Constrained Online Two-stage Stochastic Optimization: New Algorithms via Adversarial Learning

Jiashuo Jiang

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 propose a general algorithmic framework that derives 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 our framework, we obtain new results under various settings. When the model parameter at each period is drawn from identical distributions, we derive *state-of-art* $O(\sqrt{T})$ regret that improves previous bounds under special cases. Our algorithm is also robust to adversarial corruptions of model parameter realizations. When the model parameters are drawn from unknown non-stationary distributions and we are given prior estimates of the distributions, we develop new algorithm from our framework with a regret $O(W_T + \sqrt{T})$, where W_T measures the total inaccuracy of the prior estimates.

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 θ 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 θ. 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 θ, 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 θ_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 θ_t, we also need to satisfy a long-term global constraint of the following form: $\frac{1}{T} \cdot \sum_{t=1}^{T} \boldsymbol{x}_t \in \mathcal{B}(\boldsymbol{C}, \boldsymbol{\theta})$, where $\mathcal{B}(\boldsymbol{C}, \boldsymbol{\theta})$ is a set that depends on the entire sequence $\boldsymbol{\theta} = (\theta_t)_{t=1}^{T}$ and $\boldsymbol{C} = (\boldsymbol{c}_t)_{t=1}^{T}$. 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.

Motivating example. The online problem is motivated from an example in supply chain and 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.

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, bandits-with-knapsacks (BwK) problem, and the more general online convex optimization (OCO) problem. 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. We review the literature on these problems in Section 1.2. For our model, if we assume the second-stage decision step is de-activated, i.e., at each period t, the second-stage decision $\boldsymbol{x}_t$ is fixed as long as the first-stage decision $\boldsymbol{c}_t$ is determined and the model uncertainty θ_t is realized, then, our model reduces to the bandits-based model described above. On the other hand, if the first-stage decision step is de-activated, i.e., $\mathcal{C}$ is a set of a fixed point, then, our model reduces to the type-based model described above.

From a different perspective, our online problem is also motivated from the computational challenge faced by existing approaches for the classical two-stage stochastic optimization problem. By examining the literature, most approaches for two-stage stochastic optimization, i.e., sample average approximation method (Shapiro and Homem-de Mello 2000) and stochastic approximation method (Nemirovski et al. 2009), would require to solve the second-stage optimization problem for one or multiple times at every iteration, where the constraint $\boldsymbol{x}_t \in \mathcal{B}(\boldsymbol{c}_t, \theta_t)$ is involved. However, when the set $\mathcal{B}(\boldsymbol{c}, \theta)$ possesses complex structures, the second-stage optimization step can be computationally burdensome. Therefore, alternatively, instead of requiring $\boldsymbol{x}_t \in \mathcal{B}(\boldsymbol{c}_t, \theta_t)$ for each $t \in [T]$ in our online problem, we require only the long-run average constraint $\frac{1}{T} \cdot \sum_{t=1}^{T} \boldsymbol{x}_t \in \mathcal{B}(\boldsymbol{C}, \boldsymbol{\theta})$ is satisfied. In other words, we allow the decision maker to violate the constraint set $\mathcal{B}(\boldsymbol{c}_t, \theta_t)$ for some rounds, but the entire sequence of decisions must fit the constraints at the very end. In this way, we are trading the regret for the online problem for computational efficiency of the classical two-stage problem, which is analogous to trading regret for efficiency for OCO problem (Mahdavi et al. 2012).

1.1. Our Approach and Results

Our main approach relies on regarding one long-term constraint as an *expert*, and we develop a general framework to reduce the problem of satisfying the long-term constraints as a procedure for finding the "best" expert. To be specific, at each period t, we assign a probability $\alpha_{i,t}$ to each long-term constraint i that describes the set $\mathcal{B}(\boldsymbol{C}, \boldsymbol{\theta})$, where the summation of $\alpha_{i,t}$ over i equals 1 for each t. We show that if we are applying an adversarial learning algorithm to tackle the expert problem induced by satisfying the long-term constraints, which implies the rule to update $\alpha_{i,t}$, then the gap of constraints violation can be bounded by the regret bound of the adversarial learning algorithm. On the other hand, if we let $\lambda_{i,t} = \mu \cdot \alpha_{i,t}$ where μ is some scaling factor, then we can regard $\lambda_{i,t}$ as the *Lagrangian dual variable* for the long-term constraint i and we can apply another adversarial learning algorithm to minimize the accumulated Lagrangian value over the entire horizon, where the decision variable now is the first-stage decision. After both the first-stage decision and the Lagrangian dual variable are determined for the current period, we show that it is sufficient to decide the second-stage decision by solving a simple optimization problem. Meanwhile, the solved second-stage decision also serves as *stochastic outcome* as input to the two adversarial learning algorithms, which help us update $\alpha_{i,t}$ and the first-stage decision for the next period. This framework is general and turns out to be new for the two-stage stochastic optimization, which enables us to obtain several new results under various settings that we now describe.

We consider the stationary setting as a starter, where the distribution P_t is identical for each t. By applying the above framework, we derive an online algorithm, which we name *Doubly Adversarial Learning* (DAL) algorithm, that satisfies the long-term constraints and achieve a regret bound of $O((G+F)\cdot\sqrt{T})+O(\sqrt{T\cdot\log m})$, where G denotes the diameter of the feasible set for the first-stage actions, F is a constant depending on the gradients of the objective function and the constraint functions that define the set $\mathcal{B}$, and m denotes the number of constraints that characterize $\mathcal{B}$. This the *first* sublinear regret bound for the constrained online two-stage stochastic optimization. Our results improve previously best results under special cases, which reveals the benefits of our adversarial learning framework. For example, for the online allocation special case where the first-stage action is de-activated, our regret bound reduces to $O(\sqrt{T\cdot\log m})$, which improves the $O(\sqrt{T\cdot m})$ regret bound established in Balseiro et al. (2022) in terms of dependency on m, where the Lagrangian dual variables for the long-term constraints are updated in Balseiro et al. (2022) by online mirror descent algorithm, instead of expert algorithms.

Another benefits of our framework that derives online algorithm from adversarial learning algorithms is that our DAL algorithm is robust to adversarial corruptions of the model parameters, i.e., the model parameter θ_t is corrupted to θ_t^c by an adversary, at each period. The adversarial corruption to a stochastic model can arise from the non-stationarity of the underlying distributions $\boldsymbol{P}$ (Jiang et al. 2020), or malicious attack and false information input to the system (Lykouris et al. 2018). And the existence of the adversarial corruptions would make our problem deviate from the stationary setting and become a non-stationary setting where P_t can be non-homogeneous for each t. We show that the same online algorithm induced by our framework can tolerate a certain amount of corruptions and still achieve sublinear regret bound. We first show that if the total number of corruptions is W, then no online algorithm can achieve a regret bound better than $\Omega(W)$. We then show that our DAL algorithm achieves the regret bound $O(W+\sqrt{T})$ in the presence of adversarial corruptions, which matches the lower bound $\Omega(W)$. The linear dependency on W of our DAL algorithm corresponds to the linear dependency established in Gupta et al. (2019) for the MAB problem.

Finally, we explore whether we can improve the regret bound $O(W+\sqrt{T})$ when P_t is non-homogeneous for t, with the help of possible additional information. In practice, historical data are usually available for the distribution P_t. Therefore, we consider a prior estimate setting where we have an estimate $\hat{P}_t$ for P_t at each time period. In this case, we show that W in the lower bound

$\Omega(W)$ can now be replaced by the total inaccuracy of the prior estimates, i.e., the accumulated gap between $\hat{P}_t$ and P_t for all t. In terms of the online algorithm, note that since we have prior estimates, there is no need to apply adversarial learning algorithm to decide the first-stage action. Instead, we can simply solve a two-stage stochastic optimization problem at each period to decide both the first-stage decision and the second-stage decision. The two-stage problem is formulated differently for different periods to handle the non-stationary of the underlying distributions. The formulation relies on the prior estimates, and as a result, our algorithm, which we name *Informative Adversarial Learning* (IAL) algorithm, naturally combines the information provided by the prior estimates 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 $\Omega(W)$.

