# Constrained Online Two-stage Stochastic Optimization: Algorithm with (and without) Predictions

Piao Hu[†], Jiashuo Jiang[†], Guodong Lyu[†], Hao Su[†]

† Hong Kong University of Science and Technology

We consider an online two-stage stochastic optimization with long-term constraints over a finite horizon of $T$ periods. At each period, we take the first-stage action, observe a model parameter realization and then take the second-stage action from a feasible set that depends both on the first-stage decision and the model parameter. We aim to minimize the cumulative objective value while guaranteeing that the long-term average second-stage decision belongs to a set. We develop online algorithms for the online two-stage problem from adversarial learning algorithms. Also, the regret bound of our algorithm cam be reduced to the regret bound of embedded adversarial learning algorithms. Based on this framework, we obtain new results under various settings. When the model parameters are drawn from unknown non-stationary distributions and we are given machine-learned predictions of the distributions, we develop a new algorithm from our framework with a regret $O(W_T + \sqrt{T})$, where $W_T$ measures the total inaccuracy of the machine-learned predictions. We then develop another algorithm that works when no machine-learned predictions are given and show the performances.

## 1. Introduction

Stochastic optimization is widely used to model the decision making problem with uncertain model parameters. In general, stochastic optimization aims to solve the problem with formulation $\min_{\boldsymbol{c}\in\mathcal{C}} \mathbb{E}_\theta[F_\theta(\boldsymbol{c})]$, where $\theta$ models parameter uncertainty and we optimize the objective on average. An important class of stochastic optimization models is the *two-stage model*, where the problem is further divided into two stages. In particular, at the first stage, we decide the first-stage decision $\boldsymbol{c}$ without knowing the exact value of $\theta$. At the second stage, after a realization of the uncertain data becomes known, an optimal second stage decision $\boldsymbol{x}$ is made by solving an optimization problem parameterized by both $\boldsymbol{c}$ and $\theta$, where one of the constraint can be formulated as $x \in \mathcal{B}(c, \theta)$. Here, $F_\theta(\boldsymbol{c})$ denotes the optimal objective value of the second-stage optimization problem. Two-stage stochastic optimization has numerous applications, including transportation, logistics, financial instruments, and supply chain, among others (Birge and Louveaux 2011).

In this paper, we focus on an "online" extension of the classical two-stage stochastic optimization over a finite horizon of $T$ periods. Subsequently, at each period $t$, we first decide the first-stage decision $\boldsymbol{c}_t \in \mathcal{C}$, then observe the value of model parameter $\theta_t$, which is assumed to be drawn from an *unknown* distribution $P_t$, and finally decide the second-stage decision $\boldsymbol{x}_t$. In addition to requiring $\boldsymbol{x}_t$ belonging to a constraint set parameterized by $\boldsymbol{c}_t$ and $\theta_t$, we also need to satisfy a long-term global constraint of the following form: $\frac{1}{T} \cdot \sum_{t=1}^{T} \boldsymbol{g}_{\theta_t}(\boldsymbol{c}_t, \boldsymbol{x}_t) \in \mathcal{B}$. We aim to optimize the total objective value over the entire horizon, and we measure the performance of our online policy by "regret", the additive loss of the online policy compared to the *optimal dynamic policy* which is aware of $P_t$ for each period $t$.

Indeed, one motivating example for our model is from supply chain management and our model has many other applications. Usually, the supply chain system of a retail company is composed of two layers. One is the centralized *warehouses* layer, and the other is a local *retail stores* layer. At each period, the company needs to decide how much inventory to be invested into the warehouses, which is the first-stage deicion $\boldsymbol{c}$, and then, after the customer demand realizes, the company needs to decide how to transfer the inventory from each warehouse to the downstream retail stores to fulfill customer demand, which is the second-stage decision $\boldsymbol{x}$. It is common in practice that over the entire horizon, the total inventory received at each retail store cannot surpass certain thresholds due to some capacity constraints, or the service level of the customers for each retail store (defined as the total number of times customer demand is fully fulfilled or the fraction of fulfilled customer demand over total demand, for each retail store) cannot be smaller than certain thresholds. These operational requirements induce long-term constraints into our model.

### 1.1. Main Results

Our main result is an online algorithm to achieve a sublinear regret for the two-stage model with non-stationary distributions, i.e. $P_t$ is non-homogeneous over $t \in [T]$. The distribution $P_t$ is assumed to be unknown for each $t \in [T]$, however, in practice, historical data are usually available for the distribution $P_t$, from which we can use some machine learning methods to form a prediction for $P_t$. Therefore, we consider a prediction setting where we have a machine-learned prediction $\hat{P}_t$ of $P_t$, for each $t \in [T]$, at the beginning of the horizon. Note that the prediction $\hat{P}_t$ can be different from the true distribution $P_t$ due to certain prediction error. We aim to develop online algorithm to incorporate these predictions to guide our decision making under the non-stationary environment, while guaranteeing that the performance of our algorithm is robust to the prediction errors. To this end, we first introduce parameter $W$ to capture the total inaccuracy of the pre-given predictions. We show that no online algorithm can achieve a regret bound better than $\Omega(W)$ compared to the offline optimum. We then develop an algorithm to match this lower bound. we introduce a dual variable for each long-term constraint and we update the dual variable at each period to control how the budget of each long-term constraint is consumed. Given the fixed dual variable, we can formulate a two-stage stochastic optimization problem from the Lagrangian dual function, and the first and second stage decision at each period can be obtained from solving this two-stage stochastic optimization problem. An adversarial learning algorithm is finally utilized to update the dual variable, based on the feedback from the solution of this two-stage stochastic optimization problem. The two-stage problem is formulated differently for different periods to handle the non-stationary of the underlying distributions. The formulation relies on the predictions, and as a result, our algorithm, which we name *Informative Adversarial Learning* (IAL) algorithm, naturally combines the information provided by the predictions into the adversarial learning algorithm of the dual variables for the long-term constraints. Our IAL algorithm achieves a regret bound $O(W+\sqrt{T})$, which matches the lower bound $\Omega(W)$.

We then extend to the setting where the predictions are absent and one has to learn the current distribution purely from the past observations. Note that if we allow the underlying distributions to be arbitrarily non-stationary (this will become the adversarial setting), no sublinear regret can be obtained. We consider an intermediate setting between stationary setting and arbitrary non-stationary setting. To be specific, we assume the underlying distributions to be stationary but we allow adversarial corruptions to the realizations. There can be in total $W$ number of corruptions. We modify the IAL algorithm in that we uses another adversarial learning algorithm such as online gradient descent to update the first-stage decision, thus handling the influence of the adversarial corruptions. Therefore, we adopt two adversarial learning algorithms to update the first-stage decision and the dual variable simultaneously, while the second-stage decision is determined afterwards. In this sense, we name our algorithm *Doubly Adversarial Learning* (DAL) algorithm and we show that it achieves a regret bound of $O(W+\sqrt{T})$, which also matches the lower bound $\Omega(W)$ in terms of the dependency on the number of corruptions.

### 1.2. Related Work

**Algorithm with predictions**. There has been a recent trend on studying the performances of online algorithms with pre-existing machine-learned predictions. In this trend, the competitive ratio bound is a popular metric to measure the performance of the algorithm, which is defined as ALG/OPT, where ALG stands for the reward of the algorithm and OPT stands for the optimality benchmark. Competitive ratio bound has been investigated in a series of papers (e.g. Mahdian et al. (2012), Antoniadis et al. (2020), Balseiro et al. (2022a), Banerjee et al. (2022), Jin and Ma (2022), Golrezaei et al. (2023)). Though the competitive ratio is a robust measure and can guarantee the performance of the algorithm even when the predictions are totally inaccurate or absent, see for example Immorlica et al. (2019), Castiglioni et al. (2022), Balseiro et al. (2023), it transfers to a regret bound that is linear in $T$. The additive bound, i.e. the regret bound, for algorithm with predictions have also been studied recently, for example, in the papers Munoz and Vassilvitskii (2017), An et al. (2023), Hao et al.. However, the previous papers focus on the bandit setting without long-term constraints, while our paper considers a general two-stage model with long-term constraints and sublinear regret bounds are derived that incorporates the prediction errors. Note that the non-stationary bandits with knapsack problem has been considered in Lyu and Cheung (2023) with deterministic predictions (rather than a distributional prediction in our setting) and algorithmic design. Our model covers the bandit with knapsack model when the second-stage decision is de-activated. The algorithm with prediction setting has also been considered in other problems such as caching Lykouris and Vassilvitskii (2021), Rohatgi (2020), online scheduling Lattanzi et al. (2020), and the secretary problem Dütting et al. (2021).

**Non-stationary online optimization.** Though the classical online convex optimization allows adversarial input, it considers static benchmark by restricting the benchmark to take a common decision at each period. In order to obtain a sublinear *dynamic regret*, we have to bound the extent of non-stationarity even for the most fundamental multi-arm-bandit problem where the long-term constraint is absent. Various papers have considered the non-stationary online optimization problem with dynamic benchmark, under different measures of non-stationarity (e.g. Zinkevich (2003), Besbes et al. (2014, 2015), Cheung et al. (2020), Celli et al. (2022)). The papers that most close to ours are Balseiro et al. (2023) and Jiang et al. (2020), where similar measures of non-stationarity are considered. However, we consider an online two-stage problem, which is more general and the algorithms are different.

**Connections with other problems.** Our online model is a natural synthesis of several widely studied models in the existing literature. Roughly speaking, we classify previous models on online learning/optimization into the following two categories: the *bandits-based* model and the *type-based* model. For the bandits-based model, we make the decision and then observe the (possibly stochastic) outcome, which can be adversarially chosen. The representative problems include multi-arm-bandits (MAB) problem, and the more general online convex optimization (OCO) problem. Captured in our model, the second-stage decision is de-activated. For the type-based model, at each period, we first observe the type of the arrival, and we are clear of the possible outcome for each action (which can be type-dependent), and then we decide the action without knowing the type of future periods. Note that in the type-based model, we usually have a global constraint such that the cumulative decision over the entire horizon belongs to a set (otherwise the problem becomes trivial, just select the myopic optimal action at each period), which corresponds to the long-term constraint in our model. The representative problems for the type-based model include online allocation problem and a special case online packing problem where the objective function is linear and $\mathcal{B}(\boldsymbol{C}, \boldsymbol{\theta})$ is a polyhedron. Captured in our model, the first-stage decision is de-activated. We review the literature on these problems in Appendix A.

## 2. Problem Formulation

We consider the online two-stage stochastic optimization problem with long-term constraints in the following general formulation. There is a finite horizon of $T$ periods and at each period $t$, the following events happen in sequence:

1 we decide the first-stage decision $\boldsymbol{c}_t \in \mathcal{C}$ and incur a cost $p(\boldsymbol{c}_t)$;

2 the type $\theta_t$ is drawn independently from an *unknown* distribution $P_t$, and the second-stage objective function $f_{\theta_t}(\cdot)$ and the constraint function $\boldsymbol{g}_{\theta_t}(\cdot) = (g_{i,\theta_t}(\cdot))_{i=1}^m$ become known.

3 we decide the second-stage decision $\boldsymbol{x}_t \in \mathcal{K}(\theta_t, \boldsymbol{c}_t)$, where $\mathcal{K}(\theta_t, \boldsymbol{c}_t)$ is a feasible set parameterized by both the type $\theta_t$ and the first-stage decision $\boldsymbol{c}_t$, and incur an objective value $f_{\theta_t}(\cdot)$.

At the end of the entire horizon, the long-term constraints $\frac{1}{T} \cdot \sum_{t=1}^T \boldsymbol{x}_t \in \mathcal{B}(\boldsymbol{C}, \boldsymbol{\theta})$ need to be satisfied, which is characterized as follows, following Mahdavi et al. (2012),

$$\frac{1}{T} \cdot \sum_{t=1}^{T} \boldsymbol{g}_{\theta_t}(\boldsymbol{c}_t, \boldsymbol{x}_t) \leq \boldsymbol{\beta}, \tag{1}$$

with $\boldsymbol{\beta} \in (0,1)^m$. We aim to minimize the objective

$$\sum_{t=1}^{T} \left( p(\boldsymbol{c}_t) + f_{\theta_t}(\boldsymbol{x}_t) \right). \tag{2}$$

Any online policy is feasible as long as $\boldsymbol{c}_t$ and $\boldsymbol{x}_t$ are *agnostic* to the future realizations while satisfying (1). The benchmark is the *dynamic optimal policy*, denoted by $\pi^*$, who is aware of the distributions $\boldsymbol{P} = (P_t)_{t=1}^T$ but still the decisions $\boldsymbol{c}_t$ and $\boldsymbol{x}_t$ have to be agnostic to future realizations. Note that this benchmark is more power than the optimal online policy we are seeking for who is unaware of the distributions $\boldsymbol{P}$, and the optimal policy can be dynamic. We are interested in developing a feasible online policy $\pi$ with known *regret* upper bound compared to the optimal policy:

$$\mathsf{Regret}(\pi, T) = \mathbb{E}_{\boldsymbol{\theta} \sim \boldsymbol{P}} \left[ \mathsf{ALG}(\pi, \boldsymbol{\theta}) - \mathsf{ALG}(\pi^*, \boldsymbol{\theta}) \right], \tag{3}$$

where $\mathsf{ALG}(\pi, \boldsymbol{\theta})$ denotes the objective value of policy $\pi$ on the sequence $\boldsymbol{\theta}$. The optimal policy can possess very complicated structures thus lacks tractability. Therefore, we develop a tractable lower bound of $\mathbb{E}_{\boldsymbol{\theta} \sim \boldsymbol{P}}[\mathsf{ALG}(\pi^*, \boldsymbol{\theta})]$. The lower bound is given by the optimization problem below.

$$\begin{aligned} \mathsf{OPT} = \min \quad & \sum_{t=1}^{T} \mathbb{E}_{\tilde{\boldsymbol{c}}_t, \tilde{\boldsymbol{x}}_t, \theta_t} \left[ p(\tilde{\boldsymbol{c}}_t) + f_{\theta_t}(\tilde{\boldsymbol{x}}_t) \right], \\ \text{s.t.} \quad & \frac{1}{T} \cdot \sum_{t=1}^{T} \mathbb{E}_{\tilde{\boldsymbol{c}}_t, \tilde{\boldsymbol{x}}_t, \theta_t} [\boldsymbol{g}_{\theta_t}(\tilde{\boldsymbol{c}}_t, \tilde{\boldsymbol{x}}_t)] \leq \boldsymbol{\beta}, \\ & \tilde{\boldsymbol{x}}_t \in \mathcal{K}(\theta_t, \tilde{\boldsymbol{c}}_t), \tilde{\boldsymbol{c}}_t \in \mathcal{C}, \forall t. \end{aligned} \tag{OPT}$$

Here, $\tilde{\boldsymbol{c}}_t$ and $\tilde{\boldsymbol{x}}_t$ are random variables for each $t \in [T]$ and the distribution of $\tilde{\boldsymbol{c}}_t$ is *independent* of $\theta_t$. Note that we only need the formulation of (OPT) to conduct our theoretical analysis. We never require to really solve the optimization problem (OPT). Clearly, if we let $\tilde{\boldsymbol{c}}_t$ (resp. $\tilde{\boldsymbol{x}}_t$) denote the *marginal distribution* of the first-stage (resp. second-stage) decision made by the optimal policy, the we would have a feasible solution to Equation (OPT) while the objective value equals $\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[\mathsf{ALG}(\pi^*,\boldsymbol{\theta})]$. This argument leads to the following lemma.

Lemma 1 **(forklore)**. $\mathsf{OPT} \le \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[\mathsf{ALG}(\pi^*,\boldsymbol{\theta})]$.

Throughout the paper, we make the following assumptions.

Assumption 1. *The following conditions are satisfied:*

a. *(convexity) The function $p(\cdot)$ is a convex function with $p(\mathbf{0})=0$ and $\mathcal{C}$ is a compact and convex set which contains $\mathbf{0}$. The set $\mathcal{A}\subset[0,1]^m$ is a convex set. Also, the functions $f_\theta(\cdot)$ is convex in $\boldsymbol{x}$ and $\boldsymbol{g}_\theta(\cdot,\cdot)$ is convex in both $\boldsymbol{c}$ and $\boldsymbol{x}$, for any $\theta$.*

b. *(compactness) The set $\mathcal{K}(\theta_t,\boldsymbol{c}_t)$ is a polyhedron given by $\mathcal{K}(\theta_t,\boldsymbol{c}_t)=\{\boldsymbol{x}\in\mathbb{R}^m:\mathbf{0}\le\boldsymbol{x}$ and $B_{\theta_t}\boldsymbol{x}\le\boldsymbol{c}_t\}\cap\mathcal{K}$ where $\mathcal{K}$ is a compact convex set.*

c. *(boundedness) For any $\boldsymbol{x}\in\mathcal{K}$ and any $\theta$, we have $f_\theta(\boldsymbol{x})\in[-1,1]$ and $\boldsymbol{g}_\theta(\boldsymbol{x})\in[0,1]^m$. Moreover, it holds that $f_\theta(\mathbf{0})=0$ and $\boldsymbol{g}_\theta(\mathbf{0},\mathbf{0})=\mathbf{0}$.*

The conditions listed in Assumption 1 are standard in the literature, which mainly ensure the convexity of the problem as well as the boundedness of the objective functions and the feasible set. Finally, in condition c, we assume that there always exists $\mathbf{0}$, which denotes a *null* action that has no influence on the accumulated objective value and the constraints.

The Lagrangian dual problem of (OPT) is given below:

$$\begin{aligned}\max_{\boldsymbol{\lambda}\ge 0}\min_{\tilde{\boldsymbol{c}}_t\in\mathcal{C}}\quad \mathsf{L}(\boldsymbol{C},\boldsymbol{\lambda}) = \sum_{t=1}^{T}\mathbb{E}\Bigg[ & p(\tilde{\boldsymbol{c}}_t)-\frac{1}{T}\cdot\sum_{i=1}^{m}\lambda_i+ \\ & \min_{\tilde{\boldsymbol{x}}_t\in\mathcal{K}(\theta_t,\tilde{\boldsymbol{c}}_t)}\left\{f_{\theta_t}(\tilde{\boldsymbol{x}}_t)+\sum_{i=1}^{m}\frac{\lambda_i\cdot g_{i,\theta_t}(\tilde{\boldsymbol{c}}_t,\tilde{\boldsymbol{x}}_t)}{T\cdot\beta_i}\right\}\Bigg].\end{aligned} \tag{Dual}$$

where $\boldsymbol{C}=(\tilde{\boldsymbol{c}}_t)_{t=1}^T$ and we note that the distribution of $\tilde{\boldsymbol{c}}_t$ is independent of $\theta_t$. From weak duality, (Dual) also serves as a lower bound of $\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[\mathsf{ALG}(\pi^*,\boldsymbol{\theta})]$. Therefore, in what follows, we compare against (Dual) to derive our regret bound.

## 3. Non-stationary Setting with Predictions

In this section, we consider the non-stationary setting where $P_t$ is *unknown* and *non-homogeneous* for each $t$. Since we are comparing against the dynamic optimal policy, our definition of regret in (3) falls into the range of *dynamic regret*. It has been known in the literature (e.g. Besbes et al. (2014)) that without additional restrictions on the true distributions $\boldsymbol{P}=(P_t)_{t=1}^T$, one cannot achieve a dynamic regret that is sublinear in $T$. Therefore, we seek for sublinear regret with the help of *additional information*. In practice, the horizon can be repeated for multiple times, from which we obtain some historical dataset over the true distributions $\boldsymbol{P} = (P_t)_{t=1}^T$. It is common to utilize some machine learning methods to learn the true distributions $\boldsymbol{P} = (P_t)_{t=1}^T$. As a result, in this section, we assume that there exists a prediction $\hat{P}_t$ of the true distribution $P_t$, for each $t \in [T]$. The predictions $\hat{\boldsymbol{P}} = (\hat{P}_t)_{t=1}^T$ are given at the beginning of the horizon and can be different from the true distributions $\boldsymbol{P} = (P_t)_{t=1}^T$ due to some prediction errors. We explore how to utilize these predictions $\hat{\boldsymbol{P}} = (\hat{P}_t)_{t=1}^T$ to make our online decisions.

