. 199–207.

Besbes, Omar, Yonatan Gur, Assaf Zeevi. 2015. Non-stationary stochastic optimization. *Operations research* **63**(5) 1227–1244.

Besbes, Omar, Assaf Zeevi. 2012. Blind network revenue management. *Operations research* **60**(6) 1537–1550.

Blanchet, Jose, Guillermo Gallego, Vineet Goyal. 2016. A markov chain approximation to choice modeling. *Operations Research* **64**(4) 886–905.

Blanchet, Jose, Yang Kang, Karthyek Murthy. 2019. Robust wasserstein profile inference and applications to machine learning. *Journal of Applied Probability* **56**(3) 830–857.

Buchbinder, Niv, Joseph Naor. 2009. Online primal-dual algorithms for covering and packing. *Mathematics of Operations Research* **34**(2) 270–286.

Bumpensanti, Pornpawee, He Wang. 2020. A re-solving heuristic with uniformly bounded loss for network revenue management. *Management Science* .

Cheung, Wang Chi, Guodong Lyu, Chung-Piaw Teo, Hai Wang. 2020. Online planning with offline simulation. *Available at SSRN 3709882* .

Cheung, Wang Chi, David Simchi-Levi, Ruihao Zhu. 2019. Non-stationary reinforcement learning: The blessing of (more) optimism. *Available at SSRN 3397818* .

Cooper, William L. 2002. Asymptotic behavior of an allocation policy for revenue management. *Operations Research* **50**(4) 720–727.

Davis, James, Guillermo Gallego, Huseyin Topaloglu. 2013. Assortment planning under the multinomial logit model with totally unimodular constraint structures. *Work in Progress* .

Devanur, Nikhil R, Kamal Jain, Balasubramanian Sivan, Christopher A Wilkens. 2019. Near optimal online algorithms and fast approximation algorithms for resource allocation problems. *Journal of the ACM (JACM)* **66**(1) 7.

Esfahani, Peyman Mohajerin, Daniel Kuhn. 2018. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming* **171**(1-2) 115–166.

Feldman, Jacob B, Huseyin Topaloglu. 2017. Revenue management under the markov chain choice model. *Operations Research* **65**(5) 1322–1342.

Ferguson, Thomas S, et al. 1989. Who solved the secretary problem? *Statistical science* **4**(3) 282–289.

Fournier, Nicolas, Arnaud Guillin. 2015. On the rate of convergence in wasserstein distance of the empirical measure. *Probability Theory and Related Fields* **162**(3) 707–738.

Galichon, Alfred. 2018. *Optimal transport methods in economics*. Princeton University Press.

Gallego, Guillermo, Huseyin Topaloglu. 2014. Constrained assortment optimization for the nested logit model. *Management Science* **60**(10) 2583–2601.

Gallego, Guillermo, Huseyin Topaloglu, et al. 2019. *Revenue management and pricing analytics*, vol. 209. Springer.

Gallego, Guillermo, Garrett Van Ryzin. 1994. Optimal dynamic pricing of inventories with stochastic demand over finite horizons. *Management science* **40**(8) 999–1020.

Garivier, Aurélien, Eric Moulines. 2008. On upper-confidence bound policies for non-stationary bandit problems. *arXiv preprint arXiv:0805.3415* .

Gibbs, Alison L, Francis Edward Su. 2002. On choosing and bounding probability metrics. *International statistical review* **70**(3) 419–435.

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

Gupta, Anupam, Marco Molinaro. 2014. How experts can solve lps online. *European Symposium on Algorithms*. Springer, 517–529.

Hall, Eric, Rebecca Willett. 2013. Dynamical models and tracking regret in online convex programming. *International Conference on Machine Learning*. PMLR, 579–587.

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

Immorlica, Nicole, Karthik Abinav Sankararaman, Robert Schapire, Aleksandrs Slivkins. 2019. Adversarial bandits with knapsacks. *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 202–219.

Jadbabaie, Ali, Alexander Rakhlin, Shahin Shahrampour, Karthik Sridharan. 2015. Online optimization: Competing with dynamic comparators. *Artificial Intelligence and Statistics*. PMLR, 398–406.

Jagabathula, Srikanth. 2014. Assortment optimization under general choice. *Available at SSRN 2512831* .

Jasin, Stefanus. 2015. Performance of an lp-based control for revenue management with unknown demand parameters. *Operations Research* **63**(4) 909–915.

Jasin, Stefanus, Sunil Kumar. 2012. A re-solving heuristic with bounded revenue loss for network revenue management with customer choice. *Mathematics of Operations Research* **37**(2) 313–345.

Jasin, Stefanus, Sunil Kumar. 2013. Analysis of deterministic lp-based booking limit and bid price controls for revenue management. *Operations Research* **61**(6) 1312–1320.

Jenatton, Rodolphe, Jim Huang, Cédric Archambeau. 2016. Adaptive algorithms for online convex optimization with long-term constraints. *International Conference on Machine Learning*. PMLR, 402–411.

Jiang, Jiashuo, Shixin Wang, Jiawei Zhang. 2019. Achieving high individual service-levels without safety stock? optimal rationing policy of pooled resources. *Optimal Rationing Policy of Pooled Resources (May 2, 2019)*. .

Jiang, Jiashuo, Jiawei Zhang. 2020. Online resource allocation with stochastic resource consumption. *arXiv preprint arXiv:2012.07933* .

Kunnumkal, Sumit, Kalyan Talluri. 2016. On a piecewise-linear approximation for network revenue management. *Mathematics of Operations Research* **41**(1) 72–91.

Lattimore, Tor, Csaba Szepesvári. 2020. *Bandit algorithms*. Cambridge University Press.

Lecarpentier, Erwan, Emmanuel Rachelson. 2019. Non-stationary markov decision processes, a worst-case approach using model-based reinforcement learning. *Advances in Neural Information Processing Systems*. 7216–7225.

Li, Guang, Paat Rusmevichientong, Huseyin Topaloglu. 2015. The d-level nested logit model: Assortment and price optimization problems. *Operations Research* **63**(2) 325–342.

Li, Xiaocheng, Chunlin Sun, Yinyu Ye. 2020. Simple and fast algorithm for binary integer and online linear programming. *arXiv preprint arXiv:2003.02513* .

Lu, Haihao, Santiago Balseiro, Vahab Mirrokni. 2020. Dual mirror descent for online allocation problems. *arXiv preprint arXiv:2002.10421* .

Ma, Yuhang, Paat Rusmevichientong, Mika Sumida, Huseyin Topaloglu. 2020. An approximation algorithm for network revenue management under nonstationary arrivals. *Operations Research* **68**(3) 834–855.

Mangasarian, Olvi L, T-H Shiau. 1987. Lipschitz continuity of solutions of linear inequalities, programs and complementarity problems. *SIAM Journal on Control and Optimization* **25**(3) 583–595.

Mehta, Aranyak, Amin Saberi, Umesh Vazirani, Vijay Vazirani. 2005. Adwords and generalized on-line matching. *46th Annual IEEE Symposium on Foundations of Computer Science (FOCS'05)*. IEEE, 264–273.

Miao, Sentao, Yining Wang, Jiawei Zhang. 2021. A general framework for resource constrained revenue management with demand learning and large action space. *Available at SSRN 3841273* .

Molinaro, Marco, Ramamoorthi Ravi. 2013. The geometry of online packing linear programs. *Mathematics of Operations Research* **39**(1) 46–59.

Neely, Michael J, Hao Yu. 2017. Online convex optimization with time-varying constraints. *arXiv preprint arXiv:1702.04783* .

Rangi, Anshuka, Massimo Franceschetti, Long Tran-Thanh. 2018. Unifying the stochastic and the adversarial bandits with knapsack. *arXiv preprint arXiv:1811.12253* .

Rawlings, James B. 2000. Tutorial overview of model predictive control. *IEEE control systems magazine* **20**(3) 38–52.

Reiman, Martin I, Qiong Wang. 2008. An asymptotically optimal policy for a quantity-based network revenue management problem. *Mathematics of Operations Research* **33**(2) 257–282.

Rusmevichientong, Paat, Zuo-Jun Max Shen, David B Shmoys. 2010. Dynamic assortment optimization with a multinomial logit choice model and capacity constraint. *Operations research* **58**(6) 1666–1680.

Russac, Yoan, Claire Vernade, Olivier Cappé. 2019. Weighted linear bandits for non-stationary environments. *Advances in Neural Information Processing Systems*. 12040–12049.

Sun, Rui, Xinshang Wang, Zijie Zhou. 2020. Near-optimal primal-dual algorithms for quantity-based network revenue management. *arXiv preprint arXiv:2011.06327* .

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

Talluri, Kalyan, Garrett Van Ryzin. 2004. Revenue management under a general discrete choice model of consumer behavior. *Management Science* **50**(1) 15–33.

Talluri, Kalyan T, Garrett J Van Ryzin. 2006. *The theory and practice of revenue management*, vol. 68. Springer Science & Business Media.

Vanderbei, Robert J, et al. 2015. *Linear programming*. Springer.

Vera, Alberto, Siddhartha Banerjee. 2020. The bayesian prophet: A low-regret framework for online decision making. *Management Science* .

Vera, Alberto, Siddhartha Banerjee, Itai Gurvich. 2019. Online allocation and pricing: Constant regret via bellman inequalities. *arXiv preprint arXiv:1906.06361* .

Villani, Cédric. 2008. *Optimal transport: old and new*, vol. 338. Springer Science & Business Media.

Wagener, Nolan, Ching-An Cheng, Jacob Sacks, Byron Boots. 2019. An online learning approach to model predictive control. *arXiv preprint arXiv:1902.08967* .

Yi, Xinlei, Xiuxian Li, Tao Yang, Lihua Xie, Tianyou Chai, Karl Johansson. 2021. Regret and cumulative constraint violation analysis for online convex optimization with long term constraints. *International Conference on Machine Learning*. PMLR, 11998–12008.

Yuan, Jianjun, Andrew Lamperski. 2018. Online convex optimization for cumulative constraints. *Advances in Neural Information Processing Systems*. 6137–6146.

Zhang, Dan, Daniel Adelman. 2009. An approximate dynamic programming approach to network revenue management with customer choice. *Transportation Science* **43**(3) 381–394.

Appendix. Proofs of Lemmas, Propositions and Theorems

A. Proofs of Section 3

A1. Proof of Lemma 1

Proof: Given the realized parameters $\mathcal{H} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_T)$, we can denote the offline optimum of (PCP) as a function of $\mathcal{H}$, namely, $\{\boldsymbol{x}_t^*(\mathcal{H})\}_{t=1}^T$. Let

$$\tilde{\boldsymbol{x}}_t(\boldsymbol{\theta}) = \mathbb{E}\left[\boldsymbol{x}_t^*(\mathcal{H}) | \boldsymbol{\theta}_t = \boldsymbol{\theta}\right]$$

where the conditional expectation is taken with respect to $\mathcal{P}_j$ for $j \neq t$. We show that $\tilde{\boldsymbol{x}}_t(\boldsymbol{\theta})$ is a feasible solution to (4). Specifically, note that for each $i = 1, \ldots, m$,

$$c_i \geq \mathbb{E}\left[\sum_{t=1}^T g_i(\boldsymbol{x}_t^*(\mathcal{H}); \boldsymbol{\theta}_t)\right] = \sum_{t=1}^T \mathbb{E}_{\boldsymbol{\theta}_t \sim \mathcal{P}_t}\left[\mathbb{E}[g_i(\boldsymbol{x}_t^*(\mathcal{H}); \boldsymbol{\theta}_t) | \boldsymbol{\theta}_t = \boldsymbol{\theta}]\right]$$

$$\geq \sum_{t=1}^T \int_{\boldsymbol{\theta} \in \Theta} g_i(\tilde{\boldsymbol{x}}_t(\boldsymbol{\theta}); \boldsymbol{\theta}) d\mathcal{P}_t(\boldsymbol{\theta}) = \sum_{t=1}^T \mathcal{P}_t g_i(\boldsymbol{x}_t; \boldsymbol{\theta})$$

where the first inequality comes from the feasibility of the optimal solution $\boldsymbol{x}_t^*(\mathcal{H})$ and the second inequality follows from that the function $g_i(\cdot; \boldsymbol{\theta}_t)$ is a convex function for each i and $\boldsymbol{\theta}_t \in \Theta$. Thus, $\{\tilde{\boldsymbol{x}}_t(\boldsymbol{\theta})\}$ is a feasible solution to (4). Similarly, we can analyze the objective function

$$\mathbb{E}[R_T^*] = \mathbb{E}\left[\sum_{t=1}^T f(\boldsymbol{x}_t^*(\mathcal{H}); \boldsymbol{\theta}_t)\right] = \sum_{t=1}^T \mathbb{E}_{\boldsymbol{\theta}_t \sim \mathcal{P}_t}\left[\mathbb{E}[f(\boldsymbol{x}_t^*(\mathcal{H}); \boldsymbol{\theta}_t) | \boldsymbol{\theta}_t = \boldsymbol{\theta}]\right]$$

$$\leq \sum_{t=1}^T \int_{\boldsymbol{\theta} \in \Theta} f(\tilde{\boldsymbol{x}}_t(\boldsymbol{\theta}); \boldsymbol{\theta}) d\mathcal{P}_t(\boldsymbol{\theta}) \leq R_T^{\mathrm{UB}}$$

where the first inequality follows from that the function $f(\cdot; \boldsymbol{\theta})$ is a concave function for any $\boldsymbol{\theta} \in \Theta$ and the last inequality comes from the optimality of R_T^{UB}. Thus we complete the proof. $\square$

A2. Proof of Proposition 1

Proof: We first prove that $\boldsymbol{p}^*$ is an optimal solution for L_t. Note that for each t, $L_t(\boldsymbol{p})$ is a convex function over $\boldsymbol{p}$ and

$$\nabla L_t(\boldsymbol{p}^*) = \boldsymbol{\gamma}_t + \mathcal{P}_t \nabla h(\boldsymbol{p}^*; \boldsymbol{\theta}_t) = \boldsymbol{\gamma}_t - \mathcal{P}_t \boldsymbol{g}(\boldsymbol{x}^*(\boldsymbol{\theta}_t); \boldsymbol{\theta}_t)$$

where $\boldsymbol{x}^*(\boldsymbol{\theta}) = \mathrm{argmax}_{\boldsymbol{x} \in \mathcal{X}}\{f(\boldsymbol{x}; \boldsymbol{\theta}) - (\boldsymbol{p}^*)^T \cdot \boldsymbol{g}(\boldsymbol{x}; \boldsymbol{\theta})\}$. With the definition of $\boldsymbol{\gamma}_t$ in (6), it follows immediately that

$$\nabla L_t(\boldsymbol{p}^*) = 0$$

which implies that $\boldsymbol{p}^*$ is a minimizer of the function $L(\cdot)$ for each t. We then prove that $L(\boldsymbol{p}^*) = \sum_{t=1}^T L_t(\boldsymbol{p}^*)$. Define the set of binding constraints $\mathcal{I}_B = \{i : p_i^* > 0, i = 1, \ldots, m\}$. From the convexity of the function $L(\boldsymbol{p})$ over $\boldsymbol{p}$, for each $i \in \mathcal{I}_B$, it holds that

$$0 = \nabla_i L(\boldsymbol{p}^*) = c_i - \sum_{t=1}^T \mathcal{P}_t \boldsymbol{g}(\boldsymbol{x}^*(\boldsymbol{\theta}_t); \boldsymbol{\theta}_t) = c_i - \sum_{t=1}^T \gamma_{t,i}$$

Thus, we have that

$$\boldsymbol{c}^\top \cdot \boldsymbol{p}^* = \sum_{t=1}^{T} \boldsymbol{\gamma}_t^\top \cdot \boldsymbol{p}^*$$

It follows immediately that $L(\boldsymbol{p}^*) = \sum_{t=1}^{T} L_t(\boldsymbol{p}^*)$. $\square$

A3. Discussions on the function h

The definition of the function $h(\boldsymbol{p}; \boldsymbol{\theta})$, described as follows for completeness, plays a critical role in deriving our results,

$$h(\boldsymbol{p}; \boldsymbol{\theta}) = \max_{\boldsymbol{x} \in \mathcal{X}} \left\{ f(\boldsymbol{x}; \boldsymbol{\theta}) - \sum_{i=1}^{n} p_i \cdot g_i(\boldsymbol{x}; \boldsymbol{\theta}) \right\}. \tag{22}$$

Throughout the paper, we assume that the optimization problem in (22) can be solved efficiently so as to obtain both its optimal solution and optimal objective value. Here, we justify this assumption with a discussion of the computational aspect of solving (22). Note that when the function h is defined in the setting with random reward and consumption given in Section 6.4 where $f, \boldsymbol{g}$ are random functions, $f, \boldsymbol{g}$ can be simply replaced by their expectations $\hat{f}, \hat{\boldsymbol{g}}$ in the definition of h in (22). In the following discussion, we use the notations of $f, \boldsymbol{g}$ and $\hat{f}, \hat{\boldsymbol{g}}$ exchangeably for simplicity.