1.2. Other Related Literature

Our model synthesizes the bandits-based model, where the representative problems include MAB problem, 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. Our model is also motivated from the capacity allocation problem in practice. 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 (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 2020), resource allocation problem (Asadpour 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. (2022)) has also been considered in the literature, where the objective and the constraint functions are allowed to be general functions.

Finally, online optimization techniques have been used in the literature to solve the capacity allocation problem that arises frequently from the application contexts such as supply chains. For example, Zhong et al. (2018) considers a capacity allocation problem with a resource pooling special structure and individual service level constraints. Their policy is motivated by Blackwell's Approachability theorem which is well-known in the online optimization literature. This approach is further generalized in Lyu et al. (2019) for a more general problem structure. Finally, Jiang et al. (2022) adopts an alternative semi-infinite linear programming formulation and derives the optimal policy from applying a stochastic gradient descent algorithm to the dual problem. However, all the existing approaches would assume the underlying distribution to be known and stationary in the corresponding infinite horizon model. In contrast, our algorithm works for the unknown distribution setting, and our algorithm even generalizes to the non-stationary setting, which can be directly applied to the capacity allocation problem with unknown and non-stationary distributions.

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 θ_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 θ_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),

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

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 *optimal policy*, denoted by π^*, 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 π 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 π 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.

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

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 θ_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 θ.*

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 θ, 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:

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

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

3. Online Algorithm for Stationary Setting

In this section, we consider the stationary setting where $P_t=P$ for each $t\in[T]$. Before formally presenting our algorithm and the corresponding regret analysis, we provide intuitions on the algorithmic design, based on relating the minimax dual problem (Dual) to the *zero-sum game* that is described as follows.

Minimizing regret in repeated zero-sum games. Our algorithm is built from the *zero-sum game*, which is a game between two players $j\in\{1,2\}$ with action sets A_1 and A_2 and payoff matrix

$H \in \mathbb{R}^{|A_1|\times|A_2|}$. At each period, player 1 choose an action $a_1 \in A_1$ and player 2 choose an action $a_2 \in A_2$ and the outcome is H_{a_1,a_2}. Player 1 receives H_{a_1,a_2} as *reward* while player 2 receives H_{a_1,a_2} as *cost*. There is an algorithm ALG_1 for player 1 to maximize the reward over the entire horizon and there is an algorithm ALG_2 for player 2 to minimize the total cost. The game is *stochastic* if the outcome H_{a_1,a_2} is drawn independently from a given distribution that depends on a_1 and a_2.

The minimax Lagrangian dual problem (Dual) defines a zero-sum game. To see this point, we first note that under the stationary setting, it is optimal to set an identical first-stage decision, i.e. $\boldsymbol{c}_t^* = \boldsymbol{c}^*$ for some $\boldsymbol{c}^*$. Then, we define

$$\begin{aligned}\bar{L}(\boldsymbol{c},\boldsymbol{\lambda},\theta) =& p(\boldsymbol{c}) - \frac{1}{T}\cdot\sum_{i=1}^{m}\lambda_i + \min_{\boldsymbol{x}\in\mathcal{K}(\theta,\boldsymbol{c})}\Bigg\{f_\theta(\boldsymbol{x})\\ &+\sum_{i=1}^{m}\frac{\lambda_i\cdot g_{i,\theta}(\boldsymbol{c},\boldsymbol{x})}{T\cdot\beta_i}\Bigg\},\end{aligned} \tag{4}$$

as the single-period decomposition of $\mathsf{L}(\boldsymbol{c},\boldsymbol{\lambda})$ with realization θ.

Lemma 2. *Under the stationary setting where $P_t = P$ for each $t \in [T]$, it holds that*

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

The zero-sum game can be described as follows. At each period t, player 1 chooses the first-stage decision $\boldsymbol{c}_t \in \mathcal{C}$, and play 2 chooses the dual variable $\boldsymbol{\lambda}_t \geq 0$. We regard $\boldsymbol{\lambda}_t$ as a distribution over the long-term constraints, after divided by a scaling factor μ. Then, the range of the dual variable for all the minimax problems (20) can be restricted to the set $\mu\cdot\Delta_m$ where $\Delta_m = \{\boldsymbol{y}\in\mathbb{R}^m_{\geq 0} : \sum_{i=1}^m y_i = 1\}$ denotes a distribution over the long-term constraints and $\mu > 0$ is a constant to be specified later. The player 2 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. For player 1, we can observe the stochastic outcome $\bar{L}(\boldsymbol{c}_t,\mu\cdot\boldsymbol{e}_{i_t},\theta_t)$, where θ_t is one sample drawn independently from the distribution P. For player 2, we can observe the stochastic outcome $\bar{L}(\boldsymbol{c}_t,\mu\cdot\boldsymbol{e}_{i_t},\theta_t)$ for the currently chosen constraint i_t. Finally, the second-stage decision $\boldsymbol{x}_t$ is determined by solving the inner minimization problem in (4) for $\boldsymbol{c} = \boldsymbol{c}_t$ and $\boldsymbol{\lambda} = \mu\cdot\boldsymbol{e}_{i_t}$. Here, $\boldsymbol{c}_t$ is determined by ALG_1 for player 1 and i_t is determined by ALG_2 for player 2.

We now further specify what is the observed outcome for player 2. Note that player 2 choose a constraint i_t as action. The corresponding outcome is $\bar{L}_{i_t}(\boldsymbol{c}_t,\mu\cdot\boldsymbol{e}_{i_t},\theta_t)$, where $\bar{L}_i(\boldsymbol{c}_t,\mu\cdot\boldsymbol{e}_{i_t},\theta_t)$ is defined as follows for each $i\in[m]$,

$$\bar{L}_i(\boldsymbol{c}_t,\mu\cdot\boldsymbol{e}_{i_t},\theta_t) = p(\boldsymbol{c}_t) - \frac{\mu}{T} + f_{\theta_t}(\boldsymbol{x}_t) + \frac{\mu\cdot g_{i,\theta_t}(\boldsymbol{c}_t,\boldsymbol{x}_t)}{T\cdot\beta_i}, \tag{5}$$

Algorithm 1 Doubly Adversarial Learning (DAL) algorithm

```
Input: the scaling factor $\mu > 0$, the adversarial learning algorithm $\mathsf{ALG}_1$ for player 1, the adversarial learning algorithm $\mathsf{ALG}_2$ for player 2.
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 $\theta_t$ and determine $\boldsymbol{x}_t$ by solving the inner problem in (4) with $\boldsymbol{c} = \boldsymbol{c}_t$ and $\boldsymbol{\lambda} = \mu \cdot \boldsymbol{e}_{i_t}$.
    3. Return $\bar{L}(\boldsymbol{c}_t, \mu \cdot \boldsymbol{e}_{i_t}, \theta_t)$ to $\mathsf{ALG}_1$.
    4. Return $\bar{L}_i(\boldsymbol{c}_t, \mu \cdot \boldsymbol{e}_{i_t}, \theta_t)$ defined in (5) 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
```

where $\boldsymbol{x}_t$ denotes the second-stage decision we made at period t. Though the action for player 2 is $\boldsymbol{e}_{i_t}$, we are actually able to obtain *additional information* for player 2, which helps the convergence of ALG_2. 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 ALG_2, we are having the *full feedback*, where the outcomes of all constraints, $\bar{L}_i$ for all $i \in [m]$, can be observed. It also worth noting that under Assumption 1, $\bar{L}(\boldsymbol{c}, \boldsymbol{\lambda}, \theta)$ is a convex functions over $\boldsymbol{c}$, for any fixed $\boldsymbol{\lambda}$ and θ.