We first derive the regret lower bound and show how the regret should depend on the possible inaccuracy of the predictions. Following Jiang et al. (2020), we measure the inaccuracy of $\hat{P}_t$ by Wasserstein distance (we refer interested readers to Section 6.3 of Jiang et al. (2020) for benefits of using Wasserstein distance in online decision making), which is defined as follows

$$W(\hat{P}_t, P_t) := \inf_{Q \in \mathcal{F}(\hat{P}_t, P_t)} \int d(\theta, \theta') dQ(\theta, \theta'), \tag{4}$$

where $d(\theta, \theta') = \|(f_\theta, \boldsymbol{g}_\theta) - (f_{\theta'}, \boldsymbol{g}_{\theta'})\|_\infty$ and $\mathcal{F}(\hat{P}_t, P_t)$ denotes the set of all joint distributions for $(\theta, \theta')$ with marginal distributions $\hat{P}_t$ and $P_t$. In the following theorem, we show that one cannot break a linear dependency on the total inaccuracy of the predictions.

THEOREM 1. *Let $W_T = \sum_{t=1}^T W(\hat{P}_t, P_t)$ be the total measure of inaccuracy, where $W(\hat{P}_t, P_t)$ is defined in (4). For any feasible online policy $\pi$ that only knows predictions, there always exists $\boldsymbol{P}$ and $\hat{\boldsymbol{P}}$ such that*

$$\mathsf{Regret}(\pi, T) = \mathbb{E}_{\boldsymbol{\theta} \sim \boldsymbol{P}}[\mathsf{ALG}(\pi, \boldsymbol{\theta}) - \mathsf{ALG}(\pi^*, \boldsymbol{\theta})] \geq \Omega\left(W_T\right),$$

*where $\pi^*$ denotes the optimal policy that knows true distributions $\boldsymbol{P}$.*

We then derive our online algorithm with a regret that matches the lower bound established above. When we have a predictions available, an efficient way would be to "greedily" select $\boldsymbol{c}_t$ with the help of predictions. In order to describe our idea, we denote by $\{\hat{\boldsymbol{\lambda}}^*, \{\hat{\boldsymbol{c}}_t^*\}_{\forall t \in [T]}\}$ one optimal solution to

$$\begin{aligned} \max_{\boldsymbol{\lambda} \geq 0} \min_{\boldsymbol{c}_t \in \mathcal{C}} \hat{L}(\boldsymbol{C}, \boldsymbol{\lambda}) = \mathbb{E}_{\boldsymbol{\theta} \sim \hat{\boldsymbol{P}}} \Bigg[ \sum_{t=1}^T \Bigg( p(\boldsymbol{c}_t) - \frac{1}{T} \cdot \sum_{i=1}^m \lambda_i \\ + \min_{\boldsymbol{x}_t \in \mathcal{K}(\theta_t, \boldsymbol{c}_t)} f_{\theta_t}(\boldsymbol{x}_t) + \sum_{i=1}^m \frac{\lambda_i \cdot g_{i,\theta_t}(\boldsymbol{c}_t, \boldsymbol{x}_t)}{T \cdot \beta_i} \Bigg) \Bigg], \end{aligned} \tag{5}$$

where $\hat{\boldsymbol{P}} = (\hat{P}_t)_{t=1}^T$. Note that the value of (5) is equivalent to the value of (Dual) if $\hat{\boldsymbol{P}} = \boldsymbol{P}$. Then, we define for each $i \in [m]$,

$$\hat{\beta}_{i,t} := \mathbb{E}_{\theta \sim \hat{P}_t}\left[g_{i,\theta}(\hat{\boldsymbol{c}}_t^*, \hat{\boldsymbol{x}}_t^*(\theta))\right], \tag{6}$$

with

$$\hat{\boldsymbol{x}}_t^*(\theta) \in \operatorname{argmin}_{\boldsymbol{x}_t \in \mathcal{K}(\theta, \hat{\boldsymbol{c}}_t^*)} f_\theta(\boldsymbol{x}_t) + \sum_{i=1}^m \frac{\hat{\lambda}_i^* \cdot g_{i,\theta}(\hat{\boldsymbol{c}}_t^*, \boldsymbol{x}_t)}{T \cdot \beta_i}.$$

Here, $\hat{\boldsymbol{\beta}}_t$ can be interpreted as the *predictions-informed* target levels to achieve, at each period $t$. To be more concrete, if $\hat{P}_t = P_t$ for each $t$, then $\hat{\boldsymbol{\beta}}$ is exactly the value of the constraint functions at period $t$ in (OPT), which is a good reference to stick to. We then define

$$\begin{aligned} \hat{L}_t(\boldsymbol{c}, \boldsymbol{\lambda}) := \mathbb{E}_{\theta \sim \hat{P}_t} \Bigg[ & p(\boldsymbol{c}) - \sum_{i=1}^m \frac{\lambda_i \cdot \hat{\beta}_{i,t}}{T \cdot \beta_i} \\ & + \min_{\boldsymbol{x} \in \mathcal{K}(\theta, \boldsymbol{c})} f_\theta(\boldsymbol{x}) + \sum_{i=1}^m \frac{\lambda_i \cdot g_{i,\theta}(\boldsymbol{c}, \boldsymbol{x})}{T \cdot \beta_i} \Bigg]. \end{aligned} \tag{7}$$

The key ingredient of our analysis is the following.

Lemma 2. *It holds that*

$$\max_{\boldsymbol{\lambda} \geq 0} \min_{\boldsymbol{c}_t \in \mathcal{C}} \hat{L}(\boldsymbol{C}, \boldsymbol{\lambda}) = \sum_{t=1}^T \max_{\boldsymbol{\lambda} \geq 0} \min_{\boldsymbol{c}_t \in \mathcal{C}} \hat{L}_t(\boldsymbol{c}_t, \boldsymbol{\lambda}), \tag{8}$$

*and moreover, letting* $(\hat{\boldsymbol{\lambda}}^*, (\hat{\boldsymbol{c}}_t^*)_{t=1}^T)$ *be the optimal solution to* $\max_{\boldsymbol{\lambda} \geq 0} \min_{\boldsymbol{c}_t \in \mathcal{C}} \hat{L}(\boldsymbol{C}, \boldsymbol{\lambda})$ *used in the definition* (6)*, then* $(\hat{\boldsymbol{\lambda}}^*, \hat{\boldsymbol{c}}_t^*)$ *is an optimal solution to* $\max_{\boldsymbol{\lambda} \geq 0} \min_{\boldsymbol{c}_t \in \mathcal{C}} \hat{L}_t(\boldsymbol{c}_t, \boldsymbol{\lambda})$ *for each* $t \in [T]$*. Also, we have*

$$\sum_{i=1}^m \sum_{t=1}^T \hat{\lambda}_{i,t}^* \hat{\beta}_i = T \cdot \sum_{i=1}^m \beta_i \cdot \hat{\lambda}_i^*, \tag{9}$$

*for each* $i \in [m]$*.*

Lemma 2 implies that given $\hat{\boldsymbol{\lambda}}^*$, we can simply minimize $\hat{L}_t(\boldsymbol{c}_t, \hat{\boldsymbol{\lambda}}^*)$ over $\boldsymbol{c}_t$ to get the first-stage decision [1]. We now describe the procedure to obtain the dual variable $\boldsymbol{\lambda}_t$ for each period $t$.

We adopt the framework of expert problem, where we regard each constraint $i$ as an expert $i$ and we use the algorithm for the expert problem to update the dual variable. We regard $\boldsymbol{\lambda}_t$ as a distribution over the long-term constraints, after divided by a scaling factor $\mu$. Then, the range of the dual variable for all the minimax problems (7) can be restricted to the set $\mu \cdot \Delta_m$ where $\Delta_m = \{\boldsymbol{y} \in \mathbb{R}_{\geq 0}^m : \sum_{i=1}^m y_i = 1\}$ denotes a distribution over the long-term constraints and $\mu > 0$ is a constant to be specified later. The player of the expert problem actually chooses one constraint $i_t$ among the long-term constraints, by setting $\boldsymbol{\lambda}_t = \mu \cdot \boldsymbol{e}_{i_t}$ where $\boldsymbol{e}_{i_t} \in \mathbb{R}^m$ is a vector with 1 as the $i_t$-th component and 0 for all other components. Clearly, what we can observe is a *stochastic* outcome. We can observe the stochastic outcome $\hat{L}(\boldsymbol{c}_t, \mu \cdot \boldsymbol{e}_{i_t}, \theta_t)$ for the currently chosen constraint $i_t$. Finally, the second-stage decision $\boldsymbol{x}_t$ is determined by solving the inner minimization problem

[1] This is a stochastic optimization problem, which can be solved by applying *stochastic gradient descent* over $\boldsymbol{c}_t$ or applying *sample average approximation* to get samples of $\theta \sim \hat{P}_t$

**Algorithm 1** Informative Adversarial Learning (IAL) algorithm

```
**Input:** the scaling factor $\mu > 0$, the adversarial learning algorithm $\mathsf{ALG}_{\mathsf{Dual}}$ for dual variable $\boldsymbol{\lambda}$, and the prior estimates $\hat{\boldsymbol{P}}$.
**Initialize:** compute $\hat{\beta}_{i,t}$ for all $i \in [m], t \in [T]$ as (6).
**for** $t = 1, \dots, T$ **do**
    1. $\mathsf{ALG}_{\mathsf{Dual}}$ returns a long-term constraint $i_t \in [m]$.
    2. Set $\boldsymbol{c}_t = \mathrm{argmin}_{\boldsymbol{c}\in\mathcal{C}} \hat{L}_t(\boldsymbol{c}, \mu \cdot \boldsymbol{e}_{i_t})$.
    3. Observe $\theta_t$ and set $\boldsymbol{x}_t$ by solving the inner problem in (7) with $\boldsymbol{c} = \boldsymbol{c}_t$ and $\boldsymbol{\lambda} = \mu \cdot \boldsymbol{e}_{i_t}$.
    4. Return to $\mathsf{ALG}_{\mathsf{Dual}}$ $\hat{L}_{i,t}(\boldsymbol{c}_t, \mu \cdot \boldsymbol{e}_{i_t}, \theta_t)$ defined as follows for each $i \in [m]$,
       $\hat{L}_{i,t}(\boldsymbol{c}_t, \mu \cdot \boldsymbol{e}_{i_t}, \theta_t) = p(\boldsymbol{c}_t) - \frac{\mu \cdot \hat{\beta}_{i,t}}{T \cdot \beta_i} + f_{\theta_t}(\boldsymbol{x}_t) + \frac{\mu \cdot g_{i,\theta_t}(\boldsymbol{c}_t, \boldsymbol{x}_t)}{T \cdot \beta_i}.$   (11)
**end for**
```

in (7) for $\boldsymbol{c} = \boldsymbol{c}_t$ and $\boldsymbol{\lambda} = \mu \cdot \boldsymbol{e}_{i_t}$. Here, $\boldsymbol{c}_t$ is determined as the minimizer of $\hat{L}_t(\boldsymbol{c}_t, \hat{\boldsymbol{\lambda}}^*)$ and $i_t$ is determined by the algorithm for the expert problem over the long-term constraints.

We now further specify what is the observed outcome for the dual player. Note that the dual player chooses a constraint $i_t$ as action. The corresponding outcome is $\hat{L}_{i_t}(\boldsymbol{c}_t, \mu \cdot \boldsymbol{e}_{i_t}, \theta_t)$, where $\hat{L}_i(\boldsymbol{c}_t, \mu \cdot \boldsymbol{e}_{i_t}, \theta_t)$ is defined as follows for each $i \in [m]$,

$$\hat{L}_i(\boldsymbol{c}_t, \mu \cdot \boldsymbol{e}_{i_t}, \theta_t) = p(\boldsymbol{c}_t) - \frac{\mu}{T} + f_{\theta_t}(\boldsymbol{x}_t) + \frac{\mu \cdot g_{i,\theta_t}(\boldsymbol{c}_t, \boldsymbol{x}_t)}{T \cdot \beta_i}, \tag{10}$$

where $\boldsymbol{x}_t$ denotes the second-stage decision we made at period $t$. Though the action for the dual player is $\boldsymbol{e}_{i_t}$, we are actually able to obtain *additional information* for the dual player, which helps the convergence of the expert algorithm. It is easy to see that we have the value of $\bar{L}_i(\boldsymbol{c}_t, \mu \cdot \boldsymbol{e}_{i_t}, \theta_t)$ at the end of period $t$, for all other constraint $i \neq i_t$. Therefore, for the expert algorithm, we are having the *full feedback*, where the outcomes of all constraints, $\hat{L}_i$ for all $i \in [m]$, can be observed. The above discussion implies that we can apply adversarially learning algorithm such as Hedge algorithm to dynamically update the dual variable $\boldsymbol{\lambda}_t$ for each period $t$. Our formal algorithm is described in Algorithm 1, which we call *Informative Adversarial Learning* (IAL) algorithm since the updates are informed by the predictions.

As shown in the next theorem, the regret of Algorithm 1 is bounded by $O(W_T + \sqrt{T})$, which matches the lower bound established in Theorem 1.

Theorem 2. *Denote by $\pi$ Algorithm 1 with input $\mu = \|\hat{\boldsymbol{\lambda}}^*\|_\infty = \alpha \cdot T$ for some constant $\alpha > 0$. Denote by $W_T = \sum_{t=1}^{T} W(\hat{P}_t, P_t)$ the total measure of inaccuracy, with $W(\hat{P}_t, P_t)$ defined in (4). Then, under Assumption 1, the regret enjoys the upper bound*

$$\begin{aligned} \mathit{Regret}(\pi, T) =& \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[\mathsf{ALG}(\pi, \boldsymbol{\theta})] - \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[\mathsf{ALG}(\pi^*, \boldsymbol{\theta})] \\ \le& \tilde{O}(\sqrt{T \cdot \log m}) + O(W_T) \end{aligned} \tag{12}$$

*if Hedge is selected as* $\mathsf{ALG}_{\mathsf{Dual}}$. *Moreover, we have*

$$\frac{1}{T}\sum_{t=1}^{T} g_{i,\theta_t}(\boldsymbol{c}_t, \boldsymbol{x}_t) - \beta_i \le \tilde{O}\left(\sqrt{\frac{\log m}{T}}\right) + O\left(\frac{W_T}{T}\right),$$

*for each $i \in [m]$.*

Notably, the inaccuracy of the prior estimates would result in a constraint violation that scales as $O(\frac{1}{\sqrt{T}} + \frac{W_T}{T})$ as shown in Theorem 2. Therefore, as long as $W_T$ scales sublinearly in $T$, which guarantees a sublinear regret bound following (12), the constraint violation of our Algorithm 1 also scales as $o(1)$, implying that the solutions generated by Algorithm 1 is asymptotically feasible. Generating asymptotically feasible solutions is standard in OCOwC literature (Jenatton et al. 2016, Neely and Yu 2017, Yuan and Lamperski 2018, Yi et al. 2021).

## 4. Extension to the Setting without Predictions

In this section, we explore the setting where there are no machine-learned predictions and the true distributions are still unknown. If the true distributions can be arbitrarily non-stationary, then the setting becomes identical to the adversarial setting and in general, one cannot achieve a sublinear regret. Therefore, we consider a setting lying between the stationary and the non-stationary setting. To be specific, we assume that the true distributions are identical, i.e., $P_t = P$ for each $t \in [T]$. However, at each period $t$, after the type $\theta_t$ is realized, there can be an adversary corrupting $\theta_t$ into $\theta_t^c$, and only the value of $\theta_t^c$ is revealed to us. The adversarial corruption to a stochastic model can arise from the non-stationarity of the underlying distributions $\boldsymbol{P}$ (e.g. Jiang et al. (2020), Balseiro et al. (2023)), or malicious attack and false information input to the system (Lykouris et al. (2018), Gupta et al. (2019)).

We first characterize the difficulty of the problem. Denote by $W(\boldsymbol{\theta})$ the total number of adversarial corruptions on sequence $\boldsymbol{\theta}$, i.e.,

$$W(\boldsymbol{\theta}) = \sum_{t=1}^{T} \mathbb{1}(\theta_t \neq \theta_t^c). \tag{13}$$

We show that the optimal regret bound scales at least $\Omega(\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[W(\boldsymbol{\theta})])$. Note that in the definition of the regret, the performance of our algorithm is the total collected reward *after* adversarial corrupted, and the benchmark is the optimal policy *with* adversarial corruptions, denoted by $\pi^*$.

Theorem 3. *Let $W(\boldsymbol{\theta})$ be the total number of adversarial corruptions on sequence $\boldsymbol{\theta}$, as defined in (13). For any feasible online policy $\pi$, there always there exists distributions $\boldsymbol{P}$ and a way to corrupt $\boldsymbol{P}$ such that*

$$\begin{aligned}\mathsf{Regret}^c(\pi,T) = &\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[\mathsf{ALG}(\pi,\boldsymbol{\theta}^c)] - \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[\mathsf{ALG}(\pi^*,\boldsymbol{\theta}^c)]\\ &\geq \Omega\left(\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[W(\boldsymbol{\theta})]\right).\end{aligned}$$

Our lower bound in Theorem 3 is in correspondence to the lower bound established in Lykouris et al. (2018) for stochastic multi-arm-bandits model.

We now derive our algorithm to achieve a regret upper bound that matches the lower bound in Theorem 3 in terms of the dependency on $W(\boldsymbol{\theta})$. Our algorithm is modified from the previous Algorithm 1. However, the caveat is that since we do not have the predictions $\hat{P}_t$ for each $t\in[T]$, we cannot directly obtain the value of $\boldsymbol{c}_t$ by minimizing over $\hat{L}_t(\boldsymbol{c},\boldsymbol{\lambda})$ in (7). One can indeed use historical samples to obtain an estimate of $P_t$ and plug the estimate into the formulation of $\hat{L}_t(\boldsymbol{c},\boldsymbol{\lambda})$ to solve it. However, the existence of the adversarial corruptions will bring new challenges to tackle with. Instead, in what follows, we seek for another approach that uses another adversarial learning algorithm to update $\boldsymbol{c}_t$.

We note that the minimax Lagrangian dual problem (Dual) admits the following reformulation, given the type realization $\theta$.

$$\begin{aligned}\bar{L}(\boldsymbol{c},\boldsymbol{\lambda},\theta) = &p(\boldsymbol{c}) - \frac{1}{T}\cdot\sum_{i=1}^{m}\lambda_i + \min_{\boldsymbol{x}\in\mathcal{K}(\theta,\boldsymbol{c})}\Bigg\{f_\theta(\boldsymbol{x})\\ &+\sum_{i=1}^{m}\frac{\lambda_i\cdot g_{i,\theta}(\boldsymbol{c},\boldsymbol{x})}{T\cdot\beta_i}\Bigg\}.\end{aligned} \tag{14}$$

We have the following lemma showing the convexity of $\bar{L}(\boldsymbol{c},\boldsymbol{\lambda},\theta)$ over $\boldsymbol{c}$.

Lemma 3. *For any $\boldsymbol{\lambda}$ and any $\theta$, $\bar{L}(\boldsymbol{c},\boldsymbol{\lambda},\theta)$ defined in (14) is a convex function over $\boldsymbol{c}$, under Assumption 1.*

This above results imply that we can apply methods from online convex optimization (OCO) to update $\boldsymbol{c}_t$ for each $t\in[T]$. The formal algorithm is presented in Algorithm 2. Our algorithm admits a double adversarial learning structure. On the one hand, we use *Online Gradient Descent* (OGD) algorithm (Zinkevich 2003) as $\mathsf{ALG}_1$ to update the first-stage decision $\boldsymbol{c}_t$. On the other hand, similar to Algorithm 1, we regard each long-term constraint as an expert and use the expert algorithm such as *Hedge* algorithm (Freund and Schapire 1997) as $\mathsf{ALG}_2$ to update the dual variable $\boldsymbol{\lambda}_t$, for each $t\in[T]$. Finally, the second-stage decision $\boldsymbol{x}_t$ is determined by solving the inner minimization problem in (14) for $\boldsymbol{c}=\boldsymbol{c}_t$, $\boldsymbol{\lambda}=\boldsymbol{\lambda}_t$ and $\theta=\theta_t^c$.