When (PCP) refers to the online linear programming problem or price-based network revenue management (NRM) problem as described in Section 2.2, the optimization problem in (22) is reduced to a simple convex optimization problem and it can indeed be solved in polynomial time by existing methods. We now consider (22) when (PCP) represents the choice-based NRM problem.

In the choice-based NRM problem, there are n products and each product i is associated with a revenue r_i. The definitions of $\boldsymbol{x}_t$ and the functions $f, \boldsymbol{g}$ are described in Section 2.2. Given the assortment s, the customer with the type θ_t chooses one product i to purchase with a probability $\eta_i(s; \theta_t)$, where $\boldsymbol{\eta}$ is specified by the choice model of the customer and is assumed to be known to the decision maker. Here, the function f refers to the expected revenue of the assortment $\boldsymbol{x}_t$ and the function g_i refers to the probability that a product i is purchased:

$$f(\boldsymbol{x}_t; \theta_t) = \sum_{s \in \mathcal{S}} \sum_{j=1}^{n} x_{t,s} \cdot r_j \cdot \eta_j(s; \theta_t) \;\; \text{and} \;\; g_i(\boldsymbol{x}_t; \theta_t) = \sum_{s \in \mathcal{S}} x_{t,s} \cdot \eta_i(s; \theta_t) \quad \forall i$$

Then, the optimization problem in (22) could be rewritten in the following equivalent formulation:

$$\max_{s \in \mathcal{S} \subset \{0,1\}^n} \sum_{i=1}^{n} (r_i - p_i) \cdot \eta_i(s; \theta_t) \tag{23}$$

With the following choice models to specify the function η_i's, the optimization problem in (23) could all be solved efficiently.

Multinomial logit model (NML): Talluri and Van Ryzin (2004) show that without further constraints, it is optimal to sort the products according to a decreasing order $r_i - p_i$ and find

an optimal revenue-order assortment among $\{1\}, \{1, 2, \}, \ldots, \{1, 2 \ldots, n\}$. Rusmevichientong et al. (2010) further propose a simple polynomial-time algorithm to compute the optimal assortment where there is a capacity constraint. Davis et al. (2013) consider the assortment problem with totally unimodular constraints and show that the constrained problem could be solved as an equivalent linear program.

Nested logit model: Li et al. (2015) consider the assortment optimization problem under a d-level nested logit model and propose an algorithm to compute the optimal assortment in polynomial time. Gallego and Topaloglu (2014) consider the setting with cardinality constraint and show that the optimal assortment can be obtained efficiently by solving a linear program. Gallego and Topaloglu (2014) further consider the setting with a capacity constraint and propose an algorithm with constant competitiveness and good empirical performance.

Markov chain based models: Blanchet et al. (2016) propose the Markov chain based model and develop polynomial-time solution algorithms. Feldman and Topaloglu (2017) propose a linear programming approach to obtain the optimal assortment under this model.

General choice model: There are also papers considering the assortment problem without exploiting the specific structures of the choice model. Jagabathula (2014) studies a local search heuristic and shows its great empirical performance. Also, Berbeglia and Joret (2020) analyze the performance of the revenue-order assortment heuristic under the general discrete choice model, which can be applied to solve (23) approximately.

A4. Proof of Lemma 2

Proof: Note that the following two properties are satisfied by the update rule (7):

(i). If $\|\boldsymbol{p}_t\|_\infty \leq q$, then we must have $\|\boldsymbol{p}_{t+1}\|_\infty \leq q+1$ by noting that for each i, the i-th component of $\boldsymbol{p}_t$, denoted as $p_{t,i}$, is nonnegative and $g_i(\cdot, \boldsymbol{\theta}_t)$ is normalized within $[0, 1]$.

(ii). If there exists i such that $p_{t,i} > q$, then we must have $p_{t+1,i} < p_{t,i}$. Specifically, when $p_{t,i} > q$, we must have that $g_i(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) = 0$, otherwise we would have that

$$f(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) - \boldsymbol{p}_t^\top \cdot \boldsymbol{g}(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) \leq f(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) - p_{t,i} \cdot g_i(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) < 0$$

which contradicts the definition of $\tilde{\boldsymbol{x}}_t$ in Algorithm 1 since we could always select $\boldsymbol{x}_t = \boldsymbol{0}$ to obtain a zero objective value as per Assumption 1. Then from (7), it holds that $p_{t+1,i} < p_{t,i}$.

Starting from $\boldsymbol{p}_1 = \boldsymbol{0}$ and iteratively applying the above two property to control the increase of $\boldsymbol{p}_t$ from $t = 1$ to T, we obtain that for the first time that one component of $\boldsymbol{p}_t$ exceeds the threshold q, it is upper bounded by $q+1$ and this component will continue to decrease until it falls below the threshold q. Thus, we have $\|\boldsymbol{p}_t\|_\infty \leq q+1$ with probability 1 for each t. $\square$

A5. Proof of Theorem 1

Proof: In IGD($\boldsymbol{\gamma}$), the true action $\boldsymbol{x}_t$ taken by the decision maker differs from the virtual action $\tilde{\boldsymbol{x}}_t$ if and only if $\boldsymbol{c}_t$ cannot fully satisfy $\boldsymbol{g}(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t)$. Thus, we have that

$$f(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) - f(\boldsymbol{x}_t; \boldsymbol{\theta}_t) \leq f(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) \cdot \mathbb{I}\{\exists i : c_{t,i} < g_i(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t)\}$$

where $c_{t,i}$ denotes the i-th component of $\boldsymbol{c}_t$ and $\mathbb{I}\{\cdot\}$ denotes the indicator function. Moreover, we know

$$\mathbb{I}\{\exists i : c_{t,i} < g_i(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t)\} \leq \sum_{i=1}^{m} \mathbb{I}\left\{\sum_{j=1}^{t} g_i(\tilde{\boldsymbol{x}}_j; \boldsymbol{\theta}_j) > c_i\right\}.$$

Recall that the maximum reward generated by consuming per unit of budget of each constraint is upper bounded by q. We have

$$f(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) \cdot \mathbb{I}\{\exists i : c_{t,i} < g_i(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t)\} \leq q \sum_{i=1}^{m} g_i(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}) \cdot \mathbb{I}\left\{\sum_{j=1}^{t} g_i(\tilde{\boldsymbol{x}}_j; \boldsymbol{\theta}_j) > c_i\right\}$$

From the fact that $\boldsymbol{g}(\cdot; \boldsymbol{\theta}_t) \in [0,1]^m$,

$$\sum_{t=1}^{T} f(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) - \sum_{t=1}^{T} f(\boldsymbol{x}_t; \boldsymbol{\theta}_t) \leq q \cdot \sum_{i=1}^{m} \sum_{t=1}^{T} g_i(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) \cdot \mathbb{I}\left\{\sum_{j=1}^{t} g_i(\tilde{\boldsymbol{x}}_j; \boldsymbol{\theta}_j) > c_i\right\}$$

$$\leq q \cdot \sum_{i=1}^{m} \left[\sum_{t=1}^{T} g_i(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) - (c_i - 1)\right]^{+}$$

which related the total collected reward by the true action $\{\boldsymbol{x}_t\}_{t=1}^{T}$ and the virtual action $\{\tilde{\boldsymbol{x}}_t\}_{t=1}^{T}$. Further from Proposition 1, we have that

$$\mathrm{Reg}_T(\pi) \leq \min_{\boldsymbol{p} \geq 0} \sum_{t=1}^{T} L_t(\boldsymbol{p}) - \mathbb{E}\left[\sum_{t=1}^{T} f(\boldsymbol{x}_t; \boldsymbol{\theta}_t)\right] \leq \underbrace{\sum_{t=1}^{T} \min_{\boldsymbol{p} \geq 0} L_t(\boldsymbol{p}) - \mathbb{E}\left[\sum_{t=1}^{T} f(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t)\right]}_{\mathrm{I}}$$

$$+ \underbrace{q \cdot \mathbb{E}\left[\sum_{i=1}^{m} \left[\sum_{t=1}^{T} g_i(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) - (c_i - 1)\right]^{+}\right]}_{\mathrm{II}}$$

We then bound the term I and term II separately to derive our regret bound.

Bound I: Note that for each t, the distribution of $\boldsymbol{p}_t$ is independent from the distribution of $\boldsymbol{\theta}_\tau$ for any $\tau \leq t$, then we have that

$$\min_{\boldsymbol{p} \geq 0} L_t(\boldsymbol{p}) \leq \mathbb{E}_{\boldsymbol{p}_t}[L_t(\boldsymbol{p}_t)] = \mathbb{E}_{\boldsymbol{p}_t}\left[\boldsymbol{\gamma}_t^\top \boldsymbol{p}_t + \mathcal{P}_t h(\boldsymbol{p}_t; \boldsymbol{\theta}_t)\right]$$

where the expectation is taken with respect to the randomness of the dual price $\boldsymbol{p}_t$. Thus, we have

$$\mathrm{I} \leq \sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{p}_t}\left[\boldsymbol{\gamma}_t^\top \boldsymbol{p}_t + \mathcal{P}_t\{h(\boldsymbol{p}_t; \boldsymbol{\theta}_t) - f(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t)\}\right]$$

From the definition of $\tilde{\boldsymbol{x}}_t$, we get that $h(\boldsymbol{p}_t; \boldsymbol{\theta}_t) - f(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) = -\boldsymbol{p}_t^\top \cdot \boldsymbol{g}(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t)$, which implies that

$$\mathrm{I} \leq \sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{p}_t} \left[\boldsymbol{p}_t^\top \cdot (\boldsymbol{\gamma}_t - \mathcal{P}_t \boldsymbol{g}(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t)) \right]$$

Note that from the update rule (7), we have that

$$\|\boldsymbol{p}_{t+1}\|_2^2 \leq \|\boldsymbol{p}_t\|_2^2 + \frac{1}{T} \cdot \|\boldsymbol{g}(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) - \boldsymbol{\gamma}_t\|_2^2 - \frac{2}{\sqrt{T}} \cdot \boldsymbol{p}_t^\top \cdot (\boldsymbol{\gamma}_t - \boldsymbol{g}(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t))$$

which implies that

$$\mathbb{E}_{\boldsymbol{p}_t} \left[\boldsymbol{p}_t^\top \cdot (\boldsymbol{\gamma}_t - \mathcal{P}_t \boldsymbol{g}(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t)) \right] \leq \frac{\sqrt{T}}{2} \cdot \left(\mathbb{E}[\|\boldsymbol{p}_t\|_2^2] - \mathbb{E}[\|\boldsymbol{p}_{t+1}\|_2^2] \right) + \frac{m}{2\sqrt{T}}$$

Thus, it holds that

$$\mathrm{I} \leq \frac{m\sqrt{T}}{2} \tag{24}$$

Bound II: Note that from the update rule (7), we have that

$$\sqrt{T} \cdot \boldsymbol{p}_{t+1} \geq \sqrt{T} \cdot \boldsymbol{p}_t + \boldsymbol{g}(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) - \boldsymbol{\gamma}_t$$

which implies that

$$\sum_{t=1}^{T} \boldsymbol{g}(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) - \boldsymbol{c} \leq \sum_{t=1}^{T} \boldsymbol{g}(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) - \sum_{t=1}^{T} \boldsymbol{\gamma}_t \leq \sqrt{T} \cdot \boldsymbol{p}_{T+1}$$

Thus, it holds that

$$\mathrm{II} = q \cdot \mathbb{E} \left[\sum_{i=1}^{m} \left[\sum_{t=1}^{T} g_i(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) - (c_i - 1) \right]^+ \right] \leq mq(q+1) \cdot \sqrt{T} + qm \tag{25}$$

We obtain the $O(\sqrt{T})$ regret bound immediately by combining (24) and (25). $\qquad\square$

A6. Proof of Proposition 2

Proof: We construct the following distribution $\mathcal{P}$ over the parameter θ. Suppose that there is a single resource. Denote the support set of $\mathcal{P}$ as $\{\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(j)}, \ldots\}$ and for each $j = 1, 2, \ldots$, we have that

$$\mathcal{P}(\theta = \theta^{(j)}) = \frac{1}{2} \cdot \frac{1}{2^j} \quad \text{and} \quad \mathcal{P}(\theta = \theta^{(0)}) = \frac{1}{2}$$

Moreover, we have the following condition over the support set $\{\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(j)}, \ldots\}$ of $\mathcal{P}$. Denote sequences of constants $\{a_j\}$, $\{b_j\}$, where

$$a_j = 1 + \frac{2j+1}{2j(j+1)}, \ b_j = 1 + 1/(j+1)$$

for $j = 1, 2, \ldots$. Clearly, we have $1 \leq \cdots \leq a_{j+1} < b_j < a_j \cdots \leq 2$. The reward function satisfies

$$0 \leq \min_{0 < x \leq 1} \frac{f(x; \theta^{(0)})}{x} \leq \max_{0 < x \leq 1} \frac{f(x; \theta^{(0)})}{x} \leq 1 \text{ and } b_j \leq \min_{0 < x \leq 1} \frac{f(x; \theta^{(j)})}{x} \leq \max_{0 < x \leq 1} \frac{f(x; \theta^{(j)})}{x} \leq a_j, \ \forall j = 1, 2, \ldots$$

There is a single resource and the resource consumption function satisfies $g(x;\theta) = x$ for each θ. We are now ready to construct the example and show the lower bound.