Lemma 3. *For any $\boldsymbol{\lambda}$ and any θ, $\bar{L}(\boldsymbol{c}, \boldsymbol{\lambda}, \theta)$ defined in (4) is a convex function over $\boldsymbol{c}$, under Assumption 1.*

Following the above discussion, in practice, we can select *Online Gradient Descent* (OGD) algorithm (Zinkevich 2003) as ALG_1 for player 1, and we can select *Hedge* algorithm (Freund and Schapire 1997) as ALG_2 for player 2, where we regard m long-term constraints as m experts. Both algorithms work for adversarial input, which corresponds to the *non-homogeneity* of the input $\boldsymbol{\lambda}_t$ of ALG_1 for player 1 and the input $\boldsymbol{c}_t$ of ALG_2 for player 2. Therefore, we name our algorithm *Doubly Adversarial Learning* (DAL) algorithm. The formal algorithm is given in Algorithm 1.

We denote by $\mathrm{Reg}_1(T, \boldsymbol{\theta})$ the regret of ALG_1 for player 1, given the sequence of realizations $\boldsymbol{\theta}$. In our setting, $\mathrm{Reg}_1(T, \boldsymbol{\theta})$ will be the standard regret bound for OGD algorithm, which we formalize in Theorem 6. Similarly, we denote by $\mathrm{Reg}_2(T, \boldsymbol{\theta})$ the regret of ALG_2 for player 2, given the sequence of realizations $\boldsymbol{\theta}$. In our setting, $\mathrm{Reg}_2(T, \boldsymbol{\theta})$ will be the standard regret bound for Hedge algorithm, which we formalize in Theorem 7. The regret of Algorithm 1 can be bounded by $\mathrm{Reg}_1(T, \boldsymbol{\theta})$ and $\mathrm{Reg}_2(T, \boldsymbol{\theta})$, as shown in the following theorem.

THEOREM 1. *Denote by π Algorithm 1 with input $\mu = T$. Then, under Assumption 1, the regret enjoys the upper bound*

$$\textit{Regret}(\pi, T) \leq \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\textit{Reg}_1(T,\boldsymbol{\theta}) + \textit{Reg}_2(T,\boldsymbol{\theta})\right]. \tag{6}$$

Moreover, if OGD is selected as ALG_1 *and Hedge is selected as* ALG_2*, the regret enjoys the upper bound*

$$\textit{Regret}(\pi, T) \leq \tilde{O}((G+F)\sqrt{T}) + \tilde{O}(\sqrt{T\cdot\log m}), \tag{7}$$

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_θ for all θ, and $\max_{i\in[m]}\{\frac{1}{\beta_i}\}$.

Notably, when the first-stage decision $\boldsymbol{c}_t$ is de-activated for each period t, i.e., $\mathcal{C}$ is a singleton, then there is no need to run ALG_1 to update $\boldsymbol{c}_t$ and our problem would reduce to the online allocation problem studied in Balseiro et al. (2022) and Jiang et al. (2020). The regret upper bound Regret(π, T) in Theorem 1 would reduce to $\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[\mathrm{Reg}_2(T,\boldsymbol{\theta})]$ and thus $\tilde{O}(\sqrt{T\cdot\log m})$ if Hedge is selected as ALG_2. In terms of the dependency on m, our regret bound improves upon the $O(\sqrt{mT})$ regret in Balseiro et al. (2022) where a mirror descent algorithm is used to update the dual variable $\boldsymbol{\lambda}_t$ and the $O(m\sqrt{T})$ regret in Jiang et al. (2020) where a stochastic gradient descent algorithm is used to update $\boldsymbol{\lambda}_t$. This improvement reveals the benefit of our Algorithm 1 that builds on a adversarial learning framework.

4. Robustness to Adversarial Corruptions

In this section, we consider our problem with the presence of adversarial corruptions. To be specific, we assume that at each period t, after the type θ_t is realized, there can be an adversary corrupting θ_t into θ_t^c, and only the value of θ_t^c is revealed to us. The adversarial corruption to a stochastic model can arise from the non-stationarity of the underlying distributions $\boldsymbol{P}$ (Jiang et al. 2020), or malicious attack and false information input to the system (Lykouris et al. 2018). We now show that our Algorithm 1 can tolerate a certain amount of adversarial corruptions and the performance guarantees would degrade smoothly as the total number of corruptions increase from 0 to T.

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.,

$$W(\boldsymbol{\theta}) = \sum_{t=1}^{T} \mathbb{1}(\theta_t \neq \theta_t^c). \tag{8}$$

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 π^*.

Theorem 2. *Let $W(\boldsymbol{\theta})$ be the total number of adversarial corruptions on sequence $\boldsymbol{\theta}$, as defined in* (8). *For any feasible online policy π, 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 2 is in correspondence to the lower bound established in Lykouris et al. (2018) for stochastic multi-arm-bandits model. We now show that our Algorithm 1 achieves a regret bound that matches the lower bound in Theorem 2 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. Note that in the implementation of Algorithm 1, θ_t is always replaced by θ_t^c which is the only type information that is revealed to us.

Theorem 3. *Denote by π Algorithm 1 with input $\mu = T$. Then, under Assumption 1 and the corrupted setting, the regret enjoys the upper bound*

$$\begin{aligned}Regret^c(\pi,T) &= \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[\mathsf{ALG}(\pi,\boldsymbol{\theta}^c)] - \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[\mathsf{ALG}(\pi^*,\boldsymbol{\theta}^c)]\\ &\leq \tilde{O}((G+F)\cdot\sqrt{T}) + \tilde{O}(\sqrt{T\cdot\log m})\\ &\quad + O(\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[W(\boldsymbol{\theta})]),\end{aligned}\tag{9}$$

if OGD is selected as ALG_1 and Hedge is selected as 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 (8).

We remark that the implementation of our Algorithm 1 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 1, 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 2 shows that still, no online policy can achieve a better dependency on the number of corruptions.

5. Improvement of Adversarial Setting with Prior Estimates

In this section, we consider the adversarial setting where P_t is *unknown* and *non-homogeneous* for each t. The adversarial setting can be captured by the setting in Section 4 if we allow an *arbitrary* number of corruptions to stationary distributions. However, following Theorem 2, we would inevitably have a linear dependency on the number of corruptions, which translates into a linear regret for the adversarial setting. Therefore, we now explore whether we can get an improved bound, with the help of *additional information*. We assume that we have a prior estimate of P_t, denoted by $\hat{P}_t$, for each $t \in [T]$. In practice, instead of letting P_t be completely unknown, we usually have multiple historical samples of P_t from which we can form an estimation of P_t (e.g. the empirical distribution over the samples). We explore how to utilize the prior estimate $\hat{P}_t$ for each $t \in [T]$ to obtain an improved bound.

We first derive the regret lower bound and show how the regret should depend on the possible inaccuracy of the prior estimates. 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

$$W(\hat{P}_t, P_t) := \inf_{Q \in \mathcal{F}(\hat{P}_t, P_t)} \int d(\theta, \theta') dQ(\theta, \theta'), \tag{10}$$

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 (θ, θ') 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 prior estimates, by modifying the proof of Theorem 2.