**Algorithm 2** Doubly Adversarial Learning (DAL) algorithm

```
Input: the scaling factor $\mu > 0$, the adversarial learning algorithm $\mathsf{ALG}_1$ for the first-stage decision, the adversarial learning algorithm $\mathsf{ALG}_2$ for the dual variable.
for $t = 1, \dots, T$ do
    1. $\mathsf{ALG}_1$ returns $\boldsymbol{c}_t$ and $\mathsf{ALG}_2$ returns $i_t \in [m]$.
    2. Observe $\theta_t^c$ and determine $\boldsymbol{x}_t$ by solving the inner problem in (14) with $\boldsymbol{c} = \boldsymbol{c}_t$ and $\boldsymbol{\lambda} = \mu \cdot \boldsymbol{e}_{i_t}$.
    3. Return $\bar{L}(\boldsymbol{c}_t, \mu \cdot \boldsymbol{e}_{i_t}, \theta_t^c)$ to $\mathsf{ALG}_1$.
    4. Return $\hat{L}_i(\boldsymbol{c}_t, \mu \cdot \boldsymbol{e}_{i_t}, \theta_t^c)$ defined in (10) for all $i \in [m]$ to $\mathsf{ALG}_2$.
    if $\frac{1}{T} \cdot \sum_{\tau=1}^{t} g_{i,\theta_\tau}(\boldsymbol{c}_\tau, \boldsymbol{x}_\tau) > \beta_i$ for some $i \in [m]$ then
        we terminate the algorithm by taking the null action $\mathbf{0}$ for both stage decision in the remaining horizon.
    end if
end for
```

We now show that our Algorithm 2 achieves a regret bound that matches the lower bound in Theorem 3 in terms of the dependency on $W(\boldsymbol{\theta})$, which is in correspondence to the linear dependency on $W(\boldsymbol{\theta})$ established in Gupta et al. (2019) for the stochastic multi-arm-bandits model.

THEOREM 4. *Denote by $\pi$ Algorithm 2 with input $\mu = T$. Then, under Assumption 1 and the corrupted setting, the regret enjoys the upper bound*

$$\begin{aligned} \mathit{Regret}^c(\pi, T) &= \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[\mathsf{ALG}(\pi, \boldsymbol{\theta}^c)] - \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[\mathsf{ALG}(\pi^*, \boldsymbol{\theta}^c)] \\ &\le \tilde{O}((G+F)\cdot\sqrt{T}) + \tilde{O}(\sqrt{T \cdot \log m}) \\ &\quad + O(\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[W(\boldsymbol{\theta})]), \end{aligned} \tag{15}$$

*if OGD is selected as $\mathsf{ALG}_1$ and Hedge is selected as $\mathsf{ALG}_2$. Here, $\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[W(\boldsymbol{\theta})]$ denotes the expectation of the total number of corruptions, with $W(\boldsymbol{\theta})$ defined in (13).*

We remark that the implementation of our Algorithm 2 in this setting is *agnostic* to the total number of adversarial corruptions and achieves the optimal dependency on the number of corruptions. This is one fascinating benefits of adopting adversarial learning algorithms as algorithmic subroutines in Algorithm 2, where the corruptions can also be incorporated as adversarial input which is handled by the learning algorithms. On the other hand, even if the number of corruptions is given to us as a prior knowledge, Theorem 3 shows that still, no online policy can achieve a better dependency on the number of corruptions.

## 5. Numerical Experiments

In this section, we conduct numerical experiments to test the performance of our algorithms empirically. We try two sets of experiments for the resource allocation problems. One deals with the resource capacity constraints (packing constraints from a mathematical optimization view), where the long-term constraints take the formulation of $\sum_{t=1}^{T} \boldsymbol{g}(\boldsymbol{x}_t, \boldsymbol{\theta}_t) \leq T \cdot \boldsymbol{\beta}$. The other set of experiments deals with the service level constraints (covering constraints from a mathematical optimization view), where the long-term constraints take the formulation of $\sum_{t=1}^{T} \boldsymbol{g}(\boldsymbol{x}_t, \boldsymbol{\theta}_t) \geq T \cdot \boldsymbol{\beta}$. Note that the service level constraints (or the fairness constraints (Kearns et al. 2018)) are widely studied in the literature (e.g. Hou et al. (2009)), and we develop our algorithms to handle this type of covering constraints. The algorithmic development for the second set is described in Section B.2 in the supplementary material. In our experiments, we set the functions $\boldsymbol{g}(\boldsymbol{x}_t, \boldsymbol{\theta}_t)$ to be linear functions over $\boldsymbol{x}_t$.

**Experiment 1.** The first set deals with packing constraints. The offline problem can be formulated as follows:

$$\begin{aligned} \max \quad & \sum_{t=1}^{T} c_t \\ \text{s.t.} \quad & \sum_{t=1}^{T} x_{i,t} \leq T \cdot \beta_i, \quad \forall i = 1, \ldots, 4 \\ & x_{i,t} \leq D_{i,t,}, \quad \forall i, \forall t \\ & \sum_{i=1}^{4} x_{i,t} \geq \min\{c_t, \sum_{i=1}^{4} D_{i,t}\}, \quad \forall t \\ & x_{i,t} \geq 0, c_t \leq C, \quad \forall i, \forall t. \end{aligned}$$

We do the numerical experiment under the following settings. Consider a resource allocation problem where there is 4 resources to be allocated. At each period, the decision maker needs to first decide a budget $c_t$ which restricts the total amount of units to be invested to each resources, and then, after the demand is realized, the decision maker needs to allocated the budget to satisfy the demand for each resource. The demand for each resource is normally distributed at each period (truncated at 0). We set the mean parameter $\mu_0 = 5$ and the standard deviation parameter $\sigma_0 = 10/3$. We set the vector $\boldsymbol{\beta} = [0.95, 0.90, 0.85, 0.80]$. We assume DAL is blind to the distributions while IAL knows the distributions. We test the performance of the IAL algorithm and DAL algorithm in the following four cases:

- a. Stationary distribution case: demands of the four resources follow the same distribution $\mathcal{N}(k_0\mu_0, \sigma_0)$ with $k_0 = 2$.
- b. Non-stationary distribution case 1: We divide the whole horizon into two intervals: $[1, T/2]$ and $[T/2+1, T]$. During each interval, we sample the demands independently following the same distribution $\mathcal{N}(k_1\mu_0, \sigma_0)$ with $k_1 = 1, 3$ for the two intervals.

- c. Non-stationary distribution case 1: We divide the whole horizon into two intervals: $[1, T/2]$ and $[T/2+1, T]$. During each interval, we sample the demands independently following the same distribution $\mathcal{N}(k_2\mu_0, \sigma_0)$ with $k_2 = 3, 1$ for the two intervals.
- d. Non-stationary distribution case 3: We divide the whole horizon into five intervals: $[1, T/5]$, $[T/5+1, 2T/5]$, $[2T/5+1, 3T/5]$, $[3T/5+1, 4T/5]$, $[4T/5+1, T]$. During each interval, we sample the demands independently following the same distribution $\mathcal{N}(k_3\mu_0, \sigma_0)$ with $k_3 = 1, 2, 3, 2, 1$ for five intervals.

**Results and interpretations.** Note that the extent of non-stationarity is the same for the non-stationary case 1 and case 2. The difference is that the distribution of the two intervals is exchanged in case 1 and case 2. The extent of non-stationarity is largest for the non-stationary case 3. The numerical results are presented in Table 1. On one hand, for the stationary case, the performances of both DAL and IAL are good, with a relative regret within 4%. On the other hand, as we can see, the performance of DAL decays from situation (a) up to situation (d), as the non-stationarity of the underlying distributions gets larger and larger. However, the performance of IAL remains relatively stable because the non-stationarity of the underlying distributions has been well incorporated in the algorithmic design, which illustrates the effectiveness of IAL in a non-stationary environment.

| | stationary case | non-stationary case 1 |
|---|---|---|
| DAL | 3.43% | 8.26% |
| IAL | 3.45% | 5.84% |
| | non-stationary case 2 | non-stationary case 3 |
| DAL | 8.26% | 11.02% |
| IAL | 5.86% | 6.54% |

**Table 1 The relative regret of DAL and IAL algorithms with capacity (packing) constraints**

**Experiment 2.** The second set considers the service level constraints, namely, the covering constraints. We consider the four cases which are exactly the same as the cases described in experiment 1. The other parameters are also set as the same as those in experiment 1, except that the long-term constraints now become $\sum_{t=1}^{T} \boldsymbol{g}(\boldsymbol{x}_t, \boldsymbol{\theta}_t) \geq T \cdot \boldsymbol{\beta}$.

**Results and interpretations.** In each of the figures, corresponding to each case, we plot 4 graphs capturing how the budget $c_t$ is determined and how the service level constraints are satisfied for each DAL and IAL during the horizon of $T = 10000$ period. As we can see, both IAL and DAL converge rather quickly under the stationary case, in that the budget $c_t$ remains stable and the target service levels are achieved after about 1000 periods, as plotted in Figure 1. However, as we add non-stationarity in Figure 2, Figure 3, and Figure 4, the performance of DAL decays in that the budget $c_t$ cannot capture the non-stationarity, and the achieved service levels are not stable. In contrast, for IAL, the budget $c_t$ changes quite quickly as long as the underlying distribution shifts. Moreover, even though the distributions changed during the horizon, the achieved service levels of IAL remain stable and the targets are reached. All these numerical results correspond to our theoretical finding and illustrate the benefits of IAL under non-stationarity.

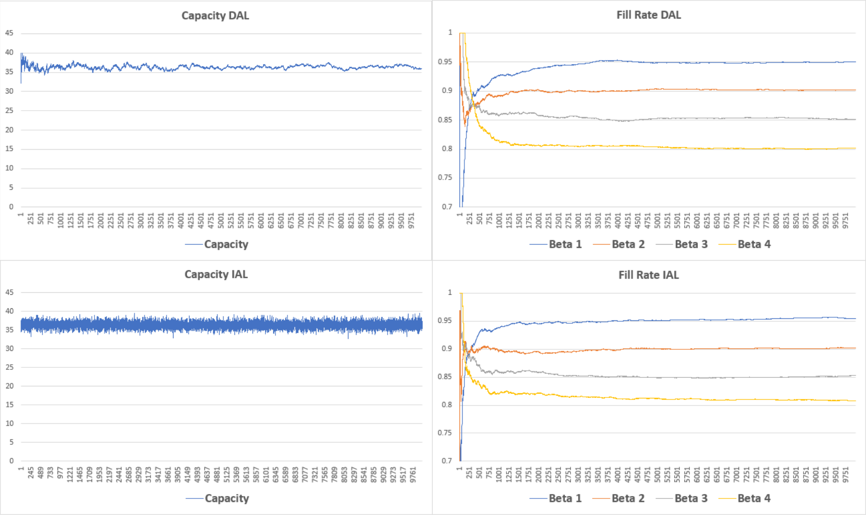

**Figure 1** **Numerical results of DAL and IAL algorithms with service level (covering) constraints for the stationary case.**

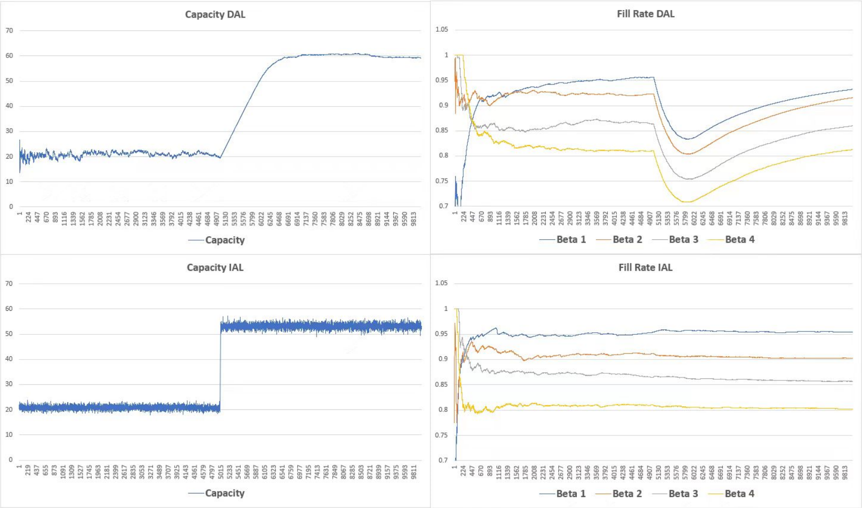

**Figure 2** **Numerical results of DAL and IAL algorithms with service level (covering) constraints for the non-stationary case 1.**

## 6. Summary

This paper proposes and studies the problem of bounding regret for online two-stage stochastic optimization with long-term constraints. The main contribution is an algorithmic framework that

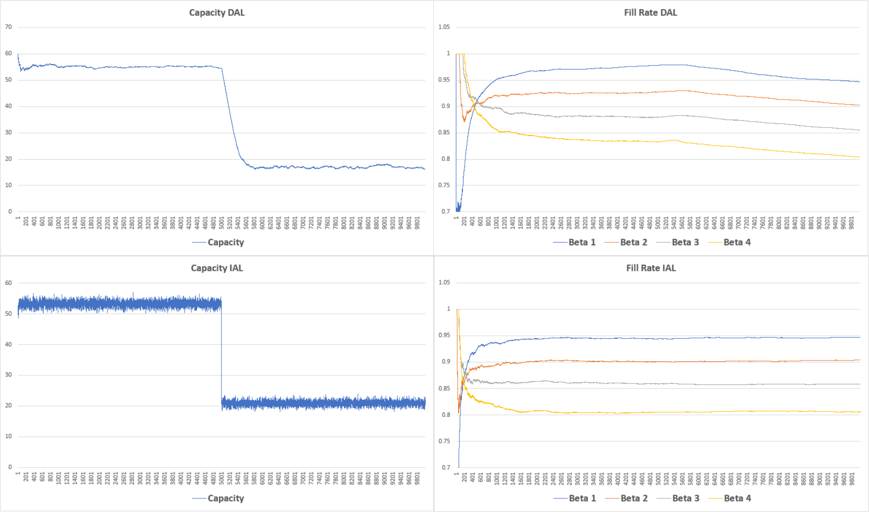

**Figure 3** Numerical results of DAL and IAL algorithms with service level (covering) constraints for the non-stationary case 2.

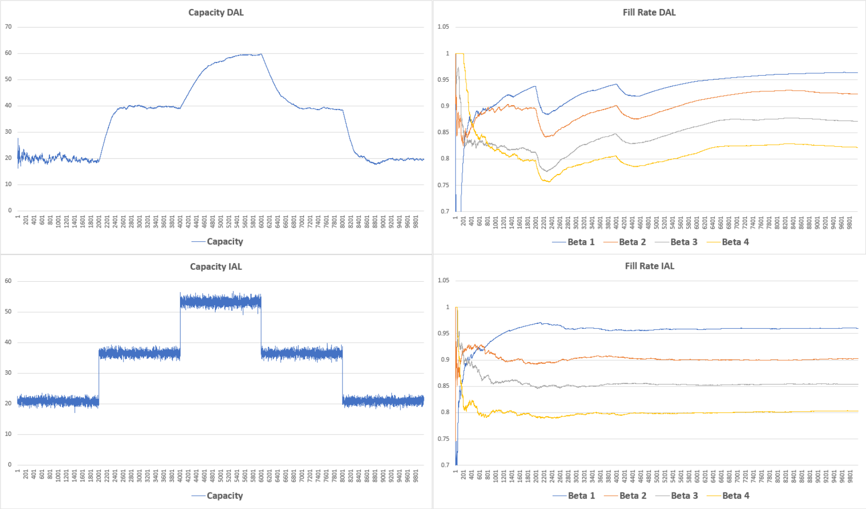

**Figure 4** Numerical results of DAL and IAL algorithms with service level (covering) constraints for the non-stationary case 3.

develops new algorithms via adversarial learning algorithms. The framework is applied to various setting. For the stationary setting, the resulted DAL algorithm is shown to achieve a sublinear regret, with a performance robust to adversarial corruptions. For the non-stationary (adversarial) setting, a modified IAL algorithm is developed, with the help of prior estimates. The sublinear regret can also be acheived by IAL algorithm as long as the cumulative inaccuracy of the prior estimates is sublinear.

## References

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

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

S. Agrawal, Z. Wang, and Y. Ye. A dynamic near-optimal algorithm for online linear programming. *Operations Research*, 62(4):876–890, 2014.

L. An, A. A. Li, B. Moseley, and R. Ravi. The newsvendor with advice. *arXiv preprint arXiv:2305.07993*, 2023.

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

A. Arlotto and I. Gurvich. Uniformly bounded regret in the multisecretary problem. *Stochastic Systems*, 2019.

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

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

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

S. Balseiro, C. Kroer, and R. Kumar. Single-leg revenue management with advice. *arXiv preprint arXiv:2202.10939*, 2022a.

S. R. Balseiro, H. Lu, and V. Mirrokni. The best of many worlds: Dual mirror descent for online allocation problems. *Operations Research*, 2022b.

S. R. Balseiro, H. Lu, and V. Mirrokni. The best of many worlds: Dual mirror descent for online allocation problems. *Operations Research*, 71(1):101–119, 2023.

S. Banerjee, V. Gkatzelis, A. Gorokh, and B. Jin. Online nash social welfare maximization with predictions. In *Proceedings of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1–19. SIAM, 2022.

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

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

J. R. Birge and F. Louveaux. *Introduction to stochastic programming*. Springer Science & Business Media, 2011.

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

M. Castiglioni, A. Celli, and C. Kroer. Online learning with knapsacks: the best of both worlds. In *International Conference on Machine Learning*, pages 2767–2783. PMLR, 2022.

A. Celli, M. Castiglioni, and C. Kroer. Best of many worlds guarantees for online learning with knapsacks. *arXiv preprint arXiv:2202.13710*, 2022.

W. C. Cheung, D. Simchi-Levi, and R. Zhu. Reinforcement learning for non-stationary markov decision processes: The blessing of (more) optimism. In *International Conference on Machine Learning*, pages 1843–1854. PMLR, 2020.

P. Dütting, S. Lattanzi, R. Paes Leme, and S. Vassilvitskii. Secretaries with advice. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 409–429, 2021.

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

Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

N. Golrezaei, P. Jaillet, and Z. Zhou. Online resource allocation with convex-set machine-learned advice. *arXiv preprint arXiv:2306.12282*, 2023.

A. Gupta and M. Molinaro. How experts can solve lps online. In *European Symposium on Algorithms*, pages 517–529. Springer, 2014.

A. Gupta, T. Koren, and K. Talwar. Better algorithms for stochastic bandits with adversarial corruptions. In *Conference on Learning Theory*, pages 1562–1578. PMLR, 2019.

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

B. Hao, R. Jain, T. Lattimore, B. Van Roy, and Z. Wen. Leveraging demonstrations to improve online learning: Quality matters.

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

I.-H. Hou, V. Borkar, and P. Kumar. *A theory of QoS for wireless*. IEEE, 2009.

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

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

R. Jenatton, J. Huang, and C. Archambeau. Adaptive algorithms for online convex optimization with long-term constraints. In *International Conference on Machine Learning*, pages 402–411. PMLR, 2016.

J. Jiang and J. Zhang. Online resource allocation with stochastic resource consumption. 11 2019. doi: 10.13140/RG.2.2.27542.09287.

J. Jiang, X. Li, and J. Zhang. Online stochastic optimization with wasserstein based non-stationarity. *arXiv preprint arXiv:2012.06961*, 2020.

B. Jin and W. Ma. Online bipartite matching with advice: Tight robustness-consistency tradeoffs for the two-stage model. *Advances in Neural Information Processing Systems*, 35:14555–14567, 2022.

M. Kearns, S. Neel, A. Roth, and Z. S. Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International conference on machine learning*, pages 2564–2572. PMLR, 2018.

T. Kesselheim, A. Tönnis, K. Radke, and B. Vöcking. Primal beats dual on online packing lps in the random-order model. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 303–312. ACM, 2014.

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

X. Li, C. Sun, and Y. Ye. Simple and fast algorithm for binary integer and online linear programming. *Advances in Neural Information Processing Systems*, 33:9412–9421, 2020.

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

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

T. Lykouris, V. Mirrokni, and R. Paes Leme. Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 114–122, 2018.

L. Lyu and W. C. Cheung. Bandits with knapsacks: advice on time-varying demands. In *International Conference on Machine Learning*, pages 23212–23238. PMLR, 2023.