Suppose there is a single resource with an initial capacity $\frac{T}{2}$, where T is the total time periods. For a fixed T, there must exists an integer k such that

$$\frac{1}{2\sqrt{T}} \le \frac{1}{2^{k+2}} \le \frac{1}{\sqrt{T}}$$

Now, we assume without loss of generality that T is set such that

$$\frac{1}{2^{k+2}} = \frac{1}{\sqrt{T}}$$

We now divide the support set $\{\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(j)}, \ldots\}$ into three subsets:

$$A_1 = \{\theta^{(j)} : 1 \le j < k\}, \ A_2 = \{\theta^{(k)}\} \text{ and } A_3 = \{\theta^{(0)}, \theta^{(j)} : j \ge k+1\}.$$

Clearly, at each time period t, we have that

$$P(\theta_t \in A_1) = \frac{1}{2} + \frac{1}{2^{k+1}}, P(\theta_t \in A_2) = \frac{1}{2^{k+1}} \text{ and } P(\theta_t \in A_3) = \frac{1}{2} - \frac{2}{2^{k+1}}.$$

Denote $\varepsilon = \frac{1}{2^{k+2}}$. We call that we encounter a type q request whenever the event $\{\theta_t \in A_q\}$ happens, where $q = 1, 2, 3$. We now denote Z_q^t as the number type q requests in the first t periods. We next introduce the event:

$$\mathcal{H}_1 = \left\{ \frac{1}{\epsilon} \le Z_2^T \le \min\{2Z_2^{T/2}, \frac{2}{\epsilon}\} \right\} = \left\{ \frac{1}{2}\mathbb{E}[Z_2^T] \le Z_2^T \le \min\{2Z_2^{T/2}, \mathbb{E}[Z_2^T]\} \right\}$$

$$\mathcal{H}_2 = \left\{ Z_1^T \ge \frac{T}{2} + \frac{2}{\epsilon} \right\} = \left\{ Z_1^T \ge \mathbb{E}[Z_1^T] + \frac{6}{\epsilon} \right\}$$

$$\mathcal{H}_3 = \left\{ Z_1^T \le \frac{T}{2} - \frac{4}{\epsilon} \right\} = \left\{ Z_1^T \le \mathbb{E}[Z_1^T] \right\}$$

It is obvious that on event $\mathcal{H}_2$, particularly on the event $\mathcal{H}_2 \cap \mathcal{H}_1$, all the resources will be consumed by type 1 requests by the offline optimum. That is, at each period t, the offline optimum will set the decision variable x_t non-zero only when the event $\{\theta_t \in A_1\}$ happens. Similarly, on event $\mathcal{H}_1 \cap \mathcal{H}_3$, the offline optimum will set the decision variable $x_t = 1$ whenever the event $\{\theta_t \in A_1\}$ or $\{\theta_t \in A_2\}$ happens. Following the same argument as Arlloto and Gurvich (2019), we know that there exists a constant $\alpha_1 > 0$, independent of ϵ, such that

$$P(\mathcal{H}_1 \cap \mathcal{H}_2) \ge \alpha_1 \text{ and } P(\mathcal{H}_1 \cap \mathcal{H}_3) \ge \alpha_1$$

Next, we consider the difference between the offline optimum and the dynamic programming policy. We denote $S_2^{T/2}$ as the amount of resource consumed by type 2 request during the first $T/2$ periods, and we consider the event

$$\mathcal{H}_4 = \left\{ S_2^{T/2} \ge \frac{Z_2^T}{4} \right\}$$

Now, on the event $\mathcal{H}_4^c \cap \mathcal{H}_1 \cap \mathcal{H}_3$, the offline optimum will set $x_t = 1$ for all type 1 and type 2 request. On the other hand, the optimal online policy consumes at most $Z_2^T/4$ amount of resource using type 2 requests during the first $T/2$ periods. Note that as induced by the event $\mathcal{H}_1$, at most $Z_2^T/2$ number of type 2 request can arrive during the last $T/2$ periods. Thus, for the optimal online policy, type 2 request can consume at most $\frac{3}{4} \cdot Z_2^T$ amount of resource, while the offline optimum consume Z_2^T amount of resource with type 2 request. Further note that the reward/size for type 2 request is at least b_k and the reward/size for type 3 request is at most a_{k+1}. This will incur a regret at least

$$\mathbb{E}[\text{Regret}] \geq \frac{b_k - a_{k+1}}{4} \cdot \mathbb{E}[Z_2^T \cdot \mathbb{1}\{\mathcal{H}_4^c \cap \mathcal{H}_1 \cap \mathcal{H}_3\}] \geq \frac{b_k - a_{k+1}}{4\epsilon} \cdot P(\mathcal{H}_4^c \cap \mathcal{H}_1 \cap \mathcal{H}_3)$$

For each arrival sample path that falls in the set $\mathcal{H}_4 \cap \mathcal{H}_1 \cap \mathcal{H}_3$, we can find another sample path in $\mathcal{H}_4 \cap \mathcal{H}_1 \cap \mathcal{H}_2$ by keeping the first $T/2$ arrivals unchanged, and for the last $T/2$ periods, replacing at most $6/\epsilon$ type 3 requests with type 1 requests. We denote this resulting set of sample path as $\mathcal{L}$. Note that for each period, the arrival probability for type 1 request and type 3 request are both bounded away from 0. We know that there exists a constant $\alpha_2 > 0$, independent of ϵ, such that

$$P(\mathcal{L}) \geq \alpha_2 \cdot P(\mathcal{H}_4 \cap \mathcal{H}_1 \cap \mathcal{H}_3).$$

Moreover, for each sample path in the set $\mathcal{L}$, since it is on the event $\mathcal{H}_4 \cap \mathcal{H}_1 \cap \mathcal{H}_2$, we know that the offline optimum will consume all the resource with type 1 request, while the optimal online policy consumes at least $\frac{Z_2^T}{4}$ amount of resource with type 2 requests. Note that the reward/size for type 1 is at least b_{k-1}, while the reward/size for type 2 is at most a_k. This will incur a regret at least

$$\mathbb{E}[\text{regret}] \geq \frac{b_{k-1} - a_k}{4} \cdot \mathbb{E}[Z_2^T \cdot \mathbb{1}\{\mathcal{L}\}] \geq \frac{b_{k-1} - a_k}{4\epsilon} \cdot P(\mathcal{L}) \geq \alpha_2 \cdot \frac{b_{k-1} - a_k}{4\epsilon} \cdot P(\mathcal{H}_4 \cap \mathcal{H}_1 \cap \mathcal{H}_3).$$

Since

$$P(\mathcal{H}_1 \cap \mathcal{H}_3) = P(\mathcal{H}_4^c \cap \mathcal{H}_1 \cap \mathcal{H}_3) + P(\mathcal{H}_4 \cap \mathcal{H}_1 \cap \mathcal{H}_3) \geq \alpha_1,$$

we know that there exists a constant $\alpha_3 > 0$, independent of ϵ, such that

$$\mathbb{E}[\text{regret}] \geq \frac{\alpha_3}{\epsilon} \cdot \min\{b_{k-1} - a_k, b_k - a_{k+1}\}.$$

Recall that $\epsilon = 1/\sqrt{T}$ and

$$\min\{b_{k-1} - a_k, b_k - a_{k+1}\} = \frac{1}{(k+1)(k+2)}$$

and that $\frac{1}{2^{k+2}} = \frac{1}{\sqrt{T}}$, i.e., $k + 2 = \frac{\log T}{2}$. We have

$$\mathbb{E}[\text{regret}] \geq \frac{\alpha_3}{4} \cdot \frac{\sqrt{T}}{(\log T)^2}$$

which completes our proof.

B. Proofs of Section 4

B1. Proof of Theorem 2

Proof: It follows directly from Proposition 2 that for any policy π, we have $\text{Reg}_T(\pi) \geq \Omega(\sqrt{T})$. Thus, it is enough to consider the $\Omega(W_T)$ part in the lower bound. We consider the following estimated problem, where the true coefficients in (PCP) are replaced by the estimates:

$$\begin{aligned}
\max \quad & x_1 + ... + x_c + x_{c+1} + ... + x_T & (26)\\
\text{s.t.} \quad & x_1 + ... + x_c + x_{c+1} + ... + x_T \leq c \\
& 0 \leq x_t \leq 1 \quad \text{for } t = 1, ..., T.
\end{aligned}$$

where $c = \frac{T}{2}$ and the prior estimate $\hat{\mathcal{P}}_t$ is simply a one-point distribution for each t. Now we consider the following two possible true problems, the distributions of which are all one-point distributions and belong to the set Ξ_P with variation budget W_T:

$$\begin{aligned}
\max \quad & x_1 + ... + x_c + \left(1 + \frac{W_T}{T}\right) x_{c+1} + ... + \left(1 + \frac{W_T}{T}\right) x_T & (27)\\
\text{s.t.} \quad & x_1 + ... + x_c + x_{c+1} + ... + x_T \leq c \\
& 0 \leq x_t \leq 1 \quad \text{for } t = 1, ..., T. \\
\max \quad & x_1 + ... + x_c + \left(1 - \frac{W_T}{T}\right) x_{c+1} + ... + \left(1 - \frac{W_T}{T}\right) x_T & (28)\\
\text{s.t.} \quad & x_1 + ... + x_c + x_{c+1} + ... + x_T \leq c \\
& 0 \leq x_t \leq 1 \quad \text{for } t = 1, ..., T.
\end{aligned}$$

where $c = \frac{T}{2}$. Denote $x_t^1(\pi)$ as the decision of any policy π at period t for scenario (27) and denote $x_t^2(\pi)$ as the decision of policy π at period t for scenario (28). Further define $T_1(\pi)$ (resp. $T_2(\pi)$) as the expected capacity consumption of policy π on scenario (27) (resp. scenario (28)) during the first $\frac{T}{2}$ time periods:

$$T_1(\pi) = \mathbb{E}\left[\sum_{t=1}^{\frac{T}{2}} x_t^1(\pi)\right] \quad \text{and} \quad T_2(\pi) = \mathbb{E}\left[\sum_{t=1}^{\frac{T}{2}} x_t^2(\pi)\right]$$

Then, we have that

$$R_T^1(\pi) = \frac{T + W_T}{2} - \frac{W_T}{T} \cdot T_1(\pi) \quad \text{and} \quad R_T^2(\pi) = \frac{T - W_T}{2} + \frac{W_T}{T} \cdot T_2(\pi)$$

where $R_T^1(\pi)$ (resp. $R_T^2(\pi)$) denotes the expected reward collected by policy π on scenario (27) (resp. scenario (28)). Thus, the regret of policy π on scenario (27) and (28) are $\frac{W_T}{T} \cdot T_1(\pi)$ and $W_T - \frac{W_T}{T} \cdot T_2(\pi)$ respectively. Further note that since the implementation of policy π at each time

period should be only dependent on the historical information and the coefficients in the estimated problem (26), we must have $T_1(\pi) = T_2(\pi)$. Thus, we have that

$$\text{Reg}_T(\pi) \geq \max\left\{\frac{W_T}{T} \cdot T_1(\pi), W_T - \frac{W_T}{T} \cdot T_1(\pi)\right\} \geq \frac{W_T}{2} = \Omega(W_T)$$

which completes our proof. $\square$

B2. Proof of Lemma 3

Proof: Due to symmetry, it is sufficient to show that for every $\boldsymbol{p}$ such that $\boldsymbol{p} \in \Omega_{\bar{p}}$,

$$L_{\mathcal{Q}_2}(\boldsymbol{p}) - L_{\mathcal{Q}_1}(\boldsymbol{p}) \leq \max\{1, \bar{p}\} \cdot W(\mathcal{Q}_1, \mathcal{Q}_2).$$

Denote $\mathcal{Q}_{1,2}^*$ as the optimal coupling of the distribution $\mathcal{Q}_1$ and $\mathcal{Q}_2$, i.e., the optimal solution to (8), and denote

$$\boldsymbol{x}^*(\boldsymbol{\theta}) = \text{argmax}_{\boldsymbol{x} \in \mathcal{X}}\left\{f(\boldsymbol{x}; \boldsymbol{\theta}) - \sum_{i=1}^m p_i \cdot g_i(\boldsymbol{x}; \boldsymbol{\theta})\right\}$$

Then for each $\boldsymbol{\theta}_1 \in \Theta$, we define

$$\hat{\boldsymbol{x}}(\boldsymbol{\theta}_1) = \int_{\boldsymbol{\theta}_2 \in \Theta} \boldsymbol{x}^*(\boldsymbol{\theta}_2) \frac{d\mathcal{Q}_{1,2}^*(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)}{d\mathcal{Q}_1(\boldsymbol{\theta}_1)}$$

where $\frac{d\mathcal{Q}_{1,2}^*(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)}{d\mathcal{Q}_1(\boldsymbol{\theta}_1)}$ is the Radon–Nikodym derivative of $\mathcal{Q}_{1,2}^*$ with respect $\mathcal{Q}_1$ and it can be interpreted as the conditional distribution of $\boldsymbol{\theta}_2$ given $\boldsymbol{\theta}_1$. Note that from the definition of $\mathcal{Q}_{1,2}^*$, we have that $\int_{\boldsymbol{\theta}_2 \in \Theta} \frac{d\mathcal{Q}_{1,2}^*(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)}{d\mathcal{Q}_1(\boldsymbol{\theta}_1)} = 1$. Thus, $\hat{\boldsymbol{x}}(\boldsymbol{\theta}_1)$ is actually a convex combination of $\{\boldsymbol{x}^*(\boldsymbol{\theta}_2)\}_{\forall \boldsymbol{\theta}_2 \in \Theta}$. Moreover, from the concavity of $f(\cdot; \boldsymbol{\theta}_1)$ and the convexity of $g_i(\cdot; \boldsymbol{\theta}_1)$ for each i, we have that

$$f(\hat{\boldsymbol{x}}(\boldsymbol{\theta}_1); \boldsymbol{\theta}_1) \geq \int_{\boldsymbol{\theta}_2 \in \Theta} f(\boldsymbol{x}^*(\boldsymbol{\theta}_2); \boldsymbol{\theta}_1) \cdot \frac{d\mathcal{Q}_{1,2}^*(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)}{d\mathcal{Q}_1(\boldsymbol{\theta}_1)}$$

and

$$g_i(\hat{\boldsymbol{x}}(\boldsymbol{\theta}_1); \boldsymbol{\theta}_1) \leq \int_{\boldsymbol{\theta}_2 \in \Theta} g_i(\boldsymbol{x}^*(\boldsymbol{\theta}_2); \boldsymbol{\theta}_1) \cdot \frac{d\mathcal{Q}_{1,2}^*(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)}{d\mathcal{Q}_1(\boldsymbol{\theta}_1)}$$