Theorem 4. *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 (10). For any feasible online policy π that only knows prior estimates, 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 π^* 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. Note that in the settings without prior estimates (Section 3 and Section 4), we employ adversarial learning algorithm for $\boldsymbol{c}_t$ to "learn" a good first-stage decision. However, when we have a prior estimate, a more efficient way would be to "greedily" select $\boldsymbol{c}_t$ with the help of prior estimates. In order to describe our idea, we denote by $\{\hat{\boldsymbol{\lambda}}^*, \{\hat{\boldsymbol{c}}_t^*\}_{\forall t \in [T]}\}$ one optimal solution to

$$\begin{aligned}\max_{\boldsymbol{\lambda} \geq 0} \min_{\boldsymbol{c}_t \in \mathcal{C}} \hat{L}(\boldsymbol{C}, \boldsymbol{\lambda}) = \mathbb{E}_{\boldsymbol{\theta} \sim \hat{\boldsymbol{P}}} \Bigg[\sum_{t=1}^T \Bigg(p(\boldsymbol{c}_t) - \frac{1}{T} \cdot \sum_{i=1}^m \lambda_i \\ + \min_{\boldsymbol{x}_t \in \mathcal{K}(\theta_t, \boldsymbol{c}_t)} f_{\theta_t}(\boldsymbol{x}_t) + \sum_{i=1}^m \frac{\lambda_i \cdot g_{i,\theta_t}(\boldsymbol{c}_t, \boldsymbol{x}_t)}{T \cdot \beta_i} \Bigg) \Bigg],\end{aligned} \tag{11}$$

where $\hat{\boldsymbol{P}} = (\hat{P}_t)_{t=1}^T$. Note that the value of (11) is equivalent to the value of (Dual) if $\hat{\boldsymbol{P}} = \boldsymbol{P}$. Then, we define for each $i \in [m]$,

$$\hat{\beta}_{i,t} := \mathbb{E}_{\theta \sim \hat{P}_t}\left[g_{i,\theta}(\hat{\boldsymbol{c}}_t^*, \hat{\boldsymbol{x}}_t^*(\theta))\right], \tag{12}$$

with

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

Here, $\hat{\boldsymbol{\beta}}_t$ can be interpreted as the *prior-estimates-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

$$\begin{aligned} \hat{L}_t(\boldsymbol{c}, \boldsymbol{\lambda}) := \mathbb{E}_{\theta \sim \hat{P}_t} \Bigg[& p(\boldsymbol{c}) - \sum_{i=1}^m \frac{\lambda_i \cdot \hat{\beta}_{i,t}}{T \cdot \beta_i} \\ & + \min_{\boldsymbol{x} \in \mathcal{K}(\theta, \boldsymbol{c})} f_\theta(\boldsymbol{x}) + \sum_{i=1}^m \frac{\lambda_i \cdot g_{i,\theta}(\boldsymbol{c}, \boldsymbol{x})}{T \cdot \beta_i} \Bigg]. \end{aligned} \tag{13}$$

The key ingredient of our analysis is the following.

LEMMA 4. *It holds that*

$$\max_{\boldsymbol{\lambda} \geq 0} \min_{\boldsymbol{c}_t \in \mathcal{C}} \hat{L}(\boldsymbol{C}, \boldsymbol{\lambda}) = \sum_{t=1}^T \max_{\boldsymbol{\lambda} \geq 0} \min_{\boldsymbol{c}_t \in \mathcal{C}} \hat{L}_t(\boldsymbol{c}_t, \boldsymbol{\lambda}), \tag{14}$$

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* (12)*, 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*

$$\sum_{i=1}^m \sum_{t=1}^T \hat{\lambda}_{i,t}^* \hat{\beta}_i = T \cdot \sum_{i=1}^m \beta_i \cdot \hat{\lambda}_i^*, \tag{15}$$

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

Lemma 4 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]. As for $\boldsymbol{\lambda}$, we still apply adversarial learning algorithm (Hedge) to dynamically update it. Our formal algorithm is described in Algorithm 2, which we call *Informative Adversarial Learning* (IAL) algorithm since the updates are informed by the prior estimates.

As shown in the next theorem, the regret of Algorithm 2 is bounded by $O(W_T + \sqrt{T})$, which matches the lower bound established in Theorem 4.

THEOREM 5. *Denote by* π *Algorithm 2 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* (10)*. Then, under Assumption 1, the regret enjoys the upper bound*

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

[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 2 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 (12).
for $t = 1, \ldots, T$ 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 (13) 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 $i \in [m]$,
       $\hat{L}_{i,t}(\boldsymbol{c}_t, \mu \cdot \boldsymbol{e}_{i_t}, \theta_t) = p(\boldsymbol{c}_t) - \frac{\mu \cdot \hat{\beta}_{i,t}}{T \cdot \beta_i} + f_{\theta_t}(\boldsymbol{x}_t) + \frac{\mu \cdot g_{i,\theta_t}(\boldsymbol{c}_t, \boldsymbol{x}_t)}{T \cdot \beta_i}.$    (16)
end for
```

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 $i \in [m]$.

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 5. Therefore, as long as W_T scales sublinearly in T, which guarantees a sublinear regret bound following (17), the constraint violation of our Algorithm 2 also scales as $o(1)$, implying that the solutions generated by Algorithm 2 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).

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 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.

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.

A. Asadpour, X. Wang, and J. Zhang. Online resource allocation with limited flexibility. *Management Science*, 66(2):642–666, 2020.

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

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.

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.

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.

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

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. *arXiv preprint arXiv:2012.07933*, 2020.

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

J. Jiang, S. Wang, and J. Zhang. Achieving high individual service levels without safety stock? optimal rationing policy of pooled resources. *Operations Research*, 2022.

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

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.

G. Lyu, W.-C. Cheung, M. C. Chou, C.-P. Teo, Z. Zheng, and Y. Zhong. Capacity allocation in flexible production networks: Theory and applications. *Management Science*, 65(11):5091–5109, 2019.

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.

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.

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

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

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

A. Shapiro and T. Homem-de Mello. On the rate of convergence of optimal solutions of monte carlo approximations of stochastic programs. *SIAM journal on optimization*, 11(1):70–86, 2000.

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.

Y. Zhong, Z. Zheng, M. C. Chou, and C.-P. Teo. Resource pooling and allocation policies to deliver differentiated service. *Management Science*, 64(4):1555–1573, 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: 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 1 and Algorithm 2, 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 (4). For example, the inner minimization problem is

$$\begin{aligned} \min \quad & f_{\theta_t}(\boldsymbol{x}) + \frac{\mu\cdot g_{i_t,\theta_t}(\boldsymbol{c},\boldsymbol{x})}{T\cdot\beta_{i_t}} \\ \text{s.t.} \quad & B_{\theta_t}\boldsymbol{x} \le \boldsymbol{c}_t \\ & \boldsymbol{x} \ge 0 \end{aligned} \tag{19}$$

and the Lagrangian dual problem of (19) is

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

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 $\|\boldsymbol{\gamma}_t^*\|_2 \le G$ 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_θ for all θ, and $\max_{i\in[m]}\{\frac{1}{\beta_i}\}$. The regret bound of OGD is as follows.

THEOREM 6 **(Theorem 1 of Zinkevich (2003)).** *If $\eta_t = \frac{1}{\sqrt{t}}$, then it holds that*

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

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_θ for all θ, and $\max_{i\in[m]}\{\frac{1}{\beta_i}\}$.

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 $t = 1, \dots, T$ 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
```

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

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

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 α, we know that there exists a constant $\delta > 0$ such that $|l_{i,t}| \le \delta$, for all $i \in [m]$ and $t \in [T]$. Here, δ depends on $\max_{i\in[m]}\{\frac{1}{\beta_i}\}$.