M. Mahdavi, R. Jin, and T. Yang. Trading regret for efficiency: online convex optimization with long term constraints. *Journal of Machine Learning Research*, 13(Sep):2503–2528, 2012.

M. Mahdian, H. Nazerzadeh, and A. Saberi. Online optimization with uncertain information. *ACM Transactions on Algorithms (TALG)*, 8(1):1–29, 2012.

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

M. Molinaro and R. Ravi. The geometry of online packing linear programs. *Mathematics of Operations Research*, 39(1):46–59, 2014.

A. Munoz and S. Vassilvitskii. Revenue optimization with approximate bid predictions. *Advances in Neural Information Processing Systems*, 30, 2017.

M. J. Neely and H. Yu. Online convex optimization with time-varying constraints. *arXiv preprint arXiv:1702.04783*, 2017.

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

D. Rohatgi. Near-optimal bounds for online caching with machine learned advice. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1834–1845. SIAM, 2020.

M. Sion. On general minimax theorems. *Pacific Journal of mathematics*, 8(1):171–176, 1958.

X. Yi, X. Li, T. Yang, L. Xie, T. Chai, and K. Johansson. Regret and cumulative constraint violation analysis for online convex optimization with long term constraints. In *International Conference on Machine Learning*, pages 11998–12008. PMLR, 2021.

J. Yuan and A. Lamperski. Online convex optimization for cumulative constraints. In *Advances in Neural Information Processing Systems*, pages 6137–6146, 2018.

M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th international conference on machine learning (icml-03)*, pages 928–936, 2003.

## Appendix A: Further Related Literature

Our model synthesizes the bandits-based model, where the representative problems include MAB problem, bandits with knapsack (BwK) problem and the more general OCO problem, and the feature-based model, where the representative problems include online allocation problem and a special case online packing problem. We now review these literature.

One representative problem for the bandits-based model is the BwK problem which reduces to MAB problem if no long-term constraints. The previous BwK results have focused on a stochastic setting (Badanidiyuru et al. 2013, Agrawal and Devanur 2014a), where a $O(\sqrt{T})$ regret bound has been derived, and an adversarial setting (e.g. Rangi et al. (2018), Immorlica et al. (2019)), where the sublinear regret is impossible to obtain and a $O(\log T)$ competitive ratio has been derived. Similar to our DAL algorithm, the algorithms in the previous literature reply on an interplay between the primal and dual LPs. To be more concrete, Agrawal and Devanur (2014a) and Agrawal and Devanur (2014b) develop a dual-based algorithms to BwK and a more general online stochastic optimization problem (belonging to bandits-based model) and analyzes the algorithm performance under further stronger conditions on the dual optimal solution. However, these learning models and algorithms are developed in the stationary (stochastic) environment, which cannot be applied to the non-stationary setting. For the non-stationary setting, the recent work Liu et al. (2022) derives sublinear regret, based on a more involved complexity measure over the non-stationarity of the underlying distributions which concerns both the temporary changes of two neighborhood distributions and the global changes of the entire distribution sequence.

Another representative problem for the bandits-based model is the OCO problem, which is one of the leading online learning frameworks (Hazan 2016). Note that the standard OCO problem generally adopts a static optimal policy as the benchmark, i.e., the decision of the benchmark needs to be the same for each period. In contrast, in our model, the benchmark is a more powerful dynamic optimal policy where the decisions are allowed to be non-homogeneous across time. Therefore, not only our model is more involved, our benchmark is also stronger. There have been results that consider a dynamic optimal policy as the benchmark for OCO (Besbes et al. 2015, Hall and Willett 2013, Jadbabaie et al. 2015), but all these works consider the unconstrained setting with no long-term constraints. For the line of works that study the problem of online convex optimization with constraints (OCOwC), existing literature would assume the constraint functions that characterize the long-term constraints are either static (Jenatton et al. 2016, Yuan and Lamperski 2018, Yi et al. 2021) or stochastically generated (Neely and Yu 2017).

For the type-based model, one representative problem is the online packing problem, where the columns and the corresponding coefficient in the objective of the underlying LP come one by one and the decision has to be made on-the-fly. The packing problem covers a wide range of applications, including secretary problem (Ferguson et al. 1989, Arlotto and Gurvich 2019), online knapsack problem (Arlotto and Xie 2020, Jiang and Zhang 2019), resource allocation problem (Li et al. 2020), network routing problem (Buchbinder and Naor 2009), matching problem (Mehta et al. 2007) etc. The problem is usually studied under either a stochastic model where the reward and size of each query is drawn independently from an unknown distribution $\mathcal{P}$, or a more general the random permutation model where the queries arrive in a random order (Molinaro and Ravi 2014, Agrawal et al. 2014, Kesselheim et al. 2014, Gupta and Molinaro 2014). The more general online allocation problem (e.g. Balseiro et al. (2022b)) has also been considered in the literature, where the objective and the constraint functions are allowed to be general functions.

## Appendix B: Extensions

### B.1. Non-convex Objective and Non-concave Constraints

Note that in Assumption 1, we require a convexity condition where the function $p(\cdot)$ is a convex function with $p(\mathbf{0})=0$ and $\mathcal{C}$ is a compact and convex set which contains $\mathbf{0}$. Also, the functions $f_\theta(\cdot)$ need to be convex in $\boldsymbol{x}$ and $\boldsymbol{g}_\theta(\cdot,\cdot)$ need to be convex in both $\boldsymbol{c}$ and $\boldsymbol{x}$, for any $\theta$. In this section, we explore whether these convexity conditions can be removed with further characterization of our model.

We now show that when $\mathcal{C}$ contains only a finite number of elements, we can remove all the convexity requirements in Assumption 1 and we still obtain a $\tilde{O}(\sqrt{T}+W_T)$ regret bound for the setting in Section 4. As for the setting in Section 3, we can obtain the same $\tilde{O}(\sqrt{T}+W_T)$ regret bound for *arbitrary* $\mathcal{C}$, as long as the minimization step $\boldsymbol{c}_t=\mathrm{argmin}_{\boldsymbol{c}\in\mathcal{C}}\hat{L}_t(\boldsymbol{c},\mu\cdot\boldsymbol{e}_{i_t})$ can be efficiently solved in Algorithm 1.

We now assume $\mathcal{C}$ has a finite support and remove all the convexity requirements in Assumption 1. In order to recover the results in Section 4, we regard each element in $\mathcal{C}$ as an expert and we apply expert algorithms to decide $\boldsymbol{c}_t$ for each period $t$. To be more concrete, note that for the first-stage decision, we only have *bandit-feedback*, i.e., we can only observe the stochastic outcome of selecting $\boldsymbol{c}_t$ instead of other elements in the set $\mathcal{C}$. Therefore, we can apply EXP3 algorithm (Auer et al. 2002) as $\mathsf{ALG}_1$. Still, following the same procedure as the proof of Theorem 4, we can reduce the regret bound of Algorithm 2 into the regret bound of $\mathsf{ALG}_1$, $\mathsf{ALG}_2$, and an additional $O(W_T)$ term if adversarial corruption exists. The above argument is formalized in the following theorem.

THEOREM 5. *Denote by $\pi$ Algorithm 2 with input $\mu=T$. Then, if $\mathcal{C}$ constains only a finite number of elements, denoted by $K$, with condition c in Assumption 1, it holds that*

$$\begin{aligned} \mathit{Regret}^c(\pi,T)=\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[\mathsf{ALG}(\pi,\boldsymbol{\theta}^c)]-\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[\mathsf{ALG}(\pi^*,\boldsymbol{\theta}^c)] \leq & \tilde{O}(\sqrt{K\cdot T})+\tilde{O}(\sqrt{T\cdot\log m}) \\ & +O(\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[W(\boldsymbol{\theta})]), \end{aligned} \tag{16}$$

*if EXP3 is selected as $\mathsf{ALG}_1$ and Hedge is selected as $\mathsf{ALG}_2$. Here, $\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[W(\boldsymbol{\theta})]$ denotes the expectation of the total number of corruptions, with $W(\boldsymbol{\theta})$ defined in (13).*

We now consider the setting in Section 3 and we show that the $\tilde{O}(\sqrt{T}+W_T)$ regret bound holds for Algorithm 1 without the convexity requirements in Assumption 1, for arbitrary $\mathcal{C}$ as long as the minimization step $\boldsymbol{c}_t=\mathrm{argmin}_{\boldsymbol{c}\in\mathcal{C}}\hat{L}_t(\boldsymbol{c},\mu\cdot\boldsymbol{e}_{i_t})$ can be efficiently solved. The key idea is to regard the first-stage decision at each period as a distribution over $\mathcal{C}$, denoted by $\tilde{\boldsymbol{c}}_t$. Then, $\mathbb{E}_{\tilde{\boldsymbol{C}}}\left[\hat{L}(\tilde{\boldsymbol{C}},\boldsymbol{\lambda})\right]$ would be a convex function over the distribution $\tilde{\boldsymbol{C}}$. Moreover, for each fixed $\boldsymbol{\lambda}$, note that selecting the worst distribution $\tilde{\boldsymbol{C}}$ to minimize $\mathbb{E}_{\tilde{\boldsymbol{C}}}\left[\hat{L}(\tilde{\boldsymbol{C}},\boldsymbol{\lambda})\right]$ can be reduced equivalent to selecting the worst deterministic $\boldsymbol{C}$ to minimize $\hat{L}(\boldsymbol{C},\boldsymbol{\lambda})$. Therefore, every step of the proof of Lemma 2 and Theorem 2 would follow by replacing $\boldsymbol{c}_t$ into the distribution $\tilde{\boldsymbol{c}}_t$ for each $t\in[T]$. We have the following result.

Theorem 6. *Denote by $\pi$ Algorithm 1 with input $\mu = \|\hat{\boldsymbol{\lambda}}^*\|_\infty = \alpha \cdot T$ for some constant $\alpha > 0$. Denote by $W_T = \sum_{t=1}^T W(\hat{P}_t, P_t)$ the total measure of inaccuracy, with $W(\hat{P}_t, P_t)$ defined in (4). Then, under condition c of Assumption 1, for arbitrary $\mathcal{C}$, the regret enjoys the upper bound*

$$\mathit{Regret}(\pi, T) = \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[\mathsf{ALG}(\pi, \boldsymbol{\theta})] - \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[\mathsf{ALG}(\pi^*, \boldsymbol{\theta})] \leq \tilde{O}(\sqrt{T \cdot \log m}) + O(W_T) \tag{17}$$

*if Hedge is selected as $\mathsf{ALG}_{\mathsf{Dual}}$. Moreover, we have*

$$\frac{1}{T}\sum_{t=1}^T g_{i,\theta_t}(\boldsymbol{c}_t, \boldsymbol{x}_t) - \beta_i \leq \tilde{O}\left(\sqrt{\frac{\log m}{T}}\right) + O\left(\frac{W_T}{T}\right), \tag{18}$$

*for each $i \in [m]$.*

## B.2. Incorporating Covering Constraints

In this section, we discuss how to incorporate covering constraints into our model. Note that in the previous section, we let the long-term constraints $\frac{1}{T} \cdot \sum_{t=1}^T \boldsymbol{x}_t \in \mathcal{B}(\boldsymbol{C}, \boldsymbol{\theta})$ be characterized as follows,

$$\frac{1}{T} \cdot \sum_{t=1}^T \boldsymbol{g}_{\theta_t}(\boldsymbol{c}_t, \boldsymbol{x}_t) \leq \boldsymbol{\beta}, \tag{19}$$

with $\boldsymbol{\beta} \in (0,1)^m$, which corresponds to packing constraints since both $\boldsymbol{g}(\cdot)$ and $\boldsymbol{\beta}$ are non-negative. For the case with covering constraints, the long-term constraints $\frac{1}{T} \cdot \sum_{t=1}^T \boldsymbol{x}_t \in \mathcal{B}(\boldsymbol{C}, \boldsymbol{\theta})$ would enjoy the following characterization

$$\frac{1}{T} \cdot \sum_{t=1}^T \boldsymbol{g}_{\theta_t}(\boldsymbol{c}_t, \boldsymbol{x}_t) \geq \boldsymbol{\beta}, \tag{20}$$

with $\boldsymbol{\beta} \in (0,1)^m$. We now show that all our previous results hold for the packing constraints (20), by simply changing the input parameter $\mu$ in Algorithm 2 and Algorithm 1. We illustrate through the stationary setting consiered in Section 4 where $P_t = P$ for each $t \in [T]$.

Now, (OPT) enjoys the following new formulation.

$$\begin{aligned} \mathsf{OPT}^{\mathsf{Covering}} = \min & \sum_{t=1}^T \mathbb{E}_{\tilde{\boldsymbol{c}}_t, \tilde{\boldsymbol{x}}_t, \theta_t}[p(\tilde{\boldsymbol{c}}_t) + f_{\theta_t}(\tilde{\boldsymbol{x}}_t)] \\ \text{s.t.} & \frac{1}{T} \cdot \sum_{t=1}^T \mathbb{E}_{\tilde{\boldsymbol{c}}_t, \tilde{\boldsymbol{x}}_t, \theta_t}[\boldsymbol{g}_{\theta_t}(\tilde{\boldsymbol{c}}_t, \tilde{\boldsymbol{x}}_t)] \geq \boldsymbol{\beta} \\ & \tilde{\boldsymbol{x}}_t \in \mathcal{K}(\theta_t, \tilde{\boldsymbol{c}}_t), \tilde{\boldsymbol{c}}_t \in \mathcal{C}, \forall t. \end{aligned} \tag{21}$$

The Lagrangian dual of $\mathsf{OPT}^{\mathsf{Covering}}$ (21) can be formulated as follows.

$$\max_{\boldsymbol{\lambda}\geq 0} \min_{\tilde{\boldsymbol{c}}_t \in \mathcal{C}} \mathsf{L}^{\mathsf{Covering}}(\boldsymbol{C}, \boldsymbol{\lambda}) = \sum_{t=1}^T \mathbb{E}\left[p(\tilde{\boldsymbol{c}}_t) + \frac{1}{T} \cdot \sum_{i=1}^m \lambda_i + \min_{\tilde{\boldsymbol{x}}_t \in \mathcal{K}(\theta_t, \tilde{\boldsymbol{c}}_t)} f_{\theta_t}(\tilde{\boldsymbol{x}}_t) - \sum_{i=1}^m \frac{\lambda_i \cdot g_{i,\theta_t}(\tilde{\boldsymbol{c}}_t, \tilde{\boldsymbol{x}}_t)}{T \cdot \beta_i}\right]. \tag{22}$$

Similar to (14), we have the following formulation for $\bar{L}(\boldsymbol{c}, \boldsymbol{\lambda}, \theta)$ as the single-period decomposition of $\mathsf{L}^{\mathsf{Covering}}(\boldsymbol{C}, \boldsymbol{\lambda})$.

$$\bar{L}^{\mathsf{Covering}}(\boldsymbol{c}, \boldsymbol{\lambda}, \theta) = p(\boldsymbol{c}) + \frac{1}{T} \cdot \sum_{i=1}^m \lambda_i + \min_{\boldsymbol{x} \in \mathcal{K}(\theta, \boldsymbol{c})} \left\{ f_\theta(\boldsymbol{x}) - \sum_{i=1}^m \frac{\lambda_i \cdot g_{i,\theta}(\boldsymbol{c}, \boldsymbol{x})}{T \cdot \beta_i} \right\}. \tag{23}$$

We can obtain the following result.

Lemma 4. *Under the stationary setting where $P_t = P$ for each $t \in [T]$, it holds that*

$$\max_{\boldsymbol{\lambda}\geq 0}\min_{\tilde{\boldsymbol{c}}_t\in\mathcal{C}} \mathbb{E}_{\tilde{\boldsymbol{C}}}[\mathsf{L}^{\mathsf{Covering}}(\boldsymbol{C},\boldsymbol{\lambda})] = \max_{\boldsymbol{\lambda}\geq 0}\min_{\tilde{\boldsymbol{c}}\in\mathcal{C}} T\cdot \mathbb{E}_{\tilde{\boldsymbol{c}},\theta\sim P}\left[\bar{L}^{\mathsf{Covering}}(\tilde{\boldsymbol{c}},\boldsymbol{\lambda},\theta)\right].$$

Therefore, we can still regard solving the dual problem (22) as a procedure of solving the repeated zero-sum games, where player 1 chooses the first-stage decision $\boldsymbol{c}_t$ and player 2 chooses one long term constraint $i_t \in [m]$. We still apply Algorithm 2 with the new definition

$$\bar{L}_i^{\mathsf{Covering}}(\boldsymbol{c}_t, \mu\cdot \boldsymbol{e}_{i_t}, \theta_t) = p(\boldsymbol{c}_t) + \frac{\mu}{T} + f_{\theta_t}(\boldsymbol{x}_t) - \frac{\mu\cdot g_{i,\theta_t}(\boldsymbol{c}_t,\boldsymbol{x}_t)}{T\cdot\beta_i}. \tag{24}$$

However, in order to incorporate covering constraints, we need to adjust the value of the scaling factor $\mu$ and we only terminate our algorithm after the entire horizon has been run out. In fact, we need an upper bound (arbitrary upper bound suffices) over the $l_1$ norm of the optimal dual variable $\boldsymbol{\lambda}^*$ of (22) to serve as the scaling factor $\mu$ and in practice, we can spend the first $\sqrt{T}$ time periods to construct an upper confidence interval of $\boldsymbol{\lambda}^*$ to serve as $\mu$ without influencing the order of the regret bound.

Theorem 7. *Denote by $\pi$ Algorithm 2 with input $\mu$ being an upper bound of $\|\boldsymbol{\lambda}^*\|_1$ where $\boldsymbol{\lambda}^*$ is the optimal dual variable of (22). Then, under Assumption 1, $\mu = \alpha\cdot T$ for some constant $\alpha > 0$ and if OGD is selected as $\mathsf{ALG}_1$ and Hedge is selected as $\mathsf{ALG}_2$, the regret enjoys the upper bound*

$$Regret^{TSC}(\pi,T) \leq \tilde{O}((G+F)\cdot\sqrt{T}) + \tilde{O}(\sqrt{T\cdot\log m}) \tag{25}$$

*and moreover, we have*

$$\beta_i - \frac{1}{T}\cdot\sum_{t=1}^{T} g_{i,\theta_t}(\boldsymbol{c}_t,\boldsymbol{x}_t) \leq \tilde{O}\left(\frac{G+F}{\alpha\sqrt{T}}\right) + \tilde{O}\left(\frac{\sqrt{\log m}}{\alpha\cdot\sqrt{T}}\right), \quad \forall i\in[m]. \tag{26}$$

## Appendix C: Regret Bounds for Adversarial Learning

In this section, we present the implementation details of two adversarial learning algorithms, OGD and Hedge, that will be used as algorithmic subroutines in our Algorithm 2 and Algorithm 1, as well as their regret analysis.