Thus, we have that

$$\begin{aligned}
L_{\mathcal{Q}_1}(\boldsymbol{p}) &= \int_{\boldsymbol{\theta}_1 \in \Theta} \max_{\boldsymbol{x} \in \mathcal{X}}\left\{f(\boldsymbol{x}; \boldsymbol{\theta}_1) - \sum_{i=1}^m p_i \cdot g_i(\boldsymbol{x}; \boldsymbol{\theta}_1)\right\} d\mathcal{Q}_1(\boldsymbol{\theta}_1) \\
&\geq \int_{\boldsymbol{\theta}_1 \in \Theta}\left\{f(\hat{\boldsymbol{x}}(\boldsymbol{\theta}_1); \boldsymbol{\theta}_1) - \sum_{i=1}^m p_i \cdot g_i(\hat{\boldsymbol{x}}(\boldsymbol{\theta}_1); \boldsymbol{\theta}_1)\right\} d\mathcal{Q}_1(\boldsymbol{\theta}_1) \\
&\geq \int_{\boldsymbol{\theta}_1 \in \Theta} \int_{\boldsymbol{\theta}_2 \in \Theta}\left\{f(\boldsymbol{x}^*(\boldsymbol{\theta}_2); \boldsymbol{\theta}_1) - \sum_{i=1}^m p_i \cdot g_i(\boldsymbol{x}^*(\boldsymbol{\theta}_2); \boldsymbol{\theta}_1)\right\} d\mathcal{Q}_{1,2}^*(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)
\end{aligned}$$

Also, note that for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$, it holds that

$$f(\boldsymbol{x}^*(\boldsymbol{\theta}_2); \boldsymbol{\theta}_1) - \sum_{i=1}^m p_i \cdot g_i(\boldsymbol{x}^*(\boldsymbol{\theta}_2); \boldsymbol{\theta}_1) \geq f(\boldsymbol{x}^*(\boldsymbol{\theta}_2); \boldsymbol{\theta}_2) - \sum_{i=1}^m p_i \cdot g_i(\boldsymbol{x}^*(\boldsymbol{\theta}_2); \boldsymbol{\theta}_2) - \max\{1, \bar{p}\} \cdot (m+1)\rho(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$$

which follows the definition of $\rho(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ in (2). Thus, we get that

$$
\begin{aligned}
L_{\mathcal{Q}_1}(\boldsymbol{p}) &\geq \int_{\boldsymbol{\theta}_1 \in \Theta} \int_{\boldsymbol{\theta}_2 \in \Theta} \left\{ f(\boldsymbol{x}^*(\boldsymbol{\theta}_2); \boldsymbol{\theta}_2) - \sum_{i=1}^{m} p_i \cdot g_i(\boldsymbol{x}^*(\boldsymbol{\theta}_2); \boldsymbol{\theta}_2) \right\} d\mathcal{Q}_{1,2}^*(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \\
&\quad - \max\{1, \bar{p}\} \cdot (m+1) \int_{\boldsymbol{\theta}_1 \in \Theta} \int_{\boldsymbol{\theta}_2 \in \Theta} \rho(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) d\mathcal{Q}_{1,2}^*(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \\
&= \int_{\boldsymbol{\theta}_2 \in \Theta} \left\{ f(\boldsymbol{x}^*(\boldsymbol{\theta}_2); \boldsymbol{\theta}_2) - \sum_{i=1}^{m} p_i \cdot g_i(\boldsymbol{x}^*(\boldsymbol{\theta}_2); \boldsymbol{\theta}_2) \right\} d\mathcal{Q}_2(\boldsymbol{\theta}_2) - \max\{1, \bar{p}\} \cdot (m+1) \mathcal{W}(\mathcal{Q}_1, \mathcal{Q}_2) \\
&= L_{\mathcal{Q}_2}(\boldsymbol{p}) - \max\{1, \bar{p}\} \cdot (m+1) \mathcal{W}(\mathcal{Q}_1, \mathcal{Q}_2)
\end{aligned}
$$

where the first equality holds by noting that $\int_{\boldsymbol{\theta}_1 \in \Theta} d\mathcal{Q}_{1,2}^*(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = d\mathcal{Q}_2(\boldsymbol{\theta}_2)$. $\qquad\square$

As a remark, we note that the proof of Lemma 3 will still go through even when the concavity and convexity of $f(\cdot; \boldsymbol{\theta})$ and $g_i(\cdot; \boldsymbol{\theta})$ do not hold. To see this, we use F to denote a distribution over the action set $\mathcal{X}$ and accordingly,

$$
\hat{f}(F; \boldsymbol{\theta}) = \int_{\boldsymbol{x} \in \mathcal{X}} \hat{f}(\boldsymbol{x}; \boldsymbol{\theta}) dF(\boldsymbol{x}) \text{ and } \hat{g}_i(F; \boldsymbol{\theta}) = \int_{\boldsymbol{x} \in \mathcal{X}} g_i(\boldsymbol{x}; \boldsymbol{\theta}) dF(\boldsymbol{x}).
$$

Then, we denote

$$
\hat{h}(\boldsymbol{p}; \boldsymbol{\theta}) := \max_{F} \left\{ \hat{f}(F; \boldsymbol{\theta}) - \boldsymbol{p}^{\top} \hat{\boldsymbol{g}}(F; \boldsymbol{\theta}) \right\}, \quad \hat{L}_{\mathcal{Q}}(\boldsymbol{p}) = \mathcal{Q} \hat{h}(\boldsymbol{p}; \boldsymbol{\theta})
$$

and

$$
\hat{\rho}(\boldsymbol{\theta}, \boldsymbol{\theta}') = \sup_{F} \|(\hat{f}(F; \boldsymbol{\theta}), \hat{\boldsymbol{g}}(F; \boldsymbol{\theta})) - (\hat{f}(F; \boldsymbol{\theta}'), \hat{\boldsymbol{g}}(F; \boldsymbol{\theta}'))\|_{\infty}, \quad \hat{\mathcal{W}}(\mathcal{Q}_1, \mathcal{Q}_2) := \inf_{\mathcal{Q}_{1,2} \in \mathcal{J}(\mathcal{Q}_1, \mathcal{Q}_2)} \int \hat{\rho}(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) d\mathcal{Q}_{1,2}(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}).
$$

Now that $\hat{f}(F; \boldsymbol{\theta})$ and $\hat{\boldsymbol{g}}(F; \boldsymbol{\theta})$ can be regarded as linear functions of $(dF(\boldsymbol{x}), \forall \boldsymbol{x} \in \mathcal{X})$, which fully characterizes the distribution F, we can apply the same procedure as the proof of Lemma 3 to show that

$$
\sup_{\boldsymbol{p} \in \Omega_{\bar{p}}} \left| \hat{L}_{\mathcal{Q}_1}(\boldsymbol{p}) - \hat{L}_{\mathcal{Q}_2}(\boldsymbol{p}) \right| \leq \max\{1, \bar{p}\} \cdot (m+1) \hat{\mathcal{W}}(\mathcal{Q}_1, \mathcal{Q}_2).
$$

On the other hand, note that

$$
\hat{h}(\boldsymbol{p}; \boldsymbol{\theta}) := \max_{F} \left\{ \hat{f}(F; \boldsymbol{\theta}) - \boldsymbol{p}^{\top} \hat{\boldsymbol{g}}(F; \boldsymbol{\theta}) \right\} = \max_{\boldsymbol{x} \in \mathcal{X}} \left\{ f(\boldsymbol{x}; \boldsymbol{\theta}) - \boldsymbol{p}^{\top} \boldsymbol{g}(\boldsymbol{x}; \boldsymbol{\theta}) \right\} = h(\boldsymbol{p}; \boldsymbol{\theta})
$$

and $\hat{\rho}(\boldsymbol{\theta}, \boldsymbol{\theta}') = \rho(\boldsymbol{\theta}, \boldsymbol{\theta}')$. We know that

$$
\left| \hat{L}_{\mathcal{Q}_1}(\boldsymbol{p}) - \hat{L}_{\mathcal{Q}_2}(\boldsymbol{p}) \right| = |L_{\mathcal{Q}_1}(\boldsymbol{p}) - L_{\mathcal{Q}_2}(\boldsymbol{p})| \text{ and } \hat{\mathcal{W}}(\mathcal{Q}_1, \mathcal{Q}_2) = \mathcal{W}(\mathcal{Q}_1, \mathcal{Q}_2).
$$

Therefore, (12) can be proved to hold for general $f(\cdot; \boldsymbol{\theta})$ and $g_i(\cdot; \boldsymbol{\theta})$ without any convexity or concavity structure.

B3. Proof of Lemma 4

Proof: Note that the following two property is satisfied by the update rule (11):

(i). If $\|\boldsymbol{p}_t\|_\infty \le q$, then we must have $\|\boldsymbol{p}_{t+1}\|_\infty \le q+1$ by noting that for each i, the i-th component of $\boldsymbol{p}_t$, denoted as $p_{t,i}$, is nonnegative and $g_i(\cdot, \boldsymbol{\theta}_t)$ is normalized within $[0, 1]$.

(ii). If there exists i such that $p_{t,i} > q$, then we must have $p_{t+1,i} < p_{t,i}$. Specifically, when $p_{t,i} > q$, we must have that $g_i(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) = 0$, otherwise we would have that

$$f(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) - \boldsymbol{p}_t^\top \cdot \boldsymbol{g}(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) \le f(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) - p_{t,i} \cdot g_i(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) < 0$$

which contradicts the definition of $\tilde{\boldsymbol{x}}_t$ in $\text{IGD}(\hat{\boldsymbol{\gamma}})$ since we could always select $\tilde{\boldsymbol{x}}_t = \boldsymbol{0}$ to obtain 0 in the objective value. Then from the non-negativity of $\hat{\boldsymbol{c}}_t$, it holds that $p_{t+1,i} < p_{t,i}$ in (11).

Starting from $\boldsymbol{p}_1 = 0$ and iteratively applying the above two property to control the increase of $\boldsymbol{p}_t$ from $t = 1$ to T, we obtain that for the first time that one component of $\boldsymbol{p}_t$ exceeds the threshold q, it is upper bounded by $q+1$ and this component will continue to decrease until it falls below the threshold q. Thus, it is obvious that we have $\|\boldsymbol{p}_t\|_\infty \le q+1$ with probability 1 for each t. $\qquad\square$

B4. Proof of Theorem 3

Similar to case of known distribution, we define the following function $\hat{L}_t(\cdot)$, based on the prior estimate $\hat{\mathcal{P}}_t$.

$$\hat{L}_t(\boldsymbol{p}) := \hat{\boldsymbol{\gamma}}_t^\top \boldsymbol{p} + \hat{\mathcal{P}}_t h(\boldsymbol{p}; \boldsymbol{\theta}). \tag{29}$$

Then we have the following relation between $\hat{L}(\cdot)$ and $\hat{L}_t(\cdot)$. As its analysis is identical to Proposition 1, we omit its proof for simplicity.

Lemma 5 *For each $t = 1, ..., T$, it holds that*

$$\hat{\boldsymbol{p}}^* \in argmin_{\boldsymbol{p} \ge 0} \hat{L}_t(\boldsymbol{p}) \tag{30}$$

where $\hat{\boldsymbol{p}}^*$ is defined in (9) *as the minimizer of the function $\hat{L}(\cdot)$. Moreover, it holds that*

$$\hat{L}(\hat{\boldsymbol{p}}^*) = \sum_{t=1}^{T} \hat{L}_t(\hat{\boldsymbol{p}}^*). \tag{31}$$

Now we proof Theorem 3 and the idea of proof is similar to Theorem 1.

Proof: From the proof of Theorem 1, we have

$$\sum_{t=1}^{T} f(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) - \sum_{t=1}^{T} f(\boldsymbol{x}_t; \boldsymbol{\theta}_t) \le q \cdot \sum_{i=1}^{m} \sum_{t=1}^{T} g_i(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) \cdot \mathbb{I}\left\{ \sum_{j=1}^{t} g_i(\tilde{\boldsymbol{x}}_j; \boldsymbol{\theta}_j) > c_i \right\}$$

$$\le q \cdot \sum_{i=1}^{m} \left[\sum_{t=1}^{T} g_i(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) - (c_i - 1) \right]^+$$

which related the total collected reward by the true action $\{\boldsymbol{x}_t\}_{t=1}^T$ and the virtual action $\{\tilde{\boldsymbol{x}}_t\}_{t=1}^T$ of the algorithm IGD($\hat{\boldsymbol{\gamma}}$). Further from Proposition 1, we have that

$$\text{Reg}_T(\pi) \leq \min_{\boldsymbol{p} \geq 0} L(\boldsymbol{p}) - \mathbb{E}\left[\sum_{t=1}^T f(\boldsymbol{x}_t; \boldsymbol{\theta}_t)\right] \leq \underbrace{\min_{\boldsymbol{p} \geq 0} L(\boldsymbol{p}) - \mathbb{E}\left[\sum_{t=1}^T f(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t)\right]}_{\text{I}}$$

$$\underbrace{+ q \cdot \mathbb{E}\left[\sum_{i=1}^m \left[\sum_{t=1}^T g_i(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) - (c_i - 1)\right]^+\right]}_{\text{II}}$$

We then bound the term I and term II separately to derive our regret bound.