Theorem 7 **(from Theorem 2 in Freund and Schapire (1997))**. *If* $\varepsilon = \sqrt{\frac{\log m}{T}}$, *then it holds that*

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

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

Appendix B: Missing Proofs for Section 3

Proof of Lemma 2. We aim to prove that

$$\max_{\boldsymbol{\lambda}\ge 0} \min_{\tilde{\boldsymbol{c}}_t\in\mathcal{C}} \mathbb{E}_{\tilde{\boldsymbol{C}}}[\mathsf{L}(\boldsymbol{C}, \boldsymbol{\lambda})] = \max_{\boldsymbol{\lambda}\ge 0} \min_{\tilde{\boldsymbol{c}}\in\mathcal{C}} T \cdot \mathbb{E}_{\tilde{\boldsymbol{c}},\theta\sim P}\left[\bar{L}(\tilde{\boldsymbol{c}}, \boldsymbol{\lambda}, \theta)\right]. \tag{20}$$

We denote by $(\boldsymbol{\lambda}^*, \tilde{\boldsymbol{c}}^*)$ one optimal solution of the right hand side (RHS) of (20), where $\tilde{\boldsymbol{c}}$ is allowed to be random. It is clear to see that $(\boldsymbol{\lambda}^*, \{\tilde{\boldsymbol{c}}_t^*\}_{\forall t\in[T]})$ with $\tilde{\boldsymbol{c}}_t^* = \tilde{\boldsymbol{c}}^*$ for each $t \in [T]$ is a feasible solution