OGD is an algorithm to be executed in a finite horizon of $T$ periods, and at each period $t$, OGD selects an action $\boldsymbol{c}_t \in \mathcal{C}$, receives an adversarial chosen cost function $h_t(\cdot)$ afterwards, and incurs a cost $h_t(\boldsymbol{c}_t)$. OGD is designed to minimize the regret

$$\mathrm{Reg}_{\mathrm{OGD}}(T) = \sum_{t=1}^{T} h_t(\boldsymbol{c}_t) - \min_{\boldsymbol{c}\in\mathcal{C}}\sum_{t=1}^{T} h_t(\boldsymbol{c}).$$

The implementation of OGD is described in Algorithm 3. In our problem, $h_t = \bar{L}_i^{\mathrm{ISP}}(\boldsymbol{c}_t, \mu\cdot\boldsymbol{e}_{i_t}, \theta_t)$ for ISP and $h_t = \bar{L}_i^{\mathrm{OSP}}(\boldsymbol{c}_t, \mu\cdot\boldsymbol{e}_{i_t}, \theta_t)$ for OSP. In order to obtain a subgradient, one can compute the optimal dual variable of the inner minimization problem of (14). For example, the inner minimization problem is

$$\begin{aligned} \min \quad & f_{\theta_t}(\boldsymbol{x}) + \frac{\mu\cdot g_{i_t,\theta_t}(\boldsymbol{c},\boldsymbol{x})}{T\cdot\beta_{i_t}} \\ \text{s.t.} \quad & B_{\theta_t}\boldsymbol{x} \leq \boldsymbol{c}_t \\ & \boldsymbol{x} \geq 0 \end{aligned} \tag{27}$$

**Algorithm 3** Online Gradient Descent (OGD) algorithm

```
Input: the step size $\eta_t$ for each $t \in [T]$.
Initially set an arbitrarily $\boldsymbol{c}_1 \in \mathcal{C}$.
for $t = 1, \ldots, T$ do
    1. Take the action $\boldsymbol{c}_t$.
    2. Observe the cost function $h_t(\cdot)$.
    3. Update action
        $\boldsymbol{c}_{t+1} = \mathcal{P}_{\mathcal{C}}\left(\boldsymbol{c}_t - \eta_t \cdot \nabla h_t(\boldsymbol{c}_t)\right)$
    where $\nabla h_t(\boldsymbol{c}_t)$ denotes a subgradient of $h_t$ at $\boldsymbol{c}_t$ and $\mathcal{P}_{\mathcal{C}}$ denotes a projection to the set $\mathcal{C}$.
end for
```

and the Lagrangian dual problem of (27) is

$$\max_{\boldsymbol{\gamma} \leq 0} \boldsymbol{\gamma}^\top \boldsymbol{c_t} + \min_{\boldsymbol{x} \geq 0} f_{\theta_t}(\boldsymbol{x}) - \boldsymbol{\gamma}^\top B_{\theta_t} \boldsymbol{x} + \frac{\mu \cdot g_{i_t, \theta_t}(\boldsymbol{c}_t, \boldsymbol{x})}{T \cdot \beta_{i_t}}.$$

The optimal dual solution $\boldsymbol{\gamma}_t^*$ can be computed from the above minimax problem and we know that

$$\nabla h_t(\boldsymbol{c}_t) = \boldsymbol{\gamma}_t^* + \frac{\mu \cdot \nabla_{\boldsymbol{c_t}} g_{i_t, \theta_t}(\boldsymbol{c}_t, \boldsymbol{x})}{T \cdot \beta_{i_t}}.$$

Clearly when $\mu = a \cdot T$ for a constant $a > 0$, we have $\|\boldsymbol{\gamma}_t^*\|_2 \leq G$ for some constant $G$ that depends only on (the upper bound of gradients of) $\{f_\theta, \boldsymbol{g}_\theta\}_{\forall \theta}$, the minimum positive element of $B_\theta$ for all $\theta$, and $\max_{i \in [m]}\{\frac{1}{\beta_i}\}$. The regret bound of OGD is as follows.

THEOREM 8 **(Theorem 1 of Zinkevich (2003))**. *If $\eta_t = \frac{1}{\sqrt{t}}$, then it holds that*

$$Reg_{OGD}(T) \leq O\left((G + F) \cdot \sqrt{T}\right)$$

*where $F$ is an upper bound of the diameter of the set $\mathcal{C}$ and $G$ is a constant that depends only on (the upper bound of gradients of) $\{f_\theta, \boldsymbol{g}_\theta\}_{\forall \theta}$, the minimum positive element of $B_\theta$ for all $\theta$, and $\max_{i \in [m]}\{\frac{1}{\beta_i}\}$.*

The Hedge algorithm is used to solve the expert problem in a finite horizon of $T$ periods. There are $m$ experts and at each period $t$, Hedge will select one expert $i_t \in [m]$ ($i_t$ can be randomly chosen), observe the reward vector $\boldsymbol{l}_t \in \mathbb{R}^m$ afterwards, and obtain an reward $l_{i_t, t}$. Hedge is designed to minimize the regret

$$\text{Reg}_{\text{Hedge}}(T) = \max_{i \in [m]} \sum_{t=1}^T l_{i,t} - \sum_{t=1}^T \mathbb{E}_{i_t}[l_{i_t, t}].$$

The Hedge algorithm is described in Algorithm 4. In our problem, for ISP, we have

$$l_{i,t} = \bar{L}_i(\boldsymbol{c}_t, \mu \cdot \boldsymbol{e}_{i_t}, \theta_t).$$

Under Assumption 1, when $\mu = \alpha \cdot T$ for a constant $\alpha$, we know that there exists a constant $\delta > 0$ such that $|l_{i,t}| \leq \delta$, for all $i \in [m]$ and $t \in [T]$. Here, $\delta$ depends on $\max_{i \in [m]}\{\frac{1}{\beta_i}\}$.

THEOREM 9 **(from Theorem 2 in Freund and Schapire (1997))**. *If $\varepsilon = \sqrt{\frac{\log m}{T}}$, then it holds that*

$$Reg_{Hedge}(T) \leq \tilde{O}(\sqrt{T \cdot \log(m)})$$

*where the constant term in $\tilde{O}(\cdot)$ depends on $\max_{i \in [m]}\{\frac{1}{\beta_i}\}$.*

**Algorithm 4** Hedge algorithm

```
Input: a parameter $\varepsilon > 0$.
Initialize: $\boldsymbol{w}_1 = \mathbf{1} \in \mathbb{R}^m$ and $\boldsymbol{y}_1 = \frac{1}{m} \cdot \boldsymbol{w}_1$.
for $t = 1, \ldots, T$ do
  1. Take the action $i_t \sim \boldsymbol{y}_t$.
  2. Observe the reward vector $\boldsymbol{l}_t$ and obtain a reward $l_{i_t,t}$.
  3. Update the weight
        $w_{i,t+1} = w_{i,t} \cdot \exp(-\varepsilon \cdot l_{i,t})$
     for each $i \in [m]$ and set
        $y_{i,t+1} = \frac{w_{i,t+1}}{\sum_{i'=1}^{m} w_{i',t+1}}$
     for each $i \in [m]$.
end for
```

## Appendix D: Missing Proofs for Section 3

*Proof of Theorem 1.* The proof is modified from the proof of Theorem 3. We consider a special case of our problem where for any $\boldsymbol{c} \in \mathcal{C}$ and any $\theta$, $\mathcal{K}(\theta, \boldsymbol{c}) = [0,1]$ (there is no need to decide the first-stage decision). There is only one long-term constraint with target $\beta = \frac{1}{2}$. Moreover, there are three possible values of $\theta$, denoted by $\{\theta^1, \theta^2, \theta^3\}$. We have $f_{\theta^1}(x) = -x$, $f_{\theta^2}(x) = -\left(1 + \frac{W_T}{T}\right)x$, $f_{\theta^3}(x) = -\left(1 - \frac{W_T}{T}\right)x$ and $g_{\theta^1}(x) = g_{\theta^2}(x) = g_{\theta^3}(x) = x$ (only one long-term constraint). The prior estimate is $\hat{P}_t = \theta^1$ with probability 1 for each $t \in [T]$, and the problem with respect to the prior estimates can be described below in (28).

$$\begin{aligned} \min \quad & -x_1 - \ldots - x_{\frac{T}{2}} - x_{\frac{T}{2}+1} - \ldots - x_T \\ \text{s.t.} \quad & x_1 + \ldots + x_{\frac{T}{2}} + x_{\frac{T}{2}+1} + \ldots + x_T \leq \frac{T}{2} \\ & 0 \leq x_t \leq 1 \quad \text{for } t = 1, \ldots, T. \end{aligned} \tag{28}$$

Now we consider the following two possible true distributions. The first possible true scenario, given in (29), is that the distribution $P_t = \theta^1$ for $t = 1, \ldots, \frac{T}{2}$ and $P_t = \theta^2$ for $t = \frac{T}{2} + 1, \ldots, T$. The second possible true scenario, given in (30), is that the distribution $P_t = \theta^1$ for $t = 1, \ldots, \frac{T}{2}$ and $P_t^c = \theta^3$ for $t = \frac{T}{2} + 1, \ldots, T$.

$$\begin{aligned} \min \quad & -x_1 - \ldots - x_{\frac{T}{2}} - \left(1 + \frac{W_T}{T}\right) x_{\frac{T}{2}+1} - \ldots - \left(1 + \frac{W_T}{T}\right) x_T \\ \text{s.t.} \quad & x_1 + \ldots + x_{\frac{T}{2}} + x_{\frac{T}{2}+1} + \ldots + x_T \leq \frac{T}{2} \\ & 0 \leq x_t \leq 1 \quad \text{for } t = 1, \ldots, T. \end{aligned} \tag{29}$$

$$\begin{aligned} \min \quad & -x_1 - \ldots - x_{\frac{T}{2}} - \left(1 - \frac{W_T}{T}\right) x_{\frac{T}{2}+1} - \ldots - \left(1 - \frac{W_T}{T}\right) x_T \\ \text{s.t.} \quad & x_1 + \ldots + x_{\frac{T}{2}} + x_{\frac{T}{2}+1} + \ldots + x_T \leq \frac{T}{2} \\ & 0 \leq x_t \leq 1 \quad \text{for } t = 1, \ldots, T. \end{aligned} \tag{30}$$

For any online policy $\pi$, denote by $x_t^1(\pi)$ the decision of the policy $\pi$ at period $t$ under the true scenario given in (29) and denote by $x_t^2(\pi)$ the decision of the policy $\pi$ at period $t$ under the true scenario (30). Further define $T_1(\pi)$ (resp. $T_2(\pi)$) as the expected capacity consumption of policy $\pi$ under the true scenario (29) (resp. true scenario (30)) 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

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

where $\mathsf{ALG}_T^1(\pi)$ (resp. $\mathsf{ALG}_T^2(\pi)$) denotes the expected reward collected by policy $\pi$ on scenario (29) (resp. scenario (30)). It is clear to see that the optimal policy $\pi^*$ who is aware of $P_t$ for each $t\in[T]$ can achieve an objective value

$$\mathsf{ALG}_T^1(\pi^*)=-\frac{T+W_T}{2} \quad \text{and} \quad \mathsf{ALG}_T^2(\pi^*)=-\frac{T}{2}.$$

Thus, the regret of policy $\pi$ on scenario (29) and (30) 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 $\pi$ at each time period should be independent of future realizations, we must have $T_1(\pi)=T_2(\pi)$ (during the first $\frac{T}{2}$ periods, the information for $\pi$ is the same for both scenarios (29) and (30)). 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. □

*Proof of Lemma 2.* Denote by $(\hat{\boldsymbol{\lambda}}^*,(\hat{\boldsymbol{c}}_t^*)_{t=1}^T)$ the optimal solution to

$$\max_{\boldsymbol{\lambda}\geq 0}\min_{\boldsymbol{c}_t\in\mathcal{C}}\hat{L}(\boldsymbol{C},\boldsymbol{\lambda}),$$

used in the definition (6). we now show that $(\hat{\boldsymbol{\lambda}}^*,\hat{\boldsymbol{c}}_t^*)$ is an optimal solution to

$$\max_{\boldsymbol{\lambda}\geq 0}\min_{\boldsymbol{c}_t\in\mathcal{C}}\hat{L}_t(\boldsymbol{c}_t,\boldsymbol{\lambda})$$

for each $t\in[T]$, which would help to complete our proof of (8). We first define

$$L_t(\boldsymbol{\lambda})=\min_{\boldsymbol{c}_t\in\mathcal{C}}\hat{L}_t(\boldsymbol{c}_t,\boldsymbol{\lambda}),$$

as a function over $\boldsymbol{\lambda}$ for each $t\in[T]$. Then, it holds that

$$\nabla L_t(\hat{\boldsymbol{\lambda}}^*)=\nabla\hat{L}_t(\hat{\boldsymbol{c}}_t^*,\hat{\boldsymbol{\lambda}}^*)=\left(-\frac{\hat{\beta}_{i,t}}{T\beta_i}+\mathbb{E}_{\theta\sim\hat{P}_t}\left[\frac{g_{i,\theta}(\hat{\boldsymbol{c}}_t^*,\hat{\boldsymbol{x}}_t^*(\theta))}{T\beta_i}\right]\right)_{\forall i\in[m]}=\mathbf{0} \tag{31}$$

The first equality of (31) follows from the fact that

$$\hat{\boldsymbol{c}}_t^*\in\text{argmin}_{\boldsymbol{c}\in\mathcal{C}}\mathbb{E}_{\theta\sim\hat{P}_t}\left[p(\boldsymbol{c})+\min_{\boldsymbol{x}\in\mathcal{K}(\theta,\boldsymbol{c})}f_\theta(\boldsymbol{x})+\sum_{i=1}^m\frac{\hat{\lambda}_i^*\cdot g_{i,\theta}(\boldsymbol{c},\boldsymbol{x})}{T\cdot\beta_i}\right] \tag{32}$$

since $(\hat{\boldsymbol{\lambda}}^*,(\hat{\boldsymbol{c}}_t^*)_{t=1}^T)$ is an optimal solution to $\max_{\boldsymbol{\lambda}\geq 0}\min_{\boldsymbol{c}_t\in\mathcal{C}}\hat{L}(\boldsymbol{C},\boldsymbol{\lambda})$. The last equality of (31) follows from the definition of $\hat{\beta}_{i,t}$ in (6). Therefore, combining (31) and (32), we know that $(\hat{\boldsymbol{\lambda}}^*,\hat{\boldsymbol{c}}_t^*)$ is an optimal solution to $\max_{\boldsymbol{\lambda}\geq 0}\min_{\boldsymbol{c}_t\in\mathcal{C}}\hat{L}_t(\boldsymbol{c}_t,\boldsymbol{\lambda})$ for each $t\in[T]$.

We now prove (8). It is sufficient to prove that

$$\hat{L}(\hat{\boldsymbol{C}}^*, \hat{\boldsymbol{\lambda}}^*) = \sum_{t=1}^{T} \hat{L}_t(\hat{\boldsymbol{c}}_t^*, \hat{\boldsymbol{\lambda}}^*) \tag{33}$$

where $\hat{\boldsymbol{C}}^* = (\hat{\boldsymbol{c}}_t^*)_{t=1}^T$. We define an index set $\mathcal{I} = \{i \in [m] : \hat{\lambda}_i^* > 0\}$. Clearly, for each $i \in \mathcal{I}$, the optimality of $\hat{\boldsymbol{\lambda}}^*$ would require that

$$\nabla_{\lambda_i} \bar{L}(\hat{\boldsymbol{C}}^*, \hat{\boldsymbol{\lambda}}^*) = -1 + \frac{1}{T\beta_i} \sum_{t=1}^{T} \mathbb{E}_{\theta \sim \hat{P}_t} \left[ g_{i,\theta}(\hat{\boldsymbol{c}}_t^*, \hat{\boldsymbol{x}}_t^*(\theta)) \right] = 0$$

which implies $\sum_{t=1}^T \hat{\beta}_{i,t} = T \cdot \beta$. Therefore, we would have

$$\sum_{i=1}^{m} \sum_{t=1}^{T} \frac{\hat{\lambda}_{i,t}^* \hat{\beta}_i}{T\beta_i} = \sum_{i=1}^{m} \hat{\lambda}_i^*.$$

which completes our proof of (9) and therefore (8). □

*Proof of Theorem 2.* The proof can be classified by the following two steps. We denote by

$$\mathsf{L}_t(\boldsymbol{c}, \boldsymbol{\lambda}) := \mathbb{E}_{\theta \sim P_t} \left[ p(\boldsymbol{c}) - \sum_{i=1}^{m} \frac{\lambda_i \cdot \hat{\beta}_{i,t}}{T \cdot \beta_i} + \min_{\boldsymbol{x} \in \mathcal{K}(\theta, \boldsymbol{c})} f_\theta(\boldsymbol{x}) + \sum_{i=1}^{m} \frac{\lambda_i \cdot g_{i,\theta}(\boldsymbol{c}, \boldsymbol{x})}{T \cdot \beta_i} \right] \tag{34}$$

for each $t \in [T]$. Our first step is to show that for any $\boldsymbol{\lambda} \geq 0$ and any $\boldsymbol{c} \in \mathcal{C}$, it holds that

$$\left| \mathsf{L}_t(\boldsymbol{c}, \boldsymbol{\lambda}) - \hat{L}_t(\boldsymbol{c}, \boldsymbol{\lambda}) \right| \leq \frac{\|\boldsymbol{\lambda}\|_1}{T\beta_{\min}} \cdot W(\hat{P}_t, P_t) \tag{35}$$

where $\beta_{\min} = \min_{i \in [m]}\{\beta_i\}$.

We now prove (35). We define

$$\hat{L}_t(\boldsymbol{c}, \boldsymbol{\lambda}, \theta) := p(\boldsymbol{c}) - \sum_{i=1}^{m} \frac{\lambda_i \cdot \hat{\beta}_{i,t}}{T \cdot \beta_i} + \min_{\boldsymbol{x} \in \mathcal{K}(\theta, \boldsymbol{c})} f_\theta(\boldsymbol{x}) + \sum_{i=1}^{m} \frac{\lambda_i \cdot g_{i,\theta}(\boldsymbol{c}, \boldsymbol{x})}{T \cdot \beta_i}.$$

It is clear to see that

$$\mathsf{L}_t(\boldsymbol{c}, \boldsymbol{\lambda}) = \mathbb{E}_{\theta \sim P_t} \left[ \hat{L}_t(\boldsymbol{c}, \boldsymbol{\lambda}, \theta) \right] \text{ and } \hat{L}_t(\boldsymbol{c}, \boldsymbol{\lambda}) = \mathbb{E}_{\theta \sim \hat{P}_t} \left[ \hat{L}_t(\boldsymbol{c}, \boldsymbol{\lambda}, \theta) \right].$$

Moreover, note that for any $\theta, \theta'$, we have

$$|\hat{L}_t(\boldsymbol{c}, \boldsymbol{\lambda}, \theta) - \hat{L}_t(\boldsymbol{c}, \boldsymbol{\lambda}, \theta')| \leq \frac{\|\boldsymbol{\lambda}\|_1}{T\beta_{\min}} \cdot d(\theta, \theta'),$$

with $d(\theta, \theta') = \|(f_\theta, \boldsymbol{g}_\theta) - (f_{\theta'}, \boldsymbol{g}_{\theta'})\|_\infty$ in the definition (4). Therefore, we have

$$\left| \mathsf{L}_t(\boldsymbol{c}, \boldsymbol{\lambda}) - \hat{L}_t(\boldsymbol{c}, \boldsymbol{\lambda}) \right| = \left| \mathbb{E}_{\theta \sim P_t} \left[ \hat{L}_t(\boldsymbol{c}, \boldsymbol{\lambda}, \theta) \right] - \mathbb{E}_{\theta' \sim \hat{P}_t} \left[ \hat{L}_t(\boldsymbol{c}, \boldsymbol{\lambda}, \theta') \right] \right| \leq \frac{\|\boldsymbol{\lambda}\|_1}{T\beta_{\min}} \cdot W(\hat{P}_t, P_t),$$

thus, complete our proof of (35).

Our second step is to bound the final regret with the help of (35). We assume without loss of generality that there always exists $i' \in [m]$ such that

$$\sum_{t=1}^{T} g_{i',\theta_t}(\boldsymbol{c}_t, \boldsymbol{x}_t) \geq T \cdot \beta_{i'}. \tag{36}$$

In fact, let there be a *dummy* constraint $i'$ such that $g_{i',\theta}(\boldsymbol{c}, \boldsymbol{x}) = \beta_{i'} = \alpha$, for arbitrary $\alpha \in (0,1)$, for any $\theta$ and $\boldsymbol{c}, \boldsymbol{x}$. Then, (36) holds.