Bound I: Note that $\|\hat{\boldsymbol{p}}^*\|_\infty \leq q$, it follows directly from Lemma 3 and Lemma 5 that

$$\min_{\boldsymbol{p} \geq \boldsymbol{0}} L(\boldsymbol{p}) \leq L(\hat{\boldsymbol{p}}^*) \leq \hat{L}(\hat{\boldsymbol{p}}^*) + \max\{q, 1\} \cdot (m+1) \cdot W_T = \sum_{t=1}^T \hat{L}_t(\hat{\boldsymbol{p}}^*) + \max\{q, 1\} \cdot (m+1) \cdot W_T$$

Note that from Lemma 4, we have that for each t, $\|\boldsymbol{p}_t\|_\infty \leq (q+1)$ with probability 1. Further note that for each t, the distribution of $\boldsymbol{p}_t$ is independent from the distribution of $\boldsymbol{\theta}_t$, then from Lemma 3 and Lemma 5, we have that

$$\hat{L}_t(\hat{\boldsymbol{p}}^*) = \min_{\boldsymbol{p} \geq \boldsymbol{0}} \hat{L}_t(\boldsymbol{p}) \leq \mathbb{E}_{\boldsymbol{p}_t}\left[\hat{L}_t(\boldsymbol{p}_t)\right] \leq \mathbb{E}_{\boldsymbol{p}_t}\left[\hat{\boldsymbol{\gamma}}_t^\top \boldsymbol{p}_t + \mathcal{P}_t h(\boldsymbol{p}_t; \boldsymbol{\theta}_t)\right] + (q+1)(m+1) \cdot W(\mathcal{P}_t, \hat{\mathcal{P}})$$

where the expectation is taken with respect to the randomness of the dual price $\boldsymbol{p}_t$. Thus, we have

$$\text{I} \leq \sum_{t=1}^T \mathbb{E}_{\boldsymbol{p}_t}\left[\hat{\boldsymbol{\gamma}}_t^\top \boldsymbol{p}_t + \mathcal{P}_t\left\{h(\boldsymbol{p}_t; \boldsymbol{\theta}_t) - f(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t)\right\}\right] + 2(q+1)(m+1) \cdot W_T.$$

From the definition of $\tilde{\boldsymbol{x}}_t$, we get that $h(\boldsymbol{p}_t; \boldsymbol{\theta}_t) - f(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) = -\boldsymbol{p}_t^\top \cdot \boldsymbol{g}(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t)$, which implies that

$$\text{I} \leq \sum_{t=1}^T \mathbb{E}_{\boldsymbol{p}_t}\left[\boldsymbol{p}_t^\top \cdot (\hat{\boldsymbol{\gamma}}_t - \mathcal{P}_t \boldsymbol{g}(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t))\right] + 2(q+1)(m+1) \cdot W_T$$

Note that from the update rule (11), we have that

$$\|\boldsymbol{p}_{t+1}\|_2^2 \leq \|\boldsymbol{p}_t\|_2^2 + \frac{1}{T} \cdot \|\boldsymbol{g}(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) - \hat{\boldsymbol{\gamma}}_t\|_2^2 - \frac{2}{\sqrt{T}} \cdot \boldsymbol{p}_t^\top \cdot (\hat{\boldsymbol{\gamma}}_t - \boldsymbol{g}(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t))$$

which implies that

$$\mathbb{E}_{\boldsymbol{p}_t}\left[\boldsymbol{p}_t^\top \cdot (\hat{\boldsymbol{\gamma}}_t - \mathcal{P}_t \boldsymbol{g}(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t))\right] \leq \frac{\sqrt{T}}{2} \cdot \left(\mathbb{E}[\|\boldsymbol{p}_t\|_2^2] - \mathbb{E}[\|\boldsymbol{p}_{t+1}\|_2^2]\right) + \frac{m}{2\sqrt{T}}$$

Thus, it holds that

$$\text{I} \leq \frac{m\sqrt{T}}{2} + 2(q+1)(m+1) \cdot W_T \tag{32}$$

Bound II: Note that from the update rule (11), we have that

$$\sqrt{T} \cdot \boldsymbol{p}_{t+1} \geq \sqrt{T} \cdot \boldsymbol{p}_t + \boldsymbol{g}(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) - \hat{\boldsymbol{\gamma}}_t$$

which implies that

$$\sum_{t=1}^{T} \boldsymbol{g}(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) - \boldsymbol{c} \leq \sum_{t=1}^{T} \boldsymbol{g}(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) - \sum_{t=1}^{T} \hat{\boldsymbol{\gamma}}_t \leq \sqrt{T} \cdot \boldsymbol{p}_{T+1}$$

Thus, it holds that

$$\text{II} = q \cdot \mathbb{E} \left[\sum_{i=1}^{m} \left[\sum_{t=1}^{T} g_i(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) - (c_i - 1) \right]^{+} \right] \leq mq(q+1) \cdot \sqrt{T} + qm \tag{33}$$

We obtain the $O(\max\{\sqrt{T}, W_T\})$ regret bound immediately by combining (32) and (33). $\quad\square$

C. Proofs of Section 5

C1. Proof of Proposition 3

Proof: We consider the implementation of any online policy π on the two scenarios (13) and (14) for $\kappa = 1$, which is replicated as follows for completeness:

$$\max \quad x_1 + ... + x_c + 2x_{c+1} + ... + 2x_T \tag{34}$$
$$\text{s.t.} \quad x_1 + ... + x_c + x_{c+1} + ... + x_T \leq c$$
$$0 \leq x_t \leq 1 \quad \text{for } t = 1, ..., T.$$
$$\max \quad x_1 + ... + x_c \tag{35}$$
$$\text{s.t.} \quad x_1 + ... + x_c + x_{c+1} + ... + x_T \leq c$$
$$0 \leq x_t \leq 1 \quad \text{for } t = 1, ..., T.$$

where $c = \frac{T}{2}$. Denote $x_t^1(\pi)$ as the decision of policy π at period t for scenario (34) and denote $x_t^2(\pi)$ as the decision of policy π at period t for scenario (35). Further define $T_1(\pi)$ (resp. $T_2(\pi)$) as the expected capacity consumption of policy π on scenario (34) (resp. scenario (35)) during the first $\frac{T}{2}$ time periods:

$$T_1(\pi) = \mathbb{E} \left[\sum_{t=1}^{\frac{T}{2}} x_t^1(\pi) \right] \quad \text{and} \quad T_2(\pi) = \mathbb{E} \left[\sum_{t=1}^{\frac{T}{2}} x_t^2(\pi) \right]$$

Then, we have that

$$R_T^1(\pi) = T - T_1(\pi) \quad \text{and} \quad R_T^2(\pi) = T_2(\pi)$$

where $R_T^1(\pi)$ (resp. $R_T^2(\pi)$) denotes the expected reward collected by policy π on scenario (34) (resp. scenario (35)). Thus, the regret of policy π on scenario (34) and (35) are $T_1(\pi)$ and $T - T_2(\pi)$ respectively. Further note that since the implementation of policy π at each time period should be independent of the future information, we must have $T_1(\pi) = T_2(\pi)$. Thus, we have that

$$\text{Reg}_T(\pi) \geq \max\{T_1(\pi), T - T_1(\pi)\} \geq \frac{T}{2} = \Omega(T)$$

which completes our proof. $\quad\square$

C2. Proof of Theorem 4

The proof of the theorem can be directly obtained from Theorem 2.

C3. Proof of Theorem 5

We first prove the following lemma, which implies that the dual variable updated in (15) is always bounded. It derivation is essentially the same as Lemma 2, so we omit its proof for simplicity.

Lemma 6 *Under Assumption 1, for each $t = 1, 2, \ldots, T$, the dual price vector satisfies $\|\boldsymbol{p}_t\|_\infty \leq q + 1$, where $\boldsymbol{p}_t$ is specified by (15) in Algorithm 2 and the constant q is defined in Assumption 1 (c).*

Now we proceed to prove Theorem 5.

Proof: From the proof of Theorem 1, we have

$$\sum_{t=1}^{T} f(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) - \sum_{t=1}^{T} f(\boldsymbol{x}_t; \boldsymbol{\theta}_t) \leq q \cdot \sum_{i=1}^{m} \left[\sum_{t=1}^{T} g_i(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) - (c_i - 1) \right]^+$$

which relates the total collected reward by the true action $\{\boldsymbol{x}_t\}_{t=1}^{T}$ and the virtual action $\{\tilde{\boldsymbol{x}}_t\}_{t=1}^{T}$. Here $[\cdot]^+$ denotes the positive part function. Furthermore, from Proposition 1 and the feasibility, we have that

$$\text{Reg}_T(\pi) \leq \min_{\boldsymbol{p} \geq \boldsymbol{0}} L(\boldsymbol{p}) - \mathbb{E}\left[\sum_{t=1}^{T} f(\boldsymbol{x}_t; \boldsymbol{\theta}_t) \right] \leq \underbrace{\min_{\boldsymbol{p} \geq \boldsymbol{0}} L(\boldsymbol{p}) - \mathbb{E}\left[\sum_{t=1}^{T} f(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) \right]}_{\text{I}}$$

$$+ \underbrace{q \cdot \mathbb{E}\left[\sum_{i=1}^{m} \left[\sum_{t=1}^{T} g_i(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) - (c_i - 1) \right]^+ \right]}_{\text{II}}$$

Next, we bound the term I and term II separately to derive our regret bound.

Bound I: We first define the following function $\bar{L}(\cdot)$:

$$\bar{L}(\boldsymbol{p}) := \frac{1}{T} \boldsymbol{p}^\top \boldsymbol{c} + \bar{\mathcal{P}}_T h(\boldsymbol{p}; \boldsymbol{\theta})$$

Note that $\hat{\mathcal{P}}_T = \frac{1}{T} \sum_{t=1}^{T} \mathcal{P}_t$, it holds that $L(\boldsymbol{p}) = T \cdot \bar{L}(\boldsymbol{p})$ for any $\boldsymbol{p}$. From Lemma 6, we know that for each t, $\|\boldsymbol{p}_t\|_\infty \leq q + 1$ with probability 1. In addition, for each t, the distribution of $\boldsymbol{p}_t$ is independent from the distribution of $\boldsymbol{\theta}_t$, then from Lemma 3, we have that

$$\min_{\boldsymbol{p} \geq \boldsymbol{0}} \bar{L}(\boldsymbol{p}) \leq \mathbb{E}_{\boldsymbol{p}_t}\left[\bar{L}(\boldsymbol{p}_t) \right] \leq \mathbb{E}_{\boldsymbol{p}_t}\left[\frac{1}{T} \boldsymbol{p}_t^\top \boldsymbol{c} + \mathcal{P}_t h(\boldsymbol{p}_t; \boldsymbol{\theta}_t) \right] + (q+1)(m+1) \cdot \mathcal{W}(\mathcal{P}_t, \bar{\mathcal{P}}_T), \qquad (36)$$

where the expectation is taken with respect to $\boldsymbol{p}_t$ in a random realization of the algorithm. Thus, we have the first term

$$\text{I} \leq \sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{p}_t}\left[\frac{1}{T} \boldsymbol{c}^\top \boldsymbol{p}_t + \mathcal{P}_t \{ h(\boldsymbol{p}_t; \boldsymbol{\theta}_t) - f(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) \} \right] + (q+1)(m+1) \cdot \mathcal{W}(\mathcal{P}_t, \bar{\mathcal{P}}_T)$$

which comes from combining (36) with the relation $L(\boldsymbol{p}) = T \cdot \bar{L}(\boldsymbol{p})$. By the definition of $\tilde{\boldsymbol{x}}_t$, $h(\boldsymbol{p}_t; \boldsymbol{\theta}_t) - f(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) = -\boldsymbol{p}_t^\top \cdot \boldsymbol{g}(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t)$, which implies that

$$\text{I} \le \sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{p}_t} \left[\boldsymbol{p}_t^\top \cdot \left(\frac{\boldsymbol{c}}{T} - \mathcal{P}_t \boldsymbol{g}(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) \right) \right] + (q+1)(m+1) \cdot \mathcal{W}(\mathcal{P}_t, \bar{\mathcal{P}}_T) \tag{37}$$

Note that from the update rule (15), we have that

$$\|\boldsymbol{p}_{t+1}\|_2^2 \le \|\boldsymbol{p}_t\|_2^2 + \frac{1}{T} \cdot \|\boldsymbol{g}(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) - \frac{\boldsymbol{c}}{T}\|_2^2 - \frac{2}{\sqrt{T}} \cdot \boldsymbol{p}_t^\top \cdot \left(\frac{\boldsymbol{c}}{T} - \boldsymbol{g}(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) \right).$$

By taking expectation with respect to both sides,

$$\mathbb{E}_{\boldsymbol{p}_t} \left[\boldsymbol{p}_t^\top \cdot \left(\frac{\boldsymbol{c}}{T} - \mathcal{P}_t \boldsymbol{g}(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) \right) \right] \le \frac{\sqrt{T}}{2} \cdot \left(\mathbb{E}[\|\boldsymbol{p}_t\|_2^2] - \mathbb{E}[\|\boldsymbol{p}_{t+1}\|_2^2] \right) + \frac{m}{2\sqrt{T}} \tag{38}$$

Plugging (38) into (37), we obtain an upper bound on Term I,

$$\text{I} \le \frac{m\sqrt{T}}{2} + (q+1)(m+1) \cdot W_T \tag{39}$$

Bound II: Note that from the update rule (15), we have

$$\sqrt{T} \cdot \boldsymbol{p}_{t+1} \ge \sqrt{T} \cdot \boldsymbol{p}_t + \boldsymbol{g}(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) - \frac{\boldsymbol{c}}{T}$$

Taking a summation with respect to both sides,

$$\sum_{t=1}^{T} \boldsymbol{g}(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) - \boldsymbol{c} \le \sqrt{T} \cdot \boldsymbol{p}_{T+1}$$

Applying Lemma 6 for a bound on $\boldsymbol{p}_{T+1}$, we obtain the upper bound for Term II,

$$\text{II} = q \cdot \mathbb{E} \left[\sum_{i=1}^{m} \left[\sum_{t=1}^{T} g_i(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) - (c_i - 1) \right]^+ \right] \le mq(q+1) \cdot \sqrt{T} + qm \tag{40}$$

We obtain the desired regret bound by combining (39) and (40).   □

D. Proofs of Section 6

D1. Proof of Theorem 6

Proof: Note that the regret can be denoted as

$$\text{Reg}_T(\mathcal{H}, \pi_{\text{Resolve}}) = \mathbb{E}_{\mathcal{H} \sim \mathcal{P}}[R_1^*(\mathbf{c}_1, \mathcal{H}(1))] - \mathbb{E}_{\mathcal{H} \sim \mathcal{P}}[\sum_{t=1}^{T} r_t \cdot x_t]$$

$$= \sum_{t=1}^{T} \underbrace{\mathbb{E}_{\mathcal{H} \sim \mathcal{P}}[R_t^*(\mathbf{c}_t, \mathcal{H}(t)) - R_{t+1}^*(\mathbf{c}_{t+1}, \mathcal{H}(t+1)) - r_t \cdot x_t]}_{\text{I}_t} \tag{41}$$

Now we denote $\{x_j^*(\mathbf{c}_t, \mathcal{H}(t))\}$ as one optimal solution to $R_t^*(\mathbf{c}_t, \mathcal{H}(t))$. Clearly, we have the following: if $\boldsymbol{\theta}_t$ is realized as $\boldsymbol{\theta}^{(j)}$, i.e., $\mathcal{H}_j(t) = \mathcal{H}_j(t+1) + 1$ and $\mathcal{H}_{j'}(t) = \mathcal{H}_{j'}(t+1)$ for other j', it holds that

$$
\begin{aligned}
R_t^*(\mathbf{c}_t, \mathcal{H}(t)) &= r_j + R_{t+1}^*(\mathbf{c}_t - \boldsymbol{a}_j, \mathcal{H}(t+1)), && \text{if } x_j^*(\mathbf{c}_t, \mathcal{H}(t)) \geq 1 \\
R_t^*(\mathbf{c}_t, \mathcal{H}(t)) &= R_{t+1}^*(\mathbf{c}_t, \mathcal{H}(t+1)), && \text{if } \mathcal{H}_j(t) - x_j^*(\mathbf{c}_t, \mathcal{H}(t)) \geq 1
\end{aligned}
\tag{42}
$$

We also denote the event $\mathcal{A}_t = \{x_j^*(\mathbf{c}_t, \mathcal{H}(t)) \geq 1\}$ and the event $\mathcal{B}_t = \{\mathcal{H}_j(t) - x_j^*(\mathbf{c}_t, \mathcal{H}(t)) \geq 1\}$.