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 $t = 1, \dots, T$ 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
```

to the left hand side (LHS) of (20). From the definition of $(\boldsymbol{\lambda}^*, \tilde{\boldsymbol{c}}^*)$, it also holds for each $t \in [T]$ that

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

which implies

$$\boldsymbol{C}^* \in \operatorname{argmin}_{\tilde{\boldsymbol{c}}_t \in \mathcal{C}} \mathbb{E}_{\tilde{\boldsymbol{C}}} [\mathsf{L}(\boldsymbol{C}, \boldsymbol{\lambda}^*)]$$

where $\boldsymbol{C}^* = (\tilde{\boldsymbol{c}}_t^*)_{t=1}^T$. Therefore, it holds that

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

We now prove the reverse direction. Denote by $(\hat{\boldsymbol{\lambda}}, \{\hat{\boldsymbol{c}}_t\}_{t=1}^T)$ one optimal solution to the LHS of (20). Clearly, it holds that for each $t \in [T]$

$$\hat{\boldsymbol{c}}_t \in \operatorname{argmin}_{\tilde{\boldsymbol{c}} \in \mathcal{C}} \mathbb{E}_{\tilde{\boldsymbol{c}}, \theta \sim P} \left[\bar{L}(\tilde{\boldsymbol{c}}, \hat{\boldsymbol{\lambda}}, \theta) \right].$$

Then, it is optimal to let $\hat{\boldsymbol{c}}_t = \hat{\boldsymbol{c}}$ for each $t \in [T]$, for some common $\hat{\boldsymbol{c}} \in \mathcal{C}$ such that

$$\hat{\boldsymbol{c}} \in \operatorname{argmin}_{\tilde{\boldsymbol{c}} \in \mathcal{C}} \mathbb{E}_{\tilde{\boldsymbol{c}}, \theta \sim P} \left[\bar{L}(\tilde{\boldsymbol{c}}, \hat{\boldsymbol{\lambda}}, \theta) \right].$$

Therefore, $(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{c}})$ is feasible to the RHS of (20) and $\hat{\boldsymbol{c}}$ solves the inner minimization problem. We have

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

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 $\alpha_1, \alpha_2 \geq 0$ 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 $\bar{L}(\boldsymbol{c}_1, \boldsymbol{\lambda}, \theta)$ (4). 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 $\bar{L}(\boldsymbol{c}_2, \boldsymbol{\lambda}, \theta)$ (4). 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

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

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 1. Denote by τ the time period that Algorithm 1 is terminated. There must be a constraint $i' \in [m]$ such that

$$\sum_{t=1}^{\tau} g_{i',\theta_t}(\boldsymbol{c}_t, \boldsymbol{x}_t) \geq T \cdot \beta_{i'}. \tag{21}$$

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

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

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

We have for any $\boldsymbol{c}$ that

$$\sum_{t=1}^{\tau} \bar{L}(\boldsymbol{c}_t, \mu \cdot \boldsymbol{e}_{i_t}, \theta_t) \leq \sum_{t=1}^{\tau} \bar{L}^{\mathrm{ISP}}(\boldsymbol{c}, \mu \cdot \boldsymbol{e}_{i_t}, \theta_t) + \mathrm{Reg}_1(\tau, \boldsymbol{\theta})$$

following regret bound of ALG_1, which implies that

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

Then, it holds that

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\sum_{t=1}^{\tau} \bar{L}(\boldsymbol{c}_t, \mu\cdot \boldsymbol{e}_{i_t}, \theta_t)\right] &\leq \mathbb{E}_{\tilde{\boldsymbol{c}}^*,\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\sum_{t=1}^{\tau} \bar{L}(\tilde{\boldsymbol{c}}^*, \mu\cdot \boldsymbol{e}_{i_t}, \theta_t)\right] + \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\text{Reg}_1(\tau, \boldsymbol{\theta})\right] \\ &\leq \tau\cdot \mathbb{E}_{\tilde{\boldsymbol{c}}^*,\theta\sim P}\left[\bar{L}(\tilde{\boldsymbol{c}}^*, \boldsymbol{\lambda}^*, \theta)\right] + \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\text{Reg}_1(\tau, \boldsymbol{\theta})\right].\end{aligned} \tag{22}$$

where the last inequality follows from the definition of the saddle-point $(\boldsymbol{\lambda}^*, \tilde{\boldsymbol{c}}^*)$. On the other hand, for any $i\in[m]$, we have

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

following the regret bound of ALG_2 (holds for arbitrary $\boldsymbol{\lambda}=\mu\cdot\boldsymbol{e}_i$). We now set $i=i'$ and we have

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

where the last inequality follows from (21). Then, it holds that

$$\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\sum_{t=1}^{\tau} \bar{L}(\boldsymbol{c}_t, \mu\cdot \boldsymbol{e}_{i_t}, \theta_t)\right] \geq \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\sum_{t=1}^{\tau}(p(\boldsymbol{c}_t)+f_{\theta_t}(\boldsymbol{x}_t))\right] + \mu\cdot\frac{T-\tau}{T} - \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\text{Reg}_2(\tau, \boldsymbol{\theta})\right]. \tag{23}$$

Combining (22) and (23), we have

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

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

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\sum_{t=1}^{\tau}(p(\boldsymbol{c}_t)+f_{\theta_t}(\boldsymbol{x}_t))\right] &\leq T\cdot\mathbb{E}_{\tilde{\boldsymbol{c}}^*,\theta\sim P}\left[\bar{L}(\boldsymbol{c}^*, \boldsymbol{\lambda}^*, \theta)\right] + \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\text{Reg}_1(\tau, \boldsymbol{\theta})\right] + \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\text{Reg}_2(\tau, \boldsymbol{\theta})\right] \\ &\leq T\cdot\mathbb{E}_{\tilde{\boldsymbol{c}}^*,\theta\sim P}\left[\bar{L}(\tilde{\boldsymbol{c}}^*, \boldsymbol{\lambda}^*, \theta)\right] + \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\text{Reg}_1(T, \boldsymbol{\theta})\right] + \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\text{Reg}_2(T, \boldsymbol{\theta})\right].\end{aligned} \tag{24}$$

which completes our proof of (6). Using Theorem 6 and Theorem 7 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

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

which completes our proof of (7). □

Appendix C: Missing Proofs for Section 4

Proof of Theorem 2. 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 θ, $\mathcal{K}(\theta,\boldsymbol{c})=[0,1]$ (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 θ, 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 (25).

$$\begin{aligned}\min\quad & -x_1-\ldots-x_{\frac{T}{2}}-x_{\frac{T}{2}+1}-\ldots-x_T\\ \text{s.t.}\quad & x_1+\ldots+x_{\frac{T}{2}}+x_{\frac{T}{2}+1}+\ldots+x_T\leq\frac{T}{2}\\ & 0\leq x_t\leq 1\quad\text{for } t=1,\ldots,T.\end{aligned}\tag{25}$$

Now we consider the following two possible adversarial corruptions. The first possible corruption, given in (26), is that the distribution $P_t^c=\theta^2$ for $t=\frac{T}{2}+1,\ldots,T$. The second possible corruption, given in (27), is that the distribution $P_t^c=\theta^3$ for $t=\frac{T}{2}+1,\ldots,T$.

$$\begin{aligned}\min\quad & -x_1-\ldots-x_{\frac{T}{2}}-\left(1+\frac{W_T}{T}\right)x_{\frac{T}{2}+1}-\ldots-\left(1+\frac{W_T}{T}\right)x_T\\ \text{s.t.}\quad & x_1+\ldots+x_{\frac{T}{2}}+x_{\frac{T}{2}+1}+\ldots+x_T\leq\frac{T}{2}\\ & 0\leq x_t\leq 1\quad\text{for } t=1,\ldots,T.\end{aligned}\tag{26}$$

$$\begin{aligned}\min\quad & -x_1-\ldots-x_{\frac{T}{2}}-\left(1-\frac{W_T}{T}\right)x_{\frac{T}{2}+1}-\ldots-\left(1-\frac{W_T}{T}\right)x_T\\ \text{s.t.}\quad & x_1+\ldots+x_{\frac{T}{2}}+x_{\frac{T}{2}+1}+\ldots+x_T\leq\frac{T}{2}\\ & 0\leq x_t\leq 1\quad\text{for } t=1,\ldots,T.\end{aligned}\tag{27}$$

For any online policy π, denote by $x_t^1(\pi)$ the decision of the policy π at period t under corruption scenario given in (26) and denote by $x_t^2(\pi)$ the decision of the policy π at period t under corruption scenario (27). Further define $T_1(\pi)$ (resp. $T_2(\pi)$) as the expected capacity consumption of policy π under corruption scenario (26) (resp. corruption scenario (27)) 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 π on scenario (26) (resp. scenario (27)). It is clear to see that the optimal policy π^* 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 π on scenario (26) and (27) are $\frac{W_T}{T}\cdot T_1(\pi)$ and $W_T - \frac{W_T}{T}\cdot T_2(\pi)$ respectively. Further note that since the implementation of policy π at each time period should be 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 π is the same for both scenarios (26) and (27)). 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 Theorem 3. The proof follows a similar procedure as the proof of Theorem 1. Denote by τ the time period that Algorithm 1 is terminated. There must be a constraint $i'\in[m]$ such that

$$\sum_{t=1}^{\tau} g_{i',\theta_t^c}(\boldsymbol{c}_t,\boldsymbol{x}_t) \geq T\cdot\beta_{i'}. \tag{28}$$

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 θ 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 τ 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 ALG_1. Then, it holds that