Let $(\hat{\boldsymbol{\lambda}}^*, (\hat{\boldsymbol{c}}_t^*)_{t=1}^T)$ be the optimal solution to $\max_{\boldsymbol{\lambda}\geq 0}\min_{\boldsymbol{c}_t\in\mathcal{C}}\hat{L}(\boldsymbol{C},\boldsymbol{\lambda})$ used in the definition (6). Then, it holds that

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\sum_{t=1}^{T}\hat{L}_t(\boldsymbol{c}_t,\mu\cdot\boldsymbol{e}_{i_t},\theta_t)\right] &= \sum_{t=1}^{T}\mathbb{E}_{\boldsymbol{c}_t,i_t}\left[\mathsf{L}_t(\boldsymbol{c}_t,\mu\cdot\boldsymbol{e}_{i_t})\right] \\ &\leq \sum_{t=1}^{T}\mathbb{E}_{\boldsymbol{c}_t,i_t}\left[\hat{L}_t(\boldsymbol{c}_t,\mu\cdot\boldsymbol{e}_{i_t})\right]+\frac{\mu\cdot W_T}{T\beta_{\min}} = \sum_{t=1}^{T}\mathbb{E}_{i_t}\left[\min_{\boldsymbol{c}\in\mathcal{C}}\hat{L}_t(\boldsymbol{c},\mu\cdot\boldsymbol{e}_{i_t})\right]+\frac{\mu\cdot W_T}{T\beta_{\min}} \\ &\leq \sum_{t=1}^{T}\min_{\boldsymbol{c}\in\mathcal{C}}\hat{L}_t(\boldsymbol{c},\hat{\boldsymbol{\lambda}}^*)+\frac{\mu\cdot W_T}{T\beta_{\min}} = \sum_{t=1}^{T}\hat{L}_t(\hat{\boldsymbol{c}}_t^*,\hat{\boldsymbol{\lambda}}^*)+\frac{\mu\cdot W_T}{T\beta_{\min}}\end{aligned} \tag{37}$$

where the first inequality follows from the definition of $\boldsymbol{c}_t$, the second inequality follows from Lemma 2, and the last equality follows from the definition of $\hat{\boldsymbol{c}}_t^*$.

On the other hand, for any $i\in[m]$, we have

$$\sum_{t=1}^{T}\hat{L}_t(\boldsymbol{c}_t,\mu\cdot\boldsymbol{e}_{i_t},\theta_t)\geq\sum_{t=1}^{T}\hat{L}_{i,t}(\boldsymbol{c}_t,\mu\cdot\boldsymbol{e}_{i_t},\theta_t)-\mathrm{Reg}(T,\boldsymbol{\theta})$$

with $\hat{L}_{i,t}$ defined in (11), where $\mathrm{Reg}(\tau,\boldsymbol{\theta})$ denotes the regret bound of $\mathsf{ALG}_{\mathsf{Dual}}$ (holds for arbitrary $\boldsymbol{\lambda}=\mu\cdot\boldsymbol{e}_i$). We now denote by

$$i^*=\mathrm{argmax}_{i\in[m]}\{\frac{1}{T}\cdot\sum_{t=1}^{T}g_{i,\theta_t}(\boldsymbol{c}_t,\boldsymbol{x}_t)-\beta_i\}.$$

We also denote by

$$d_T(\mathcal{A},\boldsymbol{\theta})=\max_{i\in[m]}\{\frac{1}{T}\cdot\sum_{t=1}^{T}g_{i,\theta_t}(\boldsymbol{c}_t,\boldsymbol{x}_t)-\beta_i\}.$$

From (36), we must have $d_T(\mathcal{A},\boldsymbol{\theta})\geq 0$. We now set $i=i^*$ and we have

$$\begin{aligned}\sum_{t=1}^{T}\hat{L}_t(\boldsymbol{c}_t,\mu\cdot\boldsymbol{e}_{i_t},\theta_t) &\geq \sum_{t=1}^{T}\hat{L}_{i^*,t}(\boldsymbol{c}_t,\mu\cdot\boldsymbol{e}_{i_t},\theta_t)-\mathrm{Reg}(T,\boldsymbol{\theta}) \\ &= \sum_{t=1}^{T}(p(\boldsymbol{c}_t)+f_{\theta_t}(\boldsymbol{x}_t))-\mu\cdot\frac{\sum_{t=1}^{T}\hat{\beta}_{i^*,t}}{T\beta_{i^*}}+\sum_{t=1}^{T}\frac{\mu\cdot g_{i^*,\theta_t}(\boldsymbol{c}_t,\boldsymbol{x}_t)}{T\cdot\beta_{i^*}}-\mathrm{Reg}(T,\boldsymbol{\theta})\end{aligned} \tag{38}$$

From the construction of $\hat{\beta}_{i,t}$, we know that $\sum_{t=1}^{T}\hat{\beta}_{i,t}\leq T\cdot\beta_i$ for each $i\in[m]$. To see this point, we note that if we define $\hat{L}(\boldsymbol{\lambda})=\min_{\boldsymbol{c}_t\in\mathcal{C}}\hat{L}(\boldsymbol{C},\boldsymbol{\lambda})$, then, for each $i\in[m]$,

$$\nabla_{\lambda_i}\hat{L}(\hat{\boldsymbol{\lambda}}^*)=-1+\mathbb{E}_{\boldsymbol{\theta}\sim\hat{P}}\left[\sum_{t=1}^{T}\frac{g_{i,\theta_t}(\hat{\boldsymbol{c}}_t^*,\hat{\boldsymbol{x}}^*(\theta_t))}{T\beta_i}\right]=-1+\frac{\sum_{t=1}^{T}\hat{\beta}_{i,t}}{T\cdot\beta_i}\leq 0. \tag{39}$$

Otherwise, $\nabla_{\lambda_i}\hat{L}(\hat{\boldsymbol{\lambda}}^*)>0$ would violate the optimality of $\hat{\boldsymbol{\lambda}}^*$ to $\max_{\boldsymbol{\lambda}\geq 0}\hat{L}(\boldsymbol{\lambda})$.

Plugging (39) into (38), we get

$$\begin{aligned}\sum_{t=1}^{T}\hat{L}_t(\boldsymbol{c}_t,\mu\cdot\boldsymbol{e}_{i_t},\theta_t) &\geq \sum_{t=1}^{T}(p(\boldsymbol{c}_t)+f_{\theta_t}(\boldsymbol{x}_t))-\mu+\sum_{t=1}^{T}\frac{\mu\cdot g_{i^*,\theta_t}(\boldsymbol{c}_t,\boldsymbol{x}_t)}{T\cdot\beta_{i^*}}-\mathrm{Reg}(T,\boldsymbol{\theta}) \\ &\geq \sum_{t=1}^{T}(p(\boldsymbol{c}_t)+f_{\theta_t}(\boldsymbol{x}_t))+\frac{\mu}{\beta_{\min}}\cdot d_T(\mathcal{A},\boldsymbol{\theta})-\mathrm{Reg}(T,\boldsymbol{\theta})\end{aligned} \tag{40}$$

Combining (37) and (40), we have

$$\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\sum_{t=1}^{T}(p(\boldsymbol{c}_t)+f_{\theta_t}(\boldsymbol{x}_t))\right]\leq\sum_{t=1}^{T}\hat{L}_t(\hat{\boldsymbol{c}}_t^*,\hat{\boldsymbol{\lambda}}^*)+\frac{\mu\cdot W_T}{T\beta_{\min}}+\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\mathrm{Reg}(T,\boldsymbol{\theta})\right]-\frac{\mu}{\beta_{\min}}\cdot\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[d_T(\mathcal{A},\boldsymbol{\theta})\right]. \tag{41}$$

Denote by $(\boldsymbol{\lambda}^*, \tilde{\boldsymbol{C}}^*)$ the optimal *saddle-point* solution to

$$\max_{\boldsymbol{\lambda}\geq 0} \min_{\tilde{\boldsymbol{C}}\in\mathcal{C}} \mathbb{E}_{\tilde{\boldsymbol{C}}}[\mathsf{L}(\tilde{\boldsymbol{C}}, \boldsymbol{\lambda})] = \min_{\tilde{\boldsymbol{C}}\in\mathcal{C}} \max_{\boldsymbol{\lambda}\geq 0} \mathbb{E}_{\tilde{\boldsymbol{C}}}[\mathsf{L}(\tilde{\boldsymbol{C}}, \boldsymbol{\lambda})].$$

Then, it holds that

$$\begin{aligned}\sum_{t=1}^{T} \hat{L}_t(\hat{\boldsymbol{c}}_t^*, \hat{\boldsymbol{\lambda}}^*) &\leq \sum_{t=1}^{T} \mathbb{E}_{\tilde{\boldsymbol{c}}_t}[\hat{L}_t(\tilde{\boldsymbol{c}}_t^*, \hat{\boldsymbol{\lambda}}^*)] \leq \sum_{t=1}^{T} \mathbb{E}_{\tilde{\boldsymbol{c}}_t}[\mathsf{L}_t(\tilde{\boldsymbol{c}}_t^*, \hat{\boldsymbol{\lambda}}^*)] + \frac{\|\hat{\boldsymbol{\lambda}}^*\|_1 \cdot W_T}{T\beta_{\min}} = \mathbb{E}_{\tilde{\boldsymbol{C}}}[\mathsf{L}(\tilde{\boldsymbol{C}}^*, \hat{\boldsymbol{\lambda}}^*)] + \frac{\|\hat{\boldsymbol{\lambda}}^*\|_1 \cdot W_T}{T\beta_{\min}} \\ &\leq \mathbb{E}_{\tilde{\boldsymbol{C}}}[\mathsf{L}(\tilde{\boldsymbol{C}}^*, \hat{\boldsymbol{\lambda}}^*)] + \frac{\|\hat{\boldsymbol{\lambda}}^*\|_1 \cdot W_T}{T\beta_{\min}} = \mathsf{OPT} + \frac{\|\hat{\boldsymbol{\lambda}}^*\|_1 \cdot W_T}{T\beta_{\min}},\end{aligned} \tag{42}$$

where the first inequality follows from definition of $(\hat{\boldsymbol{\lambda}}^*, (\hat{\boldsymbol{c}}_t)_{t=1}^T)$, the second inequality follows from (35), the first equality follows from (9) and the third inequality follows from the saddle-point condition of $(\boldsymbol{\lambda}^*, \tilde{\boldsymbol{C}}^*)$.

Plugging (42) into (41), we have

$$\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\sum_{t=1}^{T}(p(\boldsymbol{c}_t) + f_{\theta_t}(\boldsymbol{x}_t))\right] \leq \mathsf{OPT} + \frac{2\mu W_T}{T\beta_{\min}} + \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\mathrm{Reg}(T, \boldsymbol{\theta})\right] - \frac{\mu}{\beta_{\min}} \cdot \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[d_T(\mathcal{A}, \boldsymbol{\theta})\right]. \tag{43}$$

with $\mu = \|\hat{\boldsymbol{\lambda}}^*\|_1$. From the non-negativity of $d_T(\mathcal{A}, \boldsymbol{\theta})$, we have

$$\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\sum_{t=1}^{T}(p(\boldsymbol{c}_t) + f_{\theta_t}(\boldsymbol{x}_t))\right] \leq \mathsf{OPT} + \frac{2\mu W_T}{T\beta_{\min}} + \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\mathrm{Reg}(T, \boldsymbol{\theta})\right].$$

Using Theorem 9 to bound $\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\mathrm{Reg}(T, \boldsymbol{\theta})\right]$, we have

$$\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\sum_{t=1}^{T}(p(\boldsymbol{c}_t) + f_{\theta_t}(\boldsymbol{x}_t))\right] \leq \mathsf{OPT} + \frac{2\mu W_T}{T\beta_{\min}} + \tilde{O}(\sqrt{T \cdot \log m})$$

which completes our proof of (12) by noting that $\mu = \alpha \cdot T$ for some constant $\alpha > 0$.

We note that $(\boldsymbol{c}_t, \boldsymbol{x}_t)_{t=1}^T$ defines a feasible solution to

$$\begin{aligned}\mathsf{OPT}^{\delta} = \min &\sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{c}_t, \boldsymbol{x}_t, \theta_t \sim P_t}\left[p(\boldsymbol{c}_t) + f_{\theta_t}(\boldsymbol{x}_t)\right] \\ \text{s.t.} &\frac{1}{T} \cdot \sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{c}_t, \boldsymbol{x}_t, \theta_t \sim P_t}[\boldsymbol{g}_{\theta_t}(\boldsymbol{c}_t, \boldsymbol{x}_t)] \leq \boldsymbol{\beta} + \delta \cdot \boldsymbol{e} \\ &\boldsymbol{x}_t \in \mathcal{K}(\theta_t, \boldsymbol{c}_t), \boldsymbol{c}_t \in \mathcal{C}, \forall t.\end{aligned} \tag{44}$$

with $\delta = \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[d_T(\mathcal{A}, \boldsymbol{\theta})]$ and $\boldsymbol{e}$ denotes a vector of all ones. We define another optimization problem by changing $P_t$ into $\hat{P}_t$ for each $t$,

$$\begin{aligned}\hat{\mathsf{OPT}}^{\hat{\delta}} = \min &\sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{c}_t, \boldsymbol{x}_t, \theta_t \sim \hat{P}_t}\left[p(\boldsymbol{c}_t) + f_{\theta_t}(\boldsymbol{x}_t)\right] \\ \text{s.t.} &\frac{1}{T} \cdot \sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{c}_t, \boldsymbol{x}_t, \theta_t \sim \hat{P}_t}[\boldsymbol{g}_{\theta_t}(\boldsymbol{c}_t, \boldsymbol{x}_t)] \leq \boldsymbol{\beta} + \hat{\delta} \cdot \boldsymbol{e} \\ &\boldsymbol{x}_t \in \mathcal{K}(\theta_t, \boldsymbol{c}_t), \boldsymbol{c}_t \in \mathcal{C}, \forall t.\end{aligned} \tag{45}$$

We have the following result that bounds the gap between $\mathsf{OPT}^{\delta}$ and $\hat{\mathsf{OPT}}^{\hat{\delta}}$, for some specific $\delta$ and $\hat{\delta}$.

Claim 1. *For any $\delta \geq 0$, if we set $\hat{\delta} = \delta + \frac{W_T}{T}$, then we have*

$$\hat{\mathsf{OPT}}^{\hat{\delta}} \leq \mathsf{OPT}^{\delta} + W_T.$$

If we regard $\hat{\mathsf{OPT}}^{\hat{\delta}}$ as a function over $\hat{\delta}$, then $\hat{\mathsf{OPT}}^{\hat{\delta}}$ is clearly a convex function over $\hat{\delta}$, where the proof follows the same spirit as the proof of Lemma 3. Moreover, note that

$$\frac{d\hat{\mathsf{OPT}}^{\hat{\delta}=0}}{d\hat{\delta}} = \|\hat{\boldsymbol{\lambda}}^*\|_1 \leq \mu.$$

We have

$$\hat{\mathsf{OPT}} = \hat{\mathsf{OPT}}^0 \leq \hat{\mathsf{OPT}}^{\hat{\delta}} + \mu \cdot \hat{\delta}$$

for any $\hat{\delta} \geq 0$. On the other hand, we have

$$\hat{\mathsf{OPT}} = \max_{\boldsymbol{\lambda}\geq 0} \min_{\boldsymbol{c}_t \in \mathcal{C}} \hat{L}(\boldsymbol{C}, \boldsymbol{\lambda}) = \sum_{t=1}^{T} \hat{L}_t(\hat{\boldsymbol{c}}_t^*, \hat{\boldsymbol{\lambda}}^*) \tag{46}$$

where the last equality follows from Lemma 2. Therefore, we now assume that $\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[d_T(\mathcal{A}, \boldsymbol{\theta})] \geq \frac{W_T}{T}$, otherwise (**??**) directly holds. We have

$$\begin{aligned}
\sum_{t=1}^{T} \hat{L}_t(\hat{\boldsymbol{c}}_t^*, \hat{\boldsymbol{\lambda}}^*) = \hat{\mathsf{OPT}} &\leq \hat{\mathsf{OPT}}^{\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[d_T(\mathcal{A},\boldsymbol{\theta})]+\frac{W_T}{T}} + \mu \cdot \left(\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[d_T(\mathcal{A}, \boldsymbol{\theta})] + \frac{W_T}{T}\right) \\
&\leq \mathsf{OPT}^{\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[d_T(\mathcal{A},\boldsymbol{\theta})]} + \mu \cdot \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[d_T(\mathcal{A}, \boldsymbol{\theta})] + W_T \cdot (1 + \frac{\mu}{T}) \\
&\leq \sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[p(\boldsymbol{c}_t) + f_{\theta_t}(\boldsymbol{x}_t)\right] + \mu \cdot \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[d_T(\mathcal{A}, \boldsymbol{\theta})] + W_T \cdot (1 + \frac{\mu}{T})
\end{aligned} \tag{47}$$

where the second inequality follows from Claim 1 by setting $\delta = \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[d_T(\mathcal{A}, \boldsymbol{\theta})]$. Plugging (47) into (41), we have

$$\begin{aligned}
\sum_{t=1}^{T} \hat{L}_t(\hat{\boldsymbol{c}}_t^*, \hat{\boldsymbol{\lambda}}^*) \leq \sum_{t=1}^{T} \hat{L}_t(\hat{\boldsymbol{c}}_t^*, \hat{\boldsymbol{\lambda}}^*) + \frac{\mu \cdot W_T}{T\beta_{\min}} + \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\mathrm{Reg}(T, \boldsymbol{\theta})\right] - \frac{\mu}{\beta_{\min}} \cdot \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[d_T(\mathcal{A}, \boldsymbol{\theta})\right] \\
+ \mu \cdot \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[d_T(\mathcal{A}, \boldsymbol{\theta})] + W_T \cdot (1 + \frac{\mu}{T}),
\end{aligned}$$

which implies

$$\mu \cdot \left(\frac{1}{\beta_{\min}} - 1\right) \cdot \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[d_T(\mathcal{A}, \boldsymbol{\theta})] \leq \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\mathrm{Reg}(T, \boldsymbol{\theta})\right] + \frac{2\mu W_T}{T\beta_{\min}} + +W_T \cdot (1 + \frac{\mu}{T}).$$

which completes our final proof by noting that $\mu = \alpha \cdot T$ for some constant $\alpha > 0$. □

*Proof of Claim 1.* Denote by $(\tilde{\boldsymbol{c}}_t, \tilde{\boldsymbol{x}}_t)_{t=1}^T$ one optimal solution to $\mathsf{OPT}^{\delta}$. Then, from the definition of $W_T$, we have that

$$\frac{1}{T} \cdot \sum_{t=1}^{T} \mathbb{E}_{\tilde{\boldsymbol{c}}_t, \tilde{\boldsymbol{x}}_t, \theta_t \sim \hat{P}_t}[\boldsymbol{g}_{\theta_t}(\boldsymbol{c}_t, \boldsymbol{x}_t)] \leq \frac{1}{T} \cdot \sum_{t=1}^{T} \mathbb{E}_{\tilde{\boldsymbol{c}}_t, \tilde{\boldsymbol{x}}_t, \theta_t \sim P_t}[\boldsymbol{g}_{\theta_t}(\boldsymbol{c}_t, \boldsymbol{x}_t)] + \frac{W_T}{T}.$$

Therefore, we know that $(\tilde{\boldsymbol{c}}_t, \tilde{\boldsymbol{x}}_t)_{t=1}^T$ is a feasible solution to $\hat{\mathsf{OPT}}^{\hat{\delta}}$. On the other hand, from the definition of $W_T$, we know that

$$\hat{\mathsf{OPT}}^{\hat{\delta}} \leq \sum_{t=1}^{T} \mathbb{E}_{\tilde{\boldsymbol{c}}_t, \tilde{\boldsymbol{x}}_t, \theta_t \sim \hat{P}_t}\left[p(\tilde{\boldsymbol{c}}_t) + f_{\theta_t}(\tilde{\boldsymbol{x}}_t)\right] \leq \sum_{t=1}^{T} \mathbb{E}_{\tilde{\boldsymbol{c}}_t, \tilde{\boldsymbol{x}}_t, \theta_t \sim \hat{P}_t}\left[p(\tilde{\boldsymbol{c}}_t) + f_{\theta_t}(\tilde{\boldsymbol{x}}_t)\right] + W_T = \mathsf{OPT}^{\delta} + W_T$$

by noting that the distribution of $\tilde{\boldsymbol{c}}_t$ must be independent of $\theta_t$ for each $t$, which completes our proof. □

## Appendix E: Missing Proofs for Section 4

*Proof of Theorem 3.* The proof is modified from the proof of Theorem 2 in Jiang et al. (2020). We let $W_T = \mathbb{E}_{\boldsymbol{\theta}\sim \boldsymbol{P}}[W(\boldsymbol{\theta})]$. We consider a special case of our problem where for any $\boldsymbol{c}\in\mathcal{C}$ and any $\theta$, $\mathcal{K}(\theta,\boldsymbol{c})=[0,1]$ (there is no need to decide the first-stage decision). There is only one long-term constraint with target $\beta=\frac{1}{2}$. Moreover, there are three possible values of $\theta$, denoted by $\{\theta^1,\theta^2,\theta^3\}$. We have $f_{\theta^1}(x)=-x$, $f_{\theta^2}(x)=-\left(1+\frac{W_T}{T}\right)x$, $f_{\theta^3}(x)=-\left(1-\frac{W_T}{T}\right)x$ and $g_{\theta^1}(x)=g_{\theta^2}(x)=g_{\theta^3}(x)=x$ (only one long-term constraint). The true distribution is $P_t=\theta^1$ with probability 1 for each $t\in[T]$, and the problem with respect to the true distributions can be described below in (48).