In the rest of the proof, we mainly analyze the term I_t in (41). We first focus on the case when $t \leq T + 1 - \frac{4(\alpha\beta_2 + \beta_2)}{p_{\min}} \cdot W_T - \frac{4}{p_{\min}}$, where W_T is the deviation budget from the estimates $\{\hat{\mathcal{P}}_t\}_{t=1}^T$ to the true distributions $\{\mathcal{P}_t\}_{t=1}^T$, $p_{\min}$ is given in Assumption 2, and α, β_2 are parameters to be determined later.

Case I: $\hat{x}_{j_t}(\mathbf{c}_t) \geq \frac{1}{2} \cdot \mathbb{E}_{\mathcal{H} \sim \hat{\mathcal{P}}}[\mathcal{H}_{j_t}(t)]$. Then we have $x_t = 1$, and from (42), we have

$$
\begin{aligned}
\mathrm{I}_t &= \mathbb{E}_{\mathcal{H} \sim \mathcal{P}}[R_t^*(\mathbf{c}_t, \mathcal{H}(t)) - R_{t+1}^*(\mathbf{c}_{t+1}, \mathcal{H}(t+1)) - r_t \cdot x_t] \\
&= \mathbb{E}\left[\mathbb{E}[r_{j_t} - r_{j_t} \cdot x_t + R_{t+1}^*(\mathbf{c}_t - \boldsymbol{a}_{j_t}, \mathcal{H}(t+1)) - R_{t+1}^*(\mathbf{c}_t - \boldsymbol{a}_{j_t} \cdot x_t, \mathcal{H}(t+1)) | \mathcal{A}_t]\right] \\
&\quad + \mathbb{E}\left[\mathbb{E}[-r_{j_t} \cdot x_t + R_{t+1}^*(\mathbf{c}_t, \mathcal{H}(t+1)) - R_{t+1}^*(\mathbf{c}_t - \boldsymbol{a}_{j_t} \cdot x_t, \mathcal{H}(t+1)) | \mathcal{A}_t^c]\right] \\
&= \mathbb{E}\left[\mathbb{E}[-r_{j_t} \cdot x_t + R_{t+1}^*(\mathbf{c}_t, \mathcal{H}(t+1)) - R_{t+1}^*(\mathbf{c}_t - \boldsymbol{a}_{j_t} \cdot x_t, \mathcal{H}(t+1)) | \mathcal{A}_t^c]\right] \\
&\leq P(\mathcal{A}_t^c)
\end{aligned}
$$

where $\mathcal{A}_t^c$ denotes the complementary event of $\mathcal{A}_t$. We now bound $P(\mathcal{A}_t^c)$.

From Theorem 2.4 in (Mangasarian and Shiau, 1987), there exists a constant $\alpha > 0$, which depends only on $\{(r_1, \boldsymbol{a}_1), \ldots, (r_n, \boldsymbol{a}_n)\}$, such that

$$
\|\boldsymbol{x}^*(\mathbf{c}_t, \mathcal{H}(t)) - \hat{\boldsymbol{x}}(\mathbf{c}_t)\|_\infty \leq \alpha \cdot \|\mathcal{H}(t) - \mathbb{E}_{\mathcal{H} \sim \hat{\mathcal{P}}}[\mathcal{H}(t)]\|_\infty.
$$

Moreover, from Hoeffdings' inequality, we know that

$$
P\left(\|\mathcal{H}(t) - \mathbb{E}_{\mathcal{H} \sim \hat{\mathcal{P}}}[\mathcal{H}(t)]\|_\infty \leq \frac{p_{\min} \cdot (T - t + 1)}{4\alpha}\right) \geq 1 - \exp(-\beta_1 \cdot (T - t + 1))
$$

for some constant $\beta_1 > 0$, and

$$
\|\mathbb{E}_{\mathcal{H} \sim \hat{\mathcal{P}}}[\mathcal{H}(t)] - \mathbb{E}_{\mathcal{H} \sim \mathcal{P}}[\mathcal{H}(t)]\|_\infty \leq \beta_2 \cdot W_T.
$$

Thus, we have that

$$
P\left(\|x^*(\mathbf{c}_t, \mathcal{H}(t)) - \hat{x}(\mathbf{c}_t)\|_\infty \leq \frac{p_{\min} \cdot (T - t + 1)}{4} + \alpha\beta_2 \cdot W_T\right) \geq 1 - \exp(-\beta_1 \cdot (T - t + 1))
$$

Since we have

$$
\hat{x}_{j_t}(\mathbf{c}_t) \geq \frac{1}{2} \cdot \mathbb{E}_{\mathcal{H} \sim \hat{\mathcal{P}}}[\mathcal{H}_{j_t}(t)] \geq \frac{1}{2} \cdot p_{\min} \cdot (T - t + 1) - \beta_2 \cdot W_T,
$$

we have that

$$P\left(x_{j_t}^*(\mathbf{c}_t, \mathcal{H}(t)) \geq \frac{p_{\min} \cdot (T - t + 1)}{4} - (\alpha\beta_2 + \beta_2) \cdot W_T \geq 1\right) \geq 1 - \exp(-\beta_1 \cdot (T - t + 1)).$$

when $t \leq T + 1 - \frac{4(\alpha\beta_2 + \beta_2)}{p_{\min}} \cdot W_T - \frac{4}{p_{\min}}$.

Thus, when $t \leq T + 1 - \frac{4(\alpha + 1)}{p_{\min}} \cdot W_T - \frac{4}{p_{\min}}$, we know that $P(\mathcal{A}_t^c) \leq \exp(-\beta_1 \cdot (T - t + 1))$.

Case II: $\hat{x}_{j_t}(\mathbf{c}_t) < \frac{1}{2} \cdot \mathbb{E}_{\mathcal{H} \sim \hat{\mathcal{P}}}[\mathcal{H}_{j_t}(t)]$. Then we have $x_t = 0$, and from (42), we have

$$\begin{aligned}
\mathrm{I}_t =& \mathbb{E}[R_t^*(\mathbf{c}_t, \mathcal{H}(t)) - R_{t+1}^*(\mathbf{c}_{t+1}, \mathcal{H}(t+1)) - r_t \cdot x_t] \\
=& \mathbb{E}\left[\mathbb{E}[-r_{j_t} \cdot x_t + R_{t+1}^*(\mathbf{c}_t, \mathcal{H}(t+1)) - R_{t+1}^*(\mathbf{c}_t - \mathbf{a}_{j_t} \cdot x_t, \mathcal{H}(t+1))|\mathcal{B}_t]\right] \\
&+ \mathbb{E}\left[\mathbb{E}[r_{j_t} - r_{j_t} \cdot x_t + R_{t+1}^*(\mathbf{c}_t - \mathbf{a}_{j_t}, \mathcal{H}(t+1)) - R_{t+1}^*(\mathbf{c}_t - \mathbf{a}_{j_t} \cdot x_t, \mathcal{H}(t+1))|\mathcal{B}_t^c]\right] \\
=& \mathbb{E}\left[\mathbb{E}[r_{j_t} + R_{t+1}^*(\mathbf{c}_t - \mathbf{a}_{j_t}, \mathcal{H}(t+1)) - R_{t+1}^*(\mathbf{c}_t, \mathcal{H}(t+1))|\mathcal{B}_t^c]\right] \\
\leq& P(\mathcal{B}_t^c)
\end{aligned}$$

Following the same approach, we have that $P(\mathcal{B}_t^c) \leq \exp(-\beta_1 \cdot (T - t + 1))$ when $t \leq T + 1 - \frac{4(\alpha\beta_2 + \beta_2)}{p_{\min}} \cdot W_T - \frac{4}{p_{\min}}$.

Thus, on both cases, we conclude that

$$\mathrm{I}_t \leq \exp(-\beta_1 \cdot (T - t + 1))$$

when $t \leq T + 1 - \frac{4(\alpha\beta_2 + \beta_2)}{p_{\min}} \cdot W_T - \frac{4}{p_{\min}}$. Thus, from (41), we have that

$$\begin{aligned}
\mathrm{Reg}_T(\mathcal{H}, \pi_{\mathrm{Resolve}}) \leq& \sum_{t=1}^{T+1-\frac{4(\alpha\beta_2+\beta_2)}{p_{\min}} \cdot W_T - \frac{4}{p_{\min}}} \mathrm{I}_t + \frac{4(\alpha\beta_2 + \beta_2)}{p_{\min}} \cdot W_T + \frac{4}{p_{\min}} \\
\leq& \sum_{t=1}^{T} \exp(-\beta_1 \cdot (T - t + 1)) + \frac{4(\alpha\beta_2 + \beta_2)}{p_{\min}} \cdot W_T + \frac{4}{p_{\min}} \\
=& \max\{O(1), O(W_T)\}
\end{aligned}$$

which completes our proof. $\square$

Remark: Note that Vera and Banerjee (2020); Banerjee and Freund (2020a,b) utilize the condition of $p_{\min}$ in Assumption 2 to control the non-stationarity of the distribution $\mathcal{P}_t$. To be specific, suppose the support set $\Theta = \{\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(n)}\}$ and for each $j = 1, \ldots, n$, we denote $H_j(t)$ as the number of times that $\boldsymbol{\theta}_\tau$ is realized as $\boldsymbol{\theta}_j$ for $\tau = t, \ldots, T$. Denote by $\mathcal{H}(t) = (\mathcal{H}_1(t), \ldots, \mathcal{H}_n(t))$. Then this concentration condition of $p_{\min}$ in Vera and Banerjee (2020); Banerjee and Freund (2020a,b) essentially requires that

$$P\left(\|\mathcal{H}(t) - \mathbb{E}[\mathcal{H}(t)]\| \geq \frac{\mathbb{E}[\mathcal{H}_j(t)]}{2\kappa_j^1}\right) \leq \frac{c_j}{(T-t)^2}, \quad \forall t \leq T - \kappa_j^2, \quad \forall j = 1, \ldots, n \qquad (43)$$

for some constants κ_j^1, κ_j^2 and c_j.

We now show through the following example for which our WBNB is sublinear in T and the example can not covered by the above concentration condition. The example can be constructed as follows. For $t = 1, \ldots, T - \sqrt{T}$, we have $\mathcal{P}_t = \mathcal{P}_1$ and for $t = T - \sqrt{T} + 1, \ldots, T$, we have $\mathcal{P}_t = \mathcal{P}_2$, where $\mathcal{P}_1$ and $\mathcal{P}_2$ are two different distributions and it is satisfied that $\mathcal{P}_2(\boldsymbol{\theta} = \boldsymbol{\theta}'_j) = 0$ for a j'. Clearly, (43) is not satisfied for $j = j'$, thus, the concentration condition does not hold. However, it is direct to check that the WBNB can be upper bounded by $O(\sqrt{T})$, which is sublinear in T.

D2. Analysis of the sub-optimality of static policies

D2.1. Proof of Proposition 4

Proof: We first describe a problem instance. Consider a single resource with initial capacity $c = 3T/4$. The parameter $\boldsymbol{\theta}_t = (r_t, 1)$; the functions $f(x_t, \boldsymbol{\theta}_t) = r_t \cdot x_t$ and $g(x_t, \boldsymbol{\theta}_t) = x_t$, where $x_t \in [0, 1]$ is the decision variable at time period t. Suppose that the prior estimate $\hat{\mathcal{P}}$ is given by

$$r_t = \begin{cases} 1, & \text{w.p. } \dfrac{3}{4} \\ \text{Unif}[0, 1], & \text{w.p. } \dfrac{1}{4} \end{cases}$$

for each t. Here $\text{Unif}[0, 1]$ denotes a uniform distribution over $[0, 1]$. Clearly, the deterministic upper bound under prior estimate $\hat{\mathcal{P}}$ takes a value of $\frac{3T}{4}$. Furthermore, we denote $\mathbb{E}[h_t^\pi(r)]$ as the expectation of the decision variable under the static policy π when the reward of period t is realized as r. Then, we make following claim.

Claim: There exists a $\hat{r} \in [1 - \frac{2W_T}{T}, 1 - \frac{W_T}{T}]$ such that $\sum_{t=1}^{3T/4} \mathbb{E}[h_t^\pi(\hat{r})] \leq T/4$.

Otherwise, suppose that for each $r \in [1 - \frac{2W_T}{T}, 1 - \frac{W_T}{T}]$, we have $\sum_{t=1}^{3T/4} \mathbb{E}[h_t^\pi(r)] \geq T/4$. Then, we compute the expected reward gained by the policy π from $r \in [1 - \frac{2W_T}{T}, 1 - \frac{W_T}{T}]$ during the first $3T/4$ periods, where the budget will never be violated. To be specific, note that the event $r \in [1 - \frac{2W_T}{T}, 1 - \frac{W_T}{T}]$ happens with probability $\frac{W_T}{4T}$ at each period t. Comparing with R_T^{UB}, this will cause a gap of at least $\frac{W_T}{T}$ for π to collect a reward $r \in [1 - \frac{2W_T}{T}, 1 - \frac{W_T}{T}]$. Thus, the gap of R_T^{UB} and π on $\hat{\mathcal{P}}$, caused from π obtaining reward from $r \in [1 - \frac{2W_T}{T}, 1 - \frac{W_T}{T}]$ during the first $3T/4$ periods is at least

$$\frac{W_T}{T} \cdot \frac{W_T}{4T} \cdot \frac{T}{4} = C_1 \cdot T^{1/2}$$

which violates the regret upper bound. Thus, the claim is proved.

Denote by $\hat{r}$ the value described in the claim. Now we construct a true distribution $\mathcal{P} = \{\mathcal{P}_1, \ldots, \mathcal{P}_T\}$ as follows:

$$r_t = \begin{cases} \hat{r}, & \text{w.p. } \dfrac{3}{4} \\ \text{Unif}[0, 1], & \text{w.p. } \dfrac{1}{4} \end{cases}$$

Clearly, the deviation budget is upper bounded by W_T. Then, we compute the total expected reward that policy π can collect on the true distribution $\mathcal{P}$. The expected reward collected by policy π during the first $3T/4$ periods is at most

$$\frac{1}{4} \cdot \frac{1}{2} \cdot \frac{3T}{4} + \frac{3}{4} \cdot \frac{T}{4} \cdot \hat{r} \leq \frac{9T}{32}$$

where the first term in the LHS denotes the upper bound of the expected reward collected from $\text{Unif}[0, 1]$, and the second term in the LHS denotes the upper bound of the expected reward collect from $\hat{r}$, which follows from the claim. Finally, without considering the budget violation and we let π to collect every reward during the last $T/4$ periods. The policy π can collect reward at most

$$\frac{T}{4} \cdot \left(\frac{1}{2} \cdot \frac{1}{4} + \frac{3}{4}\right) = \frac{7T}{32}$$

during the last $T/4$ periods. Thus, the policy π can collect at most $T/2$ reward on the true distribution $\mathcal{P}$ during the entire horizon. However, the value of $\mathbb{E}_{\mathcal{P}}[R_T^{\text{UB}}]$ is at least $\frac{3T}{4} \cdot \hat{r} \geq \frac{3T}{4} \cdot (1 - \frac{2W_T}{T}) \geq \frac{9T}{16}$, when $W_T \leq \frac{T}{8}$ which clearly holds since W_T grows sublinearly in T. Thus, the regret of policy π on the true distribution $\mathcal{P}$ is at least $\frac{T}{16}$.