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\sum_{t=1}^{\tau}\bar{L}(\boldsymbol{c}_t,\mu\cdot\boldsymbol{e}_{i_t},\theta_t^c)\right] &\leq \mathbb{E}_{\tilde{\boldsymbol{c}}^*,\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\sum_{t=1}^{\tau}\bar{L}(\tilde{\boldsymbol{c}}^*,\mu\cdot\boldsymbol{e}_{i_t},\theta_t^c)\right] + \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\mathrm{Reg}_1(\tau,\boldsymbol{\theta}^c)\right] \\ &\leq \tau\cdot\mathbb{E}_{\tilde{\boldsymbol{c}}^*,\theta\sim P^c}\left[\bar{L}(\tilde{\boldsymbol{c}}^*,\boldsymbol{\lambda}^*,\theta)\right] + \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\mathrm{Reg}_1(\tau,\boldsymbol{\theta}^c)\right] \\ &\leq \tau\cdot\mathbb{E}_{\tilde{\boldsymbol{c}}^*,\theta\sim P}\left[\bar{L}(\tilde{\boldsymbol{c}}^*,\boldsymbol{\lambda}^*,\theta)\right] + \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\mathrm{Reg}_1(\tau,\boldsymbol{\theta}^c)\right] + O(\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[W(\boldsymbol{\theta})]).\end{aligned} \tag{29}$$

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 ALG_2 (holds for arbitrary $\boldsymbol{\lambda} = \mu \cdot \boldsymbol{e}_i$). We now set $i = i'$ and we have

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

where the last inequality follows from (28). Then, it holds that

$$\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\sum_{t=1}^{\tau} \bar{L}(\boldsymbol{c}_t, \mu \cdot \boldsymbol{e}_{i_t}, \theta_t^c)\right] \geq \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\sum_{t=1}^{\tau} \big(p(\boldsymbol{c}_t) + f_{\theta_t^c}(\boldsymbol{x}_t)\big)\right] + \mu \cdot \frac{T-\tau}{T} - \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\mathrm{Reg}_2(\tau, \boldsymbol{\theta}^c)\right]. \tag{30}$$

Combining (29) and (30), we have

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

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

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\sum_{t=1}^{\tau} \big(p(\boldsymbol{c}_t) + f_{\theta_t^c}(\boldsymbol{x}_t)\big)\right] \leq &T \cdot \mathbb{E}_{\tilde{\boldsymbol{c}}^*,\theta\sim P}\left[\bar{L}(\tilde{\boldsymbol{c}}^*, \boldsymbol{\lambda}^*, \theta)\right] + O(\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[W(\boldsymbol{\theta})]) + \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\mathrm{Reg}_1(T, \boldsymbol{\theta})\right] \\ &+ \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\mathrm{Reg}_2(T, \boldsymbol{\theta})\right] \qquad (31) \\ \leq &\mathsf{OPT}^c + O(\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[W(\boldsymbol{\theta})]) + \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\mathrm{Reg}_1(T, \boldsymbol{\theta})\right] \\ &+ \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\mathrm{Reg}_2(T, \boldsymbol{\theta})\right].\end{aligned}$$

where 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 6 and Theorem 7 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

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

which completes our proof of (9). □

Appendix D: Missing Proofs for Section 5

Proof of Theorem 4. The proof is modified from the proof of Theorem 2. We consider a special case of our problem where for any $\boldsymbol{c}\in\mathcal{C}$ and any θ, $\mathcal{K}(\theta,\boldsymbol{c})=[0,1]$ (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 θ, 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 (32).

$$\begin{aligned} \min \quad & -x_1-...-x_{\frac{T}{2}}-x_{\frac{T}{2}+1}-...-x_T \\ \text{s.t.} \quad & x_1+...+x_{\frac{T}{2}}+x_{\frac{T}{2}+1}+...+x_T\leq\frac{T}{2} \\ & 0\leq x_t\leq 1 \quad \text{for } t=1,...,T. \end{aligned} \tag{32}$$

Now we consider the following two possible true distributions. The first possible true scenario, given in (33), is that the distribution $P_t=\theta^1$ for $t=1,\dots,\frac{T}{2}$ and $P_t=\theta^2$ for $t=\frac{T}{2}+1,\dots,T$. The second possible true scenario, given in (34), is that the distribution $P_t=\theta^1$ for $t=1,\dots,\frac{T}{2}$ and $P_t^c=\theta^3$ for $t=\frac{T}{2}+1,\dots,T$.

$$\begin{aligned} \min \quad & -x_1-...-x_{\frac{T}{2}}-\left(1+\frac{W_T}{T}\right)x_{\frac{T}{2}+1}-...-\left(1+\frac{W_T}{T}\right)x_T \\ \text{s.t.} \quad & x_1+...+x_{\frac{T}{2}}+x_{\frac{T}{2}+1}+...+x_T\leq\frac{T}{2} \\ & 0\leq x_t\leq 1 \quad \text{for } t=1,...,T. \end{aligned} \tag{33}$$

$$\begin{aligned} \min \quad & -x_1-...-x_{\frac{T}{2}}-\left(1-\frac{W_T}{T}\right)x_{\frac{T}{2}+1}-...-\left(1-\frac{W_T}{T}\right)x_T \\ \text{s.t.} \quad & x_1+...+x_{\frac{T}{2}}+x_{\frac{T}{2}+1}+...+x_T\leq\frac{T}{2} \\ & 0\leq x_t\leq 1 \quad \text{for } t=1,...,T. \end{aligned} \tag{34}$$

For any online policy π, denote by $x_t^1(\pi)$ the decision of the policy π at period t under the true scenario given in (33) and denote by $x_t^2(\pi)$ the decision of the policy π at period t under the true scenario (34). Further define $T_1(\pi)$ (resp. $T_2(\pi)$) as the expected capacity consumption of policy π under the true scenario (33) (resp. true scenario (34)) 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 π on scenario (33) (resp. scenario (34)). It is clear to see that the optimal policy π^* 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 π on scenario (33) and (34) are $\frac{W_T}{T} \cdot T_1(\pi)$ and $W_T - \frac{W_T}{T} \cdot T_2(\pi)$ respectively. Further note that since the implementation of policy π at each time period should be independent of future realizations, we must have $T_1(\pi) = T_2(\pi)$ (during the first $\frac{T}{2}$ periods, the information for π is the same for both scenarios (33) and (34)). 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 4. 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 (12). 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 (14). 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

$$\nabla L_t(\hat{\boldsymbol{\lambda}}^*) = \nabla \hat{L}_t(\hat{\boldsymbol{c}}_t^*, \hat{\boldsymbol{\lambda}}^*) = \left(-\frac{\hat{\beta}_{i,t}}{T\beta_i} + \mathbb{E}_{\theta \sim \hat{P}_t}\left[\frac{g_{i,\theta}(\hat{\boldsymbol{c}}_t^*, \hat{\boldsymbol{x}}_t^*(\theta))}{T\beta_i} \right] \right)_{\forall i \in [m]} = \boldsymbol{0} \tag{35}$$

The first equality of (35) follows from the fact that

$$\hat{\boldsymbol{c}}_t^* \in \mathrm{argmin}_{\boldsymbol{c} \in \mathcal{C}} \mathbb{E}_{\theta \sim \hat{P}_t}\left[p(\boldsymbol{c}) + \min_{\boldsymbol{x} \in \mathcal{K}(\theta, \boldsymbol{c})} f_\theta(\boldsymbol{x}) + \sum_{i=1}^m \frac{\hat{\lambda}_i^* \cdot g_{i,\theta}(\boldsymbol{c}, \boldsymbol{x})}{T \cdot \beta_i} \right] \tag{36}$$

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 (35) follows from the definition of $\hat{\beta}_{i,t}$ in (12). Therefore, combining (35) and (36), 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 (14). It is sufficient to prove that

$$\hat{L}(\hat{\boldsymbol{C}}^*, \hat{\boldsymbol{\lambda}}^*) = \sum_{t=1}^T \hat{L}_t(\hat{\boldsymbol{c}}_t^*, \hat{\boldsymbol{\lambda}}^*) \tag{37}$$

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 (15) and therefore (14). □

Proof of Theorem 5. The proof can be classified by the following two steps. We denote by

$$\mathsf{L}_t(\boldsymbol{c},\boldsymbol{\lambda}) := \mathbb{E}_{\theta\sim P_t}\left[p(\boldsymbol{c}) - \sum_{i=1}^m \frac{\lambda_i\cdot\hat{\beta}_{i,t}}{T\cdot\beta_i} + \min_{\boldsymbol{x}\in\mathcal{K}(\theta,\boldsymbol{c})} f_\theta(\boldsymbol{x}) + \sum_{i=1}^m \frac{\lambda_i\cdot g_{i,\theta}(\boldsymbol{c},\boldsymbol{x})}{T\cdot\beta_i}\right] \tag{38}$$

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

$$\left|\mathsf{L}_t(\boldsymbol{c},\boldsymbol{\lambda}) - \hat{L}_t(\boldsymbol{c},\boldsymbol{\lambda})\right| \leq \frac{\|\boldsymbol{\lambda}\|_1}{T\beta_{\min}}\cdot W(\hat{P}_t, P_t) \tag{39}$$

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

We now prove (39). 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 θ,θ', 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 (10). 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 (39).

Our second step is to bound the final regret with the help of (39). We assume without loss of generality that there always exists $i'\in[m]$ such that

$$\sum_{t=1}^T g_{i',\theta_t}(\boldsymbol{c}_t,\boldsymbol{x}_t) \geq T\cdot\beta_{i'}. \tag{40}$$

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 θ and $\boldsymbol{c},\boldsymbol{x}$. Then, (40) 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 (12). Then, it holds that