$$\begin{aligned}\min\quad & -x_1-\ldots-x_{\frac{T}{2}}-x_{\frac{T}{2}+1}-\ldots-x_T\\ \text{s.t.}\quad & x_1+\ldots+x_{\frac{T}{2}}+x_{\frac{T}{2}+1}+\ldots+x_T\le\frac{T}{2}\\ & 0\le x_t\le 1\quad\text{for } t=1,\ldots,T.\end{aligned}\tag{48}$$

Now we consider the following two possible adversarial corruptions. The first possible corruption, given in (49), is that the distribution $P_t^c=\theta^2$ for $t=\frac{T}{2}+1,\ldots,T$. The second possible corruption, given in (50), is that the distribution $P_t^c=\theta^3$ for $t=\frac{T}{2}+1,\ldots,T$.

$$\begin{aligned}\min\quad & -x_1-\ldots-x_{\frac{T}{2}}-\left(1+\frac{W_T}{T}\right)x_{\frac{T}{2}+1}-\ldots-\left(1+\frac{W_T}{T}\right)x_T\\ \text{s.t.}\quad & x_1+\ldots+x_{\frac{T}{2}}+x_{\frac{T}{2}+1}+\ldots+x_T\le\frac{T}{2}\\ & 0\le x_t\le 1\quad\text{for } t=1,\ldots,T.\end{aligned}\tag{49}$$

$$\begin{aligned}\min\quad & -x_1-\ldots-x_{\frac{T}{2}}-\left(1-\frac{W_T}{T}\right)x_{\frac{T}{2}+1}-\ldots-\left(1-\frac{W_T}{T}\right)x_T\\ \text{s.t.}\quad & x_1+\ldots+x_{\frac{T}{2}}+x_{\frac{T}{2}+1}+\ldots+x_T\le\frac{T}{2}\\ & 0\le x_t\le 1\quad\text{for } t=1,\ldots,T.\end{aligned}\tag{50}$$

For any online policy $\pi$, denote by $x_t^1(\pi)$ the decision of the policy $\pi$ at period $t$ under corruption scenario given in (49) and denote by $x_t^2(\pi)$ the decision of the policy $\pi$ at period $t$ under corruption scenario (50). Further define $T_1(\pi)$ (resp. $T_2(\pi)$) as the expected capacity consumption of policy $\pi$ under corruption scenario (49) (resp. corruption scenario (50)) 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

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

where $\mathsf{ALG}_T^1(\pi)$ (resp. $\mathsf{ALG}_T^2(\pi)$) denotes the expected reward collected by policy $\pi$ on scenario (49) (resp. scenario (50)). It is clear to see that the optimal policy $\pi^*$ who is aware of $P_t^c$ for each $t\in[T]$ can achieve an objective value

$$\mathsf{ALG}_T^1(\pi^*)=-\frac{T+W_T}{2}\quad\text{and}\quad \mathsf{ALG}_T^2(\pi^*)=-\frac{T}{2}.$$

Thus, the regret of policy $\pi$ on scenario (49) and (50) 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 $\pi$ at each time period should be independent of future realizations, and more importantly, should independent of corruptions in the future, we must have $T_1(\pi)=T_2(\pi)$ (during the first $\frac{T}{2}$ periods, the information for $\pi$ is the same for both scenarios (49) and (50)). Thus, we have that

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

which completes our proof. □

*Proof of Lemma 3.* Fix arbitrary $\boldsymbol{\lambda},\theta$, for any $\boldsymbol{c}_1,\boldsymbol{c}_2\in\mathcal{C}$ and any $\alpha_1,\alpha_2\geq0$ such that $\alpha_1+\alpha_2=1$, we prove

$$\alpha_1\cdot\bar{L}(\boldsymbol{c}_1,\boldsymbol{\lambda},\theta)+\alpha_2\cdot\bar{L}(\boldsymbol{c}_2,\boldsymbol{\lambda},\theta)\geq\bar{L}(\alpha_1\cdot\boldsymbol{c}_1+\alpha_2\cdot\boldsymbol{c}_2,\boldsymbol{\lambda},\theta).$$

Now, for $\bar{L}(\boldsymbol{c}_1,\boldsymbol{\lambda},\theta)$, we denote by $\boldsymbol{x}_1^*(\theta)$ one optimal solution of the inner minimization problem in the definition of $\bar{L}(\boldsymbol{c}_1,\boldsymbol{\lambda},\theta)$ (14). Similarly, for $\bar{L}(\boldsymbol{c}_2,\boldsymbol{\lambda},\theta)$, we denote by $\boldsymbol{x}_2^*(\theta)$ one optimal solution of the inner minimization problem in the definition of $\bar{L}(\boldsymbol{c}_2,\boldsymbol{\lambda},\theta)$ (14). We then define

$$\boldsymbol{x}_3^*(\theta)=\alpha_1\cdot\boldsymbol{x}_1^*(\theta)+\alpha_2\cdot\boldsymbol{x}_2^*(\theta),\quad\forall\theta.$$

Under Assumption 1, from the convexity of $f_\theta(\cdot)$ and $\boldsymbol{g}_\theta(\cdot)$, it holds that

$$\begin{aligned}&\alpha_1\cdot\left(f_\theta(\boldsymbol{x}_1^*(\theta))+\sum_{i=1}^m\frac{\lambda_i\cdot g_{i,\theta}(\boldsymbol{c}_1,\boldsymbol{x}_1^*(\theta))}{T\cdot\beta_i}\right)+\alpha_2\cdot\left(f_\theta(\boldsymbol{x}_2^*(\theta))+\sum_{i=1}^m\frac{\lambda_i\cdot g_{i,\theta}(\boldsymbol{c}_2,\boldsymbol{x}_2^*(\theta))}{T\cdot\beta_i}\right)\\&\geq f_\theta(\boldsymbol{x}_3^*(\theta))+\sum_{i=1}^m\frac{\lambda_i\cdot g_{i,\theta}(\boldsymbol{c}_3,\boldsymbol{x}_3^*(\theta))}{T\cdot\beta_i}.\end{aligned}$$

with $\boldsymbol{c}_3=\alpha_1\cdot\boldsymbol{c}_1+\alpha_2\cdot\boldsymbol{c}_2$. Given the convexity of $p(\cdot)$, we have

$$\begin{aligned}&\alpha_1\cdot\bar{L}(\boldsymbol{c}_1,\boldsymbol{\lambda},\theta)+\alpha_2\cdot\bar{L}(\boldsymbol{c}_2,\boldsymbol{\lambda},\theta)\\=&\alpha_1p(\boldsymbol{c}_1)+\alpha_2p(\boldsymbol{c}_2)-\frac{1}{T}\sum_{i=1}^m\lambda_i+\alpha_1\left(f_\theta(\boldsymbol{x}_1^*(\theta))+\sum_{i=1}^m\frac{\lambda_i\cdot g_{i,\theta}(\boldsymbol{c}_1,\boldsymbol{x}_1^*(\theta))}{T\cdot\beta_i}\right)\\&+\alpha_2\left(f_\theta(\boldsymbol{x}_2^*(\theta))+\sum_{i=1}^m\frac{\lambda_i\cdot g_{i,\theta}(\boldsymbol{c}_2,\boldsymbol{x}_2^*(\theta))}{T\cdot\beta_i}\right)\\\geq&p(\alpha_1\boldsymbol{c}_1+\alpha_2\boldsymbol{c}_2)+f_\theta(\boldsymbol{x}_3^*(\theta))+\sum_{i=1}^m\frac{\lambda_i\cdot g_{i,\theta}(\boldsymbol{c}_3,\boldsymbol{x}_3^*(\theta))}{T\cdot\beta_i}\\\geq&\bar{L}(\boldsymbol{c}_3,\boldsymbol{\lambda},\theta)\end{aligned}$$

where the last inequality follows from $\boldsymbol{x}_3^*(\theta)\in\mathcal{K}(\theta,\boldsymbol{c}_3)$, given the exact formulation of $\mathcal{K}(\theta,\boldsymbol{c})$ under Assumption 1. □

*Proof of Theorem 4.* Denote by $\tau$ the time period that Algorithm 2 is terminated. There must be a constraint $i'\in[m]$ such that

$$\sum_{t=1}^{\tau}g_{i',\theta_t^c}(\boldsymbol{c}_t,\boldsymbol{x}_t)\geq T\cdot\beta_{i'}.\tag{51}$$

Otherwise, we can assume without loss of generality that there exists a *dummy* constraint $i'$ such that $g_{i',\theta}(\boldsymbol{c}_t,\boldsymbol{x})=\beta_{i'}=\alpha$, for arbitrary $\alpha\in(0,1)$, for any $\theta$ and $\boldsymbol{c},\boldsymbol{x}$. In this case, we can set $\tau=T$.

We denote by $(\boldsymbol{\lambda}^*, \tilde{\boldsymbol{c}}^*)$ one *saddle-point* optimal solution to

$$\max_{\boldsymbol{\lambda}\geq 0}\min_{\tilde{\boldsymbol{c}}\in\mathcal{C}} \mathbb{E}_{\tilde{\boldsymbol{c}}^*,\theta\sim P^c}\left[\bar{L}(\tilde{\boldsymbol{c}},\boldsymbol{\lambda},\theta)\right] = \min_{\tilde{\boldsymbol{c}}\in\mathcal{C}}\max_{\boldsymbol{\lambda}\geq 0} \mathbb{E}_{\tilde{\boldsymbol{c}}^*,\theta\sim P^c}\left[\bar{L}(\tilde{\boldsymbol{c}},\boldsymbol{\lambda},\theta)\right]$$

where $P^c$ denotes the uniform mixture of the distributions of $\{\theta_t^c\}$ for $t=1$ to $\tau$ and the equality follows from the concavity over $\boldsymbol{\lambda}$ and the convexity over $\boldsymbol{c}$ proved in Lemma 3. We have

$$\sum_{t=1}^{\tau}\bar{L}(\boldsymbol{c}_t,\mu\cdot\boldsymbol{e}_{i_t},\theta_t^c) \leq \sum_{t=1}^{\tau}\mathbb{E}_{\tilde{\boldsymbol{c}}^*}\left[\bar{L}(\tilde{\boldsymbol{c}}^*,\mu\cdot\boldsymbol{e}_{i_t},\theta_t^c)\right] + \mathrm{Reg}_1(\tau,\boldsymbol{\theta}^c)$$

following regret bound of $\mathsf{ALG}_1$. Then, it holds that

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\sum_{t=1}^{\tau}\bar{L}(\boldsymbol{c}_t,\mu\cdot\boldsymbol{e}_{i_t},\theta_t^c)\right] &\leq \mathbb{E}_{\tilde{\boldsymbol{c}}^*,\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\sum_{t=1}^{\tau}\bar{L}(\tilde{\boldsymbol{c}}^*,\mu\cdot\boldsymbol{e}_{i_t},\theta_t^c)\right] + \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\mathrm{Reg}_1(\tau,\boldsymbol{\theta}^c)\right] \qquad (52)\\ &\leq \tau\cdot\mathbb{E}_{\tilde{\boldsymbol{c}}^*,\theta\sim P^c}\left[\bar{L}(\tilde{\boldsymbol{c}}^*,\boldsymbol{\lambda}^*,\theta)\right] + \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\mathrm{Reg}_1(\tau,\boldsymbol{\theta}^c)\right]\\ &\leq \tau\cdot\mathbb{E}_{\tilde{\boldsymbol{c}}^*,\theta\sim P}\left[\bar{L}(\tilde{\boldsymbol{c}}^*,\boldsymbol{\lambda}^*,\theta)\right] + \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\mathrm{Reg}_1(\tau,\boldsymbol{\theta}^c)\right] + O(\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[W(\boldsymbol{\theta})]).\end{aligned}$$

On the other hand, for any $i\in[m]$, we have

$$\sum_{t=1}^{\tau}\bar{L}(\boldsymbol{c}_t,\mu\cdot\boldsymbol{e}_{i_t},\theta_t^c) \geq \sum_{t=1}^{\tau}\bar{L}_i(\boldsymbol{c}_t,\mu\cdot\boldsymbol{e}_{i_t},\theta_t^c) - \mathrm{Reg}_2(\tau,\boldsymbol{\theta}^c)$$

following the regret bound of $\mathsf{ALG}_2$ (holds for arbitrary $\boldsymbol{\lambda}=\mu\cdot\boldsymbol{e}_i$). We now set $i=i'$ and we have

$$\begin{aligned}\sum_{t=1}^{\tau}\bar{L}(\boldsymbol{c}_t,\mu\cdot\boldsymbol{e}_{i_t},\theta_t^c) &\geq \sum_{t=1}^{\tau}\bar{L}_{i'}(\boldsymbol{c}_t,\mu\cdot\boldsymbol{e}_{i_t},\theta_t^c) - \mathrm{Reg}_2(\tau,\boldsymbol{\theta}^c)\\ &= \sum_{t=1}^{\tau}\left(p(\boldsymbol{c}_t)+f_{\theta_t^c}(\boldsymbol{x}_t)\right) - \mu\cdot\frac{\tau}{T} + \sum_{t=1}^{\tau}\frac{\mu\cdot g_{i',\theta_t^c}(\boldsymbol{c}_t,\boldsymbol{x}_t)}{T\cdot\beta_{i'}} - \mathrm{Reg}_2(\tau,\boldsymbol{\theta}^c)\\ &\geq \sum_{t=1}^{\tau}\left(p(\boldsymbol{c}_t)+f_{\theta_t^c}(\boldsymbol{x}_t)\right) + \mu\cdot\frac{T-\tau}{T} - \mathrm{Reg}_2(\tau,\boldsymbol{\theta}^c)\end{aligned}$$

where the last inequality follows from (51). Then, it holds that

$$\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\sum_{t=1}^{\tau}\bar{L}(\boldsymbol{c}_t,\mu\cdot\boldsymbol{e}_{i_t},\theta_t^c)\right] \geq \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\sum_{t=1}^{\tau}\left(p(\boldsymbol{c}_t)+f_{\theta_t^c}(\boldsymbol{x}_t)\right)\right] + \mu\cdot\frac{T-\tau}{T} - \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\mathrm{Reg}_2(\tau,\boldsymbol{\theta}^c)\right]. \qquad (53)$$

Combining (52) and (53), we have

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\sum_{t=1}^{\tau}\left(p(\boldsymbol{c}_t)+f_{\theta_t^c}(\boldsymbol{x}_t)\right)\right] \leq &-\mu\cdot\frac{T-\tau}{T} + \tau\cdot\mathbb{E}_{\tilde{\boldsymbol{c}}^*,\theta\sim P}\left[\bar{L}(\tilde{\boldsymbol{c}}^*,\boldsymbol{\lambda}^*,\theta)\right] + O(\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[W(\boldsymbol{\theta})])\\ &+ \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\mathrm{Reg}_1(\tau,\boldsymbol{\theta})\right] + \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\mathrm{Reg}_2(\tau,\boldsymbol{\theta})\right].\end{aligned}$$

From the boundedness conditions in Assumption 1, we have

$$\mathbb{E}_{\tilde{\boldsymbol{c}}^*,\theta\sim P}\left[\bar{L}(\tilde{\boldsymbol{c}}^*,\boldsymbol{\lambda}^*,\theta)\right] = \frac{1}{T}\cdot\mathsf{OPT} \geq -1$$

which implies that

$$-\mu\cdot\frac{T-\tau}{T} \leq (T-\tau)\cdot\mathbb{E}_{\tilde{\boldsymbol{c}}^*,\theta\sim P}\left[\bar{L}(\tilde{\boldsymbol{c}}^*,\boldsymbol{\lambda}^*,\theta)\right]$$

when $\mu=T$. Therefore, we have

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\sum_{t=1}^{\tau}\left(p(\boldsymbol{c}_t)+f_{\theta_t^c}(\boldsymbol{x}_t)\right)\right] \leq & T\cdot\mathbb{E}_{\tilde{\boldsymbol{c}}^*,\theta\sim P}\left[\bar{L}(\tilde{\boldsymbol{c}}^*,\boldsymbol{\lambda}^*,\theta)\right] + O(\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[W(\boldsymbol{\theta})]) + \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\mathrm{Reg}_1(T,\boldsymbol{\theta})\right]\\ &+ \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\mathrm{Reg}_2(T,\boldsymbol{\theta})\right] \qquad (54)\\ \leq & \mathsf{OPT}^c + O(\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[W(\boldsymbol{\theta})]) + \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\mathrm{Reg}_1(T,\boldsymbol{\theta})\right]\\ &+ \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\mathrm{Reg}_2(T,\boldsymbol{\theta})\right].\end{aligned}$$

where $\mathsf{OPT}^c$ denotes the value of the optimal policy with adversarial corruptions. It is clear to see that $|\mathsf{OPT}^c - \mathsf{OPT}| \leq O(\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[W(\boldsymbol{\theta})])$. Using Theorem 8 and Theorem 9 to bound $\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[\mathrm{Reg}_1(T,\boldsymbol{\theta})]$ and $\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[\mathrm{Reg}_2(T,\boldsymbol{\theta})]$, we have

$$\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\sum_{t=1}^{\tau}\left(p(\boldsymbol{c}_t)+f_{\theta_t^c}(\boldsymbol{x}_t)\right)\right] \leq \mathsf{OPT}^c + O((G+F)\cdot\sqrt{T}) + O(\sqrt{T\cdot\log m}) + O(\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[W(\boldsymbol{\theta})])$$

which completes our proof of (15). □

## Appendix F: Missing Proofs for Appendix B

*Proof of Theorem 5.* The proof follows a similar procedure as the proof of Theorem 4. Denote by $\tau$ the time period that Algorithm 2 is terminated. There must be a constraint $i' \in [m]$ such that

$$\sum_{t=1}^{\tau} g_{i',\theta_t^c}(\boldsymbol{c}_t,\boldsymbol{x}_t) \geq T\cdot\beta_{i'}. \tag{55}$$

Otherwise, we can assume without loss of generality that there exists a *dummy* constraint $i'$ such that $g_{i',\theta}(\boldsymbol{c}_t,\boldsymbol{x}) = \beta_{i'} = \alpha$, for arbitrary $\alpha\in(0,1)$, for any $\theta$ and $\boldsymbol{c},\boldsymbol{x}$. In this case, we can set $\tau = T$.

We denote by $(\boldsymbol{\lambda}^*, \tilde{\boldsymbol{c}}^*)$ one *saddle-point* optimal solution to

$$\max_{\boldsymbol{\lambda}\geq 0}\min_{\tilde{\boldsymbol{c}}\in\mathcal{C}} \mathbb{E}_{\tilde{\boldsymbol{c}}^*,\theta\sim P^c}\left[\bar{L}(\tilde{\boldsymbol{c}},\boldsymbol{\lambda},\theta)\right] = \min_{\tilde{\boldsymbol{c}}\in\mathcal{C}}\max_{\boldsymbol{\lambda}\geq 0} \mathbb{E}_{\tilde{\boldsymbol{c}}^*,\theta\sim P^c}\left[\bar{L}(\tilde{\boldsymbol{c}},\boldsymbol{\lambda},\theta)\right] \tag{56}$$

where $P^c$ denotes the uniform mixture of the distributions of $\{\theta_t^c\}$ for $t=1$ to $\tau$. We note that since $\tilde{\boldsymbol{c}}$ is a distribution over $K$ discrete points $\{\boldsymbol{c}^1,\ldots,\boldsymbol{c}^K\}$, $\tilde{\boldsymbol{c}}$ can be denoted by $\boldsymbol{p}_{\tilde{\boldsymbol{c}}}\in\mathbb{R}^m$ with $0\leq\boldsymbol{p}_{\tilde{\boldsymbol{c}}}$ and $\|\boldsymbol{p}_{\tilde{\boldsymbol{c}}}\|_1 = 1$. Then, we have

$$\mathbb{E}_{\tilde{\boldsymbol{c}}^*,\theta\sim P^c}\left[\bar{L}(\tilde{\boldsymbol{c}},\boldsymbol{\lambda},\theta)\right] = \langle\boldsymbol{p}_{\tilde{\boldsymbol{c}}},\bar{\boldsymbol{L}}\rangle$$

with $\bar{\boldsymbol{L}} = (\bar{L}(\boldsymbol{c}^k,\boldsymbol{\lambda},\theta))_{k=1}^K\in\mathbb{R}^K$. Therefore, $\mathbb{E}_{\tilde{\boldsymbol{c}}^*,\theta\sim P^c}\left[\bar{L}(\tilde{\boldsymbol{c}},\boldsymbol{\lambda},\theta)\right]$ can be regarded as a convex function over $\tilde{\boldsymbol{c}}$ and is clearly a concave function over $\boldsymbol{\lambda}$. The equality (56) follows from the Sion's minimax theorem (Sion 1958).