D2.2. Discussion on Bid Price Policy

The well-known bid price policy (Talluri and Van Ryzin (1998)) computes a dual optimal solution (bid price) $\hat{\boldsymbol{p}}^*$ based on the prior estimates. Specifically, the policy makes the decision $\boldsymbol{x}_t$ at each period t based on the fixed $\hat{\boldsymbol{p}}^*$ throughout the procedure:

$$\boldsymbol{x}_t \in \text{argmax}_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x}, \boldsymbol{\theta}_t) - (\hat{\boldsymbol{p}}^*)^\top \cdot \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{\theta}_t).$$

When there is no deviation between the true distributions and prior estimates, i.e., $\mathcal{P}_t = \hat{\mathcal{P}}_t$ for each t, Talluri and Van Ryzin (1998) show that bid price policy is asymptotically optimal if each period is repeated sufficiently many times and the capacity is scaled up accordingly. The following example shows that the bid price policy may fail drastically even when the deviations between the true distributions and prior estimates are small. The example follows the same spirit as the lower bound examples in our paper.

Consider the following linear program as the underlying problem (PCP) for the online stochastic optimization problem:

$$\begin{aligned}
\max \quad & x_1 + ... + x_{2c} + \frac{1}{2}x_{2c+1} + ... + \frac{1}{2}x_T && (44)\\
\text{s.t.} \quad & x_1 + ... + x_{2c} + x_{2c+1} + ... + x_T \leq c \\
& 0 \leq x_t \leq 1 \quad \text{for } t = 1, ..., T.
\end{aligned}$$

where $c = \frac{T}{3}$ and without loss of generality, we assume c is an integer. Suppose that the prior estimates for the coefficients in the objective function is larger than the true coefficients by 2ϵ for the first $\frac{T}{3}$ time periods and by ϵ for the last $\frac{2T}{3}$ time periods. Then we obtain the following linear program based on the prior estimates.

$$\max \quad (1+2\epsilon)x_1 + ... + (1+2\epsilon)x_c + (1+\epsilon)x_{c+1} + ... + (1+\epsilon)x_{2c} + (\frac{1}{2} + \epsilon)x_{2c+1} + ... + (\frac{1}{2} + \epsilon)x_T$$
$$\text{s.t.} \quad x_1 + ... + x_{2c} + x_{2c+1} + ... + x_T \leq c \tag{45}$$
$$0 \leq x_t \leq 1 \quad \text{for } t = 1, ..., T.$$

Obviously, the optimal dual solution for (45) can take any value in $(1+\epsilon, 1+2\epsilon)$. When we apply such a bid price policy to the true problem (44), the policy will set $x_t = 0$ throughout the horizon as the reward per time period under the true problem (44) is no greater than 1. Given the optimal objective value is $\frac{T}{3}$, the bid price policy will incur a regret of $\frac{T}{3}$.

As a remark, we note that the regret bound for our algorithm IGD($\hat{\gamma}$) is upper bounded by $2\epsilon T + \sqrt{T}$ which can be much smaller than $\frac{T}{3}$ for small ϵ.

D3. Proof of Theorem 7

The proof follows the same argument as that of the previous upper bounds in Theorem 3 and Theorem 5. We prove the details for completeness.

Proof: From the proof of Theorem 1, we have

$$\sum_{t=1}^{T} f(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) - \sum_{t=1}^{T} f(\boldsymbol{x}_t; \boldsymbol{\theta}_t) \leq q \cdot \sum_{i=1}^{m} \left[\sum_{t=1}^{T} g_i(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) - (c_i - 1) \right]^+$$

which relates the total collected reward by the true action $\{\boldsymbol{x}_t\}_{t=1}^{T}$ and the virtual action $\{\tilde{\boldsymbol{x}}_t\}_{t=1}^{T}$. Here $[\cdot]^+$ denotes the positive part function. Furthermore, from Proposition 1, we have that

$$\text{Reg}_T(\pi) \leq \min_{\boldsymbol{p} \geq \boldsymbol{0}} L(\boldsymbol{p}) - \mathbb{E}\left[\sum_{t=1}^{T} f(\boldsymbol{x}_t; \boldsymbol{\theta}_t) \right] \leq \underbrace{\min_{\boldsymbol{p} \geq \boldsymbol{0}} L(\boldsymbol{p}) - \mathbb{E}\left[\sum_{t=1}^{T} f(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) \right]}_{\text{I}}$$
$$+ \underbrace{q \cdot \mathbb{E}\left[\sum_{i=1}^{m} \left[\sum_{t=1}^{T} g_i(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) - (c_i - 1) \right]^+ \right]}_{\text{II}}$$

where the function $L(\cdot)$ is given in (17). Next, we bound the term I and term II separately to derive our regret bound.

Bound I: We first define the following function $\bar{L}(\cdot)$:

$$\bar{L}(\boldsymbol{p}) := \frac{1}{T}\boldsymbol{p}^\top \boldsymbol{c} + \bar{\mathcal{P}}_T h(\boldsymbol{p}; \boldsymbol{\theta}).$$

Note that $\hat{\mathcal{P}}_T = \frac{1}{T}\sum_{t=1}^{T}\mathcal{P}_t$, it holds that $L(\boldsymbol{p}) = T \cdot \bar{L}(\boldsymbol{p})$ for any $\boldsymbol{p}$. From Lemma 2, we know that for each t, $\|\boldsymbol{p}_t\|_\infty \leq q+1$ with probability 1. In addition, for each t, the distribution of $\boldsymbol{p}_t$ is independent from the distribution of $\boldsymbol{\theta}_t$, then from Lemma 3, we have that

$$\min_{\boldsymbol{p}\geq\boldsymbol{0}}\bar{L}(\boldsymbol{p}) \leq \mathbb{E}_{\boldsymbol{p}_t}\left[\bar{L}(\boldsymbol{p}_t)\right] \leq \mathbb{E}_{\boldsymbol{p}_t}\left[\frac{1}{T}\boldsymbol{p}_t^\top\boldsymbol{c} + \mathcal{P}_t h(\boldsymbol{p}_t;\boldsymbol{\theta}_t)\right] + (q+1)(m+1)\cdot\mathcal{W}(\mathcal{P}_t,\bar{\mathcal{P}}_T), \qquad (46)$$

where the expectation is taken with respect to $\boldsymbol{p}_t$ in a random realization of the algorithm. Thus, we have the first term

$$\mathrm{I} \leq \sum_{t=1}^{T}\mathbb{E}_{\boldsymbol{p}_t}\left[\frac{1}{T}\boldsymbol{c}^\top\boldsymbol{p}_t + \mathcal{P}_t\left\{h(\boldsymbol{p}_t;\boldsymbol{\theta}_t) - \hat{f}(\tilde{\boldsymbol{x}}_t;\boldsymbol{\theta}_t)\right\}\right] + (q+1)(m+1)\cdot\mathcal{W}(\mathcal{P}_t,\bar{\mathcal{P}}_T)$$

which comes from first taking expectation over $f(\tilde{\boldsymbol{x}}_t;\boldsymbol{\theta}_t)$ for given $\tilde{\boldsymbol{x}}_t,\boldsymbol{\theta}_t$, and then combining (46) with the relation $L(\boldsymbol{p}) = T \cdot \bar{L}(\boldsymbol{p})$. By the definition of $\tilde{\boldsymbol{x}}_t$, $h(\boldsymbol{p}_t;\boldsymbol{\theta}_t) - \hat{f}(\tilde{\boldsymbol{x}}_t;\boldsymbol{\theta}_t) = -\boldsymbol{p}_t^\top\cdot\hat{\boldsymbol{g}}(\tilde{\boldsymbol{x}}_t;\boldsymbol{\theta}_t)$, which implies that

$$\mathrm{I} \leq \sum_{t=1}^{T}\mathbb{E}_{\boldsymbol{p}_t}\left[\boldsymbol{p}_t^\top\cdot\left(\frac{\boldsymbol{c}}{T} - \mathcal{P}_t\hat{\boldsymbol{g}}(\tilde{\boldsymbol{x}}_t;\boldsymbol{\theta}_t)\right)\right] + (q+1)(m+1)\cdot\mathcal{W}(\mathcal{P}_t,\bar{\mathcal{P}}_T) \qquad (47)$$

Note that from the update rule (15), we have that

$$\|\boldsymbol{p}_{t+1}\|_2^2 \leq \|\boldsymbol{p}_t\|_2^2 + \frac{1}{T}\cdot\left\|\boldsymbol{g}(\tilde{\boldsymbol{x}}_t;\boldsymbol{\theta}_t) - \frac{\boldsymbol{c}}{T}\right\|_2^2 - \frac{2}{\sqrt{T}}\cdot\boldsymbol{p}_t^\top\cdot\left(\frac{\boldsymbol{c}}{T} - \boldsymbol{g}(\tilde{\boldsymbol{x}}_t;\boldsymbol{\theta}_t)\right).$$

By taking expectation with respect to both sides,

$$\mathbb{E}_{\boldsymbol{p}_t}\left[\boldsymbol{p}_t^\top\cdot\left(\frac{\boldsymbol{c}}{T} - \mathcal{P}_t\hat{\boldsymbol{g}}(\tilde{\boldsymbol{x}}_t;\boldsymbol{\theta}_t)\right)\right] \leq \frac{\sqrt{T}}{2}\cdot\left(\mathbb{E}[\|\boldsymbol{p}_t\|_2^2] - \mathbb{E}[\|\boldsymbol{p}_{t+1}\|_2^2]\right) + \frac{m}{2\sqrt{T}} \qquad (48)$$

Plugging (48) into (47), we obtain an upper bound on Term I,

$$\mathrm{I} \leq \frac{m\sqrt{T}}{2} + (q+1)(m+1)\cdot W_T \qquad (49)$$

Bound II: Note that from the update rule (15), we have

$$\sqrt{T}\cdot\boldsymbol{p}_{t+1} \geq \sqrt{T}\cdot\boldsymbol{p}_t + \boldsymbol{g}(\tilde{\boldsymbol{x}}_t;\boldsymbol{\theta}_t) - \frac{\boldsymbol{c}}{T}$$

Taking a summation with respect to both sides,

$$\sum_{t=1}^{T}\boldsymbol{g}(\tilde{\boldsymbol{x}}_t;\boldsymbol{\theta}_t) - \boldsymbol{c} \leq \sqrt{T}\cdot\boldsymbol{p}_{T+1}$$

Applying Lemma 2 for a bound on $\boldsymbol{p}_{T+1}$, we obtain the upper bound for Term II,

$$\mathrm{II} = q\cdot\mathbb{E}\left[\sum_{i=1}^{m}\left[\sum_{t=1}^{T}g_i(\tilde{\boldsymbol{x}}_t;\boldsymbol{\theta}_t) - (c_i - 1)\right]^+\right] \leq mq(q+1)\cdot\sqrt{T} + qm \qquad (50)$$

We obtain the desired regret bound by combining (49) and (50). $\quad\square$

D4. Proof of Theorem 8

The proof follows the same argument as that of the previous upper bounds in Theorem 3 and Theorem 5. We prove the details for completeness.

We first prove the following lemma, which implies that the dual variable updated in (21) is always bounded. The analysis is similar to that of Lemma 2.

Lemma 7 *Under Assumption 1, for each $l = 1, 2, \ldots, \lfloor T/K \rfloor$, the dual price vector satisfies $\|\boldsymbol{p}_l\|_\infty \leq q + \alpha_{T,K} \cdot K$, where $\boldsymbol{p}_l$ is specified by (21) in Algorithm 4 and the constant q is defined in Assumption 1 (c).*

Proof: Note that the following two properties are satisfied by the update rule (21):

(i). If $\|\boldsymbol{p}_l\|_\infty \leq q$, then we must have $\|\boldsymbol{p}_{l+1}\|_\infty \leq q + \alpha_{T,K} \cdot K$ by noting that for each i, the i-th component of $\boldsymbol{p}_l$, denoted as $p_{l,i}$, is nonnegative and $g_i(\cdot, \boldsymbol{\theta}_t)$ is normalized within $[0,1]$ for each $t = (l-1)K + 1, \ldots, lK$.

(ii). If there exists i such that $p_{l,i} > q$, then we must have $p_{l+1,i} < p_{l,i}$. Specifically, when $p_{l,i} > q$, we must have that $g_i(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) = 0$ for each $t = (l-1)K + 1, \ldots, lK$, otherwise, we would have that

$$f(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) - \boldsymbol{p}_l^\top \cdot \boldsymbol{g}(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) \leq f(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) - p_{l,i} \cdot g_i(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) < 0$$

which contradicts the definition of $\tilde{\boldsymbol{x}}_t$ in Algorithm 4 since we could always select $\boldsymbol{x}_t = \boldsymbol{0}$ to obtain a zero objective value as per Assumption 1. Then from (21), it holds that $p_{l+1,i} < p_{l,i}$.

Starting from $\boldsymbol{p}_1 = \boldsymbol{0}$ and iteratively applying the above two property to control the increase of $\boldsymbol{p}_l$ from $l = 1$ to $\lfloor T/K \rfloor$, we obtain that for the first time that one component of $\boldsymbol{p}_l$ exceeds the threshold q, it is upper bounded by $q + \alpha_{T,K} \cdot K$ and this component will continue to decrease until it falls below the threshold q. Thus, it is obvious that we have $\|\boldsymbol{p}_l\|_\infty \leq q + \alpha_{T,K} \cdot K$ with probability 1 for each l. $\square$

Now we proceed to prove Theorem 8.

Proof: From the proof of Theorem 1, we have

$$\text{Reg}_T(\pi) \leq \underbrace{\sum_{t=1}^T \min_{\boldsymbol{p} \geq 0} L_t(\boldsymbol{p}) - \mathbb{E}\left[\sum_{t=1}^T f(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t)\right]}_{\text{I}} + \underbrace{q \cdot \mathbb{E}\left[\sum_{i=1}^m \left[\sum_{t=1}^T g_i(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) - (c_i - 1)\right]^+\right]}_{\text{II}}$$

We then bound the term I and term II separately to derive our regret bound.

Bound I: Note that for each t, the distribution of $\boldsymbol{p}_l$ is independent from the distribution of $\boldsymbol{\theta}_t$ for any $t = (l-1)K + 1, \ldots, lK$, then we have that

$$\min_{\boldsymbol{p} \geq 0} L_t(\boldsymbol{p}) \leq \mathbb{E}_{\boldsymbol{p}_l}[L_t(\boldsymbol{p}_l)] = \mathbb{E}_{\boldsymbol{p}_l}\left[\boldsymbol{\gamma}_t^\top \boldsymbol{p}_l + \mathcal{P}_t h(\boldsymbol{p}_l; \boldsymbol{\theta}_t)\right]$$

where the expectation is taken with respect to the randomness of the dual price $\boldsymbol{p}_t$. Thus, we have

$$\text{I} \leq \sum_{l=1}^{\lfloor T/K \rfloor} \sum_{t=(l-1)K+1}^{lK \wedge T} \mathbb{E}_{\boldsymbol{p}_l} \left[\boldsymbol{\gamma}_t^\top \boldsymbol{p}_l + \mathcal{P}_t \{ h(\boldsymbol{p}_l; \boldsymbol{\theta}_t) - f(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) \} \right]$$

From the definition of $\tilde{\boldsymbol{x}}_t$, we get that $h(\boldsymbol{p}_t; \boldsymbol{\theta}_t) - f(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) = -\boldsymbol{p}_l^\top \cdot \boldsymbol{g}(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t)$ for $t = (l-1)K + 1, \ldots, lK$, which implies that

$$\text{I} \leq \sum_{l=1}^{\lfloor T/K \rfloor} \sum_{t=(l-1)K+1}^{lK \wedge T} \mathbb{E}_{\boldsymbol{p}_l} \left[\boldsymbol{p}_l^\top \cdot (\boldsymbol{\gamma}_t - \mathcal{P}_t \boldsymbol{g}(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t)) \right]$$

Note that from the update rule (21), we have that

$$\|\boldsymbol{p}_{l+1}\|_2^2 \leq \|\boldsymbol{p}_l\|_2^2 + \alpha_{T,K}^2 \cdot \| \sum_{t=(l-1)K+1}^{lK} \boldsymbol{g}(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) - \boldsymbol{\gamma}_t \|_2^2 - 2\alpha_{T,K} \cdot \sum_{t=(l-1)K+1}^{lK \wedge T} \boldsymbol{p}_t^\top \cdot (\boldsymbol{\gamma}_t - \boldsymbol{g}(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t))$$

which implies that

$$\sum_{t=(l-1)K+1}^{lK \wedge T} \mathbb{E}_{\boldsymbol{p}_l} \left[\boldsymbol{p}_l^\top \cdot (\boldsymbol{\gamma}_t - \mathcal{P}_t \boldsymbol{g}(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t)) \right] \leq \frac{1}{2\alpha_{T,K}} \cdot \left(\mathbb{E}[\|\boldsymbol{p}_l\|_2^2] - \mathbb{E}[\|\boldsymbol{p}_{l+1}\|_2^2] \right) + \frac{mK^2 \cdot \alpha_{T,K}}{2}$$

Thus, it holds that

$$\text{I} \leq \frac{mTK \cdot \alpha_{T,K}}{2} \tag{51}$$

Bound II: Note that from the update rule (21), we have that

$$\frac{1}{\alpha_{T,K}} \cdot \boldsymbol{p}_{l+1} \geq \frac{1}{\alpha_{T,K}} \cdot \boldsymbol{p}_l + \sum_{t=(l-1)K+1}^{lK} (\boldsymbol{g}(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) - \boldsymbol{\gamma}_t)$$

which implies that

$$\sum_{t=1}^{T} \boldsymbol{g}(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) - \boldsymbol{c} \leq \sum_{l=1}^{\lfloor T/K \rfloor} \sum_{t=(l-1)K+1}^{lK \wedge T} (\boldsymbol{g}(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) - \boldsymbol{\gamma}_t) \leq \frac{1}{\alpha_{T,K}} \cdot \boldsymbol{p}_{\lfloor T/K \rfloor + 1}$$

Thus, it holds that

$$\text{II} = q \cdot \mathbb{E} \left[\sum_{i=1}^{m} \left[\sum_{t=1}^{T} g_i(\tilde{\boldsymbol{x}}_t; \boldsymbol{\theta}_t) - (c_i - 1) \right]^+ \right] \leq mq(q + \alpha_{T,K} \cdot K) \cdot \frac{1}{\alpha_{T,K}} + qm \tag{52}$$

We obtain the $O(TK\alpha_{T,K} + \frac{1}{\alpha_{T,K}} + K)$ regret bound immediately by combining (51) and (52). $\quad\square$

E. More Numerical Results for Experiment I

		$\alpha = 0$	$\alpha = 0.01$	$\alpha = 0.02$	$\alpha = 0.04$
Upper Bound		537000	550158	559358	571775
$\beta = 0$	IGDP	520173(96.9%)	528884(96.1%)	535678(95.8%)	542302(94.9%)
	Re-solve (200)	524233(97.6%)	536703(97.6%)	543961(97.3%)	555291(97.1%)
	Re-solve (100)	523685(97.5%)	536427(97.5%)	544318(97.3%)	556507(97.3%)
	Re-solve (50)	523700(97.5%)	537134(97.6%)	544105(97.3%)	555119(97.1%)
	Re-solve (1)	527839(98.3%)	540176(98.2%)	547443(97.9%)	558640(97.7%)
$\beta = 0.01$	IGDP	518718(96.6%)	528611(96.1%)	534221(95.6%)	543138(95.0%)
	Re-solve (200)	524267(97.6%)	536168(97.6%)	545057(97.5%)	555260(97.1%)
	Re-solve (100)	524438(97.7%)	537389(97.7%)	544832(97.5%)	557295(97.5%)
	Re-solve (50)	525178(97.8%)	537462(97.7%)	545845(97.7%)	556276(97.3%)
	Re-solve (1)	529752(98.7%)	541583(98.5%)	549673(98.3%)	560308(98.0%)
$\beta = 0.02$	IGDP	518287(96.5%)	527937(96.0%)	533538(95.5%)	545654(95.5%)
	Re-solve (200)	523689(97.5%)	534737(97.3%)	543426(97.2%)	556498(97.4%)
	Re-solve (100)	525031(97.8%)	537302(97.7%)	545829(97.7%)	555397(97.3%)
	Re-solve (50)	525356(97.8%)	537465(97.8%)	544992(97.5%)	555949(97.3%)
	Re-solve (1)	529732(98.7%)	541779(98.5%)	550519(98.5%)	560719(98.2%)
$\beta = 0.04$	IGDP	518405(96.5%)	528083(96.0%)	536737(96.0%)	545405(95.5%)
	Re-solve (200)	524127(97.6%)	535318(97.4%)	544818(97.5%)	554169(97.0%)
	Re-solve (100)	523991(97.6%)	536770(97.6%)	544874(97.5%)	556560(97.5%)
	Re-solve (50)	525299(97.8%)	536481(97.6%)	545335(97.6%)	556776(97.5%)
	Re-solve (1)	530349(98.8%)	541633(98.5%)	550063(98.4%)	560365(98.1%)

Table 2   The numerical experiment based on the network revenue management problem (Jasin (2015)). The results are obtained for $T = 1000$ and 100 trails. Re-solve (a) denotes the frequency of re-solving, namely, re-solving the upper bound problem to update dual variable and $\hat{\gamma}_t$ for every a periods. For each entry b(c) of the table, b denotes the expected reward collected by the algorithm and c denotes the percentage of the expected reward of the algorithm over the upper bound.

		$\alpha = 1$	$\alpha = 1.5$	$\alpha = 2$	$\alpha = 2.5$	$\alpha = 3$
Upper Bound		282.5433	363.7044	459.7807	563.3545	670.5960
$\beta = 0$	IGDP	270.2411 (96%)	349.1769 (96%)	441.6677 (96%)	543.3373 (96%)	645.6582 (96%)
	UGD	270.3621(96%)	337.3192 (93%)	403.7044 (88%)	469.7643 (83%)	535.0654 (80%)
	FBP	270.1211 (96%)	347.4997 (96%)	439.7016 (96%)	539.9865 (96%)	642.3940 (96%)
	O2O	270.6872 (96%)	351.2534 (96%)	440.9020 (96%)	545.2106 (96%)	644.4059 (96%)
$\beta = 0.5$	IGDP	270.1595 (96%)	347.9148 (96%)	439.6166 (96%)	539.8719 (96%)	643.6777 (96%)
	UGD	270.3568 (96%)	338.6916 (93%)	405.7927 (88%)	473.6640 (84%)	540.0894 (81%)
	FBP	64.7526 (23%)	174.2642 (48%)	314.7806 (68%)	446.7646 (79%)	582.7744 (87%)
	O2O	202.3966 (72%)	289.6858 (80%)	423.5247 (92%)	536.0362 (95%)	621.4144 (93%)
$\beta = 1$	IGDP	269.8058(95%)	347.1246 (95%)	437.6279 (95%)	535.3521 (95%)	638.8322 (95%)
	UGD	269.6893 (95%)	339.1676 (93%)	408.3862 (89%)	477.2329 (84%)	544.9401 (81%)
	FBP	4.8549 (2%)	53.7881 (15%)	188.3271 (41%)	340.850 (61%)	486.4006 (73%)
	O2O	202.6792 (72%)	202.9555 (56%)	330.5638 (72%)	480.2536 (85%)	620.2584 (92%)
$\beta = 2$	IGDP	265.1512 (94%)	343.7802 (95%)	432.2275 (94%)	527.4351 (94%)	627.7440 (94%)
	UGD	265.4187 (94%)	337.4751 (93%)	410.3510 (89%)	482.8652 (86%)	554.0038 (83%)
	FBP	0.0171 (0%)	1.6210 (0.5%)	22.5201 (5%)	104.5026 (19%)	243.1575 (36%)
	O2O	203.5851 (72%)	206.4316 (57%)	210.6264 (46%)	361.1129 (64%)	514.6730 (77%)

Table 3 Computation results for Experiment I under the uniform setting: The results are reported based on 500 simulation trials. For each entry b(c) of the table, b denotes the expected reward collected by the algorithm and c denotes the percentage of the expected reward of the algorithm over the upper bound.

		$\alpha = 1$	$\alpha = 1.5$	$\alpha = 2$	$\alpha = 2.5$	$\alpha = 3$
Upper Bound		705.1450	803.5559	921.6550	1060.5567	1213.3552
$\beta = 0$	IGDP	679.9687 (96%)	771.9511 (96%)	884.2661 (96%)	1022.9439 (96%)	1171.1208 (97%)
	UGD	681.1875 (97%)	764.6037 (95%)	840.5685 (91%)	928.6610 (88%)	1018.0635 (84%)
	FBP	674.968 (96%)	768.0327 (96%)	883.1946 (96%)	1013.5706 (96%)	1165.1827 (96%)
	O2O	676.3823 (96%)	769.396 (96%)	882.5831 (96%)	1017.4399 (96%)	1159.9922(96%)
$\beta = 0.5$	IGDP	680.4068 (97%)	773.4652 (96%)	890.2161 (96%)	1023.3161 (97%)	1170.2227 (96%)
	UGD	680.931 (97%)	765.1636 (95%)	851.4681 (92%)	934.5877 (88%)	1028.4509 (85%)
	FBP	459.16 (65%)	525.5739 (66%)	619.198 (67%)	751.7758 (71%)	907.9537 (75%)
	O2O	656.6648 (93%)	712.215 (89%)	784.1766 (85%)	873.1237(82%)	996.943 (82%)
$\beta = 1$	IGDP	680.5808 (97%)	777.7524 (97%)	888.5206 (96%)	1024.4519 (97%)	1175.2438 (97%)
	UGD	681.824 (97%)	766.3631 (95%)	851.6300 (92%)	944.0707 (89%)	1035.5280 (85%)
	FBP	254.2547 (36%)	296.1898 (37%)	375.3518 (41%)	476.4165 (45%)	610.2982 (50%)
	O2O	588.1039 (83%)	744.5618 (93%)	833.9981 (90%)	916.2876 (86%)	1009.1384 (83%)
$\beta = 2$	IGDP	666.6237 (94%)	764.0318 (95%)	873.8978 (95%)	1014.7539 (96%)	1160.7693 (96%)
	UGD	667.4160 (95%)	756.7188 (94%)	851.0122 (92%)	944.1322 (89%)	1046.7647 (86%)
	FBP	51.9166 (7%)	68.2078(8%)	100.4066 (11%)	148.1866 (14%)	212.5127 (18%)
	O2O	466.4348 (66%)	535.7102 (67%)	728.6945 (79%)	957.5749 (90%)	1072.6659(88%)

Table 4   Computation results for Experiment I under the normal setting: The results are reported based on 500

simulation trials. For each entry b(c) of the table, b denotes the expected reward collected by the algorithm and c

denotes the percentage of the expected reward of the algorithm over the upper bound.

		$\alpha = 1$	$\alpha = 1.5$	$\alpha = 2$	$\alpha = 2.5$	$\alpha = 3$
Upper Bound		532.6379	630.1063	746.5027	871.63281	1010.7956
$\beta = 0$	IGDP	513.2625 (96%)	609.6643 (97%)	717.3434 (96%)	840.7550 (96%)	973.4778 (96%)
	UGD	513.9690 (97%)	592.9739 (94%)	672.7648 (90%)	758.0174 (87%)	840.9000 (83%)
	FBP	511.9639 (96%)	601.4187 (95%)	714.4849 (96%)	831.7872 (95%)	964.0174 (95%)
	O2O	512.9872 (96%)	603.5403 (96%)	713.4253 (96%)	834.2972 (96%)	970.8423 (96%)
$\beta = 0.5$	IGDP	517.2177 (97%)	607.7310 (96%)	714.3025 (96%)	843.8988 (97%)	975.2676 (97%)
	UGD	514.1228 (96%)	596.0002 (95%)	678.3624 (91%)	762.7194 (87%)	844.0139 (84%)
	FBP	372.5857 (70%)	441.8714 (70%)	542.3811 (73%)	667.7373 (77%)	821.8302 (82%)
	O2O	478.3224 (90%)	600.7489 (95%)	681.4778 (91%)	774.0669 (89%)	881.9813 (87%)
$\beta = 1$	IGDP	517.8502 (97%)	609.3739 (97%)	717.3040 (96%)	842.7028 (97%)	977.9344 (97%)
	UGD	515.1376 (97%)	598.2675 (95%)	683.0459 (92%)	769.4935 (88%)	853.1061 (85%)
	FBP	253.2191 (47%)	295.4627 (47%)	372.9051 (50%)	469.2851 (54%)	613.1751 (61%)
	O2O	391.8374 (73%)	509.6126 (81%)	674.1803 (90%)	813.8023 (93%)	912.8408 (91%)
$\beta = 2$	IGDP	507.9654 (95%)	598.2148 (95%)	713.2802 (96%)	835.5463 (96%)	963.8020 (95%)
	UGD	505.8237 (95%)	591.2771 (94%)	684.4290 (92%)	770.5260 (88%)	866.7157 (86%)
	FBP	78.3954 (15%)	96.3983 (15%)	136.5682 (18%)	197.7166 (23%)	281.8724 (28%)
	O2O	343.4216 (64%)	358.0907 (57%)	494.0880 (66%)	686.6650 (79%)	890.9912 (88%)

Table 5 Computation results for Experiment I under the mixed setting: The results are reported based on 500 simulation trials. For each entry b(c) of the table, b denotes the expected reward collected by the algorithm and c denotes the percentage of the expected reward of the algorithm over the upper bound.