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\sum_{t=1}^T \hat{L}_t(\boldsymbol{c}_t,\mu\cdot\boldsymbol{e}_{i_t},\theta_t)\right] &= \sum_{t=1}^T \mathbb{E}_{\boldsymbol{c}_t,i_t}\left[\mathsf{L}_t(\boldsymbol{c}_t,\mu\cdot\boldsymbol{e}_{i_t})\right] \\ &\leq \sum_{t=1}^T \mathbb{E}_{\boldsymbol{c}_t,i_t}\left[\hat{L}_t(\boldsymbol{c}_t,\mu\cdot\boldsymbol{e}_{i_t})\right] + \frac{\mu\cdot W_T}{T\beta_{\min}} = \sum_{t=1}^T \mathbb{E}_{i_t}\left[\min_{\boldsymbol{c}\in\mathcal{C}}\hat{L}_t(\boldsymbol{c},\mu\cdot\boldsymbol{e}_{i_t})\right] + \frac{\mu\cdot W_T}{T\beta_{\min}} \\ &\leq \sum_{t=1}^T \min_{\boldsymbol{c}\in\mathcal{C}}\hat{L}_t(\boldsymbol{c},\hat{\boldsymbol{\lambda}}^*) + \frac{\mu\cdot W_T}{T\beta_{\min}} = \sum_{t=1}^T \hat{L}_t(\hat{\boldsymbol{c}}_t^*,\hat{\boldsymbol{\lambda}}^*) + \frac{\mu\cdot W_T}{T\beta_{\min}}\end{aligned} \tag{41}$$

where the first inequality follows from the definition of $\boldsymbol{c}_t$, the second inequality follows from Lemma 4, 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 (16), 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 (40), we must have $d_T(\mathcal{A},\boldsymbol{\theta})\geq 0$. We now set $i = i^*$ and we have

$$\begin{aligned}\sum_{t=1}^T \hat{L}_t(\boldsymbol{c}_t,\mu\cdot\boldsymbol{e}_{i_t},\theta_t) &\geq \sum_{t=1}^T \hat{L}_{i^*,t}(\boldsymbol{c}_t,\mu\cdot\boldsymbol{e}_{i_t},\theta_t) - \mathrm{Reg}(T,\boldsymbol{\theta}) \\ &= \sum_{t=1}^T (p(\boldsymbol{c}_t) + f_{\theta_t}(\boldsymbol{x}_t)) - \mu\cdot\frac{\sum_{t=1}^T \hat{\beta}_{i^*,t}}{T\beta_{i^*}} + \sum_{t=1}^T \frac{\mu\cdot g_{i^*,\theta_t}(\boldsymbol{c}_t,\boldsymbol{x}_t)}{T\cdot\beta_{i^*}} - \mathrm{Reg}(T,\boldsymbol{\theta})\end{aligned} \tag{42}$$

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 $i\in[m]$,

$$\nabla_{\lambda_i}\hat{L}(\hat{\boldsymbol{\lambda}}^*) = -1 + \mathbb{E}_{\boldsymbol{\theta}\sim\hat{P}}\left[\sum_{t=1}^T \frac{g_{i,\theta_t}(\hat{\boldsymbol{c}}_t^*,\hat{\boldsymbol{x}}^*(\theta_t))}{T\beta_i}\right] = -1 + \frac{\sum_{t=1}^T \hat{\beta}_{i,t}}{T\cdot\beta_i} \leq 0. \tag{43}$$

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 (43) into (42), we get

$$\sum_{t=1}^{T} \hat{L}_t(\boldsymbol{c}_t, \mu \cdot \boldsymbol{e}_{i_t}, \theta_t) \geq \sum_{t=1}^{T} (p(\boldsymbol{c}_t) + f_{\theta_t}(\boldsymbol{x}_t)) - \mu + \sum_{t=1}^{T} \frac{\mu \cdot g_{i^*,\theta_t}(\boldsymbol{c}_t, \boldsymbol{x}_t)}{T \cdot \beta_{i^*}} - \mathrm{Reg}(T, \boldsymbol{\theta}) \tag{44}$$
$$\geq \sum_{t=1}^{T} (p(\boldsymbol{c}_t) + f_{\theta_t}(\boldsymbol{x}_t)) + \frac{\mu}{\beta_{\min}} \cdot d_T(\mathcal{A}, \boldsymbol{\theta}) - \mathrm{Reg}(T, \boldsymbol{\theta})$$

Combining (41) and (44), we have

$$\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\sum_{t=1}^{T} (p(\boldsymbol{c}_t) + f_{\theta_t}(\boldsymbol{x}_t))\right] \leq \sum_{t=1}^{T} \hat{L}_t(\hat{\boldsymbol{c}}_t^*, \hat{\boldsymbol{\lambda}}^*) + \frac{\mu \cdot W_T}{T\beta_{\min}} + \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[\mathrm{Reg}(T, \boldsymbol{\theta})] - \frac{\mu}{\beta_{\min}} \cdot \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[d_T(\mathcal{A}, \boldsymbol{\theta})]. \tag{45}$$

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

$$\sum_{t=1}^{T} \hat{L}_t(\hat{\boldsymbol{c}}_t^*, \hat{\boldsymbol{\lambda}}^*) \leq \sum_{t=1}^{T} \mathbb{E}_{\tilde{\boldsymbol{c}}_t}[\hat{L}_t(\tilde{\boldsymbol{c}}_t^*, \hat{\boldsymbol{\lambda}}^*)] \leq \sum_{t=1}^{T} \mathbb{E}_{\tilde{\boldsymbol{c}}_t}[\mathsf{L}_t(\tilde{\boldsymbol{c}}_t^*, \hat{\boldsymbol{\lambda}}^*)] + \frac{\|\hat{\boldsymbol{\lambda}}^*\|_1 \cdot W_T}{T\beta_{\min}} = \mathbb{E}_{\tilde{\boldsymbol{C}}}[\mathsf{L}(\tilde{\boldsymbol{C}}^*, \hat{\boldsymbol{\lambda}}^*)] + \frac{\|\hat{\boldsymbol{\lambda}}^*\|_1 \cdot W_T}{T\beta_{\min}}$$
$$\leq \mathbb{E}_{\tilde{\boldsymbol{C}}}[\mathsf{L}(\tilde{\boldsymbol{C}}^*, \boldsymbol{\lambda}^*)] + \frac{\|\hat{\boldsymbol{\lambda}}^*\|_1 \cdot W_T}{T\beta_{\min}} = \mathsf{OPT} + \frac{\|\hat{\boldsymbol{\lambda}}^*\|_1 \cdot W_T}{T\beta_{\min}}, \tag{46}$$

where the first inequality follows from definition of $(\hat{\boldsymbol{\lambda}}^*, (\hat{\boldsymbol{c}}_t)_{t=1}^T)$, the second inequality follows from (39), the first equality follows from (15) and the third inequality follows from the saddle-point condition of $(\boldsymbol{\lambda}^*, \tilde{\boldsymbol{C}}^*)$.

Plugging (46) into (45), we have

$$\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}\left[\sum_{t=1}^{T} (p(\boldsymbol{c}_t) + f_{\theta_t}(\boldsymbol{x}_t))\right] \leq \mathsf{OPT} + \frac{2\mu W_T}{T\beta_{\min}} + \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[\mathrm{Reg}(T, \boldsymbol{\theta})] - \frac{\mu}{\beta_{\min}} \cdot \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[d_T(\mathcal{A}, \boldsymbol{\theta})]. \tag{47}$$

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

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

Using Theorem 7 to bound $\mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{P}}[\mathrm{Reg}(T, \boldsymbol{\theta})]$, we have

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

which completes our proof of (17) 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

$$\begin{aligned} \mathsf{OPT}^{\delta} = \min \sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{c}_t, \boldsymbol{x}_t, \theta_t \sim P_t} \left[p(\boldsymbol{c}_t) + f_{\theta_t}(\boldsymbol{x}_t) \right] \\ \text{s.t.} \frac{1}{T} \cdot \sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{c}_t, \boldsymbol{x}_t, \theta_t \sim P_t} [\boldsymbol{g}_{\theta_t}(\boldsymbol{c}_t, \boldsymbol{x}_t)] \le \boldsymbol{\beta} + \delta \cdot \boldsymbol{e} \\ \boldsymbol{x}_t \in \mathcal{K}(\theta_t, \boldsymbol{c}_t), \boldsymbol{c}_t \in \mathcal{C}, \forall t. \end{aligned} \tag{48}$$

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,

$$\begin{aligned} \hat{\mathsf{OPT}}^{\hat{\delta}} = \min \sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{c}_t, \boldsymbol{x}_t, \theta_t \sim \hat{P}_t} \left[p(\boldsymbol{c}_t) + f_{\theta_t}(\boldsymbol{x}_t) \right] \\ \text{s.t.} \frac{1}{T} \cdot \sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{c}_t, \boldsymbol{x}_t, \theta_t \sim \hat{P}_t} [\boldsymbol{g}_{\theta_t}(\boldsymbol{c}_t, \boldsymbol{x}_t)] \le \boldsymbol{\beta} + \hat{\delta} \cdot \boldsymbol{e} \\ \boldsymbol{x}_t \in \mathcal{K}(\theta_t, \boldsymbol{c}_t), \boldsymbol{c}_t \in \mathcal{C}, \forall t. \end{aligned} \tag{49}$$

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

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

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

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

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

We have

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

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

$$\hat{\mathsf{OPT}} = \max_{\boldsymbol{\lambda} \ge 0} \min_{\boldsymbol{c}_t \in \mathcal{C}} \hat{L}(\boldsymbol{C}, \boldsymbol{\lambda}) = \sum_{t=1}^{T} \hat{L}_t(\hat{\boldsymbol{c}}_t^*, \hat{\boldsymbol{\lambda}}^*) \tag{50}$$

where the last equality follows from Lemma 4. Therefore, we now assume that $\mathbb{E}_{\boldsymbol{\theta} \sim \boldsymbol{P}}[d_T(\mathcal{A}, \boldsymbol{\theta})] \ge \frac{W_T}{T}$, otherwise (18) directly holds. We have

$$\begin{aligned} \sum_{t=1}^{T} \hat{L}_t(\hat{\boldsymbol{c}}_t^*, \hat{\boldsymbol{\lambda}}^*) = \hat{\mathsf{OPT}} &\le \hat{\mathsf{OPT}}^{\mathbb{E}_{\boldsymbol{\theta} \sim \boldsymbol{P}}[d_T(\mathcal{A}, \boldsymbol{\theta})] + \frac{W_T}{T}} + \mu \cdot \left(\mathbb{E}_{\boldsymbol{\theta} \sim \boldsymbol{P}}[d_T(\mathcal{A}, \boldsymbol{\theta})] + \frac{W_T}{T} \right) \\ &\le \mathsf{OPT}^{\mathbb{E}_{\boldsymbol{\theta} \sim \boldsymbol{P}}[d_T(\mathcal{A}, \boldsymbol{\theta})]} + \mu \cdot \mathbb{E}_{\boldsymbol{\theta} \sim \boldsymbol{P}}[d_T(\mathcal{A}, \boldsymbol{\theta})] + W_T \cdot (1 + \frac{\mu}{T}) \\ &\le \sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{\theta} \sim \boldsymbol{P}} \left[p(\boldsymbol{c}_t) + f_{\theta_t}(\boldsymbol{x}_t) \right] + \mu \cdot \mathbb{E}_{\boldsymbol{\theta} \sim \boldsymbol{P}}[d_T(\mathcal{A}, \boldsymbol{\theta})] + W_T \cdot (1 + \frac{\mu}{T}) \end{aligned} \tag{51}$$

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 (51) into (45), we have

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

which implies

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

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 OPT^{δ}. Then, from the definition of W_T, we have that

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

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

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

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