We have

$$\sum_{t=1}^{\tau}\bar{L}(\boldsymbol{c}_t,\mu\cdot\boldsymbol{e}_{i_t},\theta_t^c) \leq \sum_{t=1}^{\tau}\mathbb{E}_{\tilde{\boldsymbol{c}}^*}\left[\bar{L}(\tilde{\boldsymbol{c}}^*,\mu\cdot\boldsymbol{e}_{i_t},\theta_t^c)\right] + \mathrm{Reg}_1(\tau,\boldsymbol{\theta}^c)$$

following regret bound of $\mathsf{ALG}_1$. Then, it holds that

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\sum_{t=1}^{\tau}\bar{L}(\boldsymbol{c}_t,\mu\cdot\boldsymbol{e}_{i_t},\theta_t^c)\right] &\leq \mathbb{E}_{\tilde{\boldsymbol{c}}^*,\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\sum_{t=1}^{\tau}\bar{L}(\tilde{\boldsymbol{c}}^*,\mu\cdot\boldsymbol{e}_{i_t},\theta_t^c)\right] + \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[\mathrm{Reg}_1(\tau,\boldsymbol{\theta}^c)] \\ &\leq \tau\cdot\mathbb{E}_{\tilde{\boldsymbol{c}}^*,\theta\sim P^c}\left[\bar{L}(\tilde{\boldsymbol{c}}^*,\boldsymbol{\lambda}^*,\theta)\right] + \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[\mathrm{Reg}_1(\tau,\boldsymbol{\theta}^c)] \\ &\leq \tau\cdot\mathbb{E}_{\tilde{\boldsymbol{c}}^*,\theta\sim P}\left[\bar{L}(\tilde{\boldsymbol{c}}^*,\boldsymbol{\lambda}^*,\theta)\right] + \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[\mathrm{Reg}_1(\tau,\boldsymbol{\theta}^c)] + O(\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[W(\boldsymbol{\theta})]).\end{aligned} \tag{57}$$

On the other hand, for any $i\in[m]$, we have

$$\sum_{t=1}^{\tau}\bar{L}(\boldsymbol{c}_t,\mu\cdot\boldsymbol{e}_{i_t},\theta_t^c) \geq \sum_{t=1}^{\tau}\bar{L}_i(\boldsymbol{c}_t,\mu\cdot\boldsymbol{e}_{i_t},\theta_t^c) - \mathrm{Reg}_2(\tau,\boldsymbol{\theta}^c)$$

following the regret bound of $\mathsf{ALG}_2$ (holds for arbitrary $\boldsymbol{\lambda} = \mu \cdot \boldsymbol{e}_i$). We now set $i = i'$ and we have

$$\begin{aligned}\sum_{t=1}^{\tau} \bar{L}(\boldsymbol{c}_t, \mu \cdot \boldsymbol{e}_{i_t}, \theta_t^c) &\geq \sum_{t=1}^{\tau} \bar{L}_{i'}(\boldsymbol{c}_t, \mu \cdot \boldsymbol{e}_{i_t}, \theta_t^c) - \text{Reg}_2(\tau, \boldsymbol{\theta}^c) \\ &= \sum_{t=1}^{\tau} \big(p(\boldsymbol{c}_t) + f_{\theta_t^c}(\boldsymbol{x}_t)\big) - \mu \cdot \frac{\tau}{T} + \sum_{t=1}^{\tau} \frac{\mu \cdot g_{i',\theta_t^c}(\boldsymbol{c}_t, \boldsymbol{x}_t)}{T \cdot \beta_{i'}} - \text{Reg}_2(\tau, \boldsymbol{\theta}^c) \\ &\geq \sum_{t=1}^{\tau} \big(p(\boldsymbol{c}_t) + f_{\theta_t^c}(\boldsymbol{x}_t)\big) + \mu \cdot \frac{T-\tau}{T} - \text{Reg}_2(\tau, \boldsymbol{\theta}^c)\end{aligned}$$

where the last inequality follows from (55). Then, it holds that

$$\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\sum_{t=1}^{\tau} \bar{L}(\boldsymbol{c}_t, \mu \cdot \boldsymbol{e}_{i_t}, \theta_t^c)\right] \geq \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\sum_{t=1}^{\tau} \big(p(\boldsymbol{c}_t) + f_{\theta_t^c}(\boldsymbol{x}_t)\big)\right] + \mu \cdot \frac{T-\tau}{T} - \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\text{Reg}_2(\tau, \boldsymbol{\theta}^c)\right]. \tag{58}$$

Combining (57) and (58), we have

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\sum_{t=1}^{\tau} \big(p(\boldsymbol{c}_t) + f_{\theta_t^c}(\boldsymbol{x}_t)\big)\right] \leq &- \mu \cdot \frac{T-\tau}{T} + \tau \cdot \mathbb{E}_{\tilde{\boldsymbol{c}}^*, \theta\sim P}\left[\bar{L}(\tilde{\boldsymbol{c}}^*, \boldsymbol{\lambda}^*, \theta)\right] + O(\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[W(\boldsymbol{\theta})]) \\ &+ \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\text{Reg}_1(\tau, \boldsymbol{\theta})\right] + \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\text{Reg}_2(\tau, \boldsymbol{\theta})\right].\end{aligned}$$

From the boundedness condition c in Assumption 1, we have

$$\mathbb{E}_{\tilde{\boldsymbol{c}}^*, \theta\sim P}\left[\bar{L}(\tilde{\boldsymbol{c}}^*, \boldsymbol{\lambda}^*, \theta)\right] = \frac{1}{T} \cdot \mathsf{OPT} \geq -1$$

which implies that

$$-\mu \cdot \frac{T-\tau}{T} \leq (T-\tau) \cdot \mathbb{E}_{\tilde{\boldsymbol{c}}^*, \theta\sim P}\left[\bar{L}(\tilde{\boldsymbol{c}}^*, \boldsymbol{\lambda}^*, \theta)\right]$$

when $\mu = T$. Therefore, we have

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\sum_{t=1}^{\tau} \big(p(\boldsymbol{c}_t) + f_{\theta_t^c}(\boldsymbol{x}_t)\big)\right] \leq & T \cdot \mathbb{E}_{\tilde{\boldsymbol{c}}^*, \theta\sim P}\left[\bar{L}(\tilde{\boldsymbol{c}}^*, \boldsymbol{\lambda}^*, \theta)\right] + O(\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[W(\boldsymbol{\theta})]) + \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\text{Reg}_1(T, \boldsymbol{\theta})\right] \\ &+ \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\text{Reg}_2(T, \boldsymbol{\theta})\right] \\ \leq & \mathsf{OPT}^c + O(\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[W(\boldsymbol{\theta})]) + \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\text{Reg}_1(T, \boldsymbol{\theta})\right] \\ &+ \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\text{Reg}_2(T, \boldsymbol{\theta})\right].\end{aligned} \tag{59}$$

where $\mathsf{OPT}^c$ denotes the value of the optimal policy with adversarial corruptions. It is clear to see that $|\mathsf{OPT}^c - \mathsf{OPT}| \leq O(\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[W(\boldsymbol{\theta})])$. Using Theorem 3.1 in Auer et al. (2002) to bound $\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\text{Reg}_1(T, \boldsymbol{\theta})\right]$ and Theorem 9 to bound $\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\text{Reg}_2(T, \boldsymbol{\theta})\right]$, we have

$$\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\sum_{t=1}^{\tau} \big(p(\boldsymbol{c}_t) + f_{\theta_t^c}(\boldsymbol{x}_t)\big)\right] \leq \mathsf{OPT}^c + O((G+F) \cdot \sqrt{T}) + O(\sqrt{T \cdot \log m}) + O(\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[W(\boldsymbol{\theta})])$$

which completes our proof of (16). □

*Proof of Theorem 7.* From the formulation (23), the dual variable $\boldsymbol{\lambda}$ is scaled by $T$. Therefore, we know that $\mu = \alpha \cdot T$ for some constant $\alpha > 0$.

We assume without loss of generality that there always exists $i' \in [m]$ such that

$$\sum_{t=1}^{T} g_{i',\theta_t}(\boldsymbol{c}_t, \boldsymbol{x}_t) \leq T \cdot \beta_{i'}. \tag{60}$$

In fact, let there be a *dummy* constraint $i'$ such that $g_{i',\theta}(\boldsymbol{c},\boldsymbol{x})=\beta_{i'}=\alpha$, for arbitrary $\alpha\in(0,1)$, for any $\theta$ and $\boldsymbol{c},\boldsymbol{x}$. Then, (60) holds.

We denote by $(\boldsymbol{\lambda}^*,\tilde{\boldsymbol{c}}^*)$ one *saddle-point* optimal solution to

$$\max_{\boldsymbol{\lambda}\geq 0}\min_{\tilde{\boldsymbol{c}}\in\mathcal{C}}\mathbb{E}_{\tilde{\boldsymbol{c}},\theta\sim P}\left[\bar{L}^{\mathsf{Covering}}(\tilde{\boldsymbol{c}},\boldsymbol{\lambda},\theta)\right]=\min_{\tilde{\boldsymbol{c}}\in\mathcal{C}}\max_{\boldsymbol{\lambda}\geq 0}\mathbb{E}_{\tilde{\boldsymbol{c}},\theta\sim P}\left[\bar{L}^{\mathsf{Covering}}(\tilde{\boldsymbol{c}},\boldsymbol{\lambda},\theta)\right].$$

We have

$$\sum_{t=1}^{T}\bar{L}^{\mathsf{Covering}}(\boldsymbol{c}_t,\mu\cdot\boldsymbol{e}_{i_t},\theta_t)\leq\sum_{t=1}^{T}\mathbb{E}_{\tilde{\boldsymbol{c}}^*}\left[\bar{L}^{\mathsf{Covering}}(\tilde{\boldsymbol{c}}^*,\mu\cdot\boldsymbol{e}_{i_t},\theta_t)\right]+\mathrm{Reg}_1(T,\boldsymbol{\theta})$$

following regret bound of $\mathsf{ALG}_1$. Then, it holds that

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\sum_{t=1}^{T}\bar{L}^{\mathsf{Covering}}(\boldsymbol{c}_t,\mu\cdot\boldsymbol{e}_{i_t},\theta_t)\right]&\leq\mathbb{E}_{\tilde{\boldsymbol{c}}^*,\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\sum_{t=1}^{T}\bar{L}^{\mathsf{Covering}}(\tilde{\boldsymbol{c}}^*,\mu\cdot\boldsymbol{e}_{i_t},\theta_t)\right]+\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\mathrm{Reg}_1(T,\boldsymbol{\theta})\right]\\&\leq\tau\cdot\mathbb{E}_{\tilde{\boldsymbol{c}}^*,\theta\sim P}\left[\bar{L}^{\mathsf{Covering}}(\tilde{\boldsymbol{c}}^*,\boldsymbol{\lambda}^*,\theta)\right]+\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\mathrm{Reg}_1(T,\boldsymbol{\theta})\right].\end{aligned}\tag{61}$$

On the other hand, for any $i\in[m]$, we have

$$\sum_{t=1}^{T}\bar{L}^{\mathsf{Covering}}(\boldsymbol{c}_t,\mu\cdot\boldsymbol{e}_{i_t},\theta_t)\geq\sum_{t=1}^{T}\bar{L}_i^{\mathsf{Covering}}(\boldsymbol{c}_t,\mu\cdot\boldsymbol{e}_{i_t},\theta_t)-\mathrm{Reg}_2(T,\boldsymbol{\theta})$$

following the regret bound of $\mathsf{ALG}_2$ in (holds for arbitrary $\boldsymbol{\lambda}=\mu\cdot\boldsymbol{e}_i$). We now denote by

$$i^*=\mathrm{argmax}_{i\in[m]}\{\beta_i-\frac{1}{T}\cdot\sum_{t=1}^{T}g_{i,\theta_t}(\boldsymbol{c}_t,\boldsymbol{x}_t)\}.$$

We also denote by

$$d_T(\mathcal{A},\boldsymbol{\theta})=\max_{i\in[m]}\{\beta_i-\frac{1}{T}\cdot\sum_{t=1}^{T}g_{i,\theta_t}(\boldsymbol{c}_t,\boldsymbol{x}_t)\}$$

as the distance away from the target set $\mathcal{A}$. Following (60), we always have $d_T(\mathcal{A},\boldsymbol{\theta})\geq 0$. We now set $i=i^*$ and we have

$$\begin{aligned}\sum_{t=1}^{T}\bar{L}^{\mathsf{Covering}}(\boldsymbol{c}_t,\mu\cdot\boldsymbol{e}_{i_t},\theta_t)&\geq\sum_{t=1}^{T}\bar{L}_{i^*}(\boldsymbol{c}_t,\mu\cdot\boldsymbol{e}_{i_t},\theta_t)-\mathrm{Reg}_2(T,\boldsymbol{\theta})\\&=\sum_{t=1}^{T}(p(\boldsymbol{c}_t)+f_{\theta_t}(\boldsymbol{x}_t))+\mu-\sum_{t=1}^{T}\frac{\mu\cdot g_{i^*,\theta_t}(\boldsymbol{c}_t,\boldsymbol{x}_t)}{T\cdot\beta_{i^*}}-\mathrm{Reg}_2(T,\boldsymbol{\theta})\\&\geq\sum_{t=1}^{T}(p(\boldsymbol{c}_t)+f_{\theta_t}(\boldsymbol{x}_t))+\frac{\mu}{\beta_{i^*}}\cdot d_T(\mathcal{A},\boldsymbol{\theta})-\mathrm{Reg}_2(T,\boldsymbol{\theta})\end{aligned}$$

where the last inequality follows from (60). Then, it holds that

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\sum_{t=1}^{T}\bar{L}^{\mathsf{Covering}}(\boldsymbol{c}_t,\mu\cdot\boldsymbol{e}_{i_t},\theta_t)\right]\geq&\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\sum_{t=1}^{T}(p(\boldsymbol{c}_t)+f_{\theta_t}(\boldsymbol{x}_t))\right]\\&+\frac{\mu}{\beta_{\max}}\cdot\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[d_T(\mathcal{A},\boldsymbol{\theta})]-\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\mathrm{Reg}_2(T,\boldsymbol{\theta})\right].\end{aligned}\tag{62}$$

where $\beta_{\max}=\max_{i\in[m]}\{\beta_i\}$. Combining (61) and (62), we have

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\sum_{t=1}^{T}(p(\boldsymbol{c}_t)+f_{\theta_t}(\boldsymbol{x}_t))\right]+\frac{\mu}{\beta_{\max}}\cdot\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[d_T(\mathcal{A},\boldsymbol{\theta})]\leq&T\cdot\mathbb{E}_{\theta\sim P}\left[\bar{L}^{\mathsf{Covering}}(\boldsymbol{c}^*,\boldsymbol{\lambda}^*,\theta)\right]\\&+\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\mathrm{Reg}_1(T,\boldsymbol{\theta})\right]+\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\mathrm{Reg}_2(T,\boldsymbol{\theta})\right].\end{aligned}\tag{63}$$

From the non-negativity of $d_T(\mathcal{A},\boldsymbol{\theta})$, we have

$$\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\sum_{t=1}^{T}(p(\boldsymbol{c}_t)+f_{\theta_t}(\boldsymbol{x}_t))\right]\leq T\cdot\mathbb{E}_{\theta\sim P}\left[\bar{L}^{\mathsf{Covering}}(\boldsymbol{c}^*,\boldsymbol{\lambda}^*,\theta)\right]+\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\mathrm{Reg}_1(T,\boldsymbol{\theta})\right]+\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\mathrm{Reg}_2(T,\boldsymbol{\theta})\right].$$

Using Theorem 8 and Theorem 9 to bound $\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\mathrm{Reg}_1(T,\boldsymbol{\theta})\right]$ and $\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\mathrm{Reg}_2(T,\boldsymbol{\theta})\right]$, we have

$$\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\sum_{t=1}^{T}(p(\boldsymbol{c}_t)+f_{\theta_t}(\boldsymbol{x}_t))\right]\leq T\cdot\mathbb{E}_{\theta\sim P}\left[\bar{L}^{\mathsf{Covering}}(\boldsymbol{c}^*,\boldsymbol{\lambda}^*,\theta)\right]+O((G+F)\cdot\sqrt{T})+O(\sqrt{T\cdot\log m})$$

which completes our proof of (25).

We now prove (26). We note that $(\boldsymbol{c}_t,\boldsymbol{x}_t)_{t=1}^{T}$ defines a feasible solution to

$$\begin{aligned}\mathsf{OPT}^{\delta}=\min&\sum_{t=1}^{T}\mathbb{E}_{\boldsymbol{c}_t,\boldsymbol{x}_t,\theta_t}\left[p(\boldsymbol{c}_t)+f_{\theta_t}(\boldsymbol{x}_t)\right]\\ \text{s.t.}&\frac{1}{T}\cdot\sum_{t=1}^{T}\mathbb{E}_{\boldsymbol{c}_t,\boldsymbol{x}_t,\theta_t}[\boldsymbol{g}_{\theta_t}(\boldsymbol{c}_t,\boldsymbol{x}_t)]\geq\boldsymbol{\beta}-\delta\\ &\boldsymbol{x}_t\in\mathcal{K}(\theta_t,\boldsymbol{c}_t),\boldsymbol{c}_t\in\mathcal{C},\forall t.\end{aligned}\tag{64}$$

with $\delta=\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[d_T(\mathcal{A},\boldsymbol{\theta})]$. If we regard $\mathsf{OPT}^{\delta}$ as a function over $\delta$, then $\mathsf{OPT}^{\delta}$ is clearly a convex function over $\delta$, where the proof follows the same spirit as the proof of Lemma 3. Moreover, note that

$$\frac{d\mathsf{OPT}^{\delta=0}}{d\delta}=\|\boldsymbol{\lambda}^*\|_1\leq\mu.$$

We have

$$\mathsf{OPT}^{\mathsf{Covering}}=\mathsf{OPT}^{0}\leq\mathsf{OPT}^{\delta}+\mu\cdot\delta$$

for any $\delta\geq 0$. Therefore, it holds that

$$\begin{aligned}T\cdot\mathbb{E}_{\theta\sim P}\left[\bar{L}^{\mathsf{Covering}}(\boldsymbol{c}^*,\boldsymbol{\lambda}^*,\theta)\right]=\mathsf{OPT}^{\mathsf{Covering}}&\leq\mathsf{OPT}^{\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[d_T(\mathcal{A},\boldsymbol{\theta})]}+\mu\cdot\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[d_T(\mathcal{A},\boldsymbol{\theta})]\\ &\leq\sum_{t=1}^{T}\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[p(\boldsymbol{c}_t)+f_{\theta_t}(\boldsymbol{x}_t)+\mu\cdot d_T(\mathcal{A},\boldsymbol{\theta})\right]\end{aligned}\tag{65}$$

Plugging (65) into (63), we have

$$\mu\cdot\left(\frac{1}{\beta_{\max}}-1\right)\cdot\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[d_T(\mathcal{A},\boldsymbol{\theta})]\leq\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\mathrm{Reg}_1(T,\boldsymbol{\theta})\right]+\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\mathrm{Reg}_2(T,\boldsymbol{\theta})\right].$$

Our proof of (26) is completed by using Theorem 8 and Theorem 9 to bound $\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\mathrm{Reg}_1(T,\boldsymbol{\theta})\right]$ and $\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\mathrm{Reg}_2(T,\boldsymbol{\theta})\right]$